

Genome-wide study of DNA methylation in Amyotrophic Lateral Sclerosis identifies differentially methylated loci and implicates metabolic, inflammatory and cholesterol pathways

Paul J. Hop^{1,#}, Ramona A.J. Zwamborn^{1,#}, Eilis Hannon², Gemma L. Shireby², Marta F. Nabais^{2,3}, Emma M. Walker², Wouter van Rheenen¹, Joke J.F.A. van Vugt¹, Annelot M. Dekker¹, Henk-Jan Westeneng¹, Gijs H.P. Tazelaar¹, Kristel R. van Eijk¹, Matthieu Moisse^{4,5,6}, Denis Baird^{7,8}, Ahmad Al Khleifat⁹, Alfredo Iacoangeli^{9,10,11}, Nicola Ticozzi^{12,13}, Antonia Ratti^{12,14}, Jonathan Cooper-Knock¹⁵, Karen E. Morrison¹⁶, Pamela J. Shaw¹⁷, A. Nazli Basak¹⁸, Adriano Chiò^{19,20}, Andrea Calvo^{19,20}, Cristina Moglia^{19,20}, Antonio Canosa^{19,20}, Maura Brunetti¹⁹, Maurizio Grassano¹⁹, Marc Gotkine^{21,22}, Yossef Lerner^{21,22}, Michal Zabari^{21,22}, Patrick Vourc'h^{23,24}, Philippe Corcia^{25,24}, Philippe Couratier^{26,27}, Jesus S. Mora Pardina²⁸, Teresa Salas²⁹, Patrick Dion³⁰, Jay P. Ross^{30,31}, Robert D. Henderson^{32,32}, Susan Mathers³³, Pamela A. McCombe³⁴, Merrilee Needham^{35,36,37}, Garth Nicholson³⁸, Dominic B. Rowe³⁹, Roger Pamphlett⁴⁰, Karen A. Mather^{41,42}, Perminder S. Sachdev^{41,43}, Sarah Furlong³⁹, Fleur C. Garton³, Anjali K. Henders³, Tian Lin³, Shyuan T. Ngo^{44,45,34}, Frederik J. Steyn^{46,34}, Leanne Wallace³, Kelly L. Williams³⁹, BIOS Consortium*, Brain MEND Consortium*, Miguel Mitne Neto⁴⁷, Ruben J. Cauchi⁴⁸, Ian P. Blair³⁹, Matthew C. Kiernan^{49,50}, Vivian Drory^{51,52}, Monica Povedano⁵³, Mamede de Carvalho⁵⁴, Susana Pinto⁵⁴, Markus Weber⁵⁵, Guy Rouleau³⁰, Vincenzo Silani^{12,13}, John E. Landers⁵⁶, Christopher E. Shaw⁹, Peter M. Andersen⁵⁷, Allan F. McRae³, Michael A. van Es¹, R. Jeroen Pasterkamp⁵⁸, Naomi R. Wray^{3,45}, Russell L. McLaughlin⁵⁹, Orla Hardiman⁶⁰, Kevin P. Kenna^{1,58}, Ellen Tsai⁷, Heiko Runz⁷, Ammar Al-Chalabi^{9,61}, Leonard H. van den Berg¹, Philip Van Damme^{4,5,6}, Jonathan Mill^{2,+}, Jan H. Veldink^{1,+}

Affiliations

1: Department of Neurology, UMC Utrecht Brain Center, University Medical Center Utrecht, Utrecht, The Netherlands.

2: University of Exeter Medical School, College of Medicine and Health, University of Exeter, Exeter, UK.

3: Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD 4072, Australia.

- 4: KU Leuven – University of Leuven, Department of Neurosciences, Experimental Neurology, and Leuven Brain Institute (LBI), Leuven, Belgium.
- 5: VIB, Center for Brain and Disease Research, Leuven, Belgium.
- 6: University Hospitals Leuven, Department of Neurology, Leuven, Belgium.
- 7: Translational Biology, Biogen, Boston, Massachusetts, USA.
- 8: MRC Integrative Epidemiology Unit (IEU), Population Health Sciences, University of Bristol, Bristol, UK.
- 9: Maurice Wohl Clinical Neuroscience Institute, Department of Basic and Clinical Neuroscience, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, UK.
- 10: Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK.
- 11: National Institute for Health Research Biomedical Research Centre and Dementia Unit, South London and Maudsley NHS Foundation Trust and King's College London, London, UK.
- 12: Department of Neurology-Stroke Unit and Laboratory of Neuroscience, Istituto Auxologico Italiano IRCCS, Milan, Italy.
- 13: Department of Pathophysiology and Transplantation, “Dino Ferrari” Center, Università degli Studi di Milano, Milan, Italy.
- 14: Department of Medical Biotechnology and Translational Medicine, Università degli Studi di Milano, Milano, Italy.
- 15: Sheffield Institute for Translational Neuroscience (SITraN), University of Sheffield, Sheffield, UK.
- 16: School of Medicine, Dentistry, and Biomedical Sciences, Queen's University Belfast, Belfast, UK.
- 17: Sheffield Institute for Translational Neuroscience (SITraN), University of Sheffield, Sheffield, United Kingdom.
- 18: Koc University, School of Medicine, Translational Medicine Research Center, NDAL, Istanbul, Turkey.
- 19: “Rita Levi Montalcini” Department of Neuroscience, ALS Centre, University of Torino, Turin, Italy.
- 20: Azienda Ospedaliero-Universitaria Città della Salute e della Scienza, SC Neurologia 1U, Turin, Italy.
- 21: Faculty of Medicine, Hebrew University of Jerusalem, Israel.

- 22: The Agnes Ginges Center for Human Neurogenetics, Dept. of Neurology, Hadassah Medical Center, Jerusalem, Israel.
- 23: Service de Biochimie et Biologie moléculaire, CHU de Tours, Tours, France.
- 24: UMR 1253, Université de Tours, Inserm, 37044 Tours, France.
- 25: Centre de référence sur la SLA, CHU de Tours, Tours, France.
- 26: Centre de référence sur la SLA, CHRU de Limoges, Limoges, France.
- 27: UMR 1094 Université de Limoges, Inserm, 87025 Limoges, France.
- 28: ALS Unit, Hospital San Rafael, Madrid, Spain.
- 29: Department of Neurology, Hospital La Paz-Carlos III, Madrid.
- 30: Montréal Neurological Institute and Hospital, McGill University, Montréal, QC, H3A 2B4, Canada.
- 31: Department of Human Genetics, McGill University, Montréal, QC, H3A 0C7, Canada.
- 32: Department of Neurology, Royal Brisbane and Women's Hospital, Brisbane, QLD 4029, Australia.
- 33: Calvary Health Care Bethlehem, Parkdale, VIC 3195, Australia.
- 34: Centre for Clinical Research, The University of Queensland, Brisbane, QLD 4019, Australia.
- 35: Fiona Stanley Hospital, Perth, WA 6150, Australia.
- 36: Notre Dame University, Fremantle, WA 6160, Australia.
- 37: Institute for Immunology and Infectious Diseases, Murdoch University, Perth, WA 6150, Australia.
- 38: ANZAC Research Institute, Concord Repatriation General Hospital, Sydney, NSW 2139, Australia.
- 39: Centre for Motor Neuron Disease Research, Macquarie University, NSW 2109, Australia.
- 40: Discipline of Pathology and Department of Neuropathology, Brain and Mind Centre, The University of Sydney, Sydney, NSW 2050, Australia.
- 41: Centre for Healthy Brain Ageing, School of Psychiatry, University of New South Wales, Sydney, NSW 2031, Australia.
- 42: Neuroscience Research Australia Institute, Randwick, NSW 2031, Australia.
- 43: Neuropsychiatric Institute, The Prince of Wales Hospital, UNSW, Randwick, NSW 2031, Australia.
- 44: The Australian Institute for Bioengineering and Nanotechnology, The University of Queensland, Brisbane, QLD 4072, Australia.
- 45: Queensland Brain Institute, The University of Queensland, Brisbane, QLD 4072, Australia.

- 46: School of Biomedical Sciences, The University of Queensland, Brisbane, QLD 4072, Australia.
- 47: Universidade de São Paulo, São Paulo, Brazil.
- 48: Center for Molecular Medicine and Biobanking & Department of Physiology and Biochemistry, Faculty of Medicine and Surgery, University of Malta, Malta.
- 49: Brain and Mind Centre, The University of Sydney, Sydney, New South Wales, Australia.
- 50: Department of Neurology, Royal Prince Alfred Hospital, Sydney NSW 2050, Australia.
- 51: Department of Neurology, Tel-Aviv Sourasky Medical Centre, Tel-Aviv, Israel.
- 52: Sackler Faculty of Medicine, Tel-Aviv University, Tel-Aviv, Israel.
- 53: Functional Unit of Amyotrophic Lateral Sclerosis (UFELA), Service of Neurology, Bellvitge University Hospital, L'Hospitalet de Llobregat, Barcelona, Spain.
- 54: Instituto de Fisiologia, Instituto de Medicina Molecular João Lobo Antunes, Faculdade de Medicina, Universidade de Lisboa, Lisbon, Portugal.
- 55: Neuromuscular Diseases Unit/ALS Clinic, Kantonsspital St. Gallen, 9007, St. Gallen, Switzerland.
- 56: Department of Neurology, University of Massachusetts Medical School, Worcester, Massachusetts, USA.
- 57: Department of Pharmacology and Clinical Neuroscience, Umeå University, Umeå, Sweden.
- 58: Department of Translational Neuroscience, UMC Utrecht Brain Center, University Medical Center Utrecht, Utrecht, The Netherlands.
- 59: Complex Trait Genomics Laboratory, Smurfit Institute of Genetics, Trinity College Dublin, Dublin D02 PN40, Ireland.
- 60: Academic Unit of Neurology, Trinity Biomedical Sciences Institute, Trinity College Dublin, Dublin D02 PN40, Ireland.
- 61: King's College Hospital, Denmark Hill, London, UK.

Shared first authors

+ Shared last authors

* A list of authors and their affiliations appears at the end of the paper.

Correspondence

Jan H. Veldink: j.h.veldink@umcutrecht.nl

Abstract

Amyotrophic lateral sclerosis (ALS) is a fatal neurodegenerative disease with an estimated heritability of around 50%. DNA methylation patterns can serve as biomarkers of (past) exposures and disease progression, as well as providing a potential mechanism that mediates genetic or environmental risk. Here, we present a blood-based epigenome-wide association study (EWAS) meta-analysis in 10,462 samples (7,344 ALS patients and 3,118 controls), representing the largest case-control study of DNA methylation for any disease to date. We identified a total of 45 differentially methylated positions (DMPs) annotated to 42 genes, which are enriched for pathways and traits related to metabolism, cholesterol biosynthesis, and immunity. We show that DNA-methylation-based proxies for HDL-cholesterol, BMI, white blood cell (WBC) proportions and alcohol intake were independently associated with ALS. Integration of these results with our latest GWAS showed that cholesterol biosynthesis was causally related to ALS. Finally, we found that DNA methylation levels at several DMPs and blood cell proportion estimates derived from DNA methylation data, are associated with survival rate in patients, and could represent indicators of underlying disease processes.

Introduction

Amyotrophic lateral sclerosis (ALS) is a fatal neurodegenerative disorder that is characterized by progressive degeneration of motor neurons in the brain and spinal cord¹. The disease affects about 1 in 350 people, with death typically occurring within 2 to 5 years after onset. The heritability of ALS is estimated to be around 50%², showing that a considerable portion of the risk could be conferred by environmental and lifestyle risk factors. However, the identification of these factors has proven difficult due to several challenges. Lifestyle factors are typically studied using self reports, or interviews with the risk of recall and selection bias (e.g. because of selection against patients with cognitive impairment)^{3,4}. Finally, biomarkers are typically measured after onset of disease making inferences about causality virtually impossible. The combined effect of these challenges has resulted in a large body of literature with conflicting results and only a few established factors related to ALS risk or patient survival^{3,5}.

Epigenetic patterns, which regulate gene expression via modifications to DNA, histone proteins, and chromatin, can serve as biomarkers of disease progression and (past) exposures such as smoking and alcohol intake. Moreover, epigenetics patterns can provide a mechanism that mediates genetic and/or environmental risk⁶. The development of standardized assays for quantifying DNA methylation, the best characterized and most stable epigenetic modification, has enabled the systematic analysis of associations between methylomic variation and a wide range of human diseases, including cancer, schizophrenia and various neurodegenerative diseases⁶⁻⁹. In addition to a potential mechanistic role in ALS, the quantification of DNA methylation in whole blood DNA can serve as a robust quantifiable biomarker for candidate risk factors including smoking, alcohol intake, BMI, biological age, and various metabolic and inflammatory proteins¹⁰⁻¹⁶. Moreover, DNA methylation signatures have proven useful as diagnostic and prognostic biomarkers in various settings including common cancers and all-cause mortality^{17,18}.

Few studies have investigated the role of DNA methylation in ALS. Most studies have either focused on specific genomic regions such as the *C9orf72* locus^{19,20}, or were limited in terms of sample size or coverage²¹⁻²⁴. The largest epigenome-wide association study (EWAS) to date included nearly 1,500 samples, but identified only a single significant differentially methylated position (DMP)²⁵.

We present a blood-based DNA methylation study of ALS incorporating over 10,000 samples, representing the largest case-control study of DNA methylation for any disease to date. First, we implemented a stringent pipeline to meta-analyze EWAS results across multiple strata and identify DMPs associated with ALS. Second, we show how DNA methylation data can be leveraged to serve as a proxy for risk factors and exposures relevant to ALS. Third, we integrate our DNA methylation data with findings from a new genome-wide association study (GWAS) of ALS²⁶, identifying potential causal roles for pathways highlighted by our DNA methylation study. Finally, we used overlapping whole-genome-sequencing (WGS) data to assess the influence of genetic variants and leveraged extensive clinical data for survival analyses.

