Integrative proteogenomics study identifies the indirect effect of circulating proteins on 8 diseases via telomere length

Shifang Li, Meijiao Gong

Laboratory of Immunology and Vaccinology, FARAH, ULiège, Liège 4000, Belgium.

Correspondence:

Shifang Li, fruceslee@gmail.com

Laboratory of Immunology and Vaccinology, FARAH, ULiège, Liège 4000, Belgium

Abstract

Growing evidence has revealed the associations between telomere length and diseases; however, the mechanisms behind these links are not fully understood. In this study, by applying proteome-wide Mendelian randomisation (MR), multiple sensitivity analyses, colocalization, single-cell RNA-sequencing, and mediation analysis, the potential mechanisms that link 4,907 plasma proteins, telomere length, and 8 diseases were identified. Following MR analysis for the effects of plasma protein levels on telomere length, 34 proteins were found to be causally related to telomere length. Among these, 8 proteins (PSMB4, PARP1, GDI2, MAX, GMPR2, ARPC1B, ATOX1, and NUDT5) were found to be well colocalized with telomere length (posterior probability >80%). 21 mediation pairs were revealed for the indirect effect of circulating proteins on 8 diseases via telomere length. Furthermore, 35 proteins, including CD8A, were found to be influenced by telomere length among 4,907 proteins. No significant mediation pairs for circulating proteins that mediate the effects of telomere length on disease have been identified. Overall, our study provided insights into understanding the biology of telomeres and prioritized the identified proteins as prospective intervention targets for the disease.

Introduction

Telomeres are 6-nucleotide repeats that are found at the ends of chromosomes and are normally reduced as cells replicate and age. When telomeres reach a critical length, cells enter senescence; thus, telomere shortening and malfunction are one of the major characteristics of aging [1-2]. At the population level, there is significant individual variation in mean telomere length (TL), which is typically assessed in

leukocytes but also in other tissues (eg., testis and cerebellum) [2-3]. Based on the findings of epidemiological studies and Mendelian randomisation (MR) analyses, researchers found out that several diseases, including melanoma, prostate cancer, thyroid, uterine fibroid, multiple sclerosis, idiopathic pulmonary fibrosis (IPF), coronary artery disease, and hypothyroidism are associated with telomere dysfunction; however, the underlying mechanism between these associations is not fully understood [3-8]. Identifying circulating proteins that mediate these associations or determining the indirect influence of circulating proteins on diseases via TL is an approach for disentangling this association. As circulating proteins in the blood may be evaluated and, in some cases, manipulated, identifying related proteins may provide insights into developing probable therapeutic targets [9].

MR is a classical approach in genetic epidemiology research that employs single nucleotide polymorphisms (SNPs) related to the exposure variable under inquiry. Using genetic polymorphisms randomly assigned at conception as an instrumental variable (IV) for exposure, it is feasible to obtain estimates that are less susceptible to the influence of environmental confounders and reverse causation [10-11]. Researchers are currently using this strategy to uncover therapies for various diseases, including stroke, idiopathic pulmonary fibrosis (IPF), COVID-19, and ischemic heart disease [9,12-14]. In addition, by performing MR and mediation analysis, Yoshiji *et al*. have discovered nephronectin as an actionable mediator of the effect of body mass index (BMI) on COVID-19 and COL6A3 as a mediator of the effect of obesity on coronary artery disease [9,15]. Similarly, Chen *et al*. demonstrated that BMI, white matter hyperintensity, and atrial fibrillation appear to mediate the TFPI, IL6RA, and TMPRSS5 correlations with stroke [13]. The findings from these studies suggest that an MR-based strategy represents an intriguing strategy for illuminating the potential mechanisms between risk factors and disease.

In this study, we utilized proteome-wide MR using large-scale GWASs of plasma protein, multiple sensitivity analyses, colocalization, single-cell RNA-sequencing (scRNA-seq) analysis, and mediation analysis to identify the indirect effect of circulating proteins on 8 diseases via TL and uncover circulating proteins mediating

61 the effects of TL on disease.

62 **Methods**

63 **Data source for MR analysis**

64 For MR analysis, publicly available GWAS summary data were applied. We
65 selected the most complete GWAS for plasma proteins, which includes genome-wide
66 associations of genetic variants for 4,907 circulating proteins in 35,559 Icelanders
67 [16]. Prostate cancer (access ID: GCST90274714), uterine fibroids (access ID:
68 GCST009158), hypothyroidism (access ID: GCST90204167), coronary artery disease
69 (access ID: GCST005194), melanoma (access ID: GCST90011809), and thyroid
70 cancer (access ID: GCST90011813) summary statistics were retrieved from the
71 GWAS Catalog (https://www.ebi.ac.uk/gwas/). Prostate cancer included 122,188 cases
72 and 604,640 controls of European ancestry [17], uterine fibroids comprised 20,406
73 cases and 223,918 controls of European ancestry [18], hypothyroidism included
74 51,194 cases and 443,383 controls of European ancestry [19], coronary artery disease
75 comprised 34,541 cases and 261,984 controls of European ancestry [20], melanoma
76 included 6,777 cases and 410,350 controls of European ancestry [21], and thyroid
77 cancer included 762 cases and 410,350 controls of European ancestry [21]. The
78 GWAS summary statistics for multiple sclerosis (access ID: ieu-b-18) [22], TL (access
79 ID: ieu-b-4879) [8], and 102 CD8+ T cells subpopulations [23]
80 (**Supplementary Table 1**) were downloaded from the ieu open gwas project
81 (https://gwas.mrcieu.ac.uk/datasets/). The summary statistics of the large GWAS of
82 IPF were applied from https://github.com/genomicsITER/PFgenetics [24]. Details of
83 the included GWASs are summarized in **Supplementary Table 1**.

84 **IVs selection**

85 To satisfy the MR assumptions, the SNPs chosen as IVs had to show strong
86 associations with exposure. To identify SNPs strongly linked to all symptoms, a
87 genome-wide significance criterion of p<5e-08 was utilized. SNPs in the MHC region
88 were removed as well in the case of the complex linkage disequilibrium (LD)
89 structure of the MHC locus. To address bias caused by LD, we implemented SNP
90 clumping with parameters kb=10,000 and r^2=0.001, as described previously [25].

91 Furthermore, we removed palindrome SNPs and used the F-statistic to measure the
92 strength of each SNP. SNPs with F-statistics greater than 10 were deemed sufficiently
93 powerful to serve as IVs for the trait.

94 **MR and colocalization analysis**

95 The MR analysis was carried out as previously described [25]. In brief, the
96 inverse-variance weighting approach (IVs≥2) and Wald ratio (IV=1) were chosen as
97 the primary analysis methods in the present study. Sensitivity analyses were
98 conducted to ensure the robustness of the findings and to resolve any potential
99 breaches of the MR assumptions. Cochran's Q test was employed to evaluate SNP
100 heterogeneity in both the IVW and MR Egger methods. In a heterogeneity test, I^2
101 statistics were estimated using the "Isq()" function and heterogeneity *p*-value with the
102 "mr_heterogeneity()" function. Following that, a "leave one out" analysis was
103 executed to assess the impact of particular SNPs on the assessment of causal effects.
104 All MR analysis were performed in R version 4.2.2 using the "TwoSampleMR"
105 package [26]. The Benjamini-Hochberg multiple testing was used to account for
106 multiple tests. Colocalization analysis was used to assess the possibility of two traits
107 sharing a shared causal variant, as previously stated [25, 27]. A posterior probability
108 for H4 (PH4) of more than 80% was considered significant colocalization evidence.

