scientific data

OPEN

ANALYSIS

# Quality control and removal of technical variation of NMR metabolic biomarker data in ~120,000 UK Biobank participants

Scott C. Ritchie[1,2,3,4,5] ✉, Praveen Surendran[3,5,6], Savita Karthikeyan[3], Samuel A. Lambert[1,3,4,7], Thomas Bolton[3,8], Lisa Pennells[3,4], John Danesh[3,4,5,7,8,9], Emanuele Di Angelantonio[3,4,5,7,8], Adam S. Butterworth[3,4,5,7,8] & Michael Inouye[1,2,3,4,5,7,10,11]

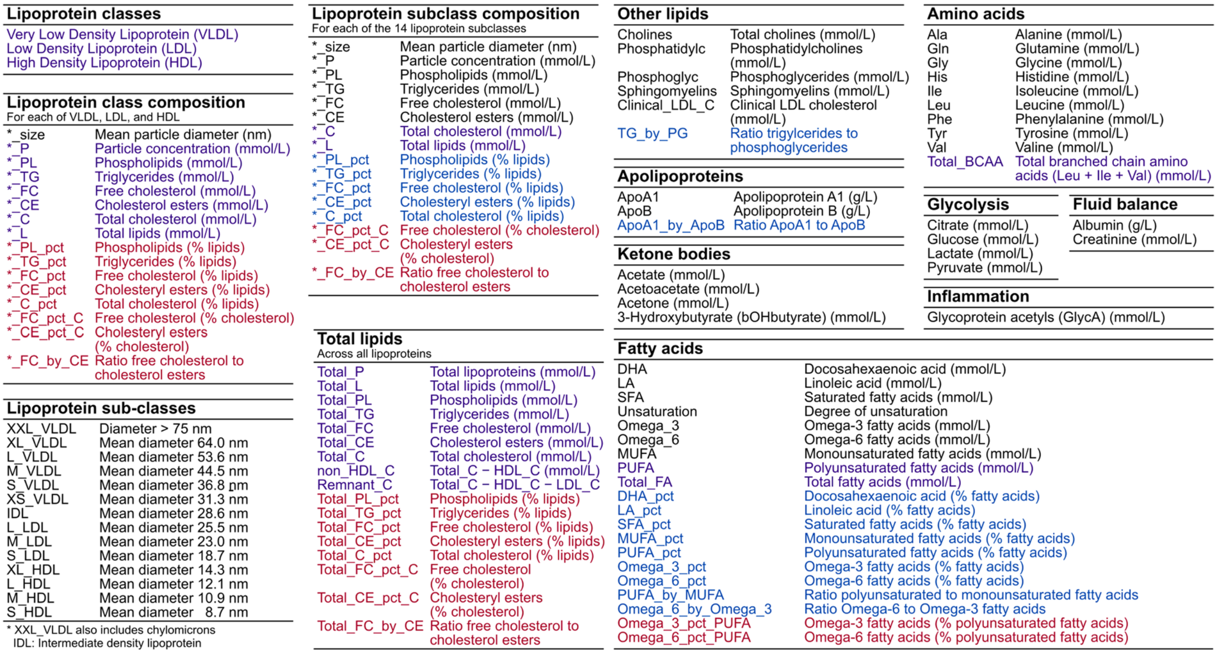
**Metabolic biomarker data quantified by nuclear magnetic resonance (NMR) spectroscopy in approximately 121,000 UK Biobank participants has recently been released as a community resource, comprising absolute concentrations and ratios of 249 circulating metabolites, lipids, and lipoprotein sub-fractions. Here we identify and characterise additional sources of unwanted technical variation influencing individual biomarkers in the data available to download from UK Biobank. These included sample preparation time, shipping plate well, spectrometer batch effects, drift over time within spectrometer, and outlier shipping plates. We developed a procedure for removing this unwanted technical variation, and demonstrate that it increases signal for genetic and epidemiological studies of the NMR metabolic biomarker data in UK Biobank. We subsequently developed an R package, ukbnmr, which we make available to the wider research community to enhance the utility of the UK Biobank NMR metabolic biomarker data and to facilitate rapid analysis.**

## Introduction

High-throughput NMR spectroscopy has enabled rapid simultaneous quantification of lipids, lipoproteins, fatty acids, and low-molecular weight metabolites including amino acids, ketone bodies, and glycolysis metabolites from a single human blood plasma sample[1,2]. Over the last decade, NMR metabolic biomarker data has been quantified in numerous cohorts each with thousands participants, helping derive new insights into the epidemiology and molecular pathogenesis of cardiovascular and metabolic diseases and examine the genetic determinants of these molecular risk factors and disease biomarkers[3].

The emergence and increasing scale of biobanks containing hundreds of thousands to millions of samples promises to enable discovery of new insights into human health and disease. To date, NMR biomarker quantification has been completed and made available by UK Biobank for 121,695 participants; showcased by Julkunen *et al.*[4]. Molecular phenotype data in large sample sizes are typically subject to the effects of unwanted technical variation (e.g. batch effects) which can obscure true biological effects and/or introduce false positive associations. Thus, a typical first step in any analysis is to identify and remove sources of unwanted variation from data. Here, we report additional quality control (QC) procedures to remove effects of technical variation present in the

[1]Cambridge Baker Systems Genomics Initiative, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK. [2]Cambridge Baker Systems Genomics Initiative, Baker Heart & Diabetes Institute, Melbourne, Victoria, Australia. [3]British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK. [4]Heart and Lung Research Institute, University of Cambridge, Cambridge, UK. [5]British Heart Foundation Centre of Research Excellence, University of Cambridge, Cambridge, UK. [6]Rutherford Fund Fellow, Department of Public Health and Primary Care, University of Cambridge, Cambridge, CB1 8RN, UK. [7]Health Data Research UK Cambridge, Wellcome Genome Campus and University of Cambridge, Cambridge, UK. [8]National Institute for Health and Care Research Blood and Transplant Research Unit in Donor Health and Genomics, University of Cambridge, Cambridge, UK. [9]Department of Human Genetics, Wellcome Sanger Institute, Hinxton, UK. [10]Department of Clinical Pathology, University of Melbourne, Parkville, Victoria, Australia. [11]The Alan Turing Institute, London, UK. ✉e-mail: sr827@medschl.cam.ac.uk



**Lipoprotein classes**

Very Low Density Lipoprotein (VLDL)
Low Density Lipoprotein (LDL)
High Density Lipoprotein (HDL)

**Lipoprotein class composition**
For each of VLDL, LDL, and HDL

| | |
|---|---|
| *_size | Mean particle diameter (nm) |
| *_P | Particle concentration (mmol/L) |
| *_PL | Phospholipids (mmol/L) |
| *_TG | Triglycerides (mmol/L) |
| *_FC | Free cholesterol (mmol/L) |
| *_CE | Cholesterol esters (mmol/L) |
| *_C | Total cholesterol (mmol/L) |
| *_L | Total lipids (mmol/L) |
| *_PL_pct | Phospholipids (% lipids) |
| *_TG_pct | Triglycerides (% lipids) |
| *_FC_pct | Free cholesterol (% lipids) |
| *_CE_pct | Cholesteryl esters (% lipids) |
| *_C_pct | Total cholesterol (% lipids) |
| *_FC_pct_C | Free cholesterol (% cholesterol) |
| *_CE_pct_C | Cholesteryl esters (% cholesterol) |
| *_FC_by_CE | Ratio free cholesterol to cholesterol esters |

**Lipoprotein sub-classes**

| | |
|---|---|
| XXL_VLDL | Diameter > 75 nm |
| XL_VLDL | Mean diameter 64.0 nm |
| L_VLDL | Mean diameter 53.6 nm |
| M_VLDL | Mean diameter 44.5 nm |
| S_VLDL | Mean diameter 36.8 nm |
| XS_VLDL | Mean diameter 31.3 nm |
| IDL | Mean diameter 28.6 nm |
| L_LDL | Mean diameter 25.5 nm |
| M_LDL | Mean diameter 23.0 nm |
| S_LDL | Mean diameter 18.7 nm |
| XL_HDL | Mean diameter 14.3 nm |
| L_HDL | Mean diameter 12.1 nm |
| M_HDL | Mean diameter 10.9 nm |
| S_HDL | Mean diameter 8.7 nm |

* XXL_VLDL also includes chylomicrons
IDL: Intermediate density lipoprotein

**Lipoprotein subclass composition**
For each of the 14 lipoprotein subclasses

| | |
|---|---|
| *_size | Mean particle diameter (nm) |
| *_P | Particle concentration (mmol/L) |
| *_PL | Phospholipids (mmol/L) |
| *_TG | Triglycerides (mmol/L) |
| *_FC | Free cholesterol (mmol/L) |
| *_CE | Cholesterol esters (mmol/L) |
| *_C | Total cholesterol (mmol/L) |
| *_L | Total lipids (mmol/L) |
| *_PL_pct | Phospholipids (% lipids) |
| *_TG_pct | Triglycerides (% lipids) |
| *_FC_pct | Free cholesterol (% lipids) |
| *_CE_pct | Cholesteryl esters (% lipids) |
| *_C_pct | Total cholesterol (% lipids) |
| *_FC_pct_C | Free cholesterol (% cholesterol) |
| *_CE_pct_C | Cholesteryl esters (% cholesterol) |
| *_FC_by_CE | Ratio free cholesterol to cholesterol esters |

**Total lipids**
Across all lipoproteins

| | |
|---|---|
| Total_P | Total lipoproteins (mmol/L) |
| Total_L | Total lipids (mmol/L) |
| Total_PL | Phospholipids (mmol/L) |
| Total_TG | Triglycerides (mmol/L) |
| Total_FC | Free cholesterol (mmol/L) |
| Total_CE | Cholesterol esters (mmol/L) |
| Total_C | Total cholesterol (mmol/L) |
| non_HDL_C | Total_C − HDL_C (mmol/L) |
| Remnant_C | Total_C − HDL_C − LDL_C |
| Total_PL_pct | Phospholipids (% lipids) |
| Total_TG_pct | Triglycerides (% lipids) |
| Total_FC_pct | Free cholesterol (% lipids) |
| Total_CE_pct | Cholesteryl esters (% lipids) |
| Total_C_pct | Total cholesterol (% lipids) |
| Total_FC_pct_C | Free cholesterol (% cholesterol) |
| Total_CE_pct_C | Cholesteryl esters (% cholesterol) |
| Total_FC_by_CE | Ratio free cholesterol to cholesterol esters |

**Other lipids**

| | |
|---|---|
| Cholines | Total cholines (mmol/L) |
| Phosphatidylc | Phosphatidylcholines (mmol/L) |
| Phosphoglyc | Phosphoglycerides (mmol/L) |
| Sphingomyelins | Sphingomyelins (mmol/L) |
| Clinical_LDL_C | Clinical LDL cholesterol (mmol/L) |
| TG_by_PG | Ratio trigylcerides to phosphoglycerides |

