

# An exposome-wide assessment of 6600 SomaScan proteins with non-genetic factors in Chinese adults

Ka Hung Chan<sup>1\*</sup>, Jonathan Clarke<sup>1\*</sup>, Maria G. Kakkoura<sup>1\*</sup>, Andri Iona<sup>1</sup>, Baihan Wang<sup>1</sup>,  
Charlotte Clarke<sup>1</sup>, Neil Wright<sup>1</sup>, Pang Yao<sup>1</sup>, Mohsen Mazidi<sup>1</sup>, Pek Kei Im<sup>1</sup>, Maryam  
Rahmati<sup>1</sup>, Christiana Kartsonaki<sup>1</sup>, Sam Morris<sup>1</sup>, Hannah Fry<sup>1</sup>, Iona Y Millwood<sup>1</sup>, Robin G  
Walters<sup>1</sup>, Yiping Chen<sup>1</sup>, Huaidong Du<sup>1</sup>, Ling Yang<sup>1</sup>, Daniel Avery<sup>1</sup>, Dan Valle Schmidt<sup>1</sup>,  
Yongmei Liu<sup>5</sup>, Canqing Yu<sup>2,3,4</sup>, Dianjianyi Sun<sup>2,3,4</sup>, Jun Lv<sup>2,3,4</sup>, Michael Hill<sup>1</sup>, Liming Li<sup>2,3,4</sup>,  
Robert Clarke<sup>1</sup>, Derrick A Bennett<sup>1†</sup>, Zhengming Chen<sup>1†</sup>, on behalf of China Kadoorie  
Biobank Collaborative Group<sup>#</sup>

1. Clinical Trial Service Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK
2. Department of Epidemiology & Biostatistics, School of Public Health, Peking University, Beijing, China
3. Peking University Center for Public Health and Epidemic Preparedness and Response, Beijing, China
4. Key Laboratory of Epidemiology of Major Diseases (Peking University), Ministry of Education, Beijing, China
5. NCDs Prevention and Control Department, Qingdao CDC, China

\* *Co-first author*; † *Co-corresponding author*; # *Members of the CKB Collaborative Group are shown in the Appendix.*

## Address for correspondence:

Assoc. Professor Derrick Bennett  
CTSU, Big Data Institute,  
Old Road Campus,  
University of Oxford  
Oxford, OX3 7LF, UK  
Tel: 44-1865-743949

or

Professor Zhengming Chen  
CTSU, Big Data Institute,  
Old Road Campus  
University of Oxford  
Oxford, OX3 7LF, UK  
Tel: 44-1865-743839

[derrick.bennett@ndph.ox.ac.uk](mailto:derrick.bennett@ndph.ox.ac.uk)

[zhengming.chen@ctsu.ox.ac.uk](mailto:zhengming.chen@ctsu.ox.ac.uk)

**Word count:** Abstract 250/250, Text 4389

(1 Table, 4 Figures and 20 Supplementary Tables and Figures)  
(CKB research tracker No.: 2023-0065; CKB data release 19.00)

**23 October 2024**

39 **Abbreviations:**

AKR1D1	3-oxo-5-beta-steroid 4-dehydrogenase
ALT	Alanine aminotransferase
ANML	Adaptive normalisation by maximum likelihood
ANTR2	Anthrax toxin receptor 2
APOF	Apolipoprotein F
BGN	Biglycan
BMI	Body mass index
BPSA	Benign prostatic-specific antigen
C1QTNF1	C1q and tumor necrosis factor related protein 1
C1QTNF3	C1q and tumor necrosis factor related protein 3
CFHR5	Complement factor H-related 5
CILP2	Cartilage intermediate layer protein 2
CKB	China Kadoorie Biobank
CO	Carbon monoxide
CRDL1	Chordin-like protein 1
CRDL2	Chordin-like protein 2
CRP	C-reactive protein
DBP	Vitamin D-binding protein
DKK2	Dickkopf-related protein 2
DJB12	DnaJ homolog subfamily B member 12
ESPN	Espin
FABP	Fatty acid binding protein
FABPA	Fatty acid binding protein, adipocyte
FIX	Coagulation factor IX
FIXab	Coagulation factor IXab
FSH	Follicle stimulating hormone
GHR	Growth hormone receptor
GPDA	Glycerol-3-phosphate dehydrogenase [NAD(+)], cytoplasmic
HBD-4	Hemoglobin subunit delta-4
HBsAg	Hepatitis B surface antigen
HBV	Hepatitis B virus
HCG	Human chorionic gonadotropin
HTRA1	Htra Serine Peptidase 1
H6ST2	Heparan-sulfate 6-O-sulfotransferase 2
DHI1	Corticosteroid 11-beta-dehydrogenase isozyme 1
IGDC4	Immunoglobulin superfamily, dcc subgroup member 4
IGFALS	Insulin-like growth factor binding protein, acid labile subunit
IGFBP-2	Insulin-like growth factor binding protein 2
IGFBP-3	Insulin-like growth factor binding protein 3
IHD	Ischemic heart disease
IL-1 R AcP	Interleukin-1 Receptor accessory protein
INHBC	Inhibin beta C chain
LH	Luteinizing hormone
LOD	Limit of detection
LPL	Lipoprotein lipase
LRP1	Ldl receptor related protein 1
MIC-1	Macrophage inhibitory cytokine 1
MMP7	Matrix metalloproteinase 7

MXRA8	Matrix remodeling associated 8
MXRA8:ECD	Matrix remodeling associated 8, extracellular domain
NCAM1	Neural cell adhesion molecule 1
NCAM-120	Neural Cell Adhesion Molecule 120 kda Isoform
NFASC	Neurofascin
NUD16	U8 snorna-decapping enzyme
PLXB2	Plexin B2
PSA	Prostate-specific antigen
PTN	Pleiotrophin
PZP	Pregnancy zone protein
QC	Quality control
RBP	Retinol-binding protein 4
RFU	Relative fluorescence units
RIDA	Rectal intestinal domain antigen
RPG	Random plasma glucose
SAP	Serum amyloid P-component
SBP	Systolic blood pressure
SCG3	Secretogranin-3
SEMA6A	Semaphorin-6A
SEM4D	Semaphorin-4D
SEM6B	Semaphorin-6B
SHBG	Sex hormone-binding globulin
sICAM-5	Soluble intercellular adhesion molecule 5
SOMAmer	Slow off-rate modified aptamers
SPINK6	Serine peptidase inhibitor, kazal type 6
TBG	Thyroxine-binding globulin
TINAL	Tubulointerstitial nephritis antigen-like 1
TTR	Transthyretin
UGDH	Udp-glucose 6-dehydrogenase
WFKN2	WAP, Kazal, immunoglobulin, Kunitz and NTR domain-containing protein 2
$\alpha$ 2AP	A2-antiplasmin

41 **Abstract** (word count: 250/250)

42 **Background:** Proteomics offer new insights into human biology and disease aetiology.  
43 Previous studies have explored the associations of SomaScan proteins with multiple non-  
44 genetic factors, but they typically involved Europeans and a limited range of factors, with no  
45 evidence from East Asia populations.

46 **Methods:** We measured plasma levels of 6,597 unique human proteins using SomaScan  
47 platform in ~2,000 participants in the China Kadoorie Biobank. Linear regression was used  
48 to examine the cross-sectional associations of 37 exposures across several different  
49 domains (e.g., socio-demographic, lifestyle, environmental, sample processing,  
50 reproductive factors, clinical measurements and frailty indices) with plasma concentrations  
51 of specific proteins, adjusting for potential confounders and multiple testing.

52 **Findings:** Overall 12 exposures were significantly associated with levels of >50 proteins,  
53 with sex (n=996), age (n=982), ambient temperature (n=802) and BMI (n=1035) showing  
54 the largest number of associations, followed by frailty indices (n=465) and clinical  
55 measurements (e.g., RPG, SBP), but not diet and physical activity which showed little  
56 associations. Many of these associations varied by sex, with a large number of age-related  
57 proteins in females also associated with menopausal status. Of the 6,597 proteins examined,  
58 43% were associated with at least one exposure, with the proportion higher for high-  
59 abundance proteins, but certain biologically-important low-abundance proteins (e.g., PSA,  
60 HBD-4) were also associated with multiple exposures. The patterns of associations  
61 appeared generally similar to those with Olink proteins.

62 **Interpretation:** In Chinese adults an exposome-wide assessment of SomaScan proteins  
63 identified a large number of associations with exposures and health-related factors,  
64 informing future research and analytic strategies.

65 **Key words:** *Exposome, Sex, Age, Frailty, Lifestyle, Protein biomarkers, Biobank, Chinese*

66 **Funding:** Wellcome Trust, UK Medical Research Council, Cancer Research UK, British  
67 Heart Foundation, National Natural Science Foundation of China, National Key Research  
68 and Development Program of China, Kadoorie Charitable Foundation, Novo Nordisk, Olink,  
69 Somalogic.

## 70 **Introduction**

71 Proteins play essential roles in all living cells, and an optimal balance of protein levels  
72 influence human health. Previous studies of individual plasma proteins have identified many  
73 easily measured biomarkers relevant for disease diagnosis and understanding of disease  
74 mechanisms, including troponin reflecting cardiac injury,<sup>1</sup> alanine transaminase (ALT)  
75 reflecting liver damage,<sup>2</sup> and C-reactive protein (CRP) reflecting systemic inflammation.<sup>3</sup>  
76 Recent advances in high-throughput proteomic assays now enable measurements of  
77 thousands of plasma proteins with high levels of accuracy.<sup>4</sup> This permits a more  
78 comprehensive investigation on the molecular mechanisms underlying disease aetiology.<sup>4</sup>

79 An estimated 70-90% of disease risk in adults could be attributed to non-genetic risk factors,  
80 collectively known as the “exposome”,<sup>5</sup> which act through multiple complex biological  
81 pathways in disease pathogenesis, with plasma proteins being likely intermediate markers  
82 of exposure or adverse effects. Recent population-based proteomics studies have assessed  
83 the associations of proteins, assayed using Olink or SomaScan platforms, with several well-  
84 established risk factors (e.g., adiposity, ageing, and smoking) and their associated biological  
85 processes.<sup>6-10</sup> However, most previous studies were conducted in European populations,<sup>7-  
86 10</sup> and did not provide systematic evaluations of impact of a wider spectrum of non-genetic  
87 factors on the plasma proteome.<sup>11,12</sup> An “exposome-wide” study of the correlates of plasma  
88 protein levels in diverse populations can inform future research priorities and guide analytical  
89 approaches (e.g., adjustment for potential confounders when studying specific exposures).