109 **Mediation analysis**

110 To evaluate the causal mediation effects, network MR with a product of
111 coefficients method was employed, as described by Yoshiji and others [9, 15]. In brief,
112 we estimated the effect of plasma protein levels on TL ($\beta_{protein\text{-}to\text{-}TL}$) and the effect of
113 TL on diseases ($\beta_{TL\text{-}to\text{-}disease}$), and then multiplied these values
114 ($\beta_{mediated}=\beta_{protein\text{-}to\text{-}TL}\times\beta_{TL\text{-}to\text{-}disease}$). Subsequently, we divided $\beta_{mediated}$ by β_{total}
115 ($\beta_{protein\text{-}to\text{-}disease}$) to estimate the proportion mediated.

116 **Single-cell RNA sequencing analysis**

117 The GSE138266 [28] and GSE174072 [29] datasets were obtained from the
118 NCBI Gene Expression Omnibus (GEO) and analyzed using the R software.
119 GSE138266 includes 12 scRNA-seq datasets from cerebrospinal fluid (CSF) in 6
120 multiple sclerosis patients and 6 control groups. The raw data of the gene expression

matrix was taken from GEO and turned into a Seurat object for downstream processing, as previously indicated [30]. In brief, doublets were detected using the "scrub.scrub_doublets()" function in Scrublet package (v2.0.2) [31] and further removed. The Harmony R package v1.0 was used for addressing batch effects after normalizing the data with log transformation and scaling [32]. The "FindClusters" function was used to cluster cells, which were then annotated according to the original literature [28, 29]. Gene sets (PSMB4, PARP1, GDI2, MAX, GMPR2, ARPC1B, ATOX1, and NUDT5) for each cell type were used for scoring, and cell scoring was performed by the "AddModuleScore" function in Seurat (v4.2.0).

Results and discussion

Identification of causal plasma proteins to TL

First, we implemented two-sample MR to determine the causal effects of plasma proteins on TL outcomes. For the purpose of this analysis, *cis*-acting protein quantitative loci (*cis*-pQTLs, pQTLs that are present inside a 1 Mb region around the gene body) from 4,907 proteins were utilized, reducing the probability of directing horizontal pleiotropy [9,12,14]. Following the *cis*-pQTL search and data harmonization, 1,428 proteins were evaluated in MR for their estimated causal effects on TL (**Figure 1A**). The F-statistics for all of the examined proteins were all greater than 10, minimizing the possibility of weak instrument bias significantly (**Supplementary Table 1**). After applying the Benjamini-Hochberg multiple testing correction (FDR<0.05) and filtering procedure (**Figure 1A**), MR revealed 34 causal proteins, including 17 positive and 18 negative associations (**Figure 1B**). For example, one standard deviation (SD) increase in genetically predicted GDI2 levels was associated with decreased TL (odds ratios (OR)=0.84, 95% confidence intervals (CI) 0.80-0.87, *p*=2.08e-15). Similarly, a one standard deviation rise in genetically predicted ARPC1B levels was linked to lower levels of TL (OR=0.88, 95%CI 0.85-0.93, *p*=7.78e-07). Higher genetically predicted levels of NUDT5 (OR[95%CI]=1.12[1.06,1.18]; *p*=3.12e-05) and PARP1 (OR[95%CI]=1.88[1.72, 2.06]; *p*=1.61e-44) have been associated with increased levels of TL. The MR-Steiger test was used to investigate potential bias from reverse causation, which supported a

151 causal direction of these 34 plasma protein levels influencing TL outcomes
152 (**Supplemental Table 1**). The 18 proteins found to be negatively linked to TL were
153 involved in the proteasome core complex, NIK/NF-kappaB signaling, Wnt signaling
154 pathway, and interleukin-1-mediated signaling network (**Figure 1C**). Furthermore, 17
155 proteins positively related to TL were shown to be enriched in cellular response to
156 starvation, RNA helicase activity, mitochondrial DNA metabolic process, and
157 antifungal humoral response (p.adj<0.05) (**Figure 1C**). Colocalization analyses were
158 then performed to assess the likelihood that protein and TL share the same causal
159 variants rather than being shared accidentally due to LD associations (**Figure 1D-1E**).
160 Interestingly, 8 proteins (PSMB4, PARP1, GDI2, MAX, GMPR2, ARPC1B, ATOX1,
161 and NUDT5) were well colocalized with TL, with posterior probability greater than
162 80% for a shared signal. Two proteins (NRBP1 and BTN3A1) have a medium
163 posterior probability of more than 70%. The other 25 proteins had a poor
164 colocalization posterior probability with TL, indicating that these interactions may be
165 impacted by LD.
166 To validate the assumption of a lack of directional pleiotropy, which can
167 reintroduce confounding, we utilized the PhenoScanner
168 (http://www.phenoscanner.medschl.cam.ac.uk/) databases to find out whether the
169 *cis*-pQTLs for 8 proteins were associated with any traits or diseases at the
170 genome-wide significant threshold of p<5e-08. Except for MAX and GMPR2, no
171 genome-wide significant association was detected for six proteins (PSMB4, PARP1,
172 GDI2, ARPC1B, ATOX1, and NUDT5). The deCODE study's lead *cis*-pQTL for
173 MAX (rs8181938) and GMPR2 (rs34354104) were associated with body-related traits
174 (termed as sub-associations), such as height, forced vital capacity, mean corpuscular
175 hemoglobin, and mean corpuscular volume, among others (**Supplementary Table 1**);
176 yet, since *cis*-acting protein quantitative loci were utilized in our study, these
177 associations may represent a case of vertical pleiotropy, minimizing the probability of
178 directional pleiotropy. Indeed, when employing *cis*-pQTL for MAX (rs8181938) and
179 GMPR2 (rs34354104) from the deCODE study as exposure and these sub-association
180 as outcomes, MR analysis indicated that both MAX and GMPR2 levels were

181 supposed to affect these sub-associations with no reverse causation (p<4.2e-08,
182 MR-Steiger test), suggesting that these sub-associations may be a case of vertical
183 pleiotropy, which does not bias MR interpretation (**Supplemental Table 1**).

184 To determine whether the 8 causative proteins had any cell type-specific
185 enrichment in whole blood leukocytes, we performed scRNA-seq analysis from GEO
186 (GSE174072) [29], which included 62,164 cells from 12 healthy individuals (**Figure**
187 **1F**). 14 cell types were annotated based on the metadata of cells in the origin paper
188 [29]. Subsequently, we compared module scores across cell populations to uncover a
189 probable cell population linked to these 8 causative proteins (**Figure 1G**). The module
190 score of proliferating lymphocytes, dendritic cells (DCs), and plasmacytoid DCs was
191 found to be considerably higher than the other cell populations (Mann-Whitney U test,
192 p<0.001), implying that these cell types may be responsible for the local production of
193 the 8 proteins.