Results

EWAS meta-analysis of ALS identifies 45 differentially methylated loci

We quantified genome-wide levels of DNA methylation in whole blood from 10,462 individuals across four strata using the Illumina 450k array (6,275 samples) and the Illumina EPIC array (4,187 samples). A total of 6,763 ALS patients and 2,943 control subjects passed our stringent quality control, which was followed by normalization of signal intensities in each stratum (**Table 1, supplementary file 1 and supplementary tables 1-5**). Samples excluded from our analyses did not show different demographic or clinical characteristics compared to the subset selected for analyses (**supplementary table 5**). We performed an EWAS in each of the four strata using two methods to adjust for known and unknown confounding. First we used a linear model adjusting for known confounders and a calibrated number of principal components (PCs) to adjust for unknown confounding factors (**Methods; supplementary figure 1**), followed by correction for residual bias and inflation in test-statistics using *bacon*²⁷ (hereafter referred to as the LB model). Second, we used MOA (mixed linear model-based omic association) as implemented in the OSCA software, in which the random effect of total genome-wide DNA methylation captures the correlation structure between probes and directly controls for the genomic inflation²⁸. Of note, the MOA algorithm did not converge for the AUS2 stratum, resulting in a total sample size of 9,459 for the MOA results. Test-statistics across strata were combined using an inverse variance-weighted fixed effects meta-analysis²⁹. Inflation of the test-statistics was well controlled in both the LB ($\lambda = 1.046$; **Figure 1c**) and the MOA results respectively ($\lambda = 0.984$; **Figure 1d**), and we observed little heterogeneity between strata (**Methods; supplementary figure 2**). Various technical sensitivity analyses indicated that the results were robust to changes in analysis strategy, including using M-values instead of β -values, using functional normalization³⁰ instead of *dasen*³¹, and excluding specific strata or experimental batches (**supplementary figure 3-4**). Finally, application of a method that we recently described³² led to the removal of likely cross-hybridizing probes, including four probes that showed high homology to the C9 repeat locus (**supplementary figure 5**). In total, 724,712 sites passed quality control, of which 332,066 were specific to the EPIC array. The LB meta-analysis resulted in 44 significant DMPs ($P < 9 \times 10^{-8}$; **Figure 1a, Table 2, supplementary table 6**) and the MOA meta-analysis resulted in 11 significant DMPs ($P < 9 \times 10^{-8}$; **Figure 1b, supplementary table 7**)³³. The significant MOA sites comprised a subset of the

significant LB sites, with the exception of cg01589155 which is annotated to the *C9orf72* locus; this site was significant in MOA ($P = 1.51 \times 10^{-8}$) and just below the significance threshold in the LB results ($P = 2.59 \times 10^{-7}$) (**supplementary figure 6**). Effect sizes were generally small and we observed both hypermethylated (51%) and hypomethylated (49%) DMPs associated with ALS (**supplementary figure 7**). Based on nearest gene mapping, these DMPs were annotated to 42 unique genes. Additionally, we annotated each site with *cis* expression associations in blood (eQTMs) calculated in an external dataset (BBMRI³⁴). This revealed that DNA methylation at 18 sites was significantly associated with the expression levels of at least one nearby gene, which included the nearest gene in 14 out of 18 sites (**Table 2, supplemental table 8**).

Table 1 - Demographic and clinical characteristics of study population:

Shown are numbers (and percentages) of samples that passed quality control.

	Project MinE		External	
	MinE 450k (N=4,474)	MinE EPIC (N=3,897)	AUS_batch1 ** (N=1,088)	AUS_batch2 ** (N=247)
Diagnosis				
Control	1,436 (32 %)	915 (23 %)	493 (45 %)	99 (40 %)
Case	3,038 (68 %)	2,982 (77 %)	595 (55 %)	148 (60 %)
Sex at birth				
Female	1,863 (42 %)	1,700 (44 %)	487 (45 %)	124 (50 %)
Male	2,611 (58 %)	2,197 (56 %)	601 (55 %)	123 (50 %)
Age (years)				
Mean (SD)	63 (± 11)	61 (± 13)	70 (± 12)	
Missing	438 (9.8%)	949 (24.4%)	77 (7.1%)	
Site of onset *				
Bulbar	861 (28 %)	739 (25 %)	173 (29 %)	36 (24 %)

Generalized	98 (3 %)	112 (4 %)	0 (0 %)	0 (0 %)
Spinal	2,023 (67 %)	2,060 (69 %)	0 (0 %)	0 (0 %)
Thoracic	10 (0 %)	5 (0 %)	0 (0 %)	0 (0 %)
Missing	46 (1.5 %)	66 (2.2 %)	422 (70.9 %)	112 (75.7 %)
Survival status *				
Alive	437 (14 %)	1,112 (37 %)	516 (87 %)	43 (29 %)
Dead	2,564 (84 %)	1,845 (62 %)	79 (13 %)	87 (59 %)
Missing	37 (1.2 %)	25 (0.8 %)	0 (0 %)	18 (12.2 %)
Survival (months) *†				
Median (Q1-Q3)	31.4 (31.4-48.9)	31.3 (21.1-47.1)	31.5 (23.6-44.4)	38.3 (25.4-66.5)
Missing	17 (0.7%)	9 (0.5%)	2 (2.5%)	1 (1.1%)
C9orf72 status *				
Expanded (≥30)	200 (7%)	155 (5 %)		
Normal	2,809 (92 %)	2,780 (93 %)		
Missing	29 (1.0 %)	47 (1.6 %)		

** data only included in case/control analyses

* case only

† dead-only

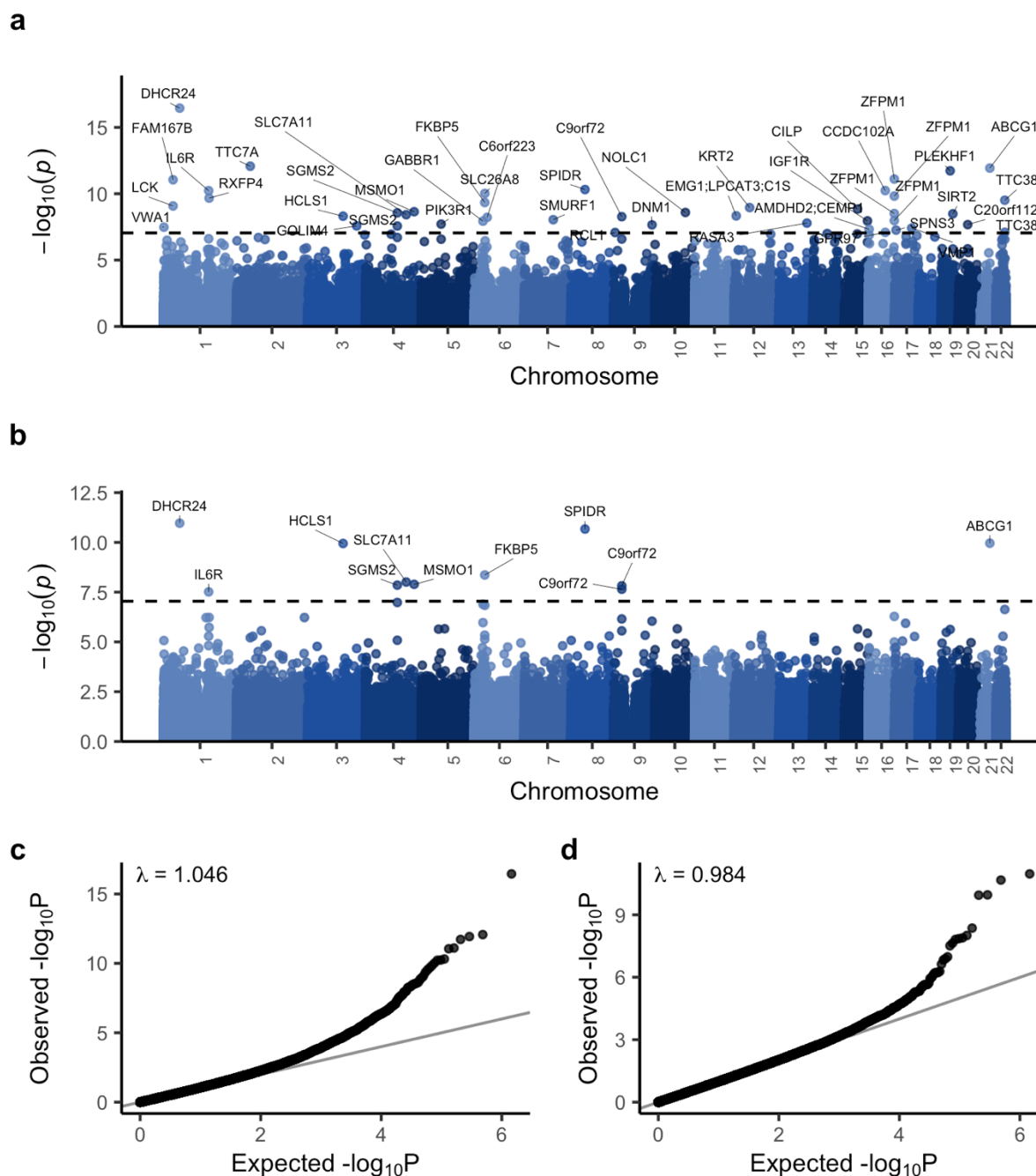


Figure 1 - EWAS meta-analysis: Epigenome-wide association study on 6,763 patients and 2,943 controls. **(a,b)** Manhattan plot comparing **(a)** LB (linear model + *bacon*) and **(b)** OSCA MOA association P -values ($-\log_{10}(P)$, y-axis) and genomic location (x-axis). The dashed line indicates the genome-wide significance threshold (9×10^{-8}). Sites were annotated with the nearest protein-coding gene in ensembl. **(c,d)** QQ-plot showing observed **(c)** LB and **(d)** OSCA MOA P -values ($-\log_{10}(P)$, y-axis) against the expected distribution under the null (x-axis).

Table 2 - Top ten significant sites:

Details of the ten most significant sites identified with the LB algorithm. Position = Chromosome:bp (GRCh37), Nearest gene = nearest gene based on Ensembl GRCh37 (75), eQTM = the most significant eQTM for the respective probe, eQTM *FDR* = *p*-value corresponding to the most significant eQTM, *FDR*-corrected for the number of tests for the respective probe, PMS = Indicates that the probe is part of the respective PMS (poly-methylation score), Trait = Overlap with significantly enriched traits from the MRC-IEU and NGDC EWAS databases (showing a maximum of five traits). Abbreviations: HGF = Hepatocyte growth factor, N.CDase = Neutral ceramidase, FGF.21 = Fibroblast growth factor 21.

Probe	Position	Nearest gene	eQTM (direction)	eQTM <i>FDR</i>	b	se	P-value	PMS	Traits
cg17901584	1: 55353706	<i>DHCR24</i>	<i>DHCR24</i> (-)	2.9×10^{-62}	0.0090	0.00110	3.6×10^{-17}	BMI, HDLchol, HGF	Hepatic fat, BMI, Metabolic trait, (serum) Triglycerides
cg06528816	2: 47242277	<i>TTC7A</i>	<i>TTC7A</i> (-)	0.13	0.0035	0.00049	8.5×10^{-13}		Allergic sensitization, Total serum IgE
cg06500161	21: 43656587	<i>ABCG1</i>	<i>ABCG1</i> (-)	1.6×10^{-25}	-0.0052	0.00073	1.2×10^{-12}	BMI, HDLchol, N.CDase, FGF.21	Hepatic fat, BMI, Metabolic trait, (serum) Triglycerides
cg14945937	19: 30162771	<i>PLEKHF1</i>	<i>PLEKHF1</i> (-)	0.02	-0.0041	0.00058	1.9×10^{-12}		
cg08940169	16: 88540241	<i>ZFPM1</i>	<i>PIEZO1</i> [†] (-)	0.08	0.0037	0.00054	7.8×10^{-12}		Allergic sensitization, Total serum IgE, Childhood asthma, Schizophrenia
cg07571745	1: 32715428	<i>FAM167B</i>	<i>CCDC28B</i> [†] (-)	0.26	-0.0033	0.00048	8.9×10^{-12}		
cg14195992	8: 48265917	<i>SPIDR</i>	<i>SPIDR</i> (-)	0.0059	-0.0053	0.00080	4.8×10^{-11}		Birth weight
cg08851837	16: 57558820	<i>CCDC102A</i>	<i>GPR56</i> [†] (+)	0.84	0.0045	0.00069	5.8×10^{-11}		
cg09257526	1: 154379696	<i>IL6R</i>	<i>ATP8B2</i> [†] (-)	0.0031	-0.0023	0.00035	5.9×10^{-11}		Alcohol consumption per day
cg15782984	6: 35993792	<i>SLC26A8</i>	<i>SLC26A8</i> (-)	0.007	-0.0046	0.00070	9.5×10^{-11}		

[†]The association between DNA methylation and the nearest gene was not significant (*FDR* > 0.05)

Sensitivity analyses indicate that ALS-associated differential methylation is not driven by genetic variation in *cis* or *trans*, riluzole use or *C9orf72* status

We performed sensitivity analyses to evaluate whether our results were driven by known biological factors associated with ALS or by genetic variation. First, we examined the effects of the *C9orf72* (C9) repeat expansion by performing an EWAS meta-analysis excluding 371 carriers of this mutation. Overall the results were highly correlated (**supplementary figure 8**), except for a subset of sites that were strongly driven by *C9orf72* carrier status. These included two genome-wide significant DMPs (cg01589155 and cg23074747) located within the C9 repeat and in a CpG island just upstream of the C9 repeat respectively. Second, to delineate whether DMPs were influenced by riluzole use, we performed an EWAS on riluzole use in ALS patients (N users = 1,803, N non-users = 451), finding no evidence of shared signals between the ALS EWAS and the riluzole EWAS (**supplementary figure 9**). Finally, we investigated whether results were driven by genetic variation. For each significant site we iteratively adjusted for all genetic variants in *cis* (<250 Kb) as detected in our overlapping whole-genome-sequencing data²⁶ (WGS, N = 7,939) and blood *trans*-mQTLs as reported in the GoDMC consortium³⁵. We found no evidence that the significant sites were driven by either genetic variants in *cis* or in *trans* (**supplementary figure 10**).

Enrichment analyses of genes annotated to ALS-associated differential DNA methylation implicate metabolic, inflammatory and cholesterol pathways

Gene set analysis. To characterize the EWAS results we performed gene set enrichment analyses based on both nearest genes and eQTLs annotated to each site^{36,37}. We considered both the default threshold used in the *methy*/GSA package ($P < 0.001$) and the stringent genome-wide significance threshold (9×10^{-8}) to select DMPs for enrichment analyses.