90 We undertook comprehensive assessment of the associations of >7000 SomaScan proteins  
91 with 37 major non-genetic risk factors in ~2000 Chinese adults in the China Kadoorie  
92 Biobank (CKB). The main objectives of the present study were to (1) systematically explore  
93 the exposure profiles of ~7000 SomaScan proteins in 2000 Chinese adults; (2) assess the  
94 profiles in relation to normalised and non-normalised SomaScan protein levels; and (3)

95 compare the findings with a parallel investigation<sup>13</sup> involving ~3000 proteins measured using  
96 an antibody-based Olink proteomics platform.

## 97 **Methods**

### 98 *Study population and design*

99 Details of the study design and participants characteristics of CKB have been reported  
100 elsewhere.<sup>11</sup> Briefly, CKB recruited ~512,000 adults aged 30-79 years from 10  
101 geographically diverse areas in 2004-2008. At baseline, trained health workers administered  
102 a comprehensive laptop-based questionnaire and recorded physical measurements,  
103 including anthropometry and blood pressure, using regularly-calibrated instruments  
104 following standardised protocols. Desktop biochemical assays included random plasma  
105 glucose (RPG) and hepatitis B surface antigen (HBsAg). A 10-mL non-fasting (with time  
106 since the last meal recorded) blood sample was collected from each participant, and then  
107 processed and stored in liquid nitrogen. The present study involved 2,026 randomly selected  
108 subcohort participants who had no prior history of cardiovascular disease, originally sampled  
109 along with 1,951 cases of incident ischaemic heart disease (IHD) for a case-cohort study.<sup>6,14</sup>

110 The CKB was approved by the Ethical Review Committee of the Chinese Center for Disease  
111 Control and Prevention (Beijing, China) and the Oxford Tropical Research Ethics Committee,  
112 University of Oxford (Oxford, UK), and all participants provided written informed consent  
113 upon recruitment.

### 114 *Proteomic assays*

115 Details of the proteomic assays in the CKB have been described elsewhere.<sup>6,14,15</sup> In brief,  
116 60µl plasma aliquots in 2D-barcode microtubes of 3,977 participants were delivered to the  
117 Somalogic Laboratory in Colorado, USA for profiling using SomaScan Assay v4.1, which  
118 covers 7,596 slow off-rate modified aptamers (SOMAmers) as protein-binding reagents, with

119 7,289 SOMAmers targeting 6,597 human proteins. Samples were randomly aliquoted into  
120 96-well plates (including 11 wells for external control samples, including 5 calibrator, 3 QC,  
121 and 3 buffer samples). The raw output of the SomaScan assay were standardised based on  
122 external control samples to account for variability in microarrays and variations within and  
123 between plates. With an optional procedure of adaptive normalisation by maximum  
124 likelihood (ANML) to an external reference, which controls for inter-sample variability, the  
125 results were provided in normalised and non-normalised values in relative fluorescence units  
126 (RFU).<sup>16</sup> The output values were further natural log-transformed in the main analysis. The  
127 limit of detection (LOD) for SOMAmers were defined with external buffer samples. Quality  
128 control (QC) checks were conducted comparing the median of QC samples on each plate  
129 to the reference, with a cross-plate QC indicator (pass/flag) assigned to each SOMAmer.  
130 The 7,289 SOMAmers targeting human proteins were mapped to proteins based on their  
131 UniProt IDs supplied by Somalogic. Details of individual proteins and their distributions by  
132 sex are shown in **eTable 1**.

### 133 *Selected baseline characteristics*

134 From the baseline questionnaire and physical measurements, we identified 37 key  
135 characteristics from six broad categories: demographics (e.g., age, sex, study area), lifestyle  
136 (e.g., alcohol, smoking, diet), environmental factors (e.g., air pollution, ambient temperature),  
137 health and wellbeing (e.g., medical history and mental health), clinical measurements (e.g.,  
138 BMI, HBV, RPG) and female reproductive factors (e.g., age at menarche, age at menopause,  
139 parity) (**eTable 2**). We also derived composite indices for (i) healthy lifestyle (ranging from  
140 0 to 5, with higher score indicating a healthier lifestyle), based on our previous publications,  
141 which takes into account smoking, alcohol intake, physical activity, dietary habits, and body  
142 shape;<sup>17-19</sup> and (ii) frailty, based on 28 variables on self-reported medical conditions,  
143 symptoms, signs, and physical measurements that capture different aspects of cumulative  
144 health status deficits as described previously<sup>20</sup> (see derivation details in **eTable 2**).

145 *Statistical analysis*

146 The present study used a comparable analytical strategy to that used for Olink proteomics  
147 in a parallel paper.<sup>13</sup> From the original subcohort of 2,026 participants, we excluded  
148 individuals with missing SomaScan proteomics data (n=4), whose blood sample with  
149 hybridization control scale factor out of range (n=7), and those with missing ambient  
150 temperature data (n=20), leaving 1,998 sub-cohort participants (1243 females, 755 males)  
151 for the main overall and sex-specific analyses.

152 Selected baseline characteristics were examined by sex, directly standardised according to  
153 age and study areas of the original CKB population structure to facilitate comparison.  
154 Multivariable linear regression was used to examine the associations of the 37 baseline  
155 characteristics with normalized plasma protein levels (in RFU), adjusting for age, age<sup>2</sup>, sex,  
156 study area, fasting time, fasting time<sup>2</sup>, outdoor temperature, outdoor temperature<sup>2</sup> and plate  
157 ID. For proteins targeted by multiple SOMAmers, the association showing the smallest p-  
158 values per exposure was selected. The analyses were repeated on non-normalized protein  
159 levels to evaluate the impact of ANML, which might have attenuated meaningful inter-  
160 individual heterogeneity. We also examined the associations (with normalised protein levels)  
161 by three classes of protein abundance, defined according to the SomaScan-supplied dilution  
162 factor for each protein, as low- ( $2 \times 10^{-1}$ ), moderate- ( $5 \times 10^{-3}$ ), and high- ( $5 \times 10^{-5}$ ) abundance.

163 Given the high correlation between many SOMAmers (or protein levels), we adopted the  
164 multiple testing correction approach described by Gadd et al.<sup>21</sup> A Bonferroni-adjusted p-  
165 value threshold was applied, based on 1,186 principal components explaining 90% of the  
166 cumulative variance in 7,335 SOMAmers (**eFigure 1** and **eTable 3**), and 21 components  
167 explaining 90% of the cumulative variance in the 32 exposures tested, plus 5 additional  
168 exposures for reproductive factors.<sup>21</sup> This adjustment was applied across all linear  
169 regression models, with a Bonferroni-adjusted p-value threshold



170 at  $0.05/(1186 \times 26) = 1.62 \times 10^{-6}$ . All statistical analyses were performed using R version  
171 4.1.2<sup>22</sup> and packages 'tidyverse' and 'ggplot2'.

## 172 *Role of the funding source*

173 The funders of the study had no role in study design, data collection, data analysis, data  
174 interpretation, or writing of the report.

## 175 **Results**

176 **Table 1** shows the distribution of the 37 baseline characteristics among the 1998  
177 participants included, overall and by sex. The mean age at baseline was 50.8 (SD 10.5)  
178 years and 62.2% were females. Males had higher prevalence of alcohol drinking (37.2% vs  
179 2.6%) and smoking (63.3% vs 2.3%) than females, but similar prevalence of self-reported  
180 poor health and prior diseases, and slightly lower healthy lifestyle index.

181 Overall, 29 exposures were significantly associated with at least one protein after  
182 Bonferroni-adjustment, with 12 showing significant associations with >50 proteins (**Table 1**;  
183 **Figure 1**). In particular, sex (n=996), age (n=982), outdoor temperature (n=802), and BMI  
184 (n=1035) had the largest numbers of significant associations, followed by several clinical  
185 measurements (e.g. SBP and RPG), health and wellbeing indicators (e.g., HBsAg status  
186 and diabetes) and the lifestyle and frailty indices (**Table 1**; **Figure 1**). In sex-specific  
187 analyses, there were more significant associations in females than in males, except for  
188 smoking ( $n_{\text{female}}=5$ ;  $n_{\text{male}}=24$ ) and alcohol consumption ( $n_{\text{female}}=2$ ;  $n_{\text{male}}=85$ ), with three of the  
189 five female reproductive factors showing significant associations with 1~58 proteins (**Table**  
190 **1**). Across the 6,597 unique human proteins from the 7,289 SOMAmers examined, 43%  
191 were associated with at least one exposure, with IGFBP-2 (n=14), BGN (n=13), FIX (n=13),  
192 FIXab (n=13), and HSP 70 (n=13) associated with the greatest number of exposures, mostly

193 with clinical measurements and demographic factors, overall and in sex-specific analyses  
194 (**Figure 2; eFigure 2**).

195 Of the 996 proteins associated with sex, the most significant associations included higher  
196 levels of leptin, FSH, and PZP in females and higher levels of PSA, HBD-4, and BPSA in  
197 males (**Figure 3ia**). Among the top 50 sex-related proteins, most were also associated with  
198 other exposures, particularly age (e.g., FSH, HCG) and BMI (e.g., leptin, FABP) (**Figure**  
199 **3ib**). Importantly, there were 180 proteins uniquely associated with sex, but not with any  
200 other exposures (**eTable 4**). Furthermore, apart from alcohol drinking and smoking where  
201 the exposure-protein associations were stronger in females, most other associations  
202 appeared either comparable or slightly stronger in males (**eFigure 3**).

