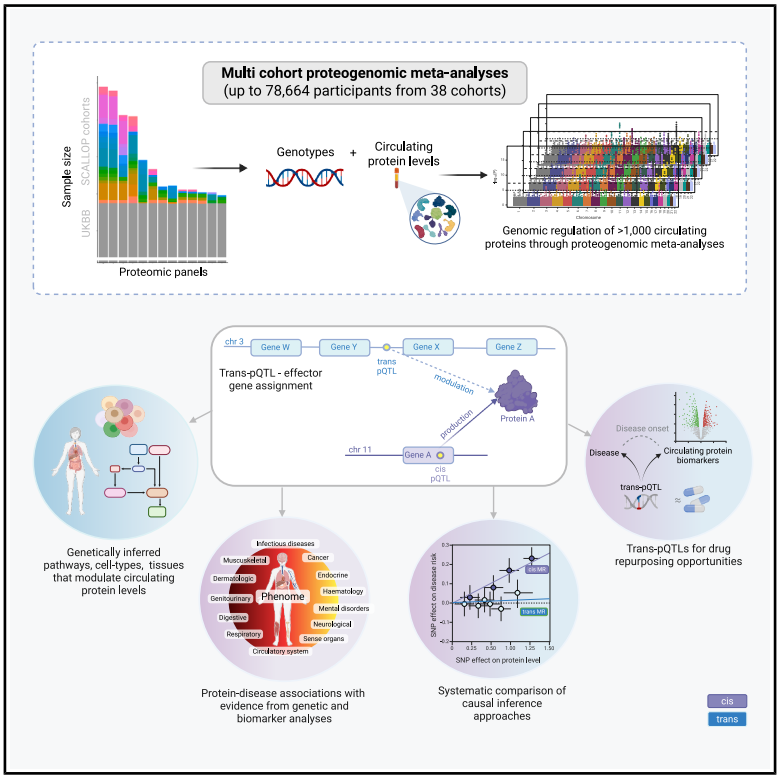
Cell

Multi-cohort proteogenomic analyses reveal genetic effects across the proteome and diseasome

Graphical abstract



Highlights

- Proteogenomic evidence from the largest multi-cohort (*n* = 38) study to date
- Identification of genes, pathways, and cell types that modulate circulating protein levels
- Circulating protein signatures with genetic evidence that inform drug repurposing
- Proteogenomic insights into mechanisms underlying diverse diseases

Authors

Mine Koprulu, Karl Smith-Byrne, Brian Richard Ferolito, ..., Anders Mälarstig, Maik Pietzner, Claudia Langenberg

Correspondence

m.koprulu@qmul.ac.uk (M.K.), maik.pietzner@bih-charite.de (M.P.), claudia.langenberg@qmul.ac.uk (C.L.)

In brief

A large-scale multi-cohort proteogenomic study identifies genetic loci influencing circulating protein levels, revealing pathways and cell types that regulate the circulating proteome and highlighting disease insights and evidence for potential therapeutic opportunities.

Koprulu et al., 2026, Cell *189*, 1–19
May 28, 2026
https://doi.org/10.1016/j.cell.2026.03.049

CellPress

Please cite this article in press as: Koprulu et al., Multi-cohort proteogenomic analyses reveal genetic effects across the proteome and diseasome, Cell (2026), https://doi.org/10.1016/j.cell.2026.03.049

Cell

CellPress
OPEN ACCESS

Article

Multi-cohort proteogenomic analyses reveal genetic effects across the proteome and diseasome

Mine Koprulu,[1,2,3,149,*] Karl Smith-Byrne,[4,149] Brian Richard Ferolito,[5] Erin Macdonald-Dunlop,[6,7,8] Jian'an Luan,[2] Åsa K. Hedman,[7,9] Chibuzor Franklin Ogamba,[4] Jurgis Kuliesius,[8] Linda Repetto,[8,10] Anna Ramisch,[11] Fahim Abbasi,[12] Johan Ärnlöv,[13,14] Themistocles L. Assimes,[15,16] BeLOVE Study Group,[17,18,19,20,21,22,23,24,25,26,27,28,29] Hanna M. Björck,[30] Sophia Björkander,[30] Morten Böttcher,[31,32] Adam Stuart Butterworth,[33,34,35,36,37]

(Author list continued on next page)

[1]Precision Healthcare University Research Institute, Queen Mary University of London, London E1 1HH, UK
[2]MRC Epidemiology Unit, School of Clinical Medicine, University of Cambridge, Cambridge CB2 0QQ, UK
[3]Computational Medicine, Berlin Institute of Health at Charité - Universitätsmedizin Berlin, Berlin 10117, Germany
[4]Cancer Epidemiology Unit, Nuffield Department of Population Health, University of Oxford, Oxford OX3 7LF, UK
[5]Million Veteran Program (MVP) Coordinating Center, Veterans Affairs Healthcare System, Boston, MA 02111, USA
[6]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm 171 65, Sweden
[7]Pfizer Research and Development, Pfizer, Stockholm 113 63, Sweden
[8]Centre for Global Health Research, Usher Institute, University of Edinburgh, Edinburgh EH16 4UX, UK
[9]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm 171 77, Sweden
[10]Estonian Genome Centre, Institute of Genomics, University of Tartu, Tartu 51010, Estonia
[11]Department of Genetic Medicine and Development, Faculty of Medicine, University of Geneva Medical School, Geneva 1211, Switzerland
[12]Division of Cardiovascular Medicine, Department of Medicine, Stanford University School of Medicine, Stanford, CA 90305-5406, USA
[13]School of Health and Social Studies, Dalarna University, Falun 79188, Sweden
[14]Division of Family Medicine and Primary Care, Department of Neurobiology, Care Sciences and Society (NVS), Karolinska Institutet, Stockholm 14183, Sweden
[15]Department of Epidemiology and Population Health, Stanford University School of Medicine, Stanford, CA 94305, USA
[16]Department of Medicine, Stanford University School of Medicine, Stanford, CA 94305, USA
[17]Berlin Institute of Health at Charité - Universitätsmedizin Berlin, Berlin 10178, Germany

(Affiliations continued on next page)

SUMMARY

Understanding the genetic regulation of circulating protein levels can provide new insights into disease mechanisms. Here, we present the largest proteogenomic study to date (*n* = 78,664 participants across 38 studies), identifying >24,000 protein quantitative trait loci (QTLs) associated with 1,116 proteins, acting near to (*n* = 5,040) or distant (*n* = 19,698) from the cognate gene. Using machine learning-guided effector gene assignment, we provide genetic evidence for pathways, cell types, and tissues that modulate circulating protein levels, highlighting N-linked glycosylation as an important regulatory pathway. We demonstrate that genetic instruments of protein production/function ("*cis*") versus modulation ("*trans*") reveal distinct phenotypic insights. We identify proteins as candidates for drug targets and engagement (e.g., plasma furin and cardiovascular diseases) by comparing *cis*-based genetic evidence with protein-disease associations. Systematic triangulation of *trans*-protein QTLs (pQTLs) with genetic and protein associations across many diseases highlights potential drug repurposing opportunities, e.g., tyrosine kinase 2 (TYK2) inhibitors for rheumatoid arthritis. Our multi-cohort meta-analyses generate proteogenomic insights into disease mechanisms and new treatment opportunities.

INTRODUCTION

Most disease-predisposing variation in the human genome resides in non-coding regions, limiting inference about causal genes and mechanisms.[1,2] Systematic functional characterization of disease-associated variants can guide the identification of clinically relevant mechanisms,[3–5] but model systems are difficult to scale or translate directly to human biology.[1–3] High-throughput, broad-coverage proteomic technologies that simultaneously target thousands of proteins in human blood

Zhengming Chen,[38] Kelly Cho,[5,39] Robert Joseph Clarke,[38] Simon Riddington Cox,[40] Kamila Czene,[41] John Danesh,[34,42,43,44,45] George Dedoussis,[46] Sölve Elmståhl,[47] Niclas Eriksson,[48] Per Eriksson,[49] Tõnu Esko,[10] Estonian Biobank Research Team,[10] Aida Ferreiro-Iglesias,[50] Paul William Franks,[1,51] Jingyuan Fu,[52,53] J. Michael Gaziano,[54,55] Mohsen Ghanbari,[56] Christian Gieger,[57,58,59] Arthur Gilly,[60] Harald Grallert,[61] Marc James Gunter,[62] Stefan Gustafsson,[63] Andreas Göteson,[64] Per Frans Leonard Hall,[65,66] Oskar Hansson,[67] Sarah Elizabeth Harris,[40] Caroline Hayward,[68] Christian Herder,[69,70,71] Natalia Hernandez-Pacheco,[30] Ziad Hijazi,[72] Robert F. Hillary,[73] Jemma Caroline Hopewell,[38]

(Author list continued on next page)

[18]Institute of Medical Informatics, Charité - Universitätsmedizin Berlin, Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin 10115, Germany
[19]Center for Stroke Research (CSB), Charité - Universitätsmedizin Berlin, Freie Universität Berlin and Humboldt-Universität Berlin, Berlin 10117, Germany
[20]Department of Neurology with experimental Neurology, Charité - Universitätsmedizin Berlin, Freie Universität Berlin and Humboldt-Universität Berlin, Berlin 10117, Germany
[21]Institute of Biometry and Clinical Epidemiology, Charité - Universitätsmedizin Berlin, Freie Universität Berlin and Humboldt-Universität Berlin, Berlin 10117, Germany
[22]Department of Endocrinology and Metabolism, Charité - Universitätsmedizin Berlin, Freie Universität Berlin and Humboldt-Universität Berlin, Berlin 10117, Germany
[23]Department of Nephrology and Medical Intensive Care, Charité - Universitätsmedizin Berlin, Freie Universität Berlin and Humboldt-Universität Berlin, Berlin 13353, Germany
[24]Department of Cardiology, Angiology and Intensive Care Medicine, Campus Virchow Klinikum, Deutsches Herzzentrum der Charité (DHZC), Berlin 13353, Germany
[25]DZHK Partner Site Berlin, DZHK (German Centre for Cardiovascular Research), Berlin 10785, Germany
[26]Biobank Technology Platform, Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin 13125, Germany
[27]Genetics and Genomics of Cardiovascular Diseases, Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin 13125, Germany
[28]Hypertension-Caused End-Organ Damage, Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin 13125, Germany
[29]Integrative Vascular Biology, Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin 13125, Germany
[30]Division of Cardiology, Center for Molecular Medicine, Karolinska University Hospital, Solna, Stockholm, Sweden
[31]Department of Clinical Medicine, Aarhus University, Aarhus 8000, Denmark
[32]Department of Cardiology, Gødstrup Hospital, Herning 7400, Denmark
[33]British Heart Foundation Centre of Research Excellence, School of Clinical Medicine, Addenbrooke's Hospital, University of Cambridge, Cambridge CB2 0BB, UK
[34]Department of Public Health and Primary Care, British Heart Foundation Cardiovascular Epidemiology Unit, University of Cambridge, Cambridge CB2 0BB, UK
[35]NIHR Blood and Transplant Research Unit in Donor Health and Behaviour, University of Cambridge, Cambridge CB2 0BB, UK
[36]Victor Phillip Dahdaleh Heart and Lung Research Institute, University of Cambridge, Cambridge CB2 0BB, UK
[37]Health Data Research UK, Wellcome Genome Campus, University of Cambridge, Hinxton CB10 1RQ, UK
[38]Nuffield Department of Population Health, University of Oxford, Oxford OX3 7LF, UK
[39]Division of Aging, Mass General Brigham, Harvard Medical School, Boston, MA 02130, USA
[40]Lothian Birth Cohorts, Department of Psychology, University of Edinburgh, Edinburgh EH8 9JZ, UK
[41]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm 171 77, Sweden
[42]British Heart Foundation Centre of Research Excellence, School of Clinical Medicine, Addenbrooke's Hospital, School of Clinical Medicine, Addenbrooke's Hospital, Cambridge CB2 0BB, UK
[43]Public Health and Primary Care, Victor Phillip Dahdaleh Heart and Lung Research Institute, Cambridge CB2 0BB, UK
[44]Health Data Research UK Cambridge, Wellcome Genome Campus, University of Cambridge, Cambridge CB2 0BB, UK
[45]Department of Human Genetics, Wellcome Sanger Institute, Cambridge CB10 1SA, UK
[46]Nutrition and Dietetics, School of Health Science and Education, Harokopio University, Athens 17671, Greece
[47]Departement of Clinical Sciences in Malmö, Lund University, Malmö 205 02, Sweden
[48]Uppsala Clinical Research Center, Uppsala University, Uppsala 751 85, Sweden
[49]Department of Medicine Solna, Karolinska Institutet, Stockholm 17177, Sweden
[50]Genomic Epidemiology Branch (GEM), International Agency for Research on Cancer (IARC), Lyon 69007, France
[51]Department of Clinical Sciences, Lund University, Helsingborg 25187, Sweden
[52]Department of Genetics, University Medical Center Groningen, Groningen 9713GZ, the Netherlands
[53]Department of Pediatrics, University Medical Center Groningen, Groningen 9713GZ, the Netherlands
[54]Department of Medicine, Brigham and Women's Hospital, Boston, MA 02120, USA
[55]Department of Medicine, VA Boston Healthcare System, Boston, MA 02132, USA
[56]Department of Epidemiology, Erasmus MC University Medical Center Rotterdam, Rotterdam 3015 GD, the Netherlands
[57]German Center for Diabetes Research (DZD), Neuherberg 85764, Germany
[58]Institute of Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg 85764, Germany

(Affiliations continued on next page)

Please cite this article in press as: Koprulu et al., Multi-cohort proteogenomic analyses reveal genetic effects across the proteome and disease, Cell (2026), https://doi.org/10.1016/j.cell.2026.03.049

CellPress
OPEN ACCESS

Shixian Hu,[74] Shih-Jen Hwang,[75,76] Christina Jern,[77,78] Åsa Johansson,[79] Lina Jonsson,[80] Anette Kalnapenkis,[10] Nicola Dorothy Kerrison,[2] Pik Fang Kho,[81,82] Lucija Klaric,[83] Leonhard Kohleick,[3] Julia Kraft,[84,85,86] Mikael Landén,[9,87] Daniel Levy,[88,89] Liming Li,[90] Lars Lind,[63] Jirong Long,[91] Niklas Mattsson-Carlgren,[92] Erik Melén,[93] Simon Kebede Merid,[93] Philipp Mertins,[94,95] Karl Michaëlsson,[96] Peter Loof Møller,[97] Federico Murgia,[98] Mette Nyegaard,[97,99]

(Author list continued on next page)

[59]Research Unit of Molecular Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg 85764, Germany
[60]Institute of Translational Genomics, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg 85764, Germany
[61]Institute of Epidemiology, Research Unit of Molecular Epidemiology, Helmholtz Zentrum München, Neuherberg 85764, Germany
[62]School of Public Health, Imperial College London, London W12 0BZ, UK
[63]Department of Medical Sciences, Uppsala University, Uppsala 75185, Sweden
[64]Department of Psychiatry and Neurochemistry, Institute of Neuroscience and Physiology at the Sahlgrenska Academy, University of Gothenburg, Gothenburg 41346, Sweden
[65]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm 171 77, Sweden
[66]Department of Oncology, Södersjukhuset, Stockholm 118 83, Sweden
[67]Clinical Memory Research Unit, Department of Clinical Sciences Malmö, Faculty of Medicine, Lund University, Lund 22100, Sweden
[68]Institute for Genetics and Cancer, University of Edinburgh, Edinburgh EH4 2XU, UK
[69]Partner Düsseldorf, German Center for Diabetes Research (DZD), Neuherberg 85764, Germany
[70]Institute for Clinical Diabetology, German Diabetes Center, Leibniz Center for Diabetes Research at Heinrich Heine University Düsseldorf, Düsseldorf 40225, Germany
[71]Division of Endocrinology and Diabetology, Medical Faculty, University Hospital Düsseldorf, Heinrich Heine University Düsseldorf, Düsseldorf 40225, Germany
[72]Department of Medical Sciences, Cardiology, Uppsala University, Uppsala 75185, Sweden
[73]Department of Psychology, University of Edinburgh, Edinburgh EH8 9JZ, UK
[74]Institute of Precision Medicine, The First Affiliated Hospital, Sun Yat-Sen University, Guangzhou 510080, China
[75]Department of Biostatistics, Boston University, Boston, MA 02215, USA
[76]Population Sciences Branch, National Heart Lung and Blood Institute, National Institute of Health, Framingham, MA 01702, USA
[77]Department of Laboratory Medicine, Institute of Biomedicine, Gothenburg 405 30, Sweden
[78]Clinical Genetics and Genomics, Sahlgrenska University Hospita, Gothenburg 413 45, Sweden
[79]Immunology, Genetics and Pathology, Uppsala University, Uppsala 74895, Sweden
[80]Institute of Neuroscience and Physiology, University of Gothenburg, Gothenburg 41345, Sweden
[81]Alnylam Pharmaceuticals, Inc., Cambridge, MA 02142, USA
[82]Cardiovascular Medicine, Stanford University, Stanford, CA 94305, USA
[83]MRC Human Genetics Unit, University of Edinburgh, Edinburgh EH4 2XU, UK
[84]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
[85]Department of Psychiatry and Psychotherapy, Charité – Universitätsmedizin Berlin, Berlin 10117, Germany
[86]German Center for Mental Health (DZPG), Partner Site Berlin/Potsdam, Berlin 10117, Germany
[87]Institute of Neuroscience and Physiology, University of Gothenburg, Gothenburg 413 45, Sweden
[88]Population Sciences Branch, National Heart, Lung, and Blood Institute of the National Institutes of Health, Framingham, MA 02461, USA
[89]Framingham Heart Study, Framingham, MA 01702, USA
[90]School of Public Health, Peking University, Beijing 100191, China
[91]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN 37213, USA
[92]Clinical Memory Research Unit, Department of Clinical Sciences Malmö, Lund University, Malmö 20213, Sweden
[93]Department of Clinical Science and Education, Karolinska Institutet, Stockholm 118 83, Sweden
[94]Core Unit Proteomics, Berlin Institute of Health at Charité - Universitätsmedizin Berlin, Berlin 13125, Germany
[95]Proteomics Platform, Max Delbrück Center for Molecular Medicine, Berlin 13125, Germany
[96]Medical Epidemiology, Department of Surgical Sciences, Uppsala University, Uppsala 751 85, Sweden
[97]Department of Health Science and Technology, Aalborg University, Gistrup 9260, Denmark
[98]Clinical Trial Service Unit, Epidemiological Studies Unit (CTSU), Nuffield Department of Population Health, University of Oxford, Oxford OX3 7LF, UK
[99]Department of Congenital Disorders, Statens Serum Institute, Copenhagen 2300, Denmark
[100]Department of Diabetes, Endocrinology & Reproductive Biology, University of Dundee, Dundee DD1 9SY, UK
[101]Department of Immunology and Inflammation, Imperial College London, London W120HS, UK
[102]Robertson Centre for Biostatistics, School of Health and Wellbeing, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow G128TB, UK
[103]Department of Public Health, University of Split, School of Medicine, Split 21000, Croatia
[104]British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge CB2 0BB, UK

(Affiliations continued on next page)

Young-Chan Park,[60] Ewan Pearson,[100] James Peters,[101] John Ross Petrie,[102] Grace Png,[60] Ozren Polašek,[103] Bram Peter Prins,[104] Stephan Ripke,[84,85,105] Michael Roden,[106,107,108] Palle Duun Rohde,[97] Saredo Said,[106,109] SCALLOP Consortium,[110] Xia Shen,[8,41,111] Jochen M. Schwenk,[112] Agneta Siegbahn,[113] J. Gustav Smith,[114,115,116,117,118,119] Tara M. Stanne,[120,121] Karsten Suhre,[122,123] Johan Sundström,[63,124] Barbara Thorand,[58,125,126] Elsa Valdes-Marquez,[38]

(Author list continued on next page)

