JASN

Plasma proteomics of renal function: a trans-ethnic meta-analysis and Mendelian randomization study

Journal:	*Journal of the American Society of Nephrology*
Manuscript ID	JASN-2020-07-1070.R1
Manuscript Type:	Original Article - Clinical Epidemiology
Date Submitted by the Author:	n/a
Complete List of Authors:	Matías-García, Pamela; Helmoltz Zentrum München Research Unit Molecular Epidemiology; Helmholtz Zentrum München Institute of Epidemiology; Technical University of Munich; German Center for Cardiovascular Research (DZHK), partner site Munich Heart Alliance Wilson, Rory; Helmoltz Zentrum München Research Unit Molecular Epidemiology; Helmholtz Zentrum München Institute of Epidemiology Guo, Qi; University of Cambridge Department of Public Health and Primary Care, Cardiovascular Epidemiology Unit Zaghlool, Shaza; Weill Cornell Medicine - Qatar, Department of Physiology and Biophysics Eales, James; University of Manchester, Institute of Cardiovascular Sciences Xu, Xiaoguang; University of Manchester, Institute of Cardiovascular Sciences Charchar, Fadi; Federation University Australia, School of Health and Life Sciences; University of Leicester, Department of Cardiovascular Sciences; University of Melbourne, Department of Physiology Dormer, John; University Hospitals of Leicester, Cellular Pathology Schlosser, Pascal; Universitätsklinikum Freiburg, Institute of Genetic Epidemiology Elhadad, Mohamed; Helmoltz Zentrum München Research Unit Molecular Epidemiology; Helmholtz Zentrum München Institute of Epidemiology Nano, Jana; Helmholtz Zentrum München Institute of Epidemiology Sharma, Sapna; Helmoltz Zentrum München Research Unit Molecular Epidemiology; Helmholtz Zentrum München Institute of Epidemiology Peters, Annette; Helmholtz Zentrum München Institute of Epidemiology Fornoni, Alessia; Diabetes Research Institute, Miller School of Medicine, University of Miami, Medicine/Nephrology Mook-Kanamori, Dennis ; Leiden University Medical Center, Department of Clinical Epidemiology Winkelmann, Juliane; Helmholtz Zentrum München, Institute of Neurogenomics; Technical University of Munich, Department of Neurogenetics and Institute of Human Genetics Danesh, John; British Heart Foundation; University of Cambridge, Centre of Research Excellence Di Angelantonio, Emanuele ; University of Cambridge, Centre of Research Excellence ; University of Cambridge Department of Public Health and Primary Care, Cardiovascular Epidemiology Unit; National Institute for Health Research, Blood and Transplant Research Unit in

	Donor Health and Genomics Ouwehand, Willem; University of Cambridge, Department of Haematology; NHS Blood and Transplant, Cambridge Biomedical Campus; Wellcome Trust Sanger Institute Watkins, Nicholas; NHS Blood and Transplant, Cambridge Biomedical Campus Roberts, David; NHS Blood and Transplant, Oxford Centre; University of Oxford Radcliffe Department of Medicine Petrera, Agnese; Helmholtz Zentrum München Deutsches Forschungszentrum für Gesundheit und Umwelt, Research Unit Protein Science and Core Facility Proteomics Graumann, Johannes; Max Planck Institute for Heart and Lung Research, Scientific Service Group Biomolecular Mass Spectrometry; DZHK, partner site Rhine-Main Koenig, Wolfgang; German Center for Cardiovascular Research (DZHK); Deutsches Herzzentrum München; University of Ulm Institute for Epidemiology and Medical Biometrics Hveem, Kristian; Norwegian University of Science and Technology Faculty of Medicine and Health Sciences, K.G. Jebsen Centre for Genetic Epidemiology, Department of Public Health and Nursing, ; Norwegian University of Science and Technology Faculty of Medicine and Health Sciences, HUNT Research Centre Jonasson, Christian; Norwegian University of Science and Technology Faculty of Medicine and Health Sciences, K.G. Jebsen Centre for Genetic Epidemiology, Department of Public Health and Nursing, ; Norwegian University of Science and Technology Faculty of Medicine and Health Sciences, HUNT Research Centre Köttgen, Anna; Universitatsklinikum Freiburg, Institute of Genetic Epidemiology; Johns Hopkins Bloomberg School of Public Health, Department of Physiology and Biophysics Butterworth, Adam; University of Cambridge Department of Public Health and Primary Care, Cardiovascular Epidemiology Unit Prunotto, Marco; University of Geneva, School of Pharmaceutical Sciences Hauck, Stefanie; Helmholtz Zentrum München Deutsches Forschungszentrum für Gesundheit und Umwelt, Research Unit Protein Science and Core Facility Proteomics Suhre, Karsten; Weill Cornell Medicine - Qatar, Department of Physiology and Biophysics Gieger, Christian; Helmoltz Zentrum München Research Unit Molecular Epidemiology; Helmholtz Zentrum München Institute of Epidemiology; German Center for Cardiovascular Research (DZHK), partner site Munich Heart Alliance Tomaszewski, Maciej; University of Manchester, Institute of Cardiovascular Sciences Teumer, Alexander; Universitätsmedizin Greifswald, Institute for Community Medicine; German Center for Cardiovascular Research (DZHK), partner site Greifswald Waldenberger, Melanie; Helmoltz Zentrum München Research Unit Molecular Epidemiology; Helmholtz Zentrum München Institute of Epidemiology; German Center for Cardiovascular Research (DZHK), partner site Munich Heart Alliance
Keywords:	Mendelian randomization, chronic kidney disease, glomerular filtration rate, Epidemiology and outcomes, plasma proteomics, biomarker discovery, testican-2, gene expression

Authors: MatÃ-as-GarcÃ-a, Pamela; Wilson, Rory; Guo, Qi; Zaghlool, Shaza; Eales, James; Xu, Xiaoguang; Charchar, Fadi; Dormer, John; Schlosser, Pascal; Elhadad, Mohamed; Nano, Jana; Sharma, Sapna; Peters, Annette; Fornoni, Alessia; Mook-Kanamori, Dennis ; Winkelmann, Juliane; Danesh, John; Di Angelantonio, Emanuele ; Ouwehand, Willem; Watkins, Nicholas; Roberts, David; Petrera, Agnese; Graumann, Johannes; Koenig, Wolfgang; Hveem, Kristian; Jonasson, Christian; KÃ¶ttgen, Anna; Butterworth, Adam; Prunotto, Marco; Hauck, Stefanie; Suhre, Karsten; Gieger, Christian; Tomaszewski, Maciej; Teumer, Alexander; Waldenberger, Melanie

Title: Plasma proteomics of renal function: a trans-ethnic meta-analysis and Mendelian randomization study

Running title: Plasma proteomics of renal function

Manuscript Type: Original Article - Clinical Epidemiology

Funders: British Heart Foundation, (Grant / Award Number: 'PG/17/35/33001','PG/19/16/34270')
National Institute for Health Research, (Grant / Award Number:)
NIHR, (Grant / Award Number:)
Qatar National Research Fund (QNRF), (Grant / Award Number: 'NPRPC11-0115-180010')
Kidney Research UK, (Grant / Award Number: 'RP_013_20190305','RP_017_20180302')
Norwegian Institute of Public Health, (Grant / Award Number:)
Nord-TrÃ¸ndelag County Council, Central Norway Health Authority, (Grant / Award Number:)
NHS, (Grant / Award Number:)
Federal Ministry of Education and Research (BMBF), (Grant / Award Number:)
Qatar Foundation, (Grant / Award Number:)
Weill Cornell Medicine in Qatar, (Grant / Award Number:)
Faculty of Medicine, Norwegian University of Science and Technology, (Grant / Award Number:)
State of Bavaria, (Grant / Award Number:)

Financial Disclosure: No The authors declare the following competing interests: C.J. has received personal fees for research consultancy work from Pfizer and Bayer outside of the submitted work, and J.D. sits on the International Cardiovascular and Metabolic Advisory Board for Novartis (since 2010); the Steering Committee of UK Biobank (since 2011); the MRC International Advisory Group (ING) member, London (since 2013); the MRC High Throughput Science â€˜Omics Panel Member, London (since 2013); the Scientific Advisory Committee for Sanofi (since 2013); the International Cardiovascular and Metabolism Research and Development Portfolio Committee for Novartis; and the Astra Zeneca Genomics Advisory Board (2018). Other authors declare no competing interests.

Study Group/Organization Name: Human Kidney Tissue Studies

Study Group Members' Names: Wojciech Wystrychowski1, Monika Szulinska2, Andrzej Antczak3, Maciej Glyda4, Robert KrÃ³l1, Joanna Zywiec5, Ewa Zukowska-Szczechowska6, Pawel Bogdanski2
1 Department of General, Vascular and Transplant Surgery, Medical University of Silesia, Katowice, Poland
2 Department of Treatment of Obesity, Metabolic Disorders and Clinical Dietetics, Poznan University of Medical Sciences, Poznan, Poland.

3 Department of Urology and Uro-oncology, Karol Marcinkowski University of Medical Sciences, Poznan, Poland
4 Department of Transplantology and General Surgery Poznan, Collegium Medicum, Nicolaus Copernicus University, Bydgoszcz, Poland
5 Department of Internal Medicine, Diabetology and Nephrology, Medical University of Silesia, Zabrze, Poland
6 Department of Health Care, Silesian Medical College, Katowice, Poland

Total number of words: 4900

Abstract: Background. Studies on the relationship between renal function and the human plasma proteome have identified several potential biomarkers. However, studies have been conducted largely in European populations, and whether the associations between plasma proteins and kidney function are causal has never been addressed.
Methods. A cross-sectional study of 993 plasma proteins and 2,882 participants of four studies of European and admixed ancestries (KORA, INTERVAL, HUNT, QMDiab) was conducted to identify trans-ethnic associations between eGFR/CKD and proteomic biomarkers. For the replicated associations, two-sample bidirectional Mendelian randomization (MR) was used to investigate potential causal relationships, followed by the analysis of gene expression in kidney.
Results. Fifty-seven plasma proteins were associated with eGFR, including two novel proteins, JAM-B and contactin-4. Nineteen of these were additionally associated with CKD. The strongest inferred causal effect was the positive effect of eGFR on testican-2, an effect in line with the known biological role of this protein and the expression of the protein-coding gene of testican-2 (SPOCK2) in renal tissue. Finally, we observed suggestive evidence of an effect of melanoma inhibitory activity (MIA), carbonic anhydrase III and cystatin-M on eGFR.
Conclusions. In a discovery-replication setting, we identified 57 proteins trans-ethnically associated with eGFR, including two potential novel biomarkers. Our findings with regard to causal relationships represent an important stepping-stone in the establishment of testican-2 as a clinically relevant physiological marker of kidney disease progression, and point to additional potential therapeutic targets warranting further investigation.

Plasma proteomics of renal function: a trans-ethnic meta-analysis and Mendelian randomization study

Significance statement

Prior studies on the plasma proteome of renal function have identified several biomarkers, but have lacked replication, were limited to European populations and/or did not investigate causality. This paper describes, firstly, the identification of plasma proteomic biomarkers in the largest cross-sectional study of renal function to date. Using four studies in a discovery-replication setting, 57 protein biomarkers trans-ethnically associated with eGFR and/or CKD were identified, two of which are novel. Investigations into causality using Mendelian randomization provide suggestive evidence for a few proteins warranting further investigation as therapeutic targets, and highlight testican-2 as a protein affected by renal function, an early milestone in its establishment as a physiological marker of kidney disease progression with potential clinical relevance.

Plasma proteomics of renal function: a trans-ethnic meta-analysis and Mendelian randomization study

Pamela R. Matías-García (ORCID iD: 0000-0002-3906-5405)[1,2,3,4]*, Rory Wilson[1,2], Qi Guo[5], Shaza B. Zaghlool[6], James M. Eales[7], Xiaoguang Xu[7], Fadi J. Charchar[8-10], John Dormer[11], Pascal Schlosser (0000-0002-8460-0462)[7,12], Mohamed A. Elhadad[1,2,4], Jana Nano[2,8,13], Sapna Sharma[1,2], Annette Peters[2,4,8,13], Alessia Fornoni[9]Fornoni[14], Dennis O. Mook-Kanamori[10]Kanamori[15], Juliane Winkelmann[11,12,16, 17], John Danesh[13, 14, 15, 16, 17, 18-23], Emanuele Di Angelantonio[13,14, 15, 16, 17,18-22], Willem H. Ouwehand[14, 19, 20, 21,19, 24-26], Nicholas A. Watkins[20]Watkins[25], David J. Roberts[15, 22, 23,20, 27, 28], Agnese Petrera (ORCID iD: 0000-0002-0583-6195)[29], Johannes Graumann (ORCID iD: 0000-0002-3015-5850)[24, 25,30, 31], Wolfgang Koenig[4,26, 27,32, 33], Kristian Hveem[28, 29,34, 35], Christian Jonasson[28, 29,34, 35], Anna Köttgen (0000-0002-4671-3714)[7,12,30,36], Adam Butterworth[5], Marco Prunotto (ORCID iD: 0000-0002-0203-0129)[31,37], Stefanie M. Hauck (ORCID iD: 0000-0002-1630-6827)[29], Karsten Suhre[6], Christian Gieger[1,2,4], Maciej Tomaszewski [7, 38], Human Kidney Tissue Studies** (names and affiliations provided after title page), Alexander Teumer[32,33,39, 40], Melanie Waldenberger (ORCID iD: 0000-0003-0583-5093)[1,2,4]*

[1] Research Unit of Molecular Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, 85764 Neuherberg, Bavaria, Germany.

[2] Institute of Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Bavaria, Germany.

[3] TUM School of Medicine, Technical University of Munich, Munich, Germany.

[4] German Center for Cardiovascular Research (DZHK), partner site Munich Heart Alliance, Munich, Germany.

[5] Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Strangeways Research Laboratory, Wort's Causeway, Cambridge, UK.

[6] Department of Physiology and Biophysics, Weill Cornell Medicine-Qatar, Doha, Qatar.

[7] Division of Cardiovascular Sciences, University of Manchester, Manchester, UK.

[8] School of Health and Life Sciences, Federation University Australia, Ballarat, VIC, Australia.

[9] Department of Cardiovascular Sciences, University of Leicester, Leicester, UK.

[10] Department of Physiology, University of Melbourne, Melbourne, VIC, Australia.

[11] University Hospitals of Leicester NHS Trust, Leicester, UK.

[127] Institute of Genetic Epidemiology, Faculty of Medicine and Medical Center-University of Freiburg, Freiburg, Germany.

[138] German Center for Diabetes Research (DZD), München-Neuherberg, Neuherberg, Germany.

[149] Katz Family Division of Nephrology and Hypertension, Department of Medicine, University of Miami Miller School of Medicine, Miami, FL, USA

[1510] Department of Clinical Epidemiology, Leiden University Medical Centre, Leiden, The Netherlands.

[1611] Institute of Neurogenomics, Helmholtz Zentrum München, German Research Centrer for Environmental Health, Neuherberg, Germany.

[1712] Department of Neurogenetics and Institute of Human Genetics, Technical University of Munich, Munich, Germany.

[1813] British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK

[1914] British Heart Foundation Centre of Research Excellence, University of Cambridge, Cambridge, UK.

[2015] National Institute for Health Research Blood and Transplant Research Unit in Donor Health and Genomics, University of Cambridge, Cambridge, UK.

[2116] National Institute for Health Research Cambridge Biomedical Research Centre, University of Cambridge and Cambridge University Hospitals, Cambridge, UK.

[2217] Health Data Research UK Cambridge, Wellcome Genome Campus and University of Cambridge, Cambridge, UK.

[23][18] Department of Human Genetics, Wellcome Sanger Institute, Hinxton, UK.

[24][19] Department of Haematology, University of Cambridge, Cambridge, UK.

[25][20] NHS Blood and Transplant, Cambridge Biomedical Campus, Long Road, Cambridge, UK.

[26][21] Wellcome Sanger Institute, Hinxton, UK.

[27][22] NHS Blood and Transplant-Oxford Centre, Level 2, John Radcliffe Hospital, Oxford, UK.

[28][23] Radcliffe Department of Medicine, University of Oxford, John Radcliffe Hospital, Oxford, UK.

[29] Research Unit Protein Science and Core Facility Proteomics, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany.

[30][24] Scientific Service Group Biomolecular Mass Spectrometry, Max Planck Institute for Heart and Lung Research, W.G. Kerckhoff Institute, Bad Nauheim, Germany.

[31][25] German Centre for Cardiovascular Research (DZHK), partner site Rhine-Main, Max Planck Institute of Heart and Lung Research, Bad Nauheim, Germany.

[32][26] Deutsches Herzzentrum München, Technical University of Munich, Munich, Germany.

[33][27] Institute of Epidemiology and Medical Biometry, University of Ulm, Ulm, Germany.

[34][28] K.G. Jebsen Centre for Genetic Epidemiology, Department of Public Health and Nursing, Faculty of Medicine and Health Science, Norwegian University of Science and Technology, Trondheim, Norway.

[35][29] HUNT Research Centre, Faculty of Medicine, Norwegian University of Science and Technology, Levanger, Norway.

[36][30] Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA.

[37][31] School of Pharmaceutical Sciences, University of Geneva, Geneva, Switzerland.

[38] Division of Medicine and Manchester Academic Health Science Centre, Manchester University NHS Foundation Trust, Manchester, UK.

[39][32] Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany.

[40][33] German Center for Cardiovascular Research (DZHK), partner site Greifswald, Greifswald, Germany.

Running title: Plasma proteomics of renal function

Abstract word count: ~~231~~239

Main text word count: ~~3,435~~4960

Corresponding author's information: * Correspondence to Pamela R. Matías-García and Melanie Waldenberger. Research Unit of Molecular Epidemiology - Institute of Epidemiology, Helmholtz Zentrum München, Ingolstädter Landstraße 1, 85764 Munich/Neuherberg, Germany; pamela.matias@helmholtz-muenchen.de, waldenberger@helmholtz-muenchen.de

Keywords: proteomics; Mendelian randomization; causal analysis; genetic variant; biomarkers; chronic kidney disease; eGFR; testican-2; genetic epidemiology; plasma proteins.

Names and affiliations from authors in the group Human Kidney Tissue Studies**

** The following investigators contributed to recruitment and/or phenotyping of human kidney gene expression studies reported in [1]: Wojciech Wystrychowski[1], Monika Szulinska[2], Andrzej Antczak[3], Maciej Glyda[4], Robert Król[1], Joanna Zywiec[5], Ewa Zukowska-Szczechowska[6], Pawel Bogdanski[2]

1 Department of General, Vascular and Transplant Surgery, Medical University of Silesia, Katowice, Poland

2 Department of Treatment of Obesity, Metabolic Disorders and Clinical Dietetics, Poznan University of Medical Sciences, Poznan, Poland.

3 Department of Urology and Uro-oncology, Karol Marcinkowski University of Medical Sciences, Poznan, Poland

4 Department of Transplantology and General Surgery Poznan, Collegium Medicum, Nicolaus Copernicus University, Bydgoszcz, Poland

5 Department of Internal Medicine, Diabetology and Nephrology, Medical University of Silesia, Zabrze, Poland

6 Department of Health Care, Silesian Medical College, Katowice, Poland

Significance statement

Prior studies on the plasma proteome of renal function have identified several biomarkers, but have lacked replication, were limited to European populations and/or did not investigate causality. This paper describes, firstly, the identification of plasma proteomic biomarkers in the largest cross-sectional study of renal function to date. Using four studies ~~total~~ in a discovery-replication setting, 57 protein biomarkers trans-ethnically associated with eGFR and/or CKD were identified, two of which are novel. Investigations into causality using Mendelian randomization provide suggestive evidence for a few proteins warranting further investigation as therapeutic targets, and highlight testican-2 as a protein affected by renal function, an early milestone in its establishment as a physiological marker of kidney disease progression with potential clinical relevance.

Abstract

Background. Studies on the relationship between renal function and the human plasma proteome have identified several potential biomarkers. However, studies have been conducted largely in European populations, and whether the associations between plasma proteins and kidney function are causal has never been addressed.

Methods. A cross-sectional study of 993 plasma proteins and 2,882 participants of four studies of European and admixed ancestries (KORA, INTERVAL, HUNT, QMDiab) was conducted to identify trans-ethnic associations between eGFR/CKD and proteomic biomarkers. For the replicated associations, two-sample bidirectional Mendelian randomization (MR) was used to investigate potential causal relationships~~.~~, followed by the analysis of gene expression in kidney.

Results. Fifty-seven plasma proteins were associated with eGFR, including two novel proteins, JAM-B and contactin-4. Nineteen of these were additionally associated with CKD. The strongest inferred causal effect ~~estimated by MR~~ was the positive effect of eGFR on testican-2, an effect in line with the known biological role of this protein and the ~~glomeruli-specific~~ expression of the protein-coding gene of testican-2 (*SPOCK2*) in renal tissue. Finally, we observed suggestive evidence of an effect of melanoma inhibitory activity (MIA), carbonic anhydrase III and cystatin-M on eGFR.

Conclusions. In a discovery-replication setting, we identified 57 proteins trans-ethnically associated with eGFR, including two potential novel biomarkers. Our findings with regard to causal relationships represent an important stepping-stone in the establishment of testican-2 as a clinically relevant physiological marker of kidney disease progression, and point to additional potential therapeutic targets warranting further investigation.

Introduction

The kidneys' ability to filter blood and maintain homeostasis is reflected in the glomerular filtration rate (GFR)[2]. Serum creatinine, a filtration marker, is freely filtered by the glomerulus[3] and its blood levels can be used to calculate estimated GFR (eGFR)[4]. Chronic kidney disease (CKD), characterized by reduced eGFR (<60 ml/min/m^2) and proteinuria, has a global prevalence of 10% to 16%[5, 6] and is expected to be increasingly common in aging populations[2]. Increased serum creatinine is not evident until ~50% of the renal filtration function is lost[7], making CKD a silent disease and creating a blind spot for early renal disease detection[8]. Its rising prevalence, in addition to the lack of therapeutic options[8], imposes a significant burden on health systems and individuals worldwide[2, 6].