203 As for the 982 proteins associated with age, DKK2, FSH, PTN, MIC-1, and CDCP1 showed  
204 the most statistically significant positive associations, whereas  $\alpha$ 2AP, IGFALS, IGDC4,  
205 CILP2, and SET showed the most significant inverse associations (**Figure 3iia**). Of the top  
206 50 (by statistical significance) age-related proteins, most were also associated with other  
207 exposures, particularly sex (e.g., FSH, PTN), BMI (e.g., CRDL1, FABPA), and urban  
208 residence (e.g., DKK2, MIC-1) (**Figure 3iib**). In sex-specific analyses, age was associated  
209 with 735 and 593 proteins in females and males, respectively (**Table 1**), with most of the  
210 357 overlapping proteins showing concordant associations (**eFigure 4**). Among the top age-  
211 associated proteins by sex, many were also associated with post-menopausal status, BMI,  
212 and urban residence in females, and with BMI and urban residence in males (**eFigure 5**).  
213 Consistently, the majority of the 58 proteins associated with post-menopausal status were  
214 also associated with age (e.g., FSH, LH, CRDL2, HCG), with some also being associated  
215 with BMI (e.g., FABPA, FABP) (**eFigure 6**).

216 HBsAg status and prevalent diabetes were associated with 468 and 257 proteins,  
217 respectively, but other health and wellbeing indicators, including self-rated health or other

218 prior medical history showed few significant associations (**Table 1**). For HBsAg status, the  
219 top positively associated proteins were DHI1, NUD16, DJB12, C1QTNF3, and AKR1D1,  
220 while the top inversely associated proteins were CFHR5, SAP, RBP, IL-1 R AcP, and  $\alpha$ 2AP  
221 (**eFigure 7**). For prevalent diabetes, the top positively associated proteins included PLXB2  
222 and SEMA6A and top inversely associated proteins were CILP2 and MXRA8:ECD (**eFigure**  
223 **7**).

224 Several clinical measurements, including BMI, SBP, DBP, heart rate, and RPG (but not  
225 height, exhaled CO or lung function) were strongly associated with multiple proteins (**Table**  
226 **1**). BMI had the highest number of associations (n=1035) among all exposures examined,  
227 with leptin, GHR, FABP, FABPA, and GPDA being the top positively associated proteins  
228 and IGFBP-2, IGFBP-1, WFKN2, SHBG, and SEZ6L being the top inversely associated  
229 proteins (**eFigure 8**). In the sex-specific analyses, leptin, GHR, FABP, FABPA, IGFBP-2,  
230 WFKN2, and SHBG were also top hits with BMI, with the same direction of association in  
231 males and females (**eFigure 9**). The proteins that had the most significant positive  
232 associations with SBP were GHR, INHBC, FIXab, FIX, and leptin, while those with the most  
233 significant inverse associations were renin, IGFBP-2, SHBG, H6ST2, and SCG3 (**eFigure**  
234 **7**). Among the top RPG-associated proteins, there were positive associations with PLXB2,  
235 SEMA6A, SEM4D, SEM6B, and NFASC and inverse associations with CILP2, MXRA8:ECD,  
236 COL15A1, SCG3, and ALB (**eFigure 7**).

237 Among the 307 proteins associated with healthy lifestyle index (**Table 1**), the most significant  
238 positive associations included TINAL, NCAM1, and NCAM-120, and the inverse  
239 associations included leptin, GPDA, and UGDH (**Figure 4ia**). The index-protein  
240 associations, as illustrated in the most strongly associated proteins, appeared to overlap  
241 with those with BMI and sex, followed by clinical measurements (e.g., SBP, DBP, RPG),  
242 and urban residence (**Figure 4ib**). The composite frailty index was associated with 465  
243 proteins overall, and 226 and 82 in females and males, respectively (**Table 1; Figure 4iia**).

244 Among the leading frailty-associated proteins, there were positive associations with CRP,  
245 ESPN, and HTRA1 and inverse associations with MXRA8:ECD, ANTR2, and SHBG (**Figure**  
246 **4iia**). Importantly, the proteins most strongly associated with frailty index overlapped with  
247 most clinical measurements (particularly BMI), age, sex, and lifestyle index, in a largely  
248 coherent manner (i.e., opposing direction of association) as expected (**Figure 4iib**).

249 Notably, we observed a general trend of higher proportion of significant associations with  
250 proteins of higher abundance (e.g., 41% proteins of high-abundance vs. 10% of low-  
251 abundance were associated with age), except for outdoor temperature (~10% for both high-  
252 and low-abundance proteins) (**eTable 5; eFigure 10**). Analysis of non-normalised protein  
253 levels showed greater number (by 1.5 to 3 times) of significant associations with age,  
254 diabetes, BMI, SBP, DBP, heart rate, RPG, menopausal status, and lifestyle and frailty  
255 indices, but considerably fewer (15% lower) significant hits with outdoor temperature  
256 (**eTable 6; eFigure 11**). Generally, the proportional increment of significant associations in  
257 the non-normalised proteins were more prominent in females than males, except for BMI  
258 and frailty index (**eTable 6**). Importantly, the effect sizes of associations related to non-  
259 normalised protein levels were smaller than for normalised levels for age, BMI, SBP, RPG,  
260 prevalent diabetes, and frailty index, but the converse was true for lifestyle index, whereas  
261 the associations with sex were highly consistent (**eFigure 12**). Across the two parallel  
262 analyses, both SomaScan and Olink platforms yielded comparable numbers of significant  
263 proteomic associations with the multiple exposures, particularly with sex, age, alcohol  
264 drinking, smoking, HBsAg status, diabetes, clinical traits, menopausal status, healthy  
265 lifestyle and frailty indices (**eFigure 13**).

## 266 **Discussion**

267 This exposome-wide analyses of almost 6,600 SomaScan proteins in ~2,000 Chinese adults  
268 demonstrated significant associations of several major non-genetic risk factors with plasma

269 levels of specific proteins. Exposures such as age, sex, ambient temperature, and BMI  
270 demonstrated the largest number of significant associations with multiple proteins, with  
271 many of these associations varying by sex. Other exposures including demographic factors,  
272 lifestyle habits, clinical measurements, and lifestyle and frailty indices were also related to a  
273 very large number of proteins. Overall, there was a larger proportion of significant  
274 associations with high-abundance proteins, but we still found associations with low-  
275 abundance but biological-important proteins not captured in the Olink immunoassays (e.g.,  
276 PSA, HBD-4, DKK2).

277 The two parallel analyses across the SomaScan and Olink<sup>13</sup> platforms yielded largely  
278 consistent findings, especially on factors associated with large number of proteins. Notably,  
279 the patterns of associations with Olink proteins were more similar to those for the non-  
280 normalised SomaScan proteins, consistent with previous investigations.<sup>15</sup> Moreover, both  
281 platforms showed directionally consistent associations, as demonstrated for instance by  
282 leptin (higher in females in both platforms) and IGFBP-2 (inversely associated with frailty  
283 index in both platforms). Furthermore, a number of proteins that are included in the  
284 SomaScan platform but not in the Olink platform were significantly associated with many  
285 exposures (e.g. age positively associated with DKK2, and frailty index inversely associated  
286 with ANTR2), demonstrating the complementary advantages of the two proteomic platforms.

287 Previous studies have investigated the SomaScan plasma proteomic profiles of various  
288 exposures.<sup>23-28</sup> However, they included primarily Western populations, used the earlier  
289 versions of the SomaScan platforms with fewer proteins measured, and typically focused on  
290 single or a small group of individual exposures.<sup>23-27</sup> The only recent broad-spectrum study  
291 that used an earlier SomaScan (v4.0) platform to explore the genetic and non-genetic  
292 predictors of 4,775 plasma proteins measured in ~8000 European adults, showed that non-  
293 modifiable factors, including genetic factors, age, and sex were the major determinants of  
294 plasma proteins (of 3,242 protein targets), which is somewhat in line with our findings (on

295 age and sex).<sup>28</sup> However, age in this European study was associated with a higher  
296 proportion of proteins than in our study,<sup>28</sup> potentially because of the larger sample size. A  
297 few studies also used the earlier Olink platform assays to investigate the exposure profiles  
298 of ~90 proteins.<sup>29-31</sup> Our study is one of the first and largest proteomic-exposome profiles  
299 studies in China and East Asia that systematically evaluates the impact of a much wider  
300 spectrum of exposures on the plasma proteome. Our analyses also identified important  
301 differences in proteomic-exposome profiles between males and females, including many  
302 proteins known to be involved in sex-specific biological processes. For example, FSH that  
303 stimulates the growth and development of follicles in females,<sup>32</sup> was higher in females than  
304 males, consistent with findings in the parallel CKB Olink proteomics study.<sup>13</sup> Furthermore,  
305 levels of PZP, an immunosuppressive protein expressed by the placenta,<sup>33</sup> were higher in  
306 females than in males, while the levels of PSA/BPSA and HBD-4 , which are expressed in  
307 prostate <sup>34</sup> and testes,<sup>35</sup> respectively, were higher in males. We also found leptin, which is  
308 known to play a key role in regulating energy balance and controlling body weight,<sup>36</sup> to be  
309 significantly higher in females, reflecting the sex-specific variations in body composition  
310 measurements,<sup>37</sup> as observed in the parallel CKB Olink proteomics study.<sup>13</sup> Interestingly,  
311 adjustment for sex in our analyses did not attenuate the associations of the sex-related  
312 proteins with other exposures, possibly reflecting other sex-independent effects of the  
313 proteins. For example, FSH was also associated with age and leptin was associated with  
314 BMI, independent of sex. Therefore, these proteins might also be involved in biological  
315 processes (metabolic, inflammatory, or ageing processes) that could be influenced by  
316 exposures like age and BMI,<sup>38</sup> beyond their associations with sex. Genetic regulation of the  
317 plasma proteome has been previously reported to be largely similar among the two sexes,<sup>39</sup>  
318 even though levels of plasma proteins were largely different between men and women,  
319 suggesting that these differences might be related mainly to differential non-genetic factor

320 between the sexes,<sup>40</sup> and highlighting the importance of performing sex-specific and sex-  
321 adjusted analyses in observational proteomic studies.

