

Supplementary Information

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

Scott C. Ritchie^{1-4,*}, Praveen Surendran^{3,5,6}, Savita Karthikeyan³, Samuel A. Lambert^{1-3,5}, Thomas Bolton^{3,7}, Lisa Pennells³, John Danesh^{3-5,7,8}, Emanuele Di Angelantonio^{3-5,7}, Adam S. Butterworth^{3-5,7}, Michael Inouye^{1-5,9,10}

Affiliations

¹Cambridge Baker Systems Genomics Initiative, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK

²Cambridge Baker Systems Genomics Initiative, Baker Heart & Diabetes Institute, Melbourne, Victoria, Australia

³British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK

⁴British Heart Foundation Centre of Research Excellence, University of Cambridge, Cambridge, UK

⁵Health Data Research UK Cambridge, Wellcome Genome Campus and University of Cambridge, Cambridge, UK

⁶Rutherford Fund Fellow, Department of Public Health and Primary Care, University of Cambridge, CB1 8RN, UK

⁷National Institute for Health Research Blood and Transplant Research Unit in Donor Health and Genomics, University of Cambridge, Cambridge, UK

⁸Department of Human Genetics, Wellcome Sanger Institute, Hinxton, UK

⁹Department of Clinical Pathology, University of Melbourne, Parkville, Victoria, Australia

¹⁰The Alan Turing Institute, London, UK

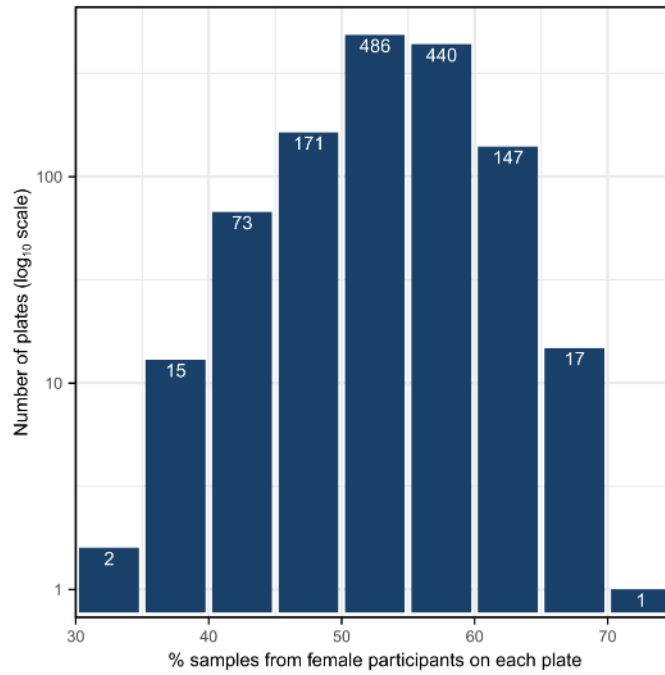
* Corresponding author: sr827@medschl.cam.ac.uk (SCR)

Table of Contents

Supplementary Figures	2
Figure S1: Allocation of UKB participant to plate is not random with respect to participant sex.	2
Figure S2: Example of intra-plate variation across plate columns.	3
Figure S3: Binning of plates within spectrometer to model drift over time.	4
Figure S4: Non-normality of biomarker concentrations requires robust linear regression.....	5
Figure S5: Obtaining absolute concentrations after removal of technical variation.....	6
Figure S6: Composite biomarkers can be re-derived without loss of information.....	7
Figure S7: Manhattan plots for GWAS of alanine and albumin after removing technical variation.....	8
Figure S8: Principal components explaining more than 1% of variation in samples.	9
Figure S9: Characteristics of the raw NMR metabolite biomarker data.....	10
Supplementary Table Legends	11
Table S1: Biomarker details.....	11
Table S2: Derivation formulae for derived biomarkers.	11
Table S3: Offsets for log transformation of biomarkers with concentrations of 0.	11
Table S4: Annotations of GWAS peaks for alanine and histidine.....	11
Table S5: Associations between biomarkers and incident coronary artery disease and stroke.....	11
Supplementary Notes	12
Unsuccessful approaches to removing technical variation.....	12
Supplementary Methods	13
Sample quality control of pre-release data.....	13
Quality control of internal control samples.....	13

45 **Supplementary Figures**

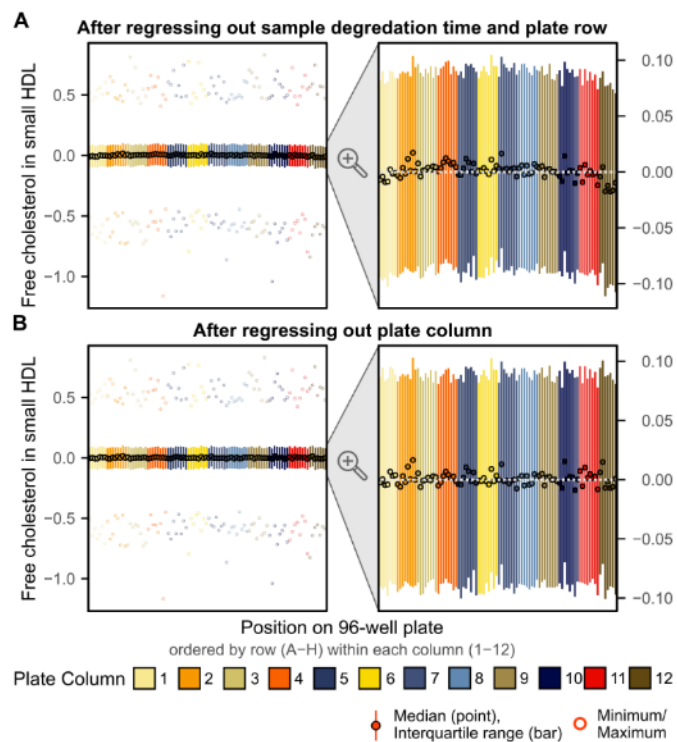
46



47

48 **Figure S1: Allocation of UKB participant to plate is not random with respect to participant sex.**

