Identification of therapeutic targets for 18 clinical diseases by integrating human plasma proteome with genome

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

The proteome is an abundant source of potential therapeutic targets, and a comprehensive evaluation of proteins as therapeutic targets for a wide range of diseases is required. By screening 4,907 plasma proteins, we conducted a systematically two-sample proteome-wide Mendelian randomisation (MR) study to uncover potential therapeutic targets for 18 clinical diseases. Following MR analysis and stringent process filtering, a total of 146 causative plasma proteins (false discovery rate<0.05) were discovered. Colocalization analysis (Posterior Probability H4>0.8) further supported the causality of three proteins (MAP2K1, GFRA1, and THBS3) in gastroesophageal reflux disease; LYPLAL1 in keratoconus; three proteins (PCSK9, ANGPTL4, and GCKR) in familial hyperlipidemia; CRAT in atopic eczema; three proteins (IRF3, CA12, and TNFRSF1B) in hypothyroidism; AIF1 in age-related hearing impairment; SCARF2 in male pattern baldness; IRF3 in basal cell carcinoma; and four proteins (RPS6KA1, ULK3, MPPED2, and BTN3A1) in prostate cancer. Interestingly, having a genetically higher circulating CA12 level is associated with a lower risk of hypothyroidism (OR=0.47, *p*-value=1.68e-05, Posterior Probability H4=0.82). Single-cell RNA sequencing analysis showed that CA12 was mainly expressed in fibroblasts and epithelial cells in patient thyroid tissue and that its expression increased in older adults. Furthermore, with a proportion of 3.8%, hypothyroidism appears to mediate the effect of IRF3 on idiopathic pulmonary fibrosis, according to mediation analysis. Overall, our research could provide new insights into the etiology of clinical diseases as well as intriguing targets for the development of screening biomarkers and therapeutic treatments.

Introduction

Plasma proteins are involved in numerous essential physiological processes, such as immunomodulation, substance transport, signaling, and homeostasis maintenance, and their dysregulation has been frequently reported in a variety of diseases, implying that these dysregulated proteins are involved in disease pathogenesis [1-2]. Given the importance of plasma proteins in disease and the fact that plasma proteins are a primary source of therapeutic targets, identifying disease-related plasma proteins can assist in deciphering disease pathophysiology and providing possible molecular targets for drug development [3-4]. Currently, large-scale proteomics study have uncovered over 18,000 protein quantitative trait loci (pQTLs) spanning over 4,800 proteins, including over 1,800 independent *cis*-pQTLs [5]. Through Mendelian randomisation (MR), these studies provide a significant data resource for systematically understanding the causal impact of plasma proteins on clinical disease. Normally, MR employs naturally randomized genetic diversity at the moment of conception as a natural experiment to uncover the causal relationship between exposure and disease, reducing the risk of reverse causation and confounding bias [6]. Proteome-wide MR has currently revealed important insights into the etiology of stroke, idiopathic pulmonary fibrosis (IPF), and COVID-19, as well as the prioritization of therapeutic targets [7-9]. Indeed, drugs containing human genetic evidence are more likely to succeed in Phase II and Phase III trials, with human genetic data supporting two-thirds of FDA-approved drugs in 2021 [10-12].

In this study, we utilized proteome-wide MR to identify circulating protein biomarkers linked with 18 clinical traits by integrating the human plasma proteome with genetic data. Colocalization analyses are conducted to determine the robustness of the proteins' instrumental variables. Furthermore, we performed single-cell sequencing data analysis and mediation analysis to identify the enriched cell types of candidate proteins in tissues as well as the potential mechanisms mediated between diseases, respectively.

Methods

Proteomic data source

A large-scale protein quantitative trait loci (pQTL) study in 35,559 Icelanders yielded the largest and most thorough genome-wide association studies (GWAS) summary datasets on 4,907 circulating proteins [5]. The detailed description of the datasets can be found in the original study [5]. In brief, the plasma protein levels were determined using the SomaScan version 4 assay from SomaLogic, and the pQTL datasets were made up of associations between genome-wide genetic variations and plasma proteins that were adjusted for age and gender using the BOLT-LMM linear mixed model. As instrumental variables in the following MR analysis, genome-wide significant and independent single nucleotide polymorphisms (SNPs) with *p*-value<5e-08 and r^2<0.001 of pQTLs were employed.

18 clinical traits data sources

To identify the potential therapeutic targets, the GWAS summary statistics from 18 diseases, including male pattern baldness (ID: GCST006661, 36,166 cases and 16,708 controls), gastroesophageal reflux disease (71,522 cases and 261,079 controls), keratoconus (ID: GCST90013442, 2,116 cases and 24,626 controls), selective IgA deficiency disease (ID: GCST003814, 1,635 cases and 4,852 controls), fibromuscular dysplasia (ID: GCST90026612, 1,556 cases and 7,100 controls), age-related hearing impairment (ID: GCST90012115, 125,688 cases and 205,071 controls), familial hyperlipidemia (ID: GCST90104003, 3,838 cases and 345,384 controls), atopic eczema (ID: GCST90027161, 22,474 cases and 774,187 controls), Brugada syndrome (ID: GCST90086158, 2,820 cases and 10,001 controls), cataract (ID: GCST009963, 21,679 cases and 387,283 controls), uterine fibroid (ID: GCST009158, 20,406 cases and 223,918 controls), amyotrophic lateral sclerosis (ID: GCST90027164, 27,205 cases and 110,881 controls), acne (ID: GCST90245818, 34,422 cases and 364,991 controls), posterior urethral valve (ID: GCST90134327, 756 cases and 4,823 controls), hypothyroidism (ID: GCST90204167, 51,194 cases and 443,383 controls), prostate cancer (ID: GCST90274714, 122,188 cases and 604,640 controls), basal cell carcinoma (ID: GCST90013410, 17,416 cases and 375,455 controls), and unipolar depression (ID: GCST006477, 66,809 individuals) were retrieved from the GWAS Catalog (https://www.ebi.ac.uk/gwas/) when an ID was provided (**Supplementary**

