Understanding Society
Working Paper Series
No. 2023 – 05
March 2023

Proteomics in Understanding Society: pre-analytic condition impacts on measurement

Anna Dearman, Meena Kumari and Yanchun Bao

University of Essex

Understanding Society

UKRI Economic and Social Research Council

Non-technical summary

Due to recent developments in laboratory methods, large numbers of proteins which are present in the blood at low concentrations can now be measured simultaneously, a technology called “proteomics”. When these proteins are measured in studies of the general population, they may provide useful insights into how social factors “get under the skin” and affect our health. This paper examines whether the method used in Understanding Society (waves 2 and 3, 2010-2012) to collect the blood samples affects the measurement of proteins. If it does, our aim is to examine whether the effect is enough to mask associations between the proteins and social factors.

In order to include as many eligible participants as possible, Understanding Society sent nurses to participants’ homes at a time convenient to them, where blood samples were collected and sent to a laboratory via the post. This meant that the samples were delayed in transit by one or more days before the liquid portion (serum) was extracted in the lab. This delay could affect the quality and protein measurements. We measured 92 proteins related to cardiovascular and metabolic health, and 92 which are related to the brain.

In this working paper, we describe the technology used to measure the proteins, and the quality checks we performed on the resulting proteomics dataset. We explore whether "pre-analytic factors", time of day of sample collection, transit delay and haemolysis (burst blood cells) have an effect on any of the protein measurements, and whether these effects interfere with our ability to find associations with other factors, using educational attainment as our example.

We find that the dataset is of high quality, with very few measurements missing. Around a tenth of the proteins were defined as “low detectability” proteins, each having at least 10% of measurements below a recommended threshold. Transit delay was related to time of day of sample collection and haemolysis and has an effect on the measurements of many proteins. Educational attainment is associated with 127 of the 184 proteins we measured including nearly half of the “low detectability” proteins. Pre-analytic factors made little difference to these associations. Our findings suggest that the method we used to collect blood samples in Understanding Society does affect measurement, but it does not impact whether we observe social variation in proteomics.

Abstract

Proteomic methods have been developed to enable measurements of low levels of proteins in blood samples. In Waves 2 and 3 (2010-2012) of Understanding Society, blood samples were collected in participants' homes, then posted to the laboratory and subsequently used for the measurement of 184 proteins. It is unknown whether the delay between blood collection and laboratory processing compromises blood sample integrity such that genuine associations between proteins and demographic factors are obscured. This paper describes the protein measurements and the quality checks performed on them. It also describes the proteins' associations with a range of factors, including pre-analytic factors (time of day of sample collection, transit delay: number of days the sample was in the post, haemolysis of sample) and demographic factors (age, sex, educational attainment, geographical region, ethnicity). A range of standard statistical tests were used. We find that protein measurements in the dataset are highly reproducible, with low levels of missingness. We observe low detectability in around one-tenth of proteins. We identify transit delay as an important source of variation, and we characterise the nature of its relationship to each individual protein. Examining educational attainment as an exemplar, we observe that of the proteins measured, 127 varied by educational attainment, including just under half of the proteins defined with low detectability. The largest associations with educational attainment are found in proteins previously described to be associated with cognitive function. While adjustment for pre-analytic factors usually improved overall model fit, association of proteins with educational attainment were largely unaltered.

Keywords: longitudinal, biosocial, health, proteomics, biomarkers

Acknowledgements: Understanding Society is an initiative funded by the Economic and Social Research Council (ESRC) and various Government Departments, with scientific leadership by the Institute for Social and Economic Research, University of Essex, and survey delivery by NatCen Social Research and Kantar Public. The research was supported by the ESRC (ES/S007253/1 & ES/ S012486/1). We would also like to thank the laboratory staff at Olink who analysed the samples.

Data availability: The Understanding Society Nurse Health Assessment data are available from the UK Data Service: University of Essex. Institute for Social and Economic Research and National Centre for Social Research, Understanding Society: Waves 2 and 3 Nurse Health Assessment, 2010- 2012 [data collection]. 5th Edition. UK Data Service. SN:7251. http://doi.org/10.5255/UKDA-SN-7251-5. Details of data cleaning can be found in the Proteomics User Guide (Quality control section 2.2.3) on the Understanding Society health assessment user guides web page: https://www. understandingsociety.ac.uk/documentation/health-assessment/user-guide

We thank Professor Paul Clarke for assisting with this project.

Contact: Meena Kumari, (mkumari@essex.ac.uk) Associate Director for Health, Biomarkers and Genetics, Institute for Social and Economic Research, University of Essex, Wivenhoe Park, Colchester, Essex, CO4 3SQ, UK.

Proteomics in Understanding Society: pre-analytic condition impacts on measurement

Anna Dearman, Meena Kumari, Yanchun Bao

Contents

1 Abstract

Proteomic methods have been developed to enable measurements of low levels of proteins in blood samples. In Waves 2 and 3 (2010-2012) of *Understanding Society*, blood samples were collected in participants' homes, then posted to the laboratory and subsequently used for the measurement of 184 proteins. It is unknown whether the delay between blood collection and laboratory processing compromises blood sample integrity such that genuine associations between proteins and demographic factors are obscured. This paper describes the protein measurements and the quality checks performed on them. It also describes the proteins' associations with a range of factors, including pre-analytic factors (time of day of sample collection, transit delay: number of days the sample was in the post, haemolysis of sample) and demographic factors (age, sex, educational attainment, geographical region, ethnicity). A range of standard statistical tests were used. We find that protein measurements in the dataset are highly reproducible, with low levels of missingness. We observe low detectability in around one-tenth of proteins. We identify transit delay as an important source of variation, and we characterise the nature of its relationship to each individual protein. Examining educational attainment as an exemplar, we observe that of the proteins measured, 127 varied by educational attainment, including just under half of the proteins defined with low detectability. The largest associations with educational attainment are found in proteins previously described to be associated with cognitive function. While adjustment for pre-analytic factors usually improved overall model fit, association of proteins with educational attainment were largely unaltered.

2 Introduction

Proteomics is the study of proteins, which are the products of gene expression, and thus important components in the biological pathways of health, disease and behaviour. The human bloodstream contains many proteins, including some which are purposefully released into the blood by cells to perform their biological role (secreted), and others which are not known to function in the blood but have simply leaked into it from other tissues. Blood levels of proteins are likely influenced by a broad range of exposures, making them useful exploratory bio-social markers.

Understanding Society, also known as the UK Household Longitudinal Study (UKHLS) is one of several longitudinal studies of UK residents which collects social, biological and clinical data to facilitate research into the social to biological transitions. In Waves 2 and 3 (2010-2012) a clinical data collection took place, giving researchers access to a range of biological and other health-related variables to study, including anthropometric measures such as waist circumference, physical performance measures such as lung function, and blood-based measures such as testosterone, lipids and DNA to enable genomic analyses.

The blood samples used for protein measurements were collected by a nurse in the participants' homes at times that were convenient to the participant. The rationale for this mode of sample collection was that the offer of a home visit would help to maintain response rates, especially in hard to reach groups. Samples were posted to the laboratory for subsequent processing, i.e. separation of the liquid fraction from the cellular fraction using a centrifuge, after which small volumes (aliquots) of cell-free liquid are dispensed into separate tubes and frozen until they are needed. A recent systematic review suggested that standard biochemical analytes are impacted by variations in sample handling and processing [1], here referred to as pre-analytic factors. There is a potential for a number of pre-analytic factors to impact protein measurements [2]. In addition to pre-analytic factors that are typically examined in survey methodology, such as interviewer or nurse effects [3], these include sample-specific factors such as time of day of sample collection, number of days in transport (transit delay) and haemolysis of the sample. Time of day of sample collection has been associated with a number of proteins and blood based analytes, for example inflammatory markers [4], due to diurnal variation (typical patterns of fluctuation throughout the day). Processing delays, which occur while the sample is in transit between being posted and arriving at the laboratory, could impact protein measurements due to ongoing biological processes [1], [2] and haemolysis [5], wherein blood cell membranes rupture, allowing intracellular components to leak into the liquid fraction, a process which could impact protein measurements via a number of mechanisms [5]. These intracellular components have been shown to interfere with protein measurements under test conditions [6], [7].

The measurement of circulating proteins at scale such that there is simultaneous assessment of a number of proteins (proteomics) is a relatively new development, with many proteins not routinely measured in population studies. Thus, there is a lack of information on protein measurements from national studies that include participants across the entire adult age range. It is therefore unclear how protein measurements might vary with participant demographic characteristics such as age, sex, educational attainment, geographical region or ethnicity. Recently, levels of proteins that reflect cardiovascular biology were shown to vary by social position in a cohort of middle-aged Danish men and women [8]. However, the majority of studies that have examined protein levels do not account for socio-economic factors and it is unclear whether pre-analytic factors would serve to obscure any apparent associations.

This paper has three aims:

Firstly, to describe the quality checks performed on the dataset; secondly, to describe the relationships between protein levels and several factors: pre-analytic (time of day of blood sampling, transit delay, and haemolysis of the sample) and demographic (age, sex, educational attainment, geographical region and ethnicity); and thirdly, to explore whether pre-analytic factors obscure the associations between proteins and demographic factors.

3 Methods

3.1 Study description

The *Understanding Society* study began in 2009 and recruited individuals from a nationally representative sample of households across the UK. It consists of subsamples, including a representative General Population Sample (GPS), legacy households from the British Household Panel Survey (BHPS) which ran from 1991-2009 and was absorbed into Understanding Society during Wave 2, and boost samples such as the Ethnic Minority Boost Sample. Information is collected from participants at one-year intervals, including their social and economic circumstances, their attitudes and beliefs, and their health.

In Waves 2 and 3 of the study (2010-2012), around five months after an interview, 58% (20,699) of eligible adult participants from the GPS and BHPS received a health assessment visit from a registered nurse. The nurses visited participants in their homes and, in addition to a range of noninvasive measures (such as blood pressure, weight, and lung function), collected non-fasting blood samples at these visits. Appointment dates and times were noted.

The blood samples were collected from 13,328 participants and dispatched using the postal service Royal Mail to the laboratory (Fisher BioServices, Bishop's Stortford, UK), where staff recorded date of receipt and processed the samples, dispensing whole blood, plasma and serum into aliquots for long term storage.

Serum aliquots from 46% (6,180) of the samples in the original blood collection were sent to the Olink Proteomics (hereafter shortened to "Olink") laboratory for the measurement of 184 proteins: 4,625 participants from Wave 2 (GPS), and 1,555 from Wave 3 (BHPS).

For more details on the nurse health assessment, sampling and biomarkers, please see the *Understanding Society* user guides available on the health assessment user guides web page: https://www.understandingsociety.ac.uk/documentation/health-assessment/user-guide.

3.2 Proteins measured

Laboratory analyses were performed by Olink. Proteins from two of their Target 96 panels - the Olink® Target 96 Cardiometabolic panel [9] and the Olink® Target 96 Neurology panel [10], together comprising 184 proteins - were measured. See table 1 for the full names and identifiers for each of the proteins.

Table 1: Proteins measured in Understanding Society

Cardiometabolic protein ID	Full name	Neurology protein ID	Full name
ANG	Angiogenin	ADAM 22	Disintegrin and metalloproteinase domain-containing protein 22
ANGPTL3	Angiopoietin-related protein 3	ADAM 23	Disintegrin and metalloproteinase domain-containing protein 23
AOC3	Membrane primary amine oxidase	Alpha-2-MRAP	Alpha-2-macroglobulin receptor-associated protein
APOM	Apolipoprotein M	BCAN	Brevican core protein
C1QTNF1	Complement C1q tumour necrosis factor-related protein 1	Beta-NGF	Beta-nerve growth factor
C2	Complement C2	BMP-4	Bone morphogenetic protein 4
CA1	Carbonic anhydrase 1	CADM3	Cell adhesion molecule 3
CA3	Carbonic anhydrase 3	CD200	OX-2 membrane glycoprotein
CA4	Carbonic anhydrase 4	CD200R1	Cell surface glycoprotein CD200 receptor 1