322 Importantly, some proteomic associations with individual lifestyle factors differed by sex. For  
323 example, TBG and sICAM-5, which have been previously associated with alcohol intake and  
324 smoking,<sup>23,41</sup> respectively, were more strongly related to the corresponding exposures in  
325 males than in females in our study. These observed sex-differences likely reflect the  
326 markedly higher prevalence (and intensity) of alcohol drinking and smoking in men than in  
327 women in the Chinese population,<sup>42,43</sup> which is also consistent with the higher number of  
328 significant hits with alcohol drinking and smoking in males. Interestingly, there were limited  
329 proteomic associations with the other individual lifestyle factors, most notably diet and  
330 physical activity, which is consistent with the previous SomaScan proteomics-exposome  
331 study in Europeans.<sup>28</sup> Lifestyle factors could influence the proteome dynamically and these  
332 dynamic protein changes are challenging to be captured fully by cross-sectional snapshots  
333 of protein levels.<sup>44</sup>

334 Age was strongly and positively associated with DKK2, which is a glycoprotein known to  
335 regulate vertebrate development and to modulate the Wnt signalling pathway.<sup>45</sup> DKK2 has  
336 been found to be involved in the development of several ageing-related diseases, including  
337 cancer.<sup>46-48</sup> We also found older age to be strongly associated with higher levels of PTN, a  
338 secreted growth factor that regulates cellular proliferation, growth, differentiation and  
339 migration.<sup>49</sup> Previous studies also reported higher levels of plasma PTN being significantly  
340 associated with chronological age.<sup>24,50</sup> Additionally, the most strongly age-associated  
341 protein was  $\alpha$ 2AP, a serine protease inhibitor responsible for inactivating plasmin.  
342 Circulating plasmin is released in the plasma during fibrinolysis, which is an important  
343 regulator of blood coagulation process, and congenital deficiency of  $\alpha$ 2AP causes a rare  
344 bleeding disorder due to increased fibrinolysis,<sup>51</sup> and an increased bleeding tendency with  
345 age in patients with heterozygous deficiency.<sup>52</sup> We also found that many other age-related

346 proteins were also independently associated with other risk factors. For example, CRDL1,  
347 a bone morphogenetic protein-4 antagonist, was associated with BMI, SBP and DBP.  
348 Moreover, in consistent with Olink proteins, we found numerous sex-specific top protein hits  
349 for age to be also associated with postmenopausal status, including FSH, which is known  
350 to be higher in menopausal women.<sup>53</sup>

351 We found a large number of proteins associated with several clinical as well as health and  
352 wellbeing-related traits. BMI was associated with the largest number of significant  
353 associations with proteins, which aligns with the relatively high protein variance attributable  
354 to measures of body composition reported previously in Europeans.<sup>28</sup> A number of these  
355 associations were previously evident in both observational and genetic analyses, such as  
356 the positive association with leptin and FABP, which is involved in the uptake, metabolism  
357 and transport of long-chain fatty acids.<sup>6,26</sup> Future studies should also consider assessing the  
358 associations of SomaScan proteins with central adiposity (e.g. waist circumference, waist-  
359 to-hip ratio), as it is known to be more strongly associated with risk of cardio-metabolic  
360 disease<sup>54,55</sup> and with Olink proteins,<sup>56</sup> compared with BMI. Prevalent diabetes and RPG  
361 were both positively associated with PLXB2 (receptor for semaphorins involved in cell-cell  
362 signalling) and inversely associated with CILP2 (involved in cartilage scaffolding), which are  
363 supported by prior observational <sup>57,58</sup> and genetic analyses.<sup>27,59</sup> In addition to many new  
364 findings that need to be explored further in subsequent studies, we also identified some  
365 previously reported associations, such as SBP with renin (enzyme involved in the renin-  
366 angiotensin-aldosterone system)<sup>60</sup> in both SomaScan and Olink studies, and HBV infection  
367 with SAP (part of the innate immune humoral arm)<sup>61</sup>.

368 The present study demonstrated a large number of proteomic associations with both the  
369 frailty <sup>20</sup> and healthy lifestyle indices.<sup>17,18</sup> The two indices capture essentially opposing  
370 dimensions of health status, and many proteins showed opposing associations across the  
371 two indices, including HTRA1, MXRA8:ECD, ANTR2, leptin, GPDA, UGDH, IGFBP-2, and



372 SHBG. HTRA1 is implicated in inhibiting active TGF- $\beta$ , which is an anti-inflammatory  
373 cytokine,<sup>62</sup> while MXRA8:ECD and ANTR2 are both involved in angiogenesis.<sup>25</sup> GPDA was  
374 previously shown to be positively associated with current smoking and alcohol drinking in  
375 the Framingham Heart Study,<sup>23</sup> which is line with our finding of an inverse association of  
376 lifestyle index with levels of this protein. Both GPDA and UGDH have also been implicated  
377 in multiple cancer types, and considered as possible therapeutic targets for treatment of  
378 cancer.<sup>63,64</sup> In addition, consistent with our findings related to the two indices, circulating  
379 IGFBP-2 and SHBG were reported to be inversely associated with obesity,<sup>65,66</sup> while a  
380 previous study showed lower circulating levels of IGFBP-3 (same protein family with IGFBP-  
381 2) to be associated with current smoking.<sup>67</sup> Both IGFBP-2 and SHBG are involved in cancer  
382 and metabolic/reproductive system disorders, respectively, and they have also been  
383 identified in the Olink CKB study. Many indices-related proteins were also associated with  
384 individual exposures, particularly BMI, other clinical measurements, sex, and age, which is  
385 an expected finding as a number of these exposures are constituents of the indices.

386 Among the many exposures examined previously, the impact of outdoor temperature on the  
387 plasma proteome has been understudied.<sup>29,68</sup> We found a large number novel significant  
388 associations that warrant further investigation. For example, SPINK6, which showed a  
389 strong positive association with ambient temperature, is a potent inhibitor of serine  
390 proteases that are essential for influenza A viruses infection in the airways.<sup>69</sup> Similarly, LRP1  
391 and MMP7, which are inversely associated with temperature, play important roles in lipid  
392 metabolism.<sup>70,71</sup> However, we found fewer significant proteomic associations of proteins with  
393 several other exposures, including household air pollution, prevalent diseases (except  
394 diabetes), mental health, height, lung function and most reproductive factors. Future  
395 proteomic studies with a larger sample size and increased proteome coverage might be  
396 needed to further clarify the relevance of the observed associations and to identify new  
397 associations.<sup>44</sup>

398 For reasons that are not fully understood, we found that use of non-normalised proteomics  
399 panels yielded an even larger number of significant associations with several major  
400 exposures (e.g., age, diabetes, and frailty and healthy lifestyle indices). This is consistent  
401 with the previous studies conducted in CKB<sup>15</sup> and with some previous studies using different  
402 approaches to proteomic profiling.<sup>72-74</sup> The normalization procedure adopted by SomaLogic  
403 standardises the overall signal of each sample to an external reference to improve data  
404 quality.<sup>75</sup> However, this process may reduce true extreme signals that are present in the  
405 general population, attenuating the number of the observed significant associations.<sup>72,76</sup>  
406 Interestingly, while most associations with the two types of data were directionally consistent,  
407 the effect sizes of the overlapping associations were smaller for non-normalised than for  
408 normalised protein levels.

409 The present study is the first and largest study in East Asia that measured the levels of 6,600  
410 plasma proteins using the SomaScan platform and conducted a comprehensive  
411 investigation of the proteomic associations with >35 exposures across different domains.  
412 This type of “exposome-wide” investigation, in parallel with that on 2923 Olink proteins,<sup>13</sup> is  
413 also important to inform future analytical approaches (e.g. confounding adjustment) when  
414 examining specific exposures. Moreover, the SomaScan platform used in CKB captured a  
415 large number of low-abundance proteins not available in the Olink panels, which may be  
416 equally important for understanding biological pathways. Comprehensive analysis of both  
417 high- and low-abundance proteins contributes to a more complete map of biological  
418 processes, providing insights into disease pathogenesis and progression. However, the  
419 current study also had several limitations. First, there were more females than males, which  
420 might bias towards finding a larger number of significant associations in females. Moreover,  
421 there was limited statistical power for performing further subgroup analyses, besides the  
422 sex-specific analyses. Second, the present study was restricted to proteins measured only  
423 in plasma and to proteins that cover a modest proportion of the human proteome. Third,

424 replication of our findings in external and independent cohorts was not possible, as currently  
425 there are limited available SomaScan proteomic data in other East Asian populations.  
426 Nevertheless, our findings are broadly consistent with the parallel analyses on Olink  
427 proteomics which included 2,168 proteins also targeted by SomaScan.<sup>13</sup> Fourth, the  
428 direction of the associations studied herein could not be confirmed due to the cross-sectional  
429 nature of the study. Finally, although the analyses were adjusted for a range of key  
430 covariates, such as age, sex, region, and fasting time, residual confounding cannot be fully  
431 excluded. Further genetic investigations could allow us to identify potential causal  
432 associations between various exposures and proteins.

433 Overall, this large proteomic study in the Chinese population showed that several exposures,  
434 particularly, age, sex, ambient temperature, BMI and composite indices of healthy lifestyle  
435 and frailty were associated with a large number of proteins that are involved in multiple  
436 pathways. The associations between major exposures and proteins also varied by sex,  
437 potentially due to sex-specific biology and sex-differential lifestyles. The findings of the  
438 present study may inform priorities for future proteomics research and further studies are  
439 warranted to replicate the current findings in independent populations and to provide  
440 information on the causal relevance of these associations.

441

442 **Acknowledgements:** The chief acknowledgment is to the participants, China Kadoorie  
443 Biobank project staff, staff of the China CDC and its regional offices for access to death and  
444 disease registries. The Chinese National Health Insurance scheme provided electronic  
445 linkage to all hospital admissions. The China Kadoorie Biobank study is jointly coordinated  
446 by the University of Oxford and the Chinese Academy of Medical Sciences.