Table 1). To the best of our knowledge, the exposure sample dataset (4,907 proteomic data) did not overlap with the outcome (18 clinical traits) we selected. Further information on the outcome dataset is listed in **Supplementary Table 1**. If standard error and effect size are missing but z-scores are included in the GWAS summary statistics file, effect size and standard error can be calculated as follows:

$$effect\ size = \frac{z}{\sqrt{2 * p(1-p) * (n+z^2)}}$$

$$standard\ error = \frac{1}{\sqrt{2 * p(1-p) * (n+z^2)}}$$

where n is the number of individuals used in the association study and p is the *p*-value for the association study.

Mendelian randomisation analysis

Based on the presence of non-overlapping samples between exposure (4,907 proteome data) and outcome (18 clinical traits), two-sample MR analysis has been performed using the TwoSampleMR, as previously described [13-14]. In brief, the data were harmonized after clumping (r^2<0.001) to exclude ambiguous SNPs with non-concordant alleles and palindromic SNPs. Following that, the Wald ratio method was used to generate effect estimates when considering a plasma protein instrumented by a single SNP, and the Inverse Variance Weighted (IVW) method was employed for proteins instrumented by two or more SNPs, followed by heterogeneity analysis. To test the consistency of the associations, additional methods such as simple mode, weighted mode, weighted median, and MR-Egger were applied. The MR-Egger intercept test was used to detect probable unbalanced pleiotropy (horizontal pleiotropy). The PhenoScanner (http://www.phenoscanner.medschl.cam.ac.uk/) databases were utilized to investigate whether variants had possible pleiotropic associations with other diseases or traits when fewer than three genetic instrumental factors were employed in MR analysis. Associations with *p*-value<5e-08 were deemed significant. In a heterogeneity test, we calculated I^2 statistics with the "Isq()" function and heterogeneity *p*-value with the "mr_heterogeneity()" function; results with an I^2> 50% and a heterogeneity *p*-value (Q_pval<0.05) were considered

heterogeneous (substantial heterogeneity) [8]. The F-statistic was calculated to quantify instrument strength, and an F-statistic greater than 10 indicated an adequately powerful instrument. To account for multiple tests, the Benjamini-Hochberg correction, which governs the false discovery rate (FDR), was implemented. The association with a *p*-value of 0.05 but a Benjamini-Hochberg adjusted *p*-value greater than 0.05 was considered nominally significant, while the association with a Benjamini-Hochberg adjusted *p*-value greater than 0.05 was considered significant. Furthermore, the Steiger filtering approach was used to verify the validity of the causal direction between the hypothesized exposure and outcomes using the directionality_test() function in the "TwoSampleMR" package [14]. In addition, we performed a leave-one-out analysis to examine whether the MR estimates could be explained by a single SNP (removing estimates where all but one leave-one-out configuration had a *p*-value<0.05). Only MR results that met all of the following criteria were chosen to avoid complex causation relationships: (1) MR-Egger regression revealed no pleiotropy (*p*-value>0.05); (2) true causal direction and Steiger *p*-value<0.05; (3) heterogeneity test $I^2<0.5$; (4) leave-one-out analysis MR *p*-value<0.05 after removing outliers; and (5) FDR<0.05.

Colocalization Analysis

Colocalization analysis was executed to figure out whether the protein and diseases shared the same causative genetic variations and to rule out confounding owing to linkage disequilibrium (LD). This was performed using the "coloc" R package [15]. The following five mutually exclusive hypotheses were tested: (1) Neither protein nor disease has a causal SNP (H0); (2) only protein has a causal SNP (H1); (3) only disease has a causal SNP (H2); (4) both protein and disease have a causal SNP, but the two causal SNPs differ (H3); and (5) both protein and disease have a causal SNP and share the same SNP (H4). To measure the support for each hypothesis, the posterior probability was adopted. A posterior probability for H4 (PH4) of more than 80% was regarded as significant evidence of colocalization.

Single-cell RNA sequencing analysis

Single-cell RNA-sequencing of 54,762 cells taken from normal thyroid tissue

147 from 7 individuals who had thyroidectomy and had verified instances of differentiated
148 thyroid carcinoma were retrieved from Gene Expression Omnibus
149 (https://www.ncbi.nlm.nih.gov/geo/, GSE182416) [16]. Seurat 4.2.0 was used for
150 processing and analyzing the single-cell dataset, as previously stated [13]. In short,
151 low-quality cells by selecting feature numbers ranging from 200 to 3,000 were
152 excluded. The data were normalized using the "LogNormalize" method using the
153 "NormalizeData" function. The "vst" selection method in the "FindVariableFeatures"
154 function yielded a total of 2,000 highly variable genes for each sample. The
155 "FindIntegrationAnchors" function was utilized to correct the batch effect and merged
156 the 7 patients for further analysis. The k-nearest neighbors were determined using
157 principal component analysis (PCA). For data visualization, the dimension reduction
158 approach UMAP (Uniform Manifold Approximation and Projection) was utilized.
159 Cell type annotation was carried out in their study based on the specific gene
160 expression [16].