We identified two main categories of enriched pathways.

First, in both the LB and MOA results we identified cholesterol/steroid biosynthesis-related pathways. These included the KEGG pathway steroid biosynthesis and the GO pathways cholesterol biosynthetic process, sterol biosynthetic process, organic hydroxy compound biosynthetic process and secondary alcohol biosynthetic process which were enriched among the MOA results (**Table 3**). In addition, we found that these and related pathways were enriched among annotated eQTLs in both the LB and MOA results (**supplementary table 9**). The enrichments were mainly driven by four DMPs: cg17901584 (DHCR24), cg05119988 (MSMO1),

cg06500161 (ABCG1) and cg06690548 (SLC7A11). Of these, cg17901584, cg05119988 and cg06500161 were strongly associated with the expression of the nearest gene in blood (DHCR24, MSMO1 and ABCG1 respectively; **Table 2, supplementary table 8**).

Second, the immune-related KEGG pathways Cytokine-cytokine receptor interaction and natural killer cell mediated cytotoxicity were enriched in the LB results (at $P < 0.001$), but not in the MOA results (**Table 3**).

Table 3 - Gene set enrichments: Details of the gene sets that were significantly enriched among the MOA and LB results based on nearest genes annotated to each site. Method = EWAS method and p -value cutoff applied to the respective EWAS test-statistics resulting in the input probes for the shown enrichment analyses, N overlap = Number of significant genes that overlap with genes in the respective pathway, N genes = Total number of genes in the pathway, FDR = FDR-controlled (False discovery rate) P -values.

Method	Database	Pathway	N overlap	N genes	FDR
LB ($p < 0.001$)	KEGG	Cytokine-cytokine receptor interaction	36	262	0.0012
	KEGG	Natural killer cell mediated cytotoxicity	22	108	0.036
MOA ($p < 0.001$)	-	-	-	-	-
LB ($p < 9 \times 10^{-8}$)	-	-	-	-	-
MOA ($p < 9 \times 10^{-8}$)	KEGG	Steroid biosynthesis	2	18	0.015
	GO BP	cholesterol biosynthetic process	3	71	0.021
	GO BP	sterol biosynthetic process	3	77	0.021
	GO BP	organic hydroxy compound biosynthetic process	4	251	0.021
	GO BP	secondary alcohol biosynthetic process	3	71	0.021

EWASdb enrichments. To further characterize the results we assessed whether the significant sites overlapped with trait-associated sites reported in publicly available EWAS databases^{38,39}. For the LB results we found a significant overlap ($FDR < 0.05$) with 23 traits in the MRC-IEU database (**Figure 2a; Table 4**) and 20 traits in the NGDC database (**supplementary figure 11, supplementary table 10-11**), with a total of 23 out of 44 probes overlapping with one or more enriched traits. For the MOA results we found a significant overlap ($FDR < 0.05$) with 20 traits in the MRC-IEU database (**supplementary figure 12, supplementary table 10-11**) and 14 traits

in the NGDC database (**supplementary figure 13, supplementary table 10-11**), with a total of 8 out of 11 probes overlapping with one or more of the enriched traits.

Among the most significant enrichments in the MRC-IEU database (all results shown in **supplementary table 10-11**) were BMI, Total Serum IgE (only enriched among the LB results), (Serum) Triglycerides, Waist circumference and HDL-cholesterol, of which all showed effect directions opposite to those found for ALS, except for HDL (**Table 4**). Using the Louvain clustering algorithm⁴⁰, we found that the overlapping traits clustered into two (MOA) to three (LB) clusters respectively. These included two connected cholesterol (including HDL-c and triglycerides) and metabolism-related (including BMI and alcohol consumption) clusters which were identified in the results from both EWAS methods. Additionally, in the LB results we identified an inflammation-related trait cluster that included traits such as total serum IgE and atopy. We found that this inflammation-related cluster was independent of the other clusters, as indicated by iterative analyses presented in **figure 2b**, showing that only the immune-related traits remained significant after excluding BMI-related probes (**supplementary figures 11-13**).

Table 4 - EWAS database enrichments: Ten most significant trait enrichments within the MRC-IEU EWAS database. *FDR* = FDR-corrected *P*-values (False discovery rate). Effect directions = indicate whether the ALS EWAS and trait EWAS effect sizes share the same direction of effect (e.g. an opposite direction of effect for Body mass index indicates that DNA methylation changes at overlapping sites associated with a *lower* BMI are also associated with a *higher* ALS risk); EWAS method = indicates whether significant sites identified with respective method were enriched for the given trait.

Trait	<i>FDR</i>	Effect directions	EWAS method
Body mass index	1.36 x 10 ⁻⁹	opposite	LB & MOA
Total serum IgE	1.93 x 10 ⁻⁷	opposite	LB
Triglycerides	4.02 x 10 ⁻⁷	opposite	LB & MOA
Serum triglycerides*	1.32 x 10 ⁻⁵	opposite	LB & MOA
Waist circumference	1.85 x 10 ⁻⁴	opposite	LB & MOA
High-density lipoprotein cholesterol	0.0013	equal	LB & MOA
Hypertriglyceridemic waist	0.0024	equal	LB & MOA
Serum high-density lipoprotein cholesterol*	0.0066	equal	LB & MOA
Fasting glucose	0.0097	opposite	LB & MOA
Atrial fibrillation	0.011	opposite	LB & MOA

*note that we adhered to the trait descriptions as provided in the database: serum, plasma and whole-blood measurements are included as distinct traits ("Triglycerides" and "High-density lipoprotein cholesterol" refer to whole-blood measurements).

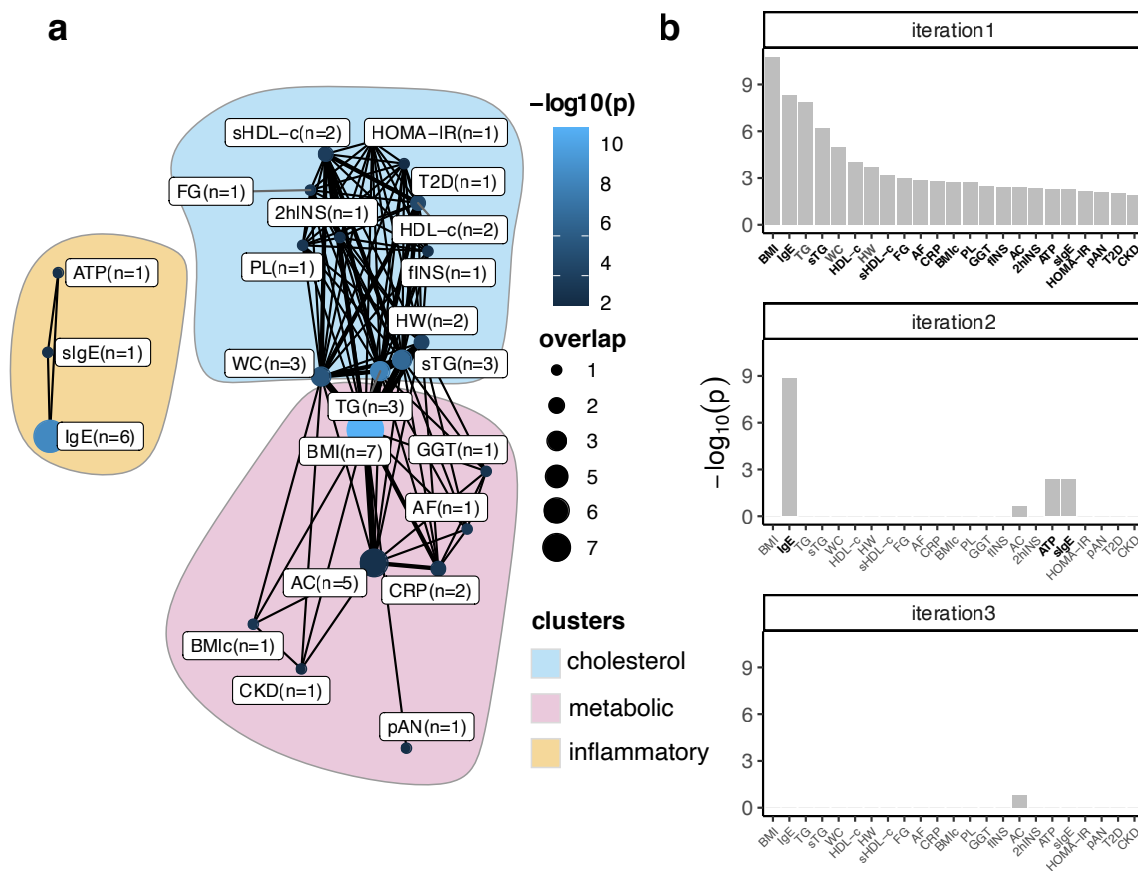


Figure 2 - EWAS database enrichments: Significant overlap between traits included in the MRC-IEU EWAS database and ALS-associated sites identified using the LB model **(a)** Network showing the traits that significantly overlap with the ALS-associated sites. Nodes indicate the overlap between ALS-associated sites and sites associated with indicated traits, with larger nodes indicating more overlap, and lighter shades of blue indicating stronger associations. Edges indicate probe overlap between the traits, with thicker lines indicating more overlapping probes. Colored surfaces indicate the clusters (cholesterol, metabolic and inflammatory) identified using the Louvain clustering algorithm. **(b)** Identification of independent clusters of traits. The first iteration shows the traits that significantly overlap with the ALS-associated probes at $FDR < 0.05$. In subsequent iterations the probes belonging to the most significant trait were excluded and enrichments tests were performed using the remaining traits. This algorithm was repeated, retaining traits that were nominally significant ($P < 0.05$, indicated in bold), until at most one trait remained significant. At the third iteration no traits remained significant, showing that both BMI and related traits (including triglycerides and HDL-c) and IgE and related traits (Atopy) show independent overlap with the ALS-associated sites.

Abbreviations: IgE = total serum IgE, TG = triglycerides, sTG= serum triglycerides, WC = waist circumference, sHDL-c = serum HDL-c, HW = Hypertriglyceridemic waist, FG = fasting glucose, AF =

atrial fibrillation, BMlc = BMI change, PL = postprandial lipemia, GGT = Gamma-glutamyl transferase, fINS=fasting insulin, AC = alcohol consumption per day, 2hINS = 2-hour insulin, ATP = atopy, sIgE = High serum IgE, pAN = Plasma adiponectin, T2D = Type II diabetes, CKD = Chronic kidney disease, HOMA-IR = Homeostatic Model Assessment of Insulin Resistance.

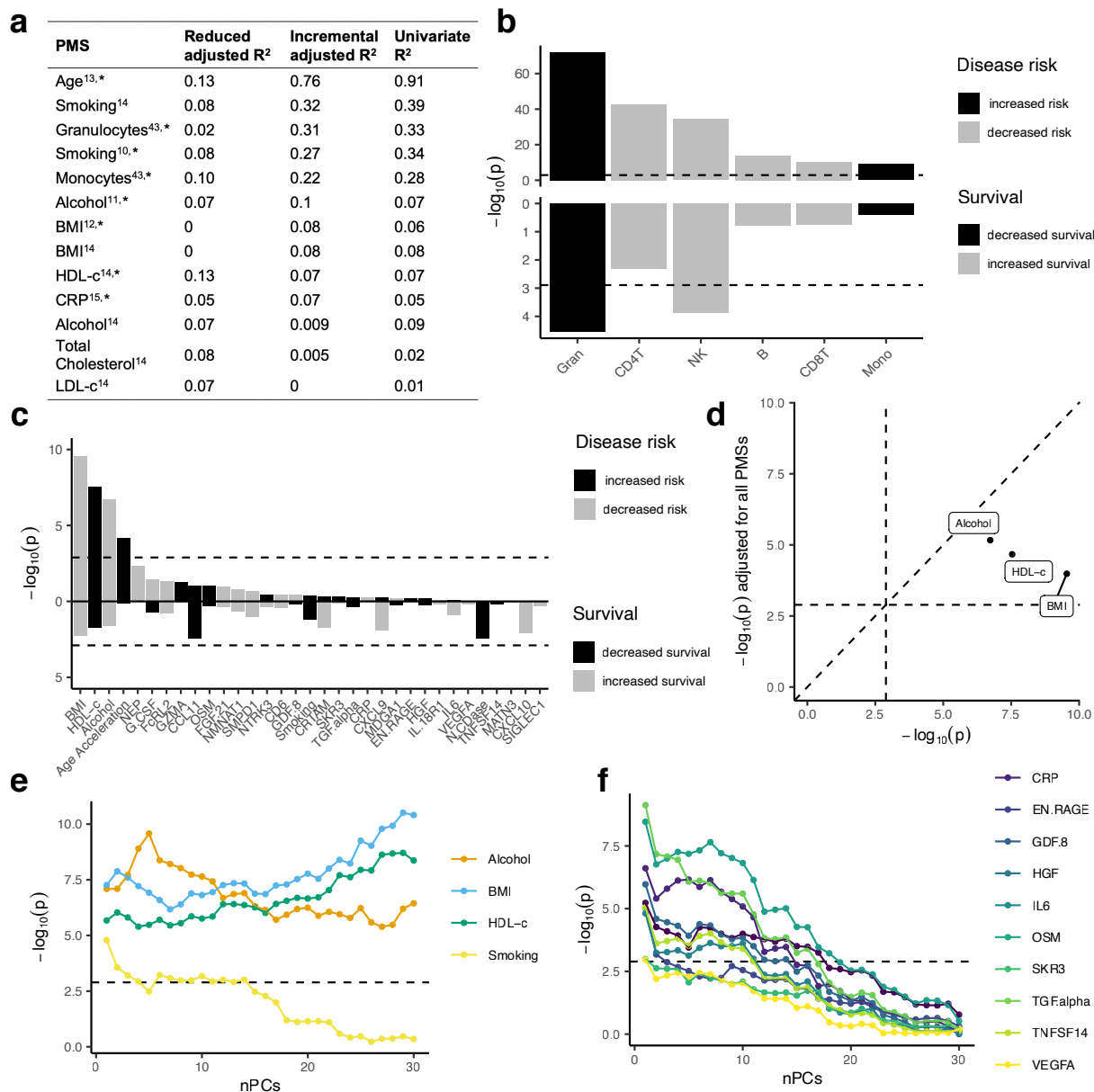


Figure 3 - Poly-methylation score analyses on disease risk and patient survival:

Poly-methylation scores (PMS) were determined as proxies for various traits, exposures, proteins and white blood cell proportions, calculated as weighted sums based on probes and weights derived from published papers respectively.

