Shared polygenic risk and causal inferences in amyotrophic lateral sclerosis

Sara Bandres-Ciga, PhD[1]*, Alastair J. Noyce, MD, PhD[2,3]*, Gibran Hemani, PhD[4], Aude Nicolas, PhD[5], Andrea Calvo, MD[6], Gabriele Mora, MD[7]; The ITALSGEN Consortium; The International ALS Genomics Consortium; Pentti J. Tienari, MD[8], David J. Stone, PhD[9], Mike A. Nalls, PhD[1,10], Andrew B. Singleton, PhD[1], Adriano Chiò, MD[6,11,12,]*, and Bryan J. Traynor, MD, PhD[5,13,]*

Affiliations

1 Molecular Genetics Section, Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD 20892, USA

2 Preventive Neurology Unit, Wolfson Institute of Preventive Medicine, Queen Mary University of London, London EC1M 6BQ, UK

3 Department of Clinical and Movement Neurosciences, University College London, Institute of Neurology, London WC1N 1PJ, UK

4 MRC Integrative Epidemiology Unit, University of Bristol, Bristol BS8 2BN, UK

5 Neuromuscular Diseases Research Section, Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD 20892, USA

6 'Rita Levi Montalcini' Department of Neuroscience, University of Turin, Turin, Italy

7 ALS Center, Istituti Clinici Scientifici Maugeri, IRCCS, Milan, Italy

8 Department of Neurology, Helsinki University Hospital and Molecular Neurology Programme, Biomedicum, University of Helsinki, Helsinki FIN-02900, Finland

9 Genetics and Pharmacogenomics, Merck Research Laboratories, Merck & Co., Inc., West

Point, PA 19486, USA

10 Data Tecnica International, Glen Echo, MD 20812, USA

11 Institute of Cognitive Sciences and Technologies, C.N.R., Rome, Italy

12 Azienda Ospedaliero Universitaria Città della Salute e della Scienza, Turin, Italy

13 Department of Neurology, Johns Hopkins University, Baltimore, MD 21287, USA

*Equal contribution

Correspondence

Bryan J. Traynor, traynorb@mail.nih.gov, Porter Neuroscience Center, 35 Convent Drive, Room 1A-213, Bethesda, MD 20892, USA

Running head: Pleiotropic and causal risk factors in ALS

Key words: amyotrophic lateral sclerosis, Mendelian randomization, LD score regression, risk factor, public resource.

Manuscript word count: 4,181

Abstract word count: 234

Title count: 10

References: 45

Figures/Tables: 5

Supplementary material -> https://github.com/SaraBandres/ALS

ABSTRACT

Objective: To identify shared polygenic risk and causal associations in amyotrophic lateral sclerosis (ALS).

Methods: Linkage disequilibrium score regression and Mendelian randomization were applied in a large-scale, data-driven manner to explore genetic correlations and causal relationships between > 700 phenotypic traits and ALS. Exposures consisted of publicly available genome-wide association studies (GWASes) summary statistics from *MR Base* and *LD-hub*. The outcome data came from the recently published ALS GWAS involving 20,806 cases and 59,804 controls. Multivariate analyses, genetic risk profiling and Bayesian colocalization analyses were also performed.

Results: We have shown via linkage disequilibrium score regression that ALS shares polygenic risk genetic factors with a number of traits and conditions, including positive correlations with smoking status and moderate levels of physical activity, and negative correlations with higher cognitive performance, higher educational attainment, and light levels of physical activity. Using Mendelian randomization, we found evidence that hyperlipidemia is a causal risk factor for ALS and localized putative functional signals within loci of interest.

Interpretation: Here we have developed a public resource (https://lng-nia.shinyapps.io/mrshiny) may be a valuable tool for the ALS community that will be expanded and updated as new data become available. Shared polygenic risk exists between ALS and educational attainment, physical activity, smoking and tenseness/restlessness. We found compelling evidence that elevated LDL cholesterol is a causal risk factor for ALS. Future randomized controlled trials should be considered as a proof of causality.

INTRODUCTION

Amyotrophic lateral sclerosis (ALS, OMIM #105400) is a progressive, fatal neurodegenerative disease. Symptom onset of ALS peaks in the mid-sixties, and most patients succumb to the disease within two to five years of becoming symptomatic [1]. The prevalence of ALS is projected to nearly double by 2040, primarily due to aging of the global population [2].

Despite considerable advances made in understanding the genetic architecture underlying ALS [3,4], the contribution of lifestyle factors and of disease-related conditions predisposing individuals to the disorder have been more difficult to elucidate. Epidemiological studies have attempted to identify risk factors and comorbidities associated with ALS, although the inability of such observational research to fully mitigate confounding effects or to exclude reverse causality has made it challenging to find replicable causes of the disease [5].

Genome-wide association studies (GWAS) have revolutionized human genetics and have led to the discovery of thousands of risk variants involved in disease etiology [6]. From the perspective of ALS research, summary statistics from hundreds of these studies have been published online in an effort to facilitate the application of current generation genomic techniques, such as linkage disequilibrium (LD) score regression testing and Mendelian randomization. Both methodologies are powerful tools to assess causality and investigate the extent to which genetic etiologies are shared across different diseases.

LD score regression and Mendelian randomization test distinct aspects of the genetic architecture underlying a disease. More specifically, LD score regression investigates whether polygenic risk contributing to a phenotype of interest might also contribute to the risk of ALS. This approach relies on the identification of shared genome-wide heritability to pinpoint overlapping polygenic genetic variation between traits (pleiotropic relationship) [7]. On the other hand, Mendelian randomization uses genetic data to assess whether an exposure exerts a causal effect on a particular outcome [8]. In contrast to LD score regression, Mendelian randomization usually focuses on genome-wide significant SNPs (causal relationship) [8]. As this analytical technique relies solely on genetic elements that remain constant over the lifespan of an individual, and that are randomized during gametogenesis, it effectively excludes reverse causality and reduces confounding to allow more reliable identification of a causal association between exposure and outcome.

Here, we implemented LD score regression and Mendelian randomization in a large-scale audit relevant to ALS. In brief, our goal was to survey curated libraries of GWAS results using LD score regression and Mendelian randomization. The former is more liberal in identifying shared variation that suggests a significant degree of shared genetic risk, whereas the latter is more conservative and attempts to pinpoint causal associations at established loci. We also created an online resource (https://lng-nia.shinyapps.io/mrshiny) that can be used by the ALS community to inform pleiotropy or causality when undertaking observational studies or pursuing disease-modifying interventions.