[105]German Center for Mental Health, DZPG, site Berlin-Potsdam, Berlin 10117, Germany
[106]Partner Düsseldorf, German Center for Diabetes Research (DZD), Neuherberg 80992, Germany
[107]Institute of Clinical Diabetology, German Diabetes Center, Düsseldorf 40225, Germany
[108]Department of Endocrinology and Diabetology, Heinrich Heine University, Düsseldorf 40225, Germany
[109]Novo Nordisk Research Centre Oxford, Novo Nordisk, Oxford OX3 7FZ, UK
[110]SCALLOP Consortium
[111]Greater Bay Area Institute of Precision Medicine, State Key Laboratory of Genetic Engineering, Center for Evolutionary Biology, School of Life Sciences, Fudan University, Shanghai 200032, China
[112]Science for Life Laboratory, Department of Protein Science, KTH Royal Institute of Technology, Solna 171 65, Sweden
[113]Department of Medical Sciences, Clinical Chemistry, Uppsala University, Uppsala 75185, Sweden
[114]Department of Molecular and Clinical Medicine, Institute of Medicine, Gothenburg University, Gothenburg 413 34, Sweden
[115]Science for Life Laboratory, Gothenburg University, Gothenburg 413 34, Sweden
[116]Department of Cardiology, Clinical Sciences, Lund University, Lund 221 84, Sweden
[117]Wallenberg Center for Molecular Medicine, Lund University Diabetes Center, Lund University, Lund 221 84, Sweden
[118]Department of Cardiology, Sahlgrenska University Hospital, Gothenburg 413 34, Sweden
[119]Department of Heart Failure and Valvular Disease, Skåne University Hospital, Lund 221 85, Sweden
[120]Department of Clinical Genetics and Genomics, Sahlgrenska University Hospital, Gothenburg 413 90, Sweden
[121]Institute of Biomedicine, Department of Laboratory Medicine, the Sahlgrenska Academy, University of Gothenburg, Gothenburg 405 30, Sweden
[122]Englander Institute for Precision Medicine, Weill Cornell Medicine, New York, NY 10021, USA
[123]Bioinformatics Core, Weill Cornell Medicine-Qatar, Doha 24144, Qatar
[124]The George Institute for Global Health, University of New South Wales, Sydney, NSW 2031, Australia
[125]Partner Munich-Neuherberg, German Center for Diabetes Research (DZD), Neuherberg 85764, Germany
[126]Institute for Medical Information Processing, Biometry and Epidemiology - IBE, Faculty of Medicine, Ludwig-Maximilians-Universität in Munich, Munich 81377, Germany
[127]Genomics Core Facility, Department of Internal Medicine, Erasmus Medical Center, Rotterdam, 3015 GD, the Netherlands
[128]Department of Orthopedics and Sports Medicine, Erasmus Medical Center, Rotterdam 3000DR, the Netherlands
[129]Population Health and Genomics, University of Dundee, Dundee DD19SY, UK
[130]Department Medical Sciences, Uppsala University, Uppsala 75285, Sweden
[131]Uppsala Clinical Research Center, Uppsala University, Uppsala 75185, Sweden
[132]BeLOVE Unit, Berlin Institute of Health at Charité - Universitätsmedizin Berlin, Berlin 10178, Germany
[133]Center for Stroke Research (CSB), Charité - Universitätsmedizin Berlin, Freie Universität Berlin and Humboldt-Universität Berlin, Berlin 12203, Germany
[134]Department of Neurology, Charité - Universitätsmedizin Berlin, Freie Universität Berlin and Humboldt-Universität Berlin, Berlin 12203, Germany
[135]Department of Gastroenterology and Hepatology, University Medical Center Groningen, University of Groningen, Groningen 9700RB, the Netherlands
[136]National Research Council, Institute of Genetic and Biomedical Research, Cagliari 09042, Italy
[137]Department of Medicine, Stanford University School of Medicine, Palo Alto, CA 94305, USA
[138]TUM School of Medicine and Health, Technical University of Munich (TUM), TUM University Hospital, Munich 81675, Germany
[139]Department of Public Health and Primary Care, British Heart Foundation Cardiovascular Epidemiology Unit, University of Cambridge, Cambridge CB1 8RN, UK
[140]Victor Phillip Dahdaleh Heart & Lung Research Institute, University of Cambridge, Cambridge CB2 0BB, UK
[141]Department of Genetics, University of Groningen, Medical Center Groningen, Groningen 9700 RB, the Netherlands
[142]Proteomics Platform, Max-Delbrück-Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin 13125, Germany
[143]Chinese Academy of Medical Sciences Oxford Institute, Nuffield Department of Medicine, University of Oxford, Oxford OX3 7BN, UK
[144]Target Discovery Institute, Nuffield Department of Medicine, University of Oxford, Oxford OX3 7FZ, UK
[145]Department of Aging, Brigham and Women's Hospital, Boston, MA 02215, USA
[146]Department of Medicine, Harvard Medical School, Boston, MA 02215, USA
[147]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm 17176, Sweden
[148]Max Planck Institute for Molecular Genetics, Berlin 14195, Germany
[149]These authors contributed equally

(Affiliations continued on next page)

Cell
Article

CellPress
OPEN ACCESS

Costanza L. Vallerga,[127] Joyce B.J. van Meurs,[127,128] Ana Viñuela,[129] Urmo Võsa,[10] Lars Wallentin,[130,131] Robin G. Walters,[38] Nicholas John Wareham,[2] Joachim Eduard Weber,[25,132,133,134] Rinse Karel Weersma,[135] James F. Wilson,[8,68] Simon Winther,[32] Summaira Yasmeen,[3] Daniela Zanetti,[136,137] Eleftheria Zeggini,[60,138] Jing Hua Zhao,[139,140] Alexandra Zhernakova,[141] Daria V. Zhernakova,[141] Matthias Ziehm,[94,142] Benedikt Mathias Kessler,[143,144] Alexandre C. Pereira,[145,146] Anders Mälarstig,[147] Maik Pietzner,[1,2,3,150,*] and Claudia Langenberg[1,2,3,148,150,151,*]
[150]These authors contributed equally
[151]Lead contact
*Correspondence: m.koprulu@qmul.ac.uk (M.K.), maik.pietzner@bih-charite.de (M.P.), claudia.langenberg@qmul.ac.uk (C.L.)
https://doi.org/10.1016/j.cell.2026.03.049

can help to link non-coding variants to disease-relevant mechanisms via proteins.[6–24] Although such proteogenomic studies have highlighted novel clinically relevant mechanisms and potential drug targets or indications,[6–24] important limitations remain. First, broad-coverage studies to date have almost exclusively relied on proximally acting *cis*-variants (i.e., *cis*-protein quantitative trait loci [*cis*-pQTLs]) to derive disease insights. However, non-coding variants are likely to map to regulatory regions directly affecting multiple genes encoded close by (as also demonstrated for gene expression QTLs[25]) or indirectly regulate proteins produced by genes that are elsewhere in the genome. Second, a better understanding of the polygenic architecture affecting diagnostic, predictive, or prognostic protein biomarkers can be important, as recently demonstrated for prostate-specific antigen and prostate cancer prediction.[26] Finally, robust and generalizable identification of pQTLs requires replication across diverse populations, which until now has rarely been pursued for broad-coverage proteomics.[18,23,24]

Here, we present multi-cohort proteogenomic meta-analyses of up to 1,161 protein targets in up to 78,664 individuals across 38 international cohorts (see STAR Methods and Figure S1A) to address these limitations and provide a conceptual advance by integrating and improving the biological interpretability of *trans*-pQTL effects, i.e., pQTLs that are not nearby the protein-encoding gene for the protein they associate with. We identify pathways, such as N-linked glycosylation, tissues, and cell types that regulate the plasma proteome based on human genetic evidence and demonstrate divergence of findings from *cis*- versus *trans*-focused genetic analyses across diseases. Systematically comparing *cis*-based gene-to-disease versus measured protein-to-disease associations in up to 1.3 million participants, we demonstrate limited convergence, with only a few examples supported by evidence from both approaches. We finally demonstrate the value of the many newly identified pleiotropic (*trans*-)pQTLs by illustrating how they can inform disease mechanisms, including druggable opportunities.

RESULTS

Multi-cohort genome-proteome-wide pQTL discovery

We performed a proteogenomic meta-analysis of 1,161 protein targets measured in blood across 38 cohorts, including up to 78,664 participants of European ancestry (Table S1A; for study design see Figure S1). We identified 14,690 regional sentinel variants (n = 1,009 *cis*-pQTLs, n = 13,681 *trans*-pQTLs) for 1,116 protein targets at a Bonferroni-corrected significance threshold. We observed at least one pQTL close to the cognate protein-coding gene (i.e., *cis*-pQTL) or elsewhere in the genome (i.e., *trans*-pQTL) for 87.1% and 94.1% of the tested protein targets (n = 1,161), respectively.

Bayesian fine-mapping[27] of the 14,690 identified genomic regions revealed 24,738 independent sets of credible variants associated with 1,116 protein targets (n = 5,040 *cis*-pQTLs and n = 19,698 *trans*-pQTLs; Figure 1). We observed a median of 4 credible sets (interquartile range [IQR]: 2–7; Table S2), including independent genetic variants within/nearby cognate protein-coding genes. We observed considerable correlation (r = 0.6) of effect sizes for the identified regional sentinel variants in cohorts of non-European ancestry (n = 3 cohorts), where overlapping information was available (3,709 pQTLs for 501 protein targets; Figure S2).

For about a third of the protein targets with at least one *cis*-pQTL, we observed evidence that at least one of the *cis* genetic signals was also colocalized with expression QTLs of the protein-coding gene in one or more of 43 tissues, indicating these as potential origins for given blood proteins[25] (see STAR Methods and Table S3). Functional annotation of identified pQTLs further replicated the previous observation of high rates of functional, e.g., missense, variants among *cis*-pQTLs (10.6%), although we observed similarly high rates of functional variants among *trans*-pQTLs (10.4%), including those associated with many protein targets. The latter may imply that broad effects on the plasma proteome require disturbance of causative gene products rather than subtle effects on gene regulation, as has been seen for most genome-wide association study (GWAS) traits.

Most fine-mapped pQTLs were "high confidence" (82.3% of the 5,040 *cis*- and 83.3% of the 19,698 *trans*-fine-mapped pQTLs), defined as (1) genome-wide significant ($p < 5 \times 10^{-8}$); (2) directionally consistent in the overall meta-analysis between SCALLOP cohorts and UK Biobank (UKBB) for *cis*- and *trans*-pQTLs; and, additionally, with (3) no or minimal heterogeneity ($p_{het} > 1 \times 10^{-4}$) for *trans*-pQTLs (Figure 1; Table S2). Of those, 278 *cis*- and 4,013 *trans*-pQTLs or their proxies ($r^2 > 0.1$) have not previously been reported for the same protein target[6–24] (Table S2). Investigating the impact of cohort-level characteristics from SCALLOP cohorts (mean cohort age, mean cohort BMI, composition of sex in the cohort, whether any cases were included in the cohort, fasting status, and blood-based sample type), we observed that the heterogeneity was mostly explained by small variations in very strong genetic effects (Tables S1 and S4).

Please cite this article in press as: Koprulu et al., Multi-cohort proteogenomic analyses reveal genetic effects across the proteome and diseasome, Cell (2026), https://doi.org/10.1016/j.cell.2026.03.049

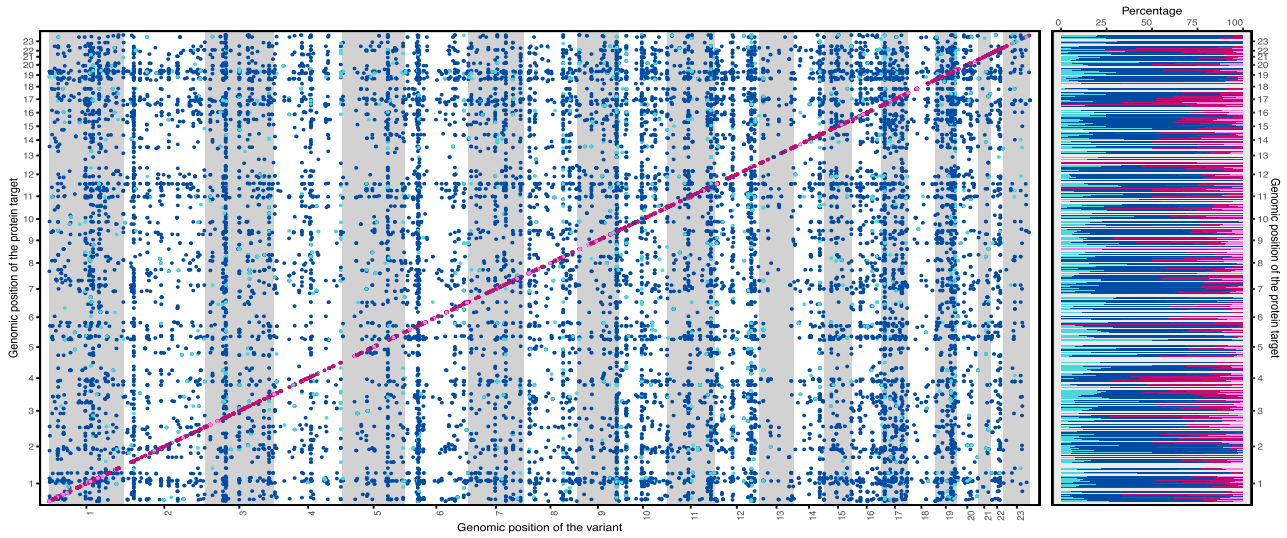


Figure 1. Fine-mapped protein quantitative trait loci in SCALLOP and UKBB meta-analyses

The left panel displays the genomic coordinates of the genetic variants and protein-encoding genes on the *x* and *y* axes, respectively. The pQTLs were colored by their category (dark pink, high-confidence *cis*-pQTLs; light pink, low-confidence *cis*-pQTLs; dark blue, high-confidence *trans*-pQTLs; light blue, low-confidence *trans*-pQTLs). The right panel displays the proportion of pQTLs in each "confidence" category (see STAR Methods) for each protein target. Relevant summary statistics can be found in Table S2.

See also Figures S1, S2, and S3.

We observed large variation among protein targets when contrasting variance explained (mean = 8.4%, IQR = 4.9%–10.4%) by loci identified in this study and the estimated remaining polygenic background (mean = 10.2%, IQR = 6.6%–13.6%; Figure S3A). Extreme examples ranged from monogenic proteins such as Fc receptor-like protein 3 (FCRL3, variance explained = 45.3%; polygenic background = 3.8%), mainly explained by identified *cis* loci, to polygenic ones with contributions from *cis* and *trans* loci, as well as the polygenic background, such as vascular endothelial growth factor receptor 2 (VEGFR-2 encoded by *KDR*, variance explained = 21.6%; polygenic background = 16.4%). pQTLs close to the protein-coding gene explained, on average, (i.e., *cis*), more variation in plasma protein levels compared with the cumulative impact of variation elsewhere in the genome (i.e., *trans*) (Figure S3A; Table S5). However, extreme outliers included intercellular adhesion molecule 2 (ICAM2), for which *cis*-pQTLs explained only 0.3% of variance compared with 52.7% explained by *trans*-pQTLs or alpha-L-fucosidase 1 (FUCA1; 6.3% by *cis*-pQTLs versus 68.4% by *trans*-pQTLs). We further observed 261 protein targets for which the linear dependency between explained variance by pQTLs and polygenic heritability estimates did not persist, providing evidence that our study may have already saturated pQTL discovery for these proteins (Figure S3B).

Understanding characteristics of protein targets under genetic regulation

A higher number of pQTLs among protein targets showed a significant association with the presence of disulfide bonds or a transmembrane domain (Figure 2; Table S6), likely explained by measuring those in the most relevant compartment (i.e., the circulation). In contrast, a higher constraint of the protein-coding gene (i.e., probability of loss-of-function intolerance [pLI]) was inversely associated with the presence and number of *cis*-pQTLs ($p = 4.8 \times 10^{-4}$) as well as the number of *trans*-pQTLs ($p = 9.9 \times 10^{-13}$).

Proteins with a higher number of *trans*-pQTLs were further characterized by those enriched for characteristics of secreted proteins,[28] such as glycosylation and sulfation, but depleted for structural features of intracellular proteins, such as zinc-finger and DNA-binding domains (Figure 2; Table S6). We note that these findings persisted upon controlling for missingness rates for a given protein; that is, our findings are unlikely to be driven by higher precision for proteins reliably measured with the assay technology (Table S6).

Effector genes at *trans*-pQTLs inform pathway and cell-type contributions

Most identified genetic regulation of the human proteome does not occur within or nearby a protein's cognate gene but rather elsewhere in the genome. Mapping effector genes in distal genomic regions to pQTLs that associate with plasma protein levels has been challenging but can benefit from prior biological knowledge, for example, in ligand-receptor pairs.[20]

We identified at least one candidate effector gene with at least medium confidence (candidate gene score > 1 out of 3) for more than half of the *trans*-pQTLs (n = 11,261) by incorporating prior biological knowledge in a machine learning framework (see STAR Methods and Table S7). This included 1,534 high-confidence assignments with a candidate gene score ≥2 (maximum: 2.704 out of 3). For two-thirds of the loci (n = 13,881), the distribution of candidate gene scores across genes indicated a single causal gene as the most likely. As a partial external validation, we observed 552 *trans*-pQTLs mapping to effector genes that encoded high-confidence protein-protein interaction partners

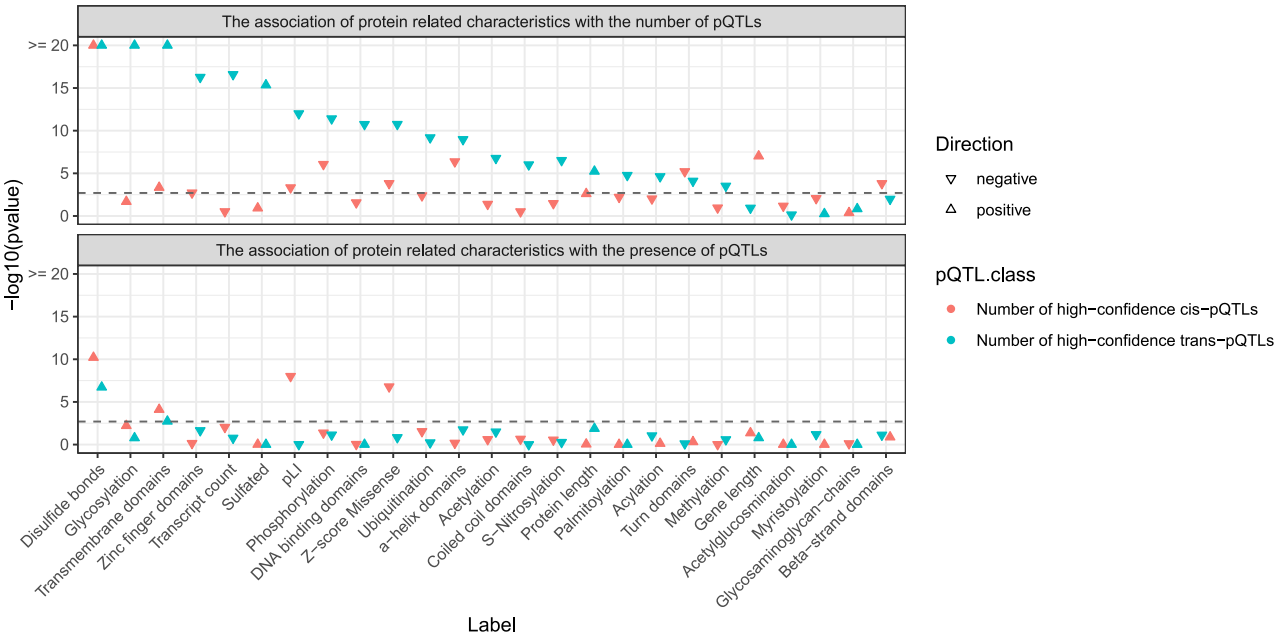


Figure 2. Protein characteristics associated with the presence and number of pQTLs based on zero-inflated Poisson regression models

Results for high-confidence *cis*- and *trans*-pQTLs are colored red and blue, respectively. The results from Poisson regression (i.e., association with the number of pQTLs with protein characteristics) and logistic regression (i.e., association with presence or lack of pQTLs with protein characteristics) are represented in the top and bottom panels, respectively. The horizontal dashed line represents the Bonferroni-corrected significance threshold ($p < 2 \times 10^{-3}$). The *p* value display was capped at $p < 1 \times 10^{-20}$ for display purposes; *p* values can be found in Table S5. Abbreviations: pLI, probability of loss-of-function intolerance.

from the STRING network, which was not part of our gene annotation pipeline.