161 **Mediation analysis**

162 The product of coefficients approach was employed to quantify the impacts of
163 proteins on IPF outcomes via hypothyroidism for proteins that causally are associated
164 with IPF and hypothyroidism, as mentioned by Yoshiji and her colleagues [8]. In brief,
165 the effect of IRF3 on hyperthyroidism was first evaluated and then multiplied by the
166 effect of hyperthyroidism on IPF. Following that, the proportion of the entire effect of
167 IRF3 on IPF mediated by hyperthyroidism was calculated by dividing the
168 hyperthyroidism-mediated effect (hyperthyroidism-to-IPF) by the overall effect
169 (IRF3-to-IPF) [8,17].

170 **Results and discussion**

171 **Identification of causal proteins associated with clinical traits**

172 The MR analysis was used to evaluate the causal effects of 4,907 plasma proteins
173 on 18 diseases by using *cis*-pQTLs (pQTLs that reside within a 1 Mb region around a
174 transcription start site of a protein-coding gene) for these proteins as instrumental
175 variables and 18 clinical traits as outcomes (**Figure 1A** and **Supplemental Table 1**).
176 Given that *cis*-pQTLs are more likely than *trans*-pQTLs to directly alter the

177 transcription or translation of their linked gene, the likelihood of directional horizontal
178 pleiotropy is considerably reduced [8]. Following MR analysis and stringent process
179 filtering (**Method** and **Figure 1A**), a total of 146 plasma proteins were found for
180 associations with 18 clinical traits (**Figure 1B** and **Supplemental Table 1**).
181 Particularly, the causal associations were found for 12 clinical traits, including male
182 pattern baldness (6 proteins), gastroesophageal reflux disease (15 proteins),
183 Keratoconus (3 proteins), selective IgA deficiency disease (2 proteins), age-related
184 hearing impairment (12 proteins), familial hyperlipidemia (8 proteins), atopic eczema
185 (4 proteins), Brugada syndrome (1 protein), amyotrophic lateral sclerosis (1 protein),
186 hypothyroidism (27 proteins), basal cell carcinoma (5 proteins), and prostate cancer
187 (62 proteins). For example, a genetic tendency to elevated LYPLAL1 was linked to an
188 increased risk of keratoconus (OR per-1-SD higher plasma protein level [95%
189 confidence interval (CI)]=2.52[1.56, 4.06]; *p*-value=0.00014). Higher levels of
190 genetically predicted CRAT were linked to an increased incidence of atopic eczema
191 (OR[95% CI]=1.63[1.27, 2.10]; *p*-value=0.00012). Genetically determined higher
192 SCARF2 levels were associated with a higher risk of male pattern baldness
193 (OR[95%CI]: 1.23[1.11, 1.36]; *p*-value=3.91e-05). It is worth noting that the
194 causative protein that is significantly linked to one clinical trait may also be linked to
195 other clinical traits, while this is nominally significant (**Figure 1B**). For example,
196 THBS3, which was associated with an increased risk of gastroesophageal reflux
197 disease (OR[95%CI]=1.31[1.15, 1.49]; *p*-value=6.2e-05), was also linked to a higher
198 risk of cataract (OR[95%CI]=1.014[1.003, 1.025]; *p*-value=0.0125) and uterine
199 fibroid (OR[95%CI]=1.32[1.05, 1.67]; *p*-value=0.016). Furthermore, CRAT, which
200 was linked to atopic eczema, was also associated with a higher risk of age-related
201 hearing impairment (OR[95%CI]=1.079[1.022, 1.139]; *p*-value=0.0059) and
202 fibromuscular dysplasia (OR[95%CI]=3.28[1.11,9.71]; *p*-value=0.031). These
203 findings indicated an opportunity for existing pharmaceutical repurposing in the
204 prevention of disease. In fact, inhibitors of interleukin-23 (IL-23) signaling, initially
205 developed for psoriasis, have been repurposed for Crohn's disease based on GWAS
206 that found links between genetic variants in the IL23R gene and Crohn's disease

207 [3,18,19]. Similarly, IL-17A signaling inhibitors established for psoriasis, rheumatoid
208 arthritis, and uveitis have been investigated and approved for ankylosing spondylitis
209 based on GWAS findings [3]. Notably, LYPLAL1, which has been connected to an
210 increased risk of keratoconus, has also been linked to a lower risk of male pattern
211 baldness (OR[95%CI]=0.87[0.77, 0.995]; *p*-value=0.043) and atopic eczema
212 (OR[95%CI]=0.74[0.60, 0.92]; *p*-value=0.0062). These findings suggested that when
213 repurposing drugs, it is essential to consider the potential effects caused by
214 multiplicity.

215 Colocalization analysis was then implemented to determine whether previously
216 identified relationships between proteins and clinical traits were driven by linkage
217 disequilibrium (LD) (**Figure 1C**). Among the 146 causative proteins, 18 showed
218 higher colocalization evidence (posterior probability PH4>0.8), whereas 6 showed
219 medium colocalization evidence (posterior probability PH4>0.7) (**Figure 1D**). For
220 example, the colocalization indicated that the genetic variations linked to CRAT,
221 LYPLAL1, and THBS3 (*cis*-pQTLs) were caused by the same genetic variants that
222 underpin the association with individual clinical traits (**Figure 1E**). For the other 122
223 causative proteins, there was no colocalization evidence (posterior probability
224 PH4<0.6), indicating that the MR findings for these proteins were likely biased by
225 LD.