Table 1: Proteins measured in Understanding Society *(continued)*

Cardiometabolic protein ID	Full name	Neurology protein ID	Full name
CCL14	C-C motif chemokine 14	CD38	ADP-ribosyl cyclase/cyclic ADP-ribose hydrolase 1
CCL18	C-C motif chemokine 18	CDH3	Cadherin-3
CCL5	C-C motif chemokine 5	CDH6	Cadherin-6
CD46	Membrane cofactor protein	CLEC10A	C-type lectin domain family 10 member A
CD59	CD59 glycoprotein	CLEC1B	C-type lectin domain family 1 member B
CDH1	Cadherin-1	CLM-1	CMRF35-like molecule 1
CES1	Liver carboxylesterase 1	CLM-6	CMRF35-like molecule 6
CFHR5	Complement factor H-related protein 5	CNTN5	Contactin-5
CHL1	Neural cell adhesion molecule L1-like protein	CPA2	Carboxypeptidase A2
CNDP1	Beta-Ala-His dipeptidase	CPM	Carboxypeptidase M
COL18A1	Collagen alpha-1(XVIII) chain	CRTAM	Cytotoxic and regulatory T-cell molecule
COMP	Cartilage oligomeric matrix protein	CTSC	Dipeptidyl peptidase 1
CR2	Complement receptor type 2	CTSS	Cathepsin S
CRTAC1	Cartilage acidic protein 1	DDR1	Epithelial discoidin domain-containing receptor 1
CST3	Cystatin-C	Dkk-4	Dickkopf-related protein 4
DEFA1	Neutrophil defensin 1	DRAXIN	Draxin
DPP4	Dipeptidyl peptidase 4	EDA2R	Tumour necrosis factor receptor superfamily member 27
EFEMP1	EGF-containing fibulin-like extracellular matrix protein 1	EFNA4	Ephrin-A4
ENG	Endoglin	EPHB6	Ephrin type-B receptor 6
F11	Coagulation factor XI	EZR	Ezrin
F7	Coagulation factor VII	FcRL2	Fc receptor-like protein 2
FAP	Prolyl endopeptidase FAP	FLRT2	Leucine-rich repeat transmembrane protein FLRT2
FCGR2A	Low affinity immunoglobulin gamma Fc region receptor II-a	G-CSF	Granulocyte colony-stimulating factor
FCGR3B	Low affinity immunoglobulin gamma Fc region receptor III-B	gal-8	Galectin-8
FCN2	Ficolin-2	GDF-8	Growth/differentiation factor 8
FETUB	Fetuin-B	GDNF	Glial cell line-derived neurotrophic factor
GAS6	Growth arrest-specific protein 6	GDNFR-alpha-3	GDNF family receptor alpha-3
GNLY	Granulysin	GFR-alpha-1	GDNF family receptor alpha-1
GP1BA	Platelet glycoprotein Ib alpha chain	GM-CSF-R-alpha	Granulocyte-macrophage colony-stimulating factor receptor subunit alpha
ICAM1	Intercellular adhesion molecule 1	GPC5	Glypican-5
ICAM3	Intercellular adhesion molecule 3	GZMA	Granzyme A
IGFBP3	Insulin-like growth factor-binding protein 3	HAGH	Hydroxyacylglutathione hydrolase; mitochondrial
IGFBP6	Insulin-like growth factor-binding protein 6	IL-5R-alpha	Interleukin-5 receptor subunit alpha
IGLC2	Immunoglobulin lambda constant 2	IL12	Interleukin-12 subunit alpha and interleukin-12 subunit beta
IL7R	Interleukin-7 receptor subunit alpha	JAM-B	Junctional adhesion molecule B
ITGAM	Integrin alpha-M	KYNU	Kynureninase
KIT	Mast/stem cell growth factor receptor Kit	LAIR-2	Leukocyte-associated immunoglobulin-like receptor 2
LCN2	Neutrophil gelatinase-associated lipocalin	LAT	Linker for activation of T-cells family member 1
LILRB1	Leukocyte immunoglobulin-like receptor subfamily B member 1	LAYN	Layilin
LILRB2	Leukocyte immunoglobulin-like receptor subfamily B member 2	LXN	Latexin

Table 1: Proteins measured in Understanding Society *(continued)*

Cardiometabolic protein ID	Full name	Neurology protein ID	Full name
LILRB5	Leukocyte immunoglobulin-like receptor subfamily B member 5	MANF	Mesencephalic astrocyte-derived neurotrophic factor
LTBP2	Latent-transforming growth factor beta-binding protein 2	MAPT	Microtubule-associated protein tau
LYVE1	Lymphatic vessel endothelial hyaluronic acid receptor 1	MATN3	Matrilin-3
MBL2	Mannose-binding protein C	MDGA1	MAM domain-containing glycosylphosphatidylinositol anchor protein 1
MEGF9	Multiple epidermal growth factor-like domains protein 9	MSR1	Macrophage scavenger receptor types I and II
MET	Hepatocyte growth factor receptor	N-CDase	Neutral ceramidase
MFAP5	Microfibrillar-associated protein 5	N2DL-2	UL16-binding protein 2
NCAM1	Neural cell adhesion molecule 1	NAAA	N-acylethanolamine-hydrolyzing acid amidase
NID1	Nidogen-1	NBL1	Neuroblastoma suppressor of tumourigenicity 1
NOTCH1	Neurogenic locus notch homolog protein 1	NCAN	Neurocan core protein
NRP1	Neuropilin-1	NEP	Neprilysin
OSMR	Oncostatin-M-specific receptor subunit beta	NMNAT1	Nicotinamide/nicotinic acid mononucleotide adenylyltransferase 1
PAM	Peptidyl-glycine alpha-amidating monooxygenase	Nr-CAM	Neuronal cell adhesion molecule
PCOLCE	Procollagen C-endopeptidase enhancer 1	NRP2	Neuropilin-2
PLA2G7	Platelet-activating factor acetylhydrolase	NTRK2	BDNF/NT-3 growth factors receptor
PLTP	Phospholipid transfer protein	NTRK3	NT-3 growth factor receptor
PLXNB2	Plexin-B2	PDGF-R-alpha	Platelet-derived growth factor receptor alpha
PRCP	Lysosomal Pro-X carboxypeptidase	PLXNB1	Plexin-B1
PROC	Vitamin K-dependent protein C	PLXNB3	Plexin-B3
PRSS2	Trypsin-2	PRTG	Protogenin
PTPRS	Receptor-type tyrosine-protein phosphatase S	PVR	Poliovirus receptor
QPCT	Glutaminyl-peptide cyclotransferase	RGMA	Repulsive guidance molecule A
REG1A	Lithostathine-1-alpha	RGMB	Repulsive guidance molecule B
REG3A	Regenerating islet-derived protein 3-alpha	ROBO2	Roundabout homolog 2
SAA4	Serum amyloid A-4 protein	RSPO1	R-spondin-1
SELL	L-selectin	SCARA5	Scavenger receptor class A member 5
SERPINA5	Plasma serine protease inhibitor	SCARB2	Lysosome membrane protein 2
SERPINA7	Thyroxine-binding globulin	SCARF2	Scavenger receptor class F member 2
SOD1	Superoxide dismutase [Cu-Zn]	sFRP-3	Secreted frizzled-related protein 3
SPARCL1	SPARC-like protein 1	Siglec-9	Sialic acid-binding Ig-like lectin 9
ST6GAL1	Beta-galactoside alpha-2;6-sialyltransferase 1	SIGLEC1	Sialoadhesin
TCN2	Transcobalamin-2	SKR3	Serine/threonine-protein kinase receptor R3
TGFBI	Transforming growth factor-beta-induced protein ig-h3	SMOC2	SPARC-related modular calcium-binding protein 2
TGFBR3	Transforming growth factor beta receptor type 3	SMPD1	Sphingomyelin phosphodiesterase
THBS4	Thrombospondin-4	SPOCK1	Testican-1
TIE1	Tyrosine-protein kinase receptor Tie-1	THY 1	Thy-1 membrane glycoprotein
TIMD4	T-cell immunoglobulin and mucin domain-containing protein 4	TMPRSS5	Transmembrane protease serine 5
TIMP1	Metalloproteinase inhibitor 1	TN-R	Tenascin-R

Table 1: Proteins measured in Understanding Society *(continued)*

Cardiometabolic protein ID	Full name	Neurology protein ID	Full name
TNC	Tenascin	TNFRSF12A	Tumour necrosis factor receptor superfamily member 12A
TNXB	Tenascin-X	TNFRSF21	Tumour necrosis factor receptor superfamily member 21
UMOD	Uromodulin	UNC5C	Netrin receptor UNC5C
VASN	Vasorin	VWC2	Brorin
VCAM1	Vascular cell adhesion protein 1	WFIKKN1	WAP Kazal immunoglobulin Kunitz and NTR domain-containing protein 1

Proteins were measured using Olink's Proximity Extension Assay (PEA), in which liquid biological samples are added to reagents that contain, for each protein target, a pair of oligonucleotide-linked antibodies which bind to the protein. This binding brings the two oligonucleotides into sufficiently close proximity for a real-time qPCR reaction to take place, yielding a numerical value (known as the "Ct value") which corresponds to the amount of protein in the sample.

The assay is performed one panel (92 proteins) at a time, in plates which contain batches of up to 88 participant samples, plus two external control samples to monitor assay performance, and six other control samples used to calibrate and normalize the measurements. There are also four internal controls in the reagents that are added to each sample, which monitor the performance at different stages of the process. In addition to standard in-house quality control procedures, reproducibility was monitored by running 70 participant samples in duplicate, i.e. twice, with one replicate in one plate and another replicate in a different plate, to which Olink were "blinded". For more information on Olink's QC procedure, see their website (https://olink.com/faq/how-is-quality-control-of-the-data-performed/). Quality control (QC) checks and data pre-processing were performed by Olink. Missing measurements were assigned as "sample failed", "assay failed" or "datapoint failed" by Olink.

The Ct value resulting from the qPCR reaction is corrected for inter-plate variability (or "batch effects") and normalized against one of the internal controls to give an arbitrary unit "NPX" (Normalized Protein eXpression) wherein one unit increase reflects a doubling in concentration. More information is available at Olink's website: https://olink.com/faq/what-is-npx/.

For more information on the biological function of each protein, please see the "Proteins" section of the *Understanding Society* proteomics user guide available on the health assessment user guides web page: https://www.understandingsociety.ac.uk/documentation/health-assessment/user-guide.

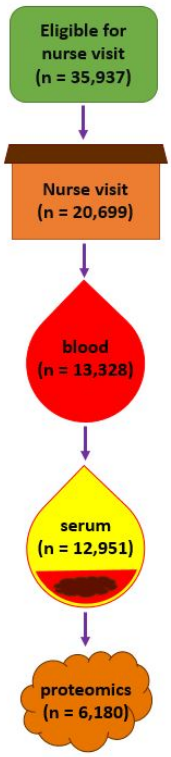

Figure 1: Samples derived from Understanding Society blood collection at waves 2 and 3 (2010-2012)

3.3 Pre-analytic factors

Nurses recorded the time of day of blood sampling; for analytic purposes, this was collapsed into a three level "time of day" variable ("morning": before midday, "afternoon": midday until 16:59, and "evening": 17:00 until midnight). "Transit delay", a continuous variable reflecting the difference, in whole days, between "interview

date" and "lab date" was calculated from dates recorded by the nurse and laboratory. Any values below 1 were set to missing. Transit delay squared was derived by exponentiating this variable to examine potential non-linear associations. Haemolysis of the sample was noted by the testing laboratory during previous biomarker testing (Newcastle upon Tyne Hospitals NHS Foundations Trust (NUTH)).

3.4 Demographic variables

Age and sex were recorded during the nurse visit. Age was scaled and centered for a mean of 0 and a standard deviation of 1. The scaled age was then exponentiated to derive a scaled-age-squared variable. Male is coded as "1" in the dataset and Female is coded as "2". Educational attainment was classified as a six-level factor based on highest qualification: "Degree", "Other higher degree", "A-level etc", "GCSE etc", "Other qualification" and "No qualification". Geographical region was measured using the Office for National Statistics (ONS) government office region category (https://www.ons.gov.uk/methodology/geography/ukgeographies/administrativegeography/england), extended to include Wales, Scotland and Northern Ireland. Ethnic group was measured with the ONS 2011 Census 17-category, self-reported ethnic group question (https://www.understandingsociety.ac.uk/documentation/mainstage/dataset-documentation/search/datafile/xwavedat) and converted to a five-level factor: "Asian", "Black", "Mixed", "Other" and "White", as per Easterbrook et al (2019) [11].

3.5 Statistical analysis

Unpaired t-tests and chi-square tests were used to assess whether the *Understanding Society* proteomics subsample differed significantly from the non-proteomics blood subsample in terms of time of day, transit delay, haemolysis, age, sex, educational attainment, geographical region and ethnicity. ANOVA was used to assess whether transit delay was associated with time of day, haemolysis, sex, educational attainment, geographical region and ethnicity, and linear regression was used to assess whether it was associated with age. An alpha of 0.05 was used to indicate significance.

For the 70 duplicated samples, we measured the reproducibility of protein measurements in accordance with Olink [12] using Pearson's correlation coefficient. Values should be positively correlated, exceeding 0.8, i.e. having a "very strong" correlation [13] .

In all of the following analyses, a Bonferroni-adjusted significance threshold of 0.05/184 was used in order to ensure the chance of inferring that at least one protein is associated remains below 0.05, despite 184 multiple comparisons. Samples with a transit delay greater than one week were excluded.