We next sought to identify the pathways, cell types, and tissues involved in the *trans* regulation of the assayed protein targets. We identified at least one significantly enriched pathway among the top-prioritized *trans*-effector genes for 431 protein targets (Table S8A). Those segregated roughly into two categories. For 101 protein targets, effector genes were enriched for at least one pathway that also involved the *cis*-protein, providing evidence that higher/lower circulating protein levels were, at least partly, a result of generally higher/lower pathway activity. For example, *trans*-effector genes for the low-density lipoprotein (LDL)-receptor were more than 85-fold enriched for members of cholesterol metabolism (false discovery rate [FDR] = 9.8×10^{-15}). In contrast, for most assayed protein targets ($n = 330$), pathways enriched among *trans*-effector genes rather pointed to distal, systemic regulators of (circulating) protein levels. These included pathways related to post-translational modifications, such as "asparagine N-linked glycosylation" ($n = 143$ protein targets; Figure 3B), as well as pathways related to platelet and blood cell biology more broadly, such as platelet aggregation ($n = 41$) (Figure 3A). N-glycosylation is the most common post-translational modification of secreted proteins, determining correct folding and secretion, as well as affecting half-life and signaling properties of protein targets.[29] We note that the most pleiotropic *trans*-effector genes mapped to elongation, branching, and capping of the glycans, modulating stability rather than earlier processes, such as folding in the endoplasmic reticulum.

We identified 95 and 97 protein targets for which *trans*-effector genes were significantly (FDR < 5%) enriched for the specific/enhanced expression of the protein-coding gene in one of 12 tissues and 29 cell types, respectively (Figure 3C; Tables S8B and S8C). This included 44 protein-tissue and 76 protein-cell-type pairs that were not the primary site of protein production of the *cis*-protein, potentially demonstrating cross-organ communication in protein homeostasis. Most prominently, effector genes with enhanced expression in the liver, specifically hepatocytes, were linked to many circulating protein targets. Effector genes expressed in immune cell populations that modulate interleukins (ILs), e.g., genes with enhanced expression in natural killer (NK) cells, were also more than 13-fold enriched (FDR = 3.4×10^{-6}) for *trans*-effector genes of the NK cell activator IL-15. Other examples included the brain-specific extracellular matrix protein tenascin R, for which *trans*-effector genes were more than 20-fold enriched (FDR = 2.8×10^{-8}) for genes with enhanced expression in endothelial cells, or neurotrophin 3, a neuronal growth factor, with *trans*-effector genes being strongly enriched for enhanced expression in alveolar type 2 cells in the lung. We also observed examples with *trans*-effector genes being enriched for multiple, different cell types for the same protein target. For example, *trans*-effector genes for IL-27 with enhanced expression in hepatocytes and Kupffer cells (resident macrophages in the liver) were linked to glycosylation and recognition of pathogens, respectively.

Molecular versus phenome-wide pleiotropy

Almost half (43.4%; $n = 4,547$) of all independent pQTLs we identified have been reported to associate with at least one

Please cite this article in press as: Koprulu et al., Multi-cohort proteogenomic analyses reveal genetic effects across the proteome and diseasome, Cell (2026), https://doi.org/10.1016/j.cell.2026.03.049

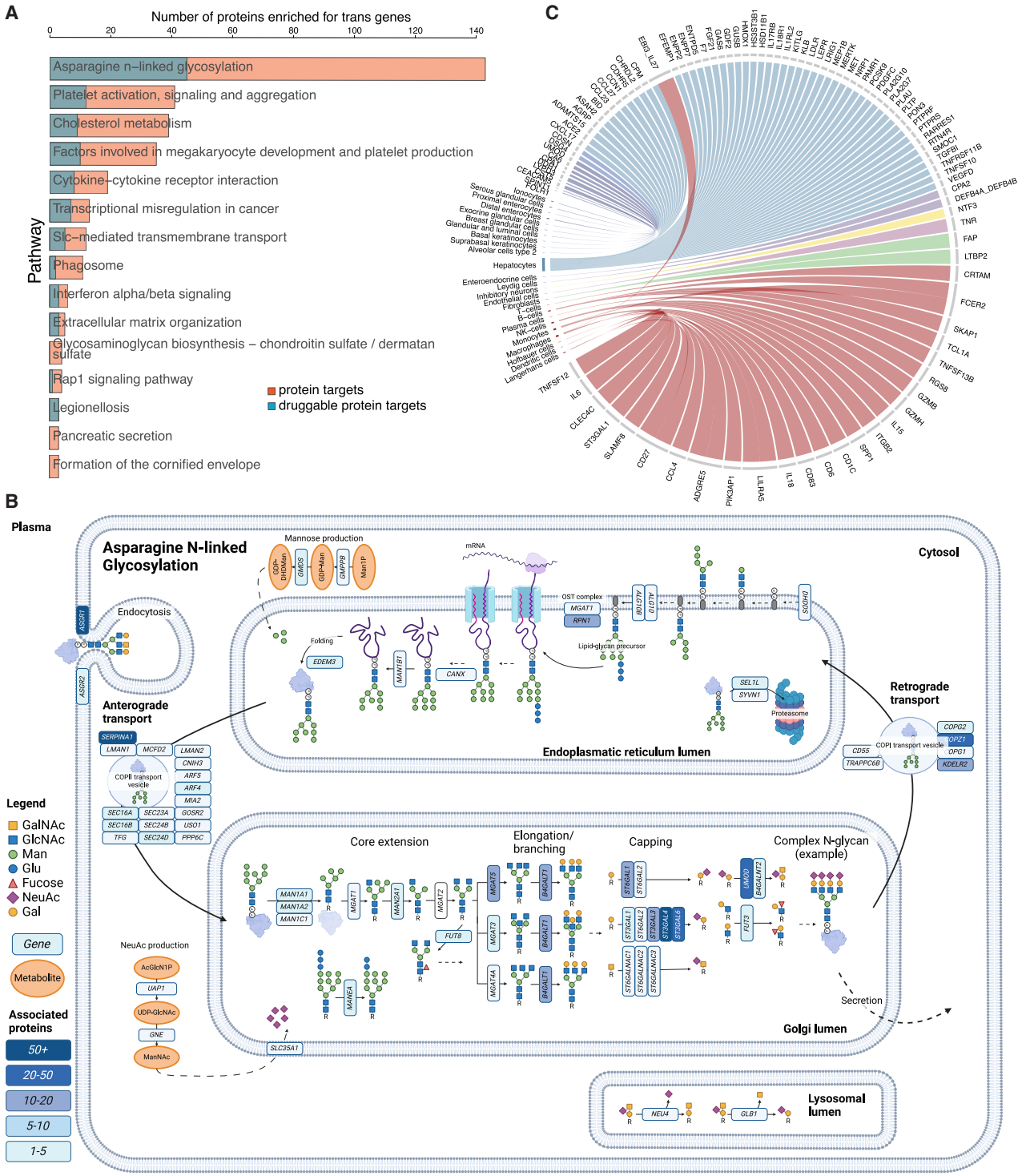


Figure 3. Effector gene assignment at *trans*-pQTLs identifies major pathways and cell types involved in plasma protein modulation

(A) Pathways most frequently enriched among top-prioritized *trans*-effector genes across all 439 protein targets with at least one significantly enriched (FDR < 5%) pathway. Proteins that are the targets of approved drugs or those in clinical development are highlighted in gray.

(B) Pathway diagram showing the role of *trans*-pQTL effector genes in "asparagine N-linked glycosylation" based on the R-HSA-446203 REACTOME pathway. Only members of the pathway that were assigned at least one *trans*-pQTL are displayed to retain readability, and colors indicate how often the gene was assigned

(legend continued on next page)

non-proteomic trait, e.g., disease susceptibility, in the GWAS catalog[30] (see STAR Methods and Figures 4A–4C). This included a >4-fold enrichment (odds ratio [OR]: 4.11; $p < 1.1 \times 10^{-230}$) of *trans*- compared with *cis*-pQTLs, demonstrating the importance of understanding the polygenic architecture of plasma proteins (Figure 4; Table S9A). We subsequently characterized the pleiotropic genetic variants into three categories: (1) "molecular pleiotropy" (associated with >5 proteins and ≤5 non-proteomic phenotypes), (2) "phenotypic pleiotropy" (associated with >5 non-proteomic phenotypes and ≤5 proteins), and (3) "unspecific pleiotropy" (associated with >5 proteins and >5 non-proteomic phenotypes) (Figures 4B and 4C).

More than half (332 out of 533) of the pQTLs that were pleiotropic at the protein level also showed phenotypic pleiotropy in the GWAS catalog. Associated effector genes were >2-fold enriched for enhanced expression in hepatocytes (OR = 2.82; FDR = 3.8×10^{-6}), aligning with the liver's role in whole-body homeostasis. This included 229 *trans*-pQTLs consistently enriched to likely act on protein complexes (OR = 1.83; $p < 2.7 \times 10^{-2}$), ligand-receptor pairs (OR = 2.73; $p < 3.8 \times 10^{-5}$), or pathway partners (OR = 7.66; $p < 3.5 \times 10^{-13}$) of associated protein targets compared with other modes of pleiotropy (Figure 4D), illustrating that part of the plasma proteome may be a surrogate of critical disease-predisposing mechanisms in tissues. For example, we identified rs10849448 as a *trans*-pQTL for eight protein targets, likely acting via lymphotoxin beta receptor (*LTBR*) (score 1.8), which encodes for LTBR, the receptor for lymphotoxin (LTA), which was among the associated protein targets. The same variant, or proxies thereof ($r^2 > 0.8$), has been reported to associate with >20 traits in the GWAS catalog, including celiac disease, chronic obstructive pulmonary disease, primary biliary cirrhosis, and different measures of lymphocytes. Although LTBR is not expressed on lymphocytes, signaling via LTBR is essential for the development of tertiary lymphoid structures that are associated with chronic inflammatory diseases.[31] Pharmacological blocking of LTBR signaling has further been shown to revert airway fibrosis in mouse models of smoking-induced chronic obstructive pulmonary disease.[32]

In general, genetic variants associated with multiple proteins might guide the interpretation of otherwise cryptic GWAS loci. We identified 285 pQTLs previously reported for non-proteomic traits in the GWAS catalog and among which associated proteins were significantly (FDR < 0.05) enriched for one or more pathways (Figure 4E; Table S9B). For example, we observed that two independent *trans*-pQTLs were 38-fold enriched (FDR < 5.5×10^{-4}) for proteins associated with keratinization and had been previously linked to the risk of acne. Although one of the putative effector genes, *ERRFI1*, had been linked to skin morphogenesis,[33] the role of the second one, *SEMA4B*, has so far been elusive,[34] and we provide evidence for a potential function in skin morphogenesis or homeostasis.

Contrasting proximal versus distal genetic regulation of plasma proteins reveals discordant phenotypic consequences

Successfully integrating proximal, i.e., *cis*-based, genetic regulation of plasma protein levels with disease risk variants has a high biological prior that the protein is involved in disease onset. We found evidence for 300 such protein-disease pairs, combining our results with >700 diseases from the FinnGen project[35] through causal inference methods for disease follow-up. However, only 73 examples were supported by both a dose-response relationship, via Mendelian randomization (MR), and statistical colocalization of genetic risk signals, emphasizing the need for complementary evidence for the genetic prioritization of candidate causal genes underlying diseases (Table S10A).

For one-third of the *cis*-based examples ($n = 115$ out of 300), we identified enough protein-specific variants acting in *trans* to systematically test for independent support by MR. We observed supporting evidence for a third ($n = 31$) of *cis*-based protein-disease examples by demonstrating a directionally consistent association when using only *trans*-pQTLs for causal inference ($p_{het} > 0.01$; Figure 5; Table S10B). This included proteins with known roles in blood, such as known pharmacological targets that act at the blood-tissue interface (e.g., proprotein convertase subtilisin/kexin type 9 [PCSK9] and coronary artery disease). However, we found no evidence for an enrichment of proteins specific to blood cells (e.g., lymphoid tissue: OR = 1.30; $p = 0.73$) or being actively secreted into blood (OR = 0.78; $p = 0.56$) among those with consistent effects in *cis* and *trans*. Proteins likely conveying risk within tissues included sulfotransferase family 2A member 1 (SULT2A1), linked to intrahepatic cholelithiasis of pregnancy (ICP) (Figure 5). Genetic variation near *SULT2A1*, which encodes a sulfotransferase with a role in the metabolism of endogenous compounds and drugs, has previously been shown to increase the risk for ICP,[36] and gallstones in general,[37] mainly attributed to increased hepatic expression and hence activity of the enzyme to produce supersaturated bile that promotes gallstone formation.[16] We provide evidence that potential distal regulators of SULT2A1 abundance, e.g., via the *trans*-pQTL rs16919533, which we assigned to *PANX1* with moderate confidence (gene score: 1.0), may also contribute to disease risk.

However, we also observed significant differences ($p_{het} < 0.01$) between results from well-powered *trans*-pQTL MRs not supporting ($n = 41$), or even opposing ($n = 14$), robust *cis*-pQTL evidence (Figure 5). For example, the well-documented effect of changes in sclerostin (SOST), a negative regulator of bone formation, on fracture risk observed with *cis*-pQTLs ($beta_{cis} = 1.34$, $p_{cis} < 5.6 \times 10^{-19}$) was null ($beta_{trans} = 0.02$, $p_{trans} = 0.84$) when considering 16 specific *trans*-pQTLs that cumulatively explained a higher amount of variation in plasma SOST levels

as an effector gene across protein targets. GalNAc, N-acetyl-D-galactosamine; GlcNAc, N-acetylglucosamine; Man, mannose; Glu, glucose; NeuAc, N-acetylneuraminic acid; Gal, galactose; created in https://BioRender.com.

(C) Chord diagram illustrating significant (FDR < 5%) enriched cell types among *trans*-effector genes across assayed protein targets. Colors are grouped by the Human Protein Atlas categories.

Relevant summary statistics can be found in Table S8.

Please cite this article in press as: Koprulu et al., Multi-cohort proteogenomic analyses reveal genetic effects across the proteome and disease, Cell (2026), https://doi.org/10.1016/j.cell.2026.03.049

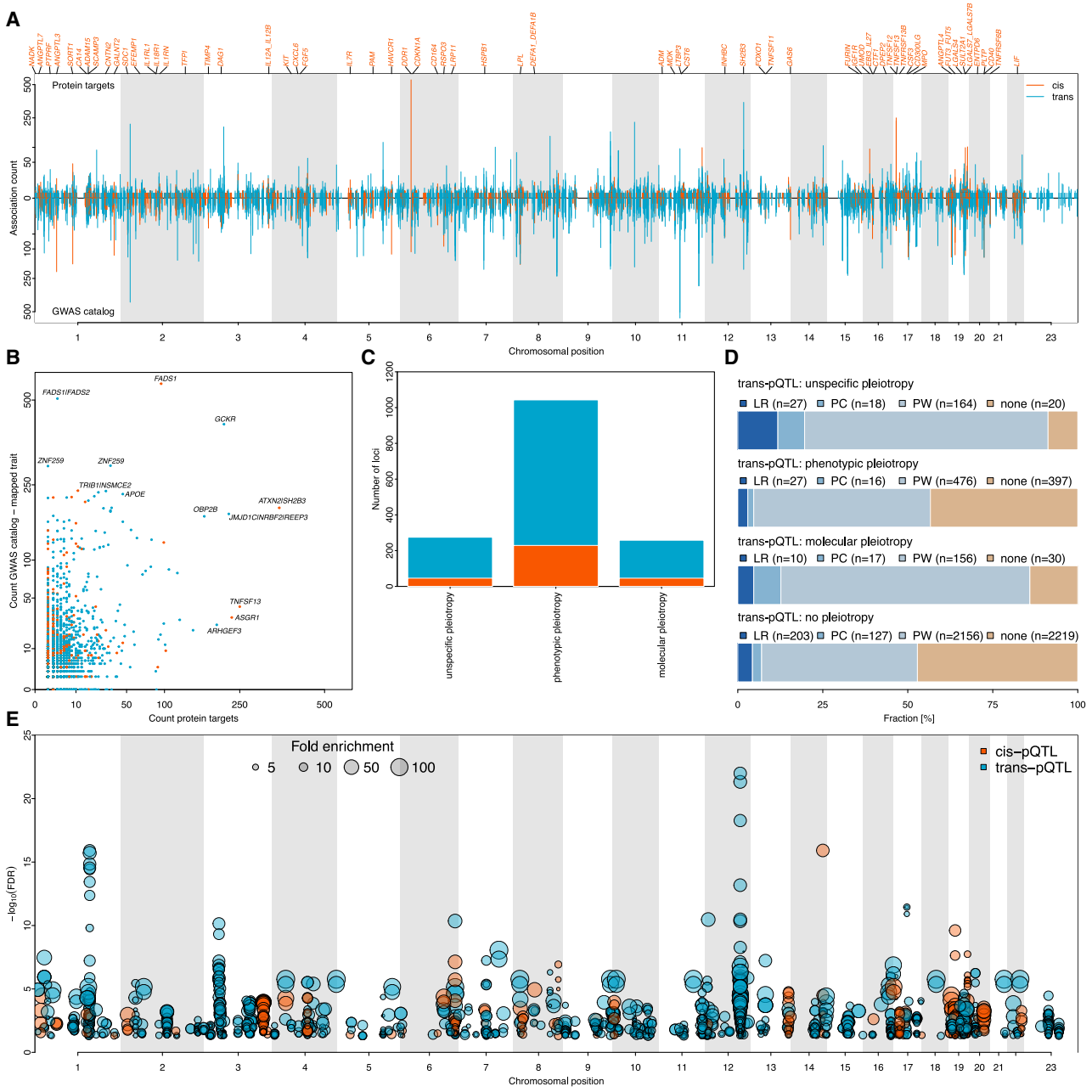


Figure 4. Molecular versus phenome-wide pleiotropy

(A) Counts of associated protein targets (upper) and associated phenotypes (lower) across 10,461 independent loci associated in the presented study. Phenotype associations were obtained from the GWAS catalog (including proxies) but by collapsing several studies of the same trait using the "mapped trait" column. Genetic variants acting as *cis*-pQTLs are shown in orange and otherwise in blue. Genes for *cis*-pQTLs associated with 10 or more phenotype categories are assigned on top of the plot. The *y* axis has been square-root transformed.

(B) Scatterplot opposing the counts from (A), with most pleiotropic effector genes for each category being annotated.

(C) Bar plot showing the number of pQTLs with evidence for different patterns of pleiotropy.

(D) Fraction of candidate effector genes for *trans*-pQTLs according to biological categories in relation to associated protein targets. LR, ligand-receptor pair; PC, protein complex; PW, pathway gene.

(E) Summary of pathway enrichment among pleiotropic pQTLs that have been reported at least once in the GWAS catalog. Dots indicate significantly enriched pathways (FDR > 5%). The size of the dot is proportional to the fold enrichment. Position along the *x* axis indicates the genomic location of the pQTL.

Please cite this article in press as: Koprulu et al., Multi-cohort proteogenomic analyses reveal genetic effects across the proteome and diseasome, Cell (2026), https://doi.org/10.1016/j.cell.2026.03.049

CellPress
OPEN ACCESS

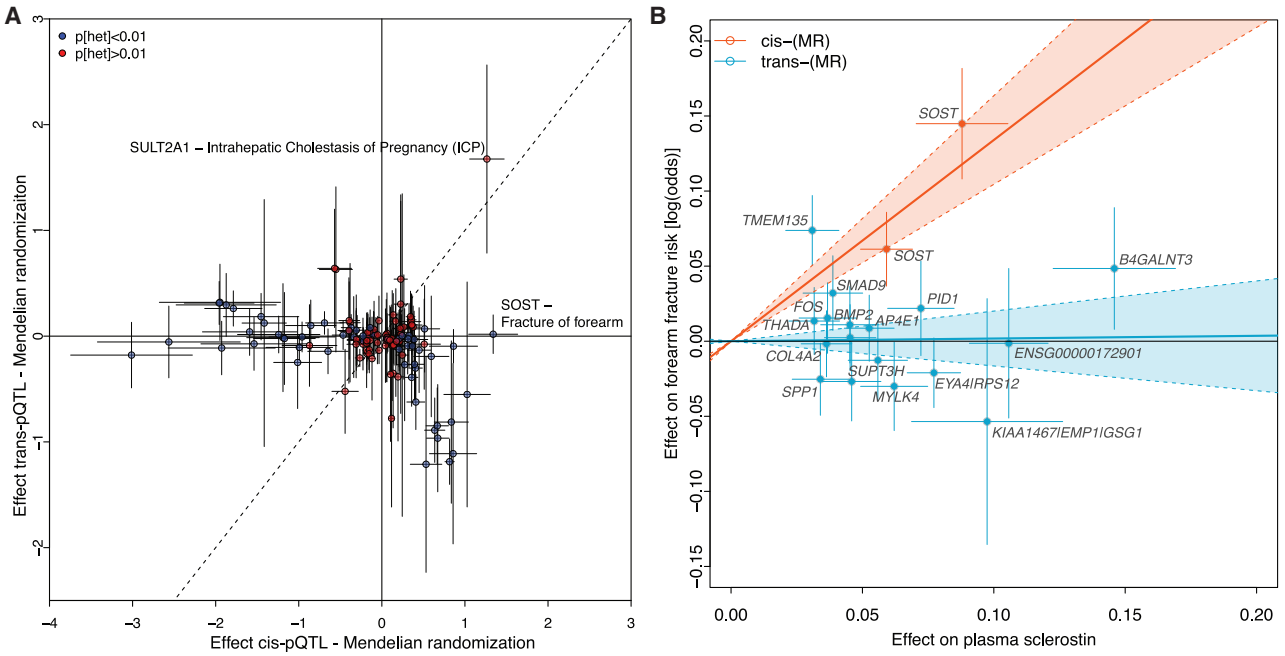


Figure 5. Effects of utilizing proximal (*cis*) and distal (*trans*) genetic variants to link proteins to disease risk

(A) Comparison of effect estimates from MR analyses using only *cis*-pQTLs (*x* axis) versus *trans*-pQTLs (*y* axis) for high-confidence protein-disease links based on *cis* findings. Colors indicate evidence for effect heterogeneity between estimates derived from *cis*- and *trans*-*pQTLs*.