226 **Genetically increased circulating CA12 level is associated with a reduced risk of**
227 **hypothyroidism**

228 Based on the MR and colocalization analysis results, the causative protein related
229 to the clinical traits was sought to further examine the role they played. In the case of
230 hypothyroidism, among the three causative proteins, a one standard deviation rise in
231 genetically predicted CA12 levels was linked with a lower risk of hypothyroidism
232 (OR=0.47, 95%CI 0.33-0.66, *p*-value=1.68e-05) (**Figure 1B, Figure 2A** and
233 **Supplementary Tables 1**). To confirm the premise of a lack of directional pleiotropy,
234 which can reintroduce confounding, we used PhenoScanner
235 (http://www.phenoscanner.medschl.cam.ac.uk/) to find out whether the CA12
236 *cis*-pQTLs were related to any characteristics or diseases at the genome-wide

237 significant threshold of *p*-value<5e-08. The deCODE study's primary *cis*-pQTL for
238 CA12 (rs7183733) was linked to dentures (**Supplementary Tables 1**). Indeed,
239 employing CA12 *cis*-pQTL as exposure and dentures as outcomes, MR analysis
240 revealed that CA12 levels were expected to impact dentures (OR=1.12, 95%CI
241 1.08-1.16, *p*-value=5.59e-12) with no reverse causation (*p*-value=1.83e-04,
242 MR-Steiger test). This finding implies that the association between CA12 and
243 dentures is a case of vertical pleiotropy, which does not contradict MR explanation.
244 GTEx v8 (https://gtexportal.org/), which contains expression data from 49 tissues and
245 838 individuals, was used to investigate the tissues in which CA12 is expressed.
246 When compared to whole blood, CA12 was highly expressed in numerous tissues,
247 including the thyroid (*p*-value<0.001) (**Figure 2B**). To better understand the cell type
248 of origin of CA12, we examined single-cell CA12 expression in human thyroid
249 tissues from thyroidectomy patients (GSE182416, available at
250 https://www.ncbi.nlm.nih.gov/geo/) [16]. CA12 was highly enriched in thyroid
251 epithelial cells and fibroblasts in single-cell sequencing as compared to other cell
252 types in adipose tissues (*p*-value<0.001) (**Figure 2C**). No difference in expression
253 between epithelial cells and fibroblasts was found (*p*-value>0.05). These findings
254 suggested that certain cell types may be accountable for local CA12 production in
255 these tissues. Furthermore, we noticed that CA12 was considerably more expressed in
256 older persons (*p*-value<0.001) (**Figure 2D**), who had a greater prevalence and
257 incidence of hypothyroidism than young ones [20]. Taken together, these findings
258 imply that CA12 may be an appropriate target for hypothyroidism.

259 In fact, the causative protein found in the study might be utilized to infer disease
260 mediator associations [7]. As an example, recent research has shown that
261 hypothyroidism has a positive causal effect on IPF [21]; however, the mechanisms
262 underlying this association are not well understood. To evaluate the indirect effect of
263 proteins on IPF outcomes via hypothyroidism, we performed a mediation analysis
264 utilizing the effect estimates from two-step MR and the overall effect from primary
265 MR. To accomplish this, the largest GWAS of IPF (2,668 cases and 8,591 controls)
266 was extracted from a recent study [22] and three proteins (IRF3, CA12, and

267 TNFRSF1B) that have a causal effect on hypothyroidism were initially used to
268 identify a causal effect on IPF using MR analysis, as described above. Surprisingly,
269 only IRF3 was found to have a causal effect on IPF (p-value<0.05/3=0.016,
270 Bonferroni-adjusted for three proteins) (OR=2.79, 95%CI 1.29-6.02, p-value=0.0088)
271 (**Figure 2E** and **Supplementary Tables 1**). Following that, the product of coefficients
272 method was utilized to assess the mediation effect, and 3.8% of the IRF3 mediation
273 affect on IPF via hypothyroidism was uncovered (**Figure 2F**).

274 Overall, our analysis identified 18 putative causative proteins for clinical traits
275 by integrating proteogenomics research. More research is needed to determine the
276 viability of the 18 discovered proteins as therapeutic targets for clinical disease
277 therapy. Furthermore, the detailed mechanism by which these proteins alter clinical
278 traits should be explored in depth using a multi-omics approach.

279 **Competing interests**

280 None declared.

281 **Author contributions**

282 Conception and design: SF. Data analyses: SF. Manuscript writing: SF. Data
283 acquisition: SF and MJ.

284 **Acknowledgments**

285 We would like to thank the IPF GWAS Collaborative Group for providing us with the
286 IPF GWAS summary data. The authors would also like to thank the other researchers
287 who contributed to developing the GWAS datasets utilized in this work for making
288 them available for research purposes.

289 **References**

290 1 Henning RH & Brundel BJJM. Proteostasis in cardiac health and disease. ***Nat***
291 ***Rev Cardiol***. 2017;14:637-53. doi:10.1038/nrcardio.2017.89.

292 2 Walker KA, Chen J, Shi L, et al. Proteomics analysis of plasma from middle-aged
293 adults identifies protein markers of dementia risk in later life. ***Sci Transl Med***. 2023,
294 15, eadf5681. doi:10.1126/scitranslmed.adf5681.

295 3 Trajanoska K, Bhérer C, Taliun D, et al. From target discovery to clinical drug
296 development with human genetics. ***Nature***. 2023;620:737–45.