For each protein, GAMLSS (Generalised Additive Models for Location Scale and Shape) and the likelihood ratio test were used to explore the relationship between transit delay and protein measurements. GAMLSS can be used to describe how the mean and variance of a variable (in this case, a protein) varies according to another variable (in this case, transit delay) by estimating values for "mu" (central tendency, mean) and "sigma" (spread, variance). In order to detect potential non-linear relationships, transit delay was synthesized into multiple dummy variables. Pairwise comparison of models that include decreasing numbers of dummy variables were performed using stepwise backward model selection: for each protein, "model 1" was specified with dummy variables for 2, 3, 4, 5, 6, and 7 days, as explanatory for both mu and sigma. "Model 2" was similar except the dummy variable for "2" was removed, effectively merging delays of 1 and 2 days. A likelihood ratio test was performed on the two models, effectively testing whether or not the merging of the first two delay lengths made a significant difference to the estimate of spread, compared to treating them as two separate levels. The GAMLSS coefficients were used to generate an estimated effect size corresponding to a further day's increase in delay (i.e. 2 days) "% change in standard deviation (SD)":

$$100 * ((SD_{TD2} - SD_{TD1}) / SD_{TD1})$$

where SD_{TD1} is the exponent of the sigma coefficient for the intercept, and SD_{TD2} is the exponent of the sum of the sigma coefficients for both the intercept and the dummy variable for 2 days' delay. This approach was repeated, each time removing another dummy variable to effectively merge the lowest few delay lengths.

In the remaining analyses, all continuous variables (protein NPX values, transit delay, transit delay squared, age and age-squared) were scaled and centered for a mean of 0 and a standard deviation of 1.

For each protein, bivariate analyses (t-test, analysis of variance (ANOVA), linear regression) and multivariate regression analyses were conducted to measure their association with a range of factors. T-tests were conducted to examine haemolysis and sex. Effect size is calculated as Cohen's *d* (mean difference/SD). Bivariate one-way ANOVAs were conducted for categorical factors: time of day, educational attainment, government office region, and ethnicity. The effect size reported is eta squared (variance explained). Linear regression analyses were used to examine transit delay, transit delay squared, age and age squared (for the squared variables, the linear counterpart was included as the sole covariate). We report the beta coefficient.

To assess the importance of adjustment for pre-analytic factors, likelihood ratio tests were used to compare bivariate linear regressions of the proteins and demographic variables with corresponding multivariate regressions (i.e. adjusted for transit delay and transit delay squared). We report the p value from the likelihood ratio tests, and the increase in R-squared as a percent of the total variance. To gauge whether adjustment for pre-analytic factors changes the magnitude and significance of proteins' associations with educational attainment, we repeated these analyses using ANOVA and ANCOVA. The measures of variance explained for linear regression (R squared) and ANOVA (eta squared) were identical (Pearson's correlation coefficient >0.99) but ANOVA benefits from providing a single p value for the educational attainment term. When specifying the ANCOVA models, educational attainment was ordered last in order to examine its marginal effect. To obtain unfixed confidence intervals, a two-sided alternative hypothesis was used.

4 Results

4.1 Participant characteristics

Pre-analytic factors and participant characteristics are shown in table 2. In the subsample of blood used for proteomic analyses, time of day of sample collection was significantly different, transit delay was significantly shorter, and haemolysis was around half as common, compared to the non-proteomics subsample. Participants with proteomics measurements were also older and more female, and differed according to geographical regions and educational attainment, when compared to the non-proteomics subsample. However, the proportions of participants in each of the ethnic groups was not significantly different.

4.2 Proteomic dataset quality checks

The duplicated samples produced reproducible measurements: Pearson's correlation coefficients were all very strong, falling between 0.86 – 1.0 for each duplicate. No samples failed in both panels. 114 samples (1.8%) failed in one panel: 103 (1.7%) failed in the neurology panel, whereas only 11 (0.2%) failed in the cardiometabolic panel. Therefore, only 0.9% of all datapoints were removed due to sample failure. Missing values *not* due to sample failure are extremely rare (0.03% of datapoints) and the vast majority result from assay failure in one protein, DEFA1, across four plates (341 datapoints; 5.5% of DEFA1 values). Only 16 values are missing due to "datapoint failure". 93.3% of datapoints across the dataset fell above the limit of detection for the corresponding protein. 146/184 proteins have a below-LOD rate less than 0.5%. 18/184 proteins (9.8%) have a high below-LOD rate (>=10%) and thus are defined as "low detectability" proteins. Thirteen of these are from the cardiometabolic panel, and five are from the neurology panel (table 3). 96.0% of the below-LOD values in the dataset are from these low detectability proteins.

The distributions of NPX values for each protein are all broadly normal, however some proteins appear to have slight skew and/or a bimodal distribution. For boxplots, see figures 2 and 3, and for histograms and q-q plots, see the proteomics user guide's "Proteins" section, available on the *Understanding Society* health assessment user guides web page: https://www.understandingsociety.ac.uk/documentation/health-assessment/user-guide.

Table 2: Comparison of pre-analytic factors and characteristics of participants that provided blood samples with and without protein data

	In protein sample	In blood sample (but not protein)		
	N=6,180	N=7,095		
Pre-analytic factors				
Time of day			**p=<0.001**	***
Morning	1,957 (31.7)	2,152 (30.3)		
Afternoon	2,261 (36.6)	2,473 (34.9)		
Evening	1,962 (31.7)	2,470 (34.8)		
Transit delay (days)	**2.5 (1.6), 1-32**	**2.6 (1.6), 1-45**	**p=<0.001**	***
Haemolysis	**182 (3.0)**	**409 (6.0)**	**p=<0.001**	***
Demographic factors				
Age (y)	**53.1 (17.7), 16-102**	**50.9 (16.7), 16-99**	**p=<0.001**	***
Men	**2,453 (36.7)**	**3,474 (49.0)**	**p=<0.001**	***
Educational attainment			**p=<0.001**	***
Degree	1,320 (21.6)	1,549 (22.1)		
Other higher degree	779 (12.8)	900 (12.8)		
A-levels, etc	1,156 (18.9)	1,353 (19.3)		
GCSEs, etc	1,216 (19.9)	1,526 (21.7)		
Other qualification	667 (10.9)	787 (11.2)		
No qualifications	968 (15.9)	911 (13.0)		
Region			**p=<0.001**	***
North East	317 (5.1)	342 (4.8)		
North West	699 (11.3)	846 (11.9)		
Yorkshire and the Humber	562 (9.1)	627 (8.8)		
East Midlands	563 (9.1)	560 (7.9)		
West Midlands	520 (8.4)	598 (8.4)		
East of England	558 (9.0)	746 (10.5)		
London	400 (6.5)	515 (7.3)		
South East	874 (14.2)	1,047 (14.8)		
South West	649 (10.5)	712 (10.0)		
Wales	501 (8.1)	463 (6.5)		
Scotland	533 (8.6)	633 (8.9)		
Northern Ireland	0 (0.0)	0 (0.0)		
Ethnicity			**p=0.53**	
Asian	155 (2.5)	188 (2.7)		
Black	61 (1.0)	68 (1.0)		
Mixed	39 (0.6)	60 (0.9)		
Other Ethnicity	15 (0.2)	23 (0.3)		
White	5,910 (95.6)	6,745 (95.1)		

Values shown are either ’mean (SD), range’ or ’n (%)’

Table 3: 'Low detectability' proteins

Protein	Below LOD (%)	Panel
CES1	54.0	cardiometabolic
DEFA1	37.2	cardiometabolic
FAP	93.8	cardiometabolic
GNLY	17.7	cardiometabolic
IL7R	23.0	cardiometabolic
ITGAM	79.4	cardiometabolic
LTBP2	98.7	cardiometabolic
MFAP5	67.4	cardiometabolic
PLA2G7	74.5	cardiometabolic
PLTP	66.8	cardiometabolic
REG3A	99.4	cardiometabolic
SOD1	21.3	cardiometabolic
UMOD	74.6	cardiometabolic
Beta-NGF	99.7	neurology
G-CSF	19.8	neurology
GDNF	88.1	neurology
LXN	85.5	neurology
MAPT	99.8	neurology