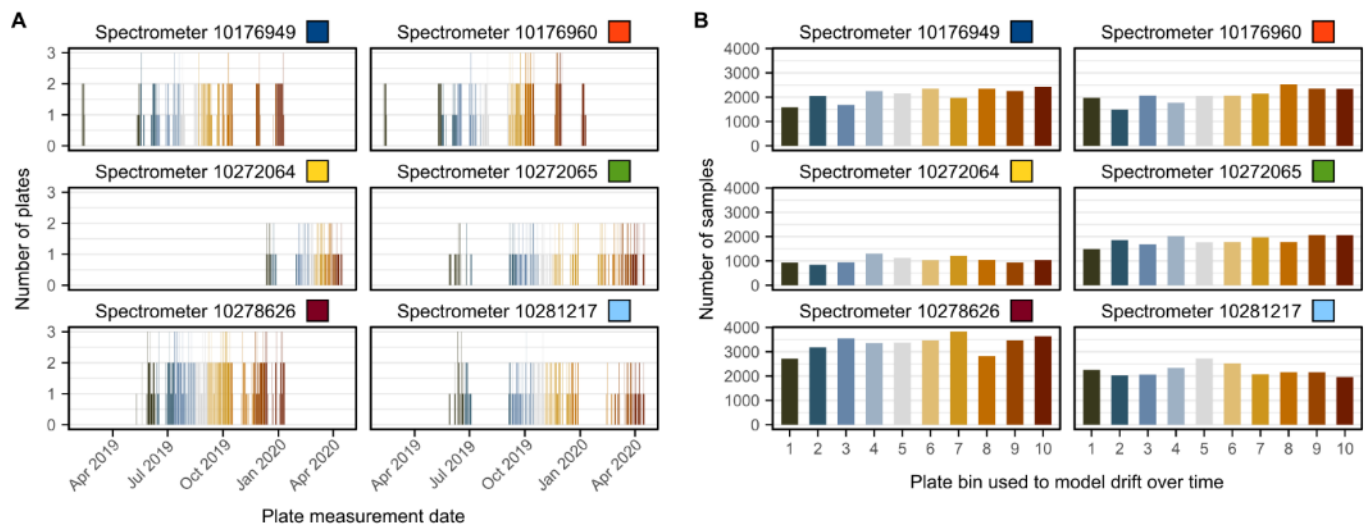
49 Histogram shows the number of plates by percentage of samples from female participants in bins of 5%. The
50 number of plates in each bin is annotated.



51

52 **Figure S2: Example of intra-plate variation across plate columns.**

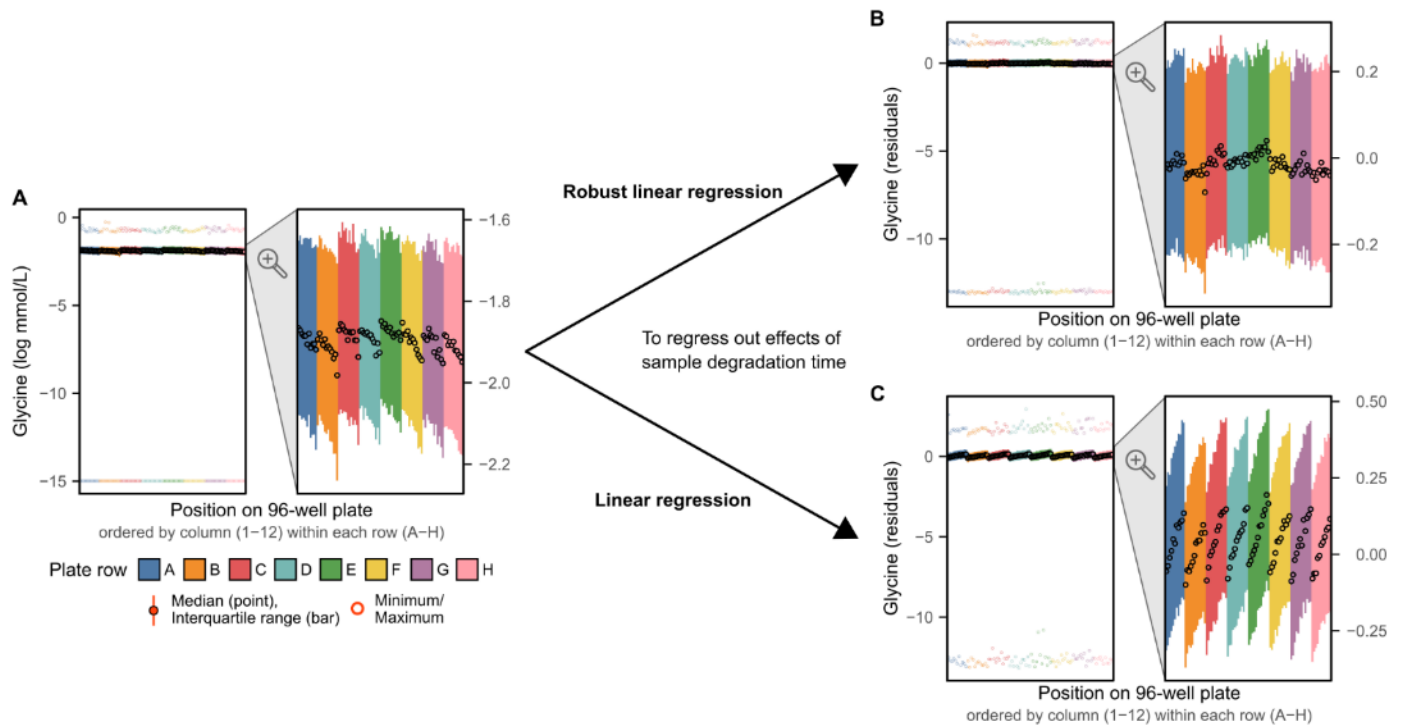
53 **A)** Shows inter-column variation for free cholesterol in small HDL after regressing out sample degradation
 54 time and plate row. The left plot shows a summary of the regression residuals (minimum, maximum,
 55 interquartile range, and median) for each plate position grouped and coloured by plate column (1–12)
 56 and ordered within each column by plate row (A–H). A zoomed in view focusing on the interquartile range is
 57 shown on the right. A grey dotted line is overlaid at $y=0$, lower values can be seen for samples measured in
 58 columns on the outside of the plate (columns 1 and 12). **B)** Shows a summary of residuals for each plate
 59 position after regressing out plate column number as a categorical variable.