194 **Mediation effect of proteins on 8 diseases via TL**

195 Previous study has demonstrated that TL has a causative effect on various
196 diseases, including an increased risk of melanoma, prostate cancer, thyroid cancer,
197 uterine fibroid, multiple sclerosis, and idiopathic pulmonary fibrosis with each
198 standard deviation increase in TL [8]. In contrast, each standard deviation increase in
199 TL was related to a lower risk of coronary artery disease and hypothyroidism [8].
200 However, the mechanisms behind these associations were primarily unresolved. To
201 investigate this, two hypotheses have been proposed: (1) TL may participate in protein
202 mediation of diseases; and (2) proteins mediate the influence of TL on diseases
203 (discussed below). To test the first hypothesis, a two-sample MR was used to
204 investigate the causal effect of 34 proteins related to TL on 8 diseases (**Supplemental**
205 **Table 1**). Following MR analysis, sensitivity analysis identified 33 substantially
206 associated associations (two associations had been removed due to failure to pass
207 leave-one-out analysis) (**Figure 1H**). Except for BTN3A1, which appears to be shared
208 as a causative variant with prostate cancer (PH4=0.807), there is no strong
209 colocalization evidence for these significant associations. Interestingly, genetically
210 higher levels of PARP1 were related to an increased incidence of melanoma

211 (OR[95%CI]: 4.37[1.96, 9.71]; p=0.00029), prostate cancer (OR[95%CI]: 1.43[1.14,
212 1.81]; p=0.0021), and multiple sclerosis (OR[95%CI]: 4.09[2.009, 8.34]; p=0.0001).
213 Furthermore, ATOX1 was found to be genetically linked to a decreased risk of
214 multiple sclerosis. Using sing-cell RNA sequencing of cerebrospinal fluid from
215 multiple sclerosis patients and control groups (GEO, GSE138266) [28], it was found
216 that PARP1 expression in B cells and plasma cells was higher in multiple sclerosis
217 patients than in control groups (p<0.05) (**Figure S2A-S2D**). In contrast, ATOX1
218 expression in mDCs was lower in the cerebrospinal fluid of multiple sclerosis patients
219 than in control groups (p=0.03). These findings suggest that PARP1 and ATOX1 may
220 represent potential therapeutic targets for multiple sclerosis.

221 Subsequently, we performed a mediation study utilizing the effect estimates from
222 two-step MR and the overall effect from primary MR to evaluate the indirect effect of
223 proteins on 8 diseases via TL (**Figure 1I**). There were a total of 21 mediation
224 pairs revealed (**Supplemental Table 1**). For example, the proportion of PARP1's TL
225 mediation effect on melanoma, prostate cancer, and multiple sclerosis is 19.1%,
226 53.2%, and 26.6%, respectively. Similarly, with values of 11%, 20.4%, and 25.6%, TL
227 appears to mediate the effect of GDI2 on hypothyroidism, prostate cancer, and uterine
228 fibroids, respectively. The indirect effect of VAT1 on IPF risk via TL accounts for
229 32% of the total. Similarly, the proportion of TMEM106A's mediation effect on IPF
230 via TL is approximately one-quarter of the entire effect (28%) (**Figure 1I**).

231 **Identification of the causal protein affected by TL**

232 Following that, we used two-sample MR to quantify the effect of TL on
233 circulating protein levels on a proteome-wide scale (**Figure 2A**). The F-statistic,
234 which measures the strength of the correlation between genetic variations and TL,
235 was 119.6, indicating that there was no evidence of instrument bias. Out of the 4,907
236 proteins examined (**Supplemental Table 1**), 40 were found to be influenced by TL
237 (FDR<0.05), suggesting that TL has a moderate effect on plasma protein levels. After
238 that, we performed heterogeneity test, directional pleiotropy test, and
239 reverse-causation test to keep only the proteins that had been significantly affected by
240 TL. We found no significant heterogeneity in 40 proteins (I^2 <50% for all). The

241 MR-Egger intercept test revealed no evidence of directed horizontal pleiotropy in 35
242 of the 40 proteins tested ($P_{\text{Egger intercept}}$>0.05). There was no evidence of reverse
243 causation using protein levels as the exposures and TL as the outcome. Importantly,
244 according to the leave-one-out analysis, these 35 associations were not driven by
245 outliers. Therefore, a total of 35 protein levels were identified as causal proteins
246 impacted by TL, with no apparent heterogeneity, directional pleiotropy, or reverse
247 causation (**Figure 2B-2C**). Among the 35 causal proteins, 24 positively proteins
248 related to TL were positively associated with antimicrobial humoral response
249 (*p*.adj=0.00035) and positive regulation of wound healing (*p*.adj=0.0038), while 11
250 negatively proteins associated with TL were negatively associated with cellular
251 protein metabolic process (*p*.adj=0.018) and negative regulation of IRE1-mediated
252 unfolded protein response (*p*.adj=0.018) (**Figure 2D**). Interestingly, among the 35
253 TL-driven proteins studied, higher levels of TL were shown to be related to higher
254 levels of plasma CD8A (OR[95%CI]: 1.25[1.15, 1.36]; *p*=1.15e-07) (**Figure 2C**).
255 Previous research has shown that the number of CD8+ T cells declines linearly with
256 age, which is consistent with this finding [33-34]. To figure out whether
257 the subpopulation of CD8+ T cells was influenced by TL, MR analysis was carried
258 out on 102 cell populations of CD8 T cells from a cohort of 3,757 Sardinians [23].
259 Interestingly, following multiple testing corrections (FDR<0.05), CD8+ Natural Killer
260 T cells, Naive CD8+ T cells, and CD28+ CD45RA+ CD8+ T cells were essentially
261 affected by TL (**Figure 2E**). For these causality associations, MR analyses using the
262 weighted median, weighted mode, and MR-Egger slope approaches produced
263 directionally consistent results, as accomplished by the inverse variance weighted
264 method (**Supplemental Table 1**).

265 **Causal effect of TL-driven proteins on 8 diseases**

266 To test the second hypothesis, that proteins mediate the impact of TL on diseases,
267 a two-sample MR was initially performed to evaluate the causal effect of 35
268 TL-driven proteins on 8 diseases. Following the *cis*-pQTL search of 35 proteins and
269 data harmonization, 27 proteins were examined in the MR analysis. After MR analysis
270 and procedure filtering, no statistical significance was observed for 27 proteins on

diseases on a level of Benjamini-Hochberg multiple testing correction. However, 4 associations were discovered at the nominally significant (uncorrected $p<0.05$) level, as validated by sensitivity and leave-one-out analysis (**Figure 2F**). It was discovered that increasing blood ST3GAL6 expression by one SD was related to an increased incidence of hypothyroidism (OR: 1.058; 95%CI: 1.015-1.103; $p=0.0076$). Per 1-SD rise in blood SPOCK3 and ELANE expression, the incidence of multiple sclerosis was reduced, with ORs (95%CI) of 0.79 (95%CI: 0.67-0.94; $p=0.0092$) and 0.502 (95%CI: 0.268-0.938, $p=0.030$), respectively. Furthermore, each 1-SD increase in blood ST3GAL6 expression was linked to a lower incidence of hypothyroidism (OR: 1.058; 95%CI: 1.015-1.103; $p=0.0075$). However, there was no evidence that these 4 TL-driven proteins mediated the effect of TL on individual diseases due to the opposite direction of the total effect ($\beta_{TL\text{-}to\text{-}disease}$) with the mediated effect ($\beta_{TL\text{-}to\text{-}protein} \times \beta_{protein\text{-}to\text{-}disease}$) (**Supplemental Table 1**).

