#### An exposome-wide assessment of 6600 SomaScan proteins 1

2

## 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>, 3 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 4 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 5 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>, 6 Yongmei Liu<sup>5</sup>, Canging 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>, 7 Robert Clarke<sup>1</sup>, Derrick A Bennett<sup>1†</sup>, Zhengming Chen<sup>1†</sup>, on behalf of China Kadoorie 8 Biobank Collaborative Group# 9 10

- 1. Clinical Trial Service Unit, Nuffield Department of Population Health, University of 11 Oxford, Oxford, UK 12
- 13 2. Department of Epidemiology & Biostatistics, School of Public Health, Peking University, Beijing, China 14
- 3. Peking University Center for Public Health and Epidemic Preparedness and Response, 15 Beiiing, China 16
- 4. Key Laboratory of Epidemiology of Major Diseases (Peking University), Ministry of 17 Education, Beijing, China 18
- 5. NCDs Prevention and Control Department, Qingdao CDC, China 19
- 20 \* Co-first author; <sup>†</sup>Co-corresponding author; *# Members of the CKB Collaborative Group*
- are shown in the Appendix. 21

#### Address for correspondence:

24	Assoc. Professor Derrick Bennett	or	Professor Zhengming Chen
25	CTSU, Big Data Institute,		CTSU, Big Data Institute,
26	Old Road Campus,		Old Road Campus
27	University of Oxford		University of Oxford
28	Oxford, OX3 7LF, UK		Oxford, OX3 7LF, UK
29	Tel: 44-1865-743949		Tel: 44-1865-743839
30	derrick.bennett@ndph.ox.ac.uk		zhengming.chen@ctsu.ox.ac.uk

31	Word count: Abstract 250/250, Text 4389
32	(1 Table, 4 Figures and 20 Supplementary Tables and Figures)
33	(CKB research tracker No.: 2023-0065; CKB data release 19.00)
34	
35	
36	
37	23 October 2024

38

22

23

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

## 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	C1g and tumor necrosis factor related protein 1
C1QTNF3	C1g and tumor necrosis factor related protein 3
CFHR5	Complement factor H-related 5
CII P2	Cartilage intermediate laver protein 2
CKB	China Kadoorie Biobank
CO	Carbon monoxide
CRDI 1	Chordin-like protein 1
CRDL2	Chordin-like protein 2
CRP	C-reactive protein
DRP	Vitamin D-binding protein
	Dickkonf-related protein 2
D IB12	Dna L homolog subfamily B member 12
ESDN	Espin
EARD	Estiv acid hinding protein
	Fatty acid binding protein adipocyte
	Conquisition factor IX
FIXab	
FINAD	
	Crowth hormone recenter
	Glowin normone receptor
	Giyceror-S-phosphate deltydrogenase [NAD(+)], cytopiasmic
	Henroylobili Suburili della-4
	Hepatitis Divinue
HBV	Hepallis B vilus
	Human chononic gonadotropin
	Hura Serine Pepudase 1
	Reparan-sullate 6-O-sullotransierase 2
	Controsteroid 11-beta-denydrogenase isozyme 1
	Immunoglobulin superramily, 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
	Ischemic heart disease
IL-1 R ACP	Interleukin-1 Receptor accessory protein
INHBC	Inhibin beta C chain
LH	Luteinizing hormone
	Limit of detection
	Lipoprotein lipase
LRP1	Ldl receptor related protein 1
MIC-1	Macrophage inhibitory cytokine 1
MMP7	Matrix metallopeptidase 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 α2AP	WAP, Kazal, immunoglobulin, Kunitz and NTR domain-containing protein 2 A2-antiplasmin
	•

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

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

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

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

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

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

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

#### 70 Introduction

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

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

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

compare the findings with a parallel investigation<sup>13</sup> involving ~3000 proteins measured using
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 elsewhere.<sup>11</sup> Briefly, CKB recruited ~512,000 adults aged 30-79 years from 10 100 geographically diverse areas in 2004-2008. At baseline, trained health workers administered 101 102 a comprehensive laptop-based questionnaire and recorded physical measurements, including anthropometry and blood pressure, using regularly-calibrated instruments 103 following standardised protocols. Desktop biochemical assays included random plasma 104 glucose (RPG) and hepatitis B surface antigen (HBsAg). A 10-mL non-fasting (with time 105 since the last meal recorded) blood sample was collected from each participant, and then 106 107 processed and stored in liquid nitrogen. The present study involved 2,026 randomly selected subcohort participants who had no prior history of cardiovascular disease, originally sampled 108 along with 1,951 cases of incident ischaemic heart disease (IHD) for a case-cohort study.<sup>6,14</sup> 109