297 doi:10.1038/s41586-023-06388-8.
298 4 Wallentin L, Eriksson N, Olszowka M, et al. Plasma proteins associated with
299 cardiovascular death in patients with chronic coronary heart disease: A retrospective
300 study. ***PLoS Med*** 2021;18:1-22. doi:10.1371/journal.pmed.1003513.
301 5 Ferkingstad E, Sulem P, Atlason BA, et al. Large-scale integration of the plasma
302 proteome with genetics and disease. ***Nat Genet*** 2021;53:1712-21.
303 doi:10.1038/s41588-021-00978-w.
304 6 Sanderson E, Glymour MM, Holmes MV, et al. Mendelian randomization. ***Nat***
305 ***Rev Methods Primers*** 2022; 2, 1-21. doi:10.1038/s43586-021-00092-5.
306 7 Chen L, Peters JE, Prins B, et al. Systematic Mendelian randomization using the
307 human plasma proteome to discover potential therapeutic targets for stroke. ***Nat***
308 ***Commun*** 2022;13:1-14. doi:10.1038/s41467-022-33675-1.
309 8 Yoshiji S, Butler-Laporte G, Lu T, et al. Proteome-wide Mendelian randomization
310 implicates nephronectin as an actionable mediator of the effect of obesity on
311 COVID-19 severity. ***Nat Metab*** 2023;5:248-64. doi:10.1038/s42255-023-00742-w.
312 9 Nakanishi T, Cerani A, Forgetta V, et al. Genetically increased circulating FUT3
313 level leads to reduced risk of idiopathic pulmonary fibrosis: A Mendelian
314 randomisation study. ***Eur Respir J*** 2022;59. doi:10.1183/13993003.03979-2020.
315 10 King EA, Wade Davis J, Degner JF. Are drug targets with genetic support twice
316 as likely to be approved? Revised estimates of the impact of genetic support for drug
317 mechanisms on the probability of drug approval. ***PLoS Genet*** 2019;15:1-20.
318 doi:10.1371/journal.pgen.1008489.
319 11 Ochoa D, Karim M, Ghoussaini M, et al. Human genetics evidence supports
320 two-thirds of the 2021 FDA-approved drugs. ***Nat Rev Drug Discov*** 2022;21:551.
321 doi:10.1038/d41573-022-00120-3.
322 12 Yoshiji S, Lu TY, Butler-Laporte G, et al, COL6A3-derived endotrophin mediates
323 the effect of obesity on coronary artery disease:an integrative proteogenomics analysis.
324 ***medRxiv*** 2023. doi: 10.1101/2023.04.19.23288706.
325 13 Li SF & Gong MJ. Therapeutic targets for haemorrhoidal disease: proteome-wide
326 Mendelian randomisation and colocalization analyses. ***medRxiv*** 2023.

327 doi:10.1101/2023.06.19.23291373.
328 14 Hemani G, Zheng J, Elsworth B, et al. The MR-base platform supports systematic
329 causal inference across the human phenome. ***Elife*** 2018;7:1-29.
330 doi:10.7554/eLife.34408.
331 15 Giambartolomei C, Vukcevic D, Schadt EE, et al. Bayesian Test for
332 Colocalisation between Pairs of Genetic Association Studies Using Summary
333 Statistics. ***PLoS Genet*** 2014;10. doi:10.1371/journal.pgen.1004383.
334 16 Hong Y, Kim HJ, Park S, et al. Single Cell Analysis of Human Thyroid Reveals
335 the Transcriptional Signatures of Aging. ***Endocrinol*** (United States) 2023;164:1-12.
336 doi:10.1210/endocr/bqad029.
337 17 Carter AR, Gill D, Davies NM, et al. Understanding the consequences of
338 education inequality on cardiovascular disease: Mendelian randomisation study. ***BMJ***
339 2019;365:1-12. doi:10.1136/bmj.l1855.
340 18 Duerr RH, Taylor KD, Brant SR, et al. A genome-wide association study
341 identifies IL23R as an inflammatory bowel disease gene. ***Science*** 2006;314:1461-3.
342 doi:10.1126/science.1135245.
343 19 Burton PR, Clayton DG, Cardon LR, et al. Association scan of 14,500
344 nonsynonymous SNPs in four diseases identifies autoimmunity variants. ***Nat Genet***
345 2007;39:1329-1337. doi: 10.1038/ng.2007.17.
346 20 Kim MI. Hypothyroidism in Older Adults. [Updated 2020 Jul 14]. In: Feingold KR,
347 Anawalt B, Blackman MR, et al., editors. Endotext [Internet]. South Dartmouth (MA):
348 MDText.com, Inc.; 2000-. Available from:
349 https://www.ncbi.nlm.nih.gov/books/NBK279005/.
350 21 Zhang Y, Zhao M, Guo P, et al. Mendelian randomisation highlights
351 hypothyroidism as a causal determinant of idiopathic pulmonary fibrosis.
352 ***eBioMedicine*** 2021;73:103669. doi:10.1016/j.ebiom.2021.103669.
353 22 Allen RJ, Guillen-Guio B, Oldham JM, et al. Genome-Wide Association Study of
354 Susceptibility to Idiopathic Pulmonary Fibrosis. ***Thorax*** 2022; 77(8): 829-833. doi:
355 10.1136/thoraxjnl-2021-218577.
356