Overall, we conducted a large-scale MR investigation that sought underlying mechanisms that link plasma proteins, TL, and 8 diseases (**Figure 3**). PSMB4, PARP1, GDI2, MAX, GMPR2, ARPC1B, ATOX1, and NUDT5 were found to be genetically linked to TL. Of interest, 21 pairs have revealed that the length of telomeres mediates the causative effect of circulating proteins on diseases. These findings provide insights into telomere biology and prioritize the identified proteins as potential intervention targets for the disease. Future research with a larger and more distinct genetic background should be used to validate these results.

Acknowledgments

The authors thank all of the researchers from the related genetic consortia and GWAS for making the data available to the public.

Disclosure

All the authors declared no competing interests.

Contributions

SF: study conception and design, data curation and analysis, and writing and reviewing of the manuscript. MJ: data curation, as well as manuscript preparation and review.

301 **References**

302 1 Allsopp RC, Vaziri H, Patterson C, et al. Telomere length predicts replicative
303 capacity of human fibroblasts. ***Proc Natl Acad Sci U S A*** 1992;89:10114-8.
304 doi:10.1073/pnas.89.21.10114

305 2 Demanelis K, Jasmine F, Chen LS, et al. Determinants of telomere length across
306 human tissues. ***Science*** 2020;369. doi:10.1126/science.aaz6876

307 3 Chakravarti D, LaBella KA, DePinho RA. Telomeres: history, health, and
308 hallmarks of aging. ***Cell*** 2021;184:306-22. doi:10.1016/j.cell.2020.12.028

309 4 Tsakiri KD, Cronkhite JT, Kuan PJ, et al. Adult-onset pulmonary fibrosis caused
310 by mutations in telomerase. ***Proc Natl Acad Sci U S A*** 2007;104:7552-7.
311 doi:10.1073/pnas.0701009104

312 5 Duckworth A, Gibbons MA, Allen RJ, et al. Telomere length and risk of
313 idiopathic pulmonary fibrosis and chronic obstructive pulmonary disease: a mendelian
314 randomisation study. ***Lancet Respir Med*** 2021;9:285-94.
315 doi:10.1016/S2213-2600(20)30364-7

316 6 Li C, Stoma S, Lotta LA, et al. Genome-wide Association Analysis in Humans
317 Links Nucleotide Metabolism to Leukocyte Telomere Length. ***Am J Hum Genet***
318 2020;106:389-404. doi:10.1016/j.ajhg.2020.02.006

319 7 Taub MA, Conomos MP, Keener R, et al. Genetic determinants of telomere
320 length from 109,122 ancestrally diverse whole-genome sequences in TOPMed. ***Cell***
321 ***Genomics*** 2022;2. doi:10.1016/j.xgen.2021.100084

322 8 Codd V, Wang Q, Allara E, et al. Polygenic basis and biomedical consequences of
323 telomere length variation. ***Nat Genet*** 2021;53:1425-33.
324 doi:10.1038/s41588-021-00944-6

325 9 Yoshiji S, Butler-Laporte G, Lu T, et al. Proteome-wide Mendelian randomization
326 implicates nephronectin as an actionable mediator of the effect of obesity on
327 COVID-19 severity. ***Nat Metab*** 2023;5:248-64. doi:10.1038/s42255-023-00742-w

328 10 Sekula P, Del Greco FM, Pattaro C, et al. Mendelian randomization as an
329 approach to assess causality using observational data. ***J Am Soc Nephrol***
330 2016;27:3253-65. doi:10.1681/ASN.2016010098

11 Skrivankova VW, Richmond RC, Woolf BAR, et al. Strengthening the reporting of observational studies in epidemiology using mendelian randomisation (STROBE-MR): Explanation and elaboration. ***BMJ*** 2021;375. doi:10.1136/bmj.n2233

12 Nakanishi T, Cerani A, Forgetta V, et al. Genetically increased circulating FUT3 level leads to reduced risk of idiopathic pulmonary fibrosis: A Mendelian randomisation study. ***Eur Respir J*** 2022;59. doi:10.1183/13993003.03979-2020

13 Chen L, Peters JE, Prins B, et al. Systematic Mendelian randomization using the human plasma proteome to discover potential therapeutic targets for stroke. ***Nat Commun*** 2022;13:1-14. doi:10.1038/s41467-022-33675-1

14 Mazidi M, Wright N, Yao P, et al. Plasma Proteomics to Identify Drug Targets for Ischemic Heart Disease. ***J Am Coll Cardiol*** 2023;82:1906-20. doi:10.1016/j.jacc.2023.09.804

15 Yoshiji S, Lu TY, Butler-Laporte G, et al. COL6A3-derived endotrophin mediates the effect of obesity on coronary artery disease:an integrative proteogenomics analysis. medRxiv 2023. doi: 10.1101/2023.04.19.23288706.

16 Ferkingstad E, Sulem P, Atlason BA, et al. Large-scale integration of the plasma proteome with genetics and disease. ***Nat Genet*** 2021;53:1712-21. doi:10.1038/s41588-021-00978-w.

17 Wang A, Shen J, Rodriguez AA, et al. Characterizing prostate cancer risk through multi-ancestry genome-wide discovery of 187 novel risk variants. ***Nat Genet*** Published Online First: 2023. doi:10.1038/s41588-023-01534-4

18 Gallagher CS, Mäkinen N, Harris HR, et al. Genome-wide association and epidemiological analyses reveal common genetic origins between uterine leiomyomata and endometriosis. ***Nat Commun*** 2019;10:1-11. doi:10.1038/s41467-019-12536-4

19 Mathieu S, Briend M, Abner E, et al. Genetic association and Mendelian randomization for hypothyroidism highlight immune molecular mechanisms. ***iScience*** 2022;25. doi:10.1016/j.isci.2022.104992