A number of biomarker research studies have been conducted in regard to early detection, diagnosis and/or progression prediction of kidney diseases[7, 8]. Early efforts in renal function proteome research focused on urine biomarkers, though most studies were small and lacked replication[7, 8]; combining multiple urinary biomarkers proved to be more successful (e.g. CKD273 classifier predicting deterioration and mortality)[7, 9]. More recent studies have focused on blood, an easily accessible tissue that reflects the metabolic status of multiple organs. Its complex proteomic profile, however, requires sensitive and reliable techniques for its study. A promising tool in this regard is SOMAscan, a platform using DNA aptamers to measure hundreds of plasma proteomic biomarkers[10]. Although this platform has been successfully used in different epidemiological settings[11-15], renal disease has not been sufficiently investigated: prior studies have either tested a limited number of proteins[16] or relied on small samples without replication[10, 17].

Moreover, prior studies on proteomic markers and renal function have not distinguished causality from correlation[18]. Mendelian randomization (MR), an instrumental variable analysis used to infer causal effects on a given outcome by relying on the random allocation of alleles at conception and genetic variation as a proxy to lifetime exposures, is an increasingly popular method used in genetic epidemiology studies to address causality[19, 20].

Here we present a cross-sectional study using a multiplexed aptamer-based proteomics platform to investigate associations between 1095 plasma proteins and GFR/CKD and other renal parameters in a discovery cohort (KORA, N=995), with replication in three independent studies of European and admixed ancestry (INTERVAL, HUNT and QMDiab; N =1,887). To better understand the biological significance of the identified proteins, we conducted enrichment, protein and gene expression analyses across tissues, as well as investigated their interconnection using protein-protein interaction (PPI) network analysis. We also investigated causal effects between eGFR and the replicated proteins using two-sample bidirectional Mendelian randomization, and further examined gene expression in kidney tissue to further ~~explore~~ investigate the proteins with ~~the strongest~~ evidence of a causal effect.

Methods

Study populations

The KORA study (Cooperative Health Research in the Region of Augsburg) is a population-based sample from the general population living in the region of Augsburg, Southern Germany. The KORA F4 survey, a follow-up of the KORA S4 prospective cohort (1999-2001), was conducted from 2006 to 2008 and included a total of 3080 participants. Clinical and demographic information, as well as peripheral blood for 'omics' analyses, were collected; details on the standardized examinations, interviews and tests conducted in the KORA study have been previously described[21, 22]. This study acted as discovery cohort in the cross-sectional association study of plasma proteins and renal function (Fig. 1A).

Included in the replication phase were the Nord-Trøndelag Health Study (HUNT), namely the third survey (HUNT3) from this population-based study from Norway with data on participants of European descent[23]; the INTERVAL Study (INTERVAL), a randomized trial assessing blood donation practices across the UK with extensive phenotyping available for 50,000 participants of European descent[24]; and the Qatar Metabolomics Study on Diabetes (QMDIAB), a cross-sectional case-control study on type 2 diabetes from participants of Arab,

South Asian and Filipino descent in Qatar[25]. Population characteristics from the four studies are shown in Table 1. Information on data availability is given in Supplemental Note 1.

Sample collection and proteomic profiling

EDTA plasma samples collected by the studies following standardized procedures were centrifuged, aliquoted and stored at -80°C[26-28]. Samples for proteomic profiling and GFR estimation were taken at the same time.

Proteomic profiling in all participating studies was done using SOMAscan (SomaLogic, Inc), an aptamer-based, affinity proteomics platform[10, 29-31]. Plasma samples from KORA, HUNT3 and INTERVAL were shipped on dry ice to SomaLogic, Boulder, CO, and proteomic profiling was performed using a SOMAscan panel of 1029 proteins for KORA[26], 3,622 for INTERVAL[27] and 5000 for HUNT3[28]. In the QMDiab cohort, the kit-based SOMAscan platform was run by the Weill Cornell Medicine - Qatar (WCM-Q) proteomics core following protocols and instrumentation provided by SomaLogic Inc., under supervision of SomaLogic personnel, to measure 1,129 proteins in plasma samples[26]. The samples were all measured by individuals blinded to the identities corresponding to the samples.

~~***Measurement of plasma proteins***~~

~~Proteomic profiling in all participating studies was done using SOMAscan (SomaLogic, Inc), an aptamer-based, affinity proteomics platform[10, 29-31].~~In summary, fluorescently labeled single-stranded synthetic nucleotides (Slow Off-rate Modified Aptamers, SOMAmers) immobilized on streptavidin-coated beads are incubated with plasma samples to capture proteins and generate SOMAmer-protein complexes. Washing steps eliminate unbound SOMAmers and unbound/non-specifically bound proteins. The next steps are biotin-labeling and photocleavage to liberate SOMAmer-protein complexes from the beads. This is followed by incubation in a buffer disrupting nonspecific interactions, recapturing the biotin-labeled protein/aptamer complexes in streptavidin-coated beads and additional washing steps to remove nonspecific SOMAmers. These are then eluted from the target proteins and quantified on custom DNA microarrays using deposited SOMAmer-complementary oligonucleotides, which produces measurements in relative fluorescence~~t~~ units (RFU) as

proxies to protein concentrations. Quality control at the sample and SOMAmer levels using control aptamers and calibrator samples was conducted by the manufacturer. Based on standard samples included on each plate, the resulting raw intensities are processed using a data analysis work flow including hybridization normalization, median signal normalization and signal calibration to control for inter-plate differences.

In KORA F4, a random subset of 1,000 participants of those with omics data was selected for proteomic profiling with the SOMAscan assay featuring 1,129 protein-specific SOMAmer probes[26] in fasting plasma samples. SOMAscan QC resulted in the exclusion of 29 proteins and one sample, and five proteins were further excluded due to cross-reactivity (publicly available communication "SSM-064_Rev_0_DCN_16-263" issued by SomaLogic), producing data on k = 1,095 proteins in 999 participants in the discovery dataset (Supplementary Table 1). Proteomics data from INTERVAL featured 2,994 proteins in 3,301 participants[27], HUNT included 1,054 proteins in 2,432 individuals[12] and ~~in~~ QMDiab 1,130 in 352 participants[26] after study-level quality control, for an overlap of 993 available proteins across datasets; further details on the proteomics profiling from the samples included in this study are described elsewhere [12, 26, 27]. Protein mapping to several identifiers was provided by the manufacturer (Supplementary Table 1).

Outcome definitions

Our first analysis is a proteome wide association study: we investigated associations between proteins and renal traits as outcomes, using linear regression models with adjustment for potential confounders. The primary outcomes studied in this analysis were estimated glomerular filtration rate from serum creatinine (eGFR) and chronic kidney disease (CKD), given their availability in all included studies.

eGFR was calculated using the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) equation with serum creatinine[4] with the R package *nephro* v1.2[32]. Serum creatinine was measured using the modified kinetic Jaffé reaction in KORA, HUNT and QMDiab (and calibrated by multiplying by 0.95[33]), and a nuclear magnetic resonance (NMR) platform (Nightingale Health) in the INTERVAL study. Pearson's correlation between serum

creatinine-based eGFR and this NMR-based eGFR variable was estimated in KORA (Suppl. Fig. 1). Chronic Kidney Disease (CKD) was defined as eGFR <60 ml/min/1.73 m^2[2,34].

We performed some analyses for outcomes available only in the discovery study. Urinary albumin and urinary creatinine were used to calculate urinary albumin-to-creatinine ratio (uACR) and its derived parameter microalbuminuria (MA, defined as uACR > 30 mg/g). eGFR decline was defined as $\log(eGFR)_{follow\text{-}up} - \log(eGFR)_{baseline}$ divided by the follow-up time, where KORA F4 (2006-2008) was used as baseline and KORA FF4 (2013-2014) its follow-up survey. Sensitivity analyses were also run using $eGFR_{cys}$ (derived from CKD-EPI equation using cystatin C[13]).

Definition of covariates

Covariates used in the regression analyses were: age at the time of examination, sex, BMI, smoking status, diabetes (yes/no), hypertension (yes/no), log-transformed triglycerides, high-density lipoprotein, and intake of lipid-lowering drugs (yes/no). See Supplemental Note ~~1~~ 2 for precise cohort-specific definitions of covariates used.

Statistical analysis

Data preprocessing and statistical analyses were conducted using the R language for statistical computing v.3.6.0[35]. Prior to statistical analysis, proteomic data was log-transformed and standardized. Linear regression was used to examine the association between protein levels and continuous kidney traits (log-transformed eGFR, urinary-to-albumin creatinine ratio, eGFR change), whereas logistic regression was used for binary kidney traits (CKD, MA). Multiple testing was accounted for using a Bonferroni correction considering the total number of investigated proteins at each stage (k = 1,095 in discovery).

Sensitivity analyses in the discovery sample included regression models with serum creatinine-based eGFR as outcome and no adjustment for BMI or diabetes, as well as models including cystatin C-based eGFR as outcome and the same set of covariates from the main model. Pearson's correlations between the regression coefficients resulting from the sensitivity and the main analyses were calculated. Interaction analyses were also

conducted for the proteins identified at discovery by adding an interaction term, considering age, sex and smoking as interactors, to the fully adjusted model (Supplemental Note 3).

For those protein-outcome pairs significantly associated in the discovery, two replications were conducted: a European replication (R1) and a replication in an admixed population (R2) (Fig. 1A). Replication was defined as $p<0.05$ and consistent direction of effect as in the discovery study. The European replication for eGFR consisted of the meta-analysis of results from HUNT and INTERVAL using ~~a modified form of~~ the Stouffer's method ~~(weighted by the sample size of each individual study)~~[36]~~, as common meta-analysis approaches using regression coefficients could not be used due to the relative nature of the SOMAscan data (measured in relative fluorescence units, RFU).~~ , a P-value combination method especially useful when raw data cannot be pooled across studies – which is the case with aptamer-based measurements, where data in relative fluorescence units (RFU) is not directly comparable across studies. Stouffer's method, also known as "inverse normal" or weighted Z-test, is a P-value combination method taking the P-values for the i-th study (pi), transforming them by the inverse normal distribution function and weighing them using the square root of the sample sizes as weights (wi). The sum is then computed, and the combined P-value is obtained using the distribution of the resulting statistic, $T = \Sigma wiH(pi)$ [36].

For CKD, only the HUNT study was used in the European replication (INTERVAL had only one case), and the admixed population replication was based on the results of QMDiab. Our final set of trans-ethnic associations (R3) were those pairs of proteins-outcomes that were replicated in both R1 and R2. Replicated eGFR-associated proteins were taken to the next stages of the analysis: proteomic target validation, enrichment analyses and Mendelian randomization.

Validation of proteomic targets

We examined the plasma levels of proteins measured by Proximity Extension Assay (PEA) technology (Olink) in a subgroup of randomly selected participants from the KORA F4 study (N = 173) ([37], DOI: 10.1021/acs.jproteome.0c00641). In brief, protein abundance was

quantified by real-time PCR using the PEA proteomic technology (Olink), producing relative quantification data reported in NPX values (normalized protein expression levels, on log2 scale); NPX values were intensity normalized with the plate median for each assay as the normalization factor, and samples and proteins that did not pass the quality control were excluded (DOI: 10.1021/acs.jproteome.0c00641). Seven of the most relevant proteins (i.e. cystatin-C, RELT, IGFBP-6, myoglobin, TNF sR-I, RGMB and FSTL3), the novel proteins here reported (i.e. JAM-B and contactin-4), as well as three of the proteins identified in the causal inference analysis (i.e. carbonic anhydrase 3, MIA, and cystatin M) were included in this subset of proteomic measurements. Of note, testican-2 was not measured in this assay. Scatterplots of the aptamer-based and PEA measurements, annotated with their Pearson's correlation and statistical significance, are shown in Suppl. Fig. 2.

Information on specificity and cross-reactivity of the aptamers was available from three independent studies[27, 38, 39] for 54 of the 57 proteins identified to be trans-ethnically associated with eGFR. Target specificity issues (i.e. comparable binding observed to a target that is not the product of the same gene) ~~was~~ were observed in four cases (ephrin-A5, IGFBP-5, hemojuvelin, and cystatin SA)[27, 38] (Supplementary Tables 2-4). Moreover, in previous studies, 23 of the 57 proteins were directly validated via mass spectrometry in blood plasma/serum, and other biological matrices[39] (Supplementary Table 2), and 49 using solution affinity measurements[27, 38] (Supplementary Tables 3-4).

Functional annotation~~, and~~ enrichment and expression ~~analysis~~analyses

Annotation was done using the R package *InterMineR* v1.6.1 [40], a tool facilitating access to data from the HumanMine release 6.0 (May 2019). DAVID v.6.8 [41] was used to look for annotations for Gene Ontology Terms (molecular function, biological process~~),~~) and pathway ~~and gene~~ information, as well as to identify publications relevant to the set of 57 replicated proteins. Gene information was retrieved from the human assembly GRCh37 (hg19) using BioMart v.4 [42].

To investigate the expression patterns of the 57 eGFR-associated proteins and their corresponding protein coding genes across tissues, we used proteomics and RNA-seq expression data from the ProteomicsDB [43, 44] and the Genotype-Tissue Expression (GTEx) database [45]. The data presented and described in this manuscript were generated on Oct. 2, 2020 through a multi-gene query on the ProteomicsDB Analytics Toolbox portal from: https://www.proteomicsdb.org/proteomicsdb/#analytics/expressionHeatmap and GTEx portal https://www.gtexportal.org/home/multiGeneQueryPage

Protein-protein interaction (PPI) network analysis

We queried STRING [46], the protein-protein interaction server, to examine the relationship between the proteins that were identified as robustly associated with eGFR across studies and ethnicities (k = 57 trans-ethnically eGFR-associated proteins). We used the set of SOMAscan proteins available across studies as background (k = 993), adding no additional interactors (proteins) to the network during the analyses, and considered a minimum required interaction score for a medium confidence (0.400).

Mendelian randomization

Mendelian Randomization (MR), an instrumental variable method used to infer causality, leverages the natural randomization inherent in the (random) assortment of genes during gamete formation to assess the effect of lifelong exposures on health outcomes [47]. Single-nucleotide polymorphisms (SNPs) are used as instrumental variables (or instruments), given that their alleles are randomly assigned to individuals prior to any exposures/outcome and that they are non-modifiable, thus minimizing the risk of reverse causation and confounding [47]. The idea behind MR is that if genetic variation produces differences mirroring the biological effects of environmental exposures that alter disease risk, then genetic variation itself should be related to disease risk by having an influence on the exposure [47, 48]. MR uses SNPs as surrogates for an exposure of interest, allowing the estimation of the effects of life-long, genetically determined “exposures” on health outcomes [47]. MR produces robust causal inference estimates if the SNPs used are valid instruments – that is, if they meet the three assumptions upon which MR relies: SNPs must be strongly associated with the exposure,

and not associated with either (measured or unmeasured) confounders or with the outcome (only potentially through the exposure) [49]. Causality in MR is thus defined as the modification of an exposure leading to a change in the outcome, where the inferred causal effects by MR do not necessarily imply the existence of a straightforward interpretation with respect to direct causal factors [48].

To investigate whether genetic liability to lower or higher eGFR causally alters plasma protein levels and *vice versa*, MR was conducted ~~Causality was assessed~~ in the set of 57 proteins whose associations with eGFR showed trans-ethnic replication. Two-sample bidirectional Mendelian randomization [19] (MR) was used to ~~estimate~~ infer the causal effect of renal function (eGFR as proxy thereof) on plasma protein levels (forward MR) and vice versa (reverse MR, Fig. 1B). Results from publicly available genome-wide association studies (GWAS) for (a) eGFR from the CKDGen consortium (meta-analysis of European-ancestry populations [50]), and (b) plasma proteins from INTERVAL[27] and AGES-Reykjavik[39] were used to perform MR using MRBase[51]. A detailed account on the MR methods, data sources and analyses conducted is available in Supplemental Note ~~34~~.

Instrument selection

In the forward MR (i.e. assessing the effect of renal filtration on protein levels), 256 SNPs associated with eGFR at genome-wide significance in the CKDGen results were selected as candidate instrumental variables (IV). These SNPs were then filtered based on their relevance to renal function (associated with BUN, a complementary renal trait, with an opposite direction of effect, N_{IV} = 47) and clumped based on linkage disequilibrium (r2 = 0.01 and Kb = 10,000) to identify independent variants (N_{IV}=41). Summary statistics on 41 SNP-eGFR associations were extracted from the CKDGen results, and its corresponding SNP-protein associations were extracted from the INTERVAL results for 47 proteins. For investigating the causal effects of eGFR on proteins, 47 eGFR-protein relationships were instrumented by 41 SNPs.

For the reverse MR (i.e. interrogating the causal effect of proteins on renal filtration), genome-wide significant cis-SNPs for ~~22~~ 28 proteins were identified in the INTERVAL results

as candidate IV and LD clumped (same criteria as forward MR). Summary statistics on SNP-protein associations for ~~21~~ 28 proteins were extracted from the INTERVAL results, and its corresponding SNP-eGFR associations were extracted from the CKDGen results. The same strategy was followed to identify instruments in the AGES-Reykjavik results~~; for the 36 proteins not found in INTERVAL;~~ SNP-protein results were extracted from this dataset for ~~seven~~ 29 proteins, and ~~their corresponding~~ SNP-eGFR results were extracted for 26 proteins from the CKDGen data. Further details of the genetic instrument selection and data harmonization process are shown in Suppl. Fig. ~~2~~3 and Supplementary Table 5. Thus for investigating the causal effect of proteins on eGFR, ~~28~~ 35 protein-eGFR relationships were instrumented by 1-~~3~~ 5 SNPs, of which 17 proteins were examined using data from both INTERVAL and AGES-Reykjavik (Suppl. Fig. 4). ~~Further details of the genetic instrument selection and data harmonization process are shown in Suppl. Fig. 2 and Supplementary Table 5.~~

Data harmonization, phenotypic variance explained and instrument specificity

Details on data harmonization, the handling of palindromic SNPs and calculating the phenotypic variance explained by the SNPs are given in Supplemental Note ~~3~~4. Harmonized datasets used in the MR analyses are available in Supplementary Table 6.

In order to look for further evidence of horizontal pleiotropy, association between our SNPs and other traits were searched for in the GWAS Catalog[52] (Suppl. Table 7).

MR and sensitivity analyses

The primary MR analysis used inverse variance weighted (IVW) regression. In this method the coefficient of the gene-outcome association is regressed on the coefficient of the gene-exposure association with the intercept constrained to zero, assuming no directional pleiotropy[53, 54]. Since IVW requires two or more SNPs, in cases where only one SNP instrumented the analysis, Wald's ratio (coefficient of the gene-outcome association divided by the gene-exposure association) was calculated instead[54].

For MR analyses instrumented by more than two SNPs, three further MR methods were used as sensitivity analyses[55]. MR-Egger regression was used to assess pleiotropy, as this

method allows for horizontal pleiotropy and provides an estimate of the unbalanced horizontal pleiotropic effects in its intercept[56]. Weighted median[57] and weighted mode MR[58], methods less sensitive to the presence of invalid instruments and to pleiotropic SNPs behaving as outliers, were also used. A number of additional analyses were run to check for outliers, directional pleiotropy and heterogeneity, as recommended [55, 59]. Details are given in Supplemental Note ~~34~~.

Causal estimates were assessed at a Bonferroni-corrected significance level, i.e. 0.05 divided by the number of proteins assessed in each MR direction (47 in forward and ~~28~~ 51 in reverse MR). Causal effects were considered robust if they were significant at Bonferroni $p < 0.05$ in the IVW or Wald estimator, and results from the pleiotropy-robust sensitivity MR analyses examined to test for violations to MR assumptions.

Expression analys~~e~~is in human kidney tissue ~~cohort of CKD patients~~

The correlation between gene expression analysis of *SPOCK2,* one of the genes coding for ~~the~~ proteins with ~~the strongest and most robust~~ evidence from the causal analysis ~~(testican-2)~~, and estimated glomerular filtration rate (eGFR), was initially calculated ~~for~~ with ~~independent~~ data from microdissected tubulointerstitial components of human renal biopsies from 26 individuals with CKD at different disease stages (I-IV)[60] (GEO accession: GSE69438). Gene expression of the protein-coding genes identified in MR (*SPOCK2*, *CA3*, *CST6*, *MIA*) and renal traits was further assessed in (a) data from *Nephroseq* v5 (N = 458), a platform of comprehensive renal disease gene expression datasets [61], and (b) human kidney tissue resource characterised by RNA-sequencing (N = 427, see Supplemental Note 5) [1].

Within *Nephroseq*, univariate correlation analyses between eGFR and gene expression were conducted separately in study pre-defined histological compartments of the human kidney (i.e. glomerular and tubulo-interstitial) in 458 available kidney samples from three datasets of patients with kidney disease (Ju et al. [60], Sampson et al. [62], and Reich et al. [63]), and one dataset of "apparently" healthy renal tissue (Rodwell et al. [64]). The correlations were meta-analysed using inverse variance weighted meta-analysis with random effects models [65], and heterogeneity was assessed using Cochran's Q test.