60

61 **Figure S3: Binning of plates within spectrometer to model drift over time.**

62 For each of the six spectrometers, plates were ordered by measurement date, and split into 10 bins containing
 63 approximately equal number of plates (plates measured on the same date were allocated to the same bin). **A)**
 64 Shows the number of plates measured on each date, coloured by bin, with each plot corresponding to a single
 65 spectrometer. Colours next to each spectrometer indicate the colours used in **Figure 3D** and **Figure 4** when
 66 plotting summaries of biomarker concentrations on each plate. **B)** Shows the total number of samples
 67 measured on plates allocated to each bin. Within each spectrometer plate bin treated as a categorical variable
 68 and the bin with the largest number of samples was used as the reference group to regress out drift over time
 69 with spectrometer (**Methods, Figure S3F**).



70

71

Figure S4: Non-normality of biomarker concentrations requires robust linear regression.

72

Example comparing outcome of regressing glycine (raw values after log transformation; **A**) on time between sample preparation and sample measurement (sample degradation time) using (**B**) robust linear regression vs. (**C**) linear regression. Each plot shows a summary of glycine across all plates at each plate position (minimum, maximum, median, and interquartile range) with plate positions coloured by plate row (A–H), and within each row organised from left to right by plate column (1–12). Zoomed in plots focus on the interquartile range. In **A**) units are log mmol/L, while **B**) and **C**) show residuals from the respective regression.