20 Van Der Harst P, Verweij N. Identification of 64 novel genetic loci provides an

361 expanded view on the genetic architecture of coronary artery disease. ***Circ Res***
362 2018;122:433-43. doi:10.1161/CIRCRESAHA.117.312086
363 21 Rashkin SR, Graff RE, Kachuri L, et al. Pan-cancer study detects genetic risk
364 variants and shared genetic basis in two large cohorts. ***Nat Commun*** 2020;11.
365 doi:10.1038/s41467-020-18246-6
366 22 International Multiple Sclerosis Genetics Consortium. Multiple sclerosis genomic
367 map implicates peripheral immune cells and microglia in susceptibility. ***Science***
368 2019;365:eaav7188. doi: 10.1126/science.aav7188
369 23 Orrù V, Steri M, Sidore C, et al. Complex genetic signatures in immune cells
370 underlie autoimmunity and inform therapy. ***Nat Genet*** 2020;52:1036-45.
371 doi:10.1038/s41588-020-0684-4
372 24 Allen RJ, Stockwell A, Oldham JM, et al. Genome-wide association study across
373 five cohorts identifies five novel loci associated with idiopathic pulmonary fibrosis.
374 ***Thorax*** 2022;77:829-33. doi:10.1136/thoraxjnl-2021-218577
375 25 Li SF & Gong MJ. Therapeutic targets for haemorrhoidal disease: proteome-wide
376 Mendelian randomisation and colocalization analyses. medRxiv 2023.
377 doi:10.1101/2023.06.19.2329137.
378 26 Hemani G, Zheng J, Elsworth B, et al. The MR-base platform supports systematic
379 causal inference across the human phenome. ***Elife*** 2018;7:1-29.
380 doi:10.7554/eLife.34408
381 27 Giambartolomei C, Vukcevic D, Schadt EE, et al. Bayesian Test for
382 Colocalisation between Pairs of Genetic Association Studies Using Summary
383 Statistics. ***PLoS Genet*** 2014;10. doi:10.1371/journal.pgen.1004383
384 28 Schafflick D, Xu CA, Hartlehnert M, et al. Integrated single cell analysis of blood
385 and cerebrospinal fluid leukocytes in multiple sclerosis. ***Nat Commun*** 2020;11:1-14.
386 doi:10.1038/s41467-019-14118-w
387 29 Wilk AJ, Lee MJ, Wei B, et al. Multi-omic profiling reveals widespread
388 dysregulation of innate immunity and hematopoiesis in COVID-19. ***J Exp Med***
389 2021;218. doi:10.1084/jem.20210582
390 30 Li SF & Gong MJ. Integrative single-cell analysis of epigenomic and

391 transcriptomic states in patients with systemic sclerosis-associated interstitial lung
392 disease. medRxiv 2023. doi: 10.1101/2023.09.05.23294979.
393 31 Wolock SL, Lopez R, Klein AM. Scrublet: Computational Identification of Cell
394 Doublets in Single-Cell Transcriptomic Data. ***Cell Syst*** 2019;8:281-291.e9.
395 doi:10.1016/j.cels.2018.11.005
396 32 Korsunsky I, Millard N, Fan J, et al. Fast, sensitive and accurate integration of
397 single-cell data with harmony. ***Nat Methods*** 2019;16:1289-96.
398 doi:10.1038/s41592-019-0619-0
399 33 Shin MS, Yim K, Moon K, et al. Dissecting alterations in human CD8+ T cells
400 with aging by high-dimensional single cell mass cytometry. ***Clin Immunol***
401 2019;200:24-30. doi:10.1016/j.clim.2019.01.005
402 34 Kumar B V., Connors TJ, Farber DL. Human T Cell Development, Localization,
403 and Function throughout Life. ***Immunity*** 2018;48:202-13.
404 doi:10.1016/j.immuni.2018.01.007

405

406 **Figure Legends**

407 **Figure 1. The causal effect of plasma proteins on TL.** (**A**) Design overview for the
408 investigation of the causal effect of proteins on TL. FDR, false discovery rate; LOOA,
409 leave-one-out analysis. (**B**) Volcano plots of MR results for the causal effect of plasma
410 proteins on TL levels. Labeled proteins are 34 proteins examined in the primary
411 analysis with a significant FDR-corrected *p*-value<0.05. (**C**) Biological function
412 enrichment analysis for the 17 positively and 18 negatively associated TL proteins. (**D**)
413 A boxplot illustrating the results of the co-localization analysis for the 34 proteins
414 described in (B). Eight proteins with high colocalization evidence (PH4>0.8) were
415 labeled. (**E**) The locus-compare scatter plot for the identified proteins' association
416 with TL. The locus-compare scatter plot compares the results of the protein
417 quantitative trait loci (pQTL) and TL genome-wide association study (GWAS),
418 indicating whether the two traits shared the same genetic variants. (**F**) UMAP
419 embeddings of scRNA-seq profiles (GSE174072) from 62,164 blood leukocytes in 12

healthy donors. (**G**) Module scores of gene sets (PSMB4, PARP1, GDI2, MAX, GMPR2, ARPC1B, ATOX1, and NUDT5) in each cell type. The *p*-value was obtained by the Mann-Whitney U test. (**H**) Causal effects of the selected proteins on diseases using MR analysis. The dots are the causal estimates on the OR scale, and the whiskers represent the 95% confidence intervals (CI) for these ORs. OR, odd ratios. (**I**) Mediation effects of selected proteins on diseases via TL. Mediation analyses were used to quantify the effects of 34 proteins on 8 diseases outcomes via TL. The global mediation-pairs were provided in Supplemental Table 1. TL, telomere length.

Figure 2. The causal effect of TL on proteins. (**A**) Design overview for the investigation of the causal effect of TL on proteins. FDR, false discovery rate; LOOA, leave-one-out analysis. (**B**) Volcano plots of MR results for the causal effect of TL on plasma proteins. Labeled proteins are 35 proteins examined in primary analysis with a significant FDR-corrected *p*-value<0.05. (**C**) Scatter plot for the causal effect of TL on the selected protein levels. (**D**) Enrichment analysis of biological functions for the 24 positively and 11 negatively influenced proteins by TL. (**E**) Causal effects of TL on 13 CD8+ T subpopulations. (**F**) Causal effects of the selected proteins on diseases. The dots are the causal estimates on the OR scale, and the whiskers represent the 95% confidence intervals for these ORs. OR, odd ratios. TL, telomere length.

Figure 3. Schematic representation of the results we obtained in the study.

Figure S1 Single-cell transcriptomic atlas of cell types in the cerebrospinal fluid of multiple sclerosis patients and controls. (**A**) UMAP visualization of 33,797 cells in cerebrospinal fluid from controls (n=6) and patients with multiple sclerosis (n=6). (**B**) A dot-plot depicting the expression of marker genes in each cell type. (**C**) Barplots showing the proportion of cell types in each sample. (**D**) The differential expression levels of PARP1 and ATOX1 between controls and MS samples. CSF, cerebrospinal fluid; MS, multiple sclerosis.