(B) Causal effect estimates and 95% confidence intervals (CIs) using *cis*- (orange) and *trans*-pQTLs for plasma SOST levels on fracture risk.

Relevant summary statistics can be found in Table S10.

(Figure 5). We observed a similar null effect, when further excluding bone mineral density-associated *trans*-pQTLs in an additional sensitivity analysis.[38]

In general, naive incorporation of *trans*-pQTLs into MR-based causal inference resulted in even larger disagreement with <20% of findings having concordant support based on *cis*-instruments, likely driven by pleiotropy not captured by statistical techniques (15 out of 96 for *cis*-/*trans*-pQTL-based MR and 6 out of 230 for *trans*-pQTL-based MR; Table S10C). The few potentially true positive findings missed by our biologically informed approach included plasma levels of von Willebrand factor and von Willebrand disease (*cis*/*trans*-pQTL MR: beta = −3.42, $p < 2.9 \times 10^{-16}$).

Limited concordance between protein-phenotype associations identified using *cis*-focused genetic and observational studies

Blood proteins with a causal role in disease development are attractive targets for drug development and as markers of target engagement. We therefore systematically triangulated prevalent and incident plasma protein biomarker studies in up to 52,164 UKBB participants (using logistic regression and survival analysis) with high-confidence genetic inference in a pan-biobank effort, including up to 1,296,701 participants (using *cis*-MR and colocalization) for 517 diseases (Table S11).

We observed 193 protein-disease pairs with high-confidence genetic support, of which less than a quarter (n = 52) showed directionally concordant—and at least nominally statistically significant—support in biomarker studies (Figure 6A). A similar number of protein-disease examples (n = 59) showed statistically significant support but with opposing effect directions. The latter might be explained by compensatory mechanisms, but it exemplifies the need to contextualize purely genetically inferred "causality." Conversely, among the 52,887 protein-disease associations passing statistical significance ($p < 9.3 \times 10^{-8}$) in the survival analysis, only 0.06% (n = 33) had directionally concordant, high-confidence support from genetic analysis, and another 36 had significant but directionally discordant evidence (Figure 6A; Table S11). The overlap between the two approaches did not improve, and even led to a loss in, convergent examples when accounting for potential confounders in survival analyses or considering prevalent cases.

For 44 protein-disease pairs, we identified coherent evidence from genetic, prospective survival, and prevalent disease analyses, suggesting that those protein biomarkers may not only be relevant for the onset of the disease but also its persistence (Figure 6B). Among those, plasma furin levels stood out, being consistently associated with hypertension, myocardial infarction, and arterial fibrillation (Figure 6). The genetic locus encoding furin was first linked to blood pressure[39] and later to cardiovascular diseases, with recent proteomic studies also confirming our results in another ancestry.[40–42] Furin is an endopeptidase expressed in almost all tissues, with expression highest in the liver and with a broad substrate profile relevant for cardiovascular disease. Our results collectively point to the role of extracellular rather than intracellular furin. Almost all previously reported furin effects were described to occur in the *trans*-Golgi network, and extracellular furin has been mostly investigated in relation to

Please cite this article in press as: Koprulu et al., Multi-cohort proteogenomic analyses reveal genetic effects across the proteome and diseasome, Cell (2026), https://doi.org/10.1016/j.cell.2026.03.049

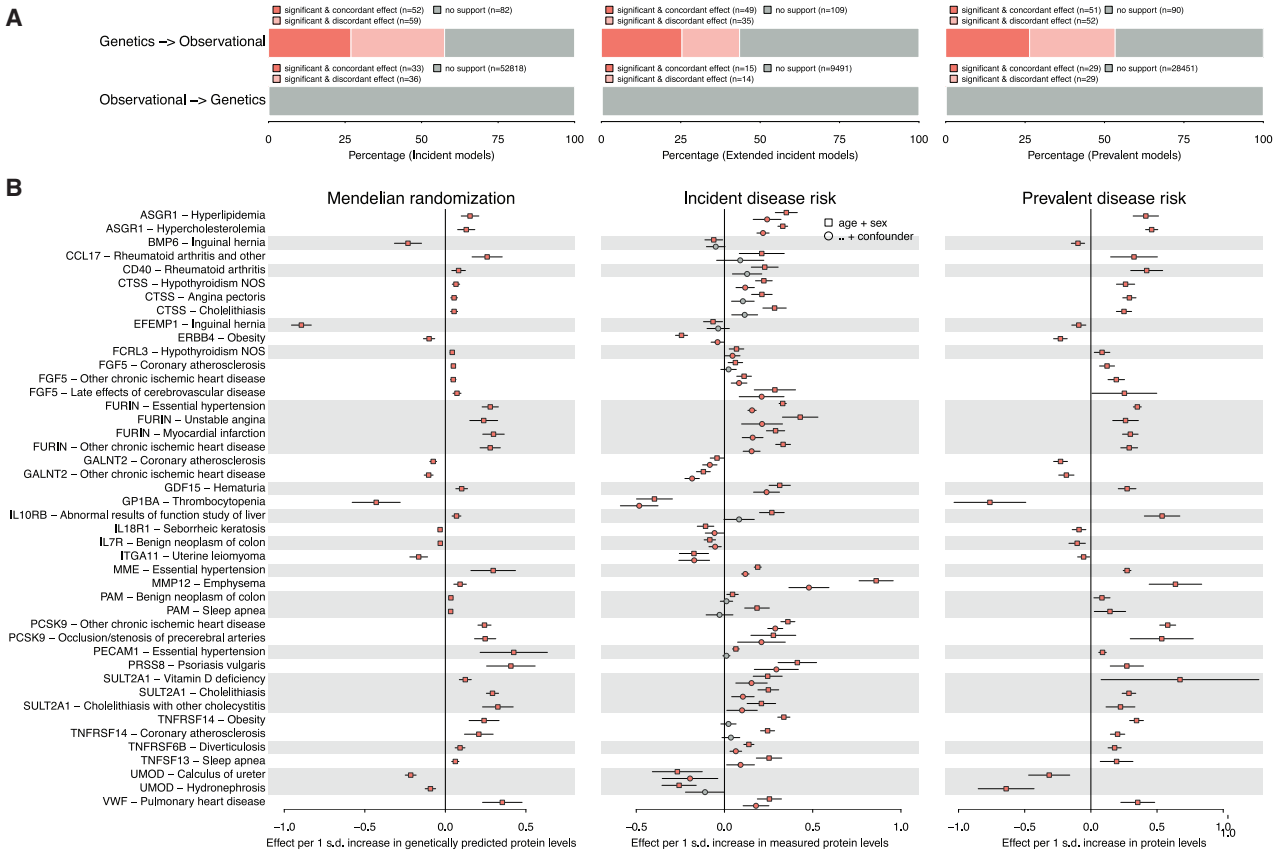


Figure 6. Concordance of genetically informed and observational protein biomarker studies

(A) Bar charts illustrating the fraction of protein-disease pairs with support from independent biomarker prioritization strategies. The upper panel illustrates starting with evidence from MR and colocalization using *cis*-pQTLs. The bottom panel illustrates starting with evidence from observational studies using survival analysis (left and middle) or prevalent disease status (right).

(B) Protein-disease examples with concordant support from genetic and (prospective) observational studies, illustrating effect estimates with 95% CIs from MR analyses (left), Cox proportional hazard models (middle, two different adjustment sets), and logistic regression models on prevalent disease status (right). Colors indicate the level of support as in (A).

Relevant summary statistics can be found in Table S11.

the cleavage of pathogen products.[43] The convergence of genetic and survival analysis (Figure 6B) may provide evidence for an extracellular role of furin in disease etiology, in contrast to the hypothesis that associations with the risk of hypertension and coronary artery disease are mediated via intracellular cleavage of brain natriuretic peptide (BNP)[44] or modulation of the LDL-receptor and cholesterol metabolism.[45] This hypothesis is supported by a lack of association between the lead *cis*-pQTL for furin (rs8027450-T) and circulating levels of both BNP (beta = 0.006; $p = 0.2$) and N-terminal prohormone of BNP (NT-proBNP, beta = 0.0058; $p = 0.3$). Pharmacological inhibition of furin in mouse models provided preliminary evidence for reduced atherosclerotic lesions and reduced vascular remodeling.[46]

trans-pQTL enrichment explains protein-disease signatures and can guide drug repurposing

The low convergence of evidence for protein-disease links from observational associations versus genetic inference from *cis*-pQTLs might, in part, be explained by circulating proteins being merely surrogates for causal disease processes within tissues. Investigating the value of many newly identified *trans*-pQTLs, we identified that >90% (280 out of 307 diseases with ≥5 significantly associated proteins) of disease biomarker signatures were significantly enriched (FDR < 0.05) for proteins associated with one or more of 170 pleiotropic pQTLs ($n = 139$ *trans*-pQTL; Table S12).

Triangulating the evidence from pQTLs with GWASs and biomarker analyses, we observed 58 examples where the protein targets associated with a given pleiotropic pQTL showed an enrichment for disease-associated plasma proteins, where the pQTL was also a GWAS locus for the same disease (Figure 6C). This supports the hypothesis that part of the circulating protein signature provides a readout of disease-predisposing processes within tissues.[47] For example, we observed a >50-fold enrichment (OR = 56.3, FDR < 9.5×10^{-20}) of proteins significantly associated with the onset of proteinuria among those proteins associated with a *trans*-pQTL, for which we prioritized *SHROOM3* as a candidate causal gene (Figure 6C;

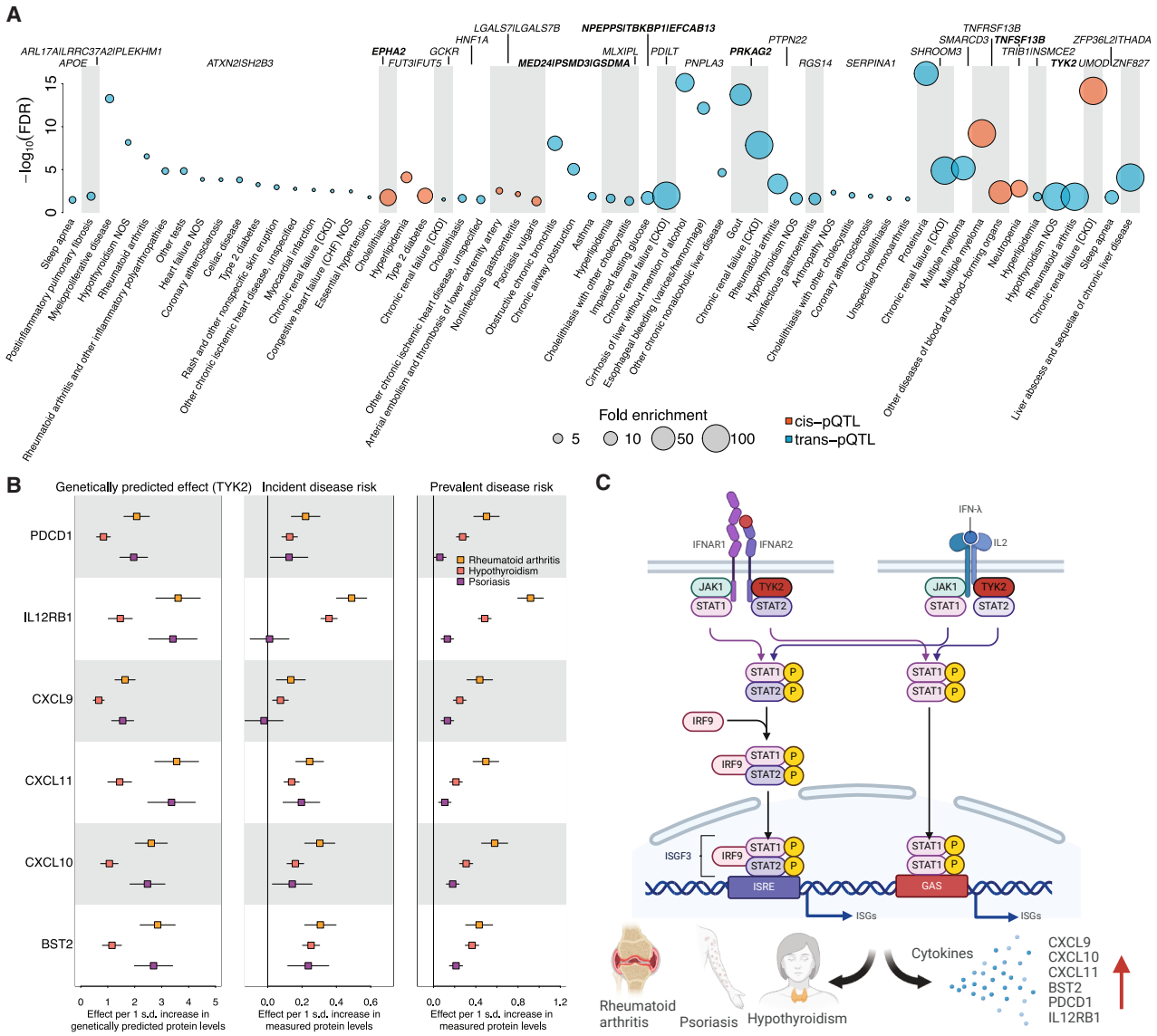


Figure 7. *trans*-pQTL enrichment explains prospective biomarker signatures and supports repurposing of TYK2 inhibitors

(A) Protein biomarker signatures significantly (FDR < 5%) enriched for pQTLs that have been reported to associate with the risk of the respective disease. The most likely effector genes are annotated. Genes that are the target of drugs are highlighted in bold.

(B) Forest plots showing associations (effect estimate with 95% CIs) between genetically predicted plasma protein levels (left; using rs34536443 in *TYK2*) and measured plasma protein levels (middle, right) with the risk/presence of psoriasis, hypothyroidism, and rheumatoid arthritis.

(C) Schematic displaying a potential role of impaired cytokine signaling due to loss of function in TYK2 for cytokine secretion and the onset of different autoimmune diseases. This figure was created using BioRender.com.

See also Figure S4.

Table S7). Knockdown of the mouse equivalent *Shroom3* has been shown to induce proteinuria by affecting podocyte foot process effacement,[48] impacting the cells forming part of the layer responsible for size-selective filtration of blood in the kidneys.

Plasma protein signatures non-randomly linked to known disease variants may further guide drug repurposing and biomarker identification for patient selection. For example, a missense variant (rs34536443, p.P1104A) in tyrosine kinase 2 (*TYK2*) acted as a *trans*-pQTL for multiple proteins (bone marrow stromal antigen 2 [BST2], C-X-C motif chemokine 9 [CXCL9], CXCL10, CXCL11, interleukin-12 receptor subunit beta-1 [IL-12RB1], and programmed cell death protein 1 [PDCD1]) (Figure 7). Higher plasma levels of all six protein targets were associated with a higher disease risk for rheumatoid arthritis, hypothyroidism, and psoriasis. There was also strong evidence for a shared genetic signal across these proteins and disease risks (HyPrColoc posterior probability [PP] = 98.9%) at the *TYK2* locus (Figures 7

and S4). Psoriasis is the approved indication for the *TYK2* (encoded at *TYK2*) inhibitor deucravacitinib, which mitigates inflammatory burden in patients,[49] supported by our genetic evidence (Figures 7 and S4). The associated circulating protein signature might therefore be indicative of early inflammatory dysregulation in patients at risk of these autoimmune diseases, and persistently elevated plasma levels among patients might help to identify those most likely to benefit from the treatment, as well as monitor treatment success, as demonstrated in a recent phase 2 trial of deucravacitinib.[50] This example illustrates the potential value of *trans*-pQTLs to guide biomarker prioritization for diverse purposes, benefiting patients as well as informing drug repurposing opportunities.

DISCUSSION

Broad-coverage, high-throughput proteomic technologies applied to large-scale patient and population cohorts offer a comprehensive molecular view that can substantially deepen our understanding of human health and disease.[6–24] Here, we present large-scale antibody-based proteogenomic meta-analyses consisting of 38 cohorts and up to 78,664 individuals to derive an expansive pQTL catalog of 24,738 fine-mapped pQTLs (including 5,040 in *cis* and 19,698 in *trans*). The multi-cohort design of our study provides consistent evidence across cohorts for >80% of the pQTLs identified in this study and also demonstrates that cohort and sample characteristics contribute only relatively minimally to observed heterogeneity of effects, in comparison to the overall strength of association, i.e., genetic effect size.

We demonstrate the importance of distal (i.e., *trans*) genetic regulation for circulating protein targets by identifying pathways, tissues, and cell types with specific, but also broader, roles. By integrating effector gene assignments across protein targets, we were able to reconstruct entire pathways relevant to plasma protein abundances through a data-driven approach based on in-human evidence. N-linked glycosylation emerged as the most frequently enriched pathway, in line with its role in the secretory pathway, involving the correct folding, trafficking, and secretion of the protein into blood or other tissues.[51–53] Enzymatic protein glycosylation is the most abundant post-translational modification, and our observation of enriched members of N- but not O-linked (a hallmark of intracellular proteins) glycosylation reflects a general finding of our study: that *trans*-effector genes are more likely to be directly linked to protein targets actively secreted into the blood.

Many studies now screen for novel biomarkers or drug targets using publicly available data, either with individual-level access to resources such as UKBB or with summary statistics from large genetic consortia, the latter being pursued purely *in silico* with strong underlying assumptions. The discrepancies in the inference arising from both approaches may explain some of the limited visible success to date. Even the high-confidence examples identified here, such as furin's role in diverse cardiovascular diseases, require follow-up before being evaluated as potential drug targets. We note that clinically useful protein biomarkers do not need to have genetic evidence to imply a causal role of the protein in disease development, but a clear and well-understood link to disease-predisposing mechanisms is essential for successful clinical application (e.g., neurofilament light chain used to diagnose neurodegenerative diseases).

By integrating different lines of orthogonal evidence, we provide clear examples of *trans*-pQTLs guiding biomarker prioritization for specific diseases. As a proof of principle, our results highlight seven *trans*-pQTLs directly supporting the use of NT-proBNP as a biomarker for (future) heart failure, with all *trans*-variants being associated with both heart failure and plasma NT-proBNP levels. Similarly, we exemplified how *trans*-pQTLs, when integrated with prospective biomarker analysis and large-scale GWASs of different diseases, can point to protein biomarkers to support drug repurposing, such as TYK2 inhibitors for rheumatoid arthritis.[54] Collectively, this provides clear conceptual advances beyond *cis*-focused causal inference techniques, which have predominantly been applied in proteogenomic studies to date, highlighting how *trans*-pQTLs can inform protein biomarker detection for diseases and drug development as well as improve our understanding of identified susceptibility loci for these diseases.

MR has become commonly used for genetically guided drug-target discovery. We observed discordant effects for well-curated *trans*-pQTL versus biologically plausible *cis*-pQTL disease associations. Although we acknowledge that limitations might remain for specific examples, our results did not reveal a systematic reason explaining the convergence or segregation of *cis* and *trans* effects on the diseases studied, and we propose a staged approach to guide the evaluation of *trans*-pQTLs in the context of MR studies. This includes biologically informed exclusion of instruments with evidence for pleiotropic effects on either the proteome or phenome, in addition to testing the alignment of *cis*- and *trans*-pQTL effect sizes and directions.