Multivariable regression analyses were conducted in the human kidney resource (N = 427) described in [1]. In brief, we constructed linear regression models with renal expression of each candidate as the response variable; while eGFR and histologically confirmed measures of structural kidney damage were used as dependent variables together with age, sex, body mass index, 3 genetic principal components, diabetes and a variable number of surrogate variables (29 for eGFR and 26 for all histology phenotypes) [1, 66]. eGFR estimation was based on circulating levels of creatinine, as reported before [1]. Histologic measures of structural integrity (glomerular sclerosis, glomerular Bowman's capsule thickening, tubular atrophy, interstitial fibrosis, interstitial inflammation and vascular lesions) were assessed microscopically and scored on a semi-quantitative scale (whereby 0 indicates no or minimal damage and 3 is consistent with the highest degree of structural injury), as reported before [67].

Results

Figure 1 illustrates the design of the present study. First, a cross-sectional association study was performed to identify proteins associated with renal function parameters in a discovery-replication setting: KORA F4 acted as the discovery, and INTERVAL, HUNT3 and QMDiab as replication studies (Fig. 1A). Replicated trans-ethnic protein associations were then assessed for causality using two-sample Mendelian randomization, using data from the largest genome-wide association (GWA) studies available for the traits of interest (CKDGen, INTERVAL and AGES-Reykjavik) (Fig. 1B).

Cross-sectional association of plasma proteins and renal function

Population characteristics of the four cohorts included in the cross-sectional association study are shown in Table 1. The largest differences between the discovery and the replication studies were observed in the age distributions, smoking habits, blood lipid levels, eGFR distribution, and CKD/diabetes prevalence.

Results from discovery study

The association between 1,095 plasma proteins and eGFR/CKD was assessed in the KORA F4 study (N = 995). A total of 80 proteins were significantly associated with eGFR ($p < 0.05/1095$). The top 3 negative associations (i.e. higher eGFR associated with lower plasma

protein levels) were observed with cystatin C (β = -0.068 [95% CI = -0.078,-0.059] change in log-transformed eGR per standard deviation increase in protein level, p = 2.63E-40), tumor necrosis factor receptor superfamily member 19L (RELT; β = -0.063 [-0.073,-0.053], p = 7.82E-33) and beta-2-microglobulin (b2-microglobulin; β = -0.059 [-0.070,-0.050], p = 6.16E-30), whereas the strongest positive association (i.e. higher eGFR associated with higher plasma protein levels) was that of testican-2 (β = 0.036 [0.026, 0.045], p = 2.066E-13) (Supplementary Table 1, Suppl. Fig. ~~3~~ 5 and ~~4A~~6A). Of note, 34 of these 80 proteins were also associated with CKD (Supplementary Table 1).

Sensitivity analyses showed that 71 of the 80 eGFR-associated-proteins identified in the main analysis were consistently associated with cystatin C-based eGFR, with a high correlation between regression coefficients (r = 0.841, p < 2.2E-16). Models with no adjustment for BMI or diabetes produced highly similar estimates to those obtained in the main analysis (r = 0.99, p < 2.2E-16 for both; Supplementary Table 8). Likewise, the exclusion of individuals with CKD (N = 38) did not significantly affect the correlation between the plasma levels of proteins and log-transformed eGFR (Suppl. Fig. 7). Interaction analyses suggested the negative associations between log-transformed eGFR and five plasma proteins (b2-Microglobulin, IGFBP-6, FSTL3, JAM-B, and renin) were accentuated with age (i.e. each additional year of age made the association stronger) (Supplementary Table 9).

To further explore proteomic associations with renal function, additional renal outcomes were assessed in the discovery cohort. eGFR change was associated with five proteins, whereas three proteins were negatively associated with uACR (epidermal growth factor receptor (EGFR), mitochondrial superoxide dismutase [Mn] (SOD2) and coagulation factor X (F10)). No proteins were significantly associated with microalbuminuria (Supplementary Table ~~9~~10).

Results from replication studies

Serum creatinine was the only available trait across all replication cohorts (Fig. 1A), thus only associations with eGFR/CKD were further explored.

The European replication (R1) was done using HUNT3 and INTERVAL (Suppl. Fig. 4B6B-C); results from this analysis confirmed the association of 62 of the 76 proteins available across studies. The second replication round (R2) was performed in QMDiab (Suppl. Fig. 4D6D), a population of admixed-ancestry; this confirmed 63 eGFR-protein associations. High correlation between z-values from the discovery and replication studies was observed (correlations ranging from 0.66 with INTERVAL to 0.93 with HUNT3, Suppl. Fig. 8). The overlap of the proteins replicated in R1 and R2 was the final set of 57 robustly replicated trans-ethnic eGFR-protein associations (R3, Supplementary Table 1011). Figure 2 shows the cross-sectional effect estimates for the top 10 protein-eGFR associations across the four cohorts; INTERVAL, a largely healthy and younger population, showed the smallest effect sizes, whereas the strongest effects were observed in HUNT3, a cohort of older individuals with lower mean eGFR and higher CKD prevalence. Two novel proteins were identified (contactin-4 and junctional adhesion molecule B, JAM-B), and all of the 57 proteins were replicated in the cystatin C-based eGFR sensitivity analysis (Supplementary Table 8).

All 34 CKD-protein associations from the discovery phase were replicated in HUNT and 20 replicated in QMDiab; these 20 were thus considered trans-ethically robust (Supplementary Table 1112). Figure 3 shows the sets of replicated proteins associated with eGFR/CKD. JAM-B, one of the novel proteins described here, is one of the 19 proteins associated with both eGFR and CKD.

Functional annotation , and enrichment and expression analysis analyses

An extended annotation file including information on GO terms, pathways (KEGG, BIOCARTA), and protein domain annotation (INTERPRO) is given in Supplementary Table 1213, and the retrieved gene information in Supplementary Table 1314. Although several pathways, biological processes and molecular functions were represented, no enrichment for Gene Ontology (GO) terms or pathways was observed.

Protein abundance based on quantitative mass spectrometry assays (from ProteomicsDB) across different tissues, cell lines, and body fluids is shown in Suppl. Fig. 9;

peptides for 15 of the 57 eGFR-associated proteins were detected in multiple human tissues and body fluids (including kidney tissue); of these ubiquitously expressed proteins, beta-2 microglobulin and glutathione S-transferase Pi are most notable (Suppl. Fig. 9). Most protein-coding genes demonstrated ubiquitous expression across virtually all human tissues, including kidney tissue, represented in the ProteomicsDB datasets, where genes like *B2M* and *GSTP1* were highly expressed across tissues in both RNA-seq and microarray expression datasets (Suppl. Fig. 10). GTEx tissue expression data was largely concordant with the observations from ProteomicsDB (Suppl. Fig. 11).

Protein-protein interaction (PPI) network analysis

We queried STRING [46], the protein-protein interaction server, to examine the relationship between the 57 replicated proteins associated with eGFR. The obtained network features 56 nodes representing proteins, connected by 60 edges representing predicted functional protein-protein associations, with an average number of interactions of 2.14 at a minimum interaction score of 0.400 (medium confidence) in the network (Suppl. Fig. 12). The expected number of edges was 36 (PPI enrichment p-value: 1.12E-04). This means the examined set of 57 eGFR-associated proteins have more interactions among themselves than what would be expected from a random set of proteins drawn from the set of proteins included in the SOMAscan studies, and thus suggest these proteins might be meaningfully biologically connected.

Mendelian randomization

To assess whether genetic susceptibility to higher or lower plasma levels of the 57 proteins identified as trans-ethnically associated with eGFR may affect this renal trait, and whether genetic liability to altered eGFR causally alters circulating levels of plasma proteins, ~~might be~~

causally affected by or involved in changes in renal function, two-sample bidirectional Mendelian randomization (MR) was conducted (Fig. 1B).

Forward MR: eGFR has an effect on testican-2

In the forward direction of the MR (i.e. assessing inferring the effect of eGFR on levels of 47 proteins), 40 SNPs explaining 1.59% of the variance of eGFR (Supplementary Table 1415) were used as genetic instruments.

Plasma levels of Seven 7 proteins were identified as causally affected by eGFR according to the IVW model (cystatin M, cathepsin H, ephrin type-B receptor 6, insulin-like growth factor-binding protein 6, testican-2, melanoma-derived growth regulatory protein, and netrin receptor UNC5C; Supplementary Table 1516 and Suppl. Fig. 13-19). This means that if eGFR is somehow altered by means of an intervention mimicking the effect of the SNP on eGFR, MR suggests the plasma levels of these proteins will change in the predicted direction. Although no evidence of directional pleiotropy, particularly influential SNPs or instrument heterogeneity was observed (with the exception of IGFB6, Suppl. Tables 1617-1819), the pleiotropy-robust sensitivity MR analyses did not provide further evidence of support causality for six of them (i.e. non-significant results), raising the possibility of suggesting the IVW findings might be being driven by undetected balanced horizontal pleiotropy[68]. In contrast, the causal effect of eGFR on testican-2 was supported by more than one MR method (weighted median β = 6.104 [95% CI, 2.808 to 9.401] unit increase in testican-2 levels per unit increase in log-transformed eGFR, p = 2.84E-04) (Table 2 and Figure 4). Furthermore, there was no evidence of outliers, influential SNPs or instrument heterogeneity (Suppl. Fig. 517).

Eleven of the SNPs instrumenting the forward MR analysis were identified as potentially pleiotropic (associations at p < 5E-08 with other related traits, Supplementary Table 7). Results from the restrictive MR conducted after excluding these pleiotropic variants were not statistically significant, but in agreement with those from the main analysis in terms of their direction and size of effect (Suppl. Table 1920).

Reverse MR: MIA, cystatin M and carbonic anhydrase III affect eGFR

In the reverse direction of the MR (i.e. assessing the effect of ~~28~~ 35 proteins on eGFR), one to ~~three~~ five cis-SNPs explaining 0.91~~1.02~~%-29.33% of phenotypic variance were used as genetic instruments (Supplementary Table ~~14~~15). Of note, no causal effect of testican-2 on eGFR was identified in this direction of the MR (Wald's ratio p = 0.053 for variant *rs1245547* from AGES-Reykjavik and p = 0.06 for variant *rs1245540* from INTERVAL) (Suppl. Table 16).

A negative effect of melanoma inhibitory activity (MIA) on eGFR ~~(i.e. higher protein levels have a negative effect on eGFR)~~ was identified in the fixed-effect IVW and weighted-median models using the 3 SNPs identified in the INTERVAL pGWAS (Table 2, Suppl. Fig. ~~6A~~20A-B), meaning MR suggests that if plasma protein levels are lowered by means of an intervention mimicking the effect of the SNP on MIA, eGFR will increase. No evidence of influential SNPs ~~was~~ observed, yet the funnel plot suggested directional pleiotropy (Supplementary Tables ~~16~~17-~~18~~19, Suppl. Fig. 20~~6~~C). One SNP instrumenting this analysis was identified as potentially pleiotropic (Supplementary Table 7). The causal effect estimated using AGES-Reykjavik pGWAS data was in agreement with this MR estimate (Suppl. Table 16).

~~A p~~Positive effects of carbonic anhydrase III and cystatin M on eGFR ~~was~~ were also identified (Wald's ratio p = 5.04E-04 and 8.41E-05, respectively, with INTERVAL pGWAS data) (Table 2). Although further sensitivity analyses were not conducted given there was only one cis-SNP available for each protein, no gene-trait associations were found for these SNPs in the GWAS Catalog[52], suggesting pleiotropic effects to be unlikely. The effect of carbonic anhydrase III on eGFR estimated using pGWAS data from AGES-Reykjavik was in line with the aforementioned estimate (Suppl. Table 16).

~~***SPOCK2 expression in renal samples from CKD patients***~~

Gene expression in kidney tissue

~~A moderate non-significant~~ An initial assessment of the expression of the gene coding for testican-2, *SPOCK2*, in tubulointerstitial components of human renal biopsies from 26 individuals with CKD[60] found no correlation ~~was found between expression of *SPOCK2* and~~ with eGFR at the time of biopsy (Pearson's r^2 = 0.25, p = 0.22) ~~on tubulointerstitial components of human renal biopsies from 26 individuals with CKD[60]~~ (Suppl. Fig. ~~7~~21). Further univariate analyses conducted with the *Nephroseq* data showed a statistically significant correlation between eGFR and *SPOCK2* gene expression in glomerular compartment/kidney cortex (r = 0.242, p = 0.033) (Suppl. Table 21). We then conducted additional analyses using RNA-sequencing-characterised human kidney transcriptomes from a resource with up to 427 individuals with matching gene expression and renal phenotype data (as described before in [1, 66, 67] and Suppl. Table 22). There was no significant association between eGFR and renal *SPOCK2* expression in this dataset (N=427, Supplemental Note 5; Suppl. Table 23). However, in the subset with information on histology phenotypes (N=283), *SPOCK2* expression was negatively associated with both tubular atrophy (β= -0.094 [-0.177, -0.011], p = 0.03) and interstitial fibrosis (β =-0.093 [-0.175, -0.010], p = 0.03), and *CST6* expression was negatively associated with glomerular sclerosis (β = -0.094 [95% CI = -0.174, -0.013], p = 0.02).

Discussion

We conducted an association study of plasma proteomics and eGFR/CKD following a discovery-replication approach involving four independent studies. We confirm known protein associations and identify two novel potential biomarkers. Furthermore, we performed two-sample bidirectional Mendelian randomization to ~~identify~~ infer ~~causal relationships~~ causality in ~~between~~ the eGFR-~~associated~~ protein associations~~proteins~~, and found evidence of causality underlying four eGFR-protein associations.

Eighty proteins were found to be associated with eGFR in our discovery analysis, with trans-ethnic replication confirming 57 of these. Nineteen of these proteins were also found to be associated with CKD. Although our analyses use serum-creatinine-based eGFR due to its

availability across studies and its utility in clinical practice, models using cystatin C-based eGFR show the results are robust to GFR estimation method. Likewise, further sensitivity analyses indicate the associations are largely independent of adjustment for BMI or diabetes. Interestingly, from the five proteins with a significant interaction with age, protein trajectories changing with age have been reported for FSTL3, IGFBP6, JAM-B and renin [69].

We identify several well-known biomarkers of renal function[10, 16, 17, 70, 71], supporting their current use as kidney function biomarkers and the validity of our analyses. Our results are also in line with those reported by other renal studies using the same aptamer-based platform and similar proteomic profiling technologies[10, 16-18]: we replicate 15 of the proteins identified in the pioneer SOMAScan study of plasma from 42 CKD patients[10], five of the proteins associated with lower baseline eGFR and 5-year eGFR decline in a study examining 80 circulating proteins in ~1,000 participants[16], and a large number of proteins reported in a recent SOMAScan study of 2,893 plasma proteins in 389 Swedish individuals[17]. Moreover, five of our proteins (TNF SR-I and –II, TAJRELT, CD55 and CCL14) were included in a signature capturing the inflammatory process underlying end-stage renal disease in diabetic cohorts[71], and another five proteins (b2-microglobulin, cystatin C, DAF, MP2K2 and testican-2) ~~are found in a set of~~ were included ~~proteins~~ in a ~~meant to reflect renal health in a~~ "stand-alone" test meant to reflect renal health [38]. Interestingly, 40% of our proteins were identified in podocyte exosome-enriched urine, suggesting their involvement in cellular functional processes underlying glomerular filter permeability[72]. A recent aptamer-based study found 126 proteins associated with baseline eGFR [18] – an overlap of 43 proteins independently reported in both their and our study, including well-known proteomic biomarkers (e.g. cystatin-C, b2-microglobulin) and testican-2, is listed in Suppl. Table 11. The protein and gene expression results for our 57 replicated proteins from ProteomicsDB and GTEx databases confirm the renal expression of these proteins. Furthermore, the ubiquitous expression across tissues and evidence from the PPI network analysis suggest their involvement in common functional pathways.

Among the 57 eGFR-associated proteins we identified, we identify two ~~have never been reported in association with eGFR nor any other renal trait~~novel eGFR-associated proteins: contactin-4, a protein involved in neuronal network development and plasticity; and junctional adhesion molecule B (JAM-B), involved in cellular junctions in epithelial tissue. Of note, the latter was also recently identified in the study by Ngo and colleagues [18], with consistent direction of effects to our findings. Glomerular filtration of proteins (determined by their molecular weight, molecular charge and shape, and their interactions with other molecules) influences their plasma levels[73, 74]. Moreover, proteins detected in urine and plasma are derived mainly from glomerular filtration of plasma proteins and epithelial cell secretion of soluble proteins[73, 74]. Contactin-4 (113.45 kDa) and JAM-B (33.21 kDa) are present in plasma but not in urine, suggesting they are either not filtered at the glomerular capillaries (perhaps due to interactions with other proteins) or filtered but later reabsorbed into blood from the tubules; this in turn suggests changes in their plasma levels may reflect changes in glomerular and tubule function[73]. Moreover, the direction of effect observed for JAM-B (higher levels significantly associated with both lower eGFR and CKD OR > 1) is supportive of this protein being related to decreased glomerular filtering.

We also identified proteins associated with complementary renal phenotypes (eGFR change, uACR and MA). eGFR decline was associated with five proteins, of which neurexin-1-beta is the only not having been previously described as related to eGFR [10, 16, 17]. Of these five proteins, DAN, TNF sR-1 and FSTL 3 were also identified as negatively associated with eGFR decline in [18]. There was no overlap between the set of proteins associated with uACR in our study and the proteins identified in a similar study[75], which may be explained by the albuminuria samples and eGFR being measured at different time points in their study[75]. No proteins were significantly associated with MA, possibly due to its low prevalence (5.9%, Table 1) in KORA.

To investigate whether genetic susceptibility to renal function (using eGFR as a proxy thereof) or plasma protein levels may have a causal effect on the other, and identify potential causal disease pathways represented by proteins, ~~better understand the biological~~

~~mechanisms in which these proteins might be involved,~~ bidirectional MR was performed to identify causal effects of proteins on eGFR and vice versa. To date, only one study using MR to assess causality of kidney function and proteomic biomarkers has been reported[76]; however, its focus, studied population and biomarker selection are markedly different to ours.

Our forward MR analyses initially suggested ~~the existence of a causal effect of~~ renal function may affect ~~on the~~ protein plasma levels of seven proteins, although sensitivity analyses seem to indicate horizontal pleiotropy may be at play for six of them, given that their estimates were compatible with the null (i.e. no causal effect). A ~~robust~~ causal ~~effect~~ association ~~of~~ between renal function ~~on~~ and plasma levels of testican-2 was ~~observed~~inferred, ~~given~~ with ~~the~~ supporting evidence ~~offered by~~ from pleiotropy-robust sensitivity MR methods allowing for violations ~~to~~ of MR assumptions. Although consistent with the main analysis, the results from the restrictive MR were not significant perhaps due to reduced statistical power deriving from (a) fewer SNPs instrumenting the analysis, and (b) the exclusion of SNPs that might be on the actual causal pathway of interest.

Testican-2 is a secreted protein of the SPARC family[77], a group of matricellular proteins (MCPs) regulating extracellular matrix (ECM)-cell interactions and ECM processing[78], and is involved in a number of biological processes (Supplementary Table ~~20~~24). Given its glomerular filtration and detection in urine[79], changes in its plasma levels may reflect changes in kidney function ~~glomerular filtration alterations~~[74], given its renal release into the bloodstream [18]. Interestingly, higher testican-2 plasma levels have also been associated with less eGFR loss over time and reduced odds of incident CKD [18]. Its protein-coding gene, *SPOCK2,* is associated with both normal maintenance of organ and tissue integrity (glomerular remodeling), as well as with wound healing and other responses to injury[80]. Enriched in human glomeruli in comparison to tubuli samples[81, 82] and other non-renal tissues[83], *SPOCK2* has been reported as a glomerular and podocyte-specific gene[82, 84], where recent evidence from immunohistochemistry and immunofluorescence of human kidney tissue confirm its glomerular expression, and podocyte-specific expression in adult human kidney samples at single-cell resolution [18]. ~~The statistically insignificant correlation~~

between eGFR and *SPOCK2* expression in tubules in our study might be thus explained by the glomerular specificity of this gene. Moreover, *SPOCK2* has been reported as one of the top downregulated genes in glomeruli from patients with diabetic kidney disease[81], which further suggests its involvement in glomerular renal disease. The association between eGFR and *SPOCK2* renal expression did not reach the level of nominal statistical significance in some experiments and this may be explained at least partly by low statistical power (i.e. in the case of the dataset of 26 CKD patients) but the directionality of the association was consistent across datasets. Moreover, the multivariable analyses showed higher scores of histologic measures of renal structural damage to be negatively associated with *SPOCK2* renal expression, and positively associated with *CST6* renal expression - the direction of these associations was again consistent with that of the effects in the cross-sectional and MR analyses. IndeedAll in all, the agreement between the positive association cross-sectional results reported by us here and by others [18], as well as our and MR resultsfindings and the association with histologic measures, indicate testican-2 and its protein-coding gene *SPOCK2* may have an active role in renal healthsupport the notion. Low of low plasma levels of testican-2 may thus be as indicative of poor renal function, suggesting testican-2 this protein to be is a physiological biomarker of kidney health and disease progression, rather than only as a filtration marker[17, 18]. Although the reverse direction of this causal association could not be tested in our analyses was not significant in our MR analyses, a causal effect of testican-2 on renal function is also possiblebiologically plausible, given the role MCPs might play in the shift from constructive ECM repair to tissue stiffening and fibrosis[78, 85], and the recent evidence of its potential *in vitro* effects on human glomerular endothelial cells (HGECs) motility [18]. Nevertheless, whether the utility of testican-2 as a biomarker is related to its potential functional effects, and mechanisms influencing its blood levels, requires further study [18].