**Apolipoproteins**

| | |
|---|---|
| ApoA1 | Apolipoprotein A1 (g/L) |
| ApoB | Apolipoprotein B (g/L) |
| ApoA1_by_ApoB | Ratio ApoA1 to ApoB |

**Ketone bodies**

Acetate (mmol/L)
Acetoacetate (mmol/L)
Acetone (mmol/L)
3-Hydroxybutyrate (bOHbutyrate) (mmol/L)

**Amino acids**

| | |
|---|---|
| Ala | Alanine (mmol/L) |
| Gln | Glutamine (mmol/L) |
| Gly | Glycine (mmol/L) |
| His | Histidine (mmol/L) |
| Ile | Isoleucine (mmol/L) |
| Leu | Leucine (mmol/L) |
| Phe | Phenylalanine (mmol/L) |
| Tyr | Tyrosine (mmol/L) |
| Val | Valine (mmol/L) |
| Total_BCAA | Total branched chain amino acids (Leu + Ile + Val) (mmol/L) |

**Glycolysis**

Citrate (mmol/L)
Glucose (mmol/L)
Lactate (mmol/L)
Pyruvate (mmol/L)

**Fluid balance**

Albumin (g/L)
Creatinine (mmol/L)

**Inflammation**

Glycoprotein acetyls (GlycA) (mmol/L)

**Fatty acids**

| | |
|---|---|
| DHA | Docosahexaenoic acid (mmol/L) |
| LA | Linoleic acid (mmol/L) |
| SFA | Saturated fatty acids (mmol/L) |
| Unsaturation | Degree of unsaturation |
| Omega_3 | Omega-3 fatty acids (mmol/L) |
| Omega_6 | Omega-6 fatty acids (mmol/L) |
| MUFA | Monounsaturated fatty acids (mmol/L) |
| PUFA | Polyunsaturated fatty acids (mmol/L) |
| Total_FA | Total fatty acids (mmol/L) |
| DHA_pct | Docosahexaenoic acid (% fatty acids) |
| LA_pct | Linoleic acid (% fatty acids) |
| SFA_pct | Saturated fatty acids (% fatty acids) |
| MUFA_pct | Monounsaturated fatty acids (% fatty acids) |
| PUFA_pct | Polyunsaturated fatty acids (% fatty acids) |
| Omega_3_pct | Omega-3 fatty acids (% fatty acids) |
| Omega_6_pct | Omega-6 fatty acids (% fatty acids) |
| PUFA_by_MUFA | Ratio polyunsaturated to monounsaturated fatty acids |
| Omega_6_by_Omega_3 | Ratio Omega-6 to Omega-3 fatty acids |
| Omega_3_pct_PUFA | Omega-3 fatty acids (% polyunsaturated fatty acids) |
| Omega_6_pct_PUFA | Omega-6 fatty acids (% polyunsaturated fatty acids) |

**Fig. 1** Summary of NMR metabolic biomarkers in UK Biobank. Short variable names, descriptions, and units for the 249 biomarkers and ratios in the original data that is presently available for download through UK Biobank, as well as the 76 additional biomarker ratios derived in this study (shown in red). Biomarkers are grouped by type, given in each table heading. The 107 biomarkers in black are those which cannot be derived from any combination of other biomarkers. The 61 biomarkers in purple are the composite biomarkers that can be derived by summing two or more of the 107 non-derived biomarkers. Shown in blue are the 81 biomarker ratios available in the original data, which can be derived from the 168 non-derived or composite biomarkers. Further details for each biomarker and ratio are provided in Table S1 and formulae for deriving the composite biomarkers and ratios are provided in Table S2.

biomarker concentrations available to download from UK Biobank. We make this procedure available a resource to the community via an R package, ukbnmr.

## Results

### Summary of NMR metabolic biomarker data available in UK Biobank.

UK Biobank is a deeply phenotyped cohort of approximately 500,000 thousand adults[5,6]. Presently, NMR metabolic biomarkers have been quantified (Nightingale Health Plc.) and made available through UK Biobank for approximately one third of randomly selected participants (**Methods**)[4]. After sample QC (**Methods**), NMR metabolic biomarker data were available for 121,657 UK Biobank participants: 117,994 at baseline assessment (2006–2010) and 5,139 at first repeat assessment (2012–2013). The NMR metabolic biomarker data comprises absolute concentrations of 168 biomarkers along with 81 biomarker ratios[1]. These primarily consist of lipids and lipoprotein sub-fractions (81%) but also include fatty acids, amino acids, ketone bodies, glycolysis related metabolites, and others. An overview of available biomarkers and ratios are provided in Fig. 1, and details are provided in Table S1. Henceforth, we refer to this as the "original" biomarker data, reflecting (i) that the data are calibrated absolute concentrations (or ratios) and not raw NMR spectra, and (ii) that prior to release the data had already been subject to quality control procedures[1] by Nightingale Health Plc. (**Methods**).

### Impact of technical variation.

Data on a range of technical factors with potential to introduce unwanted variance in NMR metabolic biomarker measurements were made available by Nightingale Health Plc. to analysts with early-access to the original NMR biomarker data. These included shipping batch, 96-well plate, well position, aliquoting robot, and aliquot tip (UK Biobank), along with spectrometer and date and time stamps for each step in the biomarker quantification pipeline (Fig. 2A, M**ethods**).

Figure 2B shows the variance in the original biomarker concentrations explained by each technical covariate, and that this technical variance is removed by the QC procedures described below (referred to henceforth as the "post-QC" biomarker data). Overall, most biomarkers were relatively robust to technical variation (Fig. 2B,C), with median of 1.5% variance explained by the most correlated technical factor with each biomarker (Fig. 2C). Within 3,169 blind duplicate samples the median coefficient of variation (CV%) was 4.55% (interquartile range [IQR]: 3.21%–6.03%) and the median coefficient of determination ($R^2$) was 0.928 (IQR: 0.886–0.947) (Table S3, **Methods**). After removal of technical variation herein, the CV% was reduced (median: 4.03%, IQR: 2.86%–5.38%) and $R^2$ increased (median: 0.937, IQR: 0.913–0.953) (Fig. 2D, Table S3).

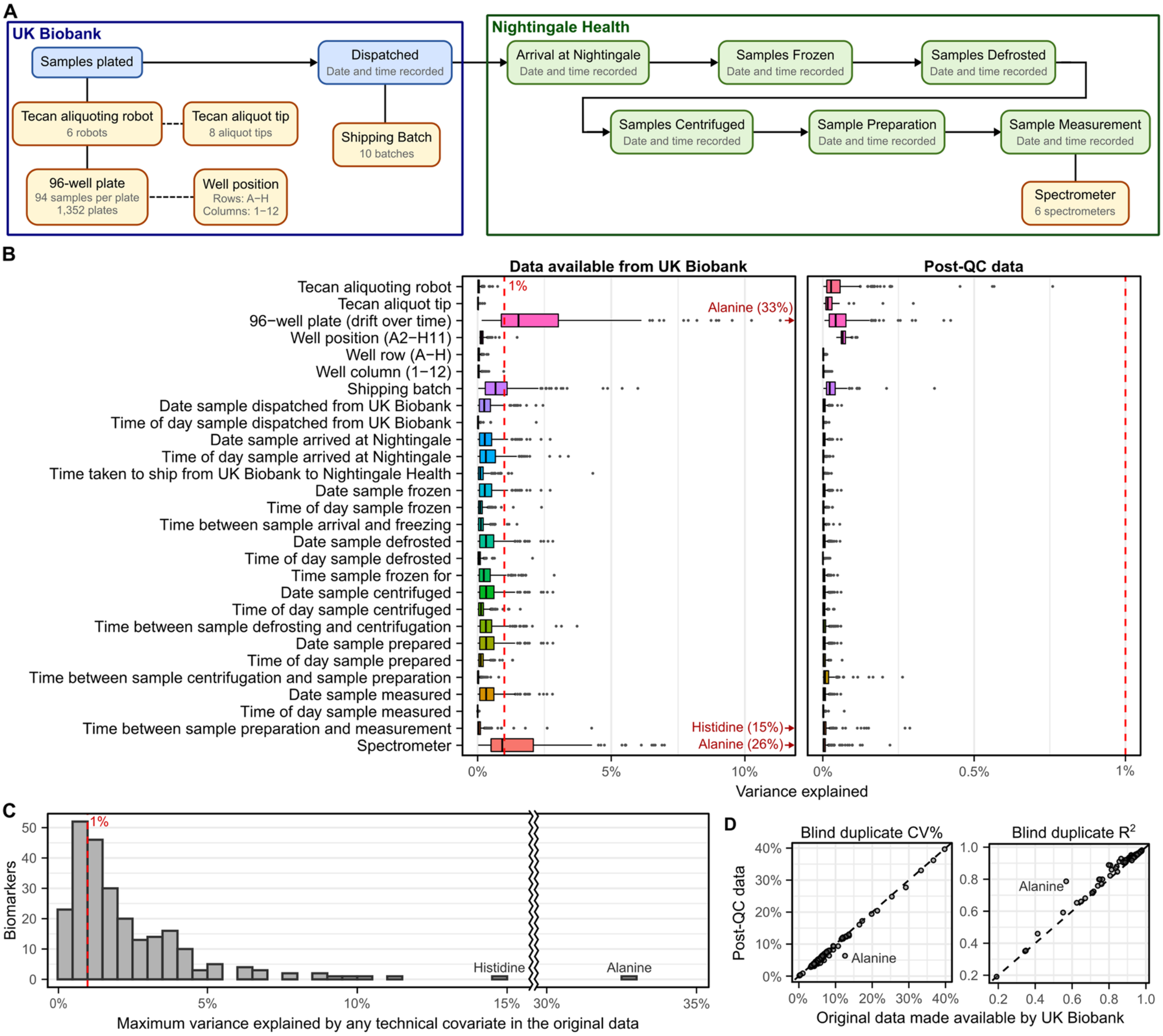


**Fig. 2** Technical covariates and their impact on biomarker concentrations. (**A**) Schematic showing the technical covariates from the UK Biobank and Nightingale Health data processing pipelines (**Methods**). UK Biobank aliquoting robot and aliquot tip were randomised with respect to each other, with respect to 96-well plate, and with respect to position on 96 well plate. Plates and shipping batch were randomised with respect to spectrometer. (**B**) Boxplots showing variance explained (**Methods**) across the 249 NMR metabolic biomarkers by each possible technical covariate in both the original dataset made available for download by UK Biobank and the dataset after applying the additional quality control procedures described here. (**C**) Histogram showing the maximum variance explained by any technical factor for each biomarker in the original data. (**D**) Comparison of coefficient of variation (CV%) and R$^2$ computed in 3,169 blind duplicate samples for each of the 107 non-derived biomarkers before and after additional quality control. CV% and R$^2$ for individual biomarkers are detailed in Table S3.