Our study demonstrates the potential of larger multi-cohort studies in expanding our understanding of the genetic regulation of the circulating proteome and of the insights into disease mechanisms and therapeutic development that proteogenomics can deliver.

Limitations of the study

All proteomic technologies currently only provide a partial coverage of the proteins detectable in circulation, including their many isoforms and post-translational modifications. Similarly, we had incomplete coverage of the range of genomic variants across cohorts, as most cohorts had array-based imputed genomic data, limiting our ability to assess rare variants and utilize statistical approaches such as fine-mapping to their full extent. Our study was conducted using predominantly European samples due to measurement availability. We believe our study provides not only a valuable multi-cohort proteogenomic resource to the community but also a template for the biological and clinical insights that can be derived from such efforts, including future studies with even greater power, allele and protein coverage, and genomic diversity.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to Claudia Langenberg (claudia.langenberg@qmul.ac.uk).

Materials availability

This study did not generate any new, unique reagents.

Data and code availability

- Access to the UKBB genomic, proteomic, and phenotype data is open to all approved health researchers (http://www.ukbiobank.ac.uk/). This research has been conducted using the UKBB resource under application no. 44448.
- Genome-wide summary statistics for all protein targets in this study will be available on https://omicscience.org/ upon publication.
- All remaining data have been accessed via publicly available links provided in the key resources table.
- Associated code and scripts for the analyses are available at https://github.com/comp-med/scallop-ukbb-ma.

CONSORTIA

The Estonian Biobank Research Team consists of Andres Metspalu, Lili Milani, Tõnu Esko, Reedik Mägi, Mait Metspalu, Mari Nelis, and Georgi Hudjashov.

ACKNOWLEDGMENTS

M.K. was funded by the Gates Cambridge Trust for her PhD. K.S.-B. is supported by Cancer Research UK (grant nos. C8221/A29017 and C16077/A29186) and by a UKRI grant no. 10063259. C.F.O. is supported by a Nuffield Department of Population Health (NDPH) studentship. J. Kuliesius was funded by the MRC Doctoral Training Programme in Precision Medicine (MR/N013166/1). J.Ä. has received funding from the Swedish Research Council (Vetenskapsrådet, grant nos. 2019-01015 and 2020-0243) and from the Swedish Heart Lung Foundation (Hjärt-Lungfonden, grant nos. 2021-0357 and 2024-0486). A.S.B. is supported by an award (SP/F/23/150048) from the British Heart Foundation (BHF) and the German Centre for Cardiovascular Research (DZHK). S.R.C. was supported by a Sir Henry Dale Fellowship, jointly funded by the Wellcome Trust and the Royal Society (221890/Z/20/Z). The recruitment of the ASAP cohort was headed by Professor Anders Franco-Cereceda. T.E. was supported by the Estonian Research Council grant PUT (PRG1291). Where authors are identified as personnel of the International Agency for Research on Cancer/World Health Organization, the authors alone are responsible for the views expressed in this article, and they do not necessarily represent the decisions, policy, or views of the International Agency for Research on Cancer/World Health Organization. P.W.F. was supported by grants from the European Commission (IMI SOPHIA – 875534), the Swedish Research Council, the Novo Nordisk Foundation, Vinnova, the Swedish Foundation for Strategic Research (LUDC-IRC, 15-0067), and Exodiab (2009-1039). J.F. is supported by a European Research Council (ERC) Consolidator grant (grant agreement no. 101001678), the Dutch Research Council (NWO) VICI grant (VI.C.202.022) and KIC grant (KICH1.LWV04.21.013), the AMMODO Science Award 2023 for Biomedical Sciences from Stichting Ammodo, and the Dutch Heart Foundation AtheroNeth project. J.M.G. is funded by the Department of Veterans Affairs. S.H. is supported by a National Institutes of Health (NIH) research grant U01AG083829. C. Hayward was funded by an MRC Human Genetics Unit program (QTL in Health and Disease) (grant U.MC_UU_00007/10). N.H.-P. was supported by a Medium-Term Research Fellowship from the European Academy of Allergy and Clinical Immunology (EAACI) and a Long-Term Research Fellowship from the European Respiratory Society (ERS) (LTRF202101-00861). J.C.H. acknowledges support from the British Heart Foundation, the National Institute for Health and Care Research Oxford Biomedical Research Centre, the British Heart Foundation Oxford Centre of Research Excellence, and the Nuffield Department of Population Health, University of Oxford, UK. Å.J. was funded by the Swedish Research Council, the Swedish Heart Lung Foundation, and the Brain Foundation. A.K. was supported by the Estonian Research Council grant PUT (PRG1291). L. Klaric was supported by an RCUK Innovation Fellowship from the National Productivity Investment Fund (MR/R026408/1). J. Long was partially supported by grant nos. R01CA247987, R01CA293996, R01MD015396, and R01CA249863. K.M. was funded by the Swedish Research Council and Olle Engkvist Foundation for proteomics analyses. M.N. was funded by Novo Nordisk Foundation (NNF21OC0071050). X.S. received a National Key Research and Development Program grant (2022YFF1202105), a National Natural Science Foundation of China (NSFC) grant (12171495), and a grant from the Swedish Research Council (Vetenskapsrådet) (2022-01309). J.S. acknowledges the Knut and Alice Wallenberg Foundation for funding the Human Protein Atlas. K.S. is supported by the Biomedical Research Program at WCMQ, a program funded by the Qatar Foundation and by Qatar National Research Fund grants NPRP11C-0115-180010 and ARG01-0420-230007. J.B.J.v.M. was funded by ReumaNL, project no. LLP34. U.V. was supported by the Estonian Research Council grant PUT (PRG1291). N.J.W. was supported by MRC grants MR/V033867/1, MC_PC_21036, and MC_UU_00006/1. S.W. is supported by the Novo Nordisk Foundation (grant no. NNF21OC0066981). We thank Maria Karaleftheri and Emmanouil Tsafantakis for their contributions to the HELIC-Manolis and HELIC-Pomak cohorts. A.Z. is funded by the Netherlands Organization for Scientific Research NWO-VICI grant VI.C.232.074, the NWO Gravitation grant ExposomeNL 024.004.017, the NWO KIC grant KICH1.LWV04.21.01, the ZonMW ME/CFS grant 10091012110015, the EU Horizon Europe Program grant INITIALISE (101094099), and the EU Horizon Europe Program grant DarkMatter (“ID-DarkMatter-NCD” [project no. 101136582]). M.Z. thanks Christin Beier, Proteomics Platform, MDC, for the Olink measurements and her technical expertise. B.M.K. was supported by the Chinese Academy of Medical Sciences (CAMS) Innovation Fund for Medical Science (CIFMS), China (grant no. 2024-I2M-2-001-1). The work was co-funded by the European Union (ERC, GenDrug, 101116072) to M.P. C.L. was supported by MRC grants MR/V033867/1 and MC_PC_21036, the DZHK (German Centre for Cardiovascular Research) site Berlin, and the European Union (ERC Consolidator Grant D-MAPS to C.L., grant no. 101231435). Views and opinions expressed herein are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We thank Laura Temperley for proofreading this manuscript to improve its clarity. We thank Werner Römisch-Margl and Gabi Kastenmüller for their support in making the results of this study publicly available through omicscience.org. This research has been conducted using the UKBB resource under application no. 44448 and uses data provided by patients and collected by the NHS as part of their care and support. Figures 3 and 7 and the graphical abstract were created using BioRender. Cohort-level acknowledgments can be found in Table S1B.

AUTHOR CONTRIBUTIONS

Performed quality control, data preparation, visualization, or statistical and bioinformatic analyses relevant to this manuscript, M.K., K.S.-B., B.R.F., E.M.-D., J. Luan, Å.K.H., C.F.O., J. Kuliesius, L.R., A.R., L. Kohleick, and M.P.; designed the analyses and drafted the manuscript, M.K., M.P., and C.L.; provided samples and conducted cohort-level quality control and data analyses, F.A., J.Ä., T.L.A., BeLOVE Study Group, H.M.B., S.B., M.B., A.S.B., Z.C., K. Cho, R.J.C., S.R.C., K. Czene, J.D., G.D., S.E., P.E., N.E., T.E., Estonian Biobank Research Team, A.F.-I., P.W.F., J.F., J.M.G., M.G., C.G., A. Gilly, H.G., M.J.G., S.G., A. Göteson, P.F.L.H., O.H., S.E.H., C. Hayward, C. Herder, N.H.-P., Z.H., R.F.H., J.C.H., S.H., S.-J.H., C.J., Å.J., L.J., A.K., N.D.K., P.F.K., L. Klaric, J. Kraft, M.L., D.L., L. Li, L. Lind, J. Long, N.M.-C., E.M., S.K.M., P.M., K.M., P.L.M., F.M., M.N., Y.-C.P., E.P., J.P., J.R.P., G.P., O.P., B.P.P., S.R., M.R., P.D.R., S.S., SCALLOP Consortium, X.S., J.M.S., A.S., J.G.S., T.M.S., K.S., J.S., B.T., E.V.-M., C.L.V., J.B.J.v.M., A.V., U.V., L.W., R.G.W., N.J.W., J.E.W., R.K.W., J.F.W., S.W., S.Y., D.Z., E.Z., J.H.Z., A.Z., D.V.Z., and M.Z.; contributed to design and supervision of key aspects of the study, B.M.K., A.C.P., A.M., M.P., and C.L.; conceptualized and designed the study, A.M. and C.L.; and all authors contributed to the interpretation of the results and critically reviewed the manuscript.

DECLARATION OF INTERESTS

E.M.-D. is an employee of Pfizer. Å.K.H. is a full-time employee of Pfizer. J.Ä. reports payment or honoraria for lectures from AstraZeneca, Boehringer Ingelheim, and Novartis and participation on advisory boards with AstraZeneca, Astellas, and Boehringer Ingelheim. A.S.B. has received grants unrelated to this work from AstraZeneca, Bayer, Biogen, BioMarin, Bioverativ, Novartis, and Sanofi. N.E. reports institutional research grants from Pfizer and Bristol-Myers Squibb/Pfizer. A. Gilly is now an employee of Regeneron Genetics Center. S.G. had previous employment at Sence Research. O.H. is an employee of Lund University and Eli Lilly, and he has previously acquired research support (for Lund University) from AVID Radiopharmaceuticals, Biogen, C2N Diagnostics, Eli Lilly, Eisai, Fujirebio, GE Healthcare, and Roche. In the past 2 years, O.H. has received consultancy/speaker fees from Alzpath, BioArctic, Biogen, Bristol-Myers Squibb, Eisai, Eli Lilly, Fujirebio, Merck, Novartis, Novo Nordisk, Roche, Sanofi, and Siemens. Z.H. is employed by Thermo Fisher Scientific. J.C.H. works at the Nuffield Department of Population Health, which receives research grants from industry that are governed by University of Oxford contracts that protect its independence and has a staff policy of not taking personal payments from industry; further details can be found at https://www.ndph.ox.ac.uk/about/independence-of-research. P.F.K. is now an employee of Alnylam Pharmaceuticals, Inc. L. Klaric is currently employed by, and has share options in, BioAge Labs. N.M.-C. has received speaker/consultancy fees from Biogen, BioArctic, Eli Lilly, Novo Nordisk, Merck, and Owkin. J.R.P. has received personal fees (via his employing institution) from Merck KGaA (lectures), research support from Merck KGaA (grant), personal fees from Novo Nordisk (lectures), personal fees from IQVIA (Boehringer Ingelheim Adjudication Committees), and travel support from Sanofi. He has also received non-financial support as co-CI of a JDRF/Breakthrough T1D-funded trial (NCT03899402) from AstraZeneca (donation of investigational medicinal product to the US site only) and Novo Nordisk (donation of investigational medicinal product to UK site only; supplementary financial support [to mitigate a budget cut during the COVID-19 pandemic]). G.P. is now an employee of Illumina. M.R. received fees for consulting, lecturing, or serving on the advisory boards of AstraZeneca, Boehringer-Ingelheim, Echosens, Eli Lilly, Madrigal, Merck-MSD, Novo Nordisk, and Target RWE and has performed investigator-initiated research with support from Boehringer-Ingelheim and Novo Nordisk to the DDZ. The research of M.R. is supported by grants from the German Research Foundation (DFG; RTG/GRK 2576), the European Community (HORIZON-HLTH-2022-STAYHLTH-02-01: Panel A) to the INTERCEPT-T2D consortium, and the Schmutzler-Stiftung. S.S. is now an employee of Novo Nordisk. Contribution to the work described in this manuscript was conducted prior to S.S.'s employment at Novo Nordisk. Novo Nordisk had no involvement in the study design, data collection, analysis, interpretation, or writing of the manuscript. X.S. is the founder of Quantix BioSciences AB, which commercializes data science and technologies for analyzing protein complex biomarkers. J.S. is a scientific advisor for ABC Labs AB and has, unrelated to this work, received speaker fees or travel support from Olink, Luminex, Illumina, Oxford Nanopore, Roche Diagnostics, and AlamarBioscience and, via KTH, conducted contract research for Capitainer and Luminex. A.S. reports research grants from Bristol-Myers Squibb/Pfizer. J.S. reports direct or indirect stock ownership in companies (Sence Research AB, Symptoms Europe AB, MinForskning AB, and Anagram kommunikation AB) providing services to companies and authorities in the health sector, including Amgen, AstraZeneca, Bayer, Boehringer, Eli Lilly, Gilead, GSK, Göteborg University, Itrim, Ipsen, Janssen, Karolinska Institutet, LIF, Linköping University, Novo Nordisk, Parexel, Pfizer, Region Stockholm, Region Uppsala, Sanofi, STRAMA, Takeda, TLV, Uppsala University, Vifor Pharma, and WeMind. R.K.W. acted as a consultant for Takeda Pharmaceuticals; received unrestricted research grants from Takeda, Johnson & Johnson, Tramedico, and Ferring; and received speaker's fees from MSD, AbbVie, and Janssen Pharmaceuticals. A.Z. received a speaker's fee from Nestle. A.M. was an employee of Pfizer R&D until September 1, 2025, but is now working full-time for Novo Nordisk A/S.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Proteomic measurements
  - Genome-wide meta-analyses of protein levels
  - Downstream analyses
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Genome-wide meta-analyses of protein levels
  - Identification of regional sentinel variants
  - Fine mapping
  - Concordance of the identified pQTLs
  - Variance explained by identified pQTLs
  - pQTLs and protein-related characteristics
  - Impact of cohort characteristics on heterogeneity
  - Characterization of variant consequences
  - Colocalization with gene expression levels
  - Machine learning-based effector gene assignment
  - Pathway, tissue, and cell-type enrichment
  - Phenotypic follow up
  - Cis-based pan-biobank phenotypic follow up
  - Comparison of observational and genetic analyses
  - Enrichment analysis anchored on pQTLs
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2026.03.049.

Received: April 24, 2025
Revised: October 9, 2025
Accepted: March 26, 2026

REFERENCES

1. Lappalainen, T., and MacArthur, D.G. (2021). From variant to function in human disease genetics. Science *373*, 1464–1468. https://doi.org/10.1126/science.abi8207.
2. Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., and Meyre, D. (2019). Benefits and limitations of genome-wide association studies. Nat. Rev. Genet. *20*, 467–484. https://doi.org/10.1038/s41576-019-0127-1.
3. Findlay, G.M. (2021). Linking genome variants to disease: scalable approaches to test the functional impact of human mutations. Hum. Mol. Genet. *30*, R187–R197. https://doi.org/10.1093/hmg/ddab219.
4. Kircher, M., Xiong, C., Martin, B., Schubach, M., Inoue, F., Bell, R.J.A., Costello, J.F., Shendure, J., and Ahituv, N. (2019). Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. Nat. Commun. *10*, 3583. https://doi.org/10.1038/s41467-019-11526-w.
5. Abell, N.S., DeGorter, M.K., Gloudemans, M.J., Greenwald, E., Smith, K.S., He, Z., and Montgomery, S.B. (2022). Multiple causal variants underlie genetic associations in humans. Science *375*, 1247–1254. https://doi.org/10.1126/science.abj5117.
6. Suhre, K., Arnold, M., Bhagwat, A.M., Cotton, R.J., Engelke, R., Raffler, J., Sarwath, H., Thareja, G., Wahl, A., deLisle, R.K., et al. (2017). Connecting genetic risk to disease end points through the human blood plasma proteome. Nat. Commun. *8*, 14357. https://doi.org/10.1038/ncomms14357.
7. Sun, B.B., Maranville, J.C., Peters, J.E., Stacey, D., Staley, J.R., Blackshaw, J., Burgess, S., Jiang, T., Paige, E., Surendran, P., et al. (2018).

Genomic atlas of the human plasma proteome. Nature *558*, 73–79. https://doi.org/10.1038/s41586-018-0175-2.