357 **Figure Legends**

358 **Figure 1 Mendelian randomisation and colocalization results.** (**A**) A diagram
359 depicting the Mendelian randomization principle and procedure filtering for MR
360 results. LOOA, leave-one-out analysis. (**B**) Shared the associations between 146
361 causative plasma proteins (related with at least one clinical trait) and 18 clinical traits
362 from MR analysis. For visualizing, only MR results showing no pleiotropy
363 (MR-Egger, *p*-value>0.5), no heterogeneity (I^2<0.5), and true causality direction
364 (Steiger test, *p*-value<0.05) were included. (**C**) The two mathematical models of
365 genetic colocalization, causation, and colocalization, as well as LD confounding. (**D**)
366 Scanning for colocalization evidence of causative plasma proteins with outcomes. The
367 size of the circle represents the posterior probability for H4, and the color of the circle
368 represents the classification of the evidence. Red indicates high evidence for
369 colocalization (PH4>0.8); yellow indicates medium evidence for colocalization
370 (PH4>0.7); and green and black indicate moderate evidence for colocalization
371 (PH4<0.7). (**E**) The locus-compare scatter plot for the CRAT, LYPLAL1, and THBS3
372 association signals. The findings of colocalization analyses for CRAT (left),
373 LYPLAL1 (center), and THBS3 (right) are shown. The marked SNP represents the
374 genetic variants used for the MR analysis.

375

376 **Figure 2 Follow-up analyses for causal proteins associated with hypothyroidism.**
377 (**A**) The locus-compare scatter plot compares the quantitative trait loci (pQTL) of
378 CA12 and the GWAS of hypothyroidism. (**B**) CA12 expression profile in human
379 tissues in GTEx v8 (https://gtexportal.org/). CA12 expression levels were displayed
380 on a log transcript per thousand plus one (TPM+1) scale. (**C**) CA12 expression
381 patterns in human thyroid tissues from thyroidectomy patients (GSE182416,
382 accessible at https://www.ncbi.nlm.nih.gov/geo/). NK cells: natural killer cells; SMCs:
383 smooth muscle cells. (**D**) A dot-plot revealing the expression of CA12 in epithelial
384 cells and fibroblasts in children and adults. (**E**) Forest plots for the effect of three
385 selected proteins on IPF. (**F**) Mediation effects of IRF3 on IPF via hypothyroidism.
386 For the effect of IRF3 on IPF mediated by hypothyroidism, the product of coefficients

method calculates the proportion mediated by multiplying $\beta_{IRF3\text{-}to\text{-}hypothyroidism}$ and $\beta_{hypothyroidism\text{-}to\text{-}IPF}$ and subsequently dividing it by $\beta_{IRF3\text{-}to\text{-}IPF}$, where $\beta_{IRF3\text{-}to\text{-}hypothyroidism}$ is the effect of IRF3 on hypothyroidism, $\beta_{hypothyroidism\text{-}to\text{-}IPF}$ is the effect of hypothyroidism on IPF, and $\beta_{IRF3\text{-}to\text{-}IPF}$ is the total effect of IRF3 on IPF.

Supplementary Table 1 The summary statistics obtained in this study.