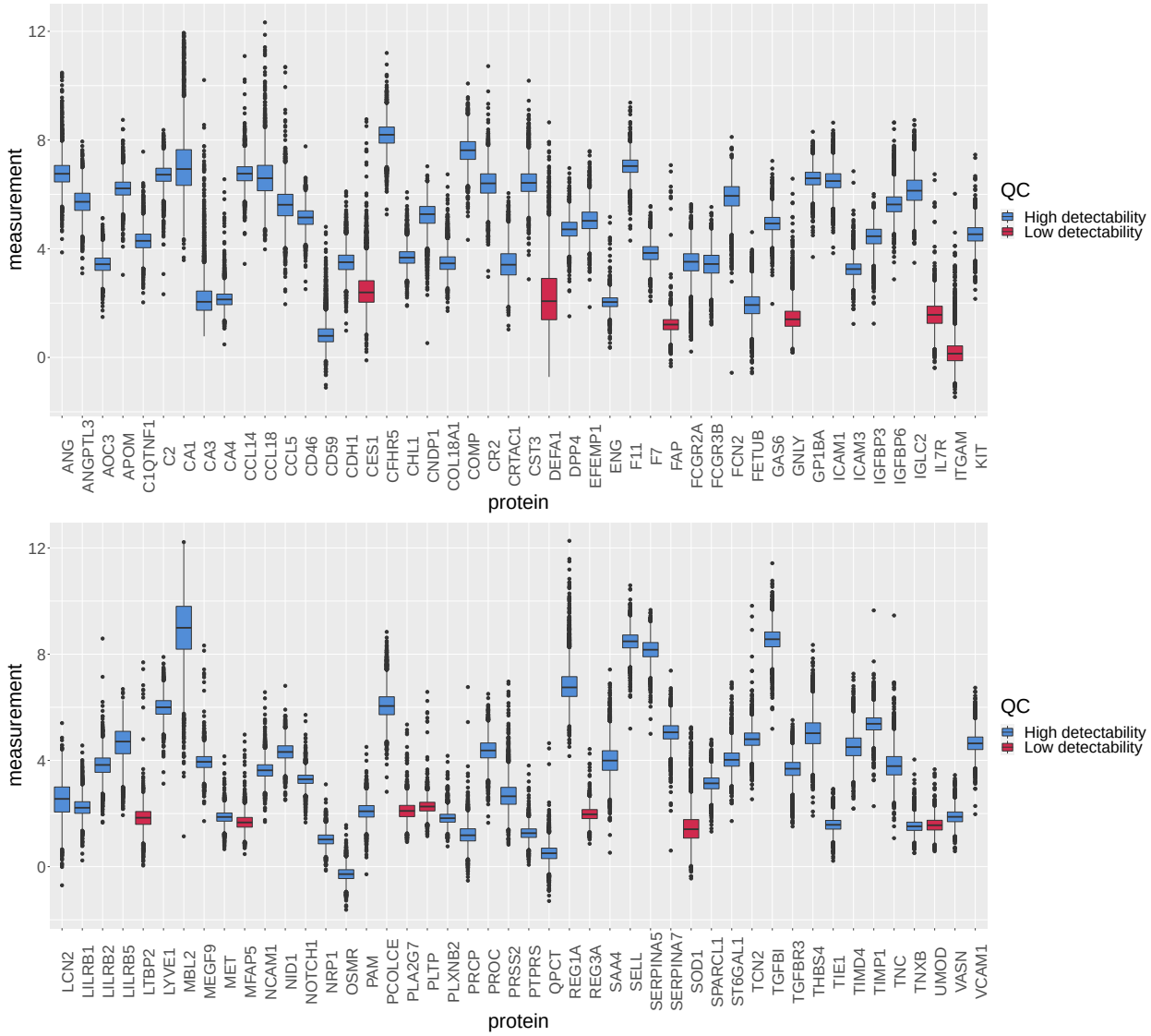

Figure 2: Cardiometabolic panel protein distributions

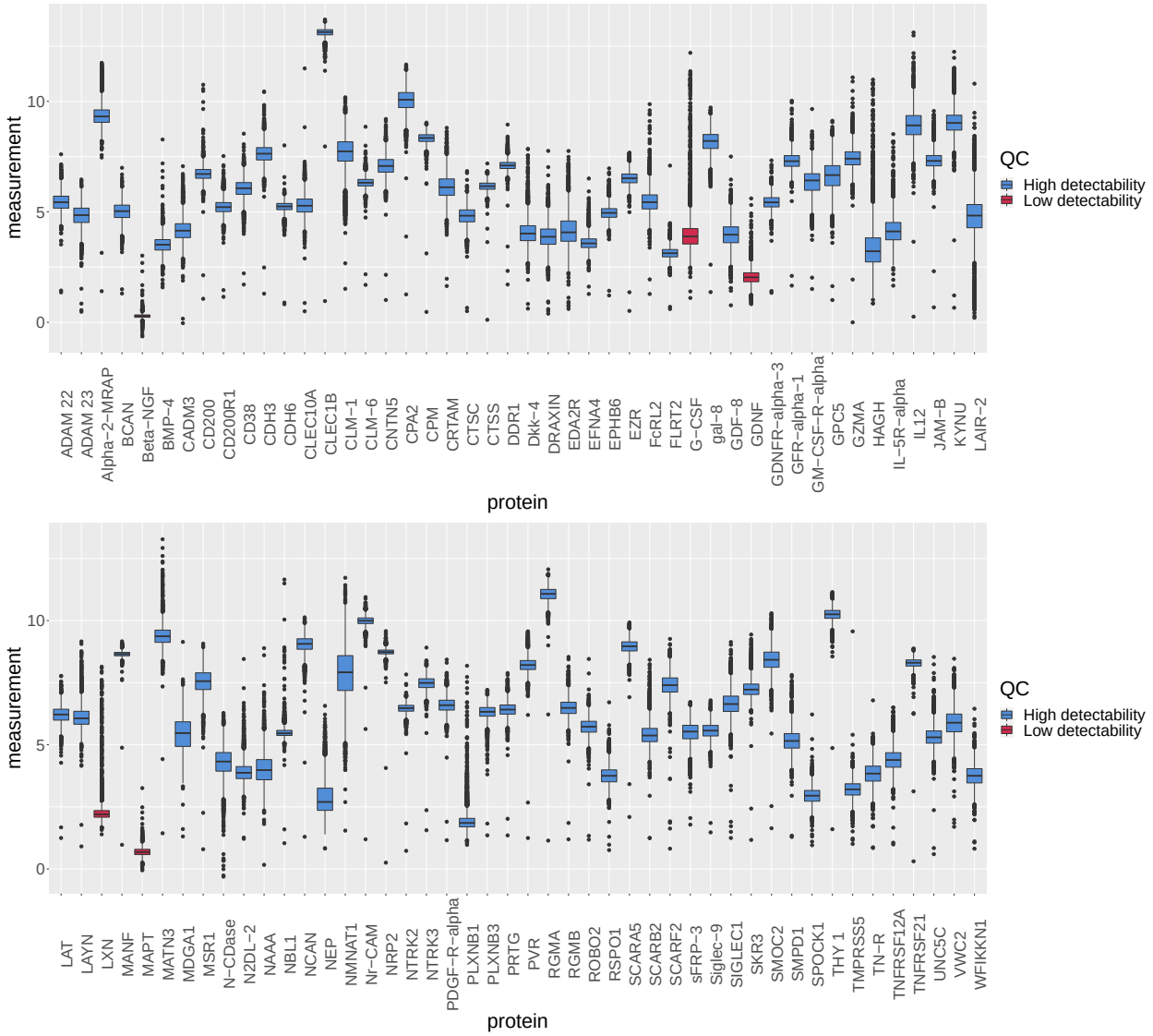

Figure 3: Neurology panel protein distributions

For each protein, we identified the minimum number of days' transit delay which introduces a statistically significant change in spread, or standard deviation - see table 4 for the number of proteins affected by each delay length. Some of the statistically significant changes in spread were arguably of small size, so we also provide counts from analyses that only take into account changes in spread that are larger than 10% or 25%. For 51 proteins, there was no significant association between transit delay and spread. At the 10% threshold this increases to 108, and at the 25% threshold this increases to 156. Appendix 1 includes a table listing the proteins in each cell of table 4, along with boxplots showing the spread of four example proteins, including one which appears to be influenced by outliers.

Table 4: Transit delay cutoffs affecting spread: protein counts

Minimum transit delay affecting spread (days)	n proteins affected		
	No SD % change cutoff	10% SD change cutoff	25% SD change cutoff
2	102	33	8
3	14	9	2
4	13	8	0
5	3	14	5
6	0	7	2
7	1	5	11
None (i.e. SD is stable)	51	108	156

4.3 Association of protein levels with pre-analytic and demographic covariates

Table 5 contains results for unadjusted T tests, ANOVA and regression analyses of pre-analytic and demographic variables, and table 6 contains results from likelihood ratio tests comparing the unadjusted analyses of demographic variables with those adjusted for pre-analytic factors transit delay and transit delay squared.

4.3.1 Unadjusted analyses of protein levels

Table 5: Table of protein associations with pre-analytic and demographic factors

	Time_of_day	Transit_delay	Transit_delay_squared	Haemolysis	Age	Age_squared	Sex	Education	Region	Ethnicity
ADAM_22	***				***	***	***	***		***
ADAM_23	***				***	***	***	***		
Alpha_2_MRAP		***		***	***		***	**	*	
ANG					***	*	***	**		
ANGPTL3	***	**	***		***		***	***		***
AOC3	***				***			***		**
APOM	*	*				***		*		
BCAN	***	**	***		***		***	***		
Beta_NGF		***								
BMP_4					***		***	**		
C1QTNF1	***				***		***	***		
C2		***			***	***	**	***		
CA1	***	***	***	***			***			
CA3		***	***	***	***	**	***	***		
CA4	***	***	***		***					

Table 5: Table of protein associations with pre-analytic and demographic factors *(continued)*

	Time_of_day	Transit_delay	Transit_delay_squared	Haemolysis	Age	Age_squared	Sex	Education	Region	Ethnicity
CADM3	***				***	***	***	***		
CCL14	***	***	***		***	***		***		***
CCL18	***		*		***	**	***	***		
CCL5	***	***					***		**	***
CD200	***				***	***		***		
CD200R1		***			***					***
CD38		***	***	***	***	***	***	***		
CD46	***	***	***		***				***	
CD59		***	***	***	***	***	***	***		
CDH1	***		**		***	***	***	***		
CDH3	***	*				***	***			***
CDH6	*	***			***	***		***		**
CES1		***			***	***	***	***		**
CFHR5							***			
CHL1					***					
CLEC10A					***	***		***		
CLEC1B	***	***	***		***				***	
CLM_1	***	***		***	***	***	***	***		***
CLM_6	*	***			***	***	**	***		***
CNDP1	***	**			***	***	*	***		
CNTN5					***		***	***		
COL18A1	***	***			***	***	***	***		***
COMP	***				***		***	***		**
CPA2	***	***			***					
CPM	***				***	***	***	***		***
CR2	***	***	***		***	***		***		*
CRTAC1	***		**		***			***		
CRTAM	***				***	***	**	***		
CST3	***				***	***	***	***		
CTSC	***	***	***	***	*		***		***	*
CTSS	***	***	***		***	***			**	
DDR1	**	***			***	**	***	***		
DEFA1	***	***	***	***	***				***	
Dkk_4	***				***	***	***	***		
DPP4					***		***		*	
DRAXIN	***	***			***	***		***		
EDA2R	***	***	***		***	***	***			***
EFEMP1	***		**		***	***		***		
EFNA4	**	***			***	***	***	***		
ENG	**	**					***			

Table 5: Table of protein associations with pre-analytic and demographic factors *(continued)*

	Time_of_day	Transit_delay	Transit_delay_squared	Haemolysis	Age	Age_squared	Sex	Education	Region	Ethnicity
EPHB6	***	**			***	***	**	***		
EZR	***	***	***	***					***	
F11						***	***			
F7	***	***			***	*	***	***		
FAP							*			
FCGR2A		***			***			***		***
FCGR3B	***	***			**		***			
FCN2					**	***	**			
FcRL2		**			***		*	***		
FETUB	***	*			***		***			**
FLRT2	***	***			***	***		***		
G_CSF		***	***	**			***			
gal_8		***	***	***	*				***	
GAS6	***				***		**	***		
GDF_8	***	***	***		***		***	***		*
GDNF	***	***		*	***		***	***		**
GDNFR_alpha_3	***				***	***		***		
GFR_alpha_1	***		***		***	***		***		***
GM_CSF_R_alpha							**	*		*
GNLY		***			***		*	***		
GP1BA	***	***	***		***					
GPC5	**	***			***			**		***
GZMA	***	***	***	***				*	*	
HAGH	***	***	*	***			***		***	
ICAM1	***				***			***		
ICAM3		***			***					
IGFBP3	***	***	**		***		***	***		*
IGFBP6	***	*			***	***	***	***		
IGLC2	***		*		***	***	***	***		
IL_5R_alpha	***	***	**		***		***	***		***
IL12	**				***	***	***	***		*
IL7R	***	***						*		
ITGAM	***	***	***	***	***				***	
JAM_B	***	**	*		***	***	***	***		
KIT	***	***	***		***	***		***		
KYNU	***	***	*	***	***	***	***	***		
LAIR_2		*			***	***		***		
LAT	***	***	**		***		***	***		**
LAYN	***		*		***	***		***		***
LCN2	***	***	***	***	***				***	

Table 5: Table of protein associations with pre-analytic and demographic factors *(continued)*

	Time_of_day	Transit_delay	Transit_delay_squared	Haemolysis	Age	Age_squared	Sex	Education	Region	Ethnicity
LILRB1	**	**			***			***		
LILRB2		***		**	***			***		
LILRB5					***		**			
LTBP2	***				***	**		***		
LXN		***	***	***			*			
LYVE1					***		***			
MANF	***	***	***		***					
MAPT		***		*						
MATN3		***				***	***			***
MBL2					**	***	***			*
MDGA1					***		***			***
MEGF9	***	*			***		*	***		
MET	*									
MFAP5	***	***			***	***	***	***		
MSR1	***		*		***		***	***		
N_CDase					***	***	***			***
N2DL_2	***	***			***	***		***		***
NAAA	***	***	*	***	*		***		***	
NBL1		***					**			
NCAM1					***	***	***			
NCAN		*			***	***	***	***		***
NEP		***				***	***			*
NID1	*					***	***	***		
NMNAT1	***	***	***	***					***	
NOTCH1		***			**					
Nr_CAM		***				***	***			
NRP1	**				***	***	***	***		
NRP2		***			***					
NTRK2		***			***	***		***		
NTRK3		*			***		***	***	***	
OSMR	***				***		*	***		
PAM					***		***	*		
PCOLCE	***		**		***		***	***		
PDGF_R_alpha	***	***			***	***	***	***		
PLA2G7		***				***	***			
PLTP					***		***			
PLXNB1		***			***			**		
PLXNB2	***				***		***	***		
PLXNB3	***	***	***		***	***	***		***	**
PRCP	***	***		*	***		***	***		

Table 5: Table of protein associations with pre-analytic and demographic factors *(continued)*

	Time_of_day	Transit_delay	Transit_delay_squared	Haemolysis	Age	Age_squared	Sex	Education	Region	Ethnicity
PROC					***	***		***		
PRSS2	*	***			***	***		***		
PRTG					***		***			
PTPRS	***	*			***	***				
PVR					***	***	*	**		***
QPCT	***	***				***	***	**		
REG1A	***		***		***	***	***	***		
REG3A					***	***		***		
RGMA	***	***	***		***	***	***	***		
RGMB	***				***	***	***	***		
ROBO2	***				***		***	***		
RSPO1	***	***	***	*	***	***		***		*
SAA4					***	**	***	***		**
SCARA5	***				***	***	***	***		
SCARB2	***	***			***	***	***	***		
SCARF2	***		***		***	***		***		***
SELL	***	***	***		***		***	***		
SERPINA5	*	***				***	***			**
SERPINA7		**					***	***		
sFRP_3	***	***			***		***	***		***
Siglec_9		***		**	***		**	**		***
SIGLEC1	***				***	***		***		
SKR3	***				***	***	***	***		
SMOC2	***				***	***	***	***		
SMPD1	***				***		***	***		
SOD1	***	***		***	***		***		***	
SPARCL1	**				***	**		***		***
SPOCK1	***	***			***	***	***			
ST6GAL1		**			***			***		
TCN2	**				***	**	**	***		
TGFBI	**				*		***			
TGFBR3	***	***			***	***	***	***		
THBS4	***				***			***		
THY_1	***	***			***	***	***	***		
TIE1										**
TIMD4	***				***	*	*	***		
TIMP1	*				***	*	***	***		*
TMPRSS5					***			***		
TN_R	***						***	*		***
TNC	**				***	***		**		

Table 5: Table of protein associations with pre-analytic and demographic factors *(continued)*

	Time_of_day	Transit_delay	Transit_delay_squared	Haemolysis	Age	Age_squared	Sex	Education	Region	Ethnicity
TNFRSF12A	***	***	***		***	***	***	***		
TNFRSF21	***				***	***		***		**
TNXB	***	***	*		***			***		***
UMOD	***				***	***	***	***		
UNC5C	***		***		***	***	***	***		***
VASN	*				***	***	***	***		
VCAM1	***				***	***	***	***		***
VWC2	***	***	***	***	***	***	***	***	***	***
WFIKKN1	**	***			***		***	***		

* p<0.05/184, ** p<0.01/184, *** p<0.001/184 from T tests, ANOVA and regression analyses of pre-analytic and demographic variables. Each row represents one protein, and each column represents a factor. Effect sizes are represented in colour form, with white corresponding to 0. There are three effect size scales: one for T-tests (haemolysis and sex), one for ANOVAs (time of day, educational attainment, geographical region and ethnicity), and one for linear regressions (transit delay, transit delay squared, age and age squared). The T test scale ranges from dark blue for negative Cohen's d values (mean difference in standardized NPX units; minimum value is -0.73) to dark red for positive values (capped at 2.00 due to high values of Cohen's d for six proteins' association with haemolysis - CA1 (3.4), CA3 (3.7), CD59 (2.1), HAGH (3.1), LXN (5.8), SOD1 (3.4) -, which would have skewed the colour scale leaving most cells almost white). Red indicates higher levels in females compared to males, or higher levels in haemolysed samples compared to non-haemolysed samples. For ANOVAs, darker grey shading rep-resents a larger value (min 5.7e-05, max 0.14). The regression coefficient scale ranges from dark blue for negative beta coefficients (minimum -0.54) to dark red for positive ones (maximum 0.77). n ranged from 5,648 to 6,050

In the unadjusted analyses, some factors are associated with the majority of the 184 proteins: age (154; 164 including age squared), educational attainment (127), time of day (126), sex (121) and transit delay (106; 118 including transit delay squared). Other factors are associated with fewer proteins: haemolysis (27), ethnicity (53) and government office region (20).