Three proteins (MIA, cystatin M and carbonic anhydrase III) were identified as potentially having a causal effect on eGFR, results which are biologically plausible given their known roles (Supplementary Table 2024). Nevertheless, the precise mechanism through which

these proteins could be exerting an effect on eGFR remains to be elucidated. Discordant directions of effect from the observational and the causal estimates (cystatin M, carbonic anhydrase III) could be explained due to differences in sample size and characteristics, reverse causation or confounding in the case of the observational estimates, or due to limitations innate to the MR methods[86, 87]. A further explanation might be that they are different effects: MR examines lifelong exposures to higher/lower protein levels, whereas results from observational studies could be reflective of acute effects[87]. The causality inferred by MR implies that if the exposure variables are lowered by means of an intervention mimicking the effect of the SNP on the exposure, the outcome variable will also change in the direction predicted by MR. That is, our results provide suggestive evidence that lifestyle or pharmacological interventions designed to improve renal function (with eGFR as a proxy thereof) increase plasma testican-2 levels. However, a limitation inherent to MR is that no information is offered on the time interval (e.g. during development) or target tissue in which such intervention would need to be delivered [88]. Likewise, current knowledge of the biological role of the proteins identified is insufficient for our findings to suggest mechanistic insights. This represents an opportunity for future research to unravel molecular mechanisms potentially underlying findings from MR and the biology of certain proteins in renal function.

The strengths of our study include the use of a multiplex proteomics platform and large sample size. This is the one of the first ~~study~~ studies to report eGFR-protein associations with replication in independent samples of diverse ancestries, adding to the robustness and generalizability of our results. Ours is ~~also~~ the first study to examine causality in proteomic associations with eGFR. The MR was conducted with the largest available GWAS results from non-overlapping European ancestry populations, thus avoiding issues derived from population stratification and sample overlap. We additionally used GWA summary statistics from a complementary renal trait (BUN) to improve the specificity of the SNPs used as genetic instruments for eGFR. We focused on cis-SNPs in the reverse MR, thus reducing the possibility of horizontal pleiotropy. Finally, the conclusions presented here are supported by the multiple sensitivity analyses conducted to test the robustness of our methods.

Our study also has several limitations. Aptamer-based proteomic methods may be affected by probe cross-reactivity and non-specific binding [89~~33~~] – yet ~~the~~ we were able to validate the aptamer-based data from 12 proteins were measured using an independent analytical method in a subset of the discovery sample. Moreover, ~~aptamer-based~~ measurements of 50 of our 57 proteins have been validated by mass spectrometry in blood plasma/serum/other biological matrices and solution affinity measurements in multiple independent studies[27, 38, 39]. This platform does not produce absolute concentrations of plasma proteins, thus limiting the interpretability of the regression coefficients. Future validation studies developing absolute quantitative assays for the detection of testican-2, as well as other proteins identified here, are warranted in order to establish reference ranges and to explore their suitability as prognostic and diagnostic biomarkers in clinical settings. Likewise, proteins undergoing post-transcriptional and post-translational modifications are not covered in this assay (although investigation of the modification of specific reagents for this purpose is underway [89]), so potentially relevant proteins undergoing such molecular modifications may have been overlooked in this study. Our findings are based on cross-sectional data, so future studies examining longitudinal changes in these proteins would be of interest. We examined linear associations with plasma proteomic levels, although non-linear trajectories might also exist for some proteins. Likewise, the age interaction effects observed warrant further investigation in future longitudinal studies. We avoided weak instrument bias in MR by selecting genome-wide significant variants, but cannot discount the possibility of having incurred selection bias in the case of the SNP-protein data. The sample size in which genetic associations with protein levels were calculated was significantly smaller than the sample used to identify genetic associations with eGFR, which likely resulted in differences in power. Moreover, despite the multi-ethnic nature of our study, Asian and African ancestries were not represented, and the generalizability of our results may not extend to these populations. Finally, a follow-up of the findings of our study in appropriate experimental models would provide additional evidence on the inferred causal associations reported here and help to unravel the molecular mechanisms underlying our findings.

In summary, we present the largest study to date examining the plasma proteomics of renal function. We use a trans-ethnic discovery-replication setting, and are the first to assess the causality underlying our associations. We identified multiple well-known markers of kidney function, and discovered two proteins not previously known to be associated with eGFR. Our findings with regard to causal relationships represent an important stepping-stone in the establishment of testican-2 as a clinically relevant physiological marker of kidney disease progression, and provide suggestive evidence for a number of proteins causally affecting eGFR warranting further investigation as potential therapeutic targets. Our results may serve as the starting point for future work on the translational role of eGFR-associated proteins as diagnostic or prognostic biomarkers of disease, potentially druggable targets, as well as for research on mechanistic insights at the tissue and single cell levels.

Author contributions

Conceived and designed the study: P.R.M.G., R.W., J.W., C.G., M.W.

Performed experiments: D.O.M.K., J.G., Ag.P.

Analyzed data or interpreted results: P.R.M.G., R.W., Q.G., S.B.Z, J.M.E., Xi.X., F.J.C., J.G., M.P., M.T., M.W.

Contributed reagents/materials/analysis tools: Q.G., S.B.Z., J.M.E., Xi.X., F.J.C., J.Do., P.S., M.A.E., S.S., A.P., A.F., D.O.M.K., J.W., J.Da., E.D.A., W.-H.-O., N.-A.-W., D.-J.-R., Ag.P., J.G., W.K., K.H., C.J., A.K., A.B., M.P., S.M.H., K.S., C.G., M.T., H.K.T.S.**, A.T., M.W.

Wrote the paper: P.R.M.G., ~~M.W~~.

Edited and improved clarity of the manuscript: R.W., Q.G., S.B.Z., J.M.E., Xi.X., F.J.C., J.Do., P.S., J.N., Ag.P., J.G., W.K., A.K.,S.M.H., M.T., A.T., M.W.

All authors discussed the results, and reviewed and approved the final manuscript.

Acknowledgements

We are grateful to all study participants of KORA, HUNT, INTERVAL and QMDiab for their invaluable contributions to these studies, as well as all members of field staff conducting the studies. An early version of the abstract was sent to the 53rd European Society of Human

Genetics (ESHG) Conference, and is listed online as P03.24.C at the ESHG 2020.2 Abstract Library (https://2020.eshg.org/index.php/abstract-library/)

Ethics approval and informed consent

All participants provided written consent, and details on institutions providing ethics approval from each study are described next:

KORA: The KORA cohort ethical approval was granted by the ethics committee of the Bavarian Medical Association (REC reference number F4: #06068)

HUNT: Regiona Committee for Medical Research Ethics (REK) and Data Inspectorate

INTERVAL: National Research Ethics Service Committee East of England - Cambridge East (Research Ethics Committee (REC) reference 11/EE/0538)

QMDiab: Institutional Review Boards of HMC and WCM-Q under research protocol number 11131/11)

Disclosures

Competing interests

The authors declare the following competing interests: C.J. has received personal fees for research consultancy work from Pfizer and Bayer outside of the submitted work, and J.D. sits on the International Cardiovascular and Metabolic Advisory Board for Novartis (since 2010); the Steering Committee of UK Biobank (since 2011); the MRC International Advisory Group (ING) member, London (since 2013); the MRC High Throughput Science 'Omics Panel Member, London (since 2013); the Scientific Advisory Committee for Sanofi (since 2013); the International Cardiovascular and Metabolism Research and Development Portfolio Committee for Novartis; and the Astra Zeneca Genomics Advisory Board (2018). Other authors declare no competing interests.

Funding

The KORA study was initiated and financed by the Helmholtz Zentrum München – German Research Center for Environmental Health, which is funded by the German Federal Ministry of Education and Research (BMBF) and by the State of Bavaria. This work was also supported by the Biomedical Research Program at Weill Cornell Medicine in Qatar, a

program funded by the Qatar Foundation. K.S. is supported by Qatar National Research Fund (QNRF) grant no. NPRPC11-0115-180010. The Nord-Trøndelag Health Study (The HUNT Study) is a collaboration between HUNT Research Centre (Faculty of Medicine, Norwegian University of Science and Technology NTNU), Nord-Trøndelag County Council, Central Norway Health Authority, and the Norwegian Institute of Public Health. The HUNT part of the project re-used protein data that was originally analysed and paid for by Somalogic Inc, CO, USA. Somalogic had no role in the design and conduct of the study; collection of phenotypic data, statistical analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication. Professor John Danesh is funded by the National Institute for Health Research [Senior Investigator Award]. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

RNA-sequencing experiments and kidney gene expression studies were supported by British Heart Foundation project grants [PG/17/35/33001 and PG/19/16/34270] and Kidney Research UK grants [RP_017_20180302 and RP_013_20190305] to M.T.

Supplemental material - Table of Contents

Supplementary File 1. Supplemental Notes, Figure and Table legends (.docx)

Supplementary File 2. Supplementary Tables (.xlsx)

Supplementary File 3. Supplementary Figures (.pdf)

Supplementary File 4. STROBE Checklist (.pdf)

References

1. Jiang, X, Eales, JM, Scannali, D, Nazgiewicz, A, Prestes, P, Maier, M, et al.: Hypertension and renin-angiotensin system blockers are not associated with expression of angiotensin-converting enzyme 2 (ACE2) in the kidney. *European Heart Journal*, 2020.

2. Eckardt, KU, Coresh, J, Devuyst, O, Johnson, RJ, Kottgen, A, Levey, AS, et al.: Evolving importance of kidney disease: from subspecialty to global health burden. *Lancet (London, England),* 382**:** 158-169, 2013.

3. Levey, AS, Inker, LA, Coresh, J: GFR estimation: from physiology to public health. *American journal of kidney diseases : the official journal of the National Kidney Foundation,* 63**:** 820-834, 2014.

4. Levey, AS, Stevens, LA, Schmid, CH, Zhang, YL, Castro, AF, 3rd, Feldman, HI, et al.: A new equation to estimate glomerular filtration rate. *Annals of internal medicine,* 150**:** 604-612, 2009.

5. Hill, NR, Fatoba, ST, Oke, JL, Hirst, JA, O'Callaghan, CA, Lasserson, DS, et al.: Global Prevalence of Chronic Kidney Disease - A Systematic Review and Meta-Analysis. *PloS one,* 11**:** e0158765, 2016.

6. Levey, AS, Atkins, R, Coresh, J, Cohen, EP, Collins, AJ, Eckardt, KU, et al.: Chronic kidney disease as a global public health problem: approaches and initiatives - a position statement from Kidney Disease Improving Global Outcomes. *Kidney Int,* 72**:** 247-259, 2007.

7. Mischak, H, Delles, C, Vlahou, A, Vanholder, R: Proteomic biomarkers in kidney disease: issues in development and implementation. *Nature reviews Nephrology,* 11**:** 221-232, 2015.

8. Sanchez-Nino, MD, Sanz, AB, Ramos, AM, Fernandez-Fernandez, B, Ortiz, A: Clinical proteomics in kidney disease as an exponential technology: heading towards the disruptive phase. *Clinical kidney journal,* 10**:** 188-191, 2017.

9. Good, DM, Zurbig, P, Argiles, A, Bauer, HW, Behrens, G, Coon, JJ, et al.: Naturally occurring human urinary peptides for use in diagnosis of chronic kidney disease. *Molecular & cellular proteomics : MCP,* 9**:** 2424-2437, 2010.

10. Gold, L, Ayers, D, Bertino, J, Bock, C, Bock, A, Brody, EN, et al.: Aptamer-based multiplexed proteomic technology for biomarker discovery. *PloS one,* 5**:** e15004, 2010.

11. Ngo, D, Sinha, S, Shen, D, Kuhn, EW, Keyes, MJ, Shi, X, et al.: Aptamer-Based Proteomic Profiling Reveals Novel Candidate Biomarkers and Pathways in Cardiovascular Disease. *Circulation,* 134**:** 270-285, 2016.

12. Ganz, P, Heidecker, B, Hveem, K, Jonasson, C, Kato, S, Segal, MR, et al.: Development and Validation of a Protein-Based Risk Score for Cardiovascular Outcomes Among Patients With Stable Coronary Heart Disease. *Jama,* 315**:** 2532-2541, 2016.

13. Inker, LA, Schmid, CH, Tighiouart, H, Eckfeldt, JH, Feldman, HI, Greene, T, et al.: Estimating glomerular filtration rate from serum creatinine and cystatin C. *The New England journal of medicine,* 367**:** 20-29, 2012.

14. Pena, MJ, Mischak, H, Heerspink, HJ: Proteomics for prediction of disease progression and response to therapy in diabetic kidney disease. *Diabetologia,* 59**:** 1819-1831, 2016.

15. Gajjala, PR, Jankowski, V, Heinze, G, Bilo, G, Zanchetti, A, Noels, H, et al.: Proteomic-Biostatistic Integrated Approach for Finding the Underlying Molecular Determinants of Hypertension in Human Plasma. *Hypertension (Dallas, Tex : 1979),* 70**:** 412-419, 2017.

16. Carlsson, AC, Ingelsson, E, Sundstrom, J, Carrero, JJ, Gustafsson, S, Feldreich, T, et al.: Use of Proteomics To Investigate Kidney Function Decline over 5 Years. *Clinical journal of the American Society of Nephrology : CJASN,* 12**:** 1226-1235, 2017.

17. Christensson, A, Ash, JA, DeLisle, RK, Gaspar, FW, Ostroff, R, Grubb, A, et al.: The Impact of the Glomerular Filtration Rate on the Human Plasma Proteome. *Proteomics Clinical applications*, 2017.

18. Ngo, D, Wen, D, Gao, Y, Keyes, MJ, Drury, ER, Katz, DH, et al.: Circulating testican-2 is a podocyte-derived marker of kidney health. *Proceedings of the National Academy of Sciences of the United States of America,* 117**:** 25026-25035, 2020.

19. Pierce, BL, Burgess, S: Efficient design for Mendelian randomization studies: subsample and 2-sample instrumental variable estimators. *American journal of epidemiology,* 178**:** 1177-1184, 2013.

20. Sekula, P, Del Greco M, F, Pattaro, C, Köttgen, A: Mendelian Randomization as an Approach to Assess Causality Using Observational Data. *Journal of the American Society of Nephrology : JASN,* 27**:** 3253-3265, 2016.

21. Holle, R, Happich, M, Lowel, H, Wichmann, HE: KORA--a research platform for population based health research. *Gesundheitswesen (Bundesverband der Arzte des Offentlichen Gesundheitsdienstes (Germany)),* 67 Suppl 1**:** S19-25, 2005.

22. Wichmann, HE, Gieger, C, Illig, T: KORA-gen--resource for population genetics, controls and a broad spectrum of disease phenotypes. *Gesundheitswesen (Bundesverband der Arzte des Offentlichen Gesundheitsdienstes (Germany)),* 67 Suppl 1**:** S26-30, 2005.

23. Krokstad, S, Langhammer, A, Hveem, K, Holmen, T, Midthjell, K, Stene, T, et al.: Cohort Profile: The HUNT Study, Norway. *International journal of epidemiology,* 42**:** 968-977, 2012.

24. Moore, C, Bolton, T, Walker, M, Kaptoge, S, Allen, D, Daynes, M, et al.: Recruitment and representativeness of blood donors in the INTERVAL randomised trial assessing varying inter-donation intervals. *Trials,* 17**:** 458, 2016.

25. Mook-Kanamori, DO, Selim, MM, Takiddin, AH, Al-Homsi, H, Al-Mahmoud, KA, Al-Obaidli, A, et al.: 1,5-Anhydroglucitol in saliva is a noninvasive marker of short-term glycemic control. *The Journal of clinical endocrinology and metabolism,* 99**:** E479-483, 2014.

26. Suhre, K, Arnold, M: Connecting genetic risk to disease end points through the human blood plasma proteome. 8**:** 14357, 2017.

27. Sun, BB, Maranville, JC, Peters, JE, Stacey, D, Staley, JR, Blackshaw, J, et al.: Genomic atlas of the human plasma proteome. *Nature,* 558**:** 73-79, 2018.

28. Nayor, M, Short, MI, Rasheed, H, Lin, H, Jonasson, C, Yang, Q, et al.: Aptamer-Based Proteomic Platform Identifies Novel Protein Predictors of Incident Heart Failure and Echocardiographic Traits. *Circulation Heart failure,* 13**:** e006749, 2020.

29. Gold, L, Walker, JJ, Wilcox, SK, Williams, S: Advances in human proteomics at high scale with the SOMAscan proteomics platform. *New biotechnology,* 29**:** 543-549, 2012.

30. Rohloff, JC, Gelinas, AD, Jarvis, TC, Ochsner, UA, Schneider, DJ, Gold, L, et al.: Nucleic Acid Ligands With Protein-like Side Chains: Modified Aptamers and Their Use as Diagnostic and Therapeutic Agents. *Molecular therapy Nucleic acids,* 3**:** e201, 2014.

31. Kraemer, S, Vaught, JD, Bock, C, Gold, L, Katilius, E, Keeney, TR, et al.: From SOMAmer-based biomarker discovery to diagnostic and clinical applications: a SOMAmer-based, streamlined multiplex proteomic assay. *PloS one,* 6**:** e26332-e26332, 2011.

32. Pattaro, C, Riegler, P, Stifter, G, Modenese, M, Minelli, C, Pramstaller, PP: Estimating the glomerular filtration rate in the general population using different equations: effects on classification and association. *Nephron Clinical practice,* 123**:** 102-111, 2013.

33. Coresh, J, Turin, TC, Matsushita, K, Sang, Y, Ballew, SH, Appel, LJ, et al.: Decline in estimated glomerular filtration rate and subsequent risk of end-stage renal disease and mortality. *Jama,* 311**:** 2518-2531, 2014.

34. K/DOQI clinical practice guidelines for chronic kidney disease: evaluation, classification, and stratification. *American journal of kidney diseases : the official journal of the National Kidney Foundation,* 39**:** S1-266, 2002.

35. R Core Team: R: A language and environment for statistical computing. Vienna, Austria, R Foundation for Statistical Computing, 2019.

36. Zaykin, DV: Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis. *Journal of evolutionary biology,* 24**:** 1836-1841, 2011.

37. Huth, C, von Toerne, C, Schederecker, F, de Las Heras Gala, T, Herder, C, Kronenberg, F, et al.: Protein markers and risk of type 2 diabetes and prediabetes: a targeted proteomics approach in the KORA F4/FF4 study. *European journal of epidemiology,* 34**:** 409-422, 2019.

38. Williams, SA, Kivimaki, M, Langenberg, C, Hingorani, AD, Casas, JP, Bouchard, C, et al.: Plasma protein patterns as comprehensive indicators of health. *Nature medicine,* 25**:** 1851-1857, 2019.

39. Emilsson, V, Ilkov, M, Lamb, JR, Finkel, N, Gudmundsson, EF, Pitts, R, et al.: Co-regulatory networks of human serum proteins link genetics to disease. *Science (New York, NY),* 361**:** 769-773, 2018.

40. Kyritsis, KA, Wang, B, Sullivan, J, Lyne, R, Micklem, G: InterMineR: an R package for InterMine databases. *Bioinformatics*, 2019.

41. Huang, DW, Sherman, BT, Lempicki, RA: Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols,* 4**:** 44-57, 2009.

42. Durinck, S, Spellman, PT, Birney, E, Huber, W: Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc,* 4**:** 1184-1191, 2009.

43. Schmidt, T, Samaras, P, Frejno, M, Gessulat, S, Barnert, M, Kienegger, H, et al.: ProteomicsDB. *Nucleic acids research,* 46**:** D1271-D1281, 2017.

44. Samaras, P, Schmidt, T, Frejno, M, Gessulat, S, Reinecke, M, Jarzab, A, et al.: ProteomicsDB: a multi-omics and multi-organism resource for life science research. *Nucleic acids research,* 48**:** D1153-D1163, 2019.

45. Consortium, GT: The Genotype-Tissue Expression (GTEx) project. *Nature genetics,* 45**:** 580-585, 2013.

46. Szklarczyk, D, Gable, AL, Lyon, D, Junge, A, Wyder, S, Huerta-Cepas, J, et al.: STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research,* 47**:** D607-d613, 2019.

47. Davey Smith, G, Hemani, G: Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Human molecular genetics,* 23**:** R89-98, 2014.

48. Smith, GD, Ebrahim, S: 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International journal of epidemiology,* 32**:** 1-22, 2003.

49. Lawlor, DA, Harbord, RM, Sterne, JA, Timpson, N, Davey Smith, G: Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in medicine,* 27**:** 1133-1163, 2008.

50. Wuttke, M, Li, Y, Li, M, Sieber, KB, Feitosa, MF, Gorski, M, et al.: A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nature genetics,* 51**:** 957-972, 2019.

51. Hemani, G, Zheng, J, Elsworth, B, Wade, KH, Haberland, V, Baird, D, et al.: The MR-Base platform supports systematic causal inference across the human phenome. *eLife,* 7, 2018.

52. Buniello, A, MacArthur, JAL, Cerezo, M, Harris, LW, Hayhurst, J, Malangone, C, et al.: The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research,* 47**:** D1005-d1012, 2019.

53. Burgess, S, Butterworth, A, Thompson, SG: Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic epidemiology,* 37**:** 658-665, 2013.

54. Teumer, A: Common Methods for Performing Mendelian Randomization. *Frontiers in cardiovascular medicine,* 5**:** 51, 2018.

55. Burgess, S, Bowden, J, Fall, T, Ingelsson, E, Thompson, SG: Sensitivity Analyses for Robust Causal Inference from Mendelian Randomization Analyses with Multiple Genetic Variants. *Epidemiology,* 28**:** 30-42, 2017.

56. Bowden, J, Davey Smith, G, Burgess, S: Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International journal of epidemiology,* 44**:** 512-525, 2015.

57. Bowden, J, Davey Smith, G, Haycock, PC, Burgess, S: Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genetic epidemiology,* 40**:** 304-314, 2016.

58. Hartwig, FP, Davey Smith, G, Bowden, J: Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *International journal of epidemiology,* 46**:** 1985-1998, 2017.

59. Verbanck, M, Chen, C-Y, Neale, B, Do, R: Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nature genetics,* 50**:** 693-698, 2018.