There were 22 (of 249) biomarkers with at least 5% of variance explained by one or more technical factors, with strongest impacts observed for inter-spectrometer differences in biomarker concentrations, along with drift over time across different plates measured within each spectrometer (Fig. 2B). Notably, 33% of the variance in alanine concentrations could be explained by drift over time within spectrometer (Fig. 2B, Fig. 3A) followed by 15% of the variance in histidine concentrations which could be explained by time between sample preparation and sample measurement (Figs. 2B,3B). Further, inter-spectrometer differences in biomarker concentrations and drift over time within spectrometer were visually apparent for nearly all biomarkers (see extended diagnostic plots on FigShare[7]). Intra-plate variation was also visually apparent for some biomarkers (see extended diagnostic plots on FigShare[7]), for example, we observed a consistent decrease in glycine concentrations from left to right across each plate (Fig. 3C).

Patterns of intra- and inter- plate variation were similar across biomarkers where present (see extended diagnostic plots on FigShare[7]). The impact of sample degradation time was consistent with ongoing cellular metabolism, with some biomarker concentrations increasing concomitant with decreases in other biomarkers in

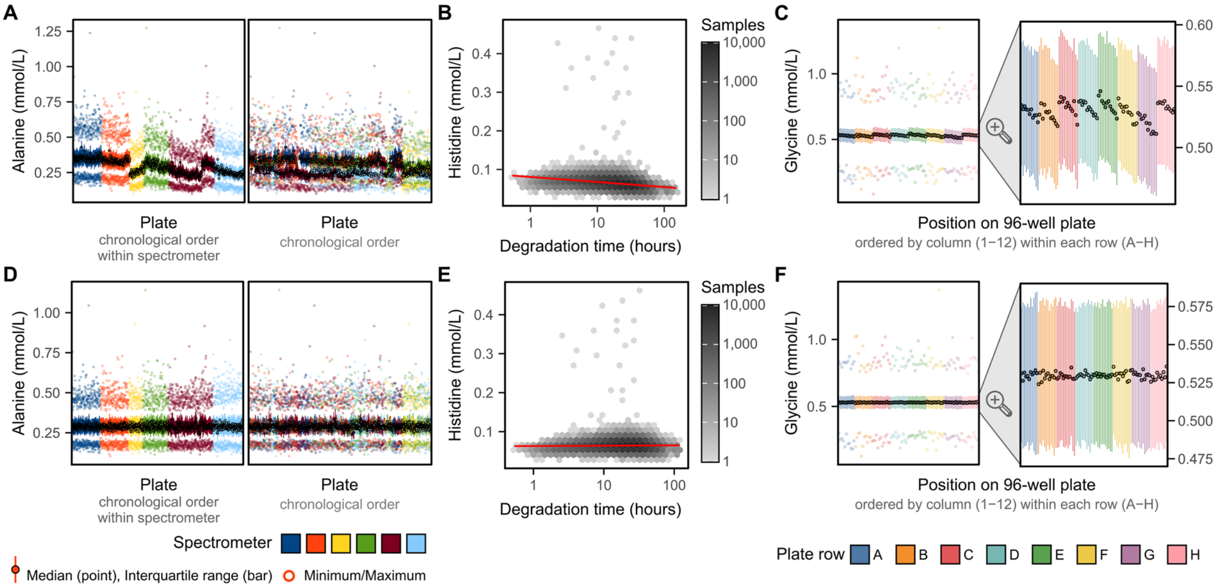


**Fig. 3** Strongest effects of technical covariates on biomarker concentrations. (**A**) Inter-plate variation in alanine concentrations attributable to drift over time within spectrometer. Each plot shows a summary of alanine concentrations on each plate (minimum, maximum, median, and interquartile range). On the left plot, plates are ordered by date of plate measurement and grouped by spectrometer. On the right plot, plates are ordered by date of measurement without grouping by spectrometer. (**B**) Histidine concentrations decrease as a function of sample degradation time (time between sample preparation and sample measurement) on a $\log_{10}$ scale. Hexagonal bins show sample counts on a $\log_{10}$ scale and red line shows the association (fit on all data points using robust linear regression). (**C**) Glycine concentrations change as a function of plate position, with concentrations decreasing as plate column (1–12) increase, and also systematically lower in plate row G. (**D**–**F**) show the same as panels (**A**–**C**) after removal of technical variation.

the same pathways. For example, sample degradation time was association with ongoing branched chain amino acid metabolism[8]; with increases in sample degradation time associating with increased alanine concentrations alongside decreased isoleucine, leucine, and valine concentrations (see extended diagnostic plots on FigShare[7]).

**Removal of technical variation.** Technical variation described above was removed using a multistep pipeline (**Methods**) detailed below. Figure 4 gives a step-by-step example, for the NMR amino acid glycine, showing how concentrations and their relationship with key technical covariates changes at each step.

A key motivating consideration at each step was to ensure adjustment for categorical technical factors did not break the samples into very small groups with potential to be non-random with respect to biological factors. For example, a simple approach to remove inter-plate variation would be to median normalise each plate. However, with only 94 samples measured per plate, this approach would likely remove not just technical variation but also biological variation of interest to downstream analysts. For example, allocation of participants to plates were not randomised with respect to participant sex (Figure S1), thus per-plate normalization would remove some sex-specific differences in metabolite concentrations, which may be of interest to downstream researchers.

With this consideration in mind, technical variation was removed through four sequential regressions (**Methods**). First, (1) the original biomarker concentrations were regressed on time between sample preparation and sample measurement on a log scale to remove any potential effects of sample degradation (Fig. 4C). Sample degradation time was fit on a log scale as its potential effects fit a pattern of exponential rather than linear decay (Fig. 3B,E). Next, (2) subsequent residuals were regressed on plate row (8 groups; rows A–H; Fig. 4D) as visual differences between plate rows remained after removing effects of sample degradation time (Fig. 4C). Third, (3) subsequent residuals were regressed on plate column (12 groups; columns 1–12; Fig. 4E) as differences between plate columns were apparent for some biomarkers (Figure S2). Finally, (4) inter-plate variation due to drift over time within spectrometer was removed (Fig. 4F) by binning plates into 10 groups by date within each of the 6 spectrometers (Figure S3) and regressing on bin as a categorical variable. At each of these steps robust linear regression (**Methods**) was used to fit robust linear regressions as we found linear regression susceptible to outliers and non-normality (Figure S4).

Prior to fitting these regressions, the original biomarker concentrations were log transformed (Fig. 4B) so that a unit decrease in biomarker levels were equivalent to a unit increase (e.g. a halving and doubling both become a 0.69 unit change on the natural log scale). A small offset was applied for biomarkers with concentrations of 0 (Table S4). Subsequent to regressing out effects of technical variation, absolute biomarker concentrations were obtained (Fig. 4G) from the regression residuals by rescaling their post-QC distributions to the distributions of the original biomarker concentrations (**Methods**, Figure S5).

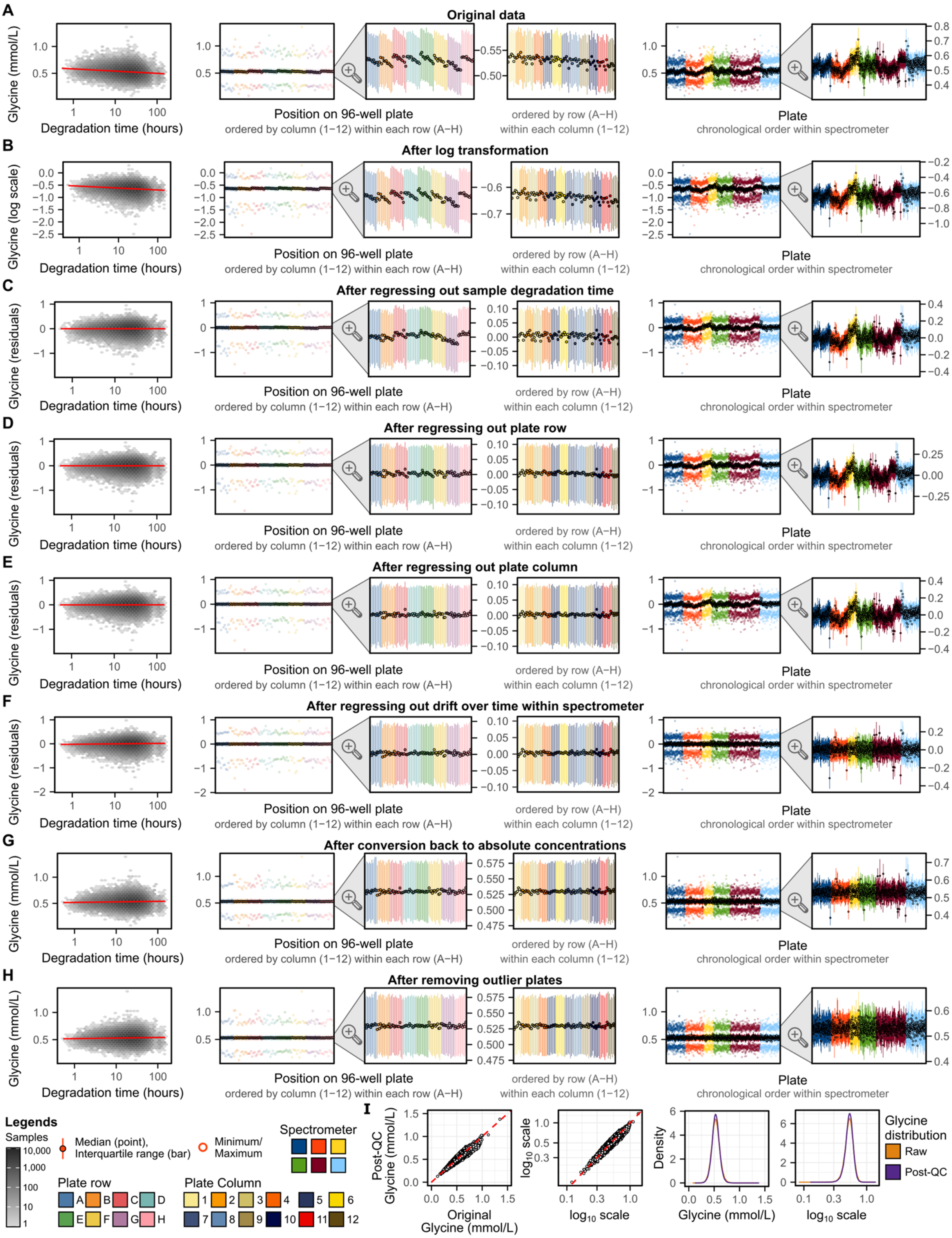


**Fig. 4** Step-by-step removal of technical variation for glycine concentrations. (**A**) Glycine concentrations in the original data made available to download through UK Biobank and their relationship with technical covariates. (**B–H**) Show how glycine levels and their relationship with technical covariates change after each step in the removal of technical variation pipeline (Fig. 1B, **Methods**). In each of **A–H**, each plot shows from left to right: (1) relationship between glycine and sample degradation time (hours between sample preparation and sample measurement) on a $\log_{10}$ scale. Hexagonal bins show sample counts on a $\log_{10}$ scale and red line shows the association (fit on all data points using robust linear regression). (2) Summary of glycine levels on each well across all plates (minimum, maximum, median, and interquartile range). Wells are grouped and coloured by plate row (**A–H**) and within each row ordered by plate column (1–12). (3) A zoomed in view showing just the