All of the haemolysis-associated proteins were associated with transit delay and/or transit delay squared. Of the 18 "low detectability" proteins, an association with educational attainment was detected for eight.

4.3.2 Associations between transit delay and other variables

Time of day was significantly associated with transit delay (p = 1.69e-33), as was haemolysis (p = 3.17e-201), age (beta coefficient = -0.084, p = 4.93e-11), education (p = 7.50e-04) and government office region (p = 1.99e-13). Neither sex nor ethnicity were associated with transit delay.

4.3.3 Adjusted analyses of protein levels

Table 6: Table of unadjusted versus adjusted model comparisons for protein-demographic factor associations

	Age_adjusted	Age_squared_adjusted	Sex_adjusted	Education_adjusted	Region_adjusted	Ethnicity_adjusted
ADAM_22	*	***				
ADAM_23						
Alpha_2_MRAP	***	***	***	***	***	***
ANG						
ANGPTL3			***	***	***	***
AOC3						
APOM	*		**			**
BCAN			***	***	***	***
Beta_NGF	***	***	***	***	***	***
BMP_4						
C1QTNF1						
C2	***	***	***	***	***	***
CA1	***	***	***	***	***	***
CA3	***	***	***	***	***	***
CA4	***	***	***	***	***	***
CADM3						
CCL14	***	***	***	***	***	***
CCL18			***		***	***
CCL5	***	***	***	***	***	***
CD200	***	***				
CD200R1	***	***	***	***	***	***
CD38	***	***	***	***	***	***
CD46	***	***	***	***	***	***
CD59	***	***	***	***	***	***
CDH1			***	*	***	***
CDH3	***	**	***	**	**	***
CDH6	***	***	***	***	***	***
CES1	***	***	***	***	***	***
CFHR5						
CHL1						
CLEC10A	***	***				

Table 6: Table of unadjusted versus adjusted model comparisons for protein-demographic factor associations *(continued)*

	Age_adjusted	Age_squared_adjusted	Sex_adjusted	Education_adjusted	Region_adjusted	Ethnicity_adjusted
CLEC1B	***	***	***	***	***	***
CLM_1	***	***	***	***	***	***
CLM_6	***	***	***	***	***	***
CNDP1			**		*	*
CNTN5						
COL18A1	***	***	***	***	***	***
COMP	***	***				
CPA2	***	***	***	***	***	***
CPM	***	***				
CR2			***	***	***	***
CRTAC1			*			*
CRTAM	***	***				
CST3			**		*	*
CTSC	***	***	***	***	***	***
CTSS	***	***	***	***	***	***
DDR1	***	***	***	***	***	***
DEFA1	***	***	***	***	***	***
Dkk_4	***	***				
DPP4						
DRAXIN	***	***	***	***	***	***
EDA2R		***	***	***	***	***
EFEMP1			***		***	***
EFNA4	***	***	***	***	***	***
ENG	*	*	*	**	*	**
EPHB6	***	***	**	***	**	**
EZR	***	***	***	***	***	***
F11						
F7	***	***	***	***	***	***
FAP						
FCGR2A	***	***	***	***	***	***
FCGR3B	***	***	***	***	***	***
FCN2		*				
FcRL2	***	***	*	**	*	*
FETUB			***	***	***	***
FLRT2	***	***	***	***	***	***
G_CSF	***	***	***	***	***	***
gal_8	***	***	***	***	***	***
GAS6						
GDF_8	**	**	***	***	***	***
GDNF	***	***	***	***	***	***

Table 6: Table of unadjusted versus adjusted model comparisons for protein-demographic factor associations *(continued)*

	Age_adjusted	Age_squared_adjusted	Sex_adjusted	Education_adjusted	Region_adjusted	Ethnicity_adjusted
GDNFR_alpha_3	***	***				
GFR_alpha_1			**		*	**
GM_CSF_R_alpha						
GNLY	***	***	***	***	***	***
GP1BA	***	***	***	***	***	***
GPC5	***	***	***	***	***	***
GZMA	***	***	***	***	***	***
HAGH	***	***	***	***	***	***
ICAM1						
ICAM3	***	***	***	***	***	***
IGFBP3	***	***	***	***	***	***
IGFBP6			***		***	***
IGLC2						
IL_5R_alpha			***	***	***	***
IL12	**	***		*		
IL7R	***	***	***	***	***	***
ITGAM	***	***	***	***	***	***
JAM_B			***		***	***
KIT	***	***	***	***	***	***
KYNU	***	***	***	***	***	***
LAIR_2	***	***		*		
LAT	***	***	***	***	***	***
LAYN	***	***				
LCN2	***	***	***	***	***	***
LILRB1	***	***	**	***	*	*
LILRB2	***	***	***	***	***	***
LILRB5						
LTBP2						
LXN	***	***	***	***	***	***
LYVE1	*	**				
MANF	***	***	***	***	***	***
MAPT	***	***	***	***	***	***
MATN3	***	**	***	***	***	***
MBL2						
MDGA1						
MEGF9			*			*
MET						
MFAP5			***	***	***	***
MSR1	***	***				
N_CDase						

Table 6: Table of unadjusted versus adjusted model comparisons for protein-demographic factor associations *(continued)*

	Age_adjusted	Age_squared_adjusted	Sex_adjusted	Education_adjusted	Region_adjusted	Ethnicity_adjusted
N2DL_2	***	***	***	***	***	***
NAAA	***	***	***	***	***	***
NBL1	***	***	***	***	***	***
NCAM1		*				
NCAN	***	***	*	**	*	*
NEP	***	***	**	***	**	***
NID1						
NMNAT1	***	***	***	***	***	***
NOTCH1	***	***	***	***	***	***
Nr_CAM	***	***	***	***	***	***
NRP1						
NRP2	***	***	***	***	***	***
NTRK2	***	***	***	***	***	***
NTRK3	***	***		*		*
OSMR						
PAM	*	*				
PCOLCE			***		**	**
PDGF_R_alpha	***	***	**	***	***	***
PLA2G7	***	***	***	***	**	***
PLTP						
PLXNB1	***	***	***	***	***	***
PLXNB2	**	**				
PLXNB3	***	***	***	***	***	***
PRCP	***	***	***	***	***	***
PROC						
PRSS2	*	**	***	***	***	***
PRTG	*	**				
PTPRS		*	**	**	**	**
PVR						
QPCT	***	***	***	***	***	***
REG1A			***	***	***	***
REG3A						
RGMA	***	***	***	***	***	***
RGMB	***	***				
ROBO2						
RSPO1	***	***	***	***	***	***
SAA4						
SCARA5			***		*	**
SCARB2	***	***	***	***	***	***
SCARF2	*	**	***		***	***

Table 6: Table of unadjusted versus adjusted model comparisons for protein-demographic factor associations *(continued)*

	Age_adjusted	Age_squared_adjusted	Sex_adjusted	Education_adjusted	Region_adjusted	Ethnicity_adjusted
SELL	*	*	***	***	***	***
SERPINA5	***	***	***	***	***	***
SERPINA7			**	**	*	*
sFRP_3			***	*	**	***
Siglec_9	***	***	***	***	***	***
SIGLEC1						
SKR3	*	**				
SMOC2	**	***				
SMPD1						
SOD1	***	***	***	***	***	***
SPARCL1						
SPOCK1	***	***	***	***	***	***
ST6GAL1	***	***	**	***	**	**
TCN2						
TGFBI						
TGFBR3			***	*	***	***
THBS4						
THY_1	***	***	***	***	***	***
TIE1						
TIMD4						
TIMP1	***	***				
TMPRSS5	***	***				
TN_R				*		
TNC		*				
TNFRSF12A	***	***	***	***	***	***
TNFRSF21	**	***				
TNXB	*	*	***	***	***	***
UMOD						
UNC5C			***		***	***
VASN						
VCAM1						
VWC2	***	***	***	***	***	***
WFIKKN1	***	***	***	***	***	***

* p<0.05/184, ** p<0.01/184, *** p<0.001/184 from likelihood ratio tests comparing unadjusted models vs models adjusting for linear and non-linear transit delay. Each row represents one protein, and each column represents a factor. Significance indicates that pre-analytic factors contributed to the model. Change in R-squared, i.e. the extra % of the variance which is explained by adding pre-analytic covariates is represented in colour form, where darker grey shading represents a larger value (min 0.0006, max 37.4). n ranged from 5,648 to 5,979.

For 142 proteins, adjustment for pre-analytic factors improved model fit for at least one of the variables studied. For 88 proteins, model fit was improved across all six variables. Proteins for which this adjustment made no improvements were: ADAM_23, ANG, AOC3, BMP_4, C1QTNF1, CADM3, CFHR5, CHL1, CNTN5, DPP4, F11, FAP, GAS6, GM_CSF_R_alpha, ICAM1, IGLC2, LILRB5, LTBP2, MBL2, MDGA1, MET, N_CDase, NID1, NRP1, OSMR, PLTP, PROC, PVR, REG3A, ROBO2, SAA4, SIGLEC1, SMPD1, SPARCL1, TCN2, TGFBI, THBS4, TIE1, TIMD4, UMOD, VASN and VCAM1. Addition of pre-analytic covariates improved model fit for educational attainment in 40 of the 57 proteins *not* associated with education in the unadjusted analyses. In an ANCOVA adjusted for pre-analytic factors, the association with educational attainment was replicated for 124 proteins, and a further seven proteins crossed our nominal threshold for significance (MAPT, ICAM3, LYVE1, MDGA1, CFHR5, CTSC and CCL5), while three proteins became non-significant (IL7R, APOM and GPC5). Effect sizes were broadly very similar as shown in figs 4-5 (top 92 proteins by effect size) and appendix 2, figs 7-8 (bottom 92 proteins).