60. Ju, W, Nair, V, Smith, S, Zhu, L, Shedden, K, Song, PXK, et al.: Tissue transcriptome-driven identification of epidermal growth factor as a chronic kidney disease biomarker. *Science translational medicine,* 7**:** 316ra193, 2015.

61. Martini, S, Eichinger, F, Nair, V, Kretzler, M: Defining human diabetic nephropathy on the molecular level: Integration of transcriptomic profiles with biological knowledge. *Reviews in Endocrine and Metabolic Disorders,* 9**:** 267-274, 2008.

62. Sampson, MG, Robertson, CC, Martini, S, Mariani, LH, Lemley, KV, Gillies, CE, et al.: Integrative Genomics Identifies Novel Associations with APOL1 Risk Genotypes in

Black NEPTUNE Subjects. *Journal of the American Society of Nephrology : JASN,* 27**:** 814-823, 2016.

63. Reich, HN, Tritchler, D, Cattran, DC, Herzenberg, AM, Eichinger, F, Boucherot, A, et al.: A molecular signature of proteinuria in glomerulonephritis. *PloS one,* 5**:** e13451, 2010.

64. Rodwell, GE, Sonu, R, Zahn, JM, Lund, J, Wilhelmy, J, Wang, L, et al.: A transcriptional profile of aging in the human kidney. *PLoS biology,* 2**:** e427, 2004.

65. Balduzzi, S, Rücker, G, Schwarzer, G: How to perform a meta-analysis with R: a practical tutorial. *Evidence Based Mental Health,* 22**:** 153-160, 2019.

66. Xu, X, Eales, JM, Akbarov, A, Guo, H, Becker, L, Talavera, D, et al.: Molecular insights into genome-wide association studies of chronic kidney disease-defining traits. *Nature communications,* 9**:** 4800, 2018.

67. Rowland, J, Akbarov, A, Eales, J, Xu, X, Dormer, JP, Guo, H, et al.: Uncovering genetic mechanisms of kidney aging through transcriptomics, genomics, and epigenomics. *Kidney Int,* 95**:** 624-635, 2019.

68. Richmond, RC, Davey Smith, G: Commentary: Orienting causal relationships between two phenotypes using bidirectional Mendelian randomization. *International journal of epidemiology,* 48**:** 907-911, 2019.

69. Lehallier, B, Gate, D, Schaum, N, Nanasi, T, Lee, SE, Yousef, H, et al.: Undulating changes in human plasma proteome profiles across the lifespan. *Nature medicine,* 25**:** 1843-1850, 2019.

70. Jovanovic, D, Krstivojevic, P, Obradovic, I, Durdevic, V, Dukanovic, L: Serum cystatin C and beta2-microglobulin as markers of glomerular filtration rate. *Renal failure,* 25**:** 123-133, 2003.

71. Niewczas, MA, Pavkov, ME, Skupien, J, Smiles, A, Md Dom, ZI, Wilson, JM, et al.: A signature of circulating inflammatory proteins and development of end-stage renal disease in diabetes. *Nature medicine,* 25**:** 805-813, 2019.

72. Prunotto, M, Farina, A, Lane, L, Pernin, A, Schifferli, J, Hochstrasser, DF, et al.: Proteomic analysis of podocyte exosome-enriched fraction from normal human urine. *Journal of proteomics,* 82**:** 193-229, 2013.

73. Jia, L, Zhang, L, Shao, C, Song, E, Sun, W, Li, M, et al.: An attempt to understand kidney's protein handling function by comparing plasma and urine proteomes. *PloS one,* 4**:** e5146, 2009.

74. Schenk, S, Schoenhals, GJ, de Souza, G, Mann, M: A high confidence, manually validated human blood plasma protein reference set. *BMC medical genomics,* 1**:** 41, 2008.

75. Carlsson, AC, Sundström, J, Carrero, JJ, Gustafsson, S, Stenemo, M, Larsson, A, et al.: Use of a proximity extension assay proteomics chip to discover new biomarkers associated with albuminuria. *European journal of preventive cardiology,* 24**:** 340-348, 2016.

76. Mohammadi-Shemirani, P, Sjaarda, J, Gerstein, HC, Treleaven, DJ, Walsh, M, Mann, JF, et al.: A Mendelian Randomization-Based Approach to Identify Early and Sensitive Diagnostic Biomarkers of Disease. *Clinical chemistry,* 65**:** 427-436, 2019.

77. Clark, CJ, Sage, EH: A prototypic matricellular protein in the tumor microenvironment--where there's SPARC, there's fire. *Journal of cellular biochemistry,* 104**:** 721-732, 2008.

78. Feng, D, Ngov, C, Henley, N, Boufaied, N, Gerarduzzi, C: Characterization of Matricellular Protein Expression Signatures in Mechanistically Diverse Mouse Models of Kidney Injury. *Scientific reports,* 9**:** 16736, 2019.

79. Marimuthu, A, O'Meally, RN, Chaerkady, R, Subbannayya, Y, Nanjappa, V, Kumar, P, et al.: A comprehensive map of the human urinary proteome. *Journal of proteome research,* 10**:** 2734-2743, 2011.

80. Francki, A, Sage, EH: SPARC and the kidney glomerulus: matricellular proteins exhibit diverse functions under normal and pathological conditions. *Trends in cardiovascular medicine,* 11**:** 32-37, 2001.

81. Woroniecka, KI, Park, AS, Mohtat, D, Thomas, DB, Pullman, JM, Susztak, K: Transcriptome analysis of human diabetic kidney disease. *Diabetes,* 60**:** 2354-2369, 2011.

82. Lindenmeyer, MT, Eichinger, F, Sen, K, Anders, HJ, Edenhofer, I, Mattinzoli, D, et al.: Systematic analysis of a novel human renal glomerulus-enriched gene expression dataset. *PloS one,* 5**:** e11545, 2010.

83. Nystrom, J, Fierlbeck, W, Granqvist, A, Kulak, SC, Ballermann, BJ: A human glomerular SAGE transcriptome database. *BMC nephrology,* 10**:** 13, 2009.

84. Ju, W, Greene, CS, Eichinger, F, Nair, V, Hodgin, JB, Bitzer, M, et al.: Defining cell-type specificity at the transcriptional level in human disease. *Genome research,* 23**:** 1862-1873, 2013.

85. Wynn, TA, Ramalingam, TR: Mechanisms of fibrosis: therapeutic translation for fibrotic disease. *Nature medicine,* 18**:** 1028-1040, 2012.

86. Haycock, PC, Burgess, S, Wade, KH, Bowden, J, Relton, C, Davey Smith, G: Best (but oft-forgotten) practices: the design, analysis, and interpretation of Mendelian randomization studies. *The American journal of clinical nutrition,* 103**:** 965-978, 2016.

87. Davies, NM, Holmes, MV, Davey Smith, G: Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *BMJ (Clinical research ed),* 362**:** k601, 2018.

88. Bretherick, AD, Canela-Xandri, O, Joshi, PK, Clark, DW, Rawlik, K, Boutin, TS, et al.: Linking protein to phenotype with Mendelian Randomization detects 38 proteins with causal roles in human diseases and traits. *PLoS genetics,* 16**:** e1008785, 2020.

89. Suhre, K, McCarthy, MI, Schwenk, JM: Genetics meets proteomics: perspectives for large population-based studies. *Nature reviews Genetics*, 2020.

Tables

Table 1. Population characteristics of association studies

Trait	KORA	HUNT3	INTERVAL	QMDIAB
N	995	930	623	334
Age (years)	59.31 (7.81)	68.94 (10.29)	47.36 (13.35)	47.10 (12.57)
Male	480 (48.2)	688 (74.0)	343 (55.1)	169 (50.6)
BMI (kg/m2)	27.78 (4.58)	28.38 (3.97)	27.16 (10.05)	29.66 (5.95)
Smoking	572 (57.5)	699 (75.16)	99 (15.89)	60 (18.0)
Serum creatinine (mg/dL)	0.85 (0.18)	0.92 (0.32)	0.70 (0.14)	0.85 (0.22)
eGFR (mL/min/1.73m2)	85.98 (14.06)	80.25 (18.75)	108.27 (16.21)	95.87 (27.32)
CKD	38 (3.8)	138 (14.8)	1 (0.16%)	30 (9.0)
UACR (mg/dL) *	5.64 (3.61, 9.94)	NA	NA	NA
MA	58 (5.9)	NA	NA	NA
HDL cholesterol (mg/dL)	57.32 (15.20)	45.12 (11.24)	74.33 (24.42)	47.58 (13.75)
Triglycerides (mg/dL) *	107 (75, 155.5)	141.27 (106.28, 194.85)	132.75 (97.35,194.70)	169 (99.20, 215.23)
Anti-hyperlipidemic medication use	142 (14.3)	NA	33 (5.30)	NA
Hypertension	397 (39.9)	355 (38.2)	48 (7.70)	103 (30.8)
Diabetes	68 (6.8)	128 (13.8)	2 (0.32)	172 (51.5)

Table legend: Measurement units are shown in parentheses in the trait column, where the absence of units means it is a categorical trait. The mean and (SD) are presented for non-skewed continuous variables, while skewed continuous variables are identified with * and median (1st, 3rd quartile) are presented. Count and (%) are shown for categorical variables. HUNT3: third survey of Nord-Trøndelag Health Study (HUNT3), INTERVAL: INTERVAL Study, QMDIAB: Qatar Metabolomics Study on Diabetes; NA: not available; BMI: body-mass index, eGFR: serum creatinine-based estimated glomerular filtration rate, CKD: chronic kidney disease, UACR: urinary albumin-to-creatinine ratio, MA: microalbuminuria, HDL: high density lipoprotein.

Table 2. Causal estimates across Mendelian randomization methods

Association		IVW	MR-Egger	Weighted median	Weighted mode	Wald ratio
eGFR → *SPOCK2*	β	5.213	8.61	6.104	6.506	--
	CI	[2.767, 7.659]	[2.405, 14.815]	[2.808, 9.401]	[1.189, 11.823]	
	p	2.95E-05	0.01	2.84E-04	0.021	
	N_{IV}	40	40	40	40	
CA3 → eGFR	β	--	--	--	--	0.007
	CI					[0.003, 0.010]
	p					5.04E-04
	N_{IV}					1
CST6 → eGFR	β	--	--	--	--	0.007
	CI					[0.004, 0.011]
	p					8.41E-05
	N_{IV}					1
MIA → eGFR	β	-0.002	-0.001	-0.002	-0.001	--
	CI	[-0.003, -0.001]	[-0.002, 0.000]	[-0.003, -0.001]	[-0.002, -0.001]	
	p	8.79E-04	0.299	2.00E-04	0.081	
	N_{IV}	3	3	3	3	

Table legend: Results from the forward MR (effect of eGFR on protein levels, i.e. eGFR → protein) are based on the 40 instruments retrieved from Wuttke, et.al. 2019[50], whereas the reverse MR (protein → eGFR) are based on the one to three instruments retrieved from the INTERVAL pGWAS reported in Sun, et.al. 2018[58]. In bold are significant p values at a Bonferroni-corrected level (0.05/47 for the forward analysis, 0.05/28 for the reverse analysis). *SPOCK2*: testican-2, *CA3*: carbonic anhydrase III, *CST6*: cystatin-M, *MIA*: melanoma-derived growth regulatory protein, IVW: Inverse-variance weighted MR; ~~MR-PRESSO: Mendelian randomization Pleiotropy RESidual Sum and Outlier;~~ β: causal estimate, SE: Standard Error, CI: 95% confidence interval of causal estimate; N_{IV}: number of SNPs used as instrumental variables.

Figures

Figure 1. Study overview

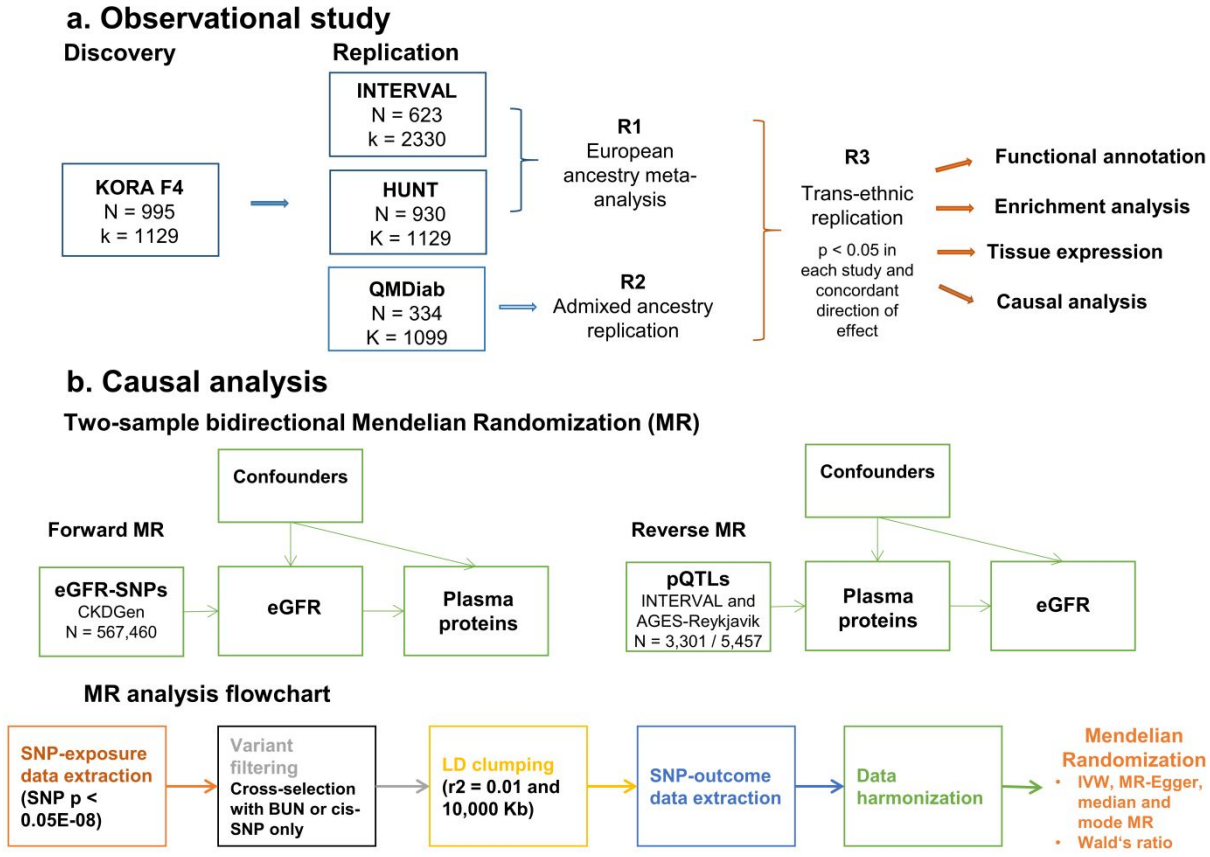


Figure 2. Cross-sectional effect estimates from the top 10 proteins associated with eGFR

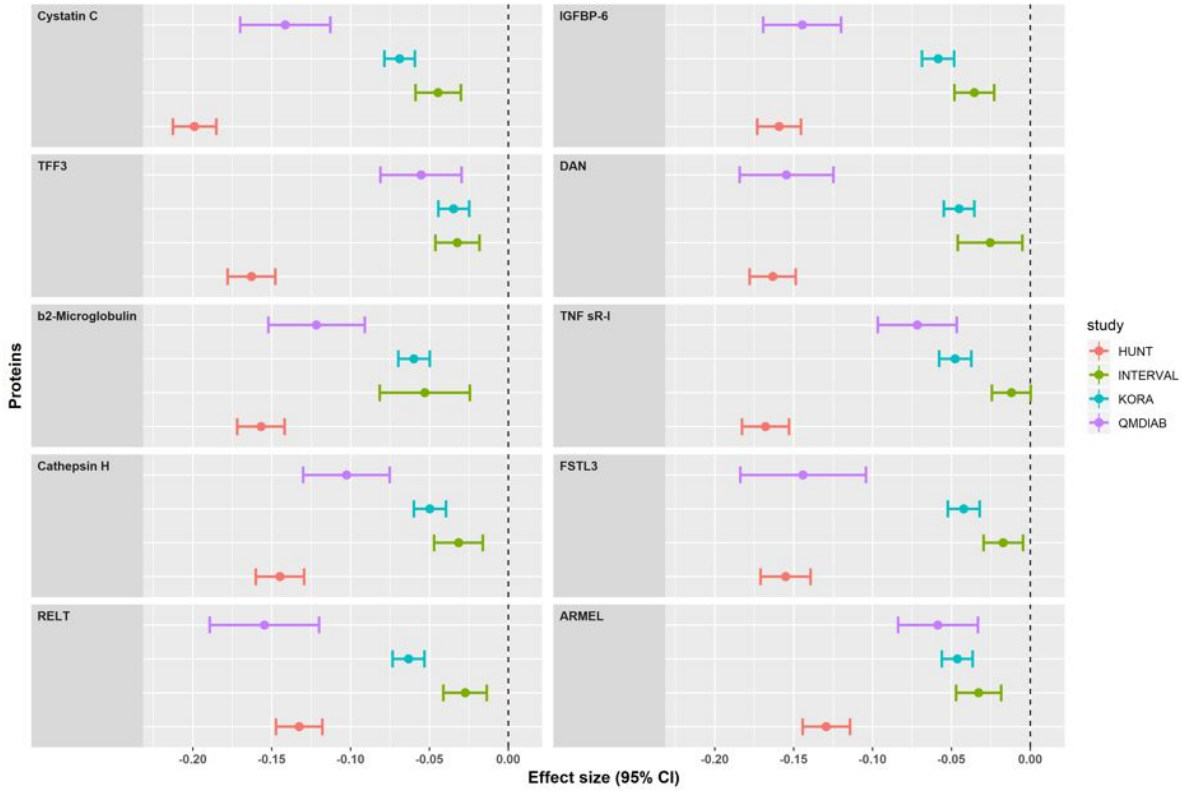


Figure 3. Overlap of associations with eGFR and CKD

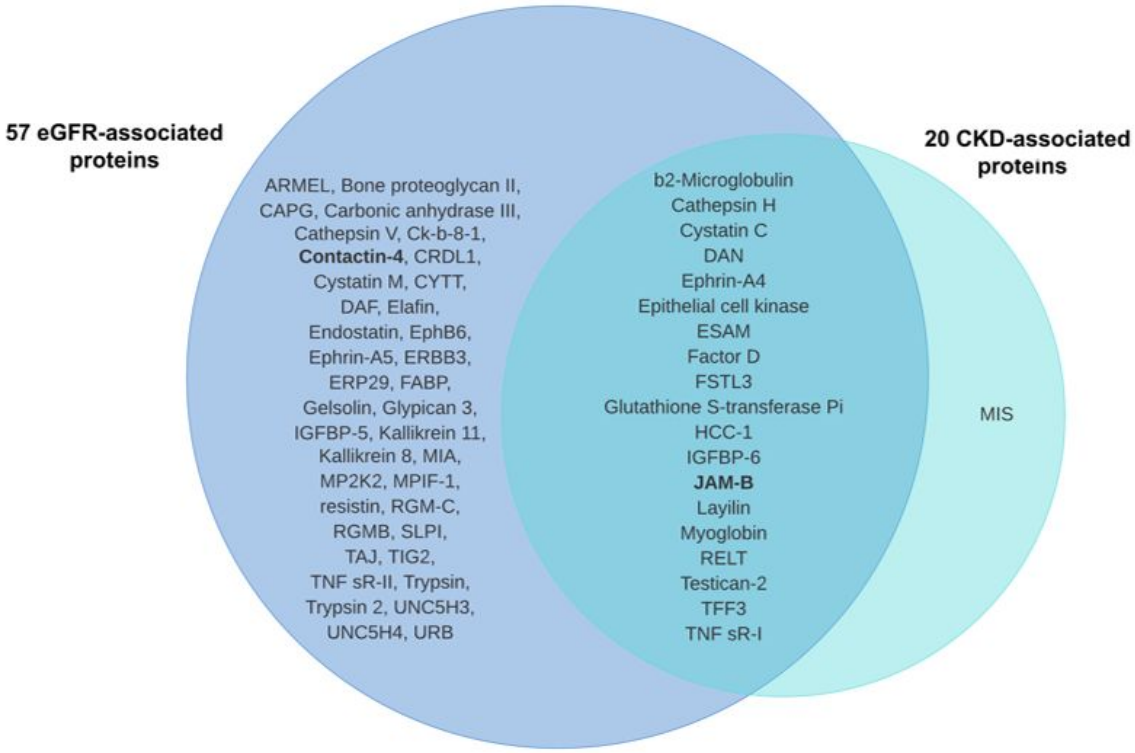


Figure 4. Results from the forward MR estimating the causal effect of eGFR on testican-2

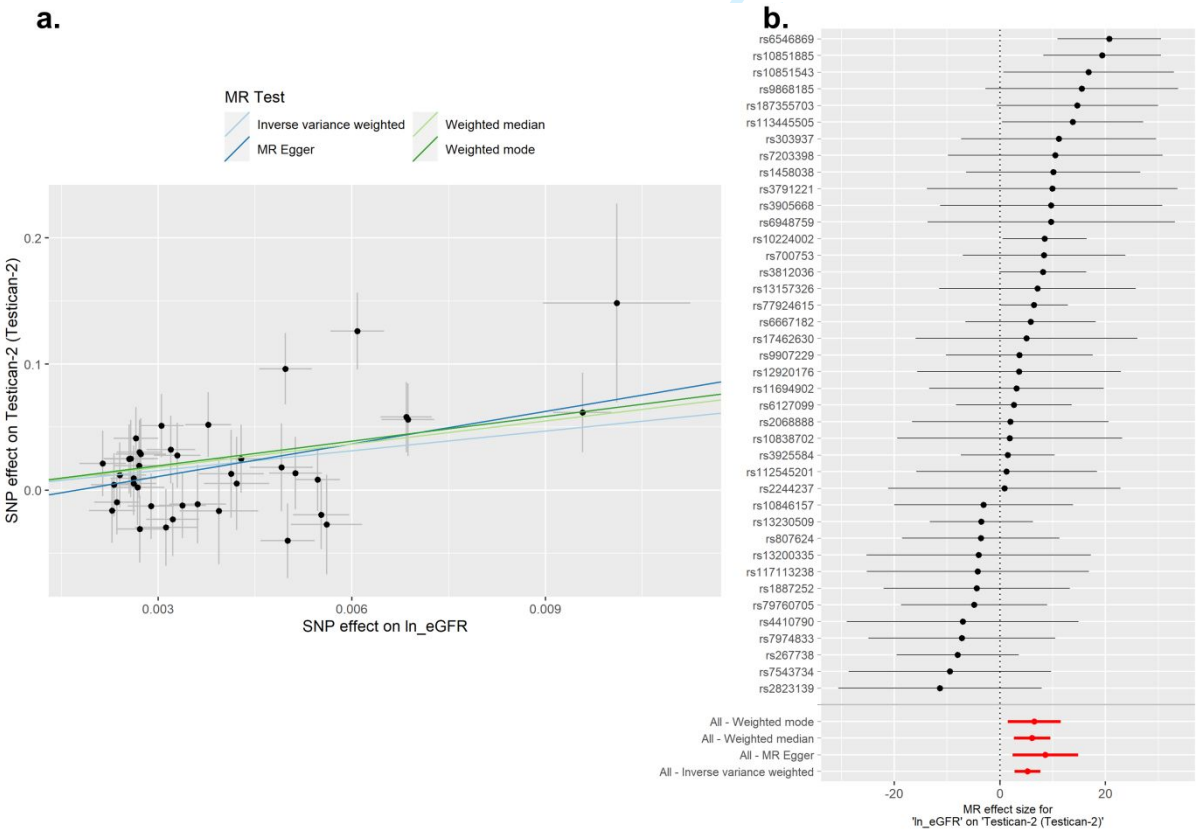


Figure legends

Fig. 1. Panel a. Cross-sectional association study. Data from 995 participants and 1129 proteins from KORA F4 was used in the discovery phase of a proteome wide association study of renal function using confounder-adjusted regression models. The replication studies were INTERVAL, HUNT and QMDiab. Three rounds of replication are shown: R1, replication based on the meta-analysis of p-values from the linear regression results of the studies with European ancestry; R2, replication based on the results of linear regression models performed in the Arab, South Asian and Filipino descent sample QMDiab; R3, identification of proteins consistently associated with eGFR across samples and ethnicities. The set of proteins identified in R3 was then functionally annotated and brought forward to the causal analysis phase. **Panel b. Causal analysis.** Two-sample bidirectional Mendelian randomization using data on participants from EA-studies in the CKDGen Consortium to instrument the forward analysis (eGFR causal to protein level) and data from INTERVAL and AGES-Reykjavik to instrument the reverse analysis (protein level causal to eGFR).Details on the data processing workflow for Mendelian randomizations are shown.