73

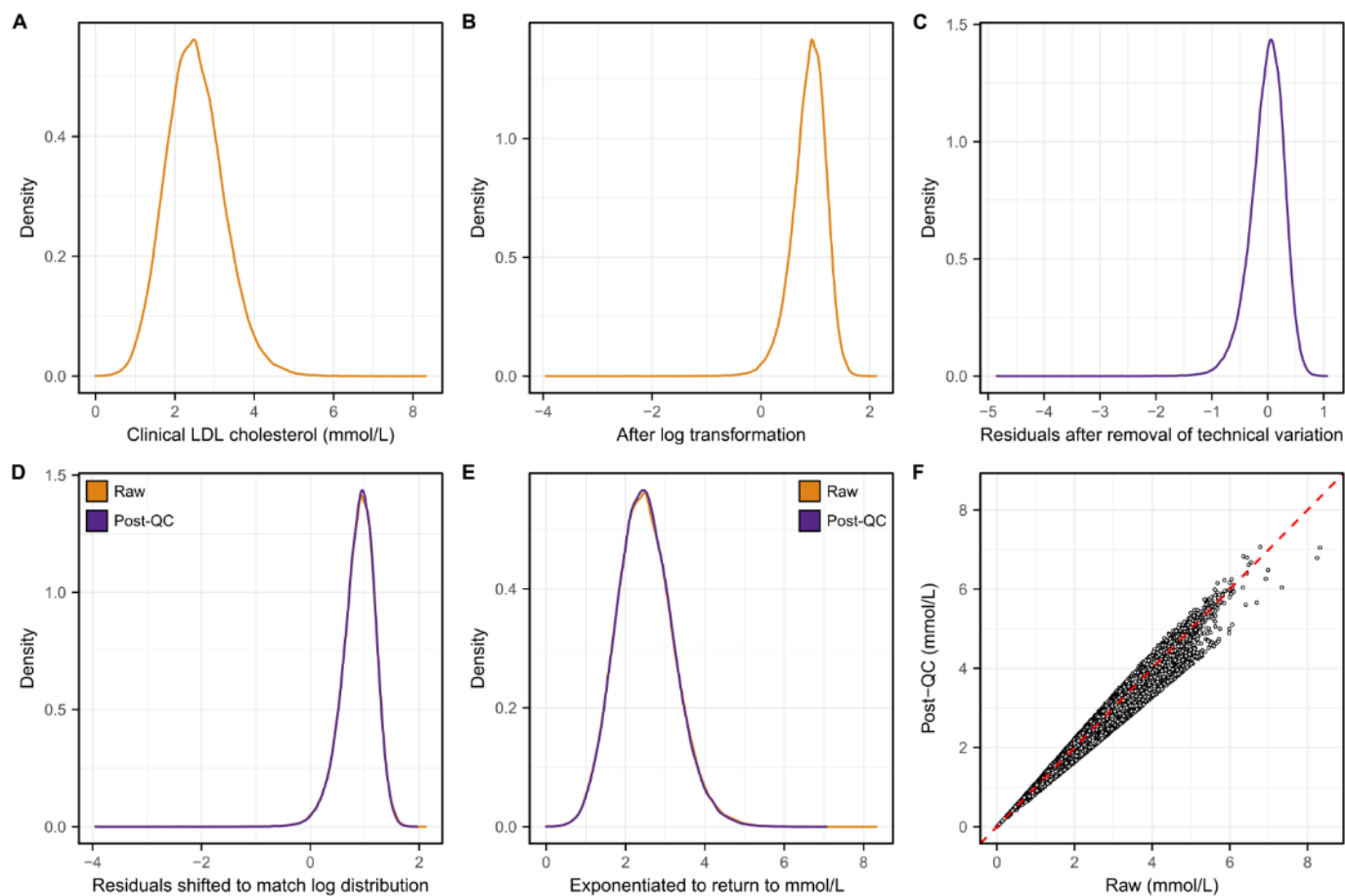
74

75

76

77

78



79

80

81

82

83

84

85

86

87

88

89

90

91

92

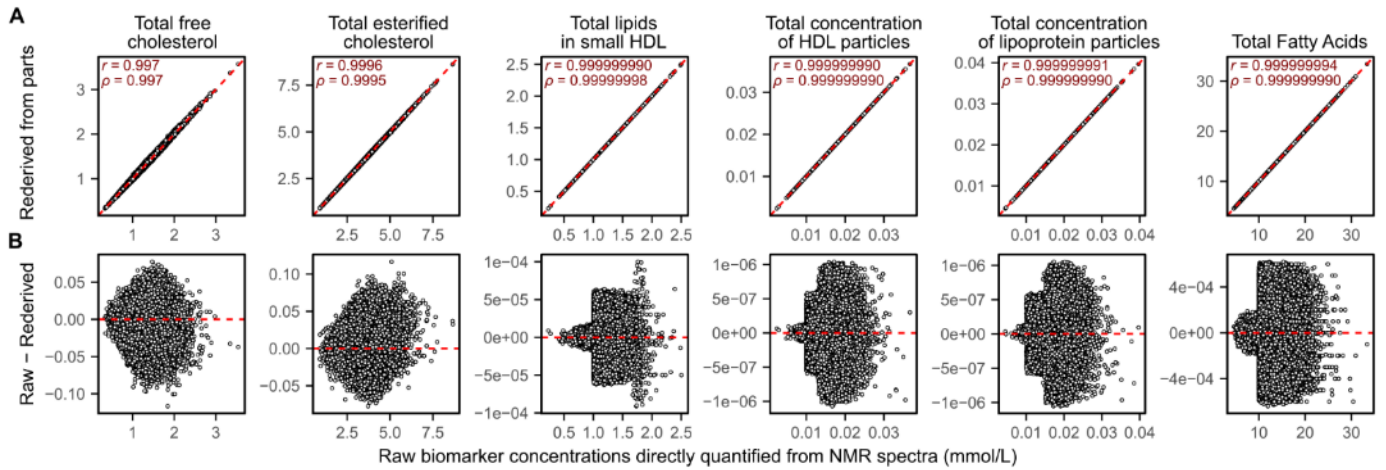
93

94

95

Figure S5: Obtaining absolute concentrations after removal of technical variation.

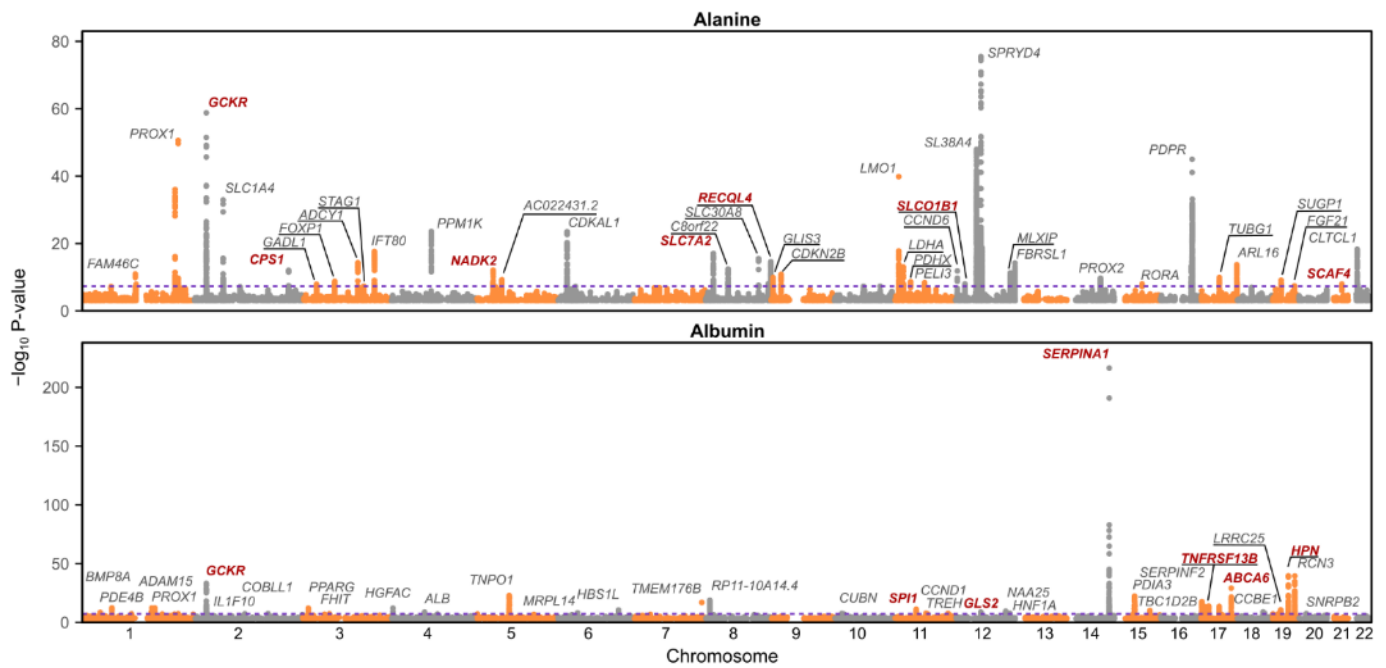
Example given using concentrations of clinical low-density lipoprotein (LDL) cholesterol. **A**) Shows the distribution in the raw data. **B**) Shows the distribution after log transformation. As some samples had a concentration of 0 mmol/L, a small offset (half the minimum non-zero value: 0.0194; **Table S3**) was applied to all concentrations to enable these samples to be included after log transformation. **C**) Shows the distribution of the residuals after regressing out in four sequential robust linear regressions sample degradation time, then plate row, then plate column, the plate drift over time within spectrometer. **D**) Shows the distribution of residuals overlaid with the distribution of log raw concentrations from **Figure S8B** after shifting the residuals distribution to have the same median as the log raw distribution. **E**) Absolute concentrations are subsequently obtained by taking the exponent of the shifted residuals and removing the originally applied offset of 0.0194 (**Table S3**) from all samples. A final offset of 4.43×10^{-5} (**Table S3**) is applied to ensure there are no negative concentrations. Here, the distribution of absolute concentrations after removal of technical variation is shown (purple) overlaid the distribution from the raw data shown in **Figure S8A**. **F**) Scatterplot comparing concentrations of clinical LDL cholesterol in the raw data (x-axis) to concentrations after removal of technical variation (y-axis). The red dashed line on the diagonal shows $y=x$, where concentrations fall if unchanged.