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

114 Proteomic assays

Details of the proteomic assays in the CKB have been described elsewhere.<sup>6,14,15</sup> In brief, 60µl plasma aliquots in 2D-barcoded microtubes of 3,977 participants were delivered to the Somalogic Laboratory in Colorado, USA for profiling using SomaScan Assay v4.1, which 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 96-well plates (including 11 wells for external control samples, including 5 calibrator, 3 QC, 120 and 3 buffer samples). The raw output of the SomaScan assay were standardised based on 121 122 external control samples to account for variability in microarrays and variations within and between plates. With an optional procedure of adaptive normalisation by maximum 123 likelihood (ANML) to an external reference, which controls for inter-sample variability, the 124 results were provided in normalised and non-normalised values in relative fluorescence units 125 (RFU).<sup>16</sup> The output values were further natural log-transformed in the main analysis. The 126 limit of detection (LOD) for SOMAmers were defined with external buffer samples. Quality 127 control (QC) checks were conducted comparing the median of QC samples on each plate 128 to the reference, with a cross-place QC indicator (pass/flag) assigned to each SOMAmer. 129 130 The 7,289 SOMAmers targeting human proteins were mapped to proteins based on their UniProt IDs supplied by Somalogic. Details of individual proteins and their distributions by 131 sex are shown in eTable 1. 132

#### 133 Selected baseline characteristics

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

#### 145 Statistical analysis

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

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

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

at  $0.05/(1186 \times 26) = 1.62 \times 10^{-6}$ . All statistical analyses were performed using R version 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**

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

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

193 with clinical measurements and demographic factors, overall and in sex-specific analyses

#### 194 (**Figure 2; eFigure 2**).

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

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

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

prior medical history showed few significant associations (Table 1). For HBsAg status, the
top positively associated proteins were DHI1, NUD16, DJB12, C1QTNF3, and AKR1D1,
while the top inversely associated proteins were CFHR5, SAP, RBP, IL-1 R AcP, and α2AP
(eFigure 7). For prevalent diabetes, the top positively associated proteins included PLXB2
and SEMA6A and top inversely associated proteins were CILP2 and MXRA8:ECD (eFigure 7).

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

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

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

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

#### 266 Discussion

This exposome-wide analyses of almost 6,600 SomaScan proteins in ~2,000 Chinese adults
 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 demonstrated the largest number of significant associations with multiple proteins, with 270 many of these associations varying by sex. Other exposures including demographic factors, 271 272 lifestyle habits, clinical measurements, and lifestyle and frailty indices were also related to a very large number of proteins. Overall, there was a larger proportion of significant 273 associations with high-abundance proteins, but we still found associations with low-274 abundance but biological-important proteins not captured in the Olink immunoassays (e.g., 275 PSA, HBD-4, DKK2). 276

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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 disease registries. The Chinese National Health Insurance scheme provided electronic 444 linkage to all hospital admissions. The China Kadoorie Biobank study is jointly coordinated 445 by the University of Oxford and the Chinese Academy of Medical Sciences. 446

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

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

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

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

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

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

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

#### 482 Figure legends

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

Three Miami plots are presented: one for female-specific analysis, one for male-specific 485 analysis, and one for overall analysis. The x-axis represents baseline characteristics 486 grouped by category, while the y-axis shows the negative logarithm of the p-value (-log10 487 p-value) for the association between each exposure and protein biomarkers. Each dot 488 represents the -log10 Bonferroni corrected p-value for these associations. For visualization 489 purposes, -log10 p-values exceeding 25 are not displayed (indicated with arrow). Positive 490 associations are shown in red, negative associations in blue, and non-significant 491 associations in grey. Analyses are adjusted for age, age<sup>2</sup>, sex, study area, fasting time, 492 fasting time<sup>2</sup>, outdoor temperature, outdoor temperature<sup>2</sup> and plate ID, where appropriate. 493 Abbreviations: BMI: Body mass index; CKB: China Kadoorie Biobank; CO: carbon-494 monoxide; DBP: Diastolic blood pressure; FEV1/FVC: Forced Expiratory Volume in 1 495 second / Forced Vital Capacity; HBV: Hepatitis B virus; MET: metabolic equivalent task; 496 RPG: random plasma glucose 497