METHODS

Outcome data

Summary statistics from our recently published GWAS of ALS involving 20,806 cases and 59,804 controls of European ancestry were used as the outcome for both LD score regression and

Mendelian randomization analyses. This study included 10,031,630 genotyped and imputed variants. Sample recruitment and genotyping quality control procedures are described elsewhere [4].

Linkage disequilibrium score regression

LD patterns across the genome enable the calculation of genetic correlations between traits. This is because the observed association for a SNP is a product of both its own contribution toward a phenotype and the association of the SNPs that are in LD with it. SNPs in regions of high LD tag a greater proportion of the genome and will show stronger associations than SNPs in regions of low LD. Using the known LD structure of a reference SNP panel, the heritability of a single phenotype or the genetic correlation of two phenotypes can be computed using LD score regression [7,9].

To study shared genetic risk via LD score regression, we used *LD Hub*, a centralized database of summary-level GWAS results across multiple diseases/traits gathered from publicly available resources [10]. LD score regression was implemented by regression of the chi-squared statistics for the genetic associations with the trait against the LD scores for genetic variants across the whole genome. Unlike Mendelian randomization, LD score regression does not assess casualty, but rather only assesses multi-directional correlations and can distinguish between population stratification and polygenicity in GWAS studies. Default settings were used in our analyses.

Mendelian randomization

Mendelian randomization is a proxy-based approach for exploring whether an exposure is causally associated with an outcome. This is done by: identifying the single nucleotide polymorphisms

(SNPs) associated with a particular exposure (for example, SNPs identified in a genome-wide association study (GWAS) as being associated with colon cancer); extracting data for those SNPs from the outcome (in this case, a large-scale GWAS of ALS [4]); harmonizing the exposure and outcome summary data; and, applying Mendelian randomization methods to test for a causal relationship between the exposure and the outcome rooted in genetic associations.

Similar to *LD-Hub*, *MR Base* database is a curated database containing summary results from 1,094 GWASes involving 889 traits [11] . These traits encompass a wide range of physiological characteristics and disease phenotypes. Each trait was tested separately as an exposure to determine if it alters risk of developing ALS. The analyses were performed using the R package *TwoSampleMR* (version 3.2.2). The instrumental variables used for each exposure/phenotype consisted of the per-allele log-odds ratio (i.e., beta estimate) and standard errors for all independent loci (i.e., SNPs) reaching genome-wide significance in the tested GWAS. Of the 1,094 GWASes with data available in *MR Base* (accessed 15th August 2018), 635 GWASes (consisting of 345 published GWASes and 290 unpublished GWASes performed on the UK Biobank, www.ukbiobank.ac.uk) were included in our analysis based on the following criteria: (i) GWAS with at least two associated SNPs with p-values less than $5.0\text{x}10^{-8}$, considering this p-value to be the generally accepted genome-wide significant threshold; (ii) SNPs present in both the exposure and outcome (ALS) datasets or when not present their linkage-disequilibrium proxies (R^2 value >= 0.8); and, (iii) independent SNPs ($R^2 < 0.001$ with any other associated SNP within 10 Mb), considered as the most stringent clumping threshold used when performing Mendelian randomization analyses.

Harmonization was undertaken to rule out strand mismatches and to ensure alignment of SNP effect sizes. Within each exposure GWAS, Wald ratios were calculated for each extracted SNP by dividing the per-allele log-odds ratio of that variant in the ALS data by the log-odds ratio of the same variant in the exposure data. We then applied a two-step approach designed to decrease the risk of false positive associations (see **Fig 1** for the workflow).

First, the inverse-variance weighted method was implemented to examine the relationship between the exposure and ALS. In this method, the Wald ratio for each SNP is weighted according to the inverse variance, and a line, constrained to pass through the origin, is fitted to the data. The slope of the line represents the pooled-effect estimate of the Wald ratios [12]. Traits were brought forward to the next stage of analysis only if the p-value of the pooled-effect estimate was less than or equal to 0.05. Next, two Mendelian randomization sensitivity tests (i.e., MR Egger and Weighted Median) were applied to those traits/GWASes passing the first phase of analysis. These sensitivity analyses evaluated core assumptions of Mendelian randomization, and traits were considered to be consistent with a causal effect when p-values were less than or equal to 0.05. Heterogeneity of effects were tested using the Cochran's Q test, quantified using the I^2 statistic, and displayed in forest plots. Steiger analyses were performed to verify that the proposed instruments were directly associated with the outcome [13] or effect estimate directionality.

We evaluated the possibility that the overall estimate was driven by a single SNP using leave-one-out analyses for each of the GWASes associated with ALS. We further explored the possibility of reverse causality by using SNPs tagging the five independent loci described in the ALS GWAS as exposure instrument variables, and the identified GWASes as the outcome. Lasso-based multivariate analysis was used to explore how each related exposure of interest (i.e., low-density lipoprotein (LDL cholesterol), self-reported cholesterol, and coronary heart disease) independently contribute to ALS.

Genetic risk score

To further test the relationship between LDL cholesterol and ALS, a cumulative genetic risk score for LDL cholesterol was calculated in a smaller subset of the samples for which individual genotype data were available, including 8,229 ALS cases and 36,329 controls [4]. The instrumental variables of interest were incorporated and weighted by beta values in the ALS GWAS. Next, a logistic regression was performed on this subset of cases and controls, regressing disease against quintile membership based on genetic risk score [14]. Odds ratios were reported comparing the lowest risk quintile (reference group) to the remaining quintiles. Genetic risk scores were also calculated for different subtypes of ALS patients (carriers of the pathogenic *C9orf72* repeat expansion, familial cases, sporadic cases, male cases and female cases). Risk profiling was adjusted for sex, age, and twenty principal components to account for population stratification.