8. Folkersen, L., Fauman, E., Sabater-Lleal, M., Strawbridge, R.J., Frånberg, M., Sennblad, B., Baldassarre, D., Veglia, F., Humphries, S.E., Rauramaa, R., et al. (2017). Mapping of 79 loci for 83 plasma protein biomarkers in cardiovascular disease. PLoS Genet. *13*, e1006706. https://doi.org/10.1371/journal.pgen.1006706.
9. Yao, C., Chen, G., Song, C., Keefe, J., Mendelson, M., Huan, T., Sun, B.B., Laser, A., Maranville, J.C., Wu, H., et al. (2018). Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. Nat. Commun. *9*, 3268. https://doi.org/10.1038/s41467-018-05512-x.
10. Gilly, A., Park, Y.C., Png, G., Barysenka, A., Fischer, I., Bjørnland, T., Southam, L., Suveges, D., Neumeyer, S., Rayner, N.W., et al. (2020). Whole-genome sequencing analysis of the cardiometabolic proteome. Nat. Commun. *11*, 6336. https://doi.org/10.1038/s41467-020-20079-2.
11. Ferkingstad, E., Sulem, P., Atlason, B.A., Sveinbjornsson, G., Magnusson, M.I., Styrmisdottir, E.L., Gunnarsdottir, K., Helgason, A., Oddsson, A., Halldorsson, B.V., et al. (2021). Large-scale integration of the plasma proteome with genetics and disease. Nat. Genet. *53*, 1712–1721. https://doi.org/10.1038/s41588-021-00978-w.
12. Gudjonsson, A., Gudmundsdottir, V., Axelsson, G.T., Gudmundsson, E.F., Jonsson, B.G., Launer, L.J., Lamb, J.R., Jennings, L.L., Aspelund, T., Emilsson, V., et al. (2022). A genome-wide association study of serum proteins reveals shared loci with common diseases. Nat. Commun. *13*, 480. https://doi.org/10.1038/s41467-021-27850-z.
13. Katz, D.H., Tahir, U.A., Bick, A.G., Pampana, A., Ngo, D., Benson, M.D., Yu, Z., Robbins, J.M., Chen, Z.Z., Cruz, D.E., et al. (2022). Whole Genome Sequence Analysis of the Plasma Proteome in Black Adults Provides Novel Insights Into Cardiovascular Disease. Circulation *145*, 357–370. https://doi.org/10.1161/CIRCULATIONAHA.121.055117.
14. Png, G., Barysenka, A., Repetto, L., Navarro, P., Shen, X., Pietzner, M., Wheeler, E., Wareham, N.J., Langenberg, C., Tsafantakis, E., et al. (2021). Mapping the serum proteome to neurological diseases using whole genome sequencing. Nat. Commun. *12*, 7042. https://doi.org/10.1038/s41467-021-27387-1.
15. Pietzner, M., Wheeler, E., Carrasco-Zanini, J., Kerrison, N.D., Oerton, E., Koprulu, M., Luan, J., Hingorani, A.D., Williams, S.A., Wareham, N.J., et al. (2021). Synergistic insights into human health from aptamer- and antibody-based proteomic profiling. Nat. Commun. *12*, 6822. https://doi.org/10.1038/s41467-021-27164-0.
16. Pietzner, M., Wheeler, E., Carrasco-Zanini, J., Cortes, A., Koprulu, M., Wörheide, M.A., Oerton, E., Cook, J., Stewart, I.D., Kerrison, N.D., et al. (2021). Mapping the proteo-genomic convergence of human diseases. Science *374*, eabj1541. https://doi.org/10.1126/science.abj1541.
17. Zhang, J., Dutta, D., Köttgen, A., Tin, A., Schlosser, P., Grams, M.E., Harvey, B., CKDGen Consortium, Yu, B., Boerwinkle, E., et al. (2022). Plasma proteome analyses in individuals of European and African ancestry identify cis-pQTLs and models for proteome-wide association studies. Nat. Genet. *54*, 593–602. https://doi.org/10.1038/s41588-022-01051-w.
18. Folkersen, L., Gustafsson, S., Wang, Q., Hansen, D.H., Hedman, Å.K., Schork, A., Page, K., Zhernakova, D.V., Wu, Y., Peters, J., et al. (2020). Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. Nat. Metab. *2*, 1135–1148. https://doi.org/10.1038/s42255-020-00287-2.
19. Koprulu, M., Carrasco-Zanini, J., Wheeler, E., Lockhart, S., Kerrison, N.D., Wareham, N.J., Pietzner, M., and Langenberg, C. (2023). Proteogenomic links to human metabolic diseases. Nat. Metab. *5*, 516–528. https://doi.org/10.1038/s42255-023-00753-7.
20. Sun, B.B., Chiou, J., Traylor, M., Benner, C., Hsu, Y.H., Richardson, T.G., Surendran, P., Mahajan, A., Robins, C., Vasquez-Grinnell, S.G., et al. (2023). Plasma proteomic associations with genetics and health in the UK Biobank. Nature *622*, 329–338. https://doi.org/10.1038/s41586-023-06592-6.
21. Eldjarn, G.H., Ferkingstad, E., Lund, S.H., Helgason, H., Magnusson, O.T., Gunnarsdottir, K., Olafsdottir, T.A., Halldorsson, B.V., Olason, P.I., Zink, F., et al. (2023). Large-scale plasma proteomics comparisons through genetics and disease associations. Nature *622*, 348–358. https://doi.org/10.1038/s41586-023-06563-x.
22. Dhindsa, R.S., Burren, O.S., Sun, B.B., Prins, B.P., Matelska, D., Wheeler, E., Mitchell, J., Oerton, E., Hristova, V.A., Smith, K.R., et al. (2023). Rare variant associations with plasma protein levels in the UK Biobank. Nature *622*, 339–347. https://doi.org/10.1038/s41586-023-06547-x.
23. Repetto, L., Chen, J., Yang, Z., Zhai, R., Timmers, P.R.H.J., Feng, X., Li, T., Yao, Y., Maslov, D., Timoshchuk, A., et al. (2024). The genetic landscape of neuro-related proteins in human plasma. Nat. Hum. Behav. *8*, 2222–2234. https://doi.org/10.1038/s41562-024-01963-z.
24. Carland, C., Png, G., Malarstig, A., Kho, P.F., Gustafsson, S., Michaelsson, K., Lind, L., Tsafantakis, E., Karaleftheri, M., Dedoussis, G., et al. (2023). Proteomic analysis of 92 circulating proteins and their effects in cardiometabolic diseases. Clin. Proteom. *20*, 31. https://doi.org/10.1186/s12014-023-09421-0.
25. GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318–1330. https://doi.org/10.1126/science.aaz1776.
26. Kachuri, L., Hoffmann, T.J., Jiang, Y., Berndt, S.I., Shelley, J.P., Schaffer, K.R., Machiela, M.J., Freedman, N.D., Huang, W.Y., Li, S.A., et al. (2023). Genetically adjusted PSA levels for prostate cancer screening. Nat. Med. *29*, 1412–1423. https://doi.org/10.1038/s41591-023-02277-9.
27. Wang, G., Sarkar, A., Carbonetto, P., and Stephens, M. (2020). A simple new approach to variable selection in regression, with application to genetic fine mapping. J. R. Stat. Soc. Series B Stat. Methodol. *82*, 1273–1300. https://doi.org/10.1111/rssb.12388.
28. Knecht, S., Eberl, H.C., Kreisz, N., Ugwu, U.J., Starikova, T., Kuster, B., and Wilhelm, S. (2023). An Introduction to Analytical Challenges, Approaches, and Applications in Mass Spectrometry-Based Secretomics. Mol. Cell. Proteomics *22*, 100636. https://doi.org/10.1016/j.mcpro.2023.100636.
29. Breitling, J., and Aebi, M. (2013). N-linked protein glycosylation in the endoplasmic reticulum. Cold Spring Harb. Perspect. Biol. *5*, a013359. https://doi.org/10.1101/cshperspect.a013359.
30. Sollis, E., Mosaku, A., Abid, A., Buniello, A., Cerezo, M., Gil, L., Groza, T., Güneş, O., Hall, P., Hayhurst, J., et al. (2023). The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. Nucleic Acids Res. *51*, D977–D985. https://doi.org/10.1093/nar/gkac1010.
31. Pitzalis, C., Jones, G.W., Bombardieri, M., and Jones, S.A. (2014). Ectopic lymphoid-like structures in infection, cancer and autoimmunity. Nat. Rev. Immunol. *14*, 447–462. https://doi.org/10.1038/nri3700.
32. Conlon, T.M., John-Schuster, G., Heide, D., Pfister, D., Lehmann, M., Hu, Y., Ertüz, Z., Lopez, M.A., Ansari, M., Strunz, M., et al. (2020). Inhibition of LTβR signalling activates WNT-induced regeneration in lung. Nature *588*, 151–156. https://doi.org/10.1038/s41586-020-2882-8.
33. Ferby, I., Reschke, M., Kudlacek, O., Knyazev, P., Pantè, G., Amann, K., Sommergruber, W., Kraut, N., Ullrich, A., Fässler, R., et al. (2006). Mig6 is a negative regulator of EGF receptor-mediated skin morphogenesis and tumor formation. Nat. Med. *12*, 568–573. https://doi.org/10.1038/nm1401.
34. Petridis, C., Navarini, A.A., Dand, N., Saklatvala, J., Baudry, D., Duckworth, M., Allen, M.H., Curtis, C.J., Lee, S.H., Burden, A.D., et al. (2018). Genome-wide meta-analysis implicates mediators of hair follicle development and morphogenesis in risk for severe acne. Nat. Commun. *9*, 5075. https://doi.org/10.1038/s41467-018-07459-5.
35. Kurki, M.I., Karjalainen, J., Palta, P., Sipilä, T.P., Kristiansson, K., Donner, K.M., Reeve, M.P., Laivuori, H., Aavikko, M., Kaunisto, M.A., et al. (2023). FinnGen provides genetic insights from a well-phenotyped isolated population. Nature *613*, 508–518. https://doi.org/10.1038/s41586-022-05473-8.

36. Dixon, P.H., Levine, A.P., Cebola, I., Chan, M.M.Y., Amin, A.S., Aich, A., Mozere, M., Maude, H., Mitchell, A.L., Zhang, J., et al. (2022). GWAS meta-analysis of intrahepatic cholestasis of pregnancy implicates multiple hepatic genes and regulatory elements. Nat. Commun. *13*, 4840. https://doi.org/10.1038/s41467-022-29931-z.

37. Joshi, A.D., Andersson, C., Buch, S., Stender, S., Noordam, R., Weng, L.C., Weeke, P.E., Auer, P.L., Boehm, B., Chen, C., et al. (2016). Four Susceptibility Loci for Gallstone Disease Identified in a Meta-analysis of Genome-Wide Association Studies. Gastroenterology *151*, 351–363.e28. https://doi.org/10.1053/j.gastro.2016.04.007.

38. Zheng, J., Wheeler, E., Pietzner, M., Andlauer, T.F.M., Yau, M.S., Hartley, A.E., Brumpton, B.M., Rasheed, H., Kemp, J.P., Frysz, M., et al. (2023). Lowering of Circulating Sclerostin May Increase Risk of Atherosclerosis and Its Risk Factors: Evidence From a Genome-Wide Association Meta-Analysis Followed by Mendelian Randomization. Arthritis Rheumatol. *75*, 1781–1792. https://doi.org/10.1002/art.42538.

39. International Consortium for Blood Pressure Genome-Wide Association Studies, Ehret, G.B., Munroe, P.B., Rice, K.M., Bochud, M., Johnson, A.D., Chasman, D.I., Smith, A.V., Tobin, M.D., Verwoert, G.C., et al. (2011). Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature *478*, 103–109. https://doi.org/10.1038/nature10405.

40. Wang, Y.K., Tang, J.N., Han, L., Liu, X.D., Shen, Y.L., Zhang, C.Y., and Liu, X.B. (2020). Elevated FURIN levels in predicting mortality and cardiovascular events in patients with acute myocardial infarction. Metabolism *111*, 154323. https://doi.org/10.1016/j.metabol.2020.154323.

41. He, Y., Ren, L., Zhang, Q., Zhang, M., Shi, J., Hu, W., Peng, H., and Zhang, Y. (2019). Serum furin as a biomarker of high blood pressure: findings from a longitudinal study in Chinese adults. Hypertens. Res. *42*, 1808–1815. https://doi.org/10.1038/s41440-019-0295-6.

42. Lind, L., Mazidi, M., Clarke, R., Bennett, D.A., and Zheng, R. (2024). Measured and genetically predicted protein levels and cardiovascular diseases in UK Biobank and China Kadoorie Biobank. Nat Cardiovasc Res. *3*, 1189–1198. https://doi.org/10.1038/s44161-024-00545-6.

43. Thomas, G. (2002). Furin at the cutting edge: from protein traffic to embryogenesis and disease. Nat. Rev. Mol. Cell Biol. *3*, 753–766. https://doi.org/10.1038/nrm934.

44. Ichiki, T., Huntley, B.K., and Burnett, J.C. (2013). BNP molecular forms and processing by the cardiac serine protease corin. Adv. Clin. Chem. *61*, 1–31. https://doi.org/10.1016/b978-0-12-407680-8.00001-4.

45. Ren, K., Jiang, T., Zheng, X.L., and Zhao, G.J. (2017). Proprotein convertase furin/PCSK3 and atherosclerosis: New insights and potential therapeutic targets. Atherosclerosis *262*, 163–170. https://doi.org/10.1016/j.atherosclerosis.2017.04.005.

46. Yakala, G.K., Cabrera-Fuentes, H.A., Crespo-Avilan, G.E., Rattanasopa, C., Burlacu, A., George, B.L., Anand, K., Mayan, D.C., Corlianò, M., Hernández-Reséndiz, S., et al. (2019). FURIN Inhibition Reduces Vascular Remodeling and Atherosclerotic Lesion Progression in Mice. Arterioscler. Thromb. Vasc. Biol. *39*, 387–401. https://doi.org/10.1161/ATVBAHA.118.311903.

47. Carrasco-Zanini, J., Pietzner, M., Davitte, J., Surendran, P., Croteau-Chonka, D.C., Robins, C., Torralbo, A., Tomlinson, C., Grünschläger, F., Fitzpatrick, N., et al. (2024). Proteomic signatures improve risk prediction for common and rare diseases. Nat. Med. *30*, 2489–2498. https://doi.org/10.1038/s41591-024-03142-z.

48. Wei, C., Banu, K., Garzon, F., Basgen, J.M., Philippe, N., Yi, Z., Liu, R., Choudhuri, J., Fribourg, M., Liu, T., et al. (2018). SHROOM3-FYN Interaction Regulates Nephrin Phosphorylation and Affects Albuminuria in Allografts. J. Am. Soc. Nephrol. *29*, 2641–2657. https://doi.org/10.1681/ASN.2018060573.

49. Hoy, S.M. (2022). Deucravacitinib: First Approval. Drugs *82*, 1671–1679. https://doi.org/10.1007/s40265-022-01796-y.

50. FitzGerald, O., Gladman, D.D., Mease, P.J., Ritchlin, C., Smolen, J.S., Gao, L., Hu, Y., Nowak, M., Banerjee, S., and Catlett, I. (2024). Phase 2 Trial of Deucravacitinib in Psoriatic Arthritis: Biomarkers Associated With Disease Activity, Pharmacodynamics, and Clinical Responses. Arthritis Rheumatol. *76*, 1397–1407. https://doi.org/10.1002/art.42921.

51. Zielinska, D.F., Gnad, F., Wiśniewski, J.R., and Mann, M. (2010). Precision mapping of an in vivo N-glycoproteome reveals rigid topological and sequence constraints. Cell *141*, 897–907. https://doi.org/10.1016/j.cell.2010.04.012.

52. Reily, C., Stewart, T.J., Renfrow, M.B., and Novak, J. (2019). Glycosylation in health and disease. Nat. Rev. Nephrol. *15*, 346–366. https://doi.org/10.1038/s41581-019-0129-4.

53. Schjoldager, K.T., Narimatsu, Y., Joshi, H.J., and Clausen, H. (2020). Global view of human protein glycosylation pathways and functions. Nat. Rev. Mol. Cell Biol. *21*, 729–749. https://doi.org/10.1038/s41580-020-00294-x.

54. Morand, E., Merola, J.F., Tanaka, Y., Gladman, D., and Fleischmann, R. (2024). TYK2: an emerging therapeutic target in rheumatic disease. Nat. Rev. Rheumatol. *20*, 232–240. https://doi.org/10.1038/s41584-024-01093-w.

55. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209. https://doi.org/10.1038/s41586-018-0579-z.

56. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443. https://doi.org/10.1038/s41586-020-2308-7.

57. UniProt Consortium (2021). UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. *49*, D480–D489. https://doi.org/10.1093/nar/gkaa1100.

58. Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J.A., Ziyatdinov, A., Benner, C., O'Dushlaine, C., Barber, M., Boutkov, B., et al. (2021). Computationally efficient whole-genome regression for quantitative and binary traits. Nat. Genet. *53*, 1097–1103. https://doi.org/10.1038/s41588-021-00870-7.

59. Willer, C.J., Li, Y., and Abecasis, G.R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics *26*, 2190–2191. https://doi.org/10.1093/bioinformatics/btq340.

60. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol. *17*, 122. https://doi.org/10.1186/s13059-016-0974-4.

61. Assarsson, E., Lundberg, M., Holmquist, G., Björkesten, J., Thorsen, S.B., Ekman, D., Eriksson, A., Rennel Dickens, E., Ohlsson, S., Edfeldt, G., et al. (2014). Homogenous 96-plex PEA immunoassay exhibiting high sensitivity, specificity, and excellent scalability. PLOS One *9*, e95192. https://doi.org/10.1371/journal.pone.0095192.

62. Zhong, W., Edfors, F., Gummesson, A., Bergström, G., Fagerberg, L., and Uhlén, M. (2021). Next generation plasma proteome profiling to monitor health and disease. Nat. Commun. *12*, 2493. https://doi.org/10.1038/s41467-021-22767-z.

63. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. *29*, 308–311. https://doi.org/10.1093/nar/29.1.308.

64. Bulik-Sullivan, B.K., Loh, P.R., Finucane, H.K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric; Genomics Consortium-Patterson, N., Daly, M.J., Price, A.L., and Neale, B.M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. *47*, 291–295. https://doi.org/10.1038/ng.3211.

65. Wallace, C. (2021). A more accurate method for colocalisation analysis allowing for multiple causal variants. PLoS Genet. *17*, e1009440. https://doi.org/10.1371/journal.pgen.1009440.

66. Yavorska, O.O., and Burgess, S. (2017). MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. Int. J. Epidemiol. *46*, 1734–1739. https://doi.org/10.1093/ije/dyx034.

67. Verma, A., Huffman, J.E., Rodriguez, A., Conery, M., Liu, M., Ho, Y.L., Kim, Y., Heise, D.A., Guare, L., Panickan, V.A., et al. (2024). Diversity and scale: Genetic architecture of 2068 traits in the VA Million Veteran Program. Science *385*, eadj1182. https://doi.org/10.1126/science.adj1182.

68. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. *12*, e1001779. https://doi.org/10.1371/journal.pmed.1001779.

69. Ochoa, D., Hercules, A., Carmona, M., Suveges, D., Baker, J., Malangone, C., Lopez, I., Miranda, A., Cruz-Castillo, C., Fumis, L., et al. (2023). The next-generation Open Targets Platform: reimagined, redesigned, rebuilt. Nucleic Acids Res. *51*, D1353–D1359. https://doi.org/10.1093/nar/gkac1046.

70. Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C., and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. PLoS Genet. *10*, e1004383. https://doi.org/10.1371/journal.pgen.1004383.

71. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res. *12*, 996–1006. https://doi.org/10.1101/gr.229102.

72. Türei, D., Valdeolivas, A., Gul, L., Palacio-Escat, N., Klein, M., Ivanova, O., Ölbei, M., Gábor, A., Theis, F., Módos, D., et al. (2021). Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. Mol. Syst. Biol. *17*, e9923. https://doi.org/10.15252/msb.20209923.

73. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. *45*, D353–D361. https://doi.org/10.1093/nar/gkw1092.

74. Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., et al. (2018). The Reactome Pathway Knowledgebase. Nucleic Acids Res. *46*, D649–D655. https://doi.org/10.1093/nar/gkx1132.

75. Stacey, D., Fauman, E.B., Ziemek, D., Sun, B.B., Harshfield, E.L., Wood, A.M., Butterworth, A.S., Suhre, K., and Paul, D.S. (2019). ProGeM: a framework for the prioritization of candidate causal genes at molecular quantitative trait loci. Nucleic Acids Res. *47*, e3. https://doi.org/10.1093/nar/gky837.

76. Karlsson, M., Zhang, C., Méar, L., Zhong, W., Digre, A., Katona, B., Sjöstedt, E., Butler, L., Odeberg, J., Dusart, P., et al. (2021). A single-cell type transcriptomics map of human tissues. Sci. Adv. *7*, eabh2169. https://doi.org/10.1126/sciadv.abh2169.

77. Verbanck, M., Chen, C.Y., Neale, B., and Do, R. (2018). Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. Nat. Genet. *50*, 693–698. https://doi.org/10.1038/s41588-018-0099-7.

78. Hemani, G., Zheng, J., Elsworth, B., Wade, K.H., Haberland, V., Baird, D., Laurin, C., Burgess, S., Bowden, J., Langdon, R., et al. (2018). The MR-Base platform supports systematic causal inference across the human phenome. eLife *7*, e34408. https://doi.org/10.7554/eLife.34408.

79. Bastarache, L. (2021). Using Phecodes for Research with the Electronic Health Record: From PheWAS to PheRS. Annu. Rev. Biomed. Data Sci. *4*, 1–19. https://doi.org/10.1146/annurev-biodatasci-122320-112352.

Please cite this article in press as: Koprulu et al., Multi-cohort proteogenomic analyses reveal genetic effects across the proteome and diseasome, Cell (2026), https://doi.org/10.1016/j.cell.2026.03.049

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Genome-wide summary statistics for 1,161 circulating proteins	This study	https://omicscience.org/
UK Biobank data	UK Biobank[55]	www.ukbiobank.ac.uk
FinnGen summary statistics	FinnGen Consortium[35]	https://www.finngen.fi/en/access_results
GTEx summary	GTEx Consortium[25]	https://gtexportal.org/home/downloads/adult-gtex/overview
gnomAD data	gnomAD Consortium[56]	https://gnomad.broadinstitute.org/data
UniProt	The UniProt Consortium[57]	http://uniprot.org/
GWAS Catalog	Sollis et al.[30]	https://www.ebi.ac.uk/gwas/
Software and algorithms		
R	The R Foundation	https://www.r-project.org/
REGENIE	Mbatchou et al.[58]	https://rgcgithub.github.io/regenie/
METAL	Willer et al.[59]	https://github.com/statgen/METAL
Variant Effect Predictor (VEP)	McLaren et al.[60]	https://www.ensembl.org/info/docs/tools/vep/

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

We performed a multi-cohort meta-analysis of genome-wide summary statistics of plasma levels of 1,194 protein targets from up to 37 cohorts (Table S1), referred to as SCALLOP meta-analyses. The participants of these cohorts were predominantly of European ancestry. Detailed information about the participant characteristics in each cohort, including mean age, mean BMI, percentage of females in the cohort and fasting status can be found on Table S1. All participants gave their informed written consent before entering the studies and each study was approved by their respective ethics committee.