(a) Explained variance of PMSs calculated in samples for which both DNA methylation data and biomarker/clinical data were available (N=800/2000). Reduced R² represents the variance explained by

the null model while the incremental R^2 represents the additional variance explained by the PMS over the null model. Lastly, the explained variance of the univariate model of the respective PMS is displayed (see Methods). The asterisk indicates that the PMS was used in the association tests.

(b,c) The upper panel shows association P -values ($-\log_{10}(P)$, y-axis) for each PMS (x-axis), **(b)** white blood cell proportions and **(c)** various traits and exposures, colored by whether a higher score is associated with increased (black) or decreased (grey) disease risk. The lower panel shows the Cox proportional hazard P -values ($-\log_{10}(P)$, y-axis) for each PMS (x-axis), colored by whether a higher score is associated with decreased (black) or increased (grey) survival respectively. The dashed line indicates the significance threshold (1.3×10^{-3}).

(d) Original P -values ($-\log_{10}(P)$, x-axis) compared to P -values after including all PMSs in the logistic regression model ($-\log_{10}(P)$, y-axis) for the significant traits/exposures.

(e, f) Associations P -values ($-\log_{10}(P)$, y-axis) upon incrementally adding principal components (PCs) to the logistic regression model.

Poly-methylation scores for BMI, HDL-c, alcohol intake and white blood cell proportions are associated with ALS

To gain further insight into potential intermediate phenotypes associated with ALS, we utilized published poly-methylation scores (PMS) as proxies for various traits and exposures, including BMI, HDL-c, LDL-c, total cholesterol, alcohol consumption, smoking, white blood cell proportions (CD4T, CD8T, monocytes, granulocytes, NK-cells), biological age and a collection of immunological and neurological proteins^{11–14,16,41–43}.

First, we performed a validation analysis for each of the PMSs for which we had relevant clinical/exposure data available (see Methods, **supplementary table 4**). We selected PMSs with an explained variance of $\geq 5\%$, as indicated by an incremental R^2 between the null model (including known covariates and control probe PCs) and the model including the respective PMS (figure 3a). Two out of the thirteen validated PMS did not meet the implemented threshold of $\geq 5\%$ (LDL-c and total cholesterol).

We found that PMSs for HDL cholesterol, monocyte cell proportion and granulocyte cell proportion were positively associated with ALS ($P < 1.3 \times 10^{-3}$; **Figure 3 b-c, supplementary table 12**), and the PMSs for alcohol intake, BMI and the other white blood cell proportions (CD4T, CD8T, NK, B-cells) were negatively associated with ALS, a result that reflects the nature of proportion data given the positive associations of other cell types ($P < 1.3 \times 10^{-3}$; **Figure 3 b-c, supplementary table 12**). Although we did find a significant association for epigenetic age acceleration (Zhang *et al.* clock¹³, adjusted for chronological age), there was significant

heterogeneity between strata (Cochran's Q test $P < 0.1/39$), **supplementary table 12**), which led us to exclude age acceleration for further consideration. Additionally, we considered the multi-tissue Horvath clock⁴⁴ and the Hannum *et al.* clock⁴⁵, but found no significant associations for either (**supplementary figure 14**).

In addition to the PC-adjusted models, we also evaluated less stringent models, showing that various immunological and neurological proteins such as CRP, IL6, TGF- α and CCL11 as well as smoking were significantly associated with ALS when PCs were excluded (**Figure 3e-f**, **supplementary figure 15**).

Conditional analyses showed that the significant PMSs were independently associated with ALS, although the HDL-c and BMI associations were attenuated after mutual adjustment (**Figure 3d**, **supplementary figure 16**). Adjustment for significant EWAS sites showed that signal is shared between several CpG sites and significant PMSs (**supplementary figure 17**), most notably, the alcohol intake association became insignificant upon adjustment for cg06690548 (*SLC7A11*) and cg18120259 (*C6orf223*), and the HDL-c association became insignificant upon adjustment for cg17901584 (*DHCR24*) and cg06500161 (*ABCG1*). We assessed whether the associations were primarily driven by carriers of the *C9orf72* repeat expansion, but found no evidence that this was the case (**supplementary figure 18**). Finally, we evaluated whether the PMS associations were primarily driven by specific strata or experimental batches by performing leave-one-out analyses. We found no evidence that one stratum or experimental batch mainly contributed to the observed signals, and as expected, the largest decrease in significance was seen when leaving out the larger experimental batches and strata (**supplementary figure 19-22**).

Survival analyses indicate that white blood cell proportions and DNA methylation at five ALS-associated DMPs are associated with disease progression

We performed multivariate Cox PH analyses on the 45 ALS-associated DMPs identified using the MOA and LB models. Test-statistics were combined using inverse variance-weighted fixed effects meta-analysis and checked for heterogeneity (Cochran's Q test $P < 0.1/45$). A total of five probes showed a significant association with survival after correcting for known confounders and PCs ($0.05/45 = P < 1.11 \times 10^{-3}$) and cross-validation between three sensitivity analyses. Effect sizes were moderate and showed both shorter and longer survival time between DNA methylation and overall survival (**supplementary table 13**).

All reported sites were not significantly affected by the addition of time-varying effects in the Cox PH model or by applying a restricted cubic spline with varying complexity to model the baseline log cumulative hazard (**supplementary figure 23**). Moreover, after adjusting for *C9orf72* carrier status in the multivariate Cox PH model, the significant sites (besides the *C9orf72* mapped probe) remained significantly associated with survival (**supplementary figure 23**). Four sites showed a significant eQTM effect with *FKBP5*, *ATP8B2*, *SPIDR*, and *DHCR24* (**Table 5**).

Table 5 - Significant sites associated with survival: Details of the sites significantly associated with survival. Position = Chromosome:bp (GRCh37), Nearest gene = nearest gene based on Ensembl GRCh37 (75), eQTM = the most significant eQTM for the respective probe, eQTM *FDR* = *p*-value corresponding to the most significant eQTM, FDR-corrected for the number of tests for the respective probe, PMS = Probe is part of the respective PMS (poly-methylation score), HR = Hazard Ratio, Trait = Overlap with significantly enriched traits from the MRC-IEU and NGDC EWAS databases (showing a maximum of five traits). Abbreviations: HGF = Hepatocyte growth factor.

Probe	Position	Nearest gene	eQTM (direction)	eQTM <i>FDR</i>	HR	P-value	PMS	Traits
cg14195992	8:48265917	<i>SPIDR</i>	<i>SPIDR</i> (-)	0.0059	0.074	4.7×10^{-7}		
cg03546163	6:35654363	<i>FKBP5</i>	<i>FKBP5</i> (-)	0.016	0.19	2.7×10^{-5}	HDLchol	Body mass index, Waist circumference, Alcohol consumption per day, Chronic kidney disease
cg09257526	1:154379696	<i>IL6R</i>	<i>ATP8B2</i> [†] (-)	0.0031	0.0048	1.3×10^{-5}		Alcohol consumption per day
cg17901584	1:55353706	<i>DHCR24</i>	<i>DHCR24</i> (-)	2.9×10^{-62}	4.62	1.0×10^{-5}	BMI, HDLchol, HGF	Hepatic fat, BMI, Metabolic trait, (serum) Triglycerides
cg01589155	9:27573532	<i>C9orf72</i>			46.99	2.0×10^{-4}		

[†]The association between DNA methylation and the nearest gene was not significant (*FDR* > 0.05)

We also assessed whether the PMSs were associated with survival, finding that a higher proportion of granulocytes was significantly associated with decreased survival and a higher proportion of natural-killer cells was associated with increased survival ($P < 1.3 \times 10^{-3}$; lower panels **Figure 3b-c, supplementary table 14**). These results were robust in different models: a Cox PH model stratified on experimental batch and with time-varying effects and a restricted cubic spline to model the baseline log cumulative hazard (**supplementary figure 24-25**). Finally, the associations were not affected upon adjustment for C9orf72 status (**supplementary figure 26**).

Discussion

Our study represents the largest case-control EWAS of any disease performed to date. Our unique dataset includes genome-wide DNA methylation data on over 10,000 individuals, with extensive clinical data and whole genome sequencing data available for the majority of samples. Following thorough quality control and extensive sensitivity analyses, we identified a total of 45 DMPs at which DNA methylation was significantly and robustly associated with ALS, with the majority representing novel findings not previously reported. Both gene set and trait enrichment analyses revealed enrichments for metabolic, cholesterol and inflammatory pathways. These findings were corroborated using a hypothesis-driven approach that leveraged published DNA methylation-based proxies for risk factors and exposures relevant to ALS. These analyses found that PMSs for BMI, HDL-cholesterol were independently associated with ALS. Moreover, we found a strong association between white blood cell PMSs and ALS, and show that an increased proportion of granulocytes is associated with worse patient survival. Finally, we show that several of the DMPs associated with ALS were also associated with survival, and could therefore be of interest as and could represent indicators of underlying disease processes.

By utilizing enrichment analyses, poly-methylation scores and survival analyses we highlight various pathways and potential disease-modifiers in ALS.

First, genes annotated to DMPs were enriched for pathways related to cholesterol biosynthesis. The main drivers of these enrichments include cg17901584 (*DHCR24*), cg06500161 (*ABCG1*), cg05119988 (*MSMO1*), and cg06690548 (*SLC7A11*). Of these, we found that DNA methylation levels at the first three sites were significantly associated with gene expression in blood. These genes are all involved in cholesterol biosynthesis and lipid transport, and DNA methylation at these sites has been robustly linked to HDL- and total cholesterol, triglyceride levels and BMI-related traits such as diabetes and hepatic fat levels^{46–52}. Moreover, both cg17901584 (*DHCR24*) and cg06500161 (*ABCG1*) are part of the HDL-c PMS, and explain a considerable part of the significant association we found between elevated HDL-cholesterol levels and ALS.

Interestingly, cg06690548 (annotated to *SLC7A11*) has also been previously associated with alcohol intake and related factors such as GGT and phosphatidylethanol levels^{11,53–55}, and the association between the alcohol PMS and ALS was primarily driven by this site. Although we did not find a significant DNA methylation-gene expression relationship (eQTM) for this site,

previous work suggests that increased DNA methylation at cg06690548 is associated with downregulation of *SLC7A11* in the brain. Interestingly, *SLC7A11* encodes xCT, a cystine-glutamate antiporter that imports cystine while exporting glutamate⁵⁶, the former being an essential precursor of glutathione, the major antioxidant in the brain. It is possible therefore, that the association found in *SLC7A11* -- and by extension the alcohol PMS -- is related to glutamate excitotoxicity and/or oxidative stress, rather than reflecting alcohol use.

Second, both the EWAS trait enrichments and PMS analyses indicate that lower BMI levels are associated with ALS. Importantly, the BMI association remained significant after adjustment for other PMSs, including those of HDL-c and alcohol intake, although these PMSs are not perfect proxies of the respective covariates. Lower BMI has been consistently reported in ALS, and may be explained by several factors including hypermetabolism⁵⁷, increased levels of physical activity prior to disease onset⁵⁸ or muscle loss and eating difficulties after onset of disease. It is also interesting to note that our findings support the hypothesis that ALS is associated with a favorable cardiovascular profile^{59,60}. Supporting findings include lowered BMI and higher HDL-c levels in ALS patients as indicated by the PMS analyses, and opposite effects at overlapping sites between ALS and traits such as BMI, triglycerides, waist circumference, fasting glucose, and type II diabetes as indicated by the trait enrichment analyses. However, since our study cannot reliably distinguish whether phenomena occur prior to or after onset of disease, we need to be cautious to conclude that a favourable cardiovascular profile is a risk factor for developing ALS. The Mendelian randomization (MR) analyses in our accompanying GWAS study²⁶ do not support a causal role for cardiovascular traits including BMI, triglycerides, blood pressure, smoking, physical activity, BMI, coronary artery disease, or stroke, but do support a causal role for cholesterol. This shows that the causal role for cholesterol in ALS might be independent from a potential cardiovascular mechanism, for example through the interplay between lipid levels and autophagy⁶¹ as illustrated by a recent study showing that high cholesterol levels lead to increased protein aggregation through autophagy impairment in mouse models of Alzheimer's disease⁶². Whether the other independently related traits (BMI and alcohol intake) are causally associated with ALS, or are indicators for ALS pathophysiology, needs further study. These relations can be complex, for example we recently showed that BMI, smoking, alcohol intake and physical activity can have a genotype and time dependent relationship with ALS⁶³.

Third, our results point towards a role of the immune system in ALS. The EWAS results were enriched for immune-related traits including IgE and allergic sensitization, and importantly, these

results were independent of predicted white blood cell proportions. Sites driving these enrichments included, among others, cg06528816 (*TTC7A*) and a cluster of three probes in the *ZFPM1* gene, both implicated in immune-related traits such as IgE⁶⁴, asthma^{65,66} and allergic sensitization⁶⁷. Our PMS analyses corroborate the role of immunity in ALS as we found that white blood cell proportions were altered in ALS, with a higher ratio of granulocytes and a lower ratio of lymphocytes in ALS patients (CD4T-cells, CD8T-cells, and NK-cells). Interestingly, a recent study showed that white blood cell alterations are shared between neurodegenerative diseases (Alzheimer's, Parkinson's disease and ALS)⁶⁸. These findings indicate that WBC alterations are not specific to ALS and are therefore more likely to be a consequence of disease, and thus their value is more likely related to disease activity rather than etiological. Indeed, in our study we found that increased granulocyte proportions are associated with worse prognosis, whereas NK-cell proportions are associated with better prognosis, indicating that WBC proportions have prognostic value. The role of immunity is further supported by our observation that various inflammatory protein PMSs such as CRP, IL6, TGF- α and CCL11 were elevated in ALS patients, although these differences became insignificant upon adjustment for principal components. Our findings are in line with previous studies that identified higher ratios of neutrophils and/or granulocytes to lymphocytes in ALS patients^{69,70}, elevated levels of inflammatory proteins⁷¹, and an association between higher neutrophil proportions and worse prognosis⁷². Although immune alterations could be part of a systemic aspect of ALS, there is evidence that suggests that the peripheral immune system contributes to neuroinflammation, the latter being an established phenomenon in ALS as well as other neurodegenerative diseases^{73,74}. Especially interesting in this regard is that recent evidence shows that mast cells infiltrate skeletal muscles at the neuromuscular junction and degranulate to help recruit neutrophils^{75,76}, which prevent reinnervation capacity and may thus be a potential mechanism causing worse prognosis. In line with this, we identified an enrichment for IgE (and related traits such as allergy and atopy), which activate mast cells, and found that increased proportions of granulocytes were associated with ALS and patient survival. Thus, these findings could be of interest for new treatments, especially given that mast cell activity can be influenced⁷⁷.