Colocalization analysis

Bayesian colocalization analysis was used as a statistical method to identify putative candidate genetic variants involved in LDL cholesterol blood levels that contribute most to the risk of developing ALS [15]. For these analyses, we considered the 78 SNPs that were significantly associated with increased LDL cholesterol and were used as relevant instrumental variables for Mendelian randomization analyses. We extracted summary statistics for those variants (as well as variants one megabase (Mb) upstream and downstream) from the LDL cholesterol GWAS and from the ALS GWAS. Bayesian colocalization was then run for each independent region as implemented in the R package coloc (https://CRAN.R-project.org/package=coloc). This analysis assessed the probability of each SNP being responsible for the change in ALS risk through variation in LDL cholesterol. We derived posterior probabilities (PPH_{0-4}) for each region and considered PPH_4 greater than 0.95 as strong evidence for colocalization under the assumption of a single causative variant per locus.

RESULTS

Large-scale linkage disequilibrium score regression analysis in ALS

LD score regression was applied to examine the genetic correlation between our recently published GWAS meta-analysis of ALS [4] and 736 phenotypes available in *LD-hub*, a centralized database of GWAS results across multiple diseases and traits (http://ldsc.broadinstitute.org/ldhub/, **Fig 1**).

Traits genetically correlated to ALS by linkage disequilibrium score regression

Our analyses identified eighteen traits that were genetically correlated to ALS after adjusting for multiple testing via false discovery rate (**Table 1**). Among these, nine traits were related to educational attainment and intelligence, indicating that higher levels of education were associated with a decreased risk of ALS (smallest adjusted p-value = 1.78x10^{-4}; regression coefficient = -0.338; 95% confidence interval (CI) = -0.46, -0.20).

Traits related to light physical activity including *walking for pleasure, walking as a mean of transport* and *light DIY physical activities* were associated with decreased risk of developing ALS (smallest adjusted p-value = 5.19x10^{-4}; regression coefficient = -0.403; 95% CI = -0.35, -0.14), whereas heavier activity levels such as *duration of moderate activity* and *performing a job that involves mainly walking or standing* were positively associated to ALS (smallest adjusted p-value = 3.09x10^{-2}; regression coefficient = 0.28; 95% CI = -0.36, -0.09).

Smoking behaviour including *Exposure to tobacco* and *being a light smoker* showed genetic correlation with ALS (smallest adjusted p-value = 1.66x10^{-3}; regression coefficient = 0.42; 95% CI = 0.23, 0.62). Detailed results for the remaining 718 non-significant traits are shown in **Table S1** and can be interactively searched at https://lng-nia.shinyapps.io/mrshiny.

Large-scale Mendelian randomization in ALS

Next, we performed Mendelian randomization to further investigate causal links between multiple phenotypic traits (exposures) and ALS (outcome, **Fig 1**). The exposures of interest consisted of 345 GWASes involving a wide range of physiological characteristics and disease phenotypes for which data was available in *MR Base* (http://www.mrbase.org/). The recently published GWAS of ALS involving 20,806 cases and 59,804 controls was used as the outcome [4].

There are no previous reports in the literature where multiple phenotypes were tested using Mendelian randomization in an unbiased, hypothesis-free manner. This raised concerns about false-positive associations and multiple-testing correction. To control for this and to confirm the validity of our findings, we replicated in an independent collection of phenotypes (290 unpublished GWASes performed on the UK Biobank, www.ukbiobank.ac.uk). Here, we only report associations that were significant across both the published and unpublished sets of GWASes. Detailed results for the 635 GWAS under study are shown in **Tables S2-S4** and can be visually explored at https://lng-nia.shinyapps.io/mrshiny.

Traits causally linked to ALS by Mendelian randomization

We identified the phenotypic traits *LDL cholesterol* and *coronary heart disease* in the published GWASes, and *self-reported high cholesterol* in the UK Biobank, as being causally linked to ALS risk (see **Table 2**, **Fig S1** for forest plots and **Table S5** for SNPs used to construct the instruments of interest. Multivariate analysis showed that the signals arising from the *coronary heart disease* and *self-reported high cholesterol* were driven by SNPs related to LDL cholesterol, revealing that both traits represent closely-related phenotypes (**Table 3**).

Leave-one-out analysis indicated that no single SNP accounted for these associations in isolation (**Fig S2** and **Table S6**). Additional analyses examining directionality, pleiotropy, and reverse causality did not indicate any violation of core Mendelian randomization assumptions for these traits (**Table S7-S8,** and **Fig S3**). We used genetic risk profiling to estimate the extent to which risk of developing ALS is attributable to LDL cholesterol. We found that individuals with the highest burden of genetic risk were 1.075 times more likely to develop ALS (95% CI, 1.001–1.15, p-value = 0.003). The increase in ALS risk associated with LDL cholesterol levels was similar across different subtypes of ALS (*C9orf72* carriers, familial ALS, sporadic ALS, male and female-only cases, **Table S9**).

Identification of functional causal variants

Bayesian colocalization analysis was performed to putatively identify the functional candidate variants that may drive risk of developing ALS through shared pathway effects of LDL cholesterol levels. We focused our efforts on the 78 independent regions associated with LDL cholesterol in the exposure GWAS. This analysis identified two independent regions with greater than 95% probability of containing a shared causal SNP (**Table S10**). Fine mapping of these regions identified two SNPs (rs182826525 within *COL4A3BP* and rs116226146 intergenic between *PPP1R2P3* and *TIMD4*) as being causally linked to ALS through an increase of LDL cholesterol levels (**Fig 2**).

DISCUSSION

We applied cutting-edge analytical techniques to genomic data across a wide range of phenotypic traits to identify factors associated with risk of developing ALS. This hypothesis-free, data-driven approach provided *prima facie* evidence supporting the existence of multiple such factors. Using genetic data to comprehensively map the risk factor landscape of ALS represents a novel approach in neurological disease. We used these results from nearly 25 million individuals (24,538,000 from published GWAS and ~ 337,159 from UKBB studies) to establish a public resource that can be accessed by other researchers to explore risk factors and shared disease mechanisms in ALS.