- Figure 2. Exposure profiles of the top 25 SomaScan protein biomarkers with most associations, overall and by sex
- The bar plots show the number of baseline associated with the 25 most frequently associated protein biomarkers after Bonferroni corrected p-value. The analyses are presented separately for females, males, and the overall. The x-axis represents the protein biomarkers, while the y-axis indicates the number of baseline characteristics associated with each protein. Bars are color-coded to represent different baseline characteristic groups. 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.

#### 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 –log10 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.
- Figures (i)b and (ii)b illustrate the top sex- and age-associated protein biomarkers, respectively, and their associations with other exposures. The width of the ribbons is inversely proportional to the p-value, indicating the strength of the association (smaller pvalues correspond to wider ribbons). The colors of the ribbons represent different baseline characteristic groups. The top protein biomarkers that are not associated with other 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.

#### 528 References

- 529 1. Westermann D, Neumann JT, Sörensen NA, Blankenberg S. High-sensitivity
- 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
- biomarkers of hepatotoxicity. *Toxicology* 2008; **245**(3): 194-205.
- 3. Dhingra R, Gona P, Nam B-H, et al. C-Reactive Protein, Inflammatory Conditions,
- and Cardiovascular Disease Risk. *Am J Med* 2007; **120**(12): 1054-62.
- Gold L, Ayers D, Bertino J, et al. Aptamer-based multiplexed proteomic technology
  for biomarker discovery. *PLoS One* 2010; **5**(12): e15004.
- 537 5. Rappaport SM, Smith MT. Enviornment and disease risks. Science 2010; 330(460-

538 1).

539 6. Yao P, Iona A, Kartsonaki C, et al. Conventional and genetic associations of

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, Gyllensten U. Protein profiling reveals

consequences of lifestyle choices on predicted biological aging. *Sci Rep* 2015; **5**: 17282.

Watanabe K, Wilmanski T, Diener C, et al. Multiomic signatures of body mass index
identify heterogeneous health phenotypes and responses to a lifestyle intervention. *Nat Med* 2023; **29**(4): 996-1008.

Huang B, Svensson P, Arnlov J, Sundstrom J, Lind L, Ingelsson E. Effects of
cigarette smoking on cardiovascular-related protein profiles in two community-based
cohort studies. *Atherosclerosis* 2016; **254**: 52-8.

11. Chen Z, Chen J, Collins R, et al. China Kadoorie Biobank of 0.5 million people:

- survey methods, baseline characteristics and long-term follow-up. *Int J Epidemiol* 2011; **40**(6): 1652-66.
- 555 12. Sudlow C, Gallacher J, Allen N, et al. UK biobank: an open access resource for
- identifying the causes of a wide range of complex diseases of middle and old age. *PLoS*
- 557 *Med* 2015; **12**(3): e1001779.
- 13. Iona A, Wang B, Clarke J, et al. An exposome-wide investigation of 2923 Olink
- 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.

- 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
- platforms in 4,000 Chinese adults. *medRxiv* 2023: 2023.12.01.23299236.
- 16. Somalogic. SomaScan® v4.0 and v4.1 Data Standardization. 2021.
- 566 <u>https://somalogic.com/tech-notes/</u>.
- 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.
- 18. Sun Q, Yu D, Fan J, et al. Healthy lifestyle and life expectancy at age 30 years in
- the Chinese population: an observational study. *Lancet Public Health* 2022; **7**(12): e994e1004.
- 19. The China Kadoorie Biobank Collaborative Group. Healthy lifestyle and life
- 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. <u>http://www.r-project.org/index.html</u>.
- 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.
- Sathyan S, Ayers E, Gao T, et al. Plasma proteomic profile of age, health span, and
  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
- to Lower Air Temperature on 71 Novel Biomarkers of Subclinical Inflammation: Results
- from the KORA F4 Study. *Environ Sci Technol* 2023; **57**(33): 12210-21.
- 30. Bao X, Borné Y, Yin S, et al. The associations of self-rated health with
- cardiovascular risk proteins: a proteomics approach. *Clin Proteomics* 2019; **16**: 40.
- 31. Pang Y, Kartsonaki C, Lv J, et al. Associations of Adiposity, Circulating Protein
- Biomarkers, and Risk of Major Vascular Diseases. *JAMA Cardiol* 2021; **6**(3): 276-86.