447 **Funding:** The funding body for the baseline survey was the Kadoorie Charitable Foundation,  
448 Hong Kong, China and the funding sources for the long-term continuation of the study  
449 include UK Wellcome Trust (202922/Z/16/Z, 104085/Z/14/Z, 088158/Z/09/Z), Chinese  
450 National Natural Science Foundation (81390540, 81390541, 81390544), and the National  
451 Key Research and Development Program of China (2016YFC0900500, 2016YFC0900501,  
452 2016YFC0900504, 2016YFC1303904). Core funding was provided to the CTSU, University  
453 of Oxford, by the British Heart Foundation, the UK Medical Research Council, and Cancer  
454 Research UK. The long-term follow-up was funded in part by the UK Wellcome Trust  
455 (212946/Z/18/Z, 202922/Z/16/Z, 104085/Z/14/Z, 088158/Z/09/Z). The proteomic assays  
456 were supported by BHF (FS/18/23/33512), Novo Nordisk, Olink, SomaLogic and NDPH.

457 **Open Access Statement:** This research was funded in whole, or in part, by the Wellcome  
458 Trust [212946/Z/18/Z, 202922/Z/16/Z, 104085/Z/14/Z, 088158/Z/09/Z]. For the purpose of  
459 Open Access, the author has applied a CC-BY public copyright license to any Author  
460 Accepted Manuscript version arising from this submission.

461 **Declaration of interests:** All authors declare no competing interests.

462 **Ethics approval:** The China Kadoorie Biobank (CKB) complies with all the required ethical  
463 standards for medical research on human subjects. Ethical approvals were granted and  
464 have been maintained by the relevant institutional ethical research committees in the UK  
465 and China.

466 **Consent to participate/publication:** All participants provided written informed consent.

467 **Author Contributions:** KHC, JC, MK, DAB, and ZC conceived and designed the study. JC  
468 conducted the statistical analyses and KHC and MK wrote the first draft of the manuscript.LL,  
469 and ZC as the members of CKB Steering Committee, designed and supervised the overall  
470 conduct of the study, including obtaining funding for the study. All other authors provided  
471 critical revision to the manuscript for important intellectual content. KHC, JC, MK, DAB and  
472 ZC are the guarantors of this work and take responsibility for the integrity and accuracy of  
473 the data analysis. DAB and ZC supervised the work. All authors contributed to the review  
474 and edit of the manuscript.

475 **Data Availability:** Data from baseline, first and second resurveys, and disease follow-up  
476 are available under the CKB Open Access Data Policy to bona fide researchers. Sharing of  
477 genotyping data is constrained by the Administrative Regulations on Human Genetic  
478 Resources of the People's Republic of China. Access to these and certain other data is  
479 available through collaboration with CKB researchers. Details of the CKB Data Sharing  
480 Policy are available at [www.ckbiobank.org](http://www.ckbiobank.org).

481

482 **Figure legends**

483 **Figure1. Exposure profile of 6597 SomaScan protein biomarkers in CKB, overall and**  
484 **by sex**

485 Three Miami plots are presented: one for female-specific analysis, one for male-specific  
486 analysis, and one for overall analysis. The x-axis represents baseline characteristics  
487 grouped by category, while the y-axis shows the negative logarithm of the p-value (-log<sub>10</sub>  
488 p-value) for the association between each exposure and protein biomarkers. Each dot  
489 represents the -log<sub>10</sub> Bonferroni corrected p-value for these associations. For visualization  
490 purposes, -log<sub>10</sub> p-values exceeding 25 are not displayed (indicated with arrow). Positive  
491 associations are shown in red, negative associations in blue, and non-significant  
492 associations in grey. Analyses are adjusted for age, age<sup>2</sup>, sex, study area, fasting time,  
493 fasting time<sup>2</sup>, outdoor temperature, outdoor temperature<sup>2</sup> and plate ID, where appropriate.  
494 Abbreviations: BMI: Body mass index; CKB: China Kadoorie Biobank; CO: carbon-  
495 monoxide; DBP: Diastolic blood pressure; FEV1/FVC: Forced Expiratory Volume in 1  
496 second / Forced Vital Capacity; HBV: Hepatitis B virus; MET: metabolic equivalent task;  
497 RPG: random plasma glucose

498 **Figure 2. Exposure profiles of the top 25 SomaScan protein biomarkers with most**  
499 **associations, overall and by sex**

500 The bar plots show the number of baseline associated with the 25 most frequently  
501 associated protein biomarkers after Bonferroni corrected p-value. The analyses are  
502 presented separately for females, males, and the overall. The x-axis represents the protein  
503 biomarkers, while the y-axis indicates the number of baseline characteristics associated with  
504 each protein. Bars are color-coded to represent different baseline characteristic groups.  
505 Analyses are adjusted for age, age<sup>2</sup>, sex, study area, fasting time, fasting time<sup>2</sup>, outdoor  
506 temperature, outdoor temperature<sup>2</sup> and plate ID, where appropriate.

507 **Figure 3. Sex- and age-associated SomaScan protein biomarkers and their**  
508 **associations with other exposures**

509 Figures (i)a and (ii)a represent the associations of sex and age, respectively, with protein  
510 biomarkers. The x-axis represents the effect size of the association between sex or age and  
511 the protein biomarkers, while the y-axis indicates the -log<sub>10</sub> p-value. Red dots denote  
512 positive Bonferroni corrected associations, blue dots denote negative Bonferroni corrected  
513 associations, and grey dots denote non-significant associations.

514 Figures (i)b and (ii)b illustrate the top sex- and age-associated protein biomarkers,  
515 respectively, and their associations with other exposures. The width of the ribbons is  
516 inversely proportional to the p-value, indicating the strength of the association (smaller p-  
517 values correspond to wider ribbons). The colors of the ribbons represent different baseline  
518 characteristic groups. The top protein biomarkers that are not associated with other  
519 exposures are not presented in the figure.

520 Analyses are adjusted for age, age<sup>2</sup>, sex, study area, fasting time, fasting time<sup>2</sup>, outdoor  
521 temperature, outdoor temperature<sup>2</sup> and plate ID, where appropriate.

522 Abbreviations: BMI: Body mass index; CO: carbon-monoxide; DBP: Diastolic blood pressure;  
523 HBV: Hepatitis B virus; RPG: random plasma glucose

524 **Figure 4. Lifestyle and frailty indices-associated SomaScan protein biomarkers and**  
525 **their associations with other exposures**

526 Same as Figure 3.

527

528 **References**

- 529 1. Westermann D, Neumann JT, Sørensen NA, Blankenberg S. High-sensitivity  
530 assays for troponin in patients with cardiac disease. *Nat Rev Cardiol* 2017; **14**(8): 472-83.
- 531 2. Ozer J, Ratner M, Shaw M, Bailey W, Schomaker S. The current state of serum  
532 biomarkers of hepatotoxicity. *Toxicology* 2008; **245**(3): 194-205.
- 533 3. Dhingra R, Gona P, Nam B-H, et al. C-Reactive Protein, Inflammatory Conditions,  
534 and Cardiovascular Disease Risk. *Am J Med* 2007; **120**(12): 1054-62.
- 535 4. Gold L, Ayers D, Bertino J, et al. Aptamer-based multiplexed proteomic technology  
536 for biomarker discovery. *PLoS One* 2010; **5**(12): e15004.
- 537 5. Rappaport SM, Smith MT. Environment and disease risks. *Science* 2010; **330**(460-  
538 1).
- 539 6. Yao P, Iona A, Kartsonaki C, et al. Conventional and genetic associations of  
540 adiposity with 1463 proteins in relatively lean Chinese adults. *Eur J Epidemiol* 2023;  
541 **38**(10): 1089-103.
- 542 7. Sun BB, Chiou J, Traylor M, et al. Plasma proteomic associations with genetics and  
543 health in the UK Biobank. *Nature* 2023; **622**(7982): 329-38.
- 544 8. Enroth S, Enroth SB, Johansson A, Gyllenstein U. Protein profiling reveals  
545 consequences of lifestyle choices on predicted biological aging. *Sci Rep* 2015; **5**: 17282.
- 546 9. Watanabe K, Wilmanski T, Diener C, et al. Multiomic signatures of body mass index  
547 identify heterogeneous health phenotypes and responses to a lifestyle intervention. *Nat*  
548 *Med* 2023; **29**(4): 996-1008.
- 549 10. Huang B, Svensson P, Arnlov J, Sundstrom J, Lind L, Ingelsson E. Effects of  
550 cigarette smoking on cardiovascular-related protein profiles in two community-based  
551 cohort studies. *Atherosclerosis* 2016; **254**: 52-8.

- 552 11. Chen Z, Chen J, Collins R, et al. China Kadoorie Biobank of 0.5 million people:  
553 survey methods, baseline characteristics and long-term follow-up. *Int J Epidemiol* 2011;  
554 **40**(6): 1652-66.
- 555 12. Sudlow C, Gallacher J, Allen N, et al. UK biobank: an open access resource for  
556 identifying the causes of a wide range of complex diseases of middle and old age. *PLoS*  
557 *Med* 2015; **12**(3): e1001779.
- 558 13. Iona A, Wang B, Clarke J, et al. An exposome-wide investigation of 2923 Olink  
559 proteins with non-genetic factors in Chinese adults. *medRxiv* 2024: 2024.10.23.24315975.
- 560 14. Mazidi M, Wright N, Yao P, et al. Plasma Proteomics to Identify Drug Targets for  
561 Ischemic Heart Disease. *J Am Coll Cardiol* 2023; **82**(20): 1906-20.
- 562 15. Wang B, Pozarickij A, Mazidi M, et al. Comparative studies of genetic and  
563 phenotypic associations for 2,168 plasma proteins measured by two affinity-based  
564 platforms in 4,000 Chinese adults. *medRxiv* 2023: 2023.12.01.23299236.
- 565 16. Somalogic. SomaScan® v4.0 and v4.1 Data Standardization. 2021.  
566 <https://somallogic.com/tech-notes/>.
- 567 17. Lv J, Yu C, Guo Y, et al. Adherence to a healthy lifestyle and the risk of type 2  
568 diabetes in Chinese adults. *Int J Epidemiol* 2017; **46**(5): 1410-20.
- 569 18. Sun Q, Yu D, Fan J, et al. Healthy lifestyle and life expectancy at age 30 years in  
570 the Chinese population: an observational study. *Lancet Public Health* 2022; **7**(12): e994-  
571 e1004.
- 572 19. The China Kadoorie Biobank Collaborative Group. Healthy lifestyle and life  
573 expectancy free of major chronic diseases at age 40 in China. *Nat Hum Behav* 2023; **7**(9):  
574 1542-50.
- 575 20. Fan J, Yu C, Guo Y, et al. Frailty index and all-cause and cause-specific mortality in  
576 Chinese adults: a prospective cohort study. *Lancet Public Health* 2020; **5**(12): e650-e60.