interquartile range and median for each well. (4) Alternative grouping of the zoomed in which wells are grouped and coloured by plate column (1–12) and within each column ordered by row (**A**–**H**). (5) Summary of glycine levels on each plate. Plates are grouped and coloured by spectrometer and within each spectrometer ordered by measurement date. (6) A zoomed in view showing just the interquartile range and median for each plate. (**I**) Compares glycine concentrations before and after removal of technical variation. From left to right: (1) glycine concentrations in the original data (x-axis) vs. glycine concentrations after removal of technical variation (y-axis) in mmol/L units, and (2) also shown on $\log_{10}$ scale (axis tick labels given in mmol/L units). (3) Distribution of Glycine concentrations in mmol/L units, and (4) also shown on $\log_{10}$ scale (axis tick labels given in mmol/L units).

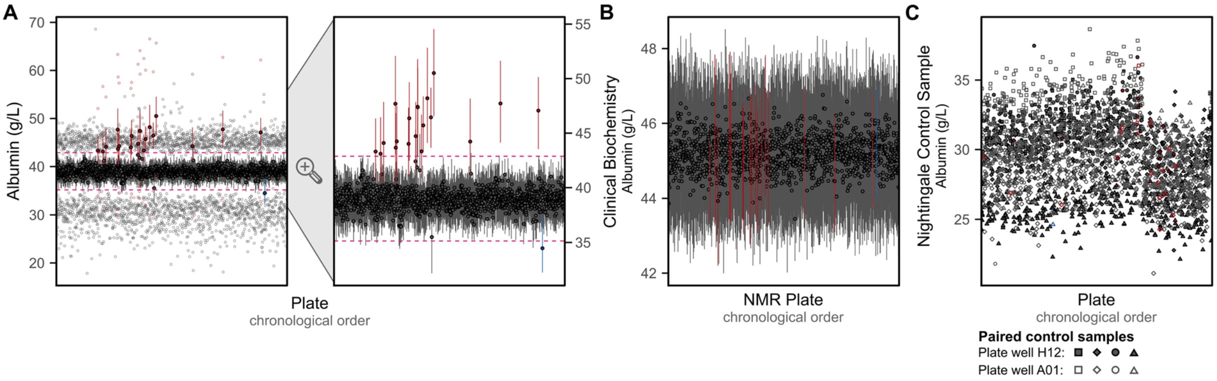


**Fig. 5** Outlier plates driven by unexplained technical variation. (**A**) Summary of albumin concentrations on each UK Biobank shipping plate after removal of technical variation. For each plate, the median, interquartile range, minimum, and maximum values are shown on the left plot, with the second plot on the right showing just the median and interquartile range for each plate. Plates identified as outliers (**Methods**) are shown in red or blue. The horizontal pink dashed line show the limits above or below which plates were tagged as outliers based on their median values. (**B**) Median and interquartile range of albumin concentrations previously measured by clinical biochemistry[9] when grouped by UK Biobank shipping plate used to send samples for NMR metabolite biomarker quantification. Plates that were identified as outliers in panel **A** are shown in red and blue. Of the 121,758 UK Biobank participants with NMR metabolite biomarkers, 107,283 also had albumin concentrations quantified by clinical biochemistry. (**C**) Albumin concentrations quantified from control samples placed on plate wells A01 and H12 by Nightingale Health Plc. for internal quality checks in their NMR metabolite biomarker quantification pipeline. Four sets of paired control samples were used across all 1,352 plates. Note inter-plate and inter-spectrometer variation has not been removed from internal control samples here. Plates that were identified as outliers in panel **A** are shown in red and blue.

After removing most visible inter-plate variation (Fig. 4F,G), we observed for many biomarkers several strong outlier plates (see extended diagnostic plots on FigShare[7]). The strongest example, for albumin concentrations, is shown in Fig. 5A. Stratification of albumin concentrations previously quantified by clinical biochemistry[9] according to the UK Biobank shipping plates supported a non-biological origin for the observed outlier plates (Fig. 5B). Investigation of control samples placed on each plate showed no deviation of control samples for outlier plates (Fig. 5C), indicating the source of this technical variation was not due to the NMR quantification pipeline but rather arose during the UK Biobank sample plating process. However, these outlier plates could not be adjusted for using any of the available technical covariates detailed in Fig. 2A. We therefore systematically identified and removed these outlier plates (**Methods**) as the final step in the pipeline to remove unwanted technical variation (Fig. 4H). Across biomarkers, these accounted for a median of 9 plates (0.66% of plates/samples), with maximum of 20 plates (1.5% plates/samples) for albumin and phosphoglycerides.

**Composite biomarkers and ratios should be re-derived after covariate adjustment.** Among the 249 NMR metabolite biomarkers available to download from UK Biobank, 81 were derived ratios and 61 were composite biomarkers—derivable as sums of two or more biomarkers (Figure S6)—while 107 were non-derivable from other biomarkers (Fig. 1, Table S1, Table S2).

Notably, we found that directly adjusting composite biomarkers or biomarker ratios for technical covariates sometimes led to different post-QC concentrations or ratios than obtained by computing the biomarker from its adjusted composite parts (Fig. 6A). In particular, technical covariates could have different effect sizes on biomarkers contributing to a composite biomarker (Fig. 6B), which could combine in a complex fashion on the composite biomarker (Fig. 6C). These differences are proportional to the effects of the adjusted covariates on each biomarker: we highlight that these differences are much larger when adjusting for age, sex, and body mass index (BMI) (Fig. 6D), which have much larger impacts on biomarker concentrations than technical covariates (Fig. 6E).

When removing the effects of technical variation above, we therefore recomputed the 61 composite biomarkers and 81 biomarker ratios after removing the effects of technical covariates from the 107 non-derivable

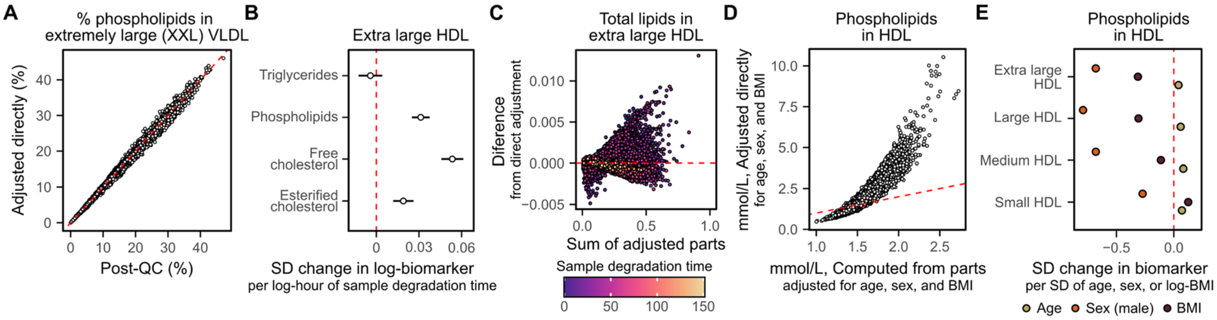


**Fig. 6** Composite biomarkers and ratios should be re-derived after removal of unwanted variation. (**A**) Adjusting the % phospholipids in XXL VLDL for technical covariates (y-axes) leads to different values than computing % phospholipids in XXL VLDL from its component biomarkers (phospholipids in XXL VLDL over total lipids in XXL VLDL, where total lipids are the sum of free cholesterol, esterified cholesterol, phospholipids, and triglycerides) after adjustment for technical covariates. (**B**,**C**) Technical covariates can have different effects on components of composite biomarkers that combine in complex ways. (**B**) Shows for each of the components of total lipids in XL HDL the association from robust linear regression between log sample degradation time and log biomarker concentrations in standard deviation units (horizontal bars show the 95% confidence intervals). (**C**) Shows the difference between total lipids in XL HDL computed from its adjusted parts to total lipids in XL HDL directly adjusted for sample degradation time (y-axis) as a function of its concentration (x-axis) with points coloured by sample degradation time (hours). (**D**,**E**) Shows how large differences can arise when adjusting for biological covariates with large effects on biomarker concentrations. (**D**) Example showing non-linear differences between phospholipids in HDL computed from its adjusted parts (x-axis) to direct adjustment of phospholipids in HDL (y-axis) of the post-QC concentrations for age, sex, and BMI. (**E**) Shows the association between age, sex, and BMI with concentrations of phospholipids in each HDL subclass from robust multivariable linear regression. Each point corresponds to the standard deviation change in the corresponding lipid per standard deviation increase in age, standard deviation increase in log transformed BMI, and when comparing male participants to female participants. 95% confidence intervals are smaller than the size of the shown points.

biomarkers. We make available code for re-deriving these biomarkers and ratios in the ukbnmr R package (**Code Availability**).

**Alternate approaches to removing technical variation.** Prior to settling on the multi-step procedure described above we tried several alternate approaches for removing the visually apparent technical variation described above (**Supplementary Note**). In particular, we sought to utilize data from eight paired internal control samples (two per 96-well plate) that are not part of the UK Biobank data release but were made available to early access researchers by Nightingale Health Plc. Two approaches were tried: (1) standardising each plate based on the differences between control samples, and (2) learning the differences between plates from these control samples using the Removal of Unwanted Variation (RUV) K-means approach designed for metabolomics data[10]. However, we found that while both methods made large changes to biomarker concentrations neither method reduced or removed visually apparent patterns of technical variation (e.g. inter-plate variation).

We also investigated whether changing the order of regressions in the multi-step procedure described above, or combining them into a single regression, impacted the removal of technical variation. We found that the multi-step procedure was largely robust to the order of regressions, with the exception of the regressions on plate row and plate column, which needed to be performed in that order to remove the visually apparent structure across well rows and columns. We further found that our multi-step procedure was equivalent (Pearson and Spearman correlation >0.9999 for all biomarkers) to a two-step procedure which (1) simultaneously corrects for sample degradation time, plate row, and plate column across all samples then (2) using those residuals to subsequently correct for spectrometer drift over time for each spectrometer separately. Combining all regressions into a single step, adjusting within each spectrometer separately for sample degradation time, plate row, plate column, and drift over time, also lead to extremely similar but not identical post-qc concentrations (Pearson and Spearman correlation >0.99 for all biomarkers).