We additionally included proteogenomic analysis conducted using genomic and proteomic data from 48,017 participants of European ancestry in UK Biobank (UKBB). Further information about the UKBB cohort and proteomic measurements can be found elsewhere.[20,55]

METHOD DETAILS

Proteomic measurements

For the 37 cohorts included in this study, antibody-based proteomic measurements were generated through at least one of the 13 Target-96 panels offered by Olink measuring 92 protein targets (Cardiometabolic, Cardiovascular II, Cardiovascular III, Cell Regulation, Development, Immuno-oncology, Inflammation, Immune Response, Metabolism, Neurology, Neuro Exploratory, Organ Damage, Oncology II). Details regarding the assay have been described in detail elsewhere.[61,62] Briefly, dual antibody based proteomic measurements are generated where each of the unique antibodies are labelled with complementary single stranded oligonucleotides, referred to as proximity extension assays (PEA).[61] Hybridization occurs between the complementary oligonucleotides when both of the unique antibodies bind to their respective protein target and come into close proximity, which can subsequently be quantified through qPCR or next-generation sequencing (NGS) methods. Relative proteomic measurements are offered as normalized protein expression (NPX) units, provided on log2 scale. Each cohort has performed quality control on their proteomic measurements such as, but not limited to, removing samples that were extreme outliers using principal-component analysis from their entire proteomic profiles. Only blood-based (i.e. plasma or serum) proteomic measurements were included in this study.

In UKBB, the proteomic measurements used in this study were generated through Olink Explore 1536 platform utilizing the same antibody-based technology but measuring a broader coverage of 1,463 protein targets. Details of proteomic measurements can be found elsewhere.[20]

Genome-wide meta-analyses of protein levels

For 1,194 unique protein assays for each protein target from all 37 cohorts, we performed additional quality control measures and retained only biallelic variants with (i) a call rate above 95%, (ii) Hardy-Weinberg p-value above $1x10^{-5}$, (iii) imputation INFO score

above 0.8, (iv) standard error of SNP-effect less than 10, (v) minor allele count (MAC) above 3 and (vi) minor allele frequency above 0.1%. We filtered out any variant which did not have an existing rsID in the dbSNP database.[63] Using all available summary statistics per protein target, we performed inverse-variance fixed-effects meta-analyses using METAL (v.2011-03-25),[59] also referred to as SCALLOP meta-analyses. The sample sizes varied from 2,297 to 31,190 depending on protein target coverage across cohorts.

In parallel, we have also conducted genome-wide association analyses for proteomic measurements of 1,463 protein targets in UK Biobank participants through REGENIE v.3.1.4.[58]

Finally, we performed an inverse-variance fixed-effects meta-analyses (with METAL[59]) for 1,161 protein targets (with overlapping UniProt ID across different antibody-based platforms offered by Olink) using summary level data from SCALLOP meta-analyses and UKB, reaching to a total sample size of up to 78,664 participants (n=17,602 – 78,864).

Downstream analyses

Following regional clumping for independent signal selection, we applied Bayesian fine-mapping to identify independent protein quantitative trait loci (pQTLs) using SuSie.[27] We characterized variance explained by significant pQTLs and heritability by polygenic background (excluding any significant loci for a given protein) through LD-score regression implemented in LDSC.[64] Leveraging the multi-cohort study design, we assessed the confidence for of each pQTL by their (i) statistical significance, (ii) directional concordance across cohorts and (iii) observed heterogeneity. We then tested the contribution of participant and cohort-level characteristics to the heterogeneity observed through a meta-regression model. We used zero-inflated Poisson regression models to test for associations between the presence and number of associated cis-/trans-pQTLs per protein and various protein characteristics.

Using the output from machine learning models we trained for trans-pQTL effector gene assignment, we tested enrichment of the assigned effector genes among pathways, tissues, cell-types. Based on assessing the overlap of pQTLs or proxies (r^2>0.8) with phenome-wide associations from GWAS catalogue,[30] we defined three categories of pleiotropy, (i) 'molecular pleiotropy' (associated with >5 proteins and $\leq$ 5 non-proteomic phenotypes), (ii) 'phenotypic pleiotropy' (associated with >5 non-proteomic phenotypes and $\leq$ 5 proteins), and (iii) 'unspescific pleitropy' (associated with >5 proteins and >5 non-proteomic phenotypes).

For phenotypic follow-up, we first assessed the link between cis-pQTLs with 835 diseases from the FinnGen[35] release 8 using fine-mapping augmented colocalization[65] and two sample MR analyses.[66] We further systematically tested for the relevance of trans-pQTLs by adopting the cis-MR workflow and testing for (i) cis and trans-pQTLs and (ii) trans-pQTLs only as genetic instruments for each protein target. To test for concordance between genetically inferred protein – disease relationships and plasma protein levels and disease onset and/or presence, we first run Cox-proportional hazard models and logistics regression models, respectively, adjusting for age, sex, and technical variables. We then integrated cis-based two-sample MR analyses from an internal pan-biobank project covering over one million individuals (MVP,[67] the Pan-UK Biobank,[68] and FinnGen[35]) and tested for concordance between the genetic and observational approaches using the relevant multiple-testing corrected significance thresholds.

Finally, we tested whether the protein association profile of pQTLs that were reported for at least one non-proteomic trait in the GWAS Catalog were significantly enriched for any pathway and whether there was a non-random overlap between protein-biomarker signatures from prospective biomarker analyses and protein signatures associated with pQTLs using Fisher's exact test. Detailed methods can be found on quantification and statistical analysis section.

QUANTIFICATION AND STATISTICAL ANALYSIS

Genome-wide meta-analyses of protein levels

We did not observe evidence for genomic inflation for (i) SCALLOP only meta-analysis (mean=1.04, IQR=1.04–1.05) or UKBB proteogenomic analyses (mean=1.09, IQR=1.07–1.10).

Using the summary statistics from SCALLOP meta-analyses and UKBB proteogenomic analysis, we performed inverse-variance fixed-effects meta-analyses using METAL (v.2011-03-25)[59] for 1,161 protein targets, mapping to the same UniProt ID. We also did not see evidence for genomic inflation across the overall meta-analyses (mean=1.08, IQR=1.06–1.09).

Genomic build GRCh37 was used throughout this study.

Identification of regional sentinel variants

Regional sentinel variant selection was performed by selecting the variant with the highest z-score within each 1 Mb window around each significant signal. The genomic regions were merged and considered as a single region if there were multiple significant variants within less than 500 kb from each other. Bonferroni corrected genome wide significance thresholds were used for each sub-study as follows, $p<4.19\times10^{-11}$ for SCALLOP meta-analyses (1,194 protein targets), $p<3.42\times10^{-11}$ for UKBB proteogenomic analyses (n=1,463 protein targets) and $p<4.31\times10^{-11}$ for the overall meta-analyses (n=1,161 overlapping protein targets).

pQTLs were defined as cis-acting if they were located within 500 kb of a protein's cognate gene and were defined as trans-acting otherwise. Where assays targeted more than one protein, *cis*-pQTLs were defined as those located within 500 kb of any potential cognate genes.

Fine mapping

We performed statistical fine-mapping on protein GWASs performed in the UKBB, as the largest single study, using the 'sum of single effects' model (SuSiE).[27] Briefly, SuSiE implements variable selection under a Bayesian framework to identify credible sets of independent variants that likely contain the true underlying causal variant. We conducted these analyses using the R package susieR (v.0.12.16) with the default parameters and priors. We used a random subset of 20,000 unrelated participants of European ancestry from the UKBB as an LD reference. Additionally, we adopted a grid search approach that iterated the maximum number of credible sets from 2 to 10, selecting the number that ensured no variants within a credible set were in LD (r2> 0.1). Credible sets were taken forward where variants additionally met genome-wide significance. We have included regional summary statistics from the overall meta-analysis for any region where it was not possible to perform SuSie on UKBB summary statistics or regions on X-chromosome. We also clumped the independent credible sets into r2 groups across all protein targets based on linkage disequilibrium (r2>0.8).

Concordance of the identified pQTLs

Variants were characterized as 'high confidence' if the pQTL was (i) at least genome-wide significant ($p<5x10^{-8}$), (ii) showed directional consistency in the overall meta-analyses between SCALLOP cohorts and UKBB for *cis*- and *trans*-pQTLs and (iii) additionally did not show substantial evidence of heterogeneity ($p_{het}>1x10^{-4}$) for *trans*-pQTLs.

We tested for independent replication of identified lead variants from credible sets by investigating for the overlap for the pQTLs or any of their proxies (r2> 0.1) in previously published affinity-based proteogenomic studies, based on a common UniProt ID for the protein targets.[6–24]

We tested the correlation between the effect size for regional sentinel variants identified in this study with cohorts of predominantly European ancestry with effect size of the pQTL in cohorts of non-European ancestry, namely Immuno-oncology panel in China Kadoorie Biobank (CKB, *n*=816 participants of Chinese ancestry), Metabolism panel in Qatar Metabolomics study on Diabetes (QMDIAB, *n*=350 participants of Indian, Filipino or Arabic ancestry) and Inflammation, Cardiovascular III, Neurology and Oncology II panels in Shanghai Women and Men's Health Study (SWMHS, *n*=548 participants of Chinese ancestry) (Figure S2).

Variance explained by identified pQTLs

We estimated the variance explained by the regional sentinel variants and heritability estimates by the polygenic background, excluding the regions harbouring any pQTLs[18] for all 1,161 protein targets included in this study. In summary, we calculated variance explained for cis and trans loci by using the formula ($2 \times \beta 2 \times f \times (1-f)$), where f is the MAF. We calculated heritability estimates by the polygenic background, by excluding the regions harbouring a pQTL, through LD-score regression (LDSC).[64]

pQTLs and protein-related characteristics

We tested the association of presence and also the number of high confidence cis- or trans- pQTLs with a wide range of protein-related characteristics, namely whether a protein (a) has sites for sulfation, glycosylation, phosphorylation, ubiquitination, s-nitrosylation, acetylation, palmitoylation, acetylglucosmination, glycosaminoglycan-chains, myristoylation, acylation and methylation, (b) has any disulfide bonds, DNA binding domains, alpha-helix domains, turn domains, transmembrane domains, zinc finger domains, coiled coil domains, beta-strand domains, or a protein's (c) transcript count as obtained from UniProt.[57] We also tested the association of the presence and also the number of high-confidence cis- or trans-pQTLs with the probability of loss of function intolerance (pLI), missense Z-score (i.e. the deviation of the observed from the expected number of missense variants), protein length and gene length as obtained from gnomAD v2.1.[56] We used 'pscl' package in R v.4.2.2 which performs zero-inflated Poisson regression characteristics by including scaled sample sizes as a covariate. In summary, the package runs two set of tests for each characteristic which are (i) Poisson regression for the count data to test the association between the number of high-confidence cis or trans pQTLs and protein-related characteristics, (ii) logistic regression to test the association between the existence of high confidence cis or trans pQTLs and protein-related characteristics. The continuous traits were normalized to have a comparable effect size estimate with the binary ones. We applied a Bonferroni-corrected significance threshold of $p<2x10^{-3}$ (corrected for the number of investigated characteristics, *n*=25). We also ran sensitivity analyses including missingness rate for proteins as a covariate in the model.

Impact of cohort characteristics on heterogeneity

To better understand cohort characteristics that might contribute to the heterogeneity, we performed meta-regression analyses in R v4.2.2 using 'metafor' package. The loci where we observed high levels of heterogeneity (heterogeneity p-value $<1x10^{-4}$) across SCALLOP cohorts from the fixed-effects meta-analyses were taken forward and a meta-regression model was fit with between the effect size of these loci from each cohort and cohort characteristics and the overall genetic effect size of the loci (from the meta-analysis) as explanatory variables. The cohort characteristics that were included were mean cohort age, sex composition of the cohort (i.e. %females), mean BMI of the cohort, blood-based tissue type (EDTA plasma, citrate plasma, or serum), fasting status and whether any disease cases were included in the cohort (Table S1).

Characterization of variant consequences

We used Variant Effect Predictor (VEP, version 99) through Ensembl to characterize the consequence of the fine-mapped variants and their proxies.[60] We used Graphql to query the Opentargets platform[69] to obtain the closest gene for each unique fine-mapped pQTL variant.

Colocalization with gene expression levels

We systematically tested for a shared genetic signal between plasma abundance of a protein and gene expression levels (eQTL) of the protein coding gene in 49 tissues from the GTEx project (v8)[25] assessed by posterior probability from statistical colocalization analysis using 'coloc' package in R 3.6.[70] The build of our study was lifted over from b37 to b38 using LiftOver,[71] to be compatible for the analyses. If any of the pQTLs did not have a corresponding genomic location in b38, they were excluded from the downstream colocalization analysis. GTEx[25] variant-gene cis-eQTL associations from each tissue were downloaded from https://console.cloud.google.com/storage/browser/gtex-resources on January 2020.

For the cis-variants, the region was defined as (±500 kb) around the protein-encoding gene of each protein-encoding gene with at least one *cis*-pQTL that were present in both datasets. For *trans*-pQTL - *cis*-eQTL colocalization analyses, the region was defined as the ±500 kb window around the *trans*-pQTL signal, regardless of the genes that reside in the region. The colocalization analysis was only performed if the region had at least suggestive evidence of association ($p<10^{-6}$) in the defined region in any of the tissues and the lead variant was in high LD with the lead *cis*-pQTL in the region ($r^2>0.6$). A prior probability of a shared signal (p12) was defined as $1x10^{-5}$ and posterior probability above 80% was used to define a high likelihood of a shared genetic signal.

Machine learning-based effector gene assignment

We implemented a staged machine learning classifier to systematically assign candidate effector genes for all *trans*-pQTLs residing outside of the MHC region. We first collated comprehensive annotations for each genetic variant or proxies thereof ($r^2>0.6$), including 1) distance to the gene body in 1Mb window, and 2) putative functional consequences based on the variant effector prediction (VEP) tool. We further systematically collated for each gene within a 1Mb window: 1) evidence for colocalization between gene and protein expression levels based GTEx version 8,[25] 2) evidence for a rare gene burden association based on Dhindsa et al. (2023),[22] 3) evidence whether any trans gene encodes for receptor/ligand or protein complex of the cis-protein based on literature evidence using the OmnipathR package v3.10.1,[72] and 4) whether any of the genes participate in the same biological pathway based on KEGG[73] and REACTOME[74] annotations.

In the absence of generalisable gold-standard variant to gene assignments, we leveraged prior biological and genomic knowledge to derive three partly distinct sets of 'putative true positive' (PTP) sets: 1) Trans genes that encode ligand – receptor pairs of those with high-confidence evidence for forming a protein complex with the cis-protein (n=540 PTPs), 2) sentinel trans-pQTLs mapping to functional variants (n=1,747 PTPs), and 3) significant gene burden results for trans genes (n=1,049 PTPs). For each of the three PTP sets, we considered all other genes within a 1Mb window in the locus as negative examples. We pruned each set to contain at each locus only one cis-protein to avoid artificially good results due to pleiotropy. We then split each of the data sets 10 times in a 7:3 ratio to obtain training and test sets separating by genomic region. For each of the ten training sets, we trained a Random Forest classifier using repeated 3-fold cross-validation implementing subsampling to account for the unbalanced data sets. This was implemented using the R caret v6.0.94 package. We used the Kappa score to select the best performing forest within each training set. Eventually, we used each of the ten Random Forest classifiers from each PTP set to assign candidate scores for all putative effector genes across the entire set of trans-pQTLs. We thereby took the median score across all ten classifiers for a PTP set and finally summed the score across all three predictions. For each of the PTP sets, we omitted features used to define true positive sets. All three classifiers showed robust performance with median Kappa values of 0.54 – 0.57. The conception of our machine learning classifier was inspired by the ProGeM framework[75] and with a rationale of integrating biologically relevant annotations such as ligand – receptor pairs or protein complexes.

Pathway, tissue, and cell-type enrichment

We obtained processed data for tissue and single cell expression from the Human Protein Atlas,[76] and used the annotations 'specific' and 'enhanced' to define a common set of genes for each tissue/cell-type with enhanced expression. We used Fisher's exact tests to test for a significant enrichment of assigned effector genes among enhanced gene expression sets. To avoid redundancies, we only took the effector gene with the highest score at each locus to test for enrichments. If required, we restricted the background of those tests to proteins covered on the Olink panels and computed the false discovery rate to account for multiple testing. We report only enriched, but not depleted tissues/cell-types, although those have been considered when performing multiple testing correction. For pathway enrichment, we used the R package gprofiler2 v0.2.2 restricting KEGG[73] and REACTOME[74] as data bases and FDR-correction.

Phenotypic follow up

We used two partially complementary approaches to systematically link *cis*-pQTLs to 835 diseases from the FinnGen release 8.[35] Firstly, we used fine-mapped protein GWAS summary statistics to colocalize credible sets in ±500kB windows around protein encoding genes, as implemented in the R package coloc with a recent augmentation to relax the single variant assumption using

the SuSiE method.[65] Colocalization was only performed if there was at least a suggestive signal for a given region for the trait being tested. We used default priors apart from implying a more stringent prior to declare a shared genetic signal (p_{12}=5x10^{-6}). To fine-map FinnGen outcomes, we used at most 5 credible sets and a random set of 30,000 UKBB participants to estimate an LD-backbone. We considered evidence of a shared signal between a protein and a FinnGen outcome, if at least one credible set was shared among both traits (PP H4 > 80%) provided the lead *cis*-pQTL was also in LD with the fine-mapped lead signal for the FinnGen trait. We used simple Wald ratios to assess effect directions, i.e., to understand whether an increase in protein levels is associated with an increase/decrease in disease risk. Secondly, we performed standard cis-based MR analysis,[66] by aggregating Wald ratios across all independently selected, genome wide significant *cis*-pQTLs and any of the FinnGen outcomes using the inverse variance weighted method. We required overlap of at least two-thirds of identified *cis*-pQTLs to avoid artificial findings. Apart from correcting for multiple testing using Bonferroni correction, we further required consistent effect directions in sensitivity analysis, including weighted median, mode, and MR-Egger regression, as well as no strong evidence for heterogeneity ($p>10^{-3}$) or pleiotropy ($p>10^{-3}$).

We further systematically tested for the relevance of *trans*-pQTLs by adapting the *cis*-MR workflow as follows. We considered all high-confidence *trans*-pQTLs associated with a protein target that were associated with less than five protein targets or phenotypes in the GWAS Catalog,[30] i.e. specific *trans*-pQTLs. We first identified instruments strongly violating MR assumptions by performing leave-one-out analysis tracking Cochran's Q statistic and omitted any instrument accounting for strong (median + 3 X SD) deviations compared to all other combinations of instruments - a procedure similar to recently proposed, but computationally more expensive, methods.[77] We also tested combining cis- and *trans*-pQTLs. Similar to the *cis*-MR, we applied the Bonferroni correction to account for multiple testing and implemented several sensitivity analyses including performing MR-Egger to assess pleiotropy, assessing directionally concordance the across different MR methods and assessing the heterogeneity among the pQTLs. We further demonstrated the need to account for pleiotropy via biological insights by running MR analyses without pleiotropy filters for trans-pQTLs.

Cis-based pan-biobank phenotypic follow up

To harmonize phenotypes between MVP,[67] the Pan-UK Biobank,[68] and FinnGen[35] (version 10), disease-based traits were mapped using codes provided by the biobanks. MVP used entirely phecodes (*n*=1,171), while the UKBB used phecodes (*n*=1,327) as well as ICD10 codes (*n*=915) to label their disease phenotypes. After restricting studies in the UKBB to those with European ancestry in their list of populations, all possible direct phecode to phecode (n=1,013) matches were made between the biobanks. UKBB traits with an ICD10 code were then mapped to phecodes using a conversion table (70), and if the derived phecode had not already been mapped to MVP in the previous step, then a match was made if possible (n = 10). For MVP phecodes that remained unmapped to UKBB, studies with case counts higher than 4,000 were reviewed and a decision was made on a case-by-case basis to map manually (n = 53) by a clinician before aligning with FinnGen. For all unmapped MVP phecodes with case counts lower than 4,000, MVP only data was used for analysis. To map MVP and UKBB phenotypes to FinnGen, we manually mapped all available R10 FinnGen phenotypes to those previously available from MVP or UKBB. All matches were double checked through clinical adjudication and comparison statistics using the number of cases per resource for each phenotype to identify significant outliers from this mapping strategy. Situations where a significant deviation from the overall relationship between MVP/UKBB/FinnGen case counts was observed were analyzed on a one-by-one case and a final harmonization decision was made. This led to FinnGen being mapped only with MVP (32 instances), only with UKBB (174 instances), and with both MVP and UKBB (501 instances).