Bhartiya D, Patel H. An overview of FSH-FSHR biology and explaining the existing
conundrums. *J Ovarian Res* 2021; **14**(1): 144.

Barqué A, Jan K, De La Fuente E, Nicholas CL, Hynes RO, Naba A. Knockout of
the gene encoding the extracellular matrix protein SNED1 results in early neonatal lethality
and craniofacial malformations. *Dev Dyn* 2021; **250**(2): 274-94.

34. Balk SP, Ko YJ, Bubley GJ. Biology of prostate-specific antigen. *J Clin Oncol* 2003;
21(2): 383-91.

35. Yamaguchi Y, Nagase T, Makita R, et al. Identification of multiple novel epididymis-

specific beta-defensin isoforms in humans and mice. *J Immunol* 2002; **169**(5): 2516-23.

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.

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.

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.

39. Bernabeu E, Canela-Xandri O, Rawlik K, Talenti A, Prendergast J, Tenesa A. Sex

differences in genetic architecture in the UK Biobank. *Nat Genet* 2021; **53**(9): 1283-9.

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

diseases. *medRxiv* 2024: 2024.02.16.24302936.

41. Williams SA, Kivimaki M, Langenberg C, et al. Plasma protein patterns as

comprehensive indicators of health. *Nat Med* 2019; **25**(12): 1851-7.

42. Im PK, Wright N, Yang L, et al. Alcohol consumption and risks of more than 200
diseases in Chinese men. *Nat Med* 2023; **29**(6): 1476-86.

43. Chan KH, Wright N, Xiao D, et al. Tobacco smoking and risks of more than 470

- diseases in China: a prospective cohort study. *Lancet Public Health* 2022; **7**(12): e1014e26.
- 44. Sun BB, Suhre K, Gibson BW. Promises and Challenges of populational
- Proteomics in Health and Disease. *Mol Cell Proteomics* 2024; **23**(7): 100786.
- 45. Krupnik VE, Sharp JD, Jiang C, et al. Functional and structural diversity of the
- human Dickkopf gene family. *Gene* 1999; **238**(2): 301-13.
- 46. Baetta R, Banfi C. Dkk (Dickkopf) Proteins. *Arterioscler Thromb Vasc Biol* 2019;
  39(7): 1330-42.
- 47. Giralt I, Gallo-Oller G, Navarro N, et al. Dickkopf Proteins and Their Role in Cancer:

A Family of Wnt Antagonists with a Dual Role. *Pharmaceuticals (Basel)* 2021; **14**(8).

48. Sebastiani P, Federico A, Morris M, et al. Protein signatures of centenarians and

their offspring suggest centenarians age slower than other humans. *Aging Cell* 2021;

639 **20**(2): e13290.

49. Wang X. Chapter Three - Pleiotrophin: Activity and mechanism. In: Makowski GS,

ed. Advances in Clinical Chemistry: Elsevier; 2020: 51-89.

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.

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
- immunity. Biochimica et Biophysica Acta (BBA) Molecular Cell Research 2017; **1864**(11,
- 649 Part B): 2118-27.

53. Hall JE. Endocrinology of the Menopause. *Endocrinol Metab Clin North Am* 2015;
44(3): 485-96.

- 52 54. Gagnon E, Pelletier W, Gobeil É, et al. Mendelian randomization prioritizes
- abdominal adiposity as an independent causal factor for liver fat accumulation and
- cardiometabolic diseases. *Commun Med (Lond)* 2022; **2**: 130.
- 55 55. Dale CE, Fatemifar G, Palmer TM, et al. Causal Associations of Adiposity and Body
- Fat Distribution With Coronary Heart Disease, Stroke Subtypes, and Type 2 Diabetes
- Mellitus: A Mendelian Randomization Analysis. *Circulation* 2017; **135**(24): 2373-88.
- 558 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.
- 58. Zaghlool SB, Halama A, Stephan N, et al. Metabolic and proteomic signatures of
- type 2 diabetes subtypes in an Arab population. *Nat Commun* 2022; **13**(1): 7121.
- 59. Saxena R, Elbers CC, Guo Y, et al. Large-scale gene-centric meta-analysis across
  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
- 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,
- biological activity, and application in diagnosis and treatment of immune-associated
- 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, Akpinar G, Canturk NZ, Kasap M. Downregulated