**Derivation of additional biomarker ratios.** We further derived 76 additional biomarker ratios not available in the original data (Fig. 1, Table S1, Table S2), and make available code for deriving these additional biomarker ratios in the ukbnmr R package (**Code Availability**).

First, we derived 20 lipid fractions that are available for the 14 lipoprotein sub-classes but not for the lipoprotein classes and total serum. For each of the 14 lipoprotein sub-classes, the NMR metabolite biomarker data includes percentages of total lipids comprised of: (1) phospholipids, (2) triglycerides, (3) free cholesterol, (4) cholesteryl esters, and (5) total cholesterol (Fig. 1). Here, we additionally derive these percentages for the three lipoprotein classes: very low density lipoprotein (VLDL), low density lipoprotein (LDL), and high density lipoprotein (HDL), as well as for total lipids in serum (Fig. 1).

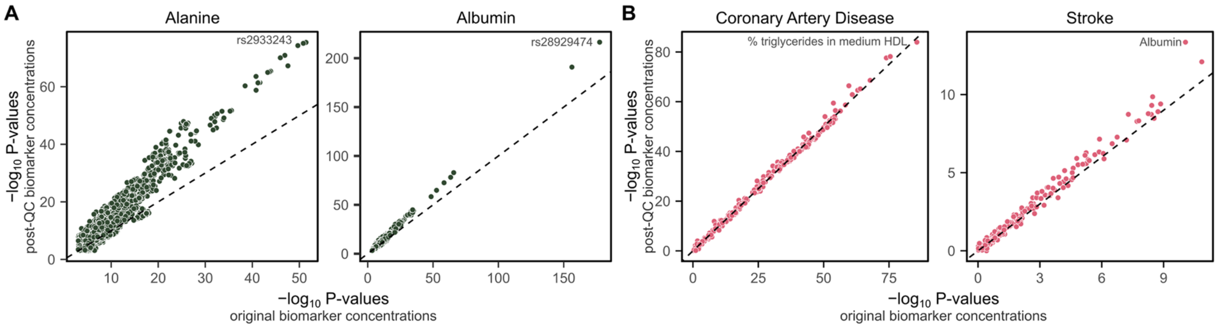


Fig. 7 Adjustment for technical covariates improves power for genetic and epidemiological studies. (**A**) Comparison of P-values from GWAS of 8.5 million common SNPs for alanine and albumin concentrations before and after removal of technical variation (**Methods**). The x-axes show $-\log_{10}$ P-values for associations between variants and concentrations prior to the additional removal of technical variation in this study, and y-axes show $-\log_{10}$ P-values for associations between variants and post-QC concentrations; after removal of technical variation. The dashed line on the diagonal shows $y = x$, where P-values are identical for both the original and post-QC concentrations. The most strongly associated variant is annotated on each plot. Manhattan plots are shown in Figure S7. The number of genome-wide significant loci ($P < 5 \times 10^{-8}$) increased from 30 to 47 and 32 to 43 for alanine and albumin respectively when removing technical variation. (**B**) Comparison of P-values from Cox proportional hazard models testing associations between incident coronary artery disease and incident stroke over 12.8 years of follow-up with original and post-QC concentrations of the 249 NMR metabolite biomarkers (**Methods**). The x-axes show $-\log_{10}$ P-values for associations between the original biomarker concentrations and incident disease, and y-axes show $-\log_{10}$ P-values for associations between post-QC biomarker concentrations and incident disease. The dashed line on the diagonal shows $y = x$, where P-values are identical for both the original and post-QC concentrations. The most strongly associated biomarker for each disease is annotated. Cox proportional hazard models were fit adjusting for age and sex. Participants with prevalent events or taking lipid lowering medication were excluded. Hazard ratios for all biomarkers are detailed in Table S6.

Next, we note that total cholesterol is composed of the sum of free cholesterol and esterified cholesterol (Table S2), thus derive cholesterol fractions (percentages of cholesterol made up of free cholesterol and esterified cholesterol) for total serum, the three lipoprotein classes, and 14 lipoprotein sub-classes (Fig. 1). We also derived the ratio of free cholesterol to cholesteryl esters as there is some evidence that this ratio may be a determinant of lipoprotein atherogenicity[11].

Finally, we note that polyunsaturated fatty acids are composed of the sum of omega-3 and omega-6 fatty acids (Table S2), thus derive the percentage of polyunsaturated fatty acids comprised of omega-3 and omega-6 fatty acids (Fig. 1).

### Quality control improves power for genetic and epidemiological studies.

Next, we examined whether removal of technical variation impacted or improved power for genetic and epidemiological studies (**Methods**).

We performed genome-wide association analyses (GWAS) (**Methods**) for original and post-QC concentrations of alanine, the biomarker most strongly affected by technical variation (Fig. 2B), and albumin, the biomarker most strongly impacted by outlier plates (Fig. 5A). We observed an increase in power for genetic associations for both biomarkers when using their post-QC concentrations with technical variation removed (Fig. 7A, Figure S7, Table S5).

A comparatively modest difference in power was observed for biomarker association scans in Cox proportional hazards models for incident coronary artery disease and incident stroke (**Methods**, Fig. 7B, Table S6). However, large differences in power are not expected in a biomarker association unless the biomarkers most strongly associated with the disease are also the ones most strongly impacted by technical variation. Notably, a large increase in power was observed for the association between albumin concentrations and incident stroke, which became the strongest association among the 249 biomarkers after removal of technical variation (Fig. 7B).

We also observed strong associations between the 76 additional biomarker ratios derived above with incident disease (Table S6), supporting their utility for future studies. For example, the percentage of LDL lipids composed of cholesterol was the biomarker with the strongest association with stroke (hazard ratio: 0.89, 95% confidence interval: 0.86–0.91, P-value: $2 \times 10^{-15}$) above and beyond that of albumin (HR: 0.83, 95% CI: 0.79–0.87, P-value: $5 \times 10^{-13}$).

### Characteristics of quality controlled NMR metabolite biomarker data.

Finally, we explore and describe basic characteristics of the post-QC data of fundamental interest to downstream analysts in Fig. 8. Specifically percentages of missing data across biomarkers (Fig. 8A) and samples (Fig. 8B), correlation between biomarkers (Fig. 8C), and sources of inter-sample variation due to common physiological and environmental confounders (Fig. 8D–F). In particular, we highlight that approximately 30% of the variation between biomarkers can be explained by a combination of sex, body mass index (BMI), and lipid lowering medication usage (Fig. 8E),

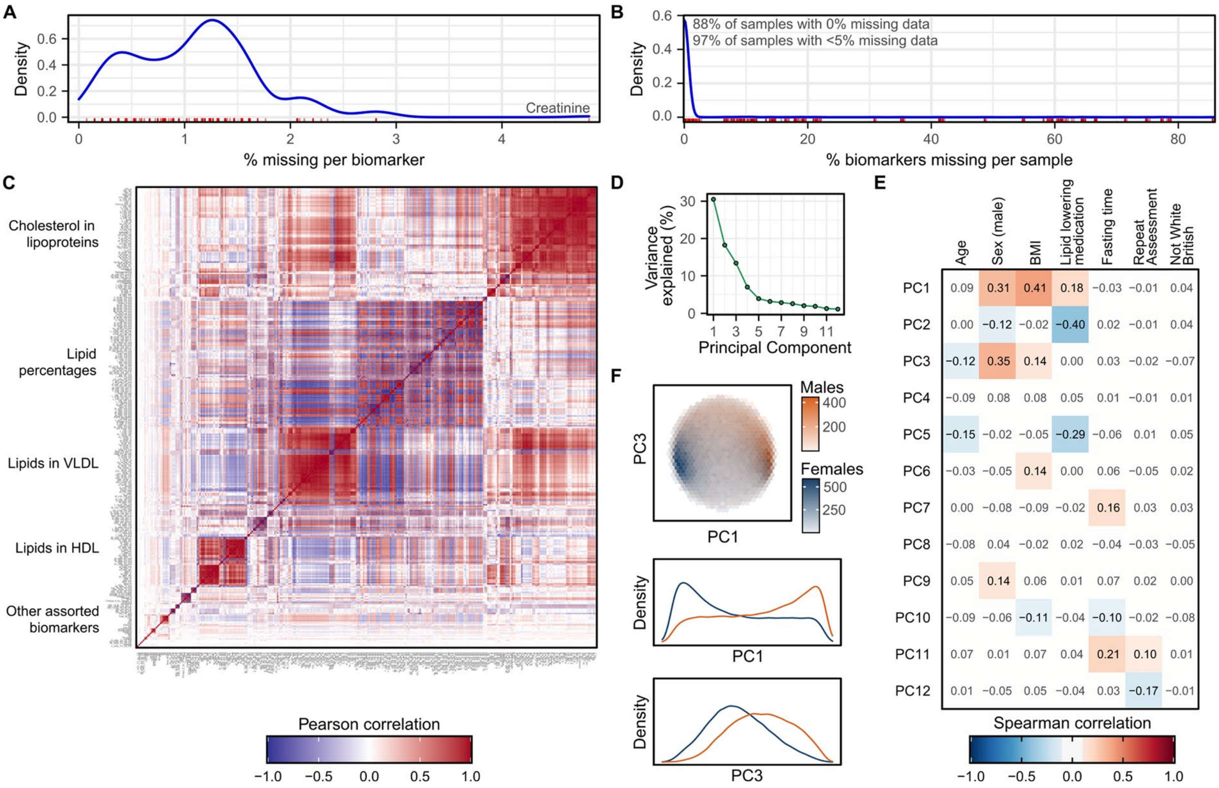


**Fig. 8** Characteristics of the NMR metabolite biomarker data after removal of technical variation. (**A**) Density plot showing distribution of missingness across the 325 NMR metabolite biomarkers (including the 76 additionally derived biomarker ratios). The rug plot in red below the distribution shows the % of samples with missing data for each biomarker. (**B**) Density plot showing distribution of missingness across the 123,023 samples. The rug plot in red below the distribution shows the % of missing biomarkers for each sample. (**C**) Pairwise Pearson correlation coefficients between the 325 biomarkers. Biomarkers are ordered by hierarchical clustering based on their topological overlap dissimiliarity[31] (**Methods**). Groups of correlated biomarkers are annotated. (**D**) Principal components (PCs) explaining more than 1% of the variation in NMR metabolite levels between samples (**Methods**). (**E**) Spearman correlation coefficients between the PCs and a selection of biological and environmental covariates. Heatmap cells are white and correlation coefficients are dulled where absolute value < 0.1. (**F**) Separation of males and females by PC1 and PC3, the two PCs most strongly correlated with sex. The first plot shows hexagonal bins of samples, coloured by sex, with the two plots below showing density plots for PC1 and PC3 stratified by sex.

and that clustering of male and female participants is visible on PC1 and PC3 (Fig. 8F, Figure S8). We also highlight that the structure of the data and relationships between biomarkers are largely unchanged from the original data (Figure S9), with the exception of biomarker and sample missingness rates which are higher in the post-QC data (Fig. 8A,B) due to the removal of outlier plates arising from unexplained technical variation (Fig. 5).