96

97 **Figure S6: Composite biomarkers can be re-derived without loss of information**

98 **A)** Compares raw biomarker concentrations (x-axes), quantified directly from the NMR spectra, to
 99 concentrations recomputed from the biomarker's parts (y-axes) for the six biomarkers most different (by
 100 Pearson and Spearman correlation) after re-derivation. The Pearson correlation coefficient (r) and Spearman
 101 rank correlation coefficient (ρ) between raw concentrations and recomputed concentrations are given in the
 102 top left of each plot. The dashed red line in each plot shows $y=x$. Formulae for each biomarker from parts are
 103 given in **Table S2**. For example, total free cholesterol can be computed by summing the concentrations of
 104 free cholesterol in the 14 lipoprotein subclasses. **B)** Shows the difference between raw and re-derived
 105 biomarker levels (y-axes) as a function of raw biomarker levels (x-axes), and that these differences are within
 106 numeric error for any given value of the biomarker.



107

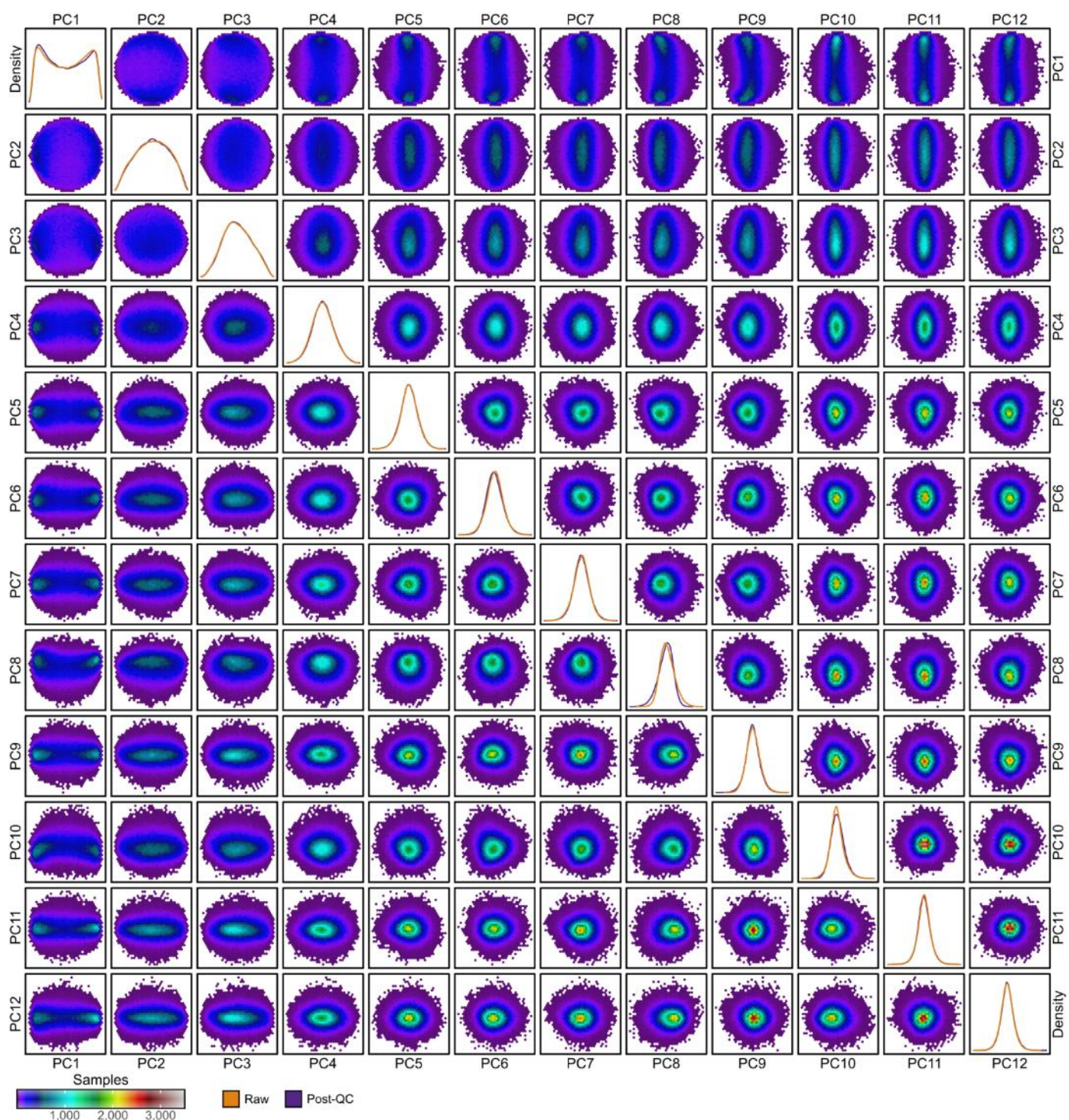
108

109

110

111

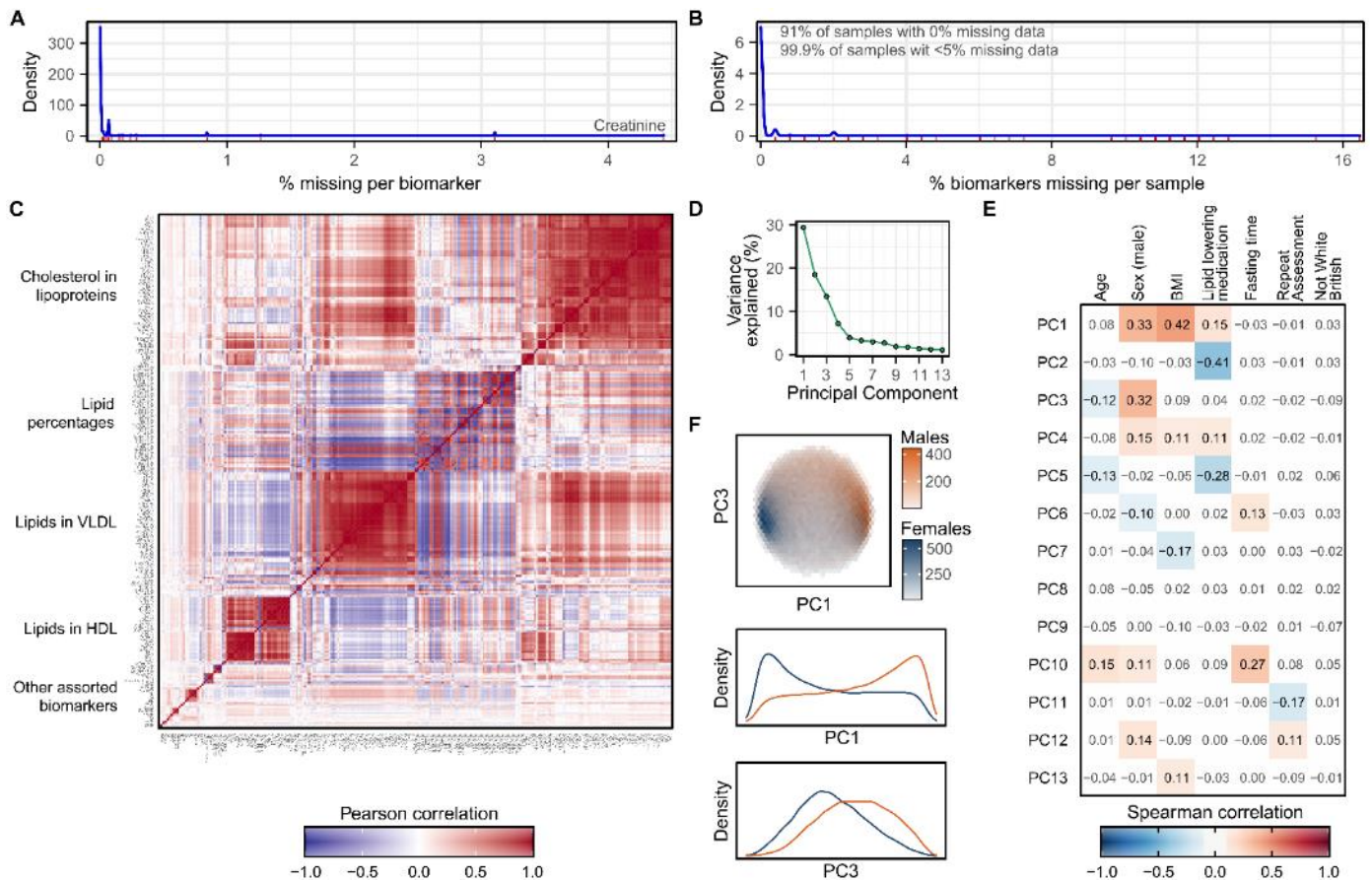
Figure S7: Manhattan plots for GWAS of alanine and albumin after removing technical variation
 Manhattan plots show $-\log_{10}$ P-values from GWAS of 8.5 million common (frequency > 1%) SNPs (Methods). P-values greater than 0.001 are omitted. Peaks are annotated with their nearest protein coding gene. Genes in red indicate protein coding SNPs. Annotations for all peaks are detailed in Table S4.