Finally, 8 out of the identified 45 sites were reported in a recent study on shared DNA methylation alterations across Parkinson's disease, Alzheimer's disease and ALS⁶⁸. Of these ALS-associated DMPs we identified associations between DNA methylation and survival within 5 of the 45 DMPs annotated to FKBP5, ATP8B2, SPIDR, and DHCR24. Interestingly, the majority (4 out of 5) of sites that we found to be associated with disease progression were

among the overlapping sites. It could therefore be speculated that these sites represent shared pathways involved in neurodegeneration and could therefore have clinical utility.

Several recent findings are relevant to our current study.

First, the largest ALS EWAS up to this date²⁵ has reported that white blood cell proportions were associated with ALS. We replicate these findings, and additionally show that WBC proportions are associated with prognosis.

Second, the DNA methylation changes that we identified in the CpG island just upstream of the *C9orf72* G₄C₂ repeat and within the repeat itself have been previously reported in carriers of the *C9orf72* repeat expansion^{78–80}. In line with this, we found that these associations were driven by the *C9orf72* carriers in our data.

Third, we do not replicate the recently reported association between epigenetic age acceleration and survival⁸¹. Importantly, in our analyses we adjusted for sampling age, as it has been shown to be crucial when studying epigenetic age acceleration⁸², especially given that age of onset affects disease progression in ALS⁸³. As we show in **supplementary figure 27** both survival and age of onset were significantly associated with age acceleration when sampling age was not accounted for, but became insignificant upon adjustment. Additionally, in our case/control analysis we observed substantial heterogeneity among strata (Cochran's Q test $P = 5.8 \times 10^{-4}$), hence our results do not support a unambiguous role for age acceleration.

One limitation of this study is that the cross-sectional nature of our study hinders inferences about causality. Mendelian randomization analyses, presented in our accompanying GWAS, can aid in identifying causal relations in observational data, as evidenced by the finding that cholesterol levels are causally related to ALS. The GWAS MR analyses (SMR⁸⁴) did not find significant evidence for a causal role of the sites identified in this study. Although this may reflect a lack of power, it could indicate that the results reflect the consequences of disease processes rather than causal mechanisms. In that case, the value of the identified DNA methylation changes would lie primarily in revealing underlying disease processes in ALS. Furthermore, our survival analyses show that the results could be of interest as potential starting points for new disease-modifying treatments. Future studies could further investigate causality in longitudinal samples, which although scarce and often small, are expected to become more readily available with the increasing number of larger-sized prospective cohorts.

We further note that we collected DNA from whole blood rather than from brain tissue, given that our aim was to identify peripheral biomarkers reflecting underlying disease processes and traits/exposures related to disease. In contrast to brain tissue, blood DNA methylation is easily accessible and allows for sampling close to disease onset. Combined with the ever growing body of literature on, and biomarkers of, various traits and exposures, it makes blood DNA methylation ideal for biomarker purposes⁸⁵. However, given the tissue-specificity of DNA methylation, further studies are needed to assess ALS-associated DNA methylation changes in the brain.

Finally, we note that although the stringent adjustment for confounding we applied by using PCs and random effects models (i.e. OSCA²⁸) is key in combating test statistic inflation in EWAS, it may have obscured biological signals of interest. For example, we show that the recent OSCA MOA algorithm results in a substantially lower (but overlapping) number of identified DMPs compared to the often used method of including principal components followed by adjustment for inflation/bias in test-statistics (termed the LB algorithm in this study). Interestingly, our results indicate that the additional DMPs identified using the LB algorithm are enriched for inflammatory pathways and traits, which corroborates previous findings that suggest that uncaptured variation can be explained by cell type heterogeneity and related immune processes^{86,87}. Similarly, we show that various immunological proteins such as CRP, IL6, TGF- α and CCL11 became insignificant upon PC adjustment. More generally, this relates to the discussion on whether to treat variables such as cell type proportions as nuisance variables in an EWAS, or view them as variables that provide valuable information in themselves^{88,89}. In this study we therefore struck a balance by opting for a two-way approach, combining a stringently corrected EWAS with a more targeted approach where we studied “confounders” such as WBC proportions, smoking and BMI as outcomes of interest, assessing them with both stringent (i.e. including PCs) as well as more lenient models.

To conclude, we present the largest case-control DNA methylation study to date, employing a comprehensive approach integrating methylomic, genomic and clinical data. We identify 45 differentially methylated sites, and by utilizing enrichment analyses, poly-methylation scores and survival analyses we highlight a potential disease-modifying role for caloric intake, cholesterol metabolism and inflammation. In our accompanying GWAS paper, we show that cholesterol

levels are causally related to ALS. Regardless of the exact etiologic role of the other identified pathways, these constitute potential targets for intervention in ALS patients.

Methods

Cohorts

We obtained DNA methylation and phenotypic/lifestyle data from 10,462 individuals (cases N = 7,344, controls N = 3,118) in fourteen different cohorts. Patients were diagnosed with definite, probable, and probable lab-supported ALS according to the revised El Escorial Criteria⁹⁰.

Population-based controls were matched for age, sex, and geographical region in a 1:2 ratio and not screened for (subclinical) signs of ALS. Detailed descriptions of each cohort are provided in the **Supplementary Note**. Experimental batches were processed in the same lab and sequenced in the same series, resulting in 44 independent batches after quality control. Strata for analyses were defined as samples within the Project MinE sequencing consortium separated by array technology (MinE 450k and MinE epic) and Australian data separated by signal intensities (AUS 1 and AUS 2).

DNA methylation profiling

For the Project Mine samples, venous blood was drawn from patients and controls from which genomic DNA was isolated using standard methods. We set the DNA concentrations at 100ng/l as measured by a fluorometer with the PicoGreen© dsDNA quantitation assay. DNA integrity was assessed using gel electrophoresis. Genomic DNA (~1 µg) was bisulfite-treated at a central site using Zymo Bisulfite Conversion Kits (Zymo Research, Orange, CA). DNA methylation was analyzed using the Infinium Methylation450k array (N=4,903) or Infinium EPIC array (N=4,300), according to the standard Infinium HD array methylation protocol (Illumina, San Diego, CA, USA).

The AUS ALS datasets were profiled as described in ref²⁵.

QC & Normalization

Sample QC. Quality control and normalization was performed separately for the four strata (MinE 450k, MinE EPIC, AUS1 and AUS2). The following metrics were used to exclude samples, the exact thresholds used for each stratum are listed in **supplementary table 1**, and the number of samples failed on each metric are listed in **supplementary table 2**. QC figures for each stratum are provided in **supplementary file 1**.

1. Median methylated and unmethylated intensities.
2. Median red/green intensity ratios calculated in type I probes.
3. Discordance between reported sex and predicted sex based on the *getSex* function in the *minfi* R package⁹¹.
4. The OP (non-polymorphic controls) and Hyb (hybridization controls) metrics as implemented in the *MethylAid* R package⁹².
5. Bisulfite conversion rate based on the *bscon* metric as implemented in the *wateRmelon* R package³¹.
6. Fraction of probes with high detection *P*-values and/or measured by a low number of beads.
7. Samples that failed on the inbreeding metric in the corresponding whole-genome-sequencing (WGS) data. Quality control of the Project MinE WGS data was performed as described earlier⁹³.
8. Genotype concordance. Briefly, we used the *omicsPrint* R package to select 413 probes that reliably measured underlying SNPs and were present in the WGS-derived SNP data⁹⁴. We performed identity-by-state (IBS) between the DNAm-inferred SNPs and the WGS-derived SNP data using the *allelesharing* function.
9. After removing samples that failed on any of the steps listed above, we performed PCA on the control probes present on the array. Samples that had values larger than 3.5 standard deviations of the mean on the first two principal components (PCs) were excluded.

Normalization. After quality control, signal intensities for all strata except AUS1 were normalized using the *dasen* function as implemented in the *wateRmelon* package³¹. Since the signal intensities in the AUS1 stratum showed dye bias (**supplementary file 1: QC figure 50**), this stratum was normalized using the *nanex* function, which is similar to the *dasen* function, except that it corrects for dye bias prior to normalizing the signal intensities.

Probe Filtering. After normalization, we set all the measurements with detection *P*-value $> 1 \times 10^{-16}$ or measured by <3 beads to missing⁹⁵. Within each stratum we then removed probes with $>5\%$ missing data.

Post-normalization filtering. After sample QC, normalization and probe filtering, we further excluded samples based on the following criteria (**supplementary table 3**):

1. Discordance between chronological (reported) age and predicted age¹³. In each stratum we regressed DNA methylation age on chronological age, and excluded samples whose residuals from this regression were more than 4.5 standard deviations from the mean.
2. We performed principal component analyses on the normalized β -values. Samples that had values larger than 4 standard deviations from the mean on the first two PCs were excluded.
3. Samples with non-MND diagnosis or missing phenotype info were excluded.
4. For each related pair of individuals (identical or first-degree) we excluded one individual to obtain a set of unrelated individuals.

EWAS

We used two approaches to test for an association between disease status and variable DNA methylation while controlling for known and unknown confounding factors.

Principal components + bacon (LB model). We performed linear regression at each site, adjusting for sex, experimental batch, predicted age, predicted white blood cell (WBC) fractions, 30 control probe PCs and a variable number of PCs derived from all measured sites. PCs were calculated by regressing the β -values on the covariates listed above, followed by performing principal component analysis on the residuals from this regression. We optimized the number of PCs included in each stratum by evaluating the sample-size normalized inflation factors (λ_{1000})⁹⁶:

$$\lambda_{1000} = 1 + (\lambda_{obs} - 1) * \left(\frac{1}{n_{cases}} + \frac{1}{n_{controls}} \right) / \left(\frac{1}{n_{cases,1000}} + \frac{1}{n_{controls,1000}} \right)$$

after employing the following linear regression at each site:

$$DNAm_i = Phenotype + Predicted\ Age + Sex + Batch + WBC_{est} + PC1 - 30_{ctrl-probes} + PC1 - m$$

where m is in $\{0,5,10,15,20,25,30\}$.

The number of PCs were chosen so that for each stratum $\lambda_{1000} \leq 1.15$. We then corrected for remaining inflation and/or bias in test-statistics of each stratum using *bacon* (implemented in the *bacon* R package)²⁷. Hereafter we refer to this model (linear model followed by *bacon*) as the LB model.

Mixed linear model. We used a mixed-linear model approach as implemented in the MOA algorithm included in the OSCA software (v0.45)²⁸. Briefly, this algorithm tests for an association between the methylation status of a CpG-site (β -value) and a trait (case/control status) while fitting all the other probes as random effects. This method is based on the assumption that including distal probes as random effects accounts for correlations induced by (unobserved) confounding factors. In addition to the random effects, we included sex, predicted age and experimental batch as fixed effects.

Meta-analysis. Test-statistics across strata were combined using an inverse variance-weighted fixed-effects meta-analysis as implemented in the *metagen* function in the *meta* package²⁹. We assessed heterogeneity using Cochran's Q test, and considered sites significantly heterogeneous if Cochran's Q test *P*-value (corrected for the number of significant sites) was < 0.1. Sites where $P < 9 \times 10^{-8}$ were considered genome-wide significant, as recommended for the EPIC array based on an empirical estimate of the independent number of tests³³.

EWAS sensitivity analyses

We performed various sensitivity analyses to assess whether the EWAS results were robust to varying analysis strategies or driven by measured biological factors.

Technical Sensitivity Analyses. The following technical sensitivity analyses were performed:

1. We performed an EWAS on *M*-values (defined as $\log_2(\frac{\beta}{1-\beta})$) instead of β -values.
2. We performed an EWAS including only the Project MinE samples (N=8371; excluding the external Australian samples).
3. We performed an EWAS where we normalized the signal intensities using functional normalization³⁰ instead of *dasen* and *nanos*³¹.
4. We performed an EWAS where we left out experimental batches that included only ALS patients and no controls.

Biological Sensitivity Analyses. The following biological sensitivity analyses were performed:

1. To assess whether results were influenced by riluzole use, we performed an EWAS on riluzole usage (yes/no) in 2,254 ALS patients for which data was available. These included 1,803 patients that used riluzole and 451 patients that did not use riluzole at time of blood draw respectively.
2. To assess whether results were driven by patients carrying the *C9orf72* (C9) repeat expansion (the most common genetic cause of ALS), we performed an EWAS where we excluded individuals carrying the C9 repeat expansion. We restricted the analysis to individuals for which C9 status was available, and who did not carry the C9 repeat expansion (N = 7,839, removed 371 C9 carriers; based on a ≥ 30 repeat cutoff as determined by ExpansionHunter⁹⁷). We compared this C9-negative EWAS to an EWAS including C9-carriers where we randomly downsampled the number of samples to match the sample size of the C9-negative EWAS.
3. To assess whether the results were driven by genetic variation, we employed the available whole genome sequencing (WGS) data (available for 7,939 samples). We considered two sets of genetic variants: (1) all variants (SNVs) in *cis* of the CpG-site (<250Kb) and (2) variants reported as *trans*-mQTLs in blood for the respective CpG in the GoDMC consortium atlas of genetic effects (for 11 DMPs at least one *trans*-mQTL was reported)³⁵. For each SNP (*j*) annotated to CpG-site (*i*), we ran the following regression:

$$DNAm_i = \text{Case/control status} + \text{Predicted Age} + \text{Sex} + \text{Batch} + WBC_{est} + PC1 - 30_{ctrl-probes - m} + PC1 - m + SNP_j$$

Where $m = 30$ for the MinE 450k stratum and $m = 15$ for the MinE EPIC stratum.