## Discussion

Technical variation is inherent to all laboratory measurements used for molecular phenotyping at biobank scale. Here, we comprehensively explored and document the sources and impacts of technical variation on the NMR metabolite biomarker data that has been made available for ~120,000 UK Biobank participants[4]. We found that the vast majority of biomarkers are relatively robust to technical variation, but that a small subset are significantly impacted by inter-spectrometer variation, within-spectrometer drift over time, and intra-plate variation. We also found that a small number of plates are unexplained outliers of non-biological origin across a swathe of biomarkers. We subsequently developed a multi-step procedure for removing this unwanted variation, which we make available as part of the ukbnmr R package as a resource to the community.

One limitation of this pipeline is that it relies on regressing out technical covariates from metadata labels, rather than calibration to quality control samples. It is therefore possible that there are other sources of technical variation that are not captured by the technical covariates examined, and that regressing on these covariates may remove biological variation of interest. Assessing whether a quality control procedure also removed biological variation is challenging and potentially intractable for a dataset as deeply phenotyped and as large as UK Biobank. However, several lines of evidence suggest our procedure is robust in this respect. First, we find that our procedure improved the reproducibility of biomarker measurements amongst blind duplicate samples; with the coefficient of variation (CV%) decreasing and coefficient of determination ($R^2$) increasing; although with the caveat that only a small fraction of samples (2.6%) could be assessed in this way. Second, we show improvement in signal for downstream genetic and phenotypic association analysis in our key examples. Third, we find that the overall correlation structure and association with key biological covariates remained unchanged after applying our procedure. However, analysts utilizing the ukbnmr package may wish to compare their key results on the original data to check for and investigate any discrepancies.

We additionally highlight several important aspects of the NMR metabolite biomarker data analysts should be aware of.

First, there are major sources of biological variation that have systematic effects on biomarker concentrations between individuals. These include major differences in most biomarker concentrations depending on participant sex, body mass index, and use of lipid lowering medication, which are expected due to the predominantly lipid content of the biomarker data.

We also highlight the strong correlation structure between biomarkers, particularly among lipid concentrations, which are typical property for this type of data[12,13]. This will directly impact choice of multiple testing strategies[14], and can present challenges for multivariable modelling[15], but can also be leveraged to increase power across multiple correlated traits[16]. While analysts may reasonably be tempted to discard such highly correlated biomarkers, we highlight Sliz *et al.*[17] as a practical example of where highly correlated lipid fractions encoded different information. We also highlight Ala-Korpela *et al.*[2] as a dissection and discussion of the relative biological importance of various lipid ratios and percentages in lipoprotein sub-fractions, and Soininen *et al.*[3] as a review of findings from epidemiological and genetic studies for the NMR metabolic biomarkers more broadly. We would also highlight that filtering to non-derived biomarkers does not mitigate challenges introduced by collinearity, as many of the non-derived biomarkers are among those with extremely strong inter-correlations.

Analysts should also be aware that no filtering or treatment of outlier samples is performed by the ukbnmr R package when removing technical variation. Many biomarkers have extreme outliers whose treatment will need to be carefully considered for statistical modelling and many biomarkers have extremely non-normal distributions. One approach to handle these can be to winsorize biomarker distributions (either before or after log transformation) or to inverse rank normalize them.

Further, analysts should also be aware that Nightingale Health Plc. has flagged some non-missing samples and biomarker measurements with QC flags (see UK Biobank fields #23651–23655 at https://biobank.ndph.ox.ac.uk/showcase/label.cgi?id=222 for sample QC flags and fields #23700–23948 at https://biobank.ndph.ox.ac.uk/showcase/label.cgi?id=221 for biomarker measurement QC flags). Sensitivity analysis to inclusion/exclusion of affected samples and measurements is recommended (see UK Biobank accompanying resources #3000 and #3004 at https://biobank.ndph.ox.ac.uk/showcase/label.cgi?id=220).

For some analysts, the post-QC dataset generated by our code in the ukbnmr R package may simply be a starting point for creating their own quality controlled datasets with further adjustment for unwanted biological variation, for example adjusting for age and sex for downstream GWAS. For these, we highlight that biomarker adjustment can introduce new unwanted variation into biomarkers which are ratios, percentages, or otherwise composed of multiple other biomarkers, e.g. with directly composite biomarkers no longer adding to the sum of their adjusted parts. If creating new post-QC datasets, we recommend filtering to the 107 non-derived biomarkers and re-deriving the rest post-adjustment, and we make publicly available methods for doing so in the R package ukbnmr. This caveat equally applies for anyone wishing to create new ratios or biomarker combinations. For example, here we additionally derive 76 lipid, cholesterol, and fatty acid percentages not present in the original Nightingale biomarker data.

In conclusion, we envision this manuscript and methods made available for removing technical variation in the ukbnmr R package will serve as a useful starting point for those who wish to use the NMR metabolite biomarker data available for ~120,000 UK Biobank participants in their future studies.

## Methods

### NMR metabolite biomarker profiling of UK Biobank.

UK Biobank is a cohort of approximately 500,000 individuals with deep phenotyping, imputed genotypes, and electronic health record linkage with written informed consent for health related research[5]. Ethics approval was obtained from the North West Multi-Center Research Ethics Committee. The current analysis was approved under UK Biobank Project 30418. UK Biobank participants were members of the general UK population between 40 to 69 years of age identified and recruited through primary care lists and who accepted an invitation to attend one of 22 assessment centres across the UK between 2006 and 2010[5]. A subset of approximately 20,000 were selected for repeat assessment between 2012 and 2013[5].

Absolute concentrations of 168 biomarkers and 81 biomarker ratios were quantified by NMR spectroscopy (Nightingale Health Plc.) from non-fasting plasma samples (UK Biobank aliquot 3) of 121,695 randomly selected UK Biobank participants as previously described[4]. These included 117,981 participants at baseline assessment and 5,141 participants at repeat assessment, among which there were 1,427 participants with measurements at both time points.

In summary, samples were randomly allocated to 96-well plates by UK Biobank and aliquoted in each well using one of six TECAN freedom EVO 150 robotic liquid handlers and one of eight 8 tips by UK Biobank. These included 3,169 blind duplicate samples: identical samples sent two (N = 3,167) or three (N = 2) times by UK Biobank to Nightingale Health Plc. with randomized sample identifiers. Nightingale Health Plc. were blinded to the identity and number of duplicate samples until after returning all data to UK Biobank. Aliquoting robot and aliquot tip were randomised with respect to each other, 96-well plate, and well position. Each plate contained a maximum of 94 samples from UK Biobank, with the remaining two wells (positions A01 and H12) reserved for internal control samples used by Nightingale Health Plc. to assess internal consistency of their biomarker quantification pipeline. In total there were eight internal control samples: four pairs, with one pair measured per plate. Plates were randomly allocated to 10 batches and shipped to Nightingale Health (Helsinki, Finland). Plates were measured on one of six spectrometers (randomised with respect to UK Biobank shipping batch, aliquoting robot, and aliquot tip) by Nightingale Health. Further details on UK Biobank sample handling can be found in UK Biobank Resource 3000 (https://biobank.ndph.ox.ac.uk/showcase/showcase/docs/nmrm_companion_doc.pdf). Further details on the Nightingale Health NMR spectroscopy quantification pipeline are detailed in Würtz *et al.*[1] and details pertaining to UK Biobank detailed in Julkunen *et al.*[4], and further companion documents can be found on the resources tab on the UK Biobank showcase page for the NMR biomarker data (https://biobank.ndph.ox.ac.uk/showcase/label.cgi?id=220).

### Sample quality control.

Pre-release data made available to early access analysts, which was used to control for technical variation, differed in sample content from the biomarker data that is now available to download from UK Biobank. In total 126,360 samples passed quality control in the pre-release data (**Supplementary Methods**). In total, 121,758 participants passed quality control: 118,047 with measurements at baseline assessment and 5,139 with measurements at first repeat assessment, including 1,428 participants with measurements at both timepoints.

For downstream analyses after removal of technical variation we filtered the pre-release data to the samples available for download from UK Biobank. Notably, the original data available for download from UK Biobank included 37 samples that did not pass sample quality control in the pre-release data due to having insufficient sample material (**Supplementary Methods**). After removing these 37 samples, the original and post-QC data used for downstream analyses contained 121,657 participants: 117,944 with measurements at baseline assessment and 5,139 with measurements at first repeat assessment, including 1,426 participants with measurements at both timepoints, with one sample per participant and timepoint. For the 3,169 blind duplicate samples, we kept for downstream analyses only the measurement that was included in the UK Biobank data release.

### Variance explained by technical covariates.

To estimate variance explained by each recorded technical covariate (Fig. 2B), linear regression were fit for each biomarker and covariate separately and the model $r^2$ obtained. Variance explained was estimated in both the original and post QC biomarker data. For the post QC data, variance explained was estimated for both the 249 Nightingale biomarkers as well as the 76 additional biomarkers derived in the post-QC dataset.

For categorical variables (well position A02–H11, well row A–H, well column 1–12, spectrometer, shipping batch, aliquoting robot, and aliquot tip) the group with the largest sample size was used as the reference group. For categorical variables with missing data (aliquot tip, N = 609 samples) the missing data were treated as a separate group, and not used as the reference group. For time of day events and date events (dispatch from UK Biobank, arrival at Nightingale Health, sample freezing, sample defrosting, sample centrifugation, sample preparation, and sample measurement) samples were split into 10 bins of equal duration and treated as a categorical variable (largest group as reference) to account for potentially non-linear effects over time. Durations between each pair of events were computed in hours (decimal) and treated as linear effects. To estimate variance explained by plate, plates were split into bins by spectrometer as described below in the removal of technical variation procedure, in this instance primarily due to the computational impracticality of fitting regression models on plate as a categorical variable (1,352 groups; regressions had not completed fitting for all biomarkers after 18 hours parallelized across 10 cores on high performance computing cluster).