450 **Supplementary Table 1. The detailed GWAS information used in our**

451 **investigation, as well as the summary statistics obtained in this study.**

Fig 1

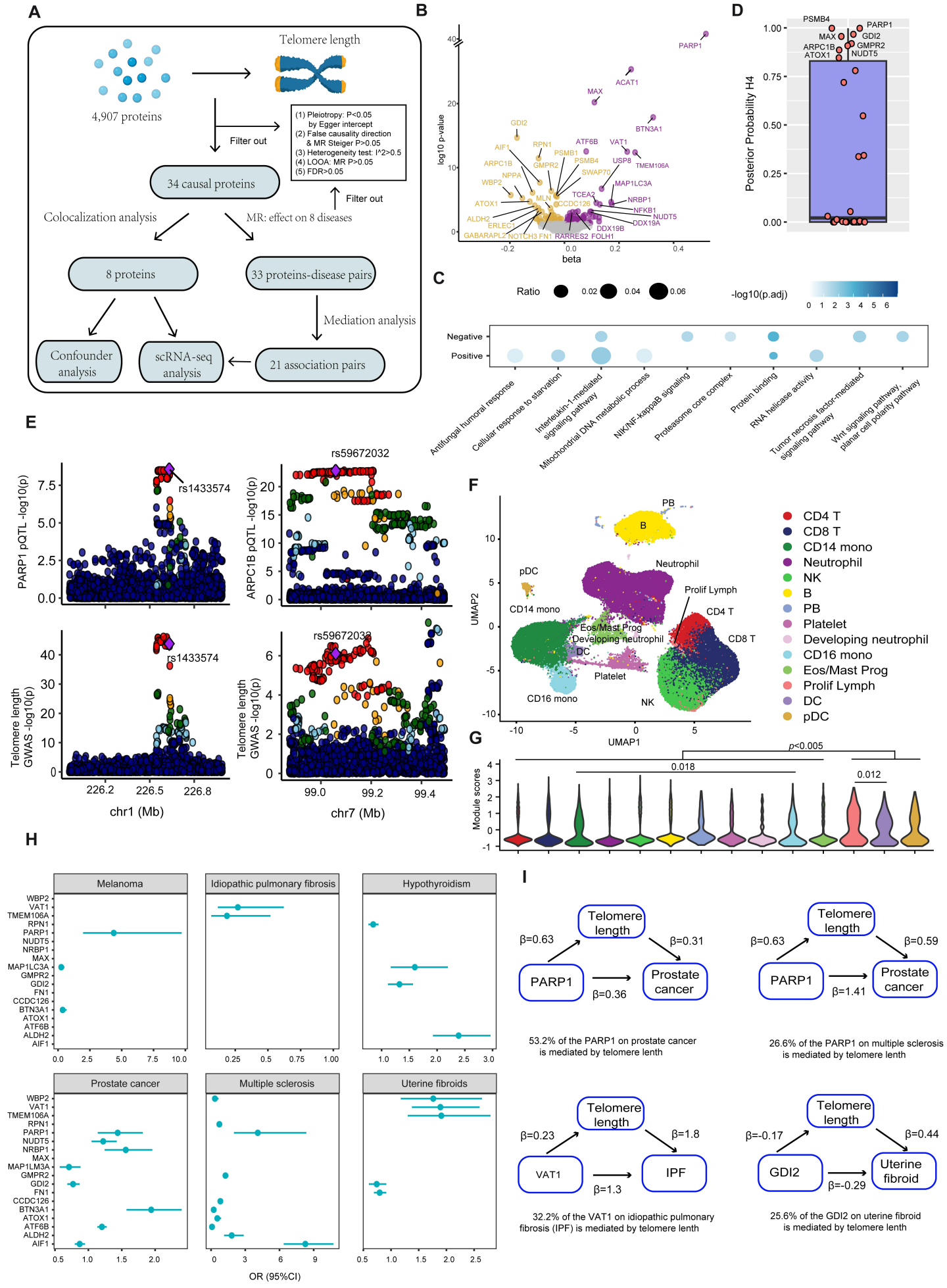

A
Telomere length
4,907 proteins
Filter out
(1) Pleiotropy: P<0.05 by Egger intercept
(2) False causality direction & MR Steiger P>0.05
(3) Heterogeneity test: I^2>0.5
(4) LOOA: MR P>0.05
(5) FDR>0.05
34 causal proteins
Filter out
Colocalization analysis
MR: effect on 8 diseases
8 proteins
33 proteins-disease pairs
Mediation analysis
Confounder analysis
scRNA-seq analysis
21 association pairs
B
log10 p-value
beta
PARP1
ACAT1
MAX
BTN3A1
GDI2
AIF1
RPN1
ATF6B
VAT1
PSMB1
PSMB4
USP8
TMEM106A
ARPC1B
GMPR2
SWAP70
NPPA
WBP2
MAP1LC3A
TCEA2
MLN
CCDC126
NRBP1
ATOX1
NFKB1
ALDH2
NUDT5
ERLEC1
DDX19A
DDX19B
GABARAPL2
NOTCH3
FN1
RARRES2
FOLH1
C
Ratio
0.02
0.04
0.06
-log10(p.adj)
Negative
Positive
Antifungal humoral response
Cellular response to starvation
Interleukin-1-mediated signaling pathway
Mitochondrial DNA metabolic process
NIK/NF-kappaB signaling
Proteasome core complex
Protein binding
RNA helicase activity
Tumor necrosis factor-mediated signaling pathway
Wnt signaling pathway, planar cell polarity pathway
D
Posterior Probability H4
PSMB4
PARP1
MAX
GDI2
ARPC1B
GMPR2
ATOX1
NUDT5
E
PARP1 pQTL -log10(p)
rs1433574
ARPC1B pQTL -log10(p)
rs59672032
Telomere length GWAS -log10(p)
chr1 (Mb)
chr7 (Mb)
F
PB
B
pDC
Neutrophil
Prolif Lymph
CD14 mono
CD4 T
Eos/Mast Prog
Developing neutrophil
CD8 T
DC
Platelet
CD16 mono
NK
UMAP1
UMAP2
CD4 T
CD8 T
CD14 mono
Neutrophil
NK
B
PB
Platelet
Developing neutrophil
CD16 mono
Eos/Mast Prog
Prolif Lymph
DC
pDC
G
p<0.005
0.018
0.012
Module scores
H
Melanoma
Idiopathic pulmonary fibrosis
Hypothyroidism
Prostate cancer
Multiple sclerosis
Uterine fibroids
WBP2
VAT1
TMEM106A
RPN1
PARP1
NUDT5
NRBP1
MAX
MAP1LM3A
GMPR2
GDI2
FN1
CCDC126
BTN3A1
ATOX1
ATF6B
ALDH2
AIF1
OR (95%CI)
I
Telomere length
β=0.63
β=0.31
PARP1
β=0.36
Prostate cancer
53.2% of the PARP1 on prostate cancer is mediated by telomere lenth
Telomere length
β=0.63
β=0.59
PARP1
β=1.41
Prostate cancer
26.6% of the PARP1 on multiple sclerosis is mediated by telomere lenth
Telomere length
β=0.23
β=1.8
VAT1
β=1.3
IPF
32.2% of the VAT1 on idiopathic pulmonary fibrosis (IPF) is mediated by telomere lenth
Telomere length
β=-0.17
β=0.44
GDI2
β=-0.29
Uterine fibroid
25.6% of the GDI2 on uterine fibroid is mediated by telomere lenth