- 577 21. Gadd DA, Hillary RF, Kuncheva Z, et al. Blood protein assessment of leading  
578 incident diseases and mortality in the UK Biobank. *Nat Aging* 2024; **4**(7): 939-48.
- 579 22. Team RC. R: A Language and Environment for Statistical Computing. Published  
580 online 2022. <http://www.r-project.org/index.html>.
- 581 23. Corlin L, Liu C, Lin H, et al. Proteomic Signatures of Lifestyle Risk Factors for  
582 Cardiovascular Disease: A Cross-Sectional Analysis of the Plasma Proteome in the  
583 Framingham Heart Study. *J Am Heart Assoc* 2021; **10**(1): e018020.
- 584 24. Sathyan S, Ayers E, Gao T, et al. Plasma proteomic profile of age, health span, and  
585 all-cause mortality in older adults. *Aging Cell* 2020; **19**(11): e13250.
- 586 25. Sathyan S, Ayers E, Gao T, Milman S, Barzilai N, Verghese J. Plasma proteomic  
587 profile of frailty. *Aging Cell* 2020; **19**(9): e13193.
- 588 26. Goudswaard LJ, Bell JA, Hughes DA, et al. Effects of adiposity on the human  
589 plasma proteome: observational and Mendelian randomisation estimates. *Int J Obes*  
590 *(Lond)* 2021; **45**(10): 2221-9.
- 591 27. Gudmundsdottir V, Zaghlool SB, Emilsson V, et al. Circulating Protein Signatures  
592 and Causal Candidates for Type 2 Diabetes. *Diabetes* 2020; **69**(8): 1843-53.
- 593 28. Carrasco-Zanini J, Wheeler E, Uluvar B, et al. Mapping biological influences on the  
594 human plasma proteome beyond the genome. *Nature Metabolism* 2024; **6**(10): 2010-23.
- 595 29. Ni W, Breitner S, Nikolaou N, et al. Effects of Short- And Medium-Term Exposures  
596 to Lower Air Temperature on 71 Novel Biomarkers of Subclinical Inflammation: Results  
597 from the KORA F4 Study. *Environ Sci Technol* 2023; **57**(33): 12210-21.
- 598 30. Bao X, Borné Y, Yin S, et al. The associations of self-rated health with  
599 cardiovascular risk proteins: a proteomics approach. *Clin Proteomics* 2019; **16**: 40.
- 600 31. Pang Y, Kartsonaki C, Lv J, et al. Associations of Adiposity, Circulating Protein  
601 Biomarkers, and Risk of Major Vascular Diseases. *JAMA Cardiol* 2021; **6**(3): 276-86.

- 602 32. Bhartiya D, Patel H. An overview of FSH-FSHR biology and explaining the existing  
603 conundrums. *J Ovarian Res* 2021; **14**(1): 144.
- 604 33. Barqué A, Jan K, De La Fuente E, Nicholas CL, Hynes RO, Naba A. Knockout of  
605 the gene encoding the extracellular matrix protein SNED1 results in early neonatal lethality  
606 and craniofacial malformations. *Dev Dyn* 2021; **250**(2): 274-94.
- 607 34. Balk SP, Ko YJ, Bublely GJ. Biology of prostate-specific antigen. *J Clin Oncol* 2003;  
608 **21**(2): 383-91.
- 609 35. Yamaguchi Y, Nagase T, Makita R, et al. Identification of multiple novel epididymis-  
610 specific beta-defensin isoforms in humans and mice. *J Immunol* 2002; **169**(5): 2516-23.
- 611 36. Saad MF, Damani S, Gingerich RL, et al. Sexual dimorphism in plasma leptin  
612 concentration. *J Clin Endocrinol Metab* 1997; **82**(2): 579-84.
- 613 37. Zillikens MC, Yazdanpanah M, Pardo LM, et al. Sex-specific genetic effects  
614 influence variation in body composition. *Diabetologia* 2008; **51**(12): 2233-41.
- 615 38. Niu L, Stinson SE, Holm LA, et al. Plasma Proteome Variation and its Genetic  
616 Determinants in Children and Adolescents. *medRxiv* 2023: 2023.03.31.23287853.
- 617 39. Bernabeu E, Canela-Xandri O, Rawlik K, Talenti A, Prendergast J, Tenesa A. Sex  
618 differences in genetic architecture in the UK Biobank. *Nat Genet* 2021; **53**(9): 1283-9.
- 619 40. Koprulu M, Wheeler E, Kerrison ND, et al. Similar and different: systematic  
620 investigation of proteogenomic variation between sexes and its relevance for human  
621 diseases. *medRxiv* 2024: 2024.02.16.24302936.
- 622 41. Williams SA, Kivimaki M, Langenberg C, et al. Plasma protein patterns as  
623 comprehensive indicators of health. *Nat Med* 2019; **25**(12): 1851-7.
- 624 42. Im PK, Wright N, Yang L, et al. Alcohol consumption and risks of more than 200  
625 diseases in Chinese men. *Nat Med* 2023; **29**(6): 1476-86.

- 626 43. Chan KH, Wright N, Xiao D, et al. Tobacco smoking and risks of more than 470  
627 diseases in China: a prospective cohort study. *Lancet Public Health* 2022; **7**(12): e1014-  
628 e26.
- 629 44. Sun BB, Suhre K, Gibson BW. Promises and Challenges of populational  
630 Proteomics in Health and Disease. *Mol Cell Proteomics* 2024; **23**(7): 100786.
- 631 45. Krupnik VE, Sharp JD, Jiang C, et al. Functional and structural diversity of the  
632 human Dickkopf gene family. *Gene* 1999; **238**(2): 301-13.
- 633 46. Baetta R, Banfi C. Dkk (Dickkopf) Proteins. *Arterioscler Thromb Vasc Biol* 2019;  
634 **39**(7): 1330-42.
- 635 47. Giralt I, Gallo-Oller G, Navarro N, et al. Dickkopf Proteins and Their Role in Cancer:  
636 A Family of Wnt Antagonists with a Dual Role. *Pharmaceuticals (Basel)* 2021; **14**(8).
- 637 48. Sebastiani P, Federico A, Morris M, et al. Protein signatures of centenarians and  
638 their offspring suggest centenarians age slower than other humans. *Aging Cell* 2021;  
639 **20**(2): e13290.
- 640 49. Wang X. Chapter Three - Pleiotrophin: Activity and mechanism. In: Makowski GS,  
641 ed. *Advances in Clinical Chemistry*: Elsevier; 2020: 51-89.
- 642 50. Menni C, Kiddle SJ, Mangino M, et al. Circulating Proteomic Signatures of  
643 Chronological Age. *J Gerontol A Biol Sci Med Sci* 2015; **70**(7): 809-16.
- 644 51. Carpenter SL, Mathew P. Alpha2-antiplasmin and its deficiency: fibrinolysis out of  
645 balance. *Haemophilia* 2008; **14**(6): 1250-4.
- 646 52. Weidmann H, Heikaus L, Long AT, Naudin C, Schlüter H, Renné T. The plasma  
647 contact system, a protease cascade at the nexus of inflammation, coagulation and  
648 immunity. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* 2017; **1864**(11,  
649 Part B): 2118-27.
- 650 53. Hall JE. Endocrinology of the Menopause. *Endocrinol Metab Clin North Am* 2015;  
651 **44**(3): 485-96.

- 652 54. Gagnon E, Pelletier W, Gobeil É, et al. Mendelian randomization prioritizes  
653 abdominal adiposity as an independent causal factor for liver fat accumulation and  
654 cardiometabolic diseases. *Commun Med (Lond)* 2022; **2**: 130.
- 655 55. Dale CE, Fatemifar G, Palmer TM, et al. Causal Associations of Adiposity and Body  
656 Fat Distribution With Coronary Heart Disease, Stroke Subtypes, and Type 2 Diabetes  
657 Mellitus: A Mendelian Randomization Analysis. *Circulation* 2017; **135**(24): 2373-88.
- 658 56. Iona A, Yao P, Pozarickij A, et al. Proteo-genomic analyses in relatively lean  
659 Chinese adults identify proteins and pathways that affect general and central adiposity  
660 levels. *Communications Biology* 2024; **7**(1): 1327.
- 661 57. Cronjé HT, Mi MY, Austin TR, et al. Plasma Proteomic Risk Markers of Incident  
662 Type 2 Diabetes Reflect Physiologically Distinct Components of Glucose-Insulin  
663 Homeostasis. *Diabetes* 2023; **72**(5): 666-73.
- 664 58. Zaghlool SB, Halama A, Stephan N, et al. Metabolic and proteomic signatures of  
665 type 2 diabetes subtypes in an Arab population. *Nat Commun* 2022; **13**(1): 7121.
- 666 59. Saxena R, Elbers CC, Guo Y, et al. Large-scale gene-centric meta-analysis across  
667 39 studies identifies type 2 diabetes loci. *Am J Hum Genet* 2012; **90**(3): 410-25.
- 668 60. Santos RAS, Oudit GY, Verano-Braga T, Canta G, Steckelings UM, Bader M. The  
669 renin-angiotensin system: going beyond the classical paradigms. *Am J Physiol Heart Circ*  
670 *Physiol* 2019; **316**(5): H958-h70.
- 671 61. Wang H, Nie Y, Sun Z, He Y, Yang J. Serum amyloid P component: Structure,  
672 biological activity, and application in diagnosis and treatment of immune-associated  
673 diseases. *Molecular Immunology* 2024; **172**: 1-8.
- 674 62. Lorenzi M, Lorenzi T, Marzetti E, et al. Association of frailty with the serine protease  
675 HtrA1 in older adults. *Exp Gerontol* 2016; **81**: 8-12.