112

113 **Figure S8: Principal components explaining more than 1% of variation in samples.**

114 Heatmaps show bivariate sample densities for each pair of the first 12 principal components (PCs) with each
 115 hexagonal cell coloured based on sample count. Heatmaps above the diagonal show the PCs computed from
 116 the raw data, and heatmaps below the diagonal show the PCs computed from the data after removal of
 117 technical variation (post-QC data). The diagonal shows density plots for the distributions of each PC in the
 118 raw and post-QC data.



119

120

121

122

123

124

125

126

127

128

129

130

131

Figure S9: Characteristics of the raw NMR metabolite biomarker data

A) Density plot showing distribution of missingness across the 249 NMR metabolite biomarkers in the raw UK Biobank data. 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 249 biomarkers, shown in the same row and column order as in **Figure 8C**. 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.

132 **Supplementary Table Legends**

133 **Table S1: Biomarker details.**

134 Biomarker: variable name used by Nightingale and column name in raw data. Description: full name or
135 description of the biomarker. Units: units of absolute concentration. Group: biomarker group as listed by
136 Nightingale. Sub-group: biomarker sub-group as listed by Nightingale. Type: indicates whether the
137 biomarker is one of the 107 non-derived biomarkers, or can be derived from these 107 non-derived
138 biomarkers. Formulae for biomarker derivation are provided in **Table S2** and code for biomarker derivation
139 is available in the ukbnmr R package (**Code Availability**). UKB Field ID: field ID for the biomarker in UK
140 Biobank. QC Flag Field ID: field ID for the corresponding biomarker measurement QC flags in UK Biobank.
141 Biomarkers without UKB Field IDs are the 76 additional derived biomarkers that are not part of the
142 Nightingale platform and not available to download from UK Biobank.

143 **Table S2: Derivation formulae for derived biomarkers.**

144 Formula for deriving each of the 142 composite biomarkers, ratios, and percentages on the Nightingale
145 platform as well as for the 76 additional derived biomarkers. For each biomarker two formulae are given: a
146 formula that uses only the 107 non-derived biomarkers, and a simplified formula which expresses each
147 biomarker in terms of its most relevant parts, in many cases using other derived biomarkers. Code for
148 biomarker derivation is available in the ukbnmr R package (**Code Availability**).

149 **Table S3: Offsets for log transformation of biomarkers with concentrations of 0.**

150 Log transformation cannot be applied to values of 0. For these biomarkers, a small offset was added to all
151 concentrations so that all values were greater than 0. The offset used for each biomarker was half of the
152 lowest non-zero concentration. When returning technical covariate adjusted residuals to absolute
153 concentrations (**Figure S8**), after inverting the log transformation and removing the log offset, a small offset
154 was required to prevent negative values. Note that these are one or more orders of magnitude smaller than
155 the minimum non-zero value.

156 **Table S4: Annotations of GWAS peaks for alanine and histidine.**

157 Lead SNPs and gene annotations for each independent LD block (**Methods**) with $P < 5 \times 10^{-8}$ for any variant
158 in GWAS of post-QC alanine and histidine concentrations. Genomic locations are on genome build GRCh37.
159 Beta indicates standard deviation change in quantile-normalized log biomarker concentrations per copy of
160 the effect allele, and SE the standard error. The frequency corresponds to the effect allele frequency in the
161 UK Biobank participants analysed (N=111,743 for alanine, N=111,575 for histidine, **Methods**). The closest
162 protein coding gene for the lead SNP (**Methods**). Where the lead SNP was located in the gene, the most
163 severe consequence and variant impact from the Ensembl variant effect predictor (VEP) are reported, along
164 with variant effect predictions from PolyPhen-2 and SIFT (**Methods**).

165 **Table S5: Associations between biomarkers and incident coronary artery disease and stroke**

166 Hazard ratios (HR), 95% confidence intervals (L95, U95), and P-values from Cox proportional hazards
167 models fit on raw and post-QC biomarker concentrations for incident coronary artery disease and incident
168 stroke over 12.8 years of follow-up (**Methods**). Models were fit adjusting for age and sex. Participants with
169 prevalent events or taking lipid lowering medication were excluded. Biomarker concentrations were log
170 transformed (logit for percentages) and standardised for model fitting: hazard ratios are per standard deviation
171 increase in log biomarker concentrations / logit biomarker percentages. Biomarkers are sorted by P-value for
172 association between post-QC concentrations and coronary artery disease.

173 **Supplementary Notes**

174 *Unsuccessful approaches to removing technical variation*

175 Several additional approaches to identifying and removing technical variation were explored without success.
176 In particular, we attempted to utilise both the internal control samples and blind duplicate participants to
177 estimate and remove inter-plate variation.

178 On each plate, internal controls were measured on opposite corners in wells A01 and H12 (**Methods**). These
179 internal control samples were measured in pairs, with a “high” control sample in well A01 and a “low” control
180 sample in well H12 (**Supplementary Methods**). In total, four pairs of internal controls were measured across
181 all plates (**Supplementary Methods**). Theoretically, concentrations for each biomarker should be identical
182 across plates for each control sample. Based on this, we attempted to use Removal of Unwanted Variation
183 (RUV) K-means (De Livera et al., 2015) to learn and remove the inter-plate variation based on the differences
184 between concentrations across plates within each set of control sample pairs. However, inspection of post-
185 QC plots (similar to those shown in **Figure 3G-I** and **Figure S3G**) showed no reduction in inter-plate
186 variation while also showing large structural changes in concentrations when comparing raw to post-QC
187 values (data not shown). We also attempted a simpler approach of estimating inter-plate variation from the
188 internal control samples as their difference from the median across all plates they were measured on, with
189 similar results (data not shown). Both approaches were also explored for identifying outlier plates in the post-
190 QC data, with similar unsuccessful results.