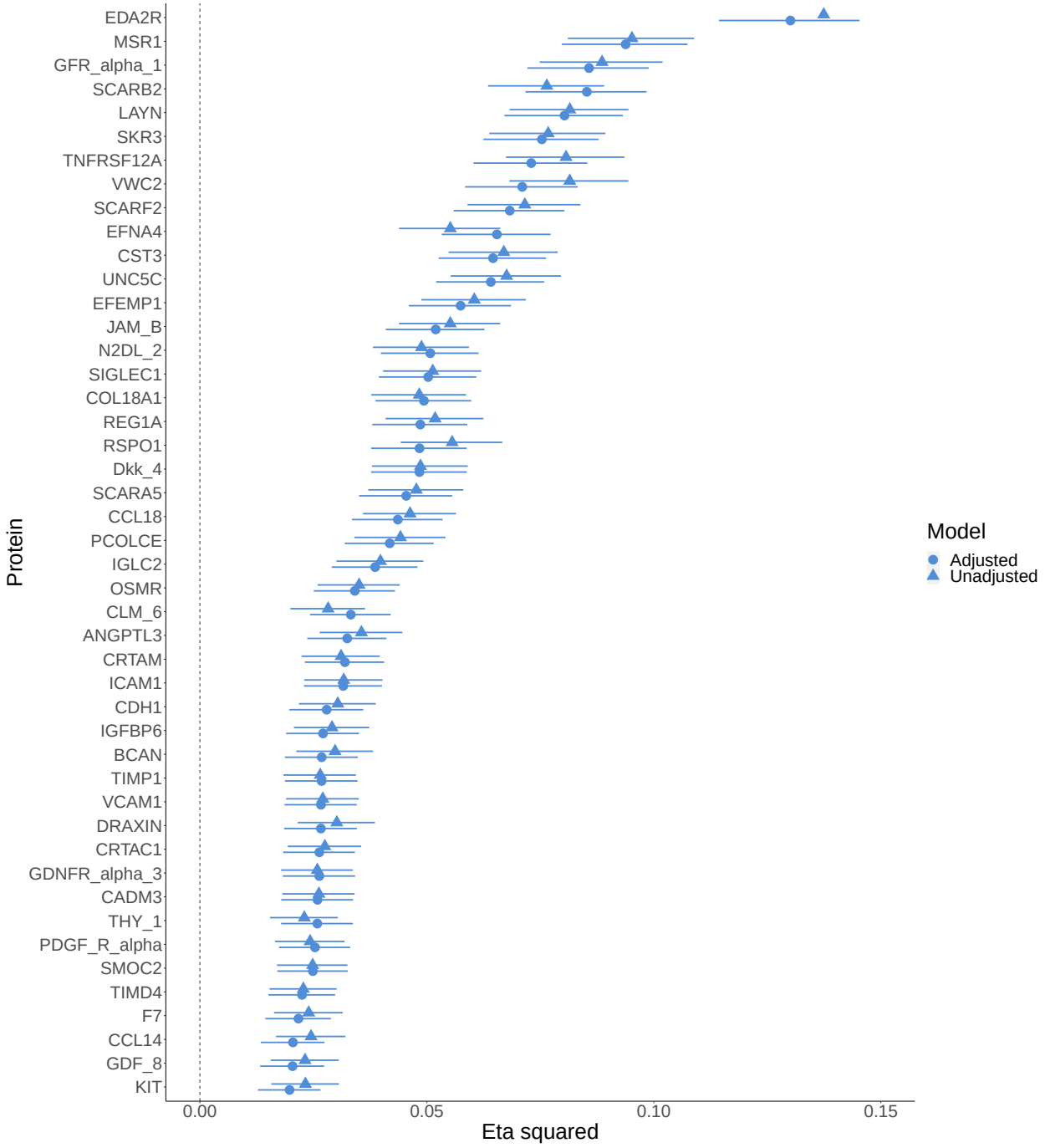

Figure 4: Associations between proteins and educational attainment

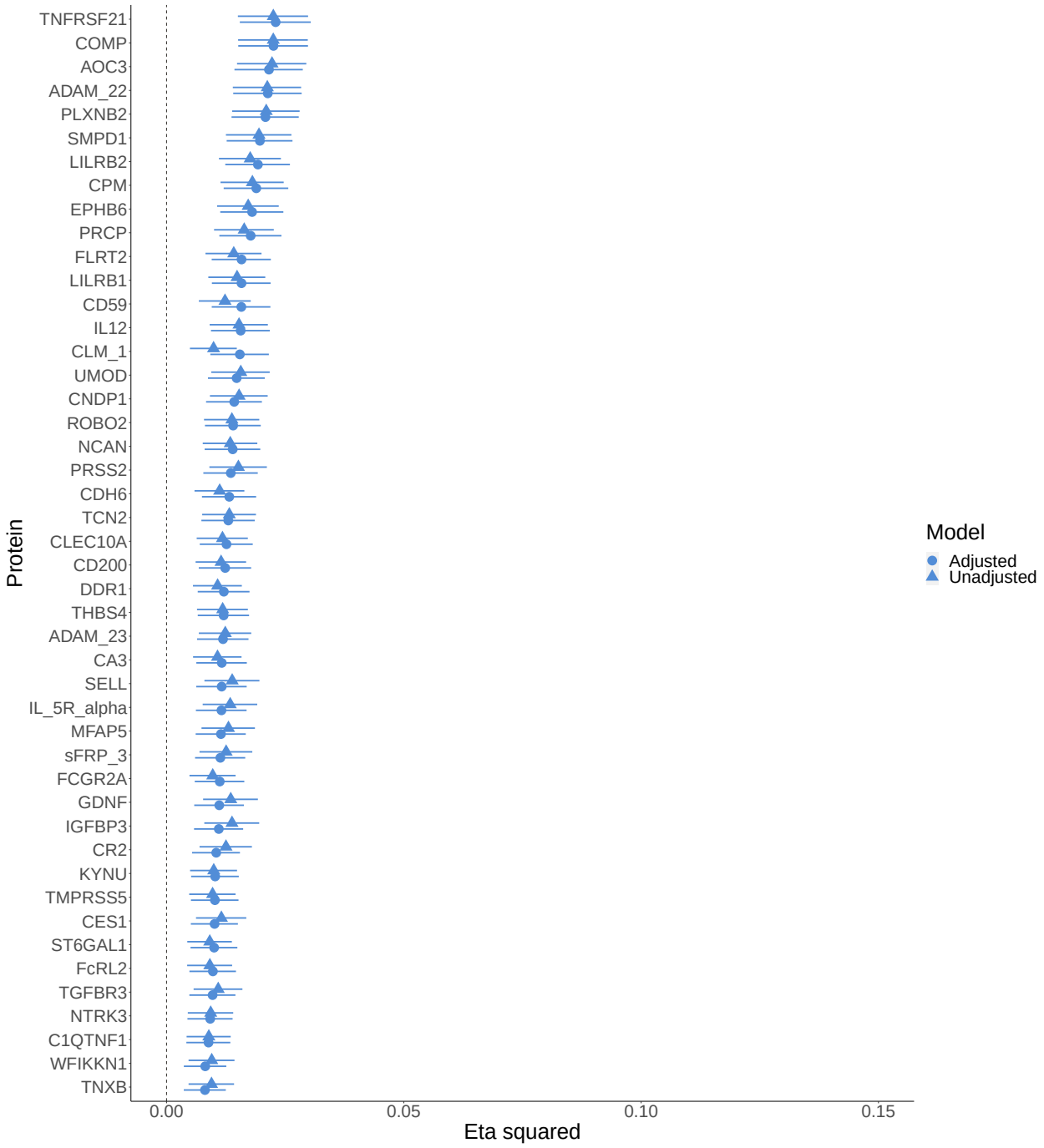

Figure 5: Associations between proteins and educational attainment

5 Discussion

We describe a high quality resource with low levels of missingness in the protein measurements. The protocol used for venous sample collection is considered sub-optimal but the association of the majority of proteins with demographic factors, including social position as measured by educational attainment, suggests that most proteins are robust to our collection protocol. We provide an overview of the key demographic and pre-analytic correlates of the protein measurements in the resource to partner the user guide provided to support use of the *Understanding Society* dataset.

Proteomic methods have been developed to enable measurements of low levels of proteins in blood samples. In the dataset, 93.3% of datapoints fell above the limit of detection for the corresponding protein suggesting that methodological issues are largely overcome. The issue of low detectability was largely confined to 18 (~10%) of the proteins, and eight of these were nevertheless found to be socially patterned, suggesting that low detectability proteins have the potential to provide new insights into the biology underlying health inequalities.

The pre-analytic factors found to be associated with protein levels and that need to be considered in analyses include time of day of sampling, transit delay and haemolysis. A high proportion of the proteins are significantly associated with a coarse "time of day" measure which may suggest diurnal variation. This could be due to two reasons: diurnal variation in protein measurements or that participant characteristics vary by time of day. There is evidence of diurnal variation in a number of analytes, including inflammatory markers [4] and lipids [14]. Dominguez-Rodriguez et al (2009) [4] review the evidence for diurnal variation in a subset of inflammatory markers, including interleukin-6 (IL-6), matrix metalloproteinases (MMPs), vascular cell adhesion protein 1 (VCAM1) and intracellular adhesion molecule 1 (ICAM1), among others. They suggest that most of these exhibit diurnal variation in either healthy people, patients with acute coronary syndrome, or both, although findings are mixed for ICAM1. The present study finds that ICAM1 and VCAM1 both vary by a coarse measure of time of day, as do two metalloproteinases ADAM 22 and ADAM 23, and the metalloproteinase inhibitor TIMP 1. Thus, our findings broadly align with those in the review. However, as the nurse visit could be scheduled at a time that suited the participant, associations may be confounded by factors which are non-randomly associated with time of day in the dataset, which may need to be accounted for in future analyses. Furthermore, we may have seen different observations if "time of day" had been operationalised using a more sensitive measure than the three-level factor used in the present study (data not shown).

In addition to time of day of sampling, transit delay and haemolysis are also potential sources of variability in protein concentrations. However, we are aware that haemolysis is not a routine metric obtained in population studies and is not measured systematically between laboratories. Haemolysis can occur during venepuncture [15], and extended processing delays could also lead to haemolysis [5]. Many proteins are linearly and/or non-linearly associated with transit delay. Proteins whose measurements increase with transit delay may be leaking from lysed blood cells over time, and those which decrease over time may have undergone degradation by proteolytic enzymes. Alternatively, ongoing biological processes may alter protein levels, or some other change could take place that leads to analytical interference in the proximity extension assay. All of the 27 proteins associated with haemolysis were also associated with transit delay and/or transit delay squared, suggesting that a proportion of the variation associated with transit delay in the dataset may be due to haemolysis. It seems that delay-related changes in central tendency of the protein measurements can be corrected by adjusting for pre-analytic covariates in any analyses. We also demonstrate that the spread of measurements changes with transit delay for some proteins, suggesting a potential reduction in precision. The exclusion of measurements from samples with transit delays above a chosen threshold may serve to reduce noise in the data. Indeed, samples with a delay greater than one week were excluded from the present study, although more stringent delays may be considered for some proteins. It is worth noting, however, that *ad hoc* data visualisations suggest that the analyses of spread may be sensitive to outliers, although this has not been systematically explored (data not shown).

We also demonstrate that transit delay is associated with a range of other factors. Due to the timed nature of post collections by the postal service, it is unsurprising that time of day is related to transit delay. The association between haemolysis and transit delay was expected [16]. Transit delay is also negatively associated with age in an expected direction, as older adults are more likely to be at home during the day and thus potentially less likely to schedule evening appointments which would lead to overnight delays in postal collection. This could also partially explain the association between transit delay and educational attainment as older people in the UK are likely to have fewer qualifications than younger people in the UK. Thus, cohort effects lead to an over-representation of retired participants in the “no qualifications” group. Transit delay is also related to geographical region, a factor which could proxy a number of different things, such as regional differences in the average postal collection time or nurse effects. A more thorough investigation would be required to explore the likely reasons for these associations.

Our findings accord with studies conducted *in vitro* in which experimentally-added hemolysate (blood cell contents) impacted on assay performance for some proteins in the cardiometabolic panel [6] and the neurology panel [7]. Most proteins were robust but the measurement of a number of proteins was sensitive to interference: CA1, CA3, CD59, SOD1, CPM, EZR, gal-8, HAGH, LXN, MANF and NMNAT1. Using a naturally-occurring indicator of haemolysis, we replicated findings for nine out of eleven, and found associations for a further 18 proteins. Our measure of haemolysis may therefore have been more sensitive than the experimental setting. However, it should be noted that the causes of haemolysis are not confined to pre-analytic problems: *in vivo* haemolysis may be linked to health and medication use [15].

A number of associations between demographic factors and protein measurements were detected, which were robust to pre-analytic factors and require further investigation.

The majority of proteins were also patterned by educational attainment, suggesting that proteins may aid the understanding of the biological pathways that underpin social differences in health [8]. While the addition of pre-analytic covariates improved model fit for 40 of the 57 proteins which were *not* associated with education in the unadjusted analyses, subsequent adjustment made little difference to the associations with educational attainment, with only ten proteins becoming significant or non-significant after adjustment. To date, studies that have examined the association of protein levels with health generally have not accounted for social factors. However our findings accord with a recent study that described the association of the Olink panel “cardiovascular” proteins - which capture biological pathways that reflect inflammation, cellular adhesion and platelet activity and intracellular signalling - with educational attainment (Shafi et al, 2002) [8] and expand these observations to Olink “cardiometabolic” and “neurology” panels. This suggests that researchers should account for measures of social position in their analyses. The findings also accord with the notion that proteomic data are likely to provide insight into the biological pathways that play a role in the widely described social inequalities in mental and physical health [17], [18], [19]. Focussing on the top ten socially patterned proteins identified in the present study, which are all from the neurology panel (EDA2R, MSR1, GFR_alpha_1, LAYN, VWC2, TNFRSF12A, SKR3, SCARB2, SCARF2 and UNC5C) we find that nine were reportedly associated with general fluid cognitive ability in an analysis exploring the relationships between plasma proteins and several cognition- and brain-related phenotypes in older cohorts (Harris et al, 2020) [20]. Cognitive ability is socially patterned [21] but Harris and colleagues did not adjust for social position, suggesting these observations require further investigation.

Our findings represent an extension of earlier analyses in a number of ways: for example, our analyses are conducted in a wider age range than in the previous publication that examined associations with educational attainment and protein levels [8]. Information in the study will enable future analyses of potential cohort effects, or non linear relationships between educational attainment and occupational exposures.

A smaller number of proteins are associated with ethnic group or are regionally patterned, which was unexpected given widely described ethnic differences in metabolic health [22] and regional differences in health in the UK [23]. These observations require further investigation.

6 Conclusion

Understanding Society's proteomics dataset is of high quality with low missingness. Transit delay is a major pre-analytic source of variability linked to our mode of collection, which initial exploratory analyses of the dataset suggest can be adequately controlled for when examining associations with age and educational attainment. Our analyses suggest that additional work is needed to understand the associations of education and other measures of social (dis)advantage on protein levels.

7 References

[1] M. Hedayati, S. A. Razavi, S. Boroomand, and S. K. Kia, "The impact of pre-analytical variations on biochemical analytes stability: A systematic review," *Journal of Clinical Laboratory Analysis*, vol. 34, p. e23551, 2020, doi: 10.1002/jcla.23551.

[2] J. Huang *et al.*, "Assessing the preanalytical variability of plasma and cerebrospinal fluid processing and its effects on inflammation-related protein biomarkers," *Molecular & Cellular Proteomics*, vol. 20, p. 100157, 2021, doi: 10.1016/j.mcpro.2021.100157.

[3] A. Cernat and J. W. Sakshaug, "Nurse effects on measurement error in household biosocial surveys," *BMC Medical Research Methodology*, vol. 20, 2020, doi: 10.1186/s12874-020-00922-2.