Reproducibility of biomarker measurements before and after removal of technical variation was assessed using data from the 3,169 blind duplicate samples. For each biomarker, outlier samples more than four times the interquartile range from the median were excluded. Within-sample coefficient of variation (CV%) was computed using the root mean square approach[18], with a small offset applied to 0 values as described below. The coefficient of determination ($R^2$) was calculated by fitting a linear regression between duplicate measurements across samples. For each pair (or triplet) of samples, the sample included in the UK Biobank data release was used as the dependent variable when fitting the linear regression.

### Removal of technical variation from known technical covariates.

The original biomarker data was filtered to 107 "non-derived" biomarkers: biomarkers that could not be expressed in terms of two or more other quantified biomarkers (Table S1, Table S2). Concentrations of these biomarkers were log transformed, with a small offset applied to biomarkers with 0 values (half the minimum non-zero value; Table S4).

Four sequential regressions were fit to remove the effects of known technical covariates (Fig. 4). Log concentrations of each biomarker were (1) regressed on log-transformed time between sample preparation and sample measurement using robust linear regression available through the MASS package in R[19,20]. Residuals were obtained, then (2) regressed on plate row as a categorical variable (8 groups; A–H) with robust linear regression using the row with the largest number of samples as the reference group (row D, N = 16,155 samples). Residuals were obtained, then (3) regressed on plate column as a categorical variable (12 groups; 1–12) with robust linear regression using the column with the largest number of samples as the reference group (column 6, N = 10,775 samples). Residuals we obtained, then (4) split into six groups by spectrometer. Plates within each spectrometer were ordered by measurement date, then split into 10 groups of approximately equal size (keeping plates measured on the same date in the same bin; Figure S3). Where measurement of samples on a plate occurred over multiple consecutive days, the plate measurement date was taken as the date on which the most samples were measured for that plate (N = 487 plates with samples measured all on the same day, N = 842 plates with samples measured over two consecutive days, N = 22 plates with samples measured over three consecutive days, N = 1 plate with samples measured over 4 consecutive days). Within each of the six spectrometer groups separately, robust linear regression was subsequently fit on plate measurement date bin as a categorical variable, using the bin with the largest number of samples as the reference group (N = 1,306–3,825 samples per reference group; Figure S3).

### Obtaining absolute concentrations for residuals.

After removing the effects of known technical covariates from the 107 non-derived biomarkers, their residuals were converted back to absolute concentrations through the procedure described below (Figure S5).

The residuals from any regression are defined as the difference between the observed independent variable (e.g. biomarker concentrations) and the parameter estimated by the regression. A resulting key property is that the distribution of the residuals is centred on 0 for this estimated parameter. In the case of robust linear regression, this parameter is an estimate of the mean that is robust to outliers[19,20]. As a consequence of the way residuals are defined, their distribution can be scaled to match the distribution of the independent variable by giving it the same estimated mean. For robust linear regression, residuals can be returned to the same scale as the observed independent variable by estimating the mean of the observed independent variable using robust linear regression and adding it to the residuals.

When adjusting for technical covariates above, the independent variable was the log transformed original concentrations of each biomarker. For each biomarker, we therefore fit a robust linear regression on the log transformed original biomarker concentrations with just an intercept term to obtain an estimate of the mean robust to outliers. We then added this estimated mean to the residuals to return the residuals to the same scale as the log transformed original concentrations. To return biomarkers to absolute concentrations, instead of log-concentrations, the exponent function was applied to inverse the log transformation. For biomarkers with concentrations of 0 in the original data, the small offset applied prior to log transformation (Table S4) was also removed. Subsequently, some samples with concentrations of 0 in the original data had negative concentrations very close to zero in the post-QC data; a small offset was applied to these biomarkers to ensure no negative concentrations (Table S4). In all cases this offset was at least one order of magnitude smaller than the smallest non-zero concentration, i.e. its impact on concentrations amounts to noise in numeric precision.

### Removal of outlier plates arising due to unexplained technical variation.

Outlier plates were subsequently identified and removed for each of the 107 non-derived biomarkers (Fig. 4H, Fig. 5). To identify outlier plates, we modelled the distribution of plate medians for each biomarker as a normal distribution. For each biomarker, we computed the mean and standard deviation of plate medians across the 1,352 plates. Acceptable limits for plate medians (Fig. 5A) were then set based on the limits of a theoretical normal distribution of 1,352 points. Plates were subsequently flagged and removed as outliers if their median concentration was greater than or less than 3.3744 standard deviations of mean of the 1,352 plate medians. A non-biological origin for outlier plates was confirmed by examining concentrations of biomarkers from previously quantified clinical biochemistry data[9] (Fig. 5B).

### Derivation of composite biomarkers, ratios, and percentages.

To create the final post-QC dataset, we subsequently recomputed the 61 composite biomarkers and 81 biomarker ratios available in the original data from the 107 non-derived biomarkers using the ukbnmr R package (**Code Availability**, formulae in Table S2). We also computed 76 additional biomarker ratios not available in the original data (Fig. 1, Table S1, Table S2).

### Comparison to direct adjustment.

In Fig. 6A–C we compared these computed post-QC biomarkers to those obtained by directly adjusting the biomarker for the technical covariates as described above. When doing so, percentage biomarkers (Table S1) were logit transformed rather than log transformed.

For Fig. 6D,E we additionally created two datasets further adjusting for age, sex, and BMI: one in which the derived biomarkers were computed from the 107 non-derived biomarkers after adjustment, and one in which the derived biomarkers were directly adjusted. In both cases the starting point was the post-QC dataset described above, and robust linear regressions were fit for each biomarker on age, sex, and log transformed BMI. Post-QC biomarker concentrations were log or logit transformed as appropriate prior to fitting robust linear regressions, and residuals rescaled to absolute concentrations as described above.

### Genome-wide association study for alanine and albumin.

Genome-wide association studies (GWASs) were performed for original and post-QC alanine and albumin (Fig. 7A, Figure S7, Table S5). GWAS were performed using the version 3 UK Biobank genotype data, which has been imputed to the 1000 genomes, UK10K, and Haplotype Reference Consortium panels[6,21] using human genome build GRCh37.

Participants were filtered to those of White British ancestry (from sample-QC file downloadable from UK Biobank in Resource 531) and filtered for relatedness (first- or second- degree relationships, kinship > 0.0884, from relatedness file downloadable from UK Biobank in Resource 531)[22]. For participants with measurements at both baseline and repeat assessment the measurement at baseline assessment was chosen. In total, 111,450 participants were included in the GWASs. Among these, 111,446 had non-missing data for the original alanine concentrations, 110,552 had non-missing data for post-QC alanine, 111,443 had non-missing data for the original albumin concentrations, and 109,803 had non-missing data for post-QC albumin.

GWAS were performed using generalized linear models using plink 2 (version 2.00a3LM AVX2 Intel; 2 Mar 2021)[23] on probabilistic dosage data extracted from UK Biobank's Oxford BGEN format files[24]. Associations were tested for 8,587,133 bi-allelic single nucleotide polymorphisms with minor allele frequency >1% and imputation INFO score >0.4. Associations were adjusted for age (UK Biobank field #21003), sex (UK Biobank field #22001), genotyping chip (UK Biobank in Resource 531), and 10 genotype PCs (UK Biobank in Resource 531) as covariates. Alanine and albumin concentrations were log-transformed prior to association testing and quantile normalized (along with covariates) by plink2.

GWAS peaks (Figure S7, Table S5) were identified by partitioning associations with $P < 5 \times 10^{-8}$ into 1,703 pre-defined linkage disequilibrium blocks that are approximately independent in populations of European ancestry[25], and taking the SNP with the smallest P-value as the lead SNP for each peak. Lead SNPs were subsequently annotated using the Ensembl variant effect predictor (VEP)[26], filtering to variants located within protein coding genes. Predictions of variant effects from PolyPhen-2 and SIFT were also obtained using the Ensembl VEP REST API (https://grch37.rest.ensembl.org/documentation/info/vep_id_post)[27,28]. For lead SNPs not located within protein coding genes, the closest protein coding gene was identified using the annotables R package[29].

**Biomarker association scan for coronary artery disease and stroke.** Cox proportional hazards models were fit using the R package survival (version 3.2–7)[30] to test associations between original and post-QC biomarker levels with incident coronary artery disease and incident stroke (Fig. 7B, Table S6). Analysis was restricted to participants with NMR metabolite biomarker data at baseline assessment and who were not already taking lipid lowering medication (N = 96,258 participants). Cox proportional hazard models were fit adjusting for age and sex (UK Biobank field #31), on standardised log transformed biomarker concentrations (logit transformation for percentages), and excluding prevalent cases. Participant age, sex, and electronic health records were obtained under UK Biobank application number 30418. After excluding prevalent cases, for coronary artery disease analyses there were 95,790 samples including 2,717 incident cases, and for stroke analyses there were 95,623 samples including 1,366 cases. Follow-up in electronic health records was truncated at 1st February 2020 to preclude any potential confounding from SARS-CoV2 exposure. The maximum and median follow-up time available in the electronic health records were 12.8 years and 10.9 years respectively.

Incident coronary artery disease was defined as the first occurring event of myocardial infarction (international classification of diseases (ICD) revision 10 codes I21–I24, and I25.2) or cardiovascular surgery (percutaneous transluminal coronary angioplasty: Office of Population Censuses and Surveys Classification of Surgical Operations and Procedures (OPSC) 4th revision codes K49, K50.1, and K75; or coronary artery bypass grafting OPSC-4 codes K40–K46). Prevalent CAD was similarly defined, with the addition of self-reported events (UKB field #6150 code "heart attack" and #20002 code 1075, UKB field #20004 code 1070, and UKB field #20004 codes 1095 and 1523). Retrospective hospital records included those using ICD revision 9 coding, for which ICD-9 codes 410–412 were used to identify previous hospitalisation with myocardial infarction.

Incident stroke was defined as the first occurring event of any stroke (ICD-10 codes I60–164) or fatal cerebrovascular events (ICD-10 codes I60–I69 and F10). Prevalent stroke was defined as any previous hospitalisation for stroke (ICD-10 codes I60–164 or ICD-9 codes 430, 431, 434, or 436) or self-reported stroke at UK Biobank assessment (UKB field #6150 code "Stroke" and #20002 codes 1081, 1086, 1491, or 1583).

**Hierarchical clustering of biomarker correlations.** For Fig. 8C pairwise Pearson correlations were computed between each pair of biomarkers in the post-QC dataset then clustered using the topological overlap dissimilarity distance metric available through the weighted gene coexpression network analysis (WGCNA) R package (version 1.69)[31,32]. Briefly, this distance metric computes the distance between biomarkers as a weighted sum of (1) the strength of correlation coefficient between a pair of biomarkers, and (2) the similarity of their correlation coefficients to all other biomarkers[32]. To ensure strong negative correlations were clustered with strong positive correlations the absolute value of the pairwise Pearson correlation was taken prior to distance calculation, and a penalization exponent of 6 applied to shrink weak correlations towards zero as recommended in the package documentation[31].