191 We also explored the possibility of using the blind duplicate samples (**Methods**) to estimate inter-plate
192 variation. However, there were several complications that led us to abandoning such an approach. First, there
193 were 5 plates containing no blind duplicate samples. Second, it was unclear how to utilize the blind duplicates
194 as each were only measured twice (or in a few cases three times), leading to each plate having a mixture of
195 different blind duplicates. Most plates (N=1,282) had four samples belonging to blind duplicates, and a small
196 number (N=18) comprised almost entirely of blind duplicate samples. Additionally there were several
197 participants (N=6) with blind duplicate samples where the multiple measurements were taken on the same
198 plate.

199 Finally, we also explored adjusting biomarker concentrations for the various sample and biomarker quality
200 control tags provided by Nightingale, for example tags indicating sample dilution. These also had the effect
201 of introducing large structural changes when comparing raw to post-QC values (e.g. shifting all Albumin
202 levels below 30 g/L up by 10 g/L; data not shown).

203 **Supplementary Methods**

204 *Sample quality control of pre-release data*

205 Pre-release raw data was provided to early access analysts in a single flat file of wide format with 126,846
206 rows: with each row corresponding to a single well on a 96 well-plate and columns corresponding to the 249
207 quantified biomarkers along with sample information.

208 First, on the advice of Nightingale Health Plc., we removed from the raw data 40 rows corresponding to
209 samples erroneously measured despite having insufficient sample material. These included 37 samples
210 present in the raw data available to download from UK Biobank (**Methods**). Next, 441 rows with missing
211 values for all 249 biomarkers were removed.

212 Next we resolved duplicate samples. In total, there were four samples measured twice (rows with the same
213 Nightingale sample identifier and UK Biobank visit annotation, measured on different plates and wells):
214 samples with Nightingale identifiers 1219449, 4478891, 2553581 and 1556057, each corresponding to an
215 anonymised UK Biobank participant at baseline assessment. From each pair of duplicates one sample was
216 kept. For samples with Nightingale identifiers 1219449 and 4478891, the rows corresponding to
217 measurements on plate 490000006107 were removed as the sample annotation information indicated these
218 were measured on spectrometer 10176949, whereas the other 92 samples on that plate were measured on
219 spectrometer 10278626. For samples with Nightingale identifiers 2553581 and 1556057, the rows
220 corresponding to measurements on plate 490000006069 were removed as they were reportedly measured 48
221 hours prior to the rest of the samples on plate 490000006069, whereas their measurements on plate
222 490000006068 occurred day as all other samples on that plate.

223 Finally, samples and biomarker measurements with quality control tag “technical error” were resolved.
224 Nightingale Health Plc. advised that all samples and biomarker measurements with this tag should be set to
225 missing. Among the 189 samples with this tag, one sample had non-missing data and was removed (the other
226 188 were removed above as they already had all missing data). Among the 2,392 biomarker measurements
227 with this tag (across 292 samples), 69 measurements (in 3 samples) were not missing in the raw data and
228 subsequently set to missing here. All samples with this quality control tag have data set to missing in the raw
229 data available to download from UK Biobank.

230 In total 126,360 samples passed quality control. These included 6,359 blind duplicate samples: samples from
231 participants sent by UK Biobank multiple times with differing sample identifiers for UK Biobank to assess
232 the internal consistency of Nightingale Health’s NMR metabolite biomarker quantification pipeline. In total,
233 121,758 participants passed quality control: 118,047 with measurements at baseline assessment and 5,139
234 with measurements at first repeat assessment, including 1,428 participants with measurements at both
235 timepoints.

236 *Quality control of internal control samples*

237 Internal control samples were placed on opposite corners of each of the 1,352 plates (wells A01 and H12) by
238 Nightingale Health to assess internal consistency of their NMR metabolite biomarker quantification pipeline.
239 Data from these internal control samples was provided to early access analysts in a single flat file of wide
240 format with 2,713 rows: with each row corresponding to a single well on a 96 well-plate and columns
241 corresponding to the 249 quantified biomarkers along with sample information. In total, 2,698 control
242 samples passed quality control.

243 First, duplicate internal control samples resolved. For plate 490000006069, there were two rows
244 corresponding to wells A01 and H12, each pair of duplicates with internal control identifiers 180827 and
245 180829 respectively. From each pair, we kept the measurement taken closest in time to the rest of the plate.
246 For plate 490000006107, there were two rows corresponding to wells A01 and H12, each pair of duplicates
247 with internal control identifiers 180829 and 180827 respectively. From each pair, we kept the measurement
248 taken on the same spectrometer (10278626) as the rest of the plate.

249 Next, several other problematic internal control samples were also identified and removed. Control samples
250 measured in well H12 on plates 490000007201 and 490000006663 were discarded, as a “high” control
251 sample was used instead of a “low” control sample: control sample 190404 was measured in well H12 on
252 these two plates but in well A01 on 213 other plates. Internal control samples 180817, 190125, and 191127
253 were also removed as they were measured on only 1-2 plates each in well A01, whereas the other 8 internal
254 control samples were measured in pairs on 213–444 plates each.

255 Finally, four failed internal control samples with extreme biomarker concentrations were removed on the
256 advice of Nightingale Health Plc: control sample 180827 on plate 490000006201 in well H12, control sample
257 180829 on plate 490000005965 in well A01, control sample 190328 on plate 490000006714 in well H12,
258 and control sample 190328 on plate 490000006802 in well H12. None of these four plates were flagged as
259 outliers for any biomarker when identifying outlier plates driven by unexplained technical variation
260 (**Methods**).