Fig 2

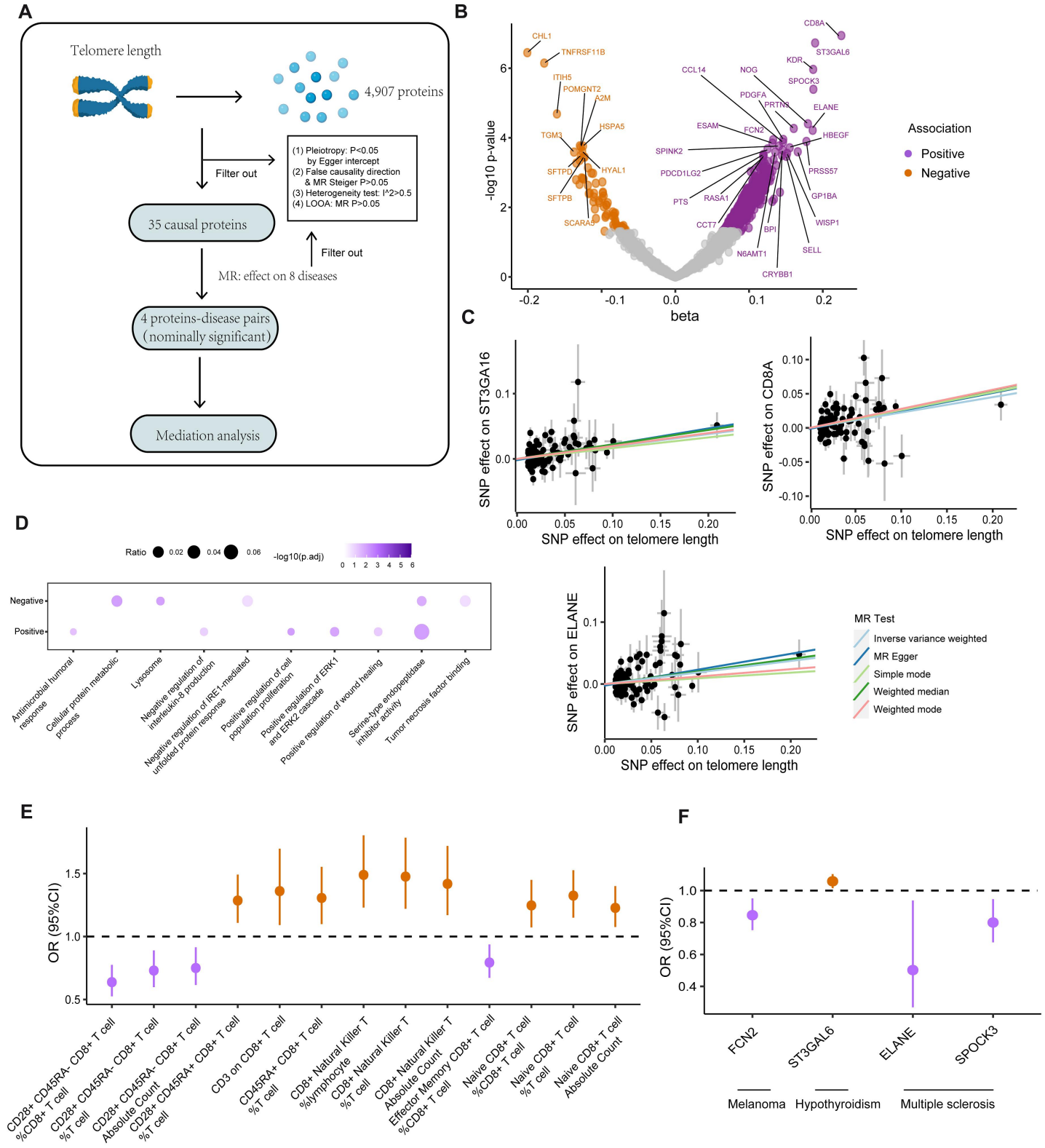

A
Telomere length
4,907 proteins
Filter out
(1) Pleiotropy: P<0.05 by Egger intercept
(2) False causality direction & MR Steiger P>0.05
(3) Heterogeneity test: I^2>0.5
(4) LOOA: MR P>0.05
35 causal proteins
Filter out
MR: effect on 8 diseases
4 proteins-disease pairs (nominally significant)
Mediation analysis
B
-log10 p-value
beta
Association
Positive
Negative
CHL1
TNFRSF11B
ITIH5
POMGNT2
A2M
HSPA5
TGM3
SFTPD
HYAL1
SFTPB
SCARA5
CD8A
ST3GAL6
KDR
SPOCK3
CCL14
NOG
PDGFA
PRTN3
ELANE
ESAM
FCN2
HBEGF
SPINK2
PDCD1LG2
PRSS57
PTS
RASA1
GP1BA
CCT7
BPI
WISP1
N6AMT1
SELL
CRYBB1
C
SNP effect on ST3GA16
SNP effect on CD8A
SNP effect on ELANE
SNP effect on telomere length
MR Test
Inverse variance weighted
MR Egger
Simple mode
Weighted median
Weighted mode
D
Ratio
-log10(p.adj)
Negative
Positive
Antimicrobial humoral response
Cellular protein metabolic process
Lysosome
Negative regulation of interleukin-8 production
Negative regulation of IRE1-mediated unfolded protein response
Positive regulation of cell population proliferation
Positive regulation of ERK1 and ERK2 cascade
Positive regulation of wound healing
Serine-type endopeptidase inhibitor activity
Tumor necrosis factor binding
E
OR (95%CI)
CD28+ CD45RA- CD8+ T cell %CD8+ T cell
CD28+ CD45RA- CD8+ T cell %T cell
CD28+ CD45RA- CD8+ T cell Absolute Count
CD28+ CD45RA+ CD8+ T cell %T cell
CD3 on CD8+ T cell
CD45RA+ CD8+ T cell %T cell
CD8+ Natural Killer T %lymphocyte
CD8+ Natural Killer T %T cell
CD8+ Natural Killer T Absolute Count
Effector Memory CD8+ T cell %CD8+ T cell
Naive CD8+ T cell %CD8+ T cell
Naive CD8+ T cell %T cell
Naive CD8+ T cell Absolute Count
F
OR (95%CI)
FCN2
ST3GAL6
ELANE
SPOCK3
Melanoma
Hypothyroidism
Multiple sclerosis