LD score regression analyses found that common genetic variation associated with higher cognitive performance is negatively correlated to ALS. At a molecular level, these findings indicate that the genetic factors driving mental ability and ALS overlap to some extent. This may have been an expected outcome given the well-known relationship between ALS and frontotemporal dementia, and our findings are consistent with previous epidemiological reports assessing the causal relationship between education and ALS [16]. Nevertheless, the large number of samples analyzed in our study and its grounding in genetics puts this risk factor on a firmer footing within the ALS field. Similar education effects have been observed in Alzheimer's disease [17], but understanding how education protects against neurodegeneration or which genetic variants are responsible for this shared risk will require additional study. One intriguing possibility is that the genetic variants responsible for ALS in middle age or in the elderly are also associated with decreased cognitive performance at a younger age. This is plausibly consistent with the observation that connectivity and grey matter volumes are altered in asymptomatic carriers of the *C9orf72* repeat expansion [18-19]**.**

Epidemiological case-control studies have extensively reported a relationship between exercise and risk of developing ALS [20,21-22], though there are conflicting results as to the level of physical activity required to increase risk [16]. Our genetic-based data demonstrate that this neuromuscular interconnection may be more complex than previously appreciated: light physical activity including '*Walking for pleasure'* or '*Light DYI activities'* was negatively associated with developing ALS, whereas more strenuous activity, such as '*duration of moderate activity*' was paradoxically correlated with ALS. Extrapolating from these observations to neuromuscular physiology, relatively low levels of exercise may exert a neuroprotective effect by preventing muscle atrophy that, in turn, supports motor neuron integrity though the neuroadaptive generation of neurotrophic factors [23]). In contrast, excessive physical activity may be detrimental to motor neurons due to excessive free radical production and/or glutamate excitotoxicity that overwhelms neuroprotective mechanisms [24,25]. Regardless, our data does not provide any insight into the effect of exercise on survival once the patient has presented with symptoms.

There is compelling epidemiological evidence showing that cigarette smoking is a key environmental risk factor for ALS [26]. Our LD data not only confirms that being a smoker is positively correlated to developing ALS, but also shows that this effect is mediated, at least in part, through shared genetic mechanisms. This is an example of the ability of this type of genomic analysis to identify pleiotropic effects, which is where a defect in a single gene can give rise to multiple, apparently unrelated phenotypes. Here we are extending the concept of pleiotropy beyond the single gene paradigm to encompass inherently complex traits driven by multiple genetic variants spread across the genome and that are typically outside of coding regions. LD score regression is not designed to identify the specific shared genetic variants responsible for both phenotypic traits, but instead focuses on establishing whether such pleiotropy exists between traits.

Using Mendelian randomization, we found compelling evidence that an alteration of lipid metabolism is a causal risk factor for ALS. We undertook sensitivity analyses to reduce the possibility of confounding in our results and replicated our findings across three different GWASes, including a large, independent cohort obtained from the UK Biobank. Furthermore, we demonstrated that the increased risk of ALS due to coronary heart disease is driven by LDL cholesterol. Though hyperlipidemia only modestly increases the risk of ALS, this effect likely operates over the lifetime of the individual, and the cumulative effect on disease risk may be substantial.

Previous epidemiological studies have explored the role of blood lipids in the pathogenesis of ALS. These observational studies have yielded controversial results with many reporting that hyperlipidemia increases disease risk, and others suggesting the opposite [27–33]. In addition to being underpowered, much of this previous research was based on blood lipid profiles obtained after diagnosis of ALS when ancillary factors may be influencing these levels [27–33].

The singular advantage of Mendelian randomization is that it is agnostic to these confounders and can be considered nature's randomized controlled trial. Based on genetic data that remain constant between the pre-symptomatic and symptomatic phases of the disease, it accurately pinpoints predisposing factors for the disease of interest. Chen *et al* recently reported a link between blood lipids and the risk of ALS using Mendelian randomization [34]. This study was performed using a lenient SNP clumping threshold involving a smaller cohort of ALS cases, and was based on the *a priori* hypothesis that blood lipids were involved in the pathogenesis of ALS. Our work extends this preliminary report by definitely applying Mendelian randomization across a large number of phenotypic traits in an unbiased fashion, by replicating our findings in an independent cohort (UK Biobank), and by delineating the specific aspects of lipid metabolism relevant to the pathogenesis of ALS.

Circulating blood cholesterol are multifunctional molecules, involved primarily in energy generation, as precursors or cofactors for signaling molecules, and in neuronal development and function [35]. Dysregulation of cholesterol homeostasis in the brain has been linked to many neurodegenerative diseases such as Huntington's disease, Parkinson's disease, Niemann-Pick disease type C, and, most notably, Alzheimer's disease [36]. The generation and clearance of β-amyloid protein is regulated by cholesterol, and drugs that inhibit cholesterol synthesis lower this protein within neurons 37, as is the more recent finding that the two secretory forms (APPα and APP β) of amyloid precursor protein (APP) have opposing associations with β-amyloid generation, cholesterol biosynthesis, and LDL receptor levels 38. The identification of the cholesterol transport protein apolipoprotein E as a major genetic risk factor for Alzheimer's disease is also consistent with a role for cholesterol in the pathogenesis of neurodegenerative disease [39-40]. Despite this, the molecular mechanisms by which altered lipid metabolism leads to neuron degeneration are unclear.

An important question arising from our analysis centers on why LDL can causally affect ALS, while at the same time LDL levels are not genetically correlated with ALS under the LD Score regression model. This apparent divergence is because the variants linked to these two traits are not pleiotropic, and again highlights the fact that Mendelian randomization and LD regression analysis investigate different aspects of the genetic architecture underlying diseases. Mendelian randomization allows us to compare two groups of people that differ by the genetic variants of interest and therefore by any modifiable factor to which those genetic variants relate. In this case, genetic variants that are associated with LDL metabolism affect LDL levels and a ratio measure is calculated to determine how much this estimated change in LDL level would predispose individuals to ALS. If substantial pleiotropy were present, we would find that the same genetic variants that affect LDL metabolism also increase the risk of ALS by themselves (i.e genetic correlation). Such pleiotropy was not observed in our data. Instead, we found that the only mechanism by which ALS risk could be increased is through an increase of LDL cholesterol levels (i.e linear association).

Our data lead us to propose that lowering blood cholesterol levels is a viable strategy for reducing risk associated with ALS. A similar approach may be effective in Alzheimer's disease where exposure to statins is associated with substantially reduced risk of dementia in observational studies [41-42]. Though the American Heart Association guidelines for treating blood cholesterol to reduce cardiovascular risk are widely implemented in the community, they primarily focus on patients over the age of 50 [43]. An alternative strategy may be to identify a younger subpopulation at increased risk of developing ALS and to institute treatment with lipid-lowering agents. This approach would initially focus on individuals with a family history of ALS or frontotemporal dementia, and on pre-symptomatic cases carrying the *C9orf72* repeat expansion; together these subtypes account for nearly one in five cases of ALS [44]. Long-term monitoring would be required to detect side-effects from the medication, and to determine effect on the age of disease onset.