Probe filtering

Probes were filtered based on the following criteria:

1. A ≥ 14 bp 3'-subsequence inexact match to the C9 repeat expansion³²
2. A ≥ 30 bp 3'-subsequence inexact off-target match to the reference genome.
3. Low mapping quality based as determined by Zhou *et al.*⁹⁸.

4. Given that we have genetic data available for most samples, we tested whether significant probes were driven by nearby (<250Kb) genetic variants, instead of removing probes containing SNPs *a priori*. Non-significant probes (used as background for enrichment analyses) were filtered based on containing a SNP with MAF > 0.01 (European population) within 5bp of the 3'-end of the probe⁹⁸.

Gene mapping & eQTM

We mapped CpG-sites to the nearest protein-coding gene within 250kb based on Ensembl (GRCh37).

We tested whether the DNA methylation levels of the significant CpG-sites were associated with the expression of nearby genes in blood using an external dataset (BBMRI; <https://www.ebi.ac.uk/ega/studies/EGAS00001001077>) for which gene expression (RNAseq) and DNA methylation data (450k array) are available for 3,251 samples³⁴. For each CpG-site, we tested for an association between DNA methylation and gene expression levels of nearby genes (<250kb), correcting for age, sex, strata, white blood cell composition and 20 PCs (10 PCs derived from gene expression data, and 10 PCs derived from the DNA methylation data). We corrected the test-statistics for bias and inflation (estimated based on the association between DNA methylation and expression levels of all genes, using the R package *bacon*²⁷).

Enrichment analyses

Gene Set Analysis. GO and KEGG enrichment analyses were performed using the *methygometh* function in the *methyGSA* R package^{37,99}. This method takes into account that the number of CpGs assigned to each gene differs by accounting for the probability of a gene being selected using Wallenius' noncentral hypergeometric distribution. We restricted the analysis to GO/KEGG categories that included at least 10 and at most 500 genes. All tested probes were used as background, and the *P*-values were adjusted for the number of categories tested using the Benjamin-Hochberg method. We considered both the default threshold used in the *methyGSA* package ($P < 0.001$) and the stringent genome-wide significance threshold (9×10^{-8}) to select DMPs for enrichment analyses. Enrichment analyses were performed using both nearest gene annotations as well as eQTM annotations. In the eQTM enrichment analyses we restricted the analysis to sites that had a significant eQTM-association after accounting for

multiple testing using a two-step approach as described previously¹⁰⁰. This resulted in 79,441 eQTM-annotated sites that were considered for the enrichment analysis.

EWASdb enrichments. We tested whether the identified sites significantly overlapped with CpG-sites associated with other traits using publicly available EWAS databases (EWASdb³⁹ and the MRC-IEU EWAS Catalog³⁸). For each trait present in the EWAS databases, we overlapped the trait-associated CpG-sites with the ALS-associated CpG-sites. Fisher's exact test was used to test for a significant enrichment ($FDR < 0.05$) relative to a background consisting of all probes included in the database. The resulting P -values were adjusted for the number of traits tested using the Benjamin-Hochberg method. We limited the analysis to traits that were associated with more than 5 and less than 5,000 CpGs and that were studied in blood-related tissues. Since the great majority of the EWASs in these databases are based on the 450k array, we limited the analyses to 450k array probes.

The Louvain algorithm as implemented in the *igraph* R package was used to identify clusters among the significantly enriched traits^{40,101}. We further identified independent clusters of traits by removing the probes linked to the most significant trait, followed by retesting for enrichment among the remaining probes. We repeated this algorithm, retaining traits that were nominally significant ($P < 0.05$), until at most one trait remained significant.

Poly-methylation scores

Derivation of poly-methylation scores. We considered the following poly-methylation scores: smoking^{14,42}, alcohol intake^{11,14}, BMI^{12,14}, HDL-c¹⁴, LDL-c¹⁴, total cholesterol¹⁴, CRP¹⁵, white blood cell proportions¹⁰² (CD8T-cells, CD4T-cells, NK-cells, Monocytes, Granulocytes, and B-cells), DNA methylation age^{13,44,45}, and a collection of 27 immunological and neurological plasma proteins^{16,103}.

The WBC proportions were estimated using the robust partial correlations (RBC) method as implemented in the *EpiDISH* R package. The Zhang *et al.* age predictions were calculated using the scripts provided by the authors (<https://github.com/qzhang314/DNAm-based-age-predictor>). The Horvath age predictions were calculated using the supplementary code⁴⁴. The *EpiSmokEr* smoking PMS were calculated using the *EpiSmokEr* R package (<https://github.com/sailalithabollepalli/EpiSmokEr>). The Liu *et al.* alcohol PMS were calculated using the *dnamalci* R package (<https://github.com/yousefi138/dnamalci>)⁴¹.

For all other PMSs, we downloaded the respective coefficients and calculated the PMS by multiplying each coefficient by the respective DNA methylation levels and summing these values:

$$PMS = \sum_i^m \beta_i \times CpG_i$$

Validation of poly-methylation scores. Intermediate phenotypes were validated in a Dutch cohort of approximately 2000 individuals (2:1 case/control ratio) of which measured biomarker data was collected for 800 individuals (**supplementary table 4**). Collected biomarker data <1 week and phenotypic data < 52 weeks after the collection of blood for DNA methylation profiling were excluded from all analyses. Furthermore, CRP values >100 mg/L were considered an indicator of significant infection¹⁰⁴ and were therefore not included for validation of the poly-methylation scores (PMS).

Linear regression models were used to identify the proportion of phenotypic variance explained by the corresponding PMS. Diagnosis, predicted age, sex, batch, cellcounts and 30 control probe PCs were considered as covariates in the null model, whereby the phenotypic measure was the dependent variable, if available:

$$Phenotype = PMS + Predicted\ Age + Sex + Batch + WBC_{est} + PC1 - 30_{ctrl-probes}$$

Incremental R² estimates were calculated between the null model and the models with the PMS of interest. The incremental R² estimates were used to determine whether the PMS increased the predictive ability above and beyond that provided by an existing model. Therefore, our analysis was limited to PMS with an incremental R² of at least 0.05 when intermediate phenotype/biomarker data was available.

Intermediate phenotypes/biomarker data were available for age (years), cigarettes (self-reported number of cigarettes in a year), alcohol (self-reported units of alcohol a week), Granulocytes (absolute counts), Monocytes (absolute counts), BMI (kg/m²), CRP (mg/L), total cholesterol (mmol/l), LDL-cholesterol (mmol/l), HDL-cholesterol (mmol/l).

Association testing. For each stratum, we tested for an association between the PMS and case/control status using logistic regression, the following regression models were used for HDL-c, BMI, alcohol intake, smoking, CRP and the immunological/neurological proteins:

$$\text{Case/control status} = \text{PMS} + \text{Predicted Age} + \text{Sex} + \text{Batch} + \text{WBC}_{est} + \text{PC1} - 30_{ctrl-probes} + \text{PC1} - m$$

Where $m = 30$ for the MinE 450k stratum, $m = 15$ for the MinE EPIC stratum, $m = 25$ for the AUS2 stratum and $m = 30$ for the AUS2 stratum.

for DNA methylation age we additionally adjusted for age at blood draw, thus representing age acceleration⁸²:

$$\text{Case/control status} = \text{Predicted Age} + \text{Age at blood draw} + \text{Sex} + \text{Batch} + \text{WBC}_{est} + \text{PC1} - 30_{ctrl-probes} + \text{PC1} - m$$

Where $m = 30$ for the MinE 450k stratum, $m = 15$ for the MinE EPIC stratum, $m = 25$ for the AUS1 stratum and $m = 30$ for the AUS2 stratum.

For the white blood cell proportions we did not adjust for array-wide principal components, given that the top principal components essentially represent white blood cell proportions⁸⁷:

$$\text{Case/control status} = \text{WBC} + \text{Predicted Age} + \text{Sex} + \text{Batch} + \text{WBC}_{est} + \text{PC1} - 30_{ctrl-probes}$$

Strata test-statistics were combined using an inverse-variance-weighted, fixed-effects meta-analysis using the *metagen* function in the *meta* R package²⁹. We corrected for the number of PMSs tested using the Bonferroni correction ($0.05/39 = P < 1.3 \times 10^{-3}$).

Survival. We used a multivariate Cox proportional hazards regression model to test for an association between survival and PMSs, adjusting for predicted age, sex, batch, estimated white blood cell proportions, 30 control probe PCs and a variable number of principal components, as described in the ‘Principal components + *bacon*’ section.

Survival analyses

Association testing. To investigate whether PMSs and the 45 significant sites were associated with survival, we applied multivariate Cox proportional hazards (PH) regression models utilizing the `coxph` function of the *survival* R package. Individuals with known overall survival, defined as the time from the date of a patient's onset to the date of death or last known follow-up were included (450k: N = 2,892, EPIC: N = 2,363). Overall survival in months was defined as outcome of interest in the multivariate Cox PH model.

For the individual sites, we used multivariate models and corrected for bias and inflation similar as described in the 'Principal components + *bacon*' section. Significant sites from the MOA and LB model were selected for survival analysis.

For the PMSs, we used multivariate models as described in the 'Poly-methylation Scores' section.

The proportional hazards (PH) assumption of the Cox model was checked using Schoenfeld and martingale residuals. Test-statistics were combined using inverse variance-weighted fixed effects meta-analysis. Sites were considered significantly heterogeneous when Cochran's Q P-values < 0.1 (corrected for the number of tests performed).

Sensitivity analyses. Sensitivity analyses were performed to assess robustness of the Cox PH output after recognition of a violation in the proportional hazards assumptions. We applied two sensitivity analyses proposed to detect and model such time-varying effects^{105,106}.

First, the multivariate cox model, as described above, was stratified by batch allowing the underlying hazard function to vary across the experimental batch levels. Moreover, the variable predicted age was added as a log transformed time-varying covariate.

Second, flexible survival regression using the Royston-Parmar (RP) spline model was performed utilizing the `flexsurvspline` function from the R package *flexsurv*. This model utilizes restricted cubic splines permitting the estimation of a continuous function instead of a step function which takes into account time varying issues encountered when using the Cox PH. Model complexity was assessed by the addition of up to five knots compared to one single knot. Lastly, we evaluated the predictive effect of *C9orf72* carrier status on survival by adjusting for C9 in the multivariate model described above.

Data availability

All summary statistics are available as supplementary data. Raw data will become available after peer review.

Acknowledgments

E.H. and J.M. were supported by Medical Research Council (MRC) grant K013807 (awarded to J.M.).

G.S. was supported by a PhD studentship from the Alzheimer's Society.

W.v.R. is supported by funding provided by the Dutch Research Council (NWO) [VENI scheme grant 09150161810018] and Prinses Beatrix Spierfond (neuromuscular fellowship grant W.F19-03).

J.J.F.A.v.V. and J.H.V. acknowledge the Prinses Beatrix Spierfonds (W.OR20-08) for funding.

A.A.K. is funded by The Motor Neurone Disease Association (MNDA) and NIHR Maudsley Biomedical Research Centre.

A.N.B. is grateful to Suna and Inan Kirac Foundation and Koc University for the excellent research environment created and for their financial support.

G.A.R. is supported by the Canadian Institutes of Health.

J.P.R. is funded by the Canadian Institutes of Health Research (FRN 159279).

R.J.P. is funded through the Gravitation program of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (BRAINSCAPES).

V.S. is supported by the Italian Ministry of Health, AriSLA, and E-Rare Joint Transnational Call.

K.P.K. is supported by funding provided by the Dutch Research Council (NWO) [VIDI grant 91719350].

D.B., E.T. and H.R. are employees of Biogen.

L.H.v.d.B. reports grants from the Netherlands ALS Foundation, grants from The Netherlands Organization for Health Research and Development (Vici scheme), grants from The European Community's Health Seventh Framework Programme (grant agreement n° 259867 (EuroMOTOR)), grants from The Netherlands Organization for Health Research and Development) the STRENGTH project, funded through the EU Joint Programme – Neurodegenerative Disease Research, JPND), during the conduct of the study.

Several authors of this publication are members of the Netherlands Neuromuscular Center (NL-NMD) and the European Reference Network for rare neuromuscular diseases EURO-NMD.

Project MinE Belgium was supported by a grant from IWT (n° 140935), the ALS Liga België, the National Lottery of Belgium and the KU Leuven Opening the Future Fund. PVD holds a senior clinical investigatorship of FWO-Vlaanderen and is supported by the E. von Behring Chair for Neuromuscular and Neurodegenerative Disorders, the ALS Liga België and the KU Leuven funds “Een Hart voor ALS”, “Laeversfonds voor ALS Onderzoek” and the “Valéry Perrier Race against ALS Fund”.

French ALS patients of the Pitié-Salpêtrière hospital (Paris) have been collected with ARSla funding support.

This work was supported by the Italian Ministry of Health (Ministero della Salute, Ricerca Sanitaria Finalizzata, grant RF-2016-02362405), the Progetti di Rilevante Interesse Nazionale program of the Ministry of Education, University and Research (grant 2017SNW5MB); the European Commission’s Health Seventh Framework Programme (FP7/2007–2013 under grant agreement 259867), and the Joint Programme–Neurodegenerative Disease Research (Strength, ALS-Care and Brain-Mend projects), granted by Italian Ministry of Education, University and Research. This study was performed under the Department of Excellence grant of the Italian Ministry of Education, University and Research to the “Rita Levi Montalcini” Department of Neuroscience, University of Torino, Italy.