- 676 63. Simsek T, Bal Albayrak MG, Akpınar G, Canturk NZ, Kasap M. Downregulated  
677 GPD1 and MAGL protein levels as potential biomarkers for the metastasis of  
678 triple-negative breast tumors to axillary lymph nodes. *Oncol Lett* 2024; **27**(1): 34.
- 679 64. Price MJ, Nguyen AD, Byemerwa JK, Flowers J, Baëta CD, Goodwin CR. UDP-  
680 glucose dehydrogenase (UGDH) in clinical oncology and cancer biology. *Oncotarget* 2023;  
681 **14**: 843-57.
- 682 65. Cooper LA, Page ST, Amory JK, Anawalt BD, Matsumoto AM. The association of  
683 obesity with sex hormone-binding globulin is stronger than the association with ageing –  
684 implications for the interpretation of total testosterone measurements. *Clinical*  
685 *Endocrinology* 2015; **83**(6): 828-33.
- 686 66. Hjortebjerg R, Kristiansen MR, Brandslund I, et al. Associations between insulin-like  
687 growth factor binding protein-2 and insulin sensitivity, metformin, and mortality in persons  
688 with T2D. *Diabetes Res Clin Pract* 2023; **205**: 110977.
- 689 67. Renehan AG, Atkin WS, O'Dwyer S T, Shalet SM. The effect of cigarette smoking  
690 use and cessation on serum insulin-like growth factors. *Br J Cancer* 2004; **91**(8): 1525-31.
- 691 68. Perry AS, Zhang K, Murthy VL, et al. Proteomics, Human Environmental Exposure,  
692 and Cardiometabolic Risk. *Circ Res* 2024.
- 693 69. Wang D, Li C, Chiu MC, et al. SPINK6 inhibits human airway serine proteases and  
694 restricts influenza virus activation. *EMBO Mol Med* 2022; **14**(1): e14485.
- 695 70. Moliere S, Jaulin A, Tomasetto CL, Dali-Youcef N. Roles of Matrix  
696 Metalloproteinases and Their Natural Inhibitors in Metabolism: Insights into Health and  
697 Disease. *Int J Mol Sci* 2023; **24**(13).
- 698 71. Rauch JN, Luna G, Guzman E, et al. LRP1 is a master regulator of tau uptake and  
699 spread. *Nature* 2020; **580**(7803): 381-5.
- 700 72. Candia J, Daya GN, Tanaka T, Ferrucci L, Walker KA. Assessment of variability in  
701 the plasma 7k SomaScan proteomics assay. *Sci Rep* 2022; **12**(1): 17147.

- 702 73. Lopez-Silva C, Surapaneni A, Coresh J, et al. Comparison of aptamer-based and  
703 antibody-based assays for protein quantification in chronic kidney disease. *CJASN* 2022;  
704 **17**(3): 350-60.
- 705 74. Pietzner M, Wheeler E, Carrasco-Zanini J, et al. Synergistic insights into human  
706 health from aptamer-and antibody-based proteomic profiling. *Nat commun* 2021; **12**(1):  
707 6822.
- 708 75. Kraemer S, Schneider DJ, Paterson C, et al. Crossing the Halfway Point: Aptamer-  
709 Based, Highly Multiplexed Assay for the Assessment of the Proteome. *J Proteome Res*  
710 2024.
- 711 76. Candia J, Cheung F, Kotliarov Y, et al. Assessment of Variability in the SOMAScan  
712 Assay. *Scientific Reports* 2017; **7**(1): 14248.
- 713

**Table 1. Baseline characteristics of participants and their associations with SomaScan protein biomarkers**

Characteristics	Mean (SD) or percentage <sup>a</sup>			No. of significant associations <sup>b</sup>		
	Female (1,243)	Male ( 755)	All (1,998)	Female	Male	All
<b>Demographics</b>						
Age	50.7 (10.2)	50.8 (11.0)	50.8 (10.5)	735	593	982
Sex	-	-	-	-	-	996
Urban residents	52.4	49.2	51.2	506	425	858
Schooling ,> 9 years	20.4	27.6	22.9	0	0	0
Employed	60.1	77.5	66.8	0	0	1
Household income ,≥ ¥20,000	43.0	47.4	44.5	0	5	7
Ownership index <sup>c</sup>	3.3 (1.3)	3.4 (1.4)	3.3 (1.3)	0	16	13
<b>Lifestyle habits</b>						
Regular alcohol drinker	2.6	37.2	25.3	2	85	99
Current smoker	2.3	63.3	15.3	5	24	36
<b>Diet</b>						
Food diversity score <sup>d</sup>	11.4 (3.3)	11.3 (3.2)	11.3 (3.3)	0	0	3
Rapeseed oil	33.3	38.8	35.4	5	0	7
Physical activity, MET-hrs/day	20.5 (13.2)	23.2 (16.3)	21.4 (14.5)	0	3	9
<b>Environmental</b>						
Outdoor temperature, °C	16.0 (10.6)	15.7 (10.9)	15.9 (10.7)	567	336	802
Clean heating fuel	45.3	44.3	45.0	0	0	0
Clean cooking fuel	49.9	36.0	44.8	3	0	0
<b>Health and wellbeing</b>						
<b>Prior physical health status</b>						
Self-rated health	8.5	8.2	8.3	0	0	1
Respiratory disease	8.3	8.0	8.2	0	0	1
Kidney/liver disease	2.1	2.5	2.2	6	19	5
HBsAg+	2.2	2.5	2.3	260	115	468
Diabetes <sup>e</sup>	7.0	5.8	6.5	166	34	257
Cancer	0.6	0.6	0.7	14	27	10
<b>Mental wellbeing</b>						
Life satisfaction	3.7	4.9	4.0	1	1	0
Mental disorder	1.1	1.5	1.2	7	9	3
<b>Clinical measurements</b>						
BMI, kg/m <sup>2</sup>	24.0 (3.5)	23.7 (3.3)	23.9 (3.4)	595	498	1035
Standing height, cm	154.5 (6.1)	165.8 (6.5)	158.7 (8.3)	3	6	22
SBP, mmHg	129.4 (22.2)	132.6 (19.9)	130.5 (21.4)	136	87	293
DBP, mmHg	77.2 (10.6)	79.7 (11.6)	78.0 (11.1)	60	74	233
Heart rate, bpm	79.4 (11.4)	78.0 (11.9)	78.8 (11.6)	61	40	181
Exhaled CO, ppm	5.0 (2.2)	11.7 (2.5)	7.5 (2.3)	1	20	12
FEV1/FVC ratio	85.1 ( 6.1)	84.9 (10.1)	85.0 (8.5)	0	4	2
RPG, mmol/L	6.1 (8.2)	5.9 (8.8)	6.0 (8.5)	251	77	343
Fasting time, hours	5.2 (5.0)	5.0 (5.0)	5.1 (5.0)	45	21	86
<b>Reproductive factors</b>						
Age at menarche, years	15.4 (2.0)	-	15.4 (2.0)	0	-	0
Age at menopause, years	39.2 (4.3)	-	39.2 (4.3)	1	-	1
Post-menopausal	54.7	-	54.7	58	-	58
Parity	99.8	-	99.8	29	-	29
Age at first live birth, years	23.9 (3.3)	-	23.9 (3.3)	0	-	0
<b>Composite scores</b>						
Lifestyle index <sup>f</sup>	3.1 (0.8)	2.2 (1.0)	2.8 (1.0)	106	96	307
Frailty index <sup>g</sup>	0.1 (0.06)	0.1 (0.06)	0.1 (0.06)	226	82	465

<sup>a</sup> Baseline characteristics adjusted for age (10-year age groups) and region (10 regions).

<sup>b</sup> Analyses are adjusted for age, age<sup>2</sup>, sex, study area, fasting time, fasting time<sup>2</sup>, outdoor temperature, outdoor temperature<sup>2</sup> and plate ID, where appropriate. Bonferroni (PCA) corrected p-value < 0.05