We conclude by saying that the reported findings should be interpreted with caution and in the context of existing evidence from other research studies using different designs, and definite conclusions should not be elaborated uniquely based on Mendelian randomization results. Future Randomised Controlled Trials should be considered as a proof of causality.

Limitations to this study

Our analyses were limited to only those GWAS studies present in two public databases, namely *LD-hub* and *MR Base*. Furthermore, the available data are focused on European populations. We envisage that future studies may expand our findings by employing larger sample sizes, greater density across the genome, and importantly Non-European populations, highlighting the utility of an ALS resource that is constantly updated as new data become available.

One of the main caveats of working with summary level data (rather than individual level data) is that there is no possibility to filter and exclude sample overlap. For Mendelian randomization analyses, we cannot exclude the possibility that samples from the same individuals were used in both the GWASes that we identified as significant exposures and in the ALS GWAS that we used as the outcome measure. Such sample overlap may bias estimates in MR and increase Type 1 error rates. We reviewed the origin of the European cohorts present in our ALS outcome and in the significantly associated exposures, and the results of this comparison are outlined in Table S11. Our data suggest that sample overlap had only a minimal effect on our results. We also performed sensitivity analyses by calculating the F-statistic parameter as described elsewhere [45]. Our results showed that two of the three GWASes of interest for which F-statistic could be calculated were considered strong instruments and are unlikely to be susceptible to bias due to overlapping samples (F-statistic for LDL cholesterol = 59.02, F-statistic for coronary heart disease =742.2).

Finally, sample overlap alone cannot account for our findings, as any sample overlap would be equally likely to occur across the diverse GWASes that we studied, and yet, we consistently identified altered lipid metabolism as a risk factor for ALS across multiple GWAS studies and across multiple populations. There is no realistic scenario in which sample overlap could have been consistently confined to just GWASes involving lipid metabolism. However, since Mendelian randomization effect estimates are often small, mandating additional follow-up on connected pathways.

A concern that might arise is to what extent hereditary cases of ALS carrying rare genetic variants might influence our analyses. One should expect that carrying large effect, rarer variants would not generally preclude the carrying of more common, small effect genetic risk factors which comprise the majority of GWAS results that were used for Mendelian randomization and LD Score regressions.

Finally, we are aware that certain bias could exist due to undetectable issues in underlying GWAS results utilized in this survey, but the fact that we have replicated our results in independent GWASes alleviates this concern.

ACKNOWLEDGMENTS

This work was supported by the Intramural Research Program of the NIH, National Institute on Aging (Z01-AG000949-02); by the National Institute of Neurological Disorders and Stroke; and by Merck & Co., Inc.. AJ Noyce is funded by the Barts Charity (Preventive Neurology Unit). Gibran Hemani is funded by the Wellcome Trust and the Royal Society [208806/Z/17/Z]. Andrea Calvo receives research support from Fondazione Vialli e Mauro onlus. Pentti J. Tienari received funding support from Helsinki University Hospital and the

Sigrid Jusélius Foundation. A. Chiò receives research support from the Italian Ministry of Health (Ricerca Finalizzata RF-2010-2309849 and RF-2016-02362405), the University of Turin (Ricerca locale ex 60% 2016 and 2017), the Joint Programme Neurodegenerative Disease Research (JPND) (BRAIN-MEND project supported by Italian Ministry of University). The research leading to these results has received funding from the European Community's Health Seventh Framework Programme (FP7/2007–2013) (grant agreement no. 259867) and support from the Target ALS Multicenter Postmortem Core and the Canadian Consortium on Neurodegeneration in Aging (ER). This study has been supported by the Italian Ministry of University under the Department of Excellence funding awarded to the 'Rita Levi Montalcini Department of Neuroscience. Bryan J. Traynor received additional support from the Center for Disease Control and Prevention, the Muscular Dystrophy Association, Microsoft Research, the Packard Center for ALS Research at Johns Hopkins, and the ALS Association.

This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the NIH, Bethesda, Maryland (http://hpc.nih.gov). This study also used genotype and clinical data from the Wellcome Trust Case Control Consortium, and from the HyperGenes Consortium. We thank the patients and research subjects who contributed samples for this study.

We thank ITALSGEN and International ALS Genetics Consortium for their contributions to this manuscript. Please see Supplemental materials for authors and affiliations.

AUTHOR CONTRIBUTIONS

S.B.C, A.J.N, M.A.N, and B.J.T contributed to the conception and design of the study. A.C, G.M, P.J.T, Ad.C, and B.J.T, S.B.C, A.J.N, G.B, M.N contributed to the acquisition and analysis of data. S.B.C, A.J.N, M.A.N, A.N, A.C, G.M, P.J.T, D.J.S, A.B.S, and Ad.C contributed to drafting a significant portion of the manuscript or figures.

CONFLICT OF INTERESTS

A.J.N receives grants from Parkinson's UK (ref G-1606), the Virginia Kieley Benefaction, the Leonard Wolfson Experimental Neurology Centre and has received honoraria from Britannia Pharmaceuticals Ltd, Profile Pharma Ltd and Global Kinetics Corporation, outside of the submitted work. Ad.C serves on scientific advisory boards for Biogen Idec, Cytokinetics, Italfarmaco, Neuraltus, and Mitsubishi Tanabe. B.J.T, P.J.T, and A.B.S hold patents on the clinical testing and therapeutic intervention for the hexanucleotide repeat expansion of *C9orf72*. All other authors declare that they have no conflicts of interest.

Commented [BCS([1]: As far as I understood we can remove this part since nothing of this is directly related to our study

REFERENCES

1. Chiò A, Logroscino G, Hardiman O, et al. Prognostic factors in ALS: A critical review. Amyotroph. Lateral Scler. 2009;10(5-6):310–323.

2. Arthur KC, Calvo A, Price TR, et al. Projected increase in amyotrophic lateral sclerosis from 2015 to 2040. Nat. Commun. 2016;7:12408.