Fig. 2. Regression coefficient estimates from the top 10 proteins identified in the cross-sectional association trans-ethnic study on eGFR. The x-axis shows the estimates and 95% CI for the regression coefficients (i.e. change in log-transformed eGR per standard deviation increase in protein level), and each panel corresponds to one protein. Estimates are color coded according to the specific study: HUNT3 in red, INTERVAL in green, KORA in blue, and QMDiab in purple. TFF3: Trefoil factor 3, RELT: Tumor necrosis factor receptor superfamily member 19L, IGFBP-6: Insulin-like growth factor-binding protein 6, DAN: Neuroblastoma suppressor of tumorigenicity 1, TNF sR-I: Tumor necrosis factor receptor superfamily member 1A, FSTL3: Follistatin-related protein 3, ARMEL: Cerebral dopamine neurotrophic factor.

Fig. 3. Results from the trans-ethnic discovery-replication observational study. Depicted in the left circle are the 57 proteins associated with eGFR, the continuous measurement of renal function; the 38 eGFR-specific proteins reflect associations along the full range of renal

function, whereas the 19 proteins also associated with CKD reflect a direct association with a clinically relevant low eGFR (<60 mL/min/1.73m^2). In bold are the two novel proteins found by this study.

Fig.4. Panel a. Scatter plot showing the individual genetic effects of the selected instrumental variables on log transformed eGFR (coefficient of the SNP-exposure association) on the x-axis and on testican-2 plasma levels (coefficient of the SNP-outcome) on the y-axis, along with their 95% CI. Each data point corresponds to an individual SNP. The lines correspond to the slopes of the different MR methods, which can be interpreted as the change in testican-2 levels per unit increase in log-transformed eGFR, and are color coded as follows: IVW-MR in light blue, MR-Egger in dark blue, weighted median in light green, weighted mode in dark green. **Panel b.** Forest plot showing the individual causal estimates of each of the 40 genetic instruments. The red points show the pooled estimates using all SNPs in the four methods. 95% CI are shown. IVW: inverse-variance-weighted; MR: Mendelian randomization.

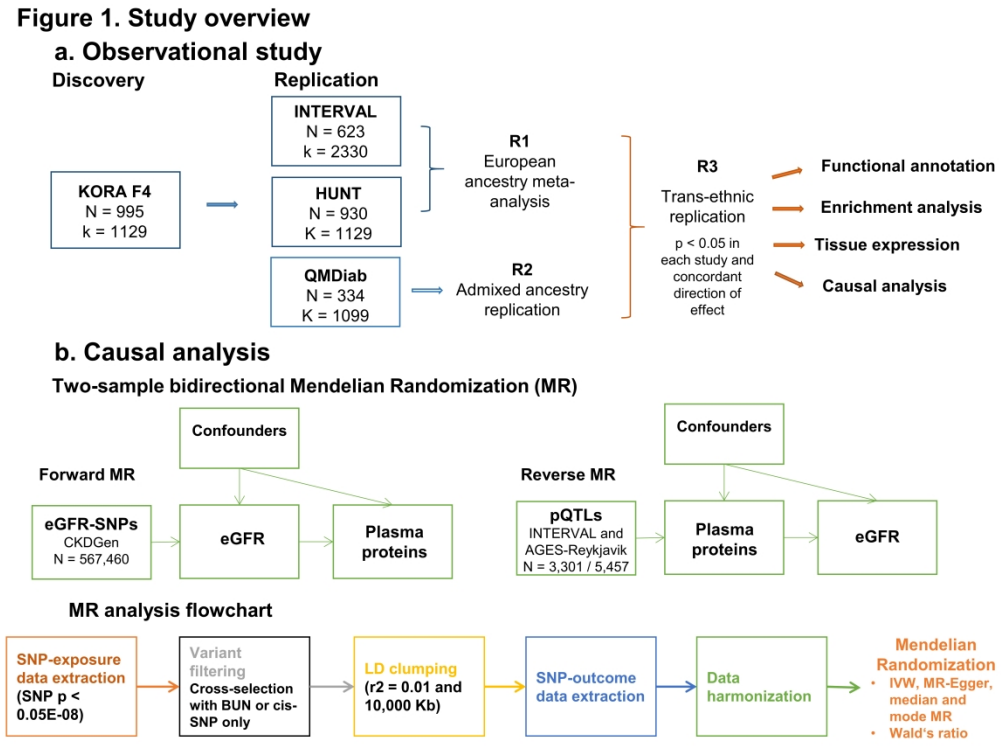


Fig. 1. Panel a. Cross-sectional association study. Data from 995 participants and 1129 proteins from KORA F4 was used in the discovery phase of a proteome wide association study of renal function using confounder-adjusted regression models. The replication studies were INTERVAL, HUNT and QMDiab. Three rounds of replication are shown: R1, replication based on the meta-analysis of p-values from the linear regression results of the studies with European ancestry; R2, replication based on the results of linear regression models performed in the Arab, South Asian and Filipino descent sample QMDiab; R3, identification of proteins consistently associated with eGFR across samples and ethnicities. The set of proteins identified in R3 was then functionally annotated and brought forward to the causal analysis phase. Panel b. Causal analysis. Two-sample bidirectional Mendelian randomization using data on participants from EA-studies in the CKDGen Consortium to instrument the forward analysis (eGFR causal to protein level) and data from INTERVAL and AGES-Reykjavik to instrument the reverse analysis (protein level causal to eGFR). Details on the data processing workflow for Mendelian randomizations are shown.

Figure 4. Results from forward MR estimating the causal effect of eGFR on testican-2

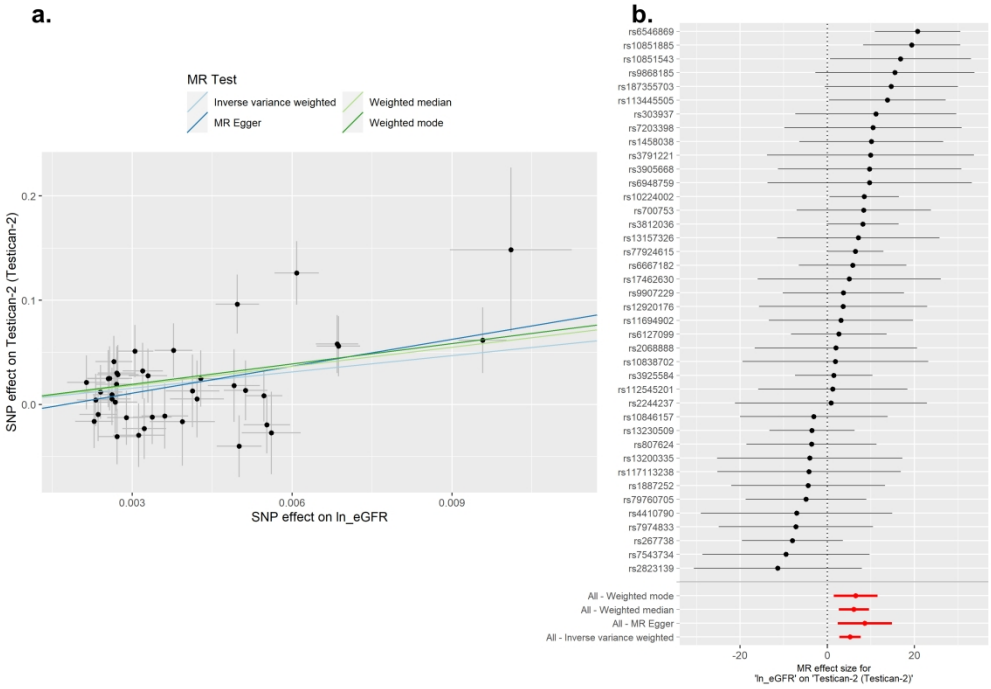


Fig.4. Panel a. Scatter plot showing the individual genetic effects of the selected instrumental variables on log transformed eGFR (coefficient of the SNP-exposure association) on the x-axis and on testican-2 plasma levels (coefficient of the SNP-outcome) on the y-axis, along with their 95% CI. Each data point corresponds to an individual SNP. The lines correspond to the slopes of the different MR methods, which can be interpreted as the change in testican-2 levels per unit increase in log-transformed eGFR, and are color coded as follows: IVW-MR in light blue, MR-Egger in dark blue, weighted median in light green, weighted mode in dark green. Panel b. Forest plot showing the individual causal estimates of each of the 40 genetic instruments. The red points show the pooled estimates using all SNPs in the four methods. 95% CI are shown. IVW: inverse-variance-weighted; MR: Mendelian randomization.

Plasma proteomics of renal function: a trans-ethnic meta-analysis and Mendelian randomization study

Supplemental Material

Supplemental Notes

Supplemental Note 1: Data availability

All data generated during this study are included in this published article and its supplementary information files. The analysis code in R is available on request. The informed consent given by the study participants does not cover posting of participant level phenotype data in public databases. Pre-existing data access policies for each of the four studies state research data requests can be submitted to each steering committee. Study-specific details regarding such requests are described in the next paragraphs:

KORA data are available upon request from KORA Project Application Self-Service Tool (https://epi.helmholtz-muenchen.de/); data requests can be submitted online and are subject to approval by the KORA Board. Data of the QMDiab study will be shared with researchers whose requests have been approved by the KORA Board.

The Nord-Trøndelag Health Study (HUNT) holds comprehensive data from more than 145,000 persons. HUNT Research Centre has been given concession to store and handle these data by the Norwegian Data Inspectorate. The key identification in the data base is the personal identification number given to all Norwegians at birth or immigration, whilst de-identified data are sent to researchers. Due to confidentiality HUNT Research Centre wants to limit storage of data outside HUNT databank, and we have restrictions for researchers for handling of HUNT data files. We have precise information on all data exported to different projects and there are no restrictions regarding data export given approval of applications to HUNT Research Centre. http://www.ntnu.edu/hunt/data

INTERVAL data are available via the European Genotype Phenotype archive (https://ega-archive.org/datasets/EGAD00001004080) and other data from the INTERVAL BioResource are available on reasonable request from helpdesk@intervalstudy.org.uk

The data supporting the findings from the kidney tissue investigations are are available upon reasonable request to the authors.

Supplemental Note 2: Covariate definition

Covariates included in the regression analyses were: age at the time of examination (years), sex (binary variable), BMI (kg/m2), smoking (current and former/never, self-reported), diabetes (fasting plasma glucose ≥ 126 mg/dl or treatment for diabetes), hypertension (systolic blood pressure ≥ 140 mm Hg or diastolic blood pressure ≥ 90 mm Hg or treatment for hypertension), log-transformed triglycerides (mg/dL), high-density lipoprotein (HDL, mg/dL), and intake of lipid-lowering drugs (yes/no, ATC code C10). There were some minor differences in the variable definitions in the replication studies: in the case of QMDiab, smoking status was a binary

variable based on cotinine detection in blood, diabetes status was assessed by casual glucose (non-fasting) ≥ 200 mg/dl (11.0 mmol/L) or treatment for diabetes or HbA1c > 6.5%, information on lipid-lowering drug intake was not available, and technical covariates (the top 3 principal components from each of the genetic and proteomic datasets) were included; in INTERVAL, self-reported medication history at a later timepoint (two years after proteomic measurement) was used to define hypertension and diabetes, and non-fasting blood samples were used for lipid measurement; in HUNT3, diabetes status was self-reported, lipid-lowering information was not available and non-fasting serum samples were used for lipid measurement.

Supplemental Note 3: Interaction analyses

We conducted interaction analyses in the discovery cohort (KORA) by adding an interaction term to the fully adjusted linear regression model:

ln(eGFR) ~ proteinx + interactor + proteinx * interactor + age + sex + BMI + smoking + diabetes + ln(triglycerides) + HDL + lipid medication intake + hypertension

where proteinx * interactor is the interaction term, and three interactors were assessed: age, sex, and smoking. One model per protein and per interactor was run. The full set of results for the 80 eGFR-associated proteins identified in KORA F4, the discovery cohort, are provided in Supplementary Table 9. The following interaction terms were significant at a Bonferroni-corrected p-value < 0.05 (0.05/997):

- 5 age*protein interactions: b2-Microglobulin, IGFBP-6, FSTL3, JAM-B, Renin

Target	Estimate	95% CI	pval	regression term
b2-Microglobulin	-0.060	-0.07,-0.05	1.12E-30	protein
b2-Microglobulin	-0.003	-0.004,-0.002	1.96E-07	protein:age
IGFBP-6	-0.058	-0.068,-0.048	5.00E-28	protein
IGFBP-6	-0.003	-0.004,-0.001	2.00E-05	protein:age
FSTL3	-0.039	-0.049,-0.029	1.46E-13	protein
FSTL3	-0.002	-0.004,-0.001	3.64E-05	protein:age
JAM-B	-0.035	-0.045,-0.026	1.27E-12	protein
JAM-B	-0.002	-0.004,-0.001	2.19E-05	protein:age
Renin	-0.024	-0.035,-0.013	1.89E-05	protein
Renin	-0.003	-0.004,-0.002	5.61E-07	protein:age

As an example, the significant interaction protein:age term from the model assessing b2-microglobulin suggests the negative association between b2-microglobulin plasma protein levels and log-transformed eGFR is accentuated with age, decreasing by 0.003 [95%CI = -0.004,-

0.002] for each additional year of age (p = 1.96E-07) (i.e. change in the slope of standardized protein relatively quantified levels for every one unit increase in age).

In summary, interaction analyses suggested the association between plasma protein levels and log-transformed eGFR varies by age: the association between protein level and log-transformed eGFR for most proteins decreased for each additional year of age. No statistically significant interactions were observed with sex or smoking. Full results for the 80 eGFR-associated proteins identified in KORA F4, the discovery cohort, from all three interaction analyses (age, sex and smoking) are provided in Suppl. Table 9.

Supplemental Note 4: Mendelian randomization analysis

Causality was assessed in the set of 57 proteins whose associations with eGFR showed trans-ethnic replication. Two-sample bidirectional Mendelian randomization[1] (MR) was used to estimate the causal effect of renal function (eGFR as proxy thereof) on plasma protein levels (forward MR) and vice versa (reverse MR, Fig. 1B). All MR analyses were conducted using MRBase as implemented in the R packages *TwoSampleMR* v.4.22 and *MRInstruments* v3.2[2].

Genetic instruments for eGFR and plasma proteins

We used publicly available genome-wide association (GWA) results for (a) eGFR from the CKDGen consortium, the largest GWA meta-analysis in European-ancestry populations[3], and (b) plasma proteins from INTERVAL[4] and AGES-Reykjavik[5], the largest studies using SOMAscan with no overlap with the studies included in the CKDGen meta-analysis. All studies had their own research protocols approved by the respective ethics committees and institutional boards, and included written informed consent from the participants.

Analyses in the CKDGen Consortium

A meta-analysis of genome-wide association (GWA) results performed in population-based samples was conducted by the CKDGen Consortium using data from 567,460 participants with European Ancestry (EA) from 85 population-based samples[3]. Data was imputed to the Haplotype Reference Consortium v1.1 or 1000 Genomes Project. Log transformed serum creatinine-based eGFR was used as the main outcome; some studies also had information on blood urea nitrogen (BUN), an alternative marker of kidney function inversely correlated with eGFR. Renal traits were adjusted for age and sex, and the residuals from these linear regression models were regressed on SNP dosage under an additive genetic model. Study site, genetic principal components, relatedness and other study-specific features were accounted for in the study-specific models. Genome-wide summary statistics were meta-analyzed using a

fixed-effects inverse-variance-weighted approach and made available at http://ckdgen.imbi.uni-freiburg.de [3].

Analyses in the INTERVAL and AGES-Reykjavik studies

In the INTERVAL Study, genotype-protein associations were determined in 3,301 participants with European ancestry[4]. Participation in the INTERVAL study was conditioned on blood donation criteria (exclusion of individuals with history of major disease and/or having recent illnesses). Genotype calling was done using the Affymetrix Axiom UK Biobank array and genotypes imputed using a combined 1000 Genomes Phase 3-UK10K reference panel[6]. Relative protein abundances were measured using an extended version of the SOMAscan assay and natural log-transformed prior to analysis[4]. Log transformed protein levels were adjusted for age, sex, duration between blood draw and processing (binary, ≤1 day/>1day), and the first three principal components of ancestry using linear regression. Residuals were then rank-inverse normalized and used in linear regression models under an additive genetic model to test association with genetic variants. Summary statistics were made available at http://www.phpc.cam.ac.uk/ceu/proteins/.

Genotype-protein associations were determined in 5,457 participants from the AGES-Reykjavik study with data imputed to the 1000 Genomes Project. Variants within a window of 150kbp up- or downstream from the protein coding genes were considered as cis SNPs. Relative concentrations of 4,137 proteins were measured using a custom-designed SOMAscan assay, and a Yeo-Johnson transformation applied prior to analysis. Protein levels were adjusted for age and sex, and linear regression analyses under an additive genetic model were conducted to test for association with genetic variants. Summary statistics for the single strongest SNP (lowest p value) per region from the significant cis-acting SNPs ($p < 5E-08$) were made available[5].

Instrument selection

A flow diagram presenting a summary of the genetic instrument selection and data harmonization process is shown in Supplementary Fig. 3 and described in detail in Supplementary Table 5.

For the forward MR, i.e. assessing the effect of renal filtration on protein levels, SNPs showing genome-wide significance with serum creatinine-based eGFR in the CKDGen results were selected as candidate instrumental variables (N = 256). In order to keep genetic signals of renal function and eliminate those more likely reflecting serum creatinine metabolism, only SNPs with a significant association with BUN (one-sided $P<0.05/256$, N=51) with effect direction

opposite of that to eGFR (N=211) were kept. 47 SNPs met both criteria and were retained for the next steps. Linkage disequilibrium (LD) clumping was conducted to identify independent variants (r2 = 0.01 and Kb = 10,000), resulting in the exclusion of 6 SNPs (rs9828976, rs28817415, rs6484504, rs2472297, rs506000, and rs111827672). Summary-level SNP-protein associations for the remaining 41 eGFR-associated SNPs were extracted from the INTERVAL GWAS results for 47 proteins.

For the reverse MR, i.e. interrogating the causal effect of proteins on renal filtration, one to 583 genome-wide significant SNPs for 37 proteins were identified in the INTERVAL GWA results. SNPs were classified as either cis-acting pQTLs (within 1Mb window from the start/end of the protein coding genes) or trans-acting pQTLs (outside the 1Mb window) based on the gene information retrieved with BioMart(Supplementary Table 6). No cis-SNPs were found for 9 proteins, whereas one to 581 cis-SNPs were found for 28 proteins. LD clumping was conducted (r2 = 0.01 and Kb = 10,000), resulting in a total of one to five independent cis-SNPs per protein. Summary-level SNP-eGFR associations for 28 proteins were then extracted from the CKDGen data (cis-SNP for UNC5H3 was not found). No SNPs were excluded in the data harmonization process, so that 28 proteins with 1 to 5 SNPs were available for MR.