We acknowledge funding from the Australian National Health and Medical Research (NHMRC) Council: 1151854, 1083187, 1173790, 1078901, 1113400, 1095215, 1176913 Enabling Grant #402703. Additional funding was provided by the Motor Neurone Disease Research Institute of Australia Ice Bucket Challenge grant for the SALSA-SGC consortium. The Older Australian Twins Study (OATS, used for controls) was facilitated through Twins Research Australia, a national resource in part supported by a Centre for Research Excellence from the Australian National Health and Medical Research Council (NHMRC 1079102). Funding for this study was awarded by the (NHMRC)/Australian Research Council Strategic Award (Grant 401162) and NHMRC grants (1405325, 1024224, 1025243, 1045325, 1085606, 568969, 1093083). We acknowledge the OATS research team: <https://cheba.unsw.edu.au/project/older-australian-twins-study>. We thank the participants and their informants for their time and generosity in contributing to this research.

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 772376 - EScORIAL).

The collaboration project is co-funded by the PPP Allowance made available by Health~Holland, Top Sector Life Sciences & Health, to stimulate public-private partnerships.

This study was supported by the ALS Foundation Netherlands.

This is an EU Joint Programme - Neurodegenerative Disease Research (JPND) project. The project is supported through the following funding organisations under the aegis of JPND - www.jpnd.eu (United Kingdom, Medical Research Council (MR/L501529/1; MR/R024804/1) and Economic and Social Research Council (ES/L008238/1)) and through the Motor Neurone Disease Association. This study represents independent research part funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. A.A-C. is supported by an NIHR Senior Investigator Award. Samples used in this research were entirely/in part obtained from the UK National DNA Bank for MND Research, funded by the MND Association and the Wellcome Trust. We would like to thank people with MND and their families for their participation in this project. We acknowledge sample management undertaken by Biobanking Solutions funded by the Medical Research Council at the Centre for Integrated Genomic Medical Research, University of Manchester.

Author Contributions

Sample ascertainment & data generation: P.J.H, R.A.J.Z, E.H, G.L.S, M.F.N, E.M.W, W.v.R, J.J.F.A.v.V, A.M.D, H-J.W, G.H.P.T, K.R.v.E, M.M, A.A.K, A.I, N.T, A.R, J.C-K, K.E.M, P.J.S, A.N.B, A.C, A. Calvo, C.M, A. Canosa, M.B, M.G, M. Gotkine, Y.L, M.Z, P.V'h, P.C, P. Couratier, J.S.M.P, T.S, P.D, J.P.R, R.D.H, S.M, P.A.M, M.N, G.N, D.B.R, R.P, K.A.M, P.S.S, S.F, F.C.G, A.K.H, T.L, S.T.N, F.J.S, L.W, K.L.W, B.C, B.M.C, M.M.N, R.J.C, I.P.B, M.C.K, V.D, M.P, M.d.C, S.P, M.W, G.R, V.S, J.E.L, C.E.S, P.M.A, A.F.M, M.A.v.E, R.J.P, N.R.W, R.L.M, O.H, K.P.K, A.A-C, L.H.v.d.B, P.V.D, J.M, J.H.V. **Whole-genome sequencing:** P.J.H, R.A.J.Z, W.v.R, J.J.F.A.v.V, A.M.D, G.H.P.T, K.R.v.E, M.M, J.C-K, K.P.K, A.A-C, L.H.v.d.B, P.V.D, J.H.V. **WGS quality-control:** P.J.H, R.A.J.Z, W.v.R, J.J.F.A.v.V, M.M, K.P.K, P.V.D, J.H.V. **Data analysis:** P.J.H, R.A.J.Z, E.H, J.M, J.H.V. **Writing the manuscript:** P.J.H, R.A.J.Z, J.M, J.H.V. **Revising manuscript:** P.J.H, R.A.J.Z, M.F.N, W.v.R, J.J.F.A.v.V, H-J.W, D.B, R.J.P, N.R.W, H.R, E.T, P.V.D, J.M, J.H.V.

Competing Interests

J.H.V. has sponsored research agreements with Biogen.

L.H.v.d.B receives personal fees from Cytokinetics, outside the submitted work.

A.A-C. has served on scientific advisory boards for Mitsubishi Tanabe Pharma, OrionPharma, Biogen Idec, Lilly, GSK, Apellis, Amylyx, and Wave Therapeutics.

A.C. serves on scientific advisory boards for Mitsubishi Tanabe, Roche, Biogen, Denali, and Cytokinetics.

Consortia

BIOS consortium (Biobank-based Integrative Omics Study)

Management Team Bastiaan T. Heijmans⁶² (Chair), Peter A.C. t Hoen⁶³, Joyce van Meurs⁶⁴, Rick Jansen⁶⁵, Lude Franke⁶⁶

Cohort collection Dorret I. Boomsma⁶⁷, René Pool⁶⁷, Jenny van Dongen⁶⁷, Joukje J. Hottenga⁶⁷ (Netherlands Twin Register); Marleen M.J. van Greevenbroek⁶⁸, Coen D.A. Stehouwer⁶⁸, Carla J.H. van der Kallen⁶⁸, Casper G. Schalkwijk⁶⁸ (Cohort study on Diabetes and Atherosclerosis Maastricht); Cisca Wijmenga⁶⁶, Lude Franke⁶⁶, Sasha Zhernakova⁶⁶, Etti F. Tigchelaar⁶⁶ (LifeLines Deep); P. Eline Slagboom⁶², Marian Beekman⁶², Joris Deelen⁶², Diana van Heemst⁶⁹ (Leiden Longevity Study); Jan H. Veldink¹, Leonard H. van den Berg¹ (Prospective ALS Study Netherlands); Cornelia M. van Duijn⁷⁰, Bert A. Hofman⁷¹, Aaron Isaacs⁷⁰, André G. Uitterlinden⁶⁴ (Rotterdam Study)

Data Generation, Joyce van Meurs⁶⁴ (Chair), P. Mila Jhamai⁶⁴, Michael Verbiest⁶⁴, H. Eka D. Suchiman⁶², Marijn Verkerk⁶⁴, Ruud van der Breggen⁶², Jeroen van Rooij⁶⁴, Nico Lakenberg⁶²

Data management and computational infrastructure Hailiang Mei⁷² (Chair), Maarten van Iterson⁶², Michiel van Galen⁶³, Jan Bot⁷³, Dasha V. Zhernakova⁶⁶, Rick Jansen⁶⁵, Peter van 't Hof⁷², Patrick Deelen⁶⁶, Irene Nooren⁷³, Peter A.C. t Hoen⁶³, Bastiaan T. Heijmans⁶², Matthijs Moed⁶²

Data Analysis Group Lude Franke⁶⁶ (Co-Chair), Martijn Vermaat⁶³, Dasha V. Zhernakova⁶⁶, René Luijk⁶², Marc Jan Bonder⁶⁶, Maarten van Iterson⁶², Patrick Deelen⁶⁶, Freerk van Dijk⁷⁴, Michiel van Galen⁶³, Wibowo Arindrarto⁷², Szymon M. Kielbasa⁷⁵, Morris A. Swertz⁷⁴, Erik W. van Zwet⁷⁵, Rick Jansen⁶⁵, Peter A.C. t Hoen⁶³ (Co-Chair), Bastiaan T. Heijmans⁶² (Co-Chair)

Affiliations

62: Molecular Epidemiology, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands.

63: Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands.

64: Department of Internal Medicine, ErasmusMC, Rotterdam, The Netherlands.

65: Department of Psychiatry, VU University Medical Center, Neuroscience Campus Amsterdam, Amsterdam, The Netherlands.

66: Department of Genetics, University of Groningen, University Medical Centre Groningen, Groningen, The Netherlands.

67: Department of Biological Psychology, VU University Amsterdam, Neuroscience Campus Amsterdam, Amsterdam, The Netherlands.

68: Department of Internal Medicine and School for Cardiovascular Diseases (CARIM), Maastricht University Medical Center, Maastricht, The Netherlands.

69: Department of Gerontology and Geriatrics, Leiden University Medical Center, Leiden, The Netherlands.

70: Department of Genetic Epidemiology, ErasmusMC, Rotterdam, The Netherlands.

71: Department of Epidemiology, ErasmusMC, Rotterdam, The Netherlands.

72: Sequence Analysis Support Core, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands.

73: SURFsara, Amsterdam, the Netherlands.

74: Genomics Coordination Center, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands.

75: Medical Statistics, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands.

BRAIN MEND Consortium

Biological Resource Analysis to Identify New MEchanisms and phenotypes in Neurodegenerative Diseases consortium.

Ammar Al-Chalabi^{9,61}, Naomi R. Wray^{3,45}, Gilbert Bensimon^{76,45}, Orla Hardiman⁶⁰, Adriano Chiò¹⁹, Jan H. Veldink¹, George Davey Smith^{77,78}, Jonathan Mill²

Affiliations

76: Département de Pharmacologie Clinique, Hôpital de la Pitié-Salpêtrière, UPMC Pharmacologie, AP-HP, Paris, France.

77: MRC Integrative Epidemiology Unit, University of Bristol, Bristol BS8 1TH, UK.

78: Population Health Science, Bristol Medical School, Bristol, Bristol BS8 1TH, UK.

References

1. van Es, M. A. *et al.* Amyotrophic lateral sclerosis. *The Lancet* **390**, 2084–2098 (2017).
2. Ryan, M., Heverin, M., McLaughlin, R. L. & Hardiman, O. Lifetime Risk and Heritability of Amyotrophic Lateral Sclerosis. *JAMA Neurol.* **76**, 1367–1374 (2019).
3. Al-Chalabi, A. & Hardiman, O. The epidemiology of ALS: a conspiracy of genes, environment and time. *Nat. Rev. Neurol.* **9**, 617–628 (2013).
4. Xue, A. *et al.* Genome-wide analyses of behavioural traits are subject to bias by misreports and longitudinal changes. *Nat. Commun.* **12**, 20211 (2021).
5. Armon, C. An Evidence-Based Medicine Approach to the Evaluation of the Role of Exogenous Risk Factors in Sporadic Amyotrophic Lateral Sclerosis. *Neuroepidemiology* **22**, 217–228 (2003).
6. Dor, Y. & Cedar, H. Principles of DNA methylation and their implications for biology and medicine. *The Lancet* **392**, 10149,777–786 (2018).
7. Hannon, E. *et al.* An integrated genetic-epigenetic analysis of schizophrenia: evidence for co-localization of genetic associations and differential DNA methylation. *Genome Biol.* **17**, 176 (2016).
8. Vallerga, C. L. *et al.* Analysis of DNA methylation associates the cystine–glutamate antiporter SLC7A11 with risk of Parkinson’s disease. *Nat. Commun.* **11**, 1238 (2020).
9. Lunnon, K. *et al.* Methylomic profiling implicates cortical deregulation of ANK1 in Alzheimer’s disease. *Nat. Neurosci.* **17**, 1164–1170 (2014).

10. Elliott, H. R. *et al.* Differences in smoking associated DNA methylation patterns in South Asians and Europeans. *Clin. Epigenetics* **6**, 4 (2014).
11. Liu, C. *et al.* A DNA methylation biomarker of alcohol consumption. *Mol. Psychiatry* **23**, 422–433 (2018).
12. Hamilton, O. K. L. *et al.* An epigenetic score for BMI based on DNA methylation correlates with poor physical health and major disease in the Lothian Birth Cohort. *Int. J. Obes.* **43**, 1795–1802 (2019).
13. Zhang, Q. *et al.* Improved precision of epigenetic clock estimates across tissues and its implication for biological ageing. *Genome Med.* **11**, 54 (2019).
14. McCartney, D. L. *et al.* Epigenetic prediction of complex traits and death. *Genome Biol.* **19**, 136 (2018).
15. Ligthart, S. *et al.* DNA methylation signatures of chronic low-grade inflammation are associated with complex diseases. *Genome Biol.* **17**, 255 (2016).
16. Gadd, D. A. *et al.* DNA methylation proxies for 16 plasma proteins predict the incidence of 7 leading causes of morbidity. *BioRxiv* (2020).
17. Marioni, R. E. *et al.* DNA methylation age of blood predicts all-cause mortality in later life. *Genome Biol.* **16**, 25 (2015).
18. Hao, X. *et al.* DNA methylation markers for diagnosis and prognosis of common cancers. *Proc. Natl. Acad. Sci.* **114**, 7414–7419 (2017).
19. Xi, Z. *et al.* Hypermethylation of the CpG island near the G 4 C 2 repeat in ALS with a C9orf72 expansion. *Am. J. Hum. Genet.* **92**, 981–989 (2013).
20. Xi, Z. *et al.* The C9orf72 repeat expansion itself is methylated in ALS and FTLN patients. *Acta Neuropathol. (Berl.)* **129**, 715–727 (2015).
21. Young, P. E., Kum Jew, S., Buckland, M. E., Pamphlett, R. & Suter, C. M. Epigenetic differences between monozygotic twins discordant for amyotrophic lateral sclerosis (ALS) provide clues to disease pathogenesis. *PLOS ONE* **12**, e0182638 (2017).