3. Chia R, Chiò A, Traynor BJ. Novel genes associated with amyotrophic lateral sclerosis: diagnostic and clinical implications. Lancet Neurol. 2018;17(1):94–102.

4. Nicolas A, Kenna KP, Renton AE, et al. Genome-wide Analyses Identify KIF5A as a Novel ALS Gene. Neuron 2018;97(6):1268–1283.e6.

5. Belbasis L, Bellou V, Evangelou E. Environmental Risk Factors and Amyotrophic Lateral Sclerosis: An Umbrella Review and Critical Assessment of Current Evidence from Systematic Reviews and Meta-Analyses of Observational Studies. Neuroepidemiology 2016;46(2):96–105.

6. Marigorta UM, Rodríguez JA, Gibson G, Navarro A. Replicability and Prediction: Lessons and Challenges from GWAS. Trends Genet. 2018;34(7):504–517.

7. Bulik-Sullivan BK, Loh P-R, Finucane HK, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. 2015;47(3):291–295.

8. Smith GD, Ebrahim S. "Mendelian randomization": can genetic epidemiology contribute to understanding environmental determinants of disease? Int. J. Epidemiol. 2003;32(1):1–22.

9. Yang X, Guo Y, Liu Y. Bayesian-inference based recommendation in online social networks [Internet]. In: 2011 Proceedings IEEE INFOCOM. 2011Available from:

http://dx.doi.org/10.1109/infcom.2011.5935224

10. Zheng J, Erzurumluoglu M, Elsworth B, et al. LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis [Internet]. 2016;Available from: http://dx.doi.org/10.1101/051094

11. Hemani G, Zheng J, Wade KH, et al. MR-Base: a platform for systematic causal inference across the phenome using billions of genetic associations [Internet]. bioRxiv 2016;078972.[cited 2018 Sep 14] Available from: https://www.biorxiv.org/content/early/2016/12/16/078972.abstract

12. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. Int. J. Epidemiol. 2015;44(2):512–525.

13. Hemani G, Tilling K, Davey Smith G. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. PLoS Genet. 2017;13(11):e1007081.

14. Bandrés-Ciga S, Price TR, Barrero FJ, et al. Genome-wide assessment of Parkinson’s disease in a Southern Spanish population. Neurobiol. Aging 2016;45:213.e3–213.e9.

15. Giambartolomei C, Vukcevic D, Schadt EE, et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. 2014;10(5):e1004383.

16. Wang M-D, Little J, Gomes J, et al. Identification of risk factors associated with onset and progression of amyotrophic lateral sclerosis using systematic review and meta-analysis. Neurotoxicology 2017;61:101–130.

17. Larsson E, Hultqvist E. Desirable places: spatial representations and educational strategies in the inner city. Br. J. Sociol. Educ. 2017;39(5):623–637.

18. Papma JM, Jiskoot LC, Panman JL, et al. Cognition and gray and white matter characteristics of presymptomatic repeat expansion. Neurology 2017;89(12):1256–1264.

19. Lee SE, Sias AC, Mandelli ML, et al. Network degeneration and dysfunction in presymptomatic expansion carriers. Neuroimage Clin 2017;14:286–297.

20. Chancellor AM, Mitchell JD, Swingler RJ. The first description of idiopathic progressive bulbar palsy. J. Neurol. Neurosurg. Psychiatry 1993;56(12):1270.

21. Gunnarsson L-G, -G. Gunnarsson L, Lindberg G, et al. Amyotrophic lateral sclerosis in Sweden in relation to occupation. Acta Neurol. Scand. 1991;83(6):394–398.

22. Chiò A, Benzi G, Dossena M, et al. Severely increased risk of amyotrophic lateral sclerosis among Italian professional football players. Brain 2005;128(Pt 3):472–476.

23. Mattson MP, Magnus T. Ageing and neuronal vulnerability. Nat. Rev. Neurosci. 2006;7(4):278–294.

24. Parker L, McGuckin TA, Leicht AS. Influence of exercise intensity on systemic oxidative stress and antioxidant capacity. Clin. Physiol. Funct. Imaging 2014;34(5):377–383.

25. Harwood CA, McDermott CJ, Shaw PJ. Physical activity as an exogenous risk factor in motor neuron disease (MND): a review of the evidence. Amyotroph. Lateral Scler. 2009;10(4):191–204.

26. Armon C. Smoking may be considered an established risk factor for sporadic ALS. Neurology 2009;73(20):1693–1698.

27. Mariosa D, Hammar N, Malmström H, et al. Blood biomarkers of carbohydrate, lipid, and apolipoprotein metabolisms and risk of amyotrophic lateral sclerosis: A more than 20-year follow-up

of the Swedish AMORIS cohort. Ann. Neurol. 2017;81(5):718–728.

28. Dupuis L, Corcia P, Fergani A, et al. Dyslipidemia is a protective factor in amyotrophic lateral sclerosis. Neurology 2008;70(13):1004–1009.

29. Goldstein MR, Mascitelli L, Pezzetta F, et al. DYSLIPIDEMIA IS A PROTECTIVE FACTOR IN AMYOTROPHIC LATERAL SCLEROSIS. Neurology 2008;71(12):956–957.

30. Schmitt F, Hussain G, Dupuis L, et al. A plural role for lipids in motor neuron diseases: energy, signaling and structure. Front. Cell. Neurosci. 2014;8:25.

31. Timmins HC, Saw W, Cheah BC, et al. Cardiometabolic health and risk of amyotrophic lateral sclerosis. Muscle Nerve 2017;56(4):721–725.

32. Kioumourtzoglou M-A, Seals RM, Gredal O, et al. Cardiovascular disease and diagnosis of amyotrophic lateral sclerosis: A population based study. Amyotroph. Lateral Scler. Frontotemporal Degener. 2016;17(7-8):548–554.

33. Chio A, Calvo A, Ilardi A, et al. Lower serum lipid levels are related to respiratory impairment in patients with ALS. Neurology 2009;73(20):1681–1685.

34. Chen X, Yazdani S, Piehl F, et al. Polygenic link between blood lipids and amyotrophic lateral sclerosis. Neurobiol. Aging 2018;67:202.e1–202.e6.