[4] A. Dominguez-Rodriguez, P. Abreu-Gonzalez, and J. C. Kaski, "Inflammatory systemic biomarkers in setting acute coronary syndromes - effects of the diurnal variation." *Current Drug Targets*, vol. 10, pp. 1001–1008, 2009, doi: 10.2174/138945009789577963.

[5] G. Lippi *et al.*, "Haemolysis: An overview of the leading cause of unsuitable specimens in clinical laboratories." *Clin Chem Lab Med.*, vol. 46, no. 6, pp. 764–772, 2008, doi: 10.1515/CCLM.2008.170.

[6] "Olink cardiometabolic validation data," Olink Proteomics, 2021. Available: https://www.olink.com/content/uploads/2021/09/olink-cardiometabolic-validation-data-v2.0.pdf

[7] "Olink neurology validation data," Olink Proteomics, 2021. Available: https://www.olink.com/content/uploads/2021/09/olink-neurology-validation-data-v2.1.pdf

[8] B. H. Shafi *et al.*, "Socioeconomic disparity in cardiovascular disease: Possible biological pathways based on a proteomic approach," *Atherosclerosis*, vol. 352, pp. 62–68, 2022, doi: 10.1016/j.atherosclerosis.2022.05.020.

[9] "Olink target 96 cardiometabolic panel," *Olink.com*. Jun. 2022. Available: https://olink.com/products-services/target/cardiometabolic-panel/

[10] "Olink target 96 neurology panel," *Olink.com*. Jun. 2022. Available: https://olink.com/products-services/target/neurology-panel/

[11] M. J. Easterbrook, T. Kuppens, and A. S. R. Manstead, "Socioeconomic status and the structure of the self-concept," *British Journal of Social Psychology*, vol. 59, no. 1, pp. 66–86, 2020, doi: https://doi.org/10.1111/bjso.12334.

[12] "Concordance and repeatability tests to evaluate lab performance," Olink Proteomics, 2022. Available: https://www.olink.com/content/uploads/2022/11/technical-note-concordance-and-repeatability-tests-to-evaluate-lab-performance-v1.0.pdf

[13] Y. Chan, "Biostatistics 104: Correlational analysis," *Singapore Med J*, vol. 44, no. 12, pp. 614–619, 2003.

[14] H. P. Sennels, H. L. Jørgensen, and J. Fahrenkrug, "Diurnal changes of biochemical metabolic markers in healthy young males – the bispebjerg study of diurnal variations," *Scandinavian Journal of Clinical and Laboratory Investigation*, vol. 75, no. 8, pp. 686–692, 2015, doi: 10.3109/00365513.2015.1080385.

[15] L. Thomas, "Haemolysis as influence & interference factor," *EJIFCC*, vol. 13, no. 4, pp. 95–98, 2002.

[16] L. Heireman, P. Van Geel, L. Musger, E. Heylen, W. Uyttenbroeck, and B. Mahieu, “Causes, consequences and management of sample hemolysis in the clinical laboratory,” *Clinical Biochemistry*, vol. 50, no. 18, pp. 1317–1322, 2017, doi: https://doi.org/10.1016/j.clinbiochem.2017.09.013.

[17] J. S. Yudkin, M. Kumari, S. E. Humphries, and V. Mohamed-Ali, “Inflammation, obesity, stress and coronary heart disease: Is interleukin-6 the link?” *Atherosclerosis*, vol. 148, no. 2, pp. 209–214, 2000, doi: https://doi.org/10.1016/S0021-9150(99)00463-3.

[18] C. Power *et al.*, “Life-course influences on health in British adults: effects of socio-economic position in childhood and adulthood,” *International Journal of Epidemiology*, vol. 36, no. 3, pp. 532–539, Jan. 2007, doi: 10.1093/ije/dyl310.

[19] G. Knies and M. Kumari, “Multimorbidity is associated with the income, education, employment and health domains of area-level deprivation in adult residents in the UK,” *Scientific Reports*, vol. 12, no. 1, p. 7280, May 2022, doi: 10.1038/s41598-022-11310-9.

[20] S. E. Harris *et al.*, “Neurology-related protein biomarkers are associated with cognitive ability and brain volume in older age,” *Nature Communications*, vol. 11, no. 800, 2020, doi: 10.1038/s41467-019-14161-7.

[21] A. Singh-Manoux, M. Richards, and M. Marmot, “Socioeconomic position across the lifecourse: How does it relate to cognitive function in mid-life?” *Annals of Epidemiology*, vol. 15, no. 8, pp. 572–578, 2005, doi: https://doi.org/10.1016/j.annepidem.2004.10.007.

[22] N. Gholap, M. Davies, K. Patel, N. Sattar, and K. Khunti, “Type 2 diabetes and cardiovascular disease in south asians,” *Primary Care Diabetes*, vol. 5, no. 1, pp. 45–56, 2011, doi: https://doi.org/10.1016/j.pcd.2010.08.002.

[23] I. E. Buchan, E. Kontopantelis, M. Sperrin, T. Chandola, and T. Doran, “North-south disparities in english mortality 1965-2015: Longitudinal population study,” *Journal of Epidemiology & Community Health*, vol. 71, no. 9, pp. 928–936, 2017, doi: 10.1136/jech-2017-209195.

8 Appendix 1

We list the proteins with a statistically significant change in spread (standard deviation, SD) after a given number of days' transit delay. The second and third column additionally impose "per cent change in SD" thresholds of 10% and 25%, respectively, wherein statistically significant changes are ignored if their magnitude falls below the threshold.

Table 7: Transit delay cutoffs affecting spread: protein names

Minimum transit delay affecting spread (days)	Proteins affected		
	No SD % change cutoff	10% SD change cutoff	25% SD change cutoff
2	Alpha_2_MRAP, ANGPTL3, APOM, BCAN, CA1, CA4, CADM3, CCL14, CCL18, CCL5, CD200, CD200R1, CD38, CD46, CDH1, CDH3, CDH6, CLEC10A, CLEC1B, CLM_1, CLM_6, CNDP1, CPA2, CPM, CR2, CST3, CTSC, CTSS, DDR1, DEFA1, DRAXIN, EDA2R, EFEMP1, EZR, F7, FCGR3B, FETUB, gal_8, GDF_8, GDNF, GFR_alpha_1, GNLY, GP1BA, GPC5, GZMA, HAGH, IGFBP3, IGFBP6, IL_5R_alpha, IL7R, ITGAM, JAM_B, KIT, KYNU, LAT, LAYN, LCN2, LTBP2, LXN, MANF, MATN3, MFAP5, MSR1, NAAA, NBL1, NCAN, NMNAT1, Nr_CAM, NRP2, NTRK2, NTRK3, PCOLCE, PDGF_R_alpha, PLXNB1, PLXNB3, PRSS2, PRTG, PVR, QPCT, REG1A, RGMA, RSPO1, SCARA5, SCARF2, SELL, SERPINA5, sFRP_3, Siglec_9, SIGLEC1, SKR3, SMOC2, SOD1, SPOCK1, TGFBR3, THY_1, TN_R, TNFRSF12A, TNFRSF21, TNXB, UNC5C, VWC2, WFIKKN1	CD200, CD38, CDH6, CLEC10A, CLEC1B, CLM_6, CPM, CTSS, DDR1, DEFA1, EZR, GDNF, GZMA, MANF, MSR1, NBL1, NCAN, Nr_CAM, NRP2, NTRK2, NTRK3, PDGF_R_alpha, PLXNB1, PRTG, PVR, REG1A, RGMA, SCARA5, sFRP_3, SKR3, TGFBR3, THY_1, TNFRSF21	CLEC1B, CPM, MANF, Nr_CAM, NRP2, NTRK2, THY_1, TNFRSF21

3	Beta_NGF, CA3, CD59, EFNA4, EPHB6, FAP, FLRT2, G_CSF, LAIR_2, MAPT, PLA2G7, SCARB2, TMPRSS5, TNC	Alpha_2_MRAP, Beta_NGF, CD59, G_CSF, gal_8, ITGAM, Siglec_9, TMPRSS5, TNC	G_CSF, NBL1
4	C2, FCGR2A, FcRL2, ICAM3, LILRB1, LILRB2, MET, N2DL_2, NOTCH1, PAM, PRCP, SAA4, SERPINA7	CA4, FAP, GNLY, LXN, PLXNB3, SAA4, SCARF2, SERPINA7	-
5	BMP_4, CES1, FCN2	BMP_4, CA3, CCL5, CES1, CPA2, FCGR2A, FCN2, FLRT2, LCN2, MATN3, MFAP5, N2DL_2, NMNAT1, QPCT	BMP_4, CLEC10A, FCN2, FLRT2, LXN
6	-	CD46, CLM_1, CTSC, LAT, NAAA, SOD1, SPOCK1	Beta_NGF, LCN2
7	IL12	CA1, EFNA4, GPC5, HAGH, IL12	CA1, CA3, CD59, CTSS, EZR, HAGH, IL12, N2DL_2, PLXNB1, REG1A, SOD1

None (i.e. SD is stable)	ADAM_22, ADAM_23, ANG, AOC3, C1QTNF1, CFHR5, CHL1, CNTN5, COL18A1, COMP, CRTAC1, CRTAM, Dkk_4, DPP4, ENG, F11, GAS6, GDNFR_alpha_3, GM_CSF_R_alpha, ICAM1, IGLC2, LILRB5, LYVE1, MBL2, MDGA1, MEGF9, N_CDase, NCAM1, NEP, NID1, NRP1, OSMR, PLTP, PLXNB2, PROC, PTPRS, REG3A, RGMB, ROBO2, SMPD1, SPARCL1, ST6GAL1, TCN2, TGFBI, THBS4, TIE1, TIMD4, TIMP1, UMOD, VASN, VCAM1	ADAM_22, ADAM_23, ANG, ANGPTL3, AOC3, APOM, BCAN, C1QTNF1, C2, CADM3, CCL14, CCL18, CD200R1, CDH1, CDH3, CFHR5, CHL1, CNDP1, CNTN5, COL18A1, COMP, CR2, CRTAC1, CRTAM, CST3, Dkk_4, DPP4, DRAXIN, EDA2R, EFEMP1, ENG, EPHB6, F11, F7, FCGR3B, FcRL2, FETUB, GAS6, GDF_8, GDNFR_alpha_3, GFR_alpha_1, GM_CSF_R_alpha, GP1BA, ICAM1, ICAM3, IGFBP3, IGFBP6, IGLC2, IL_5R_alpha, IL7R, JAM_B, KIT, KYNU, LAIR_2, LAYN, LILRB1, LILRB2, LILRB5, LTBP2, LYVE1, MAPT, MBL2, MDGA1, MEGF9, MET, N_CDase, NCAM1, NEP, NID1, NOTCH1, NRP1, OSMR, PAM, PCOLCE, PLA2G7, PLTP, PLXNB2, PRCP, PROC, PRSS2, PTPRS, REG3A, RGMB, ROBO2, RSPO1, SCARB2, SELL, SERPINA5, SIGLEC1, SMOC2, SMPD1, SPARCL1, ST6GAL1, TCN2, TGFBI, THBS4, TIE1, TIMD4, TIMP1, TN_R, TNFRSF12A, TNXB, UMOD, UNC5C, VASN, VCAM1, VWC2, WFIKKN1	ADAM_22, ADAM_23, Alpha_2_MRAP, ANG, ANGPTL3, AOC3, APOM, BCAN, C1QTNF1, C2, CA4, CADM3, CCL5, CCL14, CCL18, CD200, CD200R1, CD38, CD46, CDH1, CDH3, CDH6, CES1, CFHR5, CHL1, CLM_1, CLM_6, CNDP1, CNTN5, COL18A1, COMP, CPA2, CR2, CRTAC1, CRTAM, CST3, CTSC, DDR1, DEFA1, Dkk_4, DPP4, DRAXIN, EDA2R, EFEMP1, EFNA4, ENG, EPHB6, F11, F7, FAP, FCGR2A, FCGR3B, FcRL2, FETUB, gal_8, GAS6, GDF_8, GDNF, GDNFR_alpha_3, GFR_alpha_1, GM_CSF_R_alpha, GNLY, GP1BA, GPC5, GZMA, ICAM1, ICAM3, IGFBP3, IGFBP6, IGLC2, IL_5R_alpha, IL7R, ITGAM, JAM_B, KIT, KYNU, LAIR_2, LAT, LAYN, LILRB1, LILRB2, LILRB5, LTBP2, LYVE1, MAPT, MATN3, MBL2, MDGA1, MEGF9, MET, MFAP5, MSR1, N_CDase, NAAA, NCAM1, NCAN, NEP, NID1, NMNAT1, NOTCH1, NRP1, NTRK3, OSMR, PAM, PCOLCE, PDGF_R_alpha, PLA2G7, PLTP, PLXNB2, PLXNB3, PRCP, PROC, PRSS2, PRTG, PTPRS, PVR, QPCT, REG3A, RGMA, RGMB, ROBO2, RSPO1, SAA4, SCARA5, SCARB2, SCARF2, SELL, SERPINA5, SERPINA7, sFRP_3, SIGLEC1, Siglec_9, SKR3, SMOC2, SMPD1, SPARCL1, SPOCK1, ST6GAL1, TCN2, TGFBI, TGFBR3, THBS4, TIE1, TIMD4, TIMP1, TMPRSS5, TN_R, TNC, TNFRSF12A, TNXB, UMOD, UNC5C, VASN, VCAM1, VWC2, WFIKKN1

We also include some example boxplots of proteins whose spread changes with transit delay. The spread of CA3 changes significantly at 3 days' delay (fig 6a), compared to 1-2 days', but only by 2.4%, whereas it changes by 14.2% at 5 days' delay compared to 1-4 days', and by 89.5% at 7 days' delay compared to 1-6 days'. By contrast, ANG is consistently robust to transit delay (fig 6b) and BMP-4 (fig 6c) is robust until five days when the standard deviation changes by 29.4%. It is worth noting that, for some of the proteins, spread is drastically altered by extreme outliers; for example CLEC1B (fig 6d) which has two low outliers (values of 0.96 and 7.96; not plotted) which both have a transit delay of 2 days, contributing to an overall increase in spread of 48.7% at 2 days' delay – if these outliers are removed, this changes to an 18.8% reduction.