22. Figueroa-Romero, C. *et al.* Identification of Epigenetically Altered Genes in Sporadic Amyotrophic Lateral Sclerosis. *PLoS ONE* **7**, e52672 (2012).
23. Morahan, J. M., Yu, B., Trent, R. J. & Pamphlett, R. A genome-wide analysis of brain DNA methylation identifies new candidate genes for sporadic amyotrophic lateral sclerosis. *Amyotroph. Lateral Scler.* **10**, 418–429 (2009).
24. Tarr, I. S. *et al.* Monozygotic twins and triplets discordant for amyotrophic lateral sclerosis display differential methylation and gene expression. *Sci. Rep.* **9**, 8254 (2019).
25. Nabais, M. F. *et al.* Significant out-of-sample classification from methylation profile scoring for amyotrophic lateral sclerosis. *Npj Genomic Med.* **5**, 10 (2020).
26. van Rheenen, W. *et al.* Common and rare variant association analyses in Amyotrophic Lateral Sclerosis identify 15 risk loci with distinct genetic architectures and neuron-specific biology. *MedRxiv submitted.* (2021).
27. van Iterson, M., van Zwet, E. W. & Heijmans, B. T. Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution. *Genome Biol.* **18**, 19 (2017).
28. Zhang, F. *et al.* OSCA: a tool for omic-data-based complex trait analysis. *Genome Biol.* **20**, 107 (2019).
29. Schwarzer, G. *meta: An R package for meta-analysis.* (2019).
30. Fortin, J.-P. *et al.* Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biol.* **15**, 503 (2014).
31. Pidsley, R. *et al.* A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics* **14**, 293 (2013).
32. Hop, P. J. *et al.* Cross-reactive probes on Illumina DNA methylation arrays: a large study on ALS shows that a cautionary approach is warranted in interpreting epigenome-wide association studies. *NAR Genomics Bioinforma.* **2**, 4, lqaa105 (2020).
33. Mansell, G. *et al.* Guidance for DNA methylation studies: statistical insights from the

- Illumina EPIC array. *BMC Genomics* **20**, 366 (2019).
34. Hop, P. J. *et al.* Genome-wide identification of genes regulating DNA methylation using genetic anchors for causal inference. *Genome Biol.* **21**, 220 (2020).
 35. Min, J. L. *et al.* Genomic and phenomic insights from an atlas of genetic effects on DNA methylation. *medRxiv* (2020) doi:10.1101/2020.09.01.20180406.
 36. Ren, X. & Kuan, P. F. methylGSA: a Bioconductor package and Shiny app for DNA methylation data length bias adjustment in gene set testing. *Bioinformatics* **35**, 11,1958-1959 (2019).
 37. Phipson, B., Maksimovic, J. & Oshlack, A. missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. *Bioinformatics* 286–288 (2015) doi:10.1093/bioinformatics/btv560.
 38. Staley, J. *et al.* EWAS Catalog: The MRC-IEU catalog of epigenome-wide association studies. Accessed: June 6 2020, from <http://www.ewascatalog.org> (2020).
 39. Iiu, D. *et al.* EWASdb: epigenome-wide association study database, accessed June 6 2020. *Nucleic Acids Res.* **47**, D989–D993 (2019).
 40. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, P10008 (2008).
 41. Yousefi, P. D. *et al.* Validation and characterisation of a DNA methylation alcohol biomarker across the life course. **11**, 163 (2019).
 42. Bollepalli, S., Korhonen, T., Kaprio, J., Anders, S. & Ollikainen, M. EpiSmokEr: A robust classifier to determine smoking status from DNA methylation data. *Epigenomics* **11**, 1469–1486 (2019).
 43. Teschendorff, A. E., Breeze, C. E., Zheng, S. C. & Beck, S. A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. *BMC Bioinformatics* **18**, 105 (2017).
 44. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol.* **14**,

- R115 (2013).
45. Hannum, G. *et al.* Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. *Mol. Cell* **49**, 359–367 (2013).
 46. Dekkers, K. F. *et al.* Blood lipids influence DNA methylation in circulating cells. *Genome Biol.* **17**, 138 (2016).
 47. BIOS consortium *et al.* Epigenome-wide association study (EWAS) on lipids: the Rotterdam Study. *Clin. Epigenetics* **9**, 47 (2017).
 48. Hedman, Å. K. *et al.* Epigenetic Patterns in Blood Associated With Lipid Traits Predict Incident Coronary Heart Disease Events and Are Enriched for Results From Genome-Wide Association Studies. *Circ. Cardiovasc. Genet.* **10**, e001487 (2017).
 49. Wahl, S. *et al.* Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature* **541**, 81–86 (2017).
 50. Mendelson, M. M. *et al.* Association of Body Mass Index with DNA Methylation and Gene Expression in Blood Cells and Relations to Cardiometabolic Disease: A Mendelian Randomization Approach. *PLOS Med.* **14**, 1, e1002215 (2017).
 51. Chambers, J. C. *et al.* Epigenome-wide association of DNA methylation markers in peripheral blood from Indian Asians and Europeans with incident type 2 diabetes: a nested case-control study. *Lancet Diabetes Endocrinol.* **3**, 526–534 (2015).
 52. Ma, J. *et al.* A Peripheral Blood DNA Methylation Signature of Hepatic Fat Reveals a Potential Causal Pathway for Nonalcoholic Fatty Liver Disease. *Diabetes* **68**, 1073–1083 (2019).
 53. Wilson, L. E. *et al.* Alcohol and DNA Methylation: An Epigenome-Wide Association Study in Blood and Normal Breast Tissue. *Am. J. Epidemiol.* **188**, 1055–1065 (2019).
 54. Nano, J. *et al.* Epigenome-Wide Association Study Identifies Methylation Sites Associated With Liver Enzymes and Hepatic Steatosis. *Gastroenterology* **153**, 1096–1106 (2017).

55. Liang, X. *et al.* DNA methylation signature on phosphatidylethanol, not on self-reported alcohol consumption, predicts hazardous alcohol consumption in two distinct populations. *Mol. Psychiatry* (2020) doi:10.1038/s41380-020-0668-x.
56. Sato, H., Tamba, M., Ishii, T. & Bannai, S. Cloning and Expression of a Plasma Membrane Cystine/Glutamate Exchange Transporter Composed of Two Distinct Proteins. *J. Biol. Chem.* **274**, 11455–11458 (1999).
57. Dupuis, L., Pradat, P.-F., Ludolph, A. C. & Loeffler, J.-P. Energy metabolism in amyotrophic lateral sclerosis. *Lancet Neurol.* **10**, 75–82 (2011).
58. Harwood, C. A. *et al.* Long-term physical activity: an exogenous risk factor for sporadic amyotrophic lateral sclerosis? *Amyotroph. Lateral Scler. Front. Degener.* **17**, 377–384 (2016).
59. Seelen, M. *et al.* Prior medical conditions and the risk of amyotrophic lateral sclerosis. *J. Neurol.* **261**, 1949–1956 (2014).
60. Turner, M. R., Wotton, C., Talbot, K. & Goldacre, M. J. Cardiovascular fitness as a risk factor for amyotrophic lateral sclerosis: indirect evidence from record linkage study: Table 1. *J. Neurol. Neurosurg. Psychiatry* **83**, 395–398 (2012).
61. Singh, R. *et al.* Autophagy regulates lipid metabolism. *Nature* **458**, 1131–1135 (2009).
62. Barbero-Camps, E. *et al.* Cholesterol impairs autophagy-mediated clearance of amyloid beta while promoting its secretion. *Autophagy* **14**, 1129–1154 (2018).
63. Westeneng, H.-J. *et al.* Effects of presymptomatic lifestyle on ALS stratified by C9orf72 genotype: a longitudinal population-based study. *Lancet Neurol.* (*In press*).
64. Chen, W. *et al.* An epigenome-wide association study of total serum IgE in Hispanic children. *J. Allergy Clin. Immunol.* **140**, 2,571-577 (2017).
65. Reese, S. E. *et al.* Epigenome-wide meta-analysis of DNA methylation and childhood asthma. *J. Allergy Clin. Immunol.* **143**, 2062–2074 (2019).
66. Cardenas, A. *et al.* The nasal methylome as a biomarker of asthma and airway inflammation in children. *Nat. Commun.* **10**, 3095 (2019).

67. Peng, C. *et al.* Epigenome-wide association study reveals methylation pathways associated with childhood allergic sen. *Epigenetics* **14**, 445–466 (2019).
68. Nabais et al., M. F. Meta-analysis of genome-wide DNA methylation identifies shared associations across neurodegenerative disorders. *Genome Biol.* (*In press*).
69. Keizman, D. *et al.* Low-grade systemic inflammation in patients with amyotrophic lateral sclerosis. *Acta Neurol. Scand.* **119**, 383–389 (2009).
70. Murdock, B. J. *et al.* Increased ratio of circulating neutrophils to monocytes in amyotrophic lateral sclerosis. *Neurol. Neuroimmunol. Neuroinflammation* **3**, 4,e242 (2016).
71. Hu, Y. *et al.* Increased peripheral blood inflammatory cytokine levels in amyotrophic lateral sclerosis: a meta-analysis study. *Sci. Rep.* **7**, 1,9094 (2017).
72. Choi, S.-J. *et al.* High neutrophil-to-lymphocyte ratio predicts short survival duration in amyotrophic lateral sclerosis. *Sci. Rep.* **10**, 428 (2020).
73. Béland, L.-C. *et al.* Immunity in amyotrophic lateral sclerosis: blurred lines between excessive inflammation and inefficient immune responses. *Brain Commun.* **2**, fcaa124 (2020).
74. Ransohoff, R. M. How neuroinflammation contributes to neurodegeneration. *Science* **353**, 777–783 (2016).
75. Trias, E. *et al.* Mast cells and neutrophils mediate peripheral motor pathway degeneration in ALS. *JCI Insight* **3**, 19,e123249 (2018).
76. Trias, E. *et al.* Evidence for mast cells contributing to neuromuscular pathology in an inherited model of ALS. *JCI Insight* **2**, e95934 (2017).
77. Mora, J. S. *et al.* Masitinib as an add-on therapy to riluzole in patients with amyotrophic lateral sclerosis: a randomized clinical trial. *Amyotroph. Lateral Scler. Front. Degener.* **21**, 1-2,5-14 (2020).
78. Xi, Z. *et al.* Hypermethylation of the CpG island near the G 4 C 2 repeat in ALS with a C9orf72 expansion. *Am. J. Hum. Genet.* **92**, 981–989 (2013).

79. Xi, Z. *et al.* The C9orf72 repeat expansion itself is methylated in ALS and FTLD patients. *Acta Neuropathol. (Berl.)* **129**, 5,715-727 (2015).
80. He, J. *et al.* C9orf72 hexanucleotide repeat expansions in Chinese sporadic amyotrophic lateral sclerosis. *Neurobiol. Aging* **36**, 9,2660.e1-2660.e8 (2015).
81. Zhang, M. *et al.* DNA methylation age acceleration is associated with ALS age of onset and survival. *Acta Neuropathol. (Berl.)* **139**, 943–946 (2020).
82. El Khoury, L. Y. *et al.* Systematic underestimation of the epigenetic clock and age acceleration in older subjects. *Genome Biol.* **20**, 283 (2019).
83. Westeneng, H.-J. *et al.* Prognosis for patients with amyotrophic lateral sclerosis: development and validation of a personalised prediction model. *Lancet Neurol.* **17**, 5,423-433 (2018).
84. Hannon, E. *et al.* Leveraging DNA-Methylation Quantitative-Trait Loci to Characterize the Relationship between Methylomic Variation, Gene Expression, and Complex Traits. *Am. J. Hum. Genet.* **103**, 654–665 (2018).
85. Bakulski, K. M., Halladay, A., Hu, V. W., Mill, J. & Fallin, M. D. Epigenetic Research in Neuropsychiatric Disorders: the “Tissue Issue”. *Curr. Behav. Neurosci. Rep.* **3**, 264–274 (2016).
86. Rahmani, E. *et al.* Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. *Nat. Methods* **13**, 5,443-445 (2016).
87. Farré, P. *et al.* Concordant and discordant DNA methylation signatures of aging in human blood and brain. *Epigenetics Chromatin* **8**, 19 (2015).
88. Kong, Y., Rastogi, D., Seoighe, C., Grealley, J. M. & Suzuki, M. Insights from deconvolution of cell subtype proportions enhance the interpretation of functional genomic data. *PLOS ONE* **14**, e0215987 (2019).
89. Lappalainen, T. & Grealley, J. M. Associating cellular epigenetic models with human phenotypes. *Nat. Rev. Genet.* **18**, 441–451 (2017).

90. Brooks, B. R., Miller, R. G., Swash, M. & Munsat, T. L. El Escorial revisited: revised criteria for the diagnosis of amyotrophic lateral sclerosis. *Amyotroph. Lateral Scler. Other Motor Neuron Disord.* **1**, 293–299 (2000).
91. Aryee, M. J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 10,1363-1369 (2014).
92. van Iterson, M. *et al.* MethylAid: visual and interactive quality control of large Illumina 450k datasets. *Bioinformatics* **30**, 23,435-3437 (2014).
93. Project MinE ALS Sequencing Consortium. Project MinE: study design and pilot analyses of a large-scale whole-genome sequencing study in amyotrophic lateral sclerosis. *Eur. J. Hum. Genet.* **26**, 1537–1546 (2018).
94. van Iterson, M., Cats, D., Hop, P., BIOS Consortium & Heijmans, B. T. omicsPrint: detection of data linkage errors in multiple omics studies. *Bioinformatics* **34**, 2142–2143 (2018).
95. Lehne, B. *et al.* A coherent approach for analysis of the Illumina HumanMethylation450 BeadChip improves data quality and performance in epigenome-wide association studies. *Genome Biol* **16**, 37 (2015).
96. Freedman, M. L. *et al.* Assessing the impact of population stratification on genetic association studies. *Nat. Genet.* **36**, 388–393 (2004).
97. Dolzhenko, E. *et al.* Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res.* **27**, 1895–1903 (2017).
98. Zhou, W., Laird, P. W. & Shen, H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Res.* **45**, e22 (2017).
99. Ren, X. & Kuan, P. F. methylGSA: a Bioconductor package and Shiny app for DNA methylation data length bias adjustment in gene set testing. *Bioinformatics* **35**, 11,1958-1959 (2019).

100. Luijk, R., Goeman, J. J., Slagboom, E. P., Heijmans, B. T. & van Zwet, E. W. An alternative approach to multiple testing for methylation QTL mapping reduces the proportion of falsely identified CpGs. *Bioinformatics* **31**, 3,340-345 (2015).
101. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal Complex Syst.* 1695 (2006).
102. Zheng, S. C., Breeze, C. E., Beck, S. & Teschendorff, A. E. Identification of differentially methylated cell types in epigenome-wide association studies. *Nat. Methods* **15**, 1059–1066 (2018).
103. Hillary, R. F. *et al.* Genome and epigenome wide studies of neurological protein biomarkers in the Lothian Birth Cohort 1936. *Nat. Commun.* **10**, 3160 (2019).
104. Lobo, S. M. *et al.* C-reactive protein levels correlate with mortality and organ failure in critically ill patients. *Chest* **123**, 2043–2049 (2003).
105. Royston, P. & Parmar, M. K. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Stat. Med.* **21**, 2175–2197 (2002).
106. Zhang, Z., Reinikainen, J., Adeleke, K. A., Pieterse, M. E. & Groothuis-Oudshoorn, C. G. Time-varying covariates and coefficients in Cox regression models. *Ann. Transl. Med.* **6**, 7,121 (2018).