**Principal components analysis.** For Fig. 8D–F, Figure S8, and Figure S9D–F, principal components analysis (PCA) was performed across participants. For PCA, biomarker concentrations were first either logit transformed (percentages; Table S1) or log transformed (all other biomarkers) and standardised to have mean of 0 and unit variance. Small offsets were applied to biomarkers with 0 values to facilitate log transformation, and for percentages, small negative offsets were applied for those with values of 100%. Missing values were imputed as the biomarker mean (0) as PCA could not handle missing data. Ratios with infinite (x/0) or undefined (0/0) values were also set to missing prior to log transformation. The prcomp function in R was used for PCA.

To identify biological correlates of fitted principal components (PCs) (Figure S8E, Figure S9E), linked phenotype data was obtained for UK Biobank participants. Spearman correlation coefficients were fit between each PC and participant age, sex, BMI (UK Biobank field #21001), ethnicity (White British/other; UK Biobank field #21000), fasting time (UK Biobank field #74), and lipid lowering medication usage (UK Biobank fields #6153 and #6177). We also examined whether PCs correlated with UK Biobank assessment visit (repeat vs. baseline). Spearman correlation was used as several of biological and technical correlates were binary (yes/no) variables.

## Data availability

All data described are available through UK Biobank subject to approval from the UK Biobank access committee. See https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access for further details.

Extensive diagnostic plots showing the impact of technical variation and its removal on all biomarkers can be obtained from FigShare[7] at https://doi.org/10.6084/m9.figshare.21546576.v1.

## Code availability

Code for removing technical variation from the NMR biomarker data and for re-computing composite biomarkers and ratios is available through the R package ukbnmr, which also provides tools for extracting and processing NMR metabolite biomarker data and associated quality control tags from UK Biobank. The ukbnmr R package is available to download from CRAN at https://cran.r-project.org/web/packages/ukbnmr/ or GitHub at https://github.com/sritchie73/ukbnmr/ and is permanently archived by Zenodo[33] at https://doi.org/10.5281/zenodo.7515459.

Code underlying this paper are separately available at https://github.com/sritchie73/UKB_NMR_QC_paper/. This repository and specific release for this paper are permanently archived by Zenodo[34] at https://doi.org/10.5281/zenodo.7310524.

Received: 21 July 2022; Accepted: 10 January 2023;
Published online: 31 January 2023

## References

1. Würtz, P. *et al*. Quantitative Serum Nuclear Magnetic Resonance Metabolomics in Large-Scale Epidemiology: A Primer on -Omic Technology. *Am. J. Epidemiol.* 1–13 (2017).
2. Ala-Korpela, M., Zhao, S., Järvelin, M.-R., Mäkinen, V.-P. & Ohukainen, P. Apt interpretation of comprehensive lipoprotein data in large-scale epidemiology: disclosure of fundamental structural and metabolic relationships. *Int. J. Epidemiol.* **51**, 996–1011 (2022).
3. Soininen, P., Kangas, A. J., Würtz, P., Suna, T. & Ala-Korpela, M. Quantitative serum nuclear magnetic resonance metabolomics in cardiovascular epidemiology and genetics. *Circ. Cardiovasc. Genet.* **8**, 192–206 (2015).
4. Julkunen, H. *et al*. Atlas of plasma nuclear magnetic resonance biomarkers for health and disease in 118,461 individuals from the UK Biobank. *medRxiv* https://doi.org/10.1101/2022.06.13.22276332 (2022).
5. Sudlow, C. *et al*. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
6. Bycroft, C. *et al*. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
7. Ritchie, S. Quality control and removal of technical variation of NMR metabolic biomarker data in ~120,000 UK Biobank participants. *Figshare.* https://doi.org/10.6084/m9.figshare.21546576.v1 (2022).
8. Harper, A. E., Miller, R. H. & Block, K. P. Branched-chain amino acid metabolism. *Annu. Rev. Nutr.* **4**, 409–454 (1984).
9. Allen, N. E. *et al*. Approaches to minimising the epidemiological impact of sources of systematic and random variation that may affect biochemistry assay data in UK Biobank. *Wellcome Open Research* **5**, (2020).
10. De Livera, A. M. *et al*. Statistical methods for handling unwanted variation in metabolomics data. *Anal. Chem.* **87**, 3606–3615 (2015).
11. Bagheri, B., Alikhani, A., Mokhtari, H. & Rasouli, M. The Ratio of Unesterified/esterified Cholesterol is the Major Determinant of Atherogenicity of Lipoprotein Fractions. *Med. Arch.* **72**, 103–107 (2018).
12. Inouye, M. *et al*. Novel Loci for metabolic networks and multi-tissue expression studies reveal genes for atherosclerosis. *PLoS Genet.* **8**, e1002907 (2012).
13. Würtz, P. *et al*. Metabolite Profiling and Cardiovascular Event Risk: A Prospective Study of Three Population-Based Cohorts. *Circulation* https://doi.org/10.1161/CIRCULATIONAHA.114.013116 (2015).
14. Li, J. & Ji, L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity* **95**, 221–227 (2005).
15. Vatcheva, K. P., Lee, M., McCormick, J. B. & Rahbar, M. H. Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies. *Epidemiology* **6** (2016).
16. Nath, A. P. *et al*. Multivariate Genome-wide Association Analysis of a Cytokine Network Reveals Variants with Widespread Immune, Haematological, and Cardiometabolic Pleiotropy. *Am. J. Hum. Genet.* **105**, 1076–1090 (2019).
17. Sliz, E. *et al*. Metabolomic consequences of genetic inhibition of PCSK9 compared with statin treatment. *Circulation* **138**, 2499–2512 (2018).
18. Bland, J. M. & Altman, D. G. Measurement error proportional to the mean. *BMJ* **313**, 106 (1996).
19. Huber, P. J. *Robust Statistics*. (John Wiley & Sons, 2004).
20. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S*. 4th edn (Springer, 2002).
21. Loh, P.-R. *et al*. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).
22. Manichaikul, A. *et al*. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
23. Chang, C. C. *et al*. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
24. Band, G. & Marchini, J. BGEN: a binary file format for imputed genotype and haplotype data. *bioRxiv* 308296, https://doi.org/10.1101/308296 (2018).
25. Berisa, T. & Pickrell, J. K. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics* **32**, 283–285 (2016).
26. McLaren, W. *et al*. The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
27. Adzhubei, I. A. *et al*. A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
28. Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M. & Ng, P. C. SIFT missense predictions for genomes. *Nat. Protoc.* **11**, 1–9 (2016).
29. Steinbaugh, M., Turner, S. & Wolen, A. stephenturner/annotables: Ensembl 90. *Zenodo* https://doi.org/10.5281/zenodo.996854 (2017).
30. Therneau, T. M. & Grambsch, P. M. *Modeling Survival Data: Extending the Cox Model*. (Springer Science & Business Media, 2013).
31. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
32. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4**, Article17 (2005).
33. Ritchie, S. C. ukbnmr R package version 1.4. *Zenodo* https://doi.org/10.5281/zenodo.7515459 (2023).
34. Ritchie, S. C. Quality control and removal of technical variation of NMR metabolic biomarker data in ~120,000 UK Biobank participants. *Zenodo* https://doi.org/10.5281/zenodo.7310524 (2022).

## Acknowledgements

The authors are grateful to UK Biobank for access to data to undertake this study (Project #30418).

Nightingale Health Plc is acknowledged for early access to the UK Biobank NMR biomarker data and discussions regarding sources of experimental variation.

This work was supported by core funding from the British Heart Foundation (BHF) (RG/13/13/30194; RG/18/13/33946) and the National Institute for Health and Care Research (NIHR) Cambridge Biomedical Research Centre (BRC-1215-20014).

This work was supported by Health Data Research (HDR) UK, which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and Wellcome.

This work was performed using resources provided by the Cambridge Service for Data Driven Discovery (CSD3) operated by the University of Cambridge Research Computing Service (www.csd3.cam.ac.uk), provided by Dell EMC and Intel using Tier-2 funding from the Engineering and Physical Sciences Research Council (capital grant EP/P020259/1), and DiRAC funding from the Science and Technology Facilities Council (www.dirac.ac.uk).

S.C.R was funded by the NIHR Cambridge BRC (BRC-1215-20014) and by the BHF programme grant (RG/18/13/33946). P.S. was supported by a Rutherford Fund Fellowship from the Medical Research Council grant MR/S003746/1. S.A.L. was supported by a Canadian Institutes of Health Research postdoctoral fellowship (MFE-171279). T.B. was funded by the NIHR Blood and Transplant Research Unit in Donor Health and Genomics (NIHR BTRU-2014-10024). L.P. and S.K. were funded by a BHF Programme Grant (RG/18/13/33946). J.D. holds a British Heart Foundation Professorship and a NIHR Senior Investigator Award. M.I. is supported by the Munz Chair of Cardiovascular Prediction and Prevention and the NIHR Cambridge Biomedical Research Centre (BRC-1215-20014). MI was also supported by the UK Economic and Social Research 878 Council (ES/T013192/1).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

## Author contributions

Conceptualization: S.C.R., P.S., S.K., S.A.L., L.P., J.D., E.D.A., A.S.B. and M.I. Methodology: S.C.R. Software: S.C.R. Validation: S.C.R., P.S., S.K., S.A.L. and L.P. Formal analysis: S.C.R. Investigation: S.C.R., P.S. and S.K. Resources: S.C.R., P.S., T.B., J.D., E.D.A., A.S.B. and M.I. Data Curation: S.C.R., P.S., S.K. and T.B. Writing - Original Draft: S.C.R. Writing - Review & Editing: S.C.R., P.S., S.K., S.A.L., T.B., L.P., J.D., E.D.A., A.S.B. and M.I. Visualization: S.C.R. Revisions: S.C.R. Supervision: J.D., E.D.A., A.S.B. and M.I. Project administration: S.C.R., P.S., J.D., E.D.A., A.S.B. and M.I. Funding acquisition: J.D., E.D.A., A.S.B. and M.I.

## Competing interests

During the course of this project P.S. became a full-time employee of GSK Plc. All significant contributions to this study were made prior to this role and GSK Plc had no input to the study. J.D. serves on scientific advisory boards for AstraZeneca, Novartis, and UK Biobank, and has received multiple grants from academic, charitable and industry sources outside of the submitted work. A.B. reports institutional grants from AstraZeneca, Bayer, Biogen, BioMarin, Bioverativ, Novartis, Regeneron and Sanofi. The remaining authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41597-023-01949-y.

**Correspondence** and requests for materials should be addressed to S.C.R.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.