Fig 3

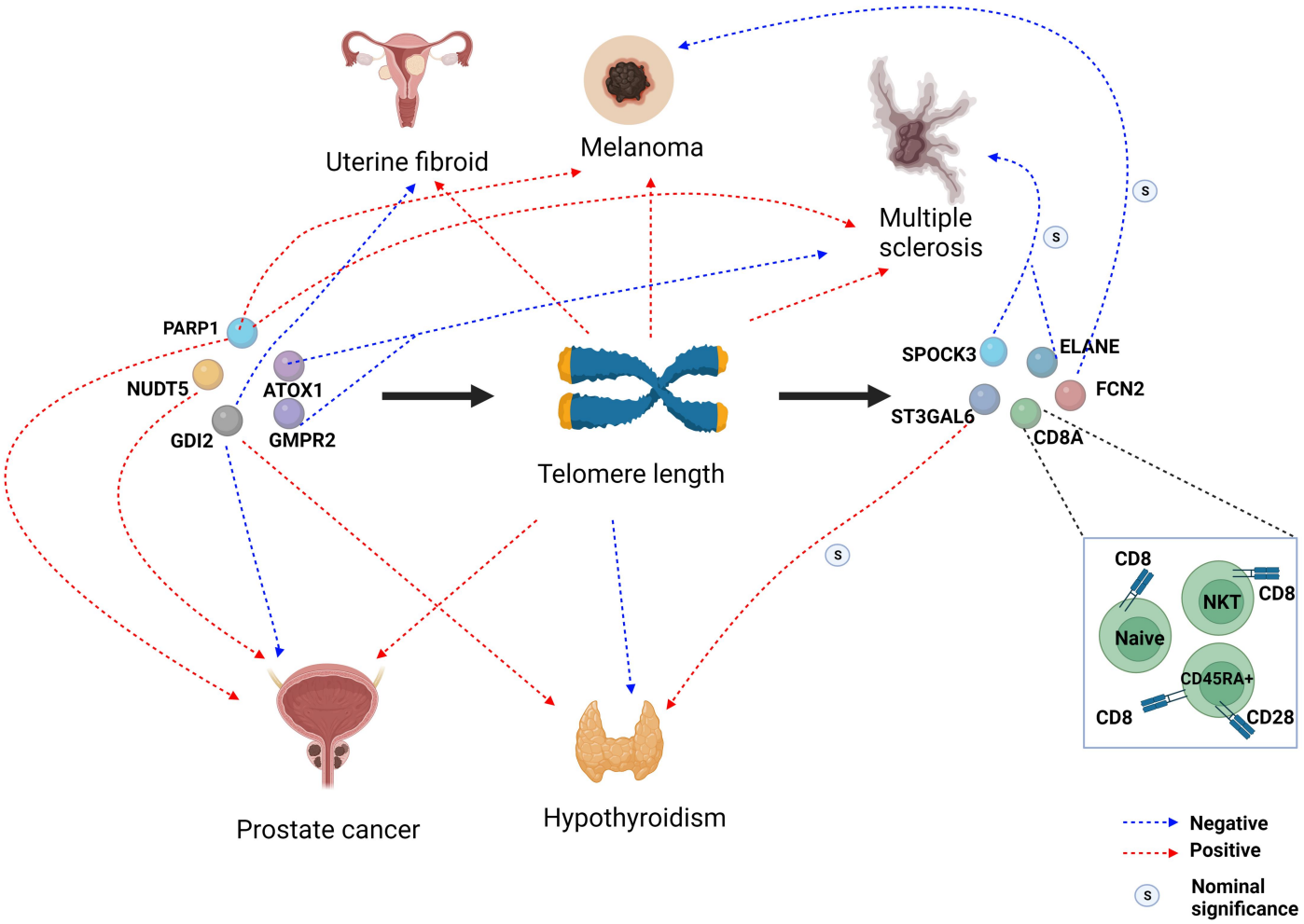
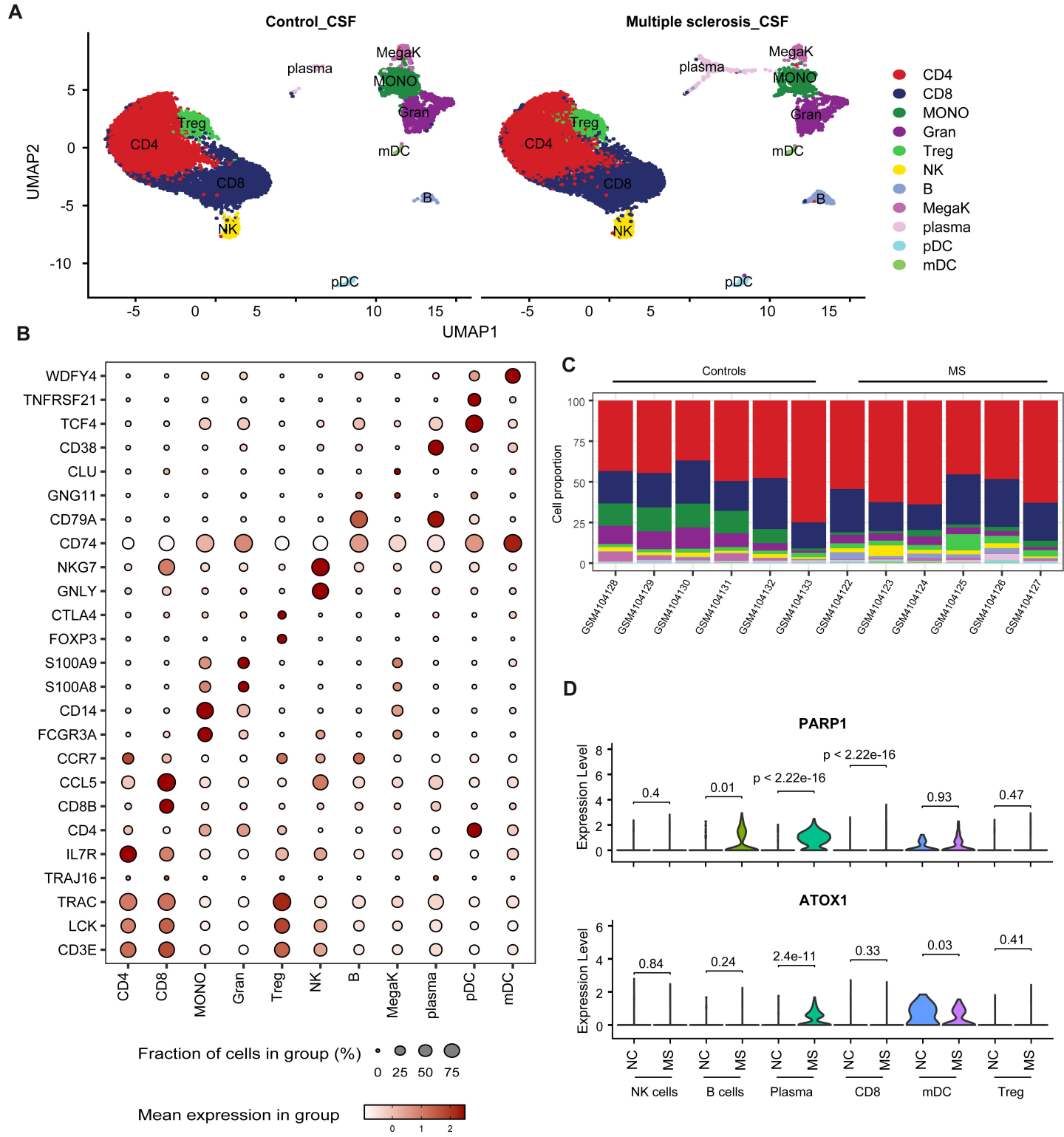

Uterine fibroid
Melanoma
Multiple sclerosis
PARP1
NUDT5
ATOX1
GDI2
GMPR2
Telomere length
SPOCK3
ELANE
FCN2
ST3GAL6
CD8A
S
CD8
NKT
CD8
Naive
CD45RA+
CD8
CD28
Prostate cancer
Hypothyroidism
Negative
Positive
Nominal significance

Fig S1


A
Control_CSF
Multiple sclerosis_CSF
UMAP2
UMAP1
CD4
CD8
Treg
NK
B
MegaK
MONO
Gran
mDC
plasma
pDC
B
WDFY4
TNFRSF21
TCF4
CD38
CLU
GNG11
CD79A
CD74
NKG7
GNLY
CTLA4
FOXP3
S100A9
S100A8
CD14
FCGR3A
CCR7
CCL5
CD8B
CD4
IL7R
TRAJ16
TRAC
LCK
CD3E
Fraction of cells in group (%)
0 25 50 75
Mean expression in group
0 1 2
C
Controls
MS
Cell proportion
GSM4104128
GSM4104129
GSM4104130
GSM4104131
GSM4104132
GSM4104133
GSM4104122
GSM4104123
GSM4104124
GSM4104125
GSM4104126
GSM4104127
D
PARP1
ATOX1
Expression Level
0.4
0.01
p < 2.22e-16
p < 2.22e-16
0.93
0.47
0.84
0.24
2.4e-11
0.33
0.03
0.41
NC
MS
NK cells
B cells
Plasma
CD8
mDC
Treg