The same strategy was followed to identify instruments in the AGES-Reykjavik GWAS . Of the 57 replicated eGFR-associated proteins, 20 were not available and eight had no genome-wide significant SNPs. Of the 29 proteins with genome-wide significant SNPs, all had at least one cis-SNP after LD clumping. Instruments for 26 proteins where then extracted from the SNP-outcome data, where SNPs for three proteins/genes (*ST6*, *CD55*, *RETN*) were not found in this dataset. One protein was further excluded in the data harmonization step due to it cis-SNP being a palindromic with intermediate allele frequency (rs28629977 for *EPHA2*). Finally, 25 proteins with 1 to 2 SNPs were available for MR.

In total, 35 unique proteins were tested in this reverse MR direction (i.e. effect of protein on eGFR). MR was performed for 11 proteins using only pGWAS data from INTERVAL and for 8 proteins using only pGWAS data from AGES-Reykjavik, whereas 16 proteins were tested using data from both datasets; the overlap between pGWAS datasets is shown in Suppl. Fig. 4. A total of 51 MR analyses were run in the reverse MR direction, thus the threshold for multiple testing correction was set at p Bonferroni = 9.43E-04 (0.05/51).

Data harmonization

Genetic effects were aligned to the exposure-increasing allele, and effect alleles, regression coefficients and effect allele frequencies were calculated correspondingly (i.e. effect allele was

coded as other allele, regression coefficient is multiplied by -1 and the new effect allele frequency is obtained after subtracting the old effect allele frequency from one[7]). Formatting and harmonization were done using the *TwoSampleMR* package[2]. Palindromic SNPs with minor allele frequency close to 50% and SNPs with incompatible alleles after data harmonization were excluded[7]. Harmonized datasets used in the MR analyses are available in Supplementary Table 8.

Phenotypic variance explained

The proportion of phenotypic variance explained by the SNPs was estimated as

$$\beta^2\left(\frac{2p\ (1-p)}{var}\right)$$

where β is the SNP effect, p is the effect allele frequency and var the variance of the sex- and age-adjusted phenotype residuals[3]. For log(eGFR) it was assumed to be 0.016 on the basis of data from 11,827 European-ancestry participants from the population-based ARIC study[8].

MR methods

Bidirectional Mendelian randomization was used to investigate the direction of the causal effects between plasma protein levels and renal function (with eGFR as a proxy thereof).

The primary analysis used inverse variance weighted (IVW) regression for analyses instrumented by two or more SNPs; in this method the coefficient of the gene-outcome association is regressed on the coefficient of the gene-exposure association with the intercept constrained to zero, assuming no directional pleiotropy [9, 10]. In cases where only one SNP instrumented the analysis, Wald's ratio (coefficient of the gene-outcome association divided by the gene-exposure association) was calculated instead [10].

To test the three assumptions upon which the validity of MR analyses depend – i.e. relevance (strength of genetic association), independence (specificity of association, i.e. no association with confounders) and exclusion restriction (only associated with outcome through the exposure)[1] – several analyses were conducted. Three further MR methods were used in sensitivity analyses for MR analyses instrumented by more than two SNPs[11] (i.e. 47 proteins in the forward MR and 2 proteins in the reverse MR, Supplementary Table 10). MR-Egger regression was used to assess pleiotropy, as this method allows for horizontal pleiotropy and provides an estimate of the unbalanced horizontal pleiotropic effects in its intercept[12]. Weighted median[13] and weighted mode MR[14], methods less sensitive to the presence of invalid instruments and to pleiotropic SNPs behaving as outliers, were also used. Consistency across causal estimates was investigated by performing single SNP analyses. Cochran's Q test was used to test for instrument heterogeneity[11] and the MR-PRESSO test was used to assess global

heterogeneity[15]. MR-PRESSO (*MR-PRESSO* v.2.2 in R) was also used to identify outlier variants (outlier test) and to obtain a causal estimate after exclusion of outliers. Leave-one-out analyses were run to identify individual SNPs potentially biasing the effect estimate, and funnel plots showing the MR estimate against their precision were done to visually check for directional pleiotropy[11].

Multiple testing was accounted for by the Bonferroni correction, defined as 0.05 divided by the number of total proteins assessed in each MR direction (i.e. 47 in forward MR, 51 in reverse MR). Causal effects were considered robust if these agreed in direction and magnitude across MR methods, were significant at Bonferroni $p < 0.05$ from the IVW estimator and had supporting evidence from MR methods in the sensitivity analyses.

Instrument specificity

In order to look for further evidence of horizontal pleiotropy, association between our SNPs and other traits were searched for in the GWAS Catalog[16]. Eleven of the SNPs instrumenting the forward MR analysis were identified as potentially pleiotropic, given their association with other traits (i.e. blood pressure, lipid dysregulation) as reported in the GWAS Catalog (reported by more than one study at $p < 5E-08$, Supplementary Table 9). Regarding the SNPs used to examine the causal effect of MIA on eGFR, one of the three (rs7937) was reported to be associated with chronic obstructive pulmonary disease (Supplementary Table 9).

Supplemental Note 5: Expression analyses in human kidney samples

Gene expression of the protein-coding genes identified in MR (*SPOCK2*, *CA3*, *CST6*, *MIA*) and renal traits was further assessed in (a) data from Nephroseq, a platform of comprehensive renal disease gene expression datasets (N = 458) [17]; and (b) RNA-sequencing-derived information on human kidney tissue (N = 427) [18].

Nephroseq analyses

We first used *Nephroseq* [17] (www.nephroseq.org) —a web-based platform for integrative data mining of comprehensive renal disease gene expression datasets—as a resource for association analysis between genes expressed in kidney and eGFR. Gene expression studies on human renal tissue with a minimum of 10 informative individuals were considered for the analyses. A total of 458 kidney samples from four eligible studies by Ju et al. [19] (261 samples), Sampson et al. [20] (92 samples), Reich et al. [21] (31 samples) and Rodwell et al. [22] (74 samples) were available for the association analysis. Rodwell et al. [22] samples were the only kidney tissues secured from patients without kidney disease. We meta-analysed the measures of association (i.e. Pearson's correlation coefficient) by inverse variance weighted meta-analysis

approach using random effect models [23]. The analysis was conducted separately for two different kidney groups. i.e. one from glomerular and cortex samples and the other from tubulointerstitial and medulla samples. Heterogeneity was examined using Cochran's Q test.

Human kidney tissue collections

We used 427 human kidney samples collected by five studies [18], namely *moleculAr analysis of human kiDney-Manchester renal tIssue pRojEct* (ADMIRE), the TRANScriptome of renaL human TissuE Study (TRANSLATE) [24-28], and its extension (TRANSLATE-T, 'zero time' pre-implantation biopsy prior to transplantation) [26], *moleculaR analysis of mEchanisms regulating gene exPression in post-ischAemic Injury to Renal allograft* (REPAIR) and *Renal gEne expreSsion and PredispOsition to cardiovascular and kidNey Disease* (RESPOND) studies.

In brief, TRANSLATE, ADMIRE and RESPOND studies collected samples from patients with unilateral kidney cancer – the specimen was taken from unaffected by cancer part of the kidney immediately after elective nephrectomy[24-28]. TRANSLATE-T and REPAIR collected pre-implantation kidney biopsies from deceased kidney donors prior to the organ transplantation[26]. The secured tissue samples from all the studies were immersed immediately in RNAlater or snap-frozen for the purpose of further molecular analysis. All patients were of white-European ethnicity. Further details on this resource of human kidney are described in [18].

Histology phenotype data

Histology samples were scored for 6 histological characteristics: glomerular sclerosis, glomerular Bowman's capsule thickening, tubular atrophy, interstitial fibrosis, interstitial inflammation and vascular lesions. All characteristics are scored from 0 to 3, 0 being no /minimal extent of the characteristic present and 3 usually representing extensive damage, as reported before [27].

Gene expression data

RNA sequencing data was generated from poly-A selected RNA samples run on Illumina sequencing instruments. Gene expression values were generated by Kallisto in units of transcripts per million (TPM). Before association testing all TPM values are transformed by log2(TPM+1), quantile normalised and standardised using the rank-based inverse normal transformation (expression units are therefore in standard deviations) [26].

Association analyses

In all association tests we adjusted for age, sex, BMI, 3 genetic principal components, diabetes and a variable number of surrogate variables (SV) (29 for eGFR and 26 for all histology phenotypes) in line with the computational pipelines developed and reported before [18]. The

surrogate variables allow us to adjust for unmeasured confounding variables, as reported before [18]. All analyses were conducted using R package *limma* with standard multivariate linear regression model (with gene expression as the response), as reported before [18].

Legends to Supplementary Tables

Supplementary Table 1. Results from cross-sectional analysis of eGFR and CKD in KORA F4

Results from discovery phase for eGFR and CKD. Provided in the Table are several protein identifiers (TargetFullName, Target, UniProt, EntrezGeneSymbol). Coefficient: linear regression coefficient, SE: standard error, 95% CI lower: lower boundary of 95% confidence interval for the regression coefficient, 95% CI upper: upper boundary of 95% confidence interval for the regression coefficient, N: sample size, Pval: unadjusted p-value, OR: odds ratio, 95% CI OR: 95% confidence interval from odds ratio. The last column shows TRUE if the protein was available in all four studies included (k = 993 in the common set of proteins).

Supplementary Table 2. Validation of proteomic targets from Emilsson et.al., 2018

Provided in the Table are several protein identifiers (TargetFullName, Target, UniProt, EntrezGeneSymbol). Columns starting with "Emilsson_" were extracted from Supplementary Tables 3 and 4 from [5]; Biological.Matrix: sample used in validation assay.

Supplementary Table 3. Validation of proteomic targets from Sun et.al., 2018

Provided in the Table are several protein identifiers (TargetFullName, Target, UniProt, EntrezGeneSymbol). Columns starting with "Sun_" were extracted from Supplementary Table 3 from [4]; further experimental details are found in [4]. The last column shows proteins flagged in terms of cross-reactivity issues: 0 for no binding observed, 1 for binding at least 10-fold weaker than target with product of same gene, 2 for binding at least 10-fold weaker than target with product of different gene, 3 for comparable binding observed to product of same gene and 4 for comparable binding observed to product of different gene.

Supplementary Table 4. Validation of proteomic targets from Williams et.al., 2019

Provided in the Table are several protein identifiers (TargetFullName, Target, UniProt, EntrezGeneSymbol). Columns starting with "Williams_" were extracted from Supplementary Table 3 from [29]; further experimental details are found in [29]. The last column shows proteins flagged in terms of cross-reactivity issues: 0 for no binding observed, 1 for binding at least 10-fold weaker than target with product of same gene, 2 for binding at least 10-fold weaker than target with product of different gene, 3 for comparable binding observed to product of same gene and 4 for comparable binding observed to product of different gene.

Supplementary Table 5. Genetic instrument selection and data harmonization (INTERVAL and AGES- Reykjavik)

Provided in the Table are several protein identifiers (TargetFullName, Target, UniProt, EntrezGeneSymbol). Significant discovery: TRUE if Bonferroni p-value < 0.05 in KORA F4, Significant R1: TRUE if replicated at p-value < 0.05 in meta-analysis and same direction of effect in studies of European Ancestry (i.e. INTERVAL and HUNT), Significant R2: TRUE if replicated at p-value < 0.05 and same direction of effect in non-European study (QMDiab), Significant R3: TRUE if replicated in R1 and R2, Significant eGFR: TRUE if replicated at p-value < 0.05 in trans-ethnic eGFR replication, Significant CKD: TRUE if replicated at p-value < 0.05 in trans-ethnic CKD replication. All columns starting with INTERVAL correspond to information extracted from the proteome GWAS reported in [4]; PGWAS: whether the protein was reported, IVs in PGWAS: number of genome-wide significant SNPs, N (extracted): number of genome-wide significant SNPs extracted from GWAS data, N (no RSID): SNPs for which no RSID was found, thus dropped from analysis, N (cis) / N (trans): total number of genome-wide significant SNPs of each type, N (LD excl.): number of SNPs excluded in LD pruning, N (cis left) / N (trans left): number of each type of SNPs left after LD pruning, N (left): total number of cis-SNPs qualifying as instrumental variables available in PGWAS data, N (cis in CKDGEN) / N (trans in CKDGEN): number of SNPs found in CKDGen data. The AGES-Reykjavik study was to identify potential instrumental variables for the 57 proteins; reported: found among aforementioned, genome-wide SNPs: TRUE if SNPs with pval < 5E-08 are available, N (cis SNPs) / N (trans): number of genome-wide significant SNPs of each type available, N (cis in CKDGEN) / N (trans in CKDGEN): SNPs found in CKDGEN data, N (cis excl. in harmonization): number of SNPs excluded in data harmonization due to it being a palindromic SNP with intermediate allele frequency, N (cis in CKDGEN final): number of cis-SNPs used in MR, study: last name of study providing GWAS data for identification of IV.

Supplementary Table 6. Harmonized summary statistics used in MR

Harmonized dataset with summary statistics from [3-5], as given in the output from the *TwoSampleMR* v.4.22 and *MRInstruments* v3.2[2] R packages.

Supplementary Table 7. Gene-trait information retrieved from GWAS Catalog for pleiotropic SNPs

SNP-trait associations obtained from the GWAS Catalog for all 57 proteins, filtered for SNPs meeting genome-wide significance and SNP-trait associations where the trait may act as a confounder between our examined exposures and outcomes (blood pressure, lipids,

cardiovascular disease, serum urate, platelet count, caffeine metabolism) reported by more than one study.

Supplementary Table 8. Sensitivity analyses with CysC-based eGFR, no adjustment for BMI and no adjustment for T2D

Results from discovery phase for cystatin C-based eGFR. Provided in the Table are several protein identifiers (TargetFullName, Target). Coefficient: linear regression coefficient, SE: standard error, 95% CI: 95% confidence interval, N: sample size, Pval: unadjusted p-value, Significant eGFR-CysC: TRUE if significant at p-value Bonferroni < 0.05 in this analysis, Significant eGFR-crea: TRUE if significant at p-value Bonferroni < 0.05 in main analysis, Significant no BMI adj.: TRUE if significant at p-value Bonferroni < 0.05 in this analysis, Significant BMI adj.: TRUE if significant at p-value Bonferroni < 0.05 in main analysis, Significant no T2D adj.: TRUE if significant at p-value Bonferroni < 0.05 in this analysis, Significant T2D adj.: TRUE if significant at p-value Bonferroni < 0.05 in main analysis.

Supplementary Table 9. Sensitivity analyses (interaction with age, sex and smoking) in KORA F4

Results from discovery phase for additional renal phenotypes. Provided in the Table are several protein identifiers (TargetFullName, Target). Analysis: interactor examined (either age, sex and smoking), Estimate: linear regression coefficient, 95%CI: 95% CI of the regression estimate, pval: unadjusted p-value, regression term: regression term from model (i.e. protein:age corresponds to interaction term with age). Significant p-values are highlighted in red.

Supplementary Table 10. Results from observational analysis of supplementary renal phenotypes (eGFR decline, log(uACR) and MA) in KORA F4

Results from discovery phase for additional renal phenotypes. Provided in the Table are several protein identifiers (TargetFullName, Target, UniProt, EntrezGeneSymbol). Coefficient: linear regression coefficient, SE: standard error, 95% CI: 95% confidence interval, N: sample size, Pval: unadjusted p-value, Pval (Bonf.): multiple testing adjusted p-value after Bonferroni. Shown in bold are significant proteins identified in this analysis.

Supplementary Table 11. Replication of cross-sectional eGFR-protein associations

Results from trans-ethnic replication phase for the 76 eGFR-associated proteins identified in the discovery sample that were available in the common set of 993 proteins. Provided in the Table are several protein identifiers (TargetFullName, Target), as well as the linear regression estimates from the four studies. Coefficient: linear regression coefficient, SE: standard error,

95% CI: 95% confidence interval, N: sample size, Pval: unadjusted p-value, Stouffer's zval: estimated effect from meta-analysis, Stouffer's pval: p-value from meta-analysis, Significant R1: TRUE if meta-analysis of studies of European ancestry (INTERVAL and HUNT) had p-value < 0.05 and the same direction of effect as the discovery, Significant R2: TRUE if replicated at p-value < 0.05 and same direction of effect as the discovery, Significant R3: TRUE if replicated in R1 and R2, Significant CKD: TRUE if replicated at p-value < 0.05 in trans-ethnic CKD replication, Previously associated with kidney function (PMID): PMIDs of literature on the protein and renal function, Reported in podocyte-exosome enriched urine (PMID: 23376485): TRUE if reported in publication, Comparison with Ngo 2020 (PMID: 32958645): "replicated" if one of 43 proteins also described in Ngo, et.al., "not reported" if one of 14 proteins discovered in our study not described in Ngo and colleagues' work, "not checked" if replication not verified. Shown in bold are those proteins showing robust trans-ethnic associations in this analysis.

Supplementary Table 12. Replication of cross-sectional CKD-protein associations

Results from trans-ethnic replication phase for the 34 CKD-associated proteins identified in the discovery sample that were available in the common set of 993 proteins. Provided in the Table are several protein identifiers (TargetFullName, Target), as well as the linear regression estimates from the four studies. Coefficient: linear regression coefficient, SE: standard error, 95% CI: 95% confidence interval, N: sample size, Pval: unadjusted p-value, Significant R1: TRUE if association replicated in HUNT, where replication is defined as p-value<0.05 and same direction of effect as discovery, Significant R2: TRUE if association replicated in QMDiab, where replication is defined as p-value<0.05 and same direction of effect as discovery, Significant R3: TRUE if replicated in R1 and R2, Significant eGFR: TRUE if replicated in trans-ethic replication. Shown in bold are significant proteins identified in this analysis.

Supplementary Table 13. Extended annotation file (DAVID)

Extended annotation on Gene Ontology terms (biological process – BP_DIRECT, cellular component – CC_DIRECT, molecular function - MF_DIRECT), pathways (BioCarta, KEGG, BBID), functional category (UP_KEYWORDS, UP_SEQ_FEATURE), protein domain/family (SMART, PIR_SUPERFAMILY), disease association (OMIM) as obtained using DAVID v6.8 [30].

Supplementary Table 14. Gene information from proteins included in MR

Provided in the Table are several protein identifiers (TargetFullName, Target, UniProt, EntrezGeneSymbol), as well as chromosome, gene start/end position, strand and starting/ending positions, annotated to GRCh37 and obtained using BioMartR.

Supplementary Table 15. Phenotypic variance explained by MR instruments

Exposure: log(eGFR) for forward MR, UniProt for reverse MR, followed by several protein identifiers. IVs: SNPs used as instruments, N IVs: number of SNPs instrumenting MR, exposure variance: variance of sex- and age-adjusted phenotype residuals estimated in KORA for plasma proteins, phenotypic variance explained by SNPs: estimated based on the SNP-phenotype effect, effect allele frequency and variance of the sex- and age-adjusted phenotype residual, N: sample size (analyses done in KORA F4).

Supplementary Table 16. Results from MR (IVW, MBE, weighted median and MR-Egger)

id.exposure: ln_eGFR in forward MR (i.e. effect of renal filtration on plasma protein levels) and EntrezGeneSymbol in reverse MR, id.outcome: EntrezGeneSymbol in forward MR (i.e. effect of renal filtration on plasma protein levels) and ln_eGFR in reverse MR, outcome: protein name in forward MR and log(eGFR) in reverse MR, exposure: log(eGFr) in forward MR and protein name in reverse MR, method: MR method used, nsnp: number of SNPs instrumenting the analysis, b: causal estimate, SE: standard error, lower CI: lower bound of 95% CI, upper CI: upper bound of 95% CI, pval: p-value, study: last name and year of publication from studies used for selection of instrumental variables, direction: forward when estimating effect of eGFR on proteins and reverse when estimating effect of proteins on eGFR.

Supplementary Table 17. Sensitivity analyses: Heterogeneity (Cochran's Q test)

id.exposure: ln_eGFR in forward MR (i.e. effect of renal filtration on plasma protein levels) and EntrezGeneSymbol in reverse MR, id.outcome: EntrezGeneSymbol in forward MR (i.e. effect of renal filtration on plasma protein levels) and ln_eGFR in reverse MR, outcome: protein name, exposure: log(eGFr) in forward MR and protein name in reverse MR, method: MR methods used, Q: Cochran's Q, Q_df: degrees of freedom, Q pval: heterogeneity p-value, direction: forward when estimating effect of eGFR on proteins and reverse when estimating effect of proteins on eGFR.

Supplementary Table 18. Sensitivity analyses: Pleiotropy in MR-Egger

id.exposure: ln_eGFR in forward MR (i.e. effect of renal filtration on plasma protein levels) and EntrezGeneSymbol in reverse MR, id.outcome: EntrezGeneSymbol in forward MR (i.e. effect of renal filtration on plasma protein levels) and ln_eGFR in reverse MR, outcome: protein name name in forward MR and log(eGFR) in reverse MR, exposure: log(eGFr) in forward MR and protein name in reverse MR, Egger intercept: intercept term in Egger regression, SE: standard error, Pval: p-value, MR direction: forward when estimating effect of eGFR on proteins and reverse when estimating effect of proteins on eGFR.

Supplementary Table 19. Sensitivity analyses: Leave-one-out analyses

id.exposure: ln_eGFR in forward MR (i.e. effect of renal filtration on plasma protein levels) and EntrezGeneSymbol in reverse MR, id.outcome: EntrezGeneSymbol in forward MR (i.e. effect of renal filtration on plasma protein levels) and ln_eGFR in reverse MR, outcome: protein name, exposure: log(eGFr) in forward MR and protein name in reverse MR, SNP: identifier of specific variant excluded in MR, N: sample size, b: causal estimate, SE: standard error, pval: p-value, direction: forward when estimating effect of eGFR on proteins and reverse when estimating effect of proteins on eGFR.