35. Pfrieger FW. Cholesterol homeostasis and function in neurons of the central nervous system. Cell. Mol. Life Sci. 2003;60(6):1158–1171.

36. Vance JE. Dysregulation of cholesterol balance in the brain: contribution to neurodegenerative diseases. Dis. Model. Mech. 2012;5(6):746–755.

37. Puglielli L, Tanzi RE, Kovacs DM. Alzheimer’s disease: the cholesterol connection. Nat. Neurosci. 2003;6(4):345–351.

38. Wang W, Mutka A-L, Zmrzljak UP, et al. Amyloid precursor protein α- and β-cleaved ectodomains exert opposing control of cholesterol homeostasis via SREBP2. The FASEB Journal 2014;28(2):849–860.

39. Corder EH, Saunders AM, Strittmatter WJ, et al. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer’s disease in late onset families. Science 1993;261(5123):921–923.

40. Canosa A, Pagani M, Brunetti M, et al. Correlation between APOE genotype and brain metabolism in ALS [Internet]. Eur. J. Neurol. 2018;Available from: http://dx.doi.org/10.1111/ene.13812

41. Zissimopoulos JM, Barthold D, Brinton RD, Joyce G. Sex and Race Differences in the Association Between Statin Use and the Incidence of Alzheimer Disease. JAMA Neurol. 2017;74(2):225–232.

42. McGuinness B, Passmore P. Can Statins Prevent or Help Treat Alzheimer’s Disease? J. Alzheimers. Dis. 2010;20(3):925–933.

43. Stone NJ, Robinson JG, Lichtenstein AH, et al. Treatment of blood cholesterol to reduce atherosclerotic cardiovascular disease risk in adults: synopsis of the 2013 American College of Cardiology/American Heart Association cholesterol guideline. Ann. Intern. Med. 2014;160(5):339–343.

44. Chia R, Chiò A, Traynor BJ. Novel genes associated with amyotrophic lateral sclerosis: diagnostic and clinical implications. Lancet Neurol. 2018;17(1):94–102.

45. Burgess S, Davies NM, Thompson SG. Bias due to participant overlap in two-sample Mendelian

randomization. Genet. Epidemiol. 2016;40(7):597–608.

FIGURES AND TABLES

SUPPLEMENTARY MATERIAL -> https://github.com/SaraBandres/ALS

Table S9. Genetic risk profiling of LDL cholesterol in ALS subtypes

Table S10. Bayesian colocalization analysis of the 78 independent regions associated to LDL cholesterol and ALS

Table S11. Summary of the cohorts included in the exposures of interest and the outcome

Description of the significant phenotypes genetically correlated with ALS and included in the UK Biobank dataset

CONSORTIA members

FIGURE LEGENDS

Figure 1. Flowchart of analysis

MR Base GWASes are available at http://www.mrbase.org/; GWAS, Genome-Wide Association study; kb, kilobases; R^2, clumping threshold; LDL, low-density lipoprotein.

Figure 2. Bayesian colocalization plots

A plot and B plot represent two independent LDL cholesterol-associated regions with posterior probability greater than 95 % of sharing a causal variant involved in ALS. Panels in column A show the region spanning chr5:73656720-75651786 where rs182826525 is likely the shared causal variant with a posterior probability of nearly 100%. Panels in column B show the region spanning chr5:155390511-157388284 where rs116226146 is likely the shared causal variant with a posterior probability of 96%. The first row displays the p-values from the LDL GWAS for each region. Color is coded by p-values in the ALS GWAS. The second row displays the P-values from the ALS GWAS for the same regions. Color is coded by P-values in the LDL GWAS. The third row shows local gene positions (with strands denoted by +/-), as well as recombination rates measured in cM/Mb. [38] The bottom row shows the posterior probabilities of a shared causal variant between LDL cholesterol and ALS.

Table 1. Linkage disequilibrium score regression results for traits genetically correlated with ALS

	Trait	Source	rg [se]	p-value	FDR p-value	h^2 [se]
EDUCATION	Fluid intelligence score	UKBB	-0.338 [0.067]	4.74E-07	1.78E-04	0.238 [0.011]
	Qualifications: Other professional qualifications eg: nursing_ teaching	UKBB	-0.257 [0.071]	3.00E-04	1.73E-02	0.047 [0.003]
	Qualifications: A levels/AS levels or equivalent	UKBB	-0.255 [0.059]	1.61E-05	1.66E-03	0.097 [0.004]
	Qualifications: College or University degree	UKBB	-0.249 [0.053]	2.77E-06	5.19E-04	0.168 [0.005]
	Qualifications: O levels/GCSEs or equivalent	UKBB	-0.238 [0.069]	5.00E-04	2.68E-02	0.049 [0.003]
	Age completed full time education	UKBB	-0.229 [0.068]	7.00E-04	3.09E-02	0.084 [0.005]
	Years of schooling 2016	27225129	-0.226 [0.059]	1.00E-04	6.83E-03	0.127 [0.004]
	Number of incorrect matches in round	UKBB	0.229 [0.06]	1.00E-04	6.83E-03	0.055 [0.003]
	Qualifications: None of the above	UKBB	0.255 [0.059]	1.49E-05	1.66E-03	0.098 [0.004]
ACTIVITY	Types of transport used (excluding work): Walk	UKBB	-0.403 [0.085]	2.11E-06	5.19E-04	0.033 [0.002]
	Types of transport used (excluding work): Public transport	UKBB	-0.402 [0.092]	1.13E-05	1.66E-03	0.022 [0.002]
	Types of physical activity in last 4 weeks: Light DIY	UKBB	-0.287 [0.07]	4.24E-05	3.54E-03	0.039 [0.002]
	Types of physical activity in last 4 weeks: Walking for pleasure	UKBB	-0.286 [0.079]	3.00E-04	1.73E-02	0.037 [0.002]
	Duration of moderate activity	UKBB	0.283 [0.084]	7.00E-04	3.09E-02	0.032 [0.002]
	Job involves mainly walking or standing	UKBB	0.216 [0.065]	9.00E-04	3.56E-02	0.08 [0.004]
SMOKING	Exposure to tobacco smoke at home	UKBB	0.42 [0.122]	6.00E-04	3.00E-02	0.012 [0.002]
	Light smokers (at least 100 smokes in lifetime)	UKBB	0.427 [0.1]	1.77E-05	1.66E-03	0.077 [0.008]
OTHER	Frequency of tenseness / restlessness in last 2 weeks	UKBB	0.227 [0.068]	9.00E-04	3.56E-02	0.044 [0.003]