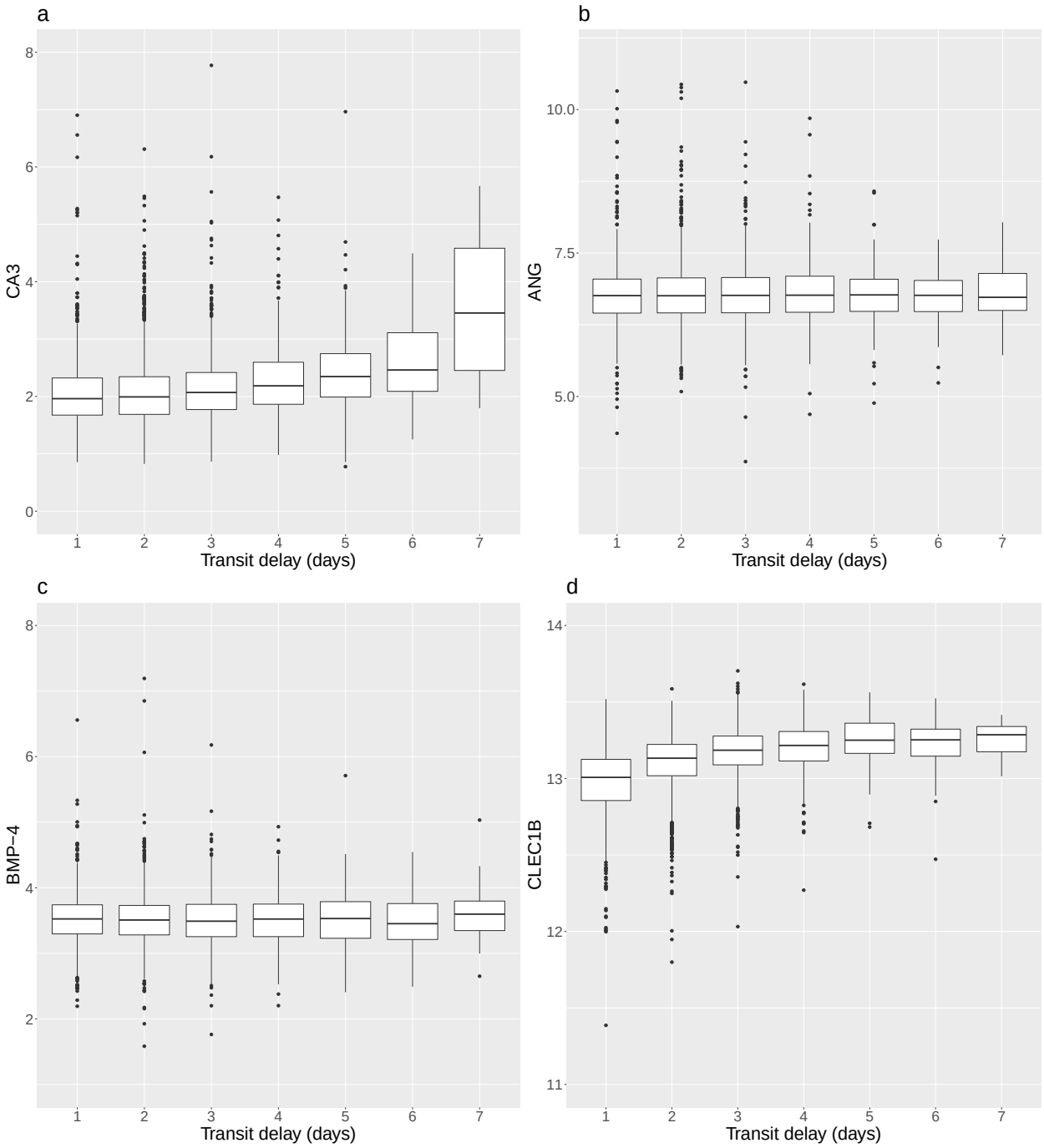

Figure 6: Example proteins whose spread is affected by transit delay

9 Appendix 2

Here we provide plots of the variance explained by educational attainment in unadjusted and adjusted models of protein levels for the 92 proteins with smaller variance explained values.

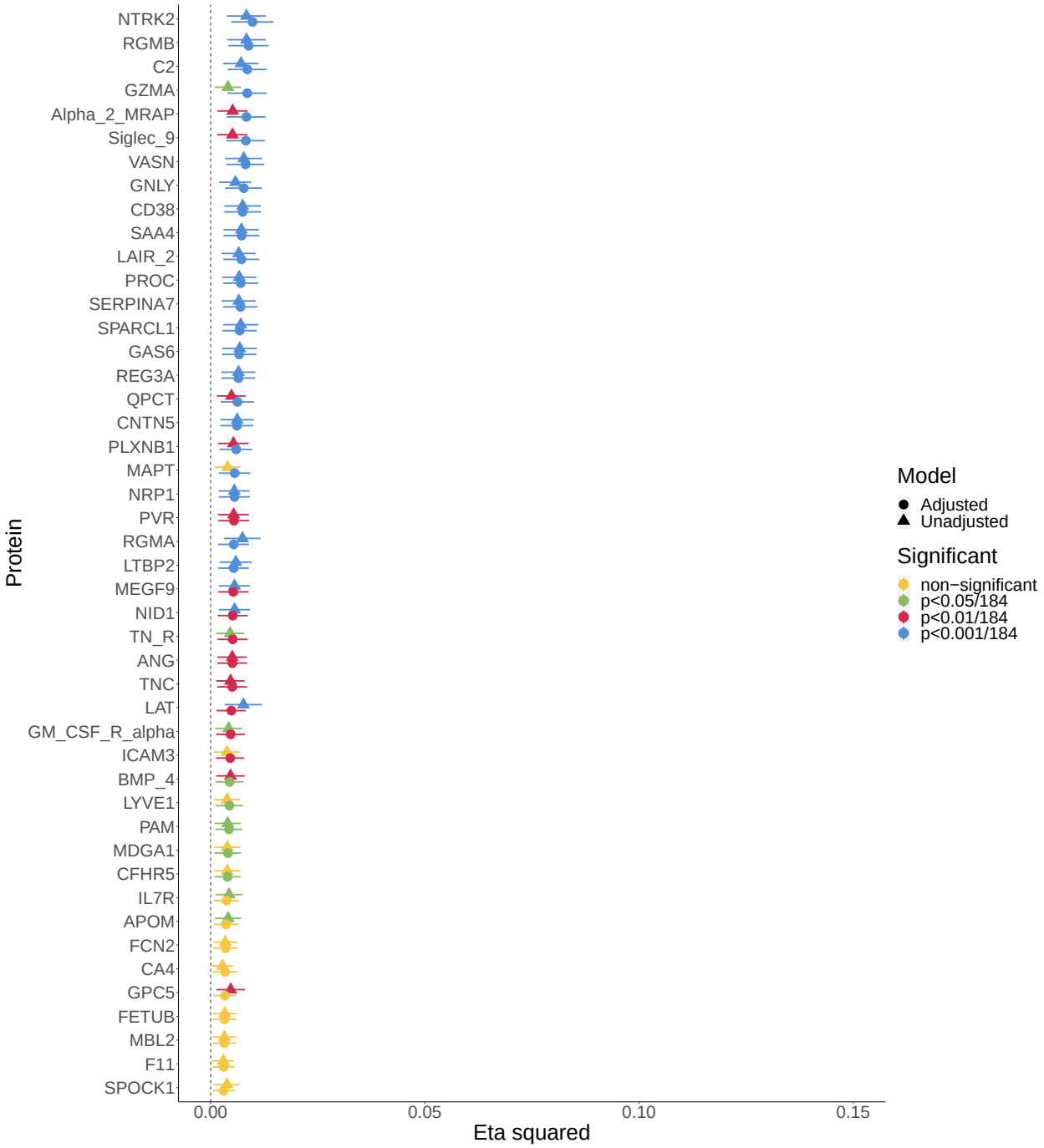

Figure 7: Associations between proteins and educational attainment

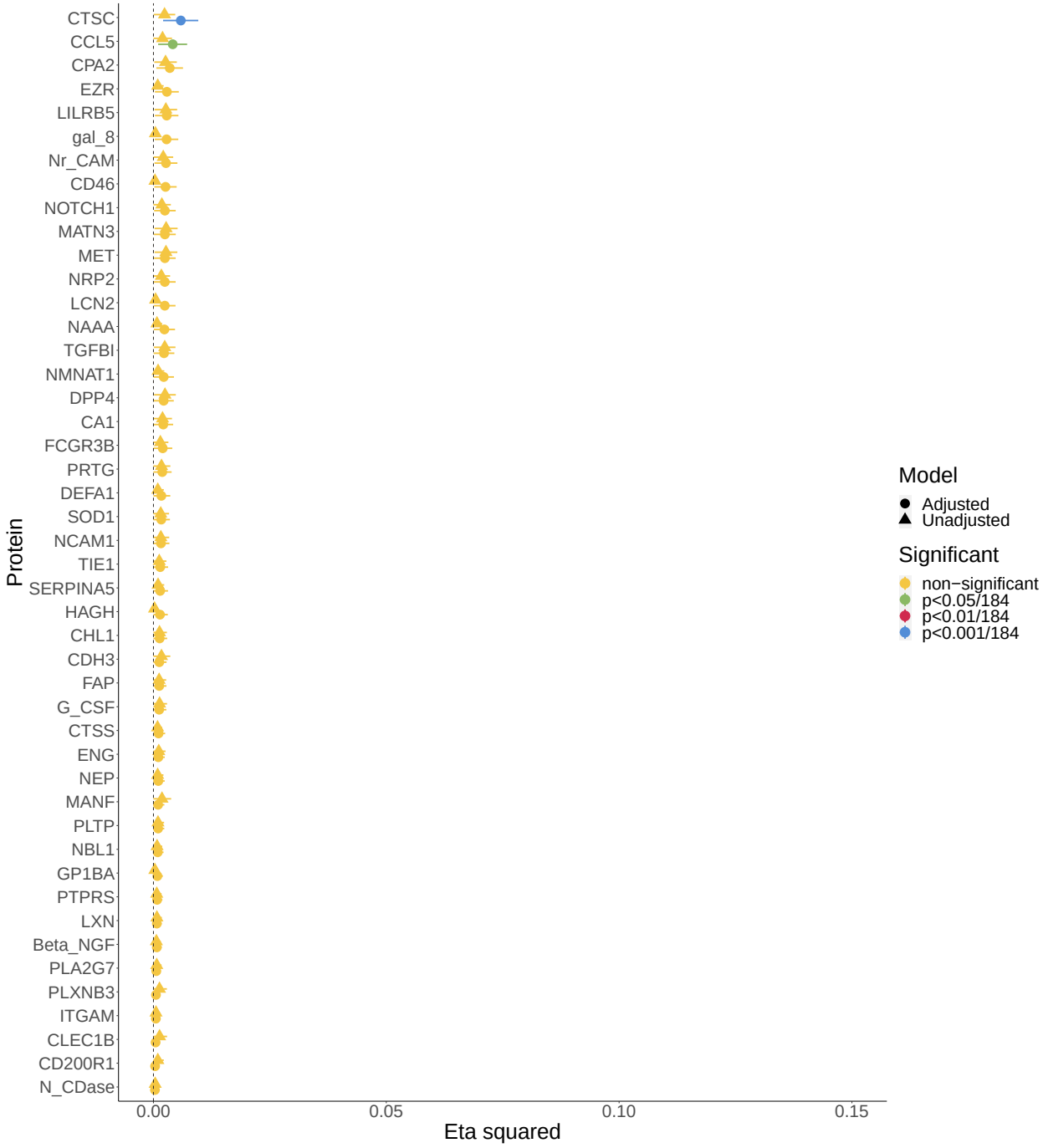

Figure 8: Associations between proteins and educational attainment