Fig 1

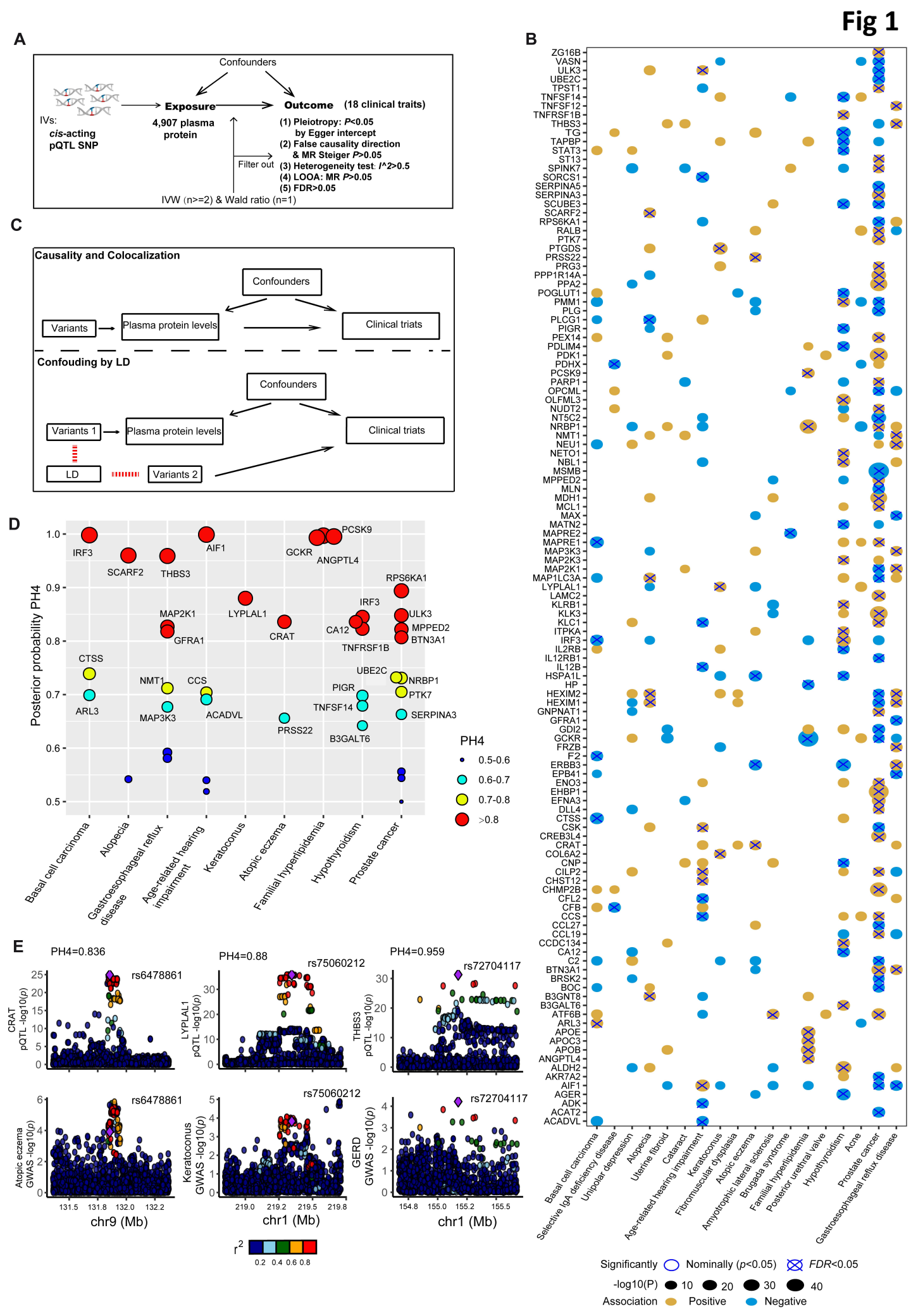
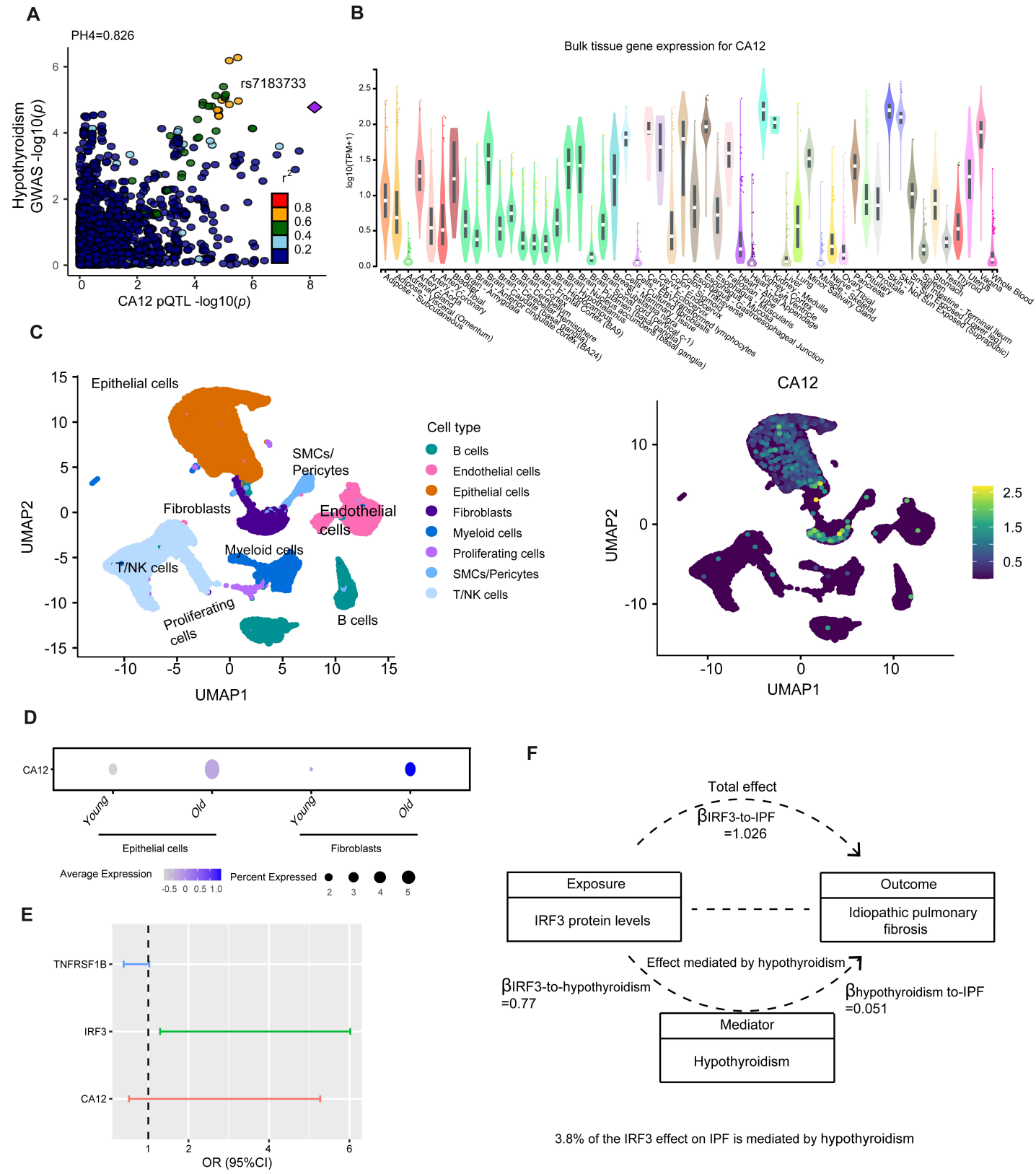