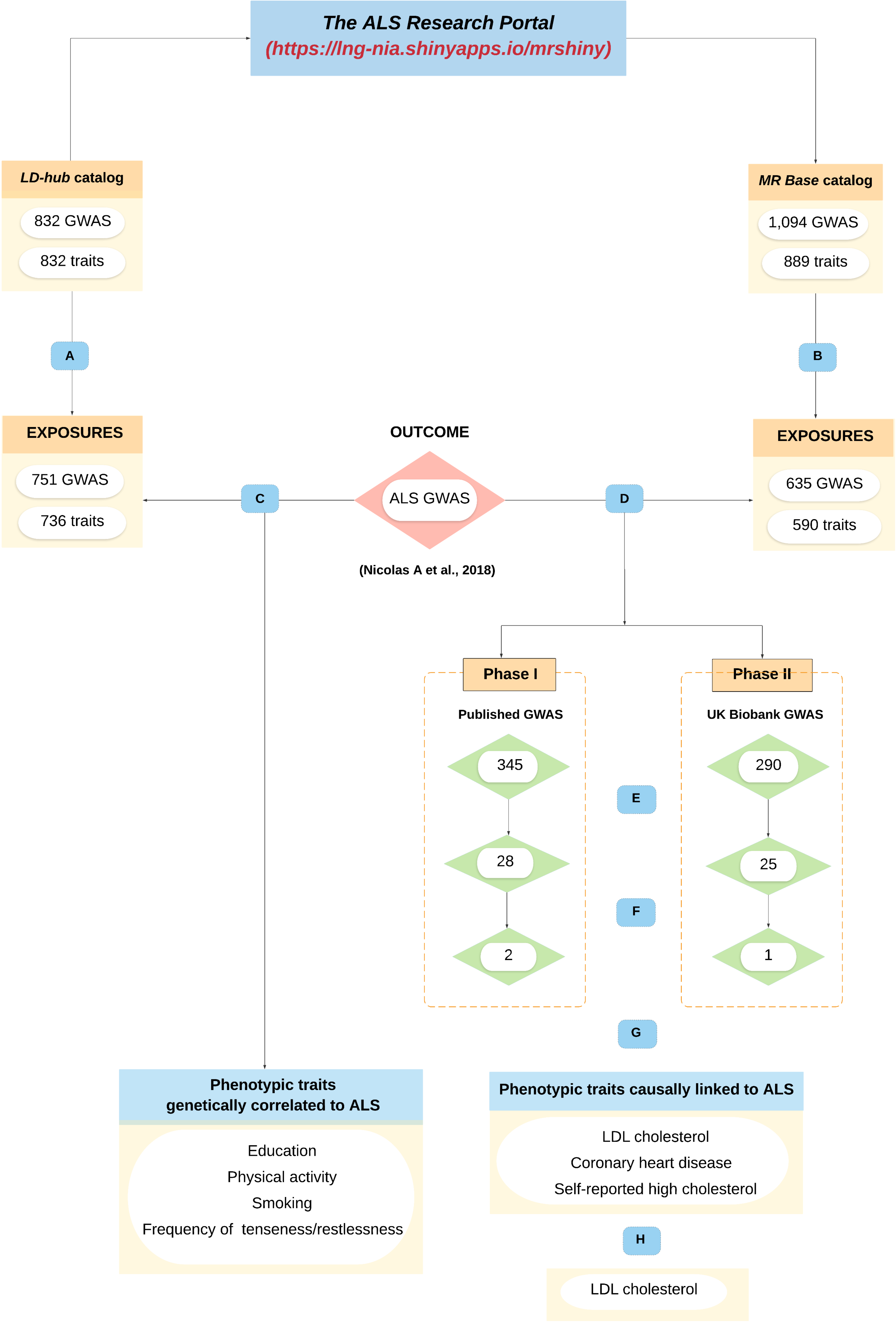
See supplementary materials for a description of the phenotypes included in the UK Biobank dataset, and Figure 1 for the number of traits screened as part of the LD score regression analysis. Source: number denotes PubMed identification numbers; UKBB, UK Biobank; rg, regression; se, standard error; FDR, false discovery rate adjusted p-value; h^2, observed narrow-sense heritability.

Table 2. Mendelian randomization results for exposures causally linked to ALS

			Inverse variance weighted		MR Egger		Weighted Median	
Exposure	**Source**	**No. of SNPs**	**OR [CI 95 %]**	**p-value**	**OR [CI 95 %]**	**p-value**	**OR [CI 95 %]**	**p-value**
LDL cholesterol id:300	24097068	78	1.116 [1.036-1.20]	0·003	1.115 [1.00-1.233]	0·054	1.108 [1.00-1.226]	0·046
Coronary heart disease id:7	26343387	37	1.063 [1.0-1.13]	0·047	1.175 [1.019-1.356]	0·032	1.116 [1.020-1.220]	0·015
Self-reported: high cholesterol id:UKB-a:108	UKBB	49	2.389 [1.48-3.842]	0·0003	2.669 [1.084-6.55]	0·038	2.110 [1.021-4.357]	0·044

id,specific code attributed to each trait by *MR Base;* No.of SNPs, number of SNPs;
CI, confidence interval; LDL, low densitity lipoprotein.

Figure 1. Flowchart of analysis


The ALS Research Portal
(https://lng-nia.shinyapps.io/mrshiny)
LD-hub catalog
832 GWAS
832 traits
MR Base catalog
1,094 GWAS
889 traits
A
B
EXPOSURES
751 GWAS
736 traits
OUTCOME
ALS GWAS
(Nicolas A et al., 2018)
C
D
EXPOSURES
635 GWAS
590 traits
Phase I
Published GWAS
345
28
2
E
F
Phase II
UK Biobank GWAS
290
25
1
G
Phenotypic traits
genetically correlated to ALS
Education
Physical activity
Smoking
Frequency of tenseness/restlessness
Phenotypic traits causally linked to ALS
LDL cholesterol
Coronary heart disease
Self-reported high cholesterol
H
LDL cholesterol