Pan-UK Biobank GWAS physical positions were converted from genome build GRCh37 to GRCh38 using LiftOver.[71] Using METAL,[59] we performed fixed effects inverse-variance weighted meta-analyses of MVP European results, UKBB European results, and FinnGen and obtained estimates of heterogeneity (option in METAL: ANALYZE HETEROGENEITY). If mapping was not possible between MVP, UKBB, or FinnGen, then only MVP results or select UKBB phenotypes were retained for further analyses. For very small p-values in MVP which were set to zero, we set to the lowest possible decimal place allowed in Python using sys.float_info.min() (2.2250738585072014e-308). To test for inflation of p-values following meta-analyses, we calculated the lambda values for all meta-analyses results. Any phenotype having a lambda value greater than 1.15 (n = 289) was rerun including the genomic control parameter in METAL. Phenotypes that were not meta-analyzed were also tested for inflated p-values. We found 81 phenotypes that needed correction, the vast majority being clinical biomarkers (75 studies). Due to highly inflated p-values, we removed eight height-based phenotypes from consideration.

Two-sample Mendelian Randomization (MR) of each of the protein-coding genes were performed against all phenotypes using instruments from SCALLOP *cis*-pQTLs. In order to determine the correct ordering of alleles between the datasets we utilized the harmonise_data() function from the TwoSampleMR[78] package in R.

We used the Wald Ratio for instruments with one genetic variant and inverse variance weighted MR for instruments with multiple genetic variants. We additionally performed MR-Egger for proteins with three or more instruments to be used as a sensitivity analysis. We tested for heterogeneity across variant-level MR estimates, using the Cochrane Q method (mr_heterogeneity option in TwoSampleMR package) and the MR-Egger intercept.

Comparison of observational and genetic analyses

To test for concordance between genetically inferred protein – disease relationships and plasma protein levels and disease onset, we performed two analysis streams. We collated electronic health records, such as primary (45% of the population) and secondary care, death certificates, or cancer registry, but also self-reported information into 1,448 medical ontology terms called 'phecodes' (which

we refer to as 'diseases' for simplicity) for all UKBB participants.[79] We kept track of the first occurrence of each disease code in any of the resources and classified participants that had any code for a given disease prior to inclusion into UKBB or reported it at recruitment as 'prevalent', and any disease occurrence thereafter as 'incident'. We then associated plasma protein levels with the risk of disease onset using Cox-proportional hazard models with age as the underlying timescale while adjusting for sex of the participant and technical variables, such as sample age and fasting time. We monitored the proportional hazard assumption using Schoenefeld residuals and applied different levels of multiple testing correction to declare significance (see below). For each disease outcome, we excluded participants with codes indicating prevalent diseases. We only tested diseases with at least 10 cases (n=765) among all 52,169 participants with overlapping proteomic and disease data. Secondly, we integrated the genetic analyses from pan-biobank genetic analyses as described above and restricted those protein–disease associations available in both pan-biobank genetic analyses and prospective analyses. For genetic discovery, we defined high confidence protein-disease associations as those passing correction for multiple testing across protein-disease pairs and with support from colocalization (PP4 > 0.8). We then sought support from prospective analyses for each high confidence protein-disease pair at $p < 0.05$. For prospective discovery, after correction for multiple testing, we investigated whether there was genetic support for each protein-disease pair at the maximum p-value where colocalization was performed ($5x10^{-6}$) and PP4 > 0.8.

Enrichment analysis anchored on pQTLs

We performed two different sets of enrichment analysis utilizing pQTLs. Firstly, for pQTLs reported for at least one non-proteomic trait in the GWAS Catalog,[30] we tested whether significantly associated protein targets were enriched for pathways using the same settings as described above using the Gprofiler software but setting the background for enrichment testing to the proteins captured in our study. Secondly, we tested for a non-random overlap between protein biomarker signatures from prospective biomarker analyses and protein signatures associated with pQTLs using Fisher's exact test. We used the proteins included in all analyses as a background and proteins that associated with disease onset and the trans-pQTL as foreground. We tested only for enrichments if at least five proteins were associated with the trans-pQTL. We linked enrichment analysis to prospective biomarker analysis further by mapping GWAS Catalog entries to phecodes to assign convergence.

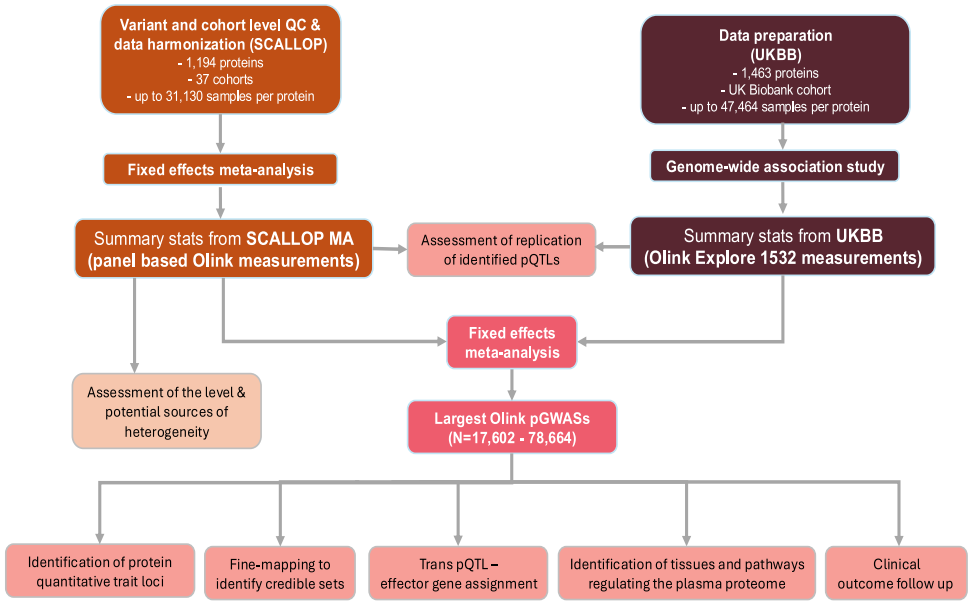
ADDITIONAL RESOURCES

Genome-wide summary statistics for all protein targets in this study will be available on https://omicscience.org/ upon publication. Associated code and scripts for the analyses are available at https://github.com/comp-med/scallop-ukbb-ma.

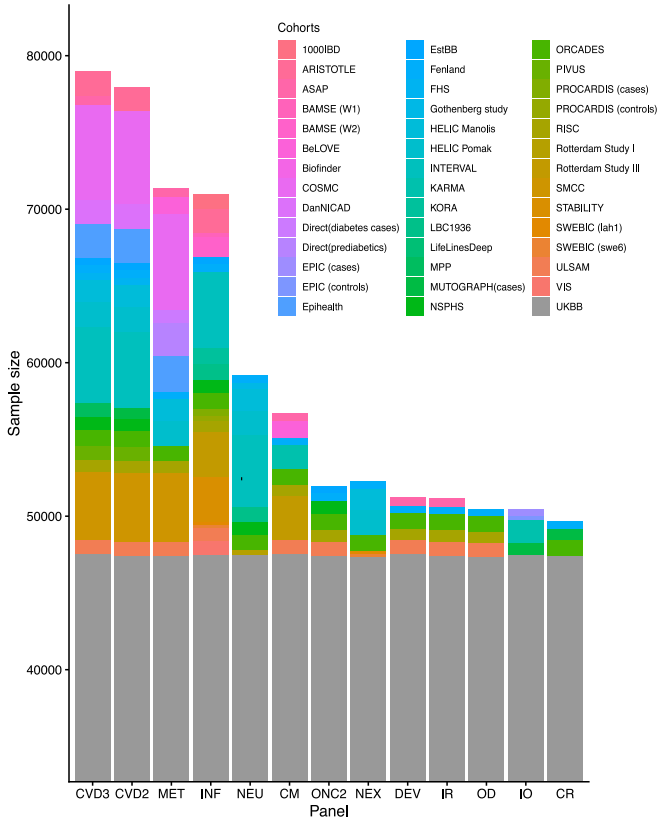
CellPress
OPEN ACCESS

Supplemental figures

A



B



(legend on next page)

Figure S1. Overview of the study, related to Figure 1

(A) Study design.

(B) Distribution of sample sizes across cohorts and panels. The axis representing UKBB samples has been split for better visibility, and the *y* axis scale starts from 35,000 to 80,000.

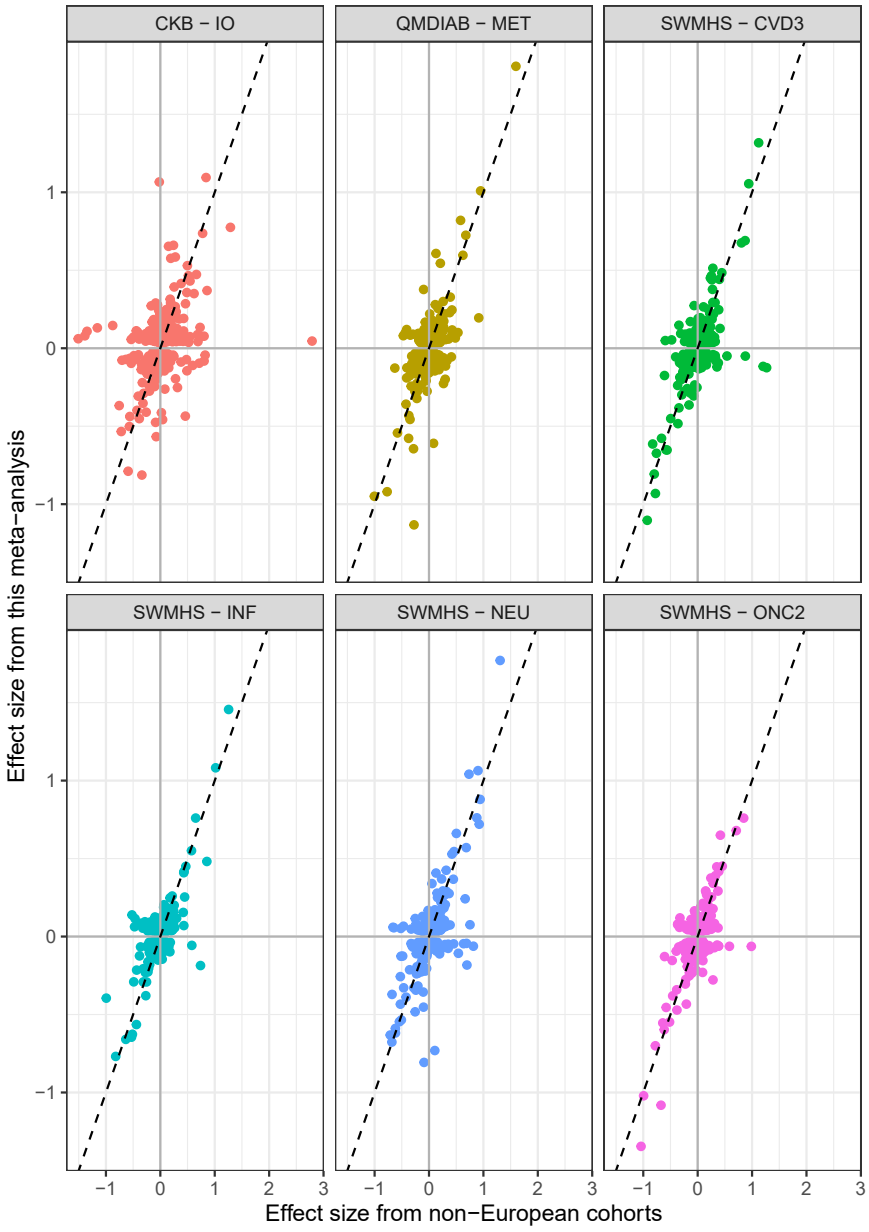


Figure S2. Correlation between the effect sizes of regional sentinel variants identified in this meta-analysis of participants mostly from European ancestry and cohorts of non-European ancestry for overlapping variant-protein pairs, related to Figure 1

The cohort-panel pairs visualized are the immuno-oncology panel in the China Kadoorie Biobank (CKB, *n* = 816 participants of Chinese ancestry), metabolism panel in the Qatar Metabolomics Study on Diabetes (QMDIAB, *n* = 350 participants of Indian, Filipino, or Arabic ancestry), and inflammation, cardiovascular III, neurology, and oncology II panels in the Shanghai Women and Men's Health Study (SWMHS, *n* = 548 participants of Chinese ancestry).

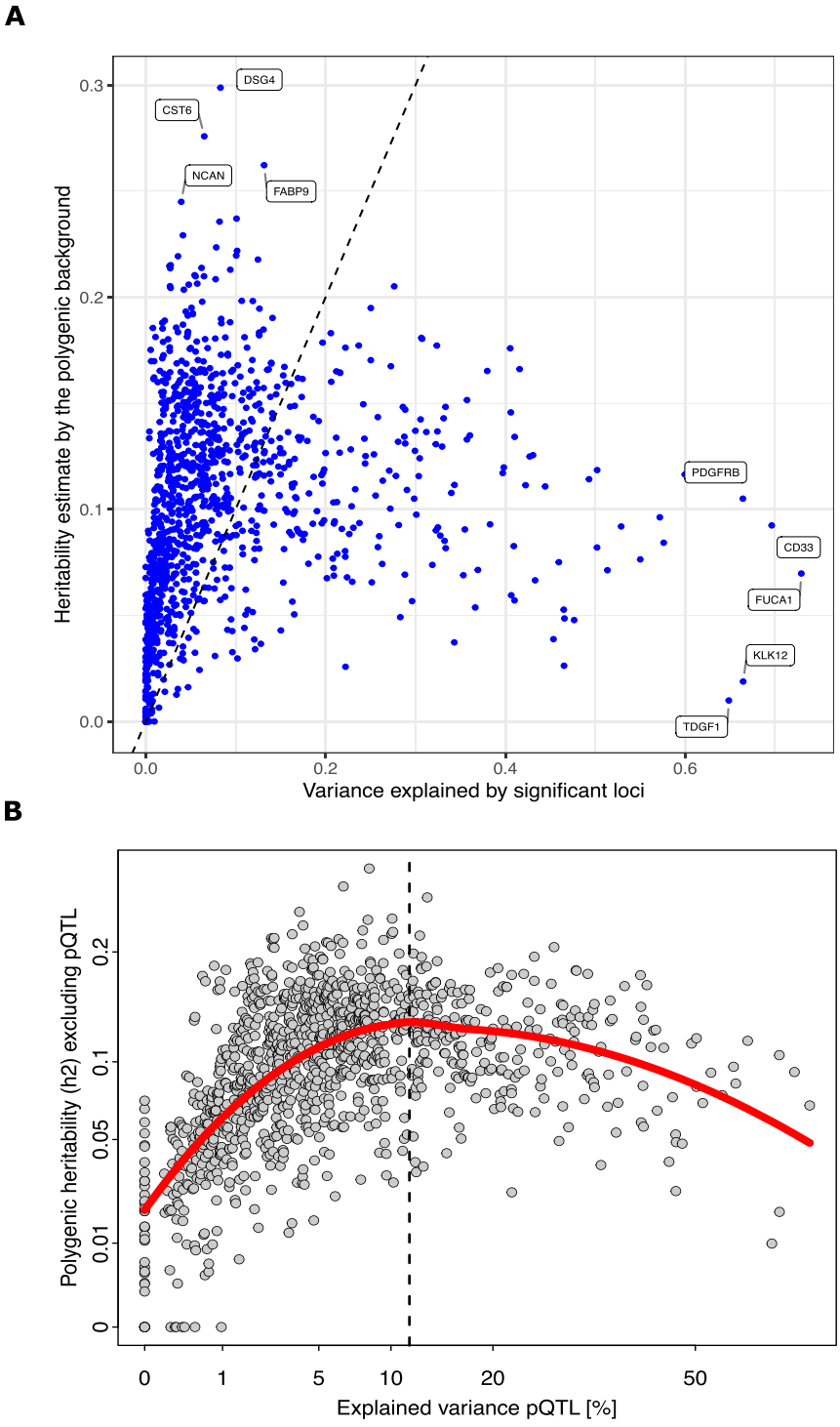


Figure S3. Variance and heritability estimates of 1,161 protein targets included in this study, related to Figure 1

(A) Variance explained by significant loci and heritability estimates excluding regions harboring any pQTLs for all 1,161 proteins included in this study.

(B) Scatterplot of protein targets with their square-root-transformed variance explained by significant loci plotted on the *x* axis and square-root-transformed heritability estimates, excluding regions harboring any pQTLs, plotted with a LOESS (locally estimated scatterplot smoothing) curve in red.

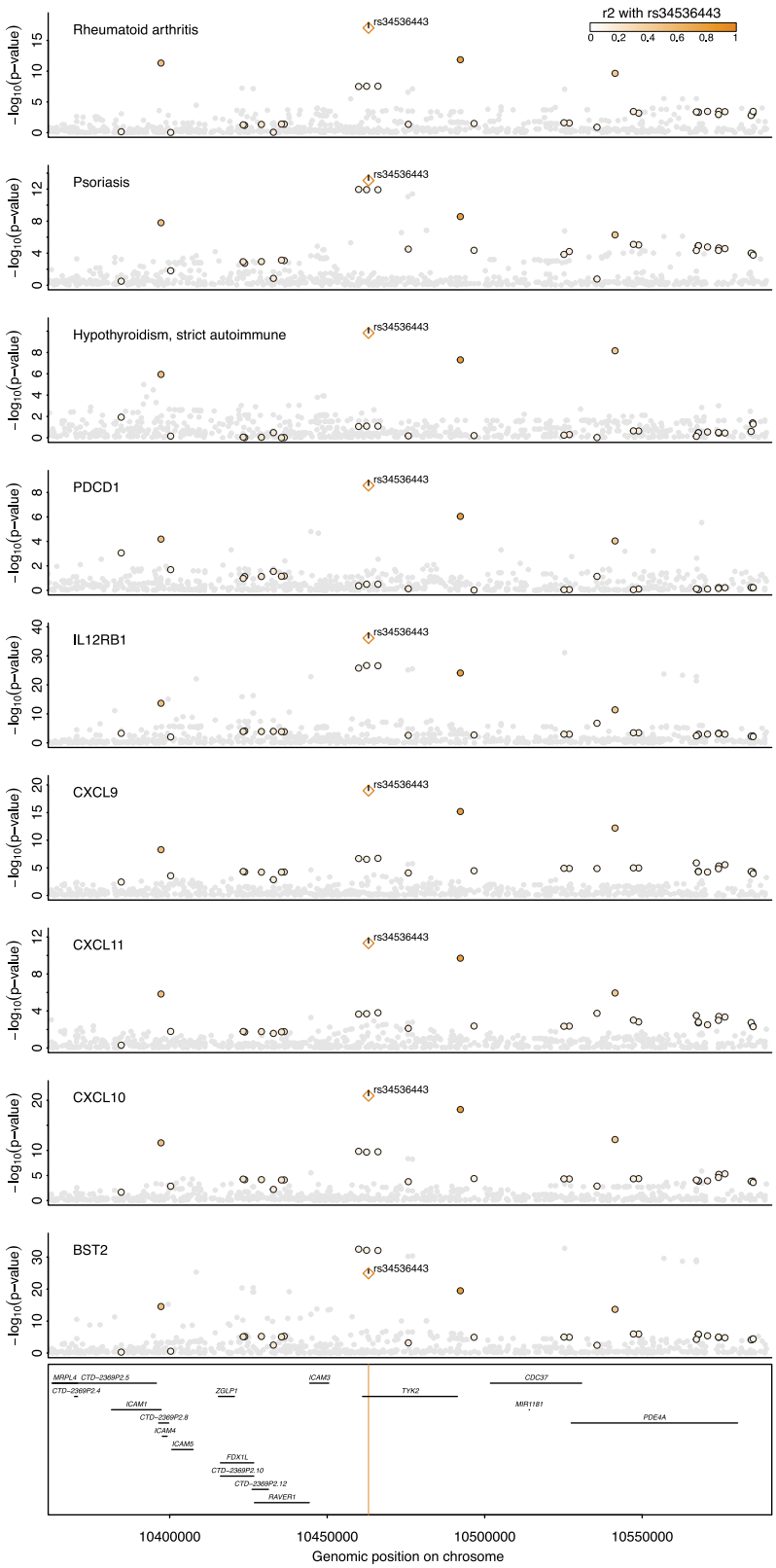


Figure S4. LocusZoom plot for *trans*-pQTL (rs34536443) and its association with multiple proteins (BST2, CXCL9, CXCL10, CXCL11, IL-12RB1, and PDCD1) and disease risk for rheumatoid arthritis, hypothyroidism, and psoriasis (HyPerColoc PP = 98.9%), related to Figure 7