- 677 GPD1 and MAGL protein levels as potential biomarkers for the metastasis of
- 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-
- 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
- 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
- growth factor binding protein-2 and insulin sensitivity, metformin, and mortality in persons
  with T2D. *Diabetes Res Clin Pract* 2023; **205**: 110977.
- 689 67. Renehan AG, Atkin WS, O'Dwyer ST, Shalet SM. The effect of cigarette smoking
- 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,
- 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).
- Rauch JN, Luna G, Guzman E, et al. LRP1 is a master regulator of tau uptake and
  spread. *Nature* 2020; **580**(7803): 381-5.
- 700 72. Candia J, Daya GN, Tanaka T, Ferrucci L, Walker KA. Assessment of variability in
- 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

- antibody-based assays for protein quantification in chronic kidney disease. *CJASN* 2022; **17**(3): 350-60.
- 705 74. Pietzner M, Wheeler E, Carrasco-Zanini J, et al. Synergistic insights into human
- 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-
- 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.

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

	Mean (SD) or percentage <sup>a</sup>			No. of significant associations b		
Characteristics	Female (1,243)	Male ( 755)	All (1,998)	Female	Male	All
Demographics						
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
Lifestyle habits						
Regular alcohol drinker	2.6	37.2	25.3	2	85	99
Current smoker	2.3	63.3	15.3	5	24	36
Diet						
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
Environmental						
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
Health and wellbeing						
Prior physical health status						
Self-rated health	8.5	8.2	8.3	0	0	1
Respiratory disease	8.3	8.0	8.2	0	0	1
Kidnev/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	.10	257
Cancer	0.6	0.6	0.7	14	27	10
Mental wellbeing	0.0	010	0.1			
Life satisfaction	37	4 9	4 0	1	1	0
Mental disorder	1 1	1.5	12	7	9	3
Clinical measurements					0	U U
BMI kg/m <sup>2</sup>	24.0 (3.5)	23 7 (3 3)	23 9 (3 4)	595	498	1035
Standing beight cm	154 5 (6.1)	165 8 (6 5)	158 7 (8 3)	3	-56	22
SBP mmHa	129 4 (22 2)	132 6 (19 9)	130 5 (21 4)	136	87	293
DBP mmHa	77.2 (10.6)	79 7 (11.6)	78.0 (11.1)	60	74	233
Heart rate hom	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)	75(23)	1	20	12
EEV1/EVC ratio	85.1 ( 6.1)	84.9 (10.1)	85.0 (8.5)	0	20	2
RPG_mmol/l	61(82)	5 9 (8 8)	6.0 (8.5)	251	77	3/3
Easting time, hours	5.2 (5.0)	5.0 (5.0)	5.1 (5.0)	251	21	86
Penroductive factors	0.2 (0.0)	0.0 (0.0)	0.1 (0.0)	-10	21	00
Ago at monarcho, voars	15 4 (2 0)		15 4 (2 0)	٥		0
	30 2 (A 2)	-	30 2 (1 3)	1	-	1
Age at menopause, years	53.2 (4.3)	_	53.2 (4.3)	59	_	59
Parity	04.7 00.9	-	04.7 00.9	00 20	-	00 20
A de at first live birth years	22 0 (2 2)	-	0.55 23 0 (2 3)	29	-	29
Composite scores	23.8 (3.3)	-	23.8 (3.3)	U	-	U
Lifestyle index f	2 1 (0 0)	2 2 (1 0)	2 0 (1 0)	106	06	207
	0.1 (0.0)	2.2 (1.0)	2.0 (1.0)	100	90	307
Fraility muex 9	0.1 (0.06)	0.1 (0.06)	0.1 (0.06)	220	ŏΖ	400

<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

c 6-point index of qualitative measures of living standards

d 24-point index of frequency of intake in 12 food groups

e Self-reported and screen detected

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

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



sex







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