A
Confounders
IVs:
cis-acting
pQTL SNP
Exposure
4,907 plasma
protein
Outcome
(18 clinical traits)
(1) Pleiotropy: P<0.05
by Egger intercept
(2) False causality direction
& MR Steiger P>0.05
(3) Heterogeneity test: I^2>0.5
(4) LOOA: MR P>0.05
(5) FDR>0.05
Filter out
IVW (n>=2) & Wald ratio (n=1)
B
C
Causality and Colocalization
Confounders
Variants
Plasma protein levels
Clinical triats
Confouding by LD
Confounders
Variants 1
Plasma protein levels
Clinical triats
LD
Variants 2
D
Posterior probability PH4
1.0
0.9
0.8
0.7
0.6
0.5
IRF3
SCARF2
THBS3
AIF1
GCKR
ANGPTL4
PCSK9
RPS6KA1
LYPLAL1
IRF3
ULK3
MAP2K1
GFRA1
CRAT
CA12
MPPED2
BTN3A1
TNFRSF1B
CTSS
UBE2C
NRBP1
NMT1
CCS
PIGR
PTK7
ARL3
MAP3K3
ACADVL
TNFSF14
SERPINA3
PRSS22
B3GALT6
PH4
0.5-0.6
0.6-0.7
0.7-0.8
>0.8
Basal cell carcinoma
Alopecia
Gastroesophageal reflux disease
Age-related hearing impairment
Keratoconus
Atopic eczema
Familial hyperlipidemia
Hypothyroidism
Prostate cancer
E
PH4=0.836
rs6478861
CRAT pQTL -log10(p)
PH4=0.88
rs75060212
LYPLAL1 pQTL -log10(p)
PH4=0.959
rs72704117
THBS3 pQTL -log10(p)
rs6478861
Atopic eczema GWAS -log10(p)
rs75060212
Keratoconus GWAS -log10(p)
rs72704117
GERD GWAS -log10(p)
131.5
131.8
132.0
132.2
chr9 (Mb)
219.0
219.2
219.5
219.8
chr1 (Mb)
154.8
155.0
155.2
155.5
chr1 (Mb)
r2
0.2
0.4
0.6
0.8
ZG16B
VASN
ULK3
UBE2C
TPST1
TNFSF14
TNFSF12
TNFRSF1B
THBS3
TG
TAPBP
STAT3
ST13
SPINK7
SORCS1
SERPINA5
SERPINA3
SCUBE3
SCARF2
RPS6KA1
RALB
PTK7
PTGDS
PRSS22
PRG3
PPP1R14A
PPA2
POGLUT1
PMM1
PLG
PLCG1
PIGR
PEX14
PDLIM4
PDK1
PDHX
PCSK9
PARP1
OPCML
OLFML3
NUDT2
NT5C2
NRBP1
NMT1
NEU1
NETO1
NBL1
MSMB
MPPED2
MLN
MDH1
MCL1
MAX
MATN2
MAPRE2
MAPRE1
MAP3K3
MAP2K3
MAP2K1
MAP1LC3A
LYPLAL1
LAMC2
KLRB1
KLK3
KLC1
ITPKA
IRF3
IL2RB
IL12RB1
IL12B
HSPA1L
HP
HEXIM2
HEXIM1
GNPNAT1
GFRA1
GDI2
GCKR
FRZB
F2
ERBB3
EPB41
ENO3
EHBP1
EFNA3
DLL4
CTSS
CSK
CREB3L4
CRAT
COL6A2
CNP
CILP2
CHST12
CHMP2B
CFL2
CFB
CCS
CCL27
CCL19
CCDC134
CA12
C2
BTN3A1
BRSK2
BOC
B3GNT8
B3GALT6
ATF6B
ARL3
APOE
APOC3
APOB
ANGPTL4
ALDH2
AKR7A2
AIF1
AGER
ADK
ACAT2
ACADVL
Basal cell carcinoma
Selective IgA deficiency disease
Unipolar depression
Alopecia
Uterine fibroid
Cataract
Age-related hearing impairment
Keratoconus
Fibromuscular dysplasia
Atopic eczema
Amyotrophic lateral sclerosis
Brugada syndrome
Familial hyperlipidemia
Posterior urethral valve
Hypothyroidism
Acne
Prostate cancer
Gastroesophageal reflux disease
Significantly
Nominally (p<0.05)
FDR<0.05
-log10(P)
10
20
30
40
Association
Positive
Negative

Fig 2


A
PH4=0.826
rs7183733
r2
0.8
0.6
0.4
0.2
Hypothyroidism GWAS -log10(p)
CA12 pQTL -log10(p)
B
Bulk tissue gene expression for CA12
log10(TPM+1)
CA12
C
Epithelial cells
SMCs/ Pericytes
Fibroblasts
Endothelial cells
Myeloid cells
T/NK cells
Proliferating cells
B cells
Cell type
B cells
Endothelial cells
Epithelial cells
Fibroblasts
Myeloid cells
Proliferating cells
SMCs/Pericytes
T/NK cells
UMAP1
UMAP2
D
CA12
Young
Old
Young
Old
Epithelial cells
Fibroblasts
Average Expression
Percent Expressed
E
TNFRSF1B
IRF3
CA12
OR (95%CI)
F
Total effect
βIRF3-to-IPF =1.026
Exposure
IRF3 protein levels
Outcome
Idiopathic pulmonary fibrosis
Effect mediated by hypothyroidism
βIRF3-to-hypothyroidism =0.77
βhypothyroidism to-IPF =0.051
Mediator
Hypothyroidism
3.8% of the IRF3 effect on IPF is mediated by hypothyroidism