<sup>c</sup> 6-point index of qualitative measures of living standards

<sup>d</sup> 24-point index of frequency of intake in 12 food groups

<sup>e</sup> Self-reported and screen detected

<sup>f</sup> 5-point index of low-risk lifestyle characteristics

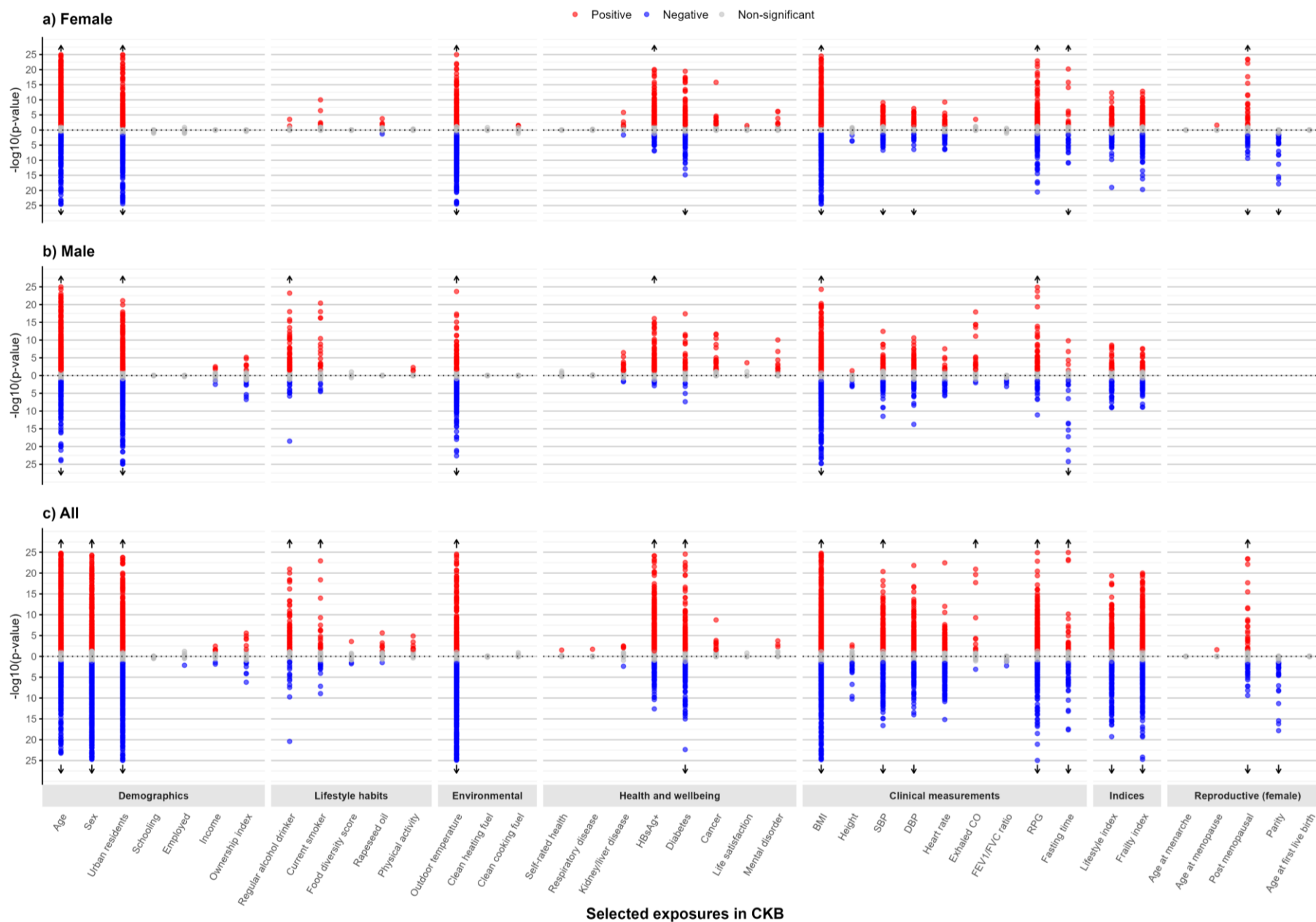
<sup>g</sup> 28-point index of accumulation of health deficits and physical activity

Abbreviations: BMI: Body mass index; CO: carbon monoxide; DBP: Diastolic blood pressure;

FEV1/FVC: Forced Expiratory Volume in 1 second / Forced Vital Capacity; HBsAg+: Hepatitis B virus surface antigen seropositive;

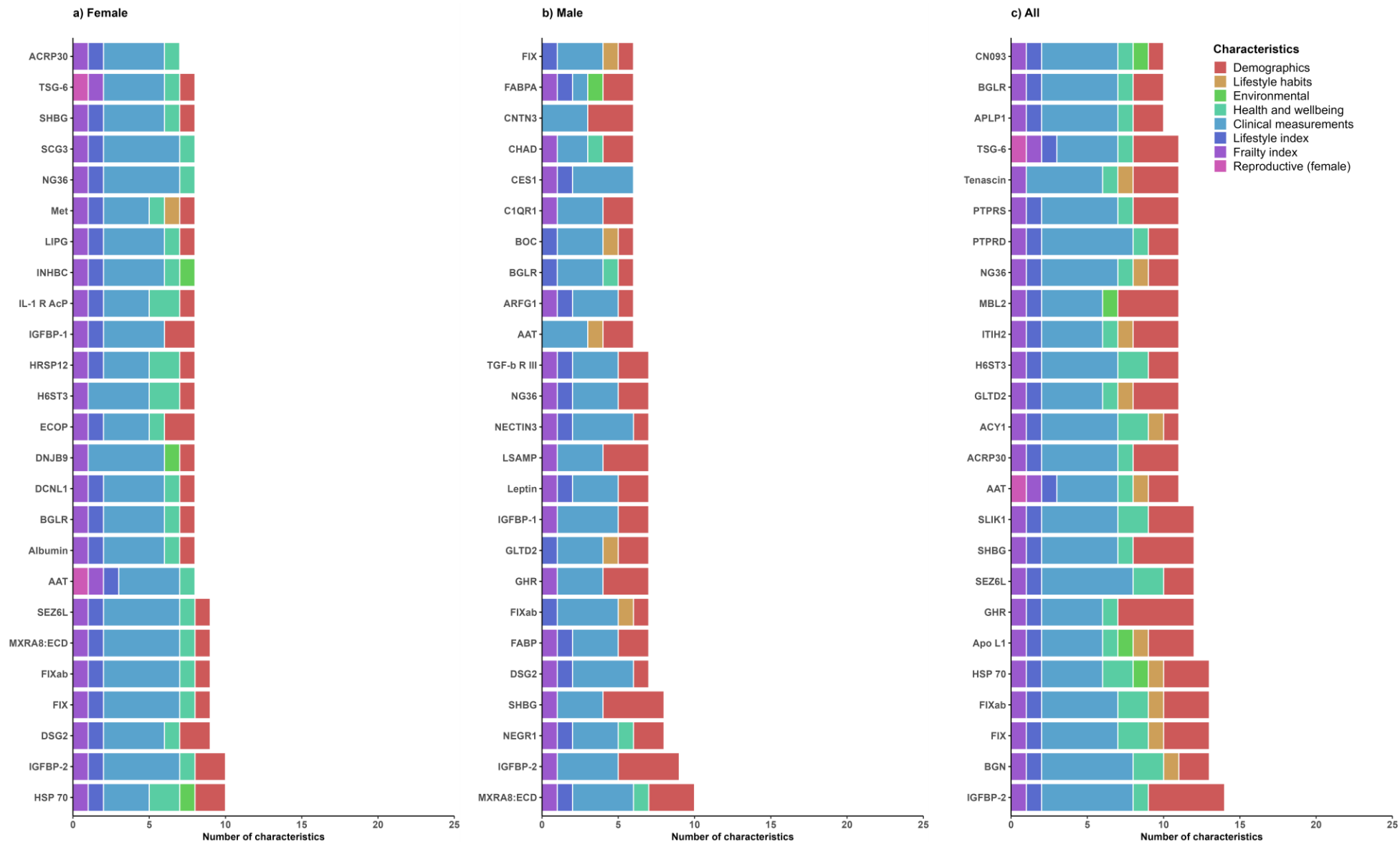
MET: metabolic equivalent of task; RPG: random plasma glucose

Figure 1. Exposure profiles of 6597 SomaScan protein biomarkers in CKB, overall and by sex



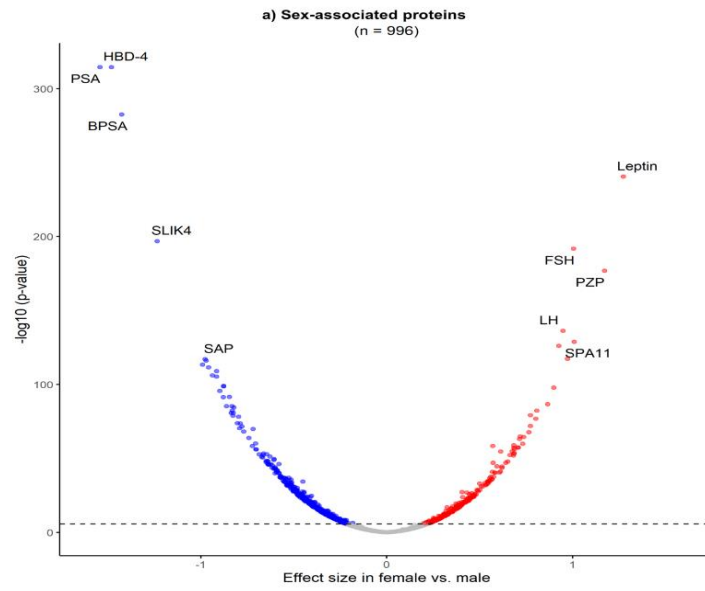


**Figure 2. Exposure profiles by characteristics type of the top 25 SomaScan protein biomarkers with most associations, overall and by sex**

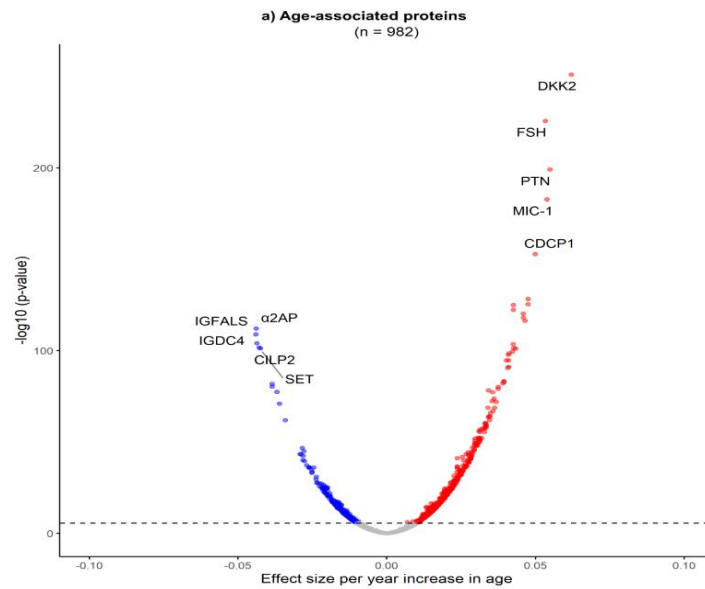


**Figure 3. Sex- and age-associated SomaScan protein biomarkers and their associations with other exposures**

**(i) Sex**



**(ii) Age**



**Figure 4. Lifestyle and frailty indices-associated SomaScan protein biomarkers and their associations with other exposures**