Supplementary Table 20. Sensitivity analyses: Results from restrictive MR

Comparison of MR results obtained in the main forward analysis (instrumented by 40 SNPs) and sensitivity restrictive MR (after exclusion of eleven SNPs) for multiple methods; id.outcome: EntrezGeneSymbol corresponding to protein coding gene, b: causal estimate, se: standard error, pval: p-value. Highlighted in green are results meeting statistical significance after Bonferroni correction (0.05/47), and highlighted in yellow results meeting nominal significance (pval < 0.05).

Supplementary Table 21. Results from correlation analyses between gene expression and eGFR from Nephroseq datasets

Using datasets curated by *Nephroseq* [17] (www.nephroseq.org), the analysis assessing renal gene expression of the four proteins identified in MR and eGFR was conducted separately for two different kidney groups: one from glomerular and cortex samples and the other from tubulointerstitial and medulla samples, shown on the boxes on the left (dataset_glom_cortex and dataset_tubint_medulla). A total of 458 kidney samples from four eligible studies by Ju et al. [19] (261 samples), Sampson et al. [20] (92 samples), Reich et al. [21] (31 samples) and Rodwell et al. [22] (74 samples) were available for the association analysis. The tables on the right show gene_symbol: Entrez Gene symbol, meta_beta: meta-analytic measure of association (i.e. Pearson's correlation coefficient) obtained by inverse variance weighted meta-analysis approach using random effect models [23], mean_se: standard error of estimate, meta_pval: pvalue of meta-analytic estimate, meta_het: Cochran's Q test p value. Highlighted in bold are gene/proteins where a significant (p<0.05) effect was observed.

Supplementary Table 22. Clinical characteristics of studies included in gene expression analyses

Basic clinical characteristics from studies included in the human kidney resource (up to N = 427 kidney samples) [18].

Supplementary Table 23. Results from multivariate regression analyses on gene expression, eGFR and histological characteristic scoring from human kidney resource

Regression analyses on the association between renal traits (eGFR, histologic scores) and gene expression of *SPOCK2, MIA, CST6* and *CA3.* Regression models included adjustment for age, sex, BMI, 3 genetic principal components, diabetes and a variable number of surrogate variables (29 for eGFR and 26 for all histology phenotypes). Beta: regression estimate, SE: standard error, 95% CI lower and upper: lower and upper bounds of the 95% CI of the regression estimate, P-value: pvalue from regression, N: sample size.

Supplementary Table 24. Description and biological roles of selected proteins

Protein, description, MW: molecular weight, glomerular filtration / detection in urine: whether the protein is filtrated at the glomeruli and if it has ever been reported in urine, biological role, relevance to kidney function. References used in this table appear at the end of this document.

Legends to Supplementary Figures

Supplementary Figure 1. Correlation of serum creatinine variables in KORA F4

Panel A) Quadrant I (top right) shows correlation between NMR-based serum creatinine variable (same as available in INTERVAL) and the standardized Jaffe reaction serum creatinine variable. Quadrants II (top left) and IV (bottom right) show density plot of standardized Jaffe reaction serum creatinine variable and NMR-based variable. Quadrant III (bottom left) shows a scatterplot of both variables. Panel B) Quadrant I (top right) shows correlation between NMR-based eGFR (same as available in INTERVAL) and the standard eGFR variable. Quadrants II (top left) and IV (bottom right) show density plot of both variables. Quadrant III (bottom left) shows a scatterplot of both variables.

Supplementary Figure 2. Correlation between aptamer-based and other measurements for proteins in KORA F4

Scatterplots showing the distribution of the aptamer-based and proximity extension assay measurements of 12 plasma proteins in a subset of the population-based sample studied in the discovery step (KORA F4, N = 174). The correlation between both measurements (Pearson's correlation) and its statistical significance are shown on each plot. Log-transformed normalized data in relative fluorescence units (RFU) from the aptamer-based platform is shown on the x-axis, and log-transformed normalized data in normalized protein expression (NPX) is shown on the y-axis.

Supplementary Figure 3. Genetic instrument selection and data harmonization

Flow diagram showing genetic instrument selection and data harmonization process. Five general steps were taken (identification of genome-wide significant SNPs, additional filtering, LD clumping, extraction of SNP-outcome data, and data harmonization). Number of SNPs identified, excluded and kept for analysis are provided for each step are given.

Supplementary Figure 4. Protein overlap in pGWAS datasets used in reverse direction of MR

Venn diagram showing the sets of proteins identified in the pGWAS datasets from INTERVAL and AGES-Reykjavik studies. The left circle shows the set of proteins for which the MR analysis was performed using INTERVAL pGWAS summary statistics, whereas the right circle shows the set of proteins for which the MR analysis was performed using AGES-Reykjavik summary statistics. In total, 35 unique proteins were tested in the reverse direction of the MR analyses (i.e. effect of protein on eGFR). MR was performed for 11 proteins using only pGWAS data from

INTERVAL and for 8 proteins using only pGWAS data from AGES-Reykjavik, whereas 16 proteins were tested using data from both datasets.

Supplementary Figure 5. Proteins and log(eGFR) distribution in discovery dataset
Scatterplots showing log-transformed estimated glomerular filtration rate (eGFR) on the x-axis and protein plasma concentrations in relative-fluorescence units (RFU) on the y-axis. Pearson's correlation coefficients and their corresponding p-values are shown in the plot. Shown are the top 10 proteins. Protein-log(eGFR) observations are color coded in agreement with chronic kidney disease (CKD) staging [31]: G1 as normal or high GFR (GFR > 90 mL/min) in blue, G2 as mild CKD (GFR = 60-89 mL/min) in green, and G3 as moderate CKD (GFR = 30-59 mL/min) in red. ARMEL: Cerebral dopamine neurotrophic factor, TNF SR-I: Tumor necrosis factor receptor superfamily member 1A, DAN: Neuroblastoma suppressor of tumorigenicity 1, RGMB: RGM domain family member B, FSTL3: Follistatin-related protein 3, JAM-B: Junctional adhesion molecule B.

Supplementary Figure 6. Cross sectional results for eGFR-protein associations across studies
Volcano plots from all included studies: panel A) data from N = 995 individuals from KORA F4, panel B) data from N = 930 individuals from HUNT3, panel C) data from N = 623 individuals from INTERVAL, and panel D) data from N = 334 individuals from QMDiab. The linear regression coefficient for the protein term is shown on the x-axis, and statistical significance as –log(pvalue) on the y-axis. Labeled and colored in red are significant associations at $p < 0.05$ in each individual study.

Supplementary Figure 7. Proteins and log(eGFR) distribution in discovery dataset after CKD exclusion
Scatterplots showing log-transformed estimated glomerular filtration rate (eGFR) on the x-axis and protein plasma concentrations in relative-fluorescence units (RFU) on the y-axis in the discovery dataset (KORA F4) after exclusion of 38 individuals with CKD. Pearson's correlation coefficients and their corresponding p-values are shown in the plot. Shown are the top 10 proteins. Protein-log(eGFR) observations are color coded in agreement with chronic kidney disease (CKD) staging [31]: G1 as normal or high GFR (GFR > 90 mL/min) in blue, G2 as mild CKD (GFR = 60-89 mL/min) in green, and G3 as moderate CKD (GFR = 30-59 mL/min) in red. ARMEL: Cerebral dopamine neurotrophic factor, TNF SR-I: Tumor necrosis factor receptor superfamily member 1A, DAN: Neuroblastoma suppressor of tumorigenicity 1, RGMB: RGM

domain family member B, FSTL3: Follistatin-related protein 3, JAM-B: Junctional adhesion molecule B.

Supplementary Figure 8. Correlation between Z-values for eGFR-protein associations across studies

Correlation between z-values (regression coefficient divided by its standard error) across all included studies was calculated for the set of proteins significantly associated with eGFR in the discovery study (k=76 proteins at p < Bonferroni in discovery). The plot shows pairwise scatterplots and correlations between Z-values between each pair of studies, as well as four plots of the density of the z-value distribution of each study in the diagonal of the plot matrix.

Supplementary Figure 9. Tissue expression of 57 eGFR-associated proteins (ProteomeDB)

Tissue expression of 57 eGFR-associated proteins (ProteomeDB). Heatmap showing expression of proteins as rows and biological sources as columns, respectively. The dendrograms show the results of hierarchical clustering of proteins and biological sources. Protein expression values were produced by the MS1 quantification technique and expression values estimated by the iBAQ approach. Highlighted with yellow is the column corresponding to kidney tissue. The data presented in this figure were generated through a multi-protein query using the UniProt IDs on the ProteomicsDB Analytics Toolbox portal on Oct. 2, 2020 from: https://www.proteomicsdb.org/proteomicsdb/#analytics/expressionHeatmap

Supplementary Figure 10. Tissue expression of 56 eGFR-associated protein coding genes (ProteomeDB)

Heatmap showing expression of 56 proteins-coding genes (the set of 57 proteins identified in as trans-ethnically associated with eGFR are the product of 56 genes) as rows and biological sources as columns, respectively. The dendrograms show the results of hierarchical clustering of proteins and biological sources. Panel A shows data from RNA-seq experiments, and panel B data from microarray experiments. Highlighted with yellow is the column corresponding to kidney tissue. The data presented in this figure were generated through a multi-protein query using the UniProt IDs on the ProteomicsDB Analytics Toolbox portal on Oct. 2, 2020 from: https://www.proteomicsdb.org/proteomicsdb/#analytics/expressionHeatmap

Supplementary Figure 11. Expression of 56 eGFR-associated protein coding genes across tissues (GTEx)

Gene expression of 57 eGFR-associated proteins across tissues (GTEx). This heatmap provides a qualitative measure of relative expression across human tissues, shown in Transcripts Per Million (TPM) (https://doi.org/10.1093/bioinformatics/btp692). Values across tissues may not be compared due to the differences in sample normalization across the diverse set of tissues represented in GTEx (https://doi.org/10.1016/j.cell.2012.10.012). The columns corresponding to cortex and medulla kidney tissue samples are highlighted in orange. The data presented in this plot was generated on Oct. 2, 2020 through a multi-gene query on the GTEx portal https://www.gtexportal.org/home/multiGeneQueryPage

Supplementary Figure 12. Protein-protein interaction network of 57 replicated eGFR-associated proteins

Protein-protein interaction network of 57 replicated eGFR-associated proteins. Colored proteins are query proteins, no specific color-coding is applied. Known interactions: cyan blue edges are retrieved from curated databases, pink edges are experimentally determined; predicted interactions: bright green edges are retrieved from gene neighborhood, red edges from gene fusions and navy blue from gene co-occurrence; others: golden edges are retrieved from text mining, black from co-expression and light purple from protein homology.

Supplementary Figures 13-19. Forward MR results for effects of eGFR on proteins.

Panel A) Forest plot showing IVW causal estimates following a leave-one-out approach (ie. presented is the casual estimate using the leave-one-out approach, when the given SNP is not included in the analysis). Panel B) Funnel plot showing the ratio estimate for each variant on the x-axis and its square root precision on the y-axis, where asymmetry suggests directional pleiotropy. Vertical lines represent the causal estimates obtained in each method, color coded as in panel C.

Supplementary Figure 20. Reverse MR analysis for MIA-eGFR

Panel A) Forest plot showing individual contributions (x-axis) of each SNP instrumenting the analysis (y-axis), followed by the pooled MR estimates obtained with all MR methods. IVW: inverse variance-weighted MR. Panel B) Scatter plot of summary data estimates for the associations of 40 SNPs with log(eGFR) (x-axis) and plasma proteins (y-axis). The lines correspond to the slopes from the IVW (shown in light blue), weighted median (shown in light green), MR-Egger (shown in dark blue) and weighted mode (shown in dark green). Panel C) Funnel plot showing IVW causal estimates following a leave-one-out approach.

Supplementary Figure 21. *SPOCK2* gene expression in renal tissue from 26 CKD patients

Scatterplot showing gene expression from microdissected tubulointerstitial components of human renal biopsies from 26 individuals with CKD at different disease stages (I-IV)[19]. eGFR is shown in the x-axis and renal *SPOCK2* gene expression in the y-axis. Stages of chronic kidney disease are color coded following the KDIGO's GFR categories: stage 1: normal or high GFR (≥90 ml/min/ 1.73m^2); stage II: mildly decreased (60-29 ml/min/ 1.73m^2); stage III: mild to moderately decreased (59-30 ml/min/ 1.73m^2); stage IV: severely decreased (15-29 ml/min/ 1.73m^2); stage V: kidney failure (15 ml/min/ 1.73m^2). Shown in blue is the regression line corresponding to the eGFR ~ *SPOCK2* expression model.

References

1. Pierce, BL, Burgess, S: Efficient design for Mendelian randomization studies: subsample and 2-sample instrumental variable estimators. *American journal of epidemiology,* 178**:** 1177-1184, 2013.
2. Hemani, G, Zheng, J, Elsworth, B, Wade, KH, Haberland, V, Baird, D, et al.: The MR-Base platform supports systematic causal inference across the human phenome. *eLife,* 7, 2018.
3. Wuttke, M, Li, Y, Li, M, Sieber, KB, Feitosa, MF, Gorski, M, et al.: A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nature genetics,* 51**:** 957-972, 2019.
4. Sun, BB, Maranville, JC, Peters, JE, Stacey, D, Staley, JR, Blackshaw, J, et al.: Genomic atlas of the human plasma proteome. *Nature,* 558**:** 73-79, 2018.
5. Emilsson, V, Ilkov, M, Lamb, JR, Finkel, N, Gudmundsson, EF, Pitts, R, et al.: Co-regulatory networks of human serum proteins link genetics to disease. *Science (New York, NY),* 361**:** 769-773, 2018.
6. Astle, WJ, Elding, H, Jiang, T, Allen, D, Ruklisa, D, Mann, AL, et al.: The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell,* 167**:** 1415-1429.e1419, 2016.
7. Hartwig, FP, Davies, NM, Hemani, G, Davey Smith, G: Two-sample Mendelian randomization: avoiding the downsides of a powerful, widely applicable but potentially fallible technique. *International journal of epidemiology,* 45**:** 1717-1726, 2016.
8. Gorski, M, van der Most, PJ, Teumer, A, Chu, AY, Li, M, Mijatovic, V, et al.: 1000 Genomes-based meta-analysis identifies 10 novel loci for kidney function. *Scientific reports,* 7**:** 45040, 2017.
9. Burgess, S, Butterworth, A, Thompson, SG: Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic epidemiology,* 37**:** 658-665, 2013.
10. Teumer, A: Common Methods for Performing Mendelian Randomization. *Frontiers in cardiovascular medicine,* 5**:** 51, 2018.
11. Burgess, S, Bowden, J, Fall, T, Ingelsson, E, Thompson, SG: Sensitivity Analyses for Robust Causal Inference from Mendelian Randomization Analyses with Multiple Genetic Variants. *Epidemiology,* 28**:** 30-42, 2017.

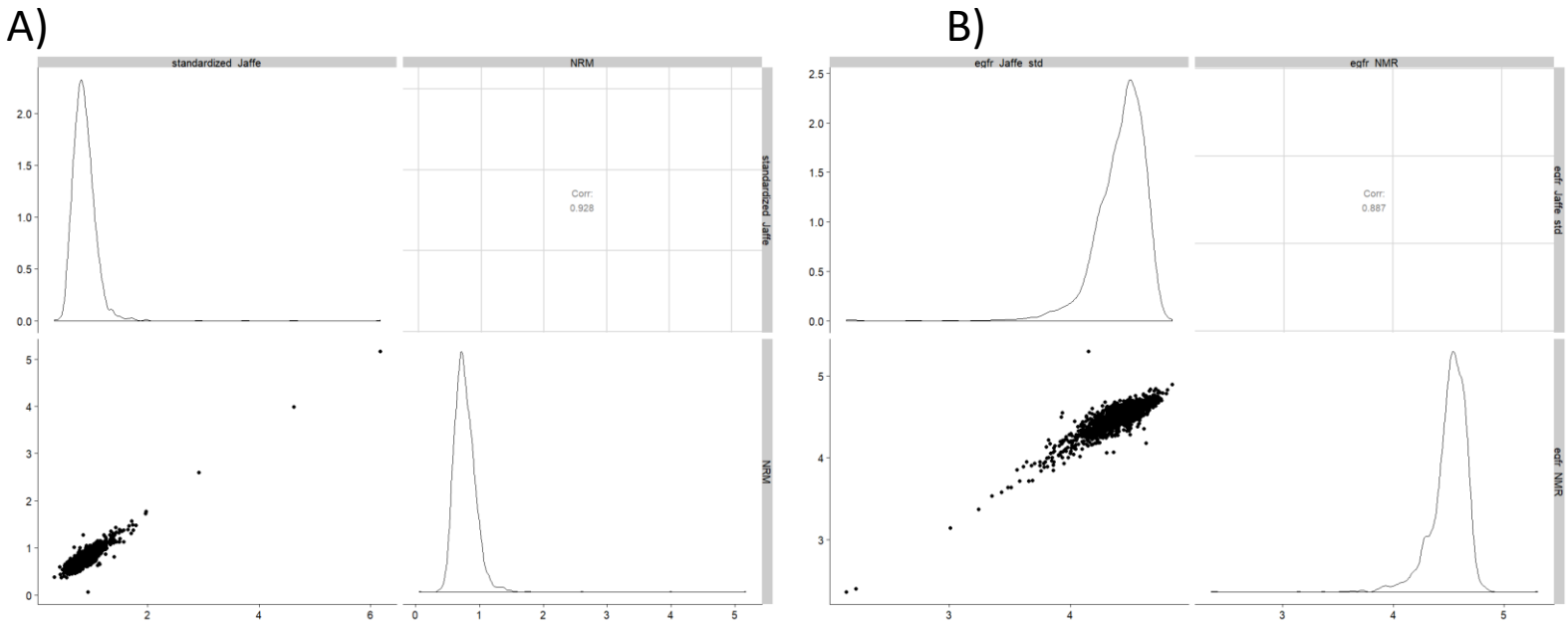
12. Bowden, J, Davey Smith, G, Burgess, S: Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International journal of epidemiology,* 44**:** 512-525, 2015.
13. Bowden, J, Davey Smith, G, Haycock, PC, Burgess, S: Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genetic epidemiology,* 40**:** 304-314, 2016.
14. Hartwig, FP, Davey Smith, G, Bowden, J: Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *International journal of epidemiology,* 46**:** 1985-1998, 2017.
15. Verbanck, M, Chen, C-Y, Neale, B, Do, R: Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nature genetics,* 50**:** 693-698, 2018.
16. Buniello, A, MacArthur, JAL, Cerezo, M, Harris, LW, Hayhurst, J, Malangone, C, et al.: The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research,* 47**:** D1005-d1012, 2019.
17. Martini, S, Eichinger, F, Nair, V, Kretzler, M: Defining human diabetic nephropathy on the molecular level: Integration of transcriptomic profiles with biological knowledge. *Reviews in Endocrine and Metabolic Disorders,* 9**:** 267-274, 2008.
18. Jiang, X, Eales, JM, Scannali, D, Nazgiewicz, A, Prestes, P, Maier, M, et al.: Hypertension and renin-angiotensin system blockers are not associated with expression of angiotensin-converting enzyme 2 (ACE2) in the kidney. *European Heart Journal*, 2020.
19. Ju, W, Nair, V, Smith, S, Zhu, L, Shedden, K, Song, PXK, et al.: Tissue transcriptome-driven identification of epidermal growth factor as a chronic kidney disease biomarker. *Science translational medicine,* 7**:** 316ra193, 2015.
20. Sampson, MG, Robertson, CC, Martini, S, Mariani, LH, Lemley, KV, Gillies, CE, et al.: Integrative Genomics Identifies Novel Associations with APOL1 Risk Genotypes in Black NEPTUNE Subjects. *Journal of the American Society of Nephrology : JASN,* 27**:** 814-823, 2016.
21. Reich, HN, Tritchler, D, Cattran, DC, Herzenberg, AM, Eichinger, F, Boucherot, A, et al.: A molecular signature of proteinuria in glomerulonephritis. *PloS one,* 5**:** e13451, 2010.
22. Rodwell, GE, Sonu, R, Zahn, JM, Lund, J, Wilhelmy, J, Wang, L, et al.: A transcriptional profile of aging in the human kidney. *PLoS biology,* 2**:** e427, 2004.
23. Balduzzi, S, Rücker, G, Schwarzer, G: How to perform a meta-analysis with R: a practical tutorial. *Evidence Based Mental Health,* 22**:** 153-160, 2019.

24. Marques, FZ, Romaine, SP, Denniff, M, Eales, J, Dormer, J, Garrelds, IM, et al.: Signatures of miR-181a on the Renal Transcriptome and Blood Pressure. *Molecular medicine (Cambridge, Mass),* 21**:** 739-748, 2015.
25. Tomaszewski, M, Eales, J, Denniff, M, Myers, S, Chew, GS, Nelson, CP, et al.: Renal Mechanisms of Association between Fibroblast Growth Factor 1 and Blood Pressure. *Journal of the American Society of Nephrology : JASN,* 26**:** 3151-3160, 2015.
26. Xu, X, Eales, JM, Akbarov, A, Guo, H, Becker, L, Talavera, D, et al.: Molecular insights into genome-wide association studies of chronic kidney disease-defining traits. *Nature communications,* 9**:** 4800, 2018.
27. Rowland, J, Akbarov, A, Eales, J, Xu, X, Dormer, JP, Guo, H, et al.: Uncovering genetic mechanisms of kidney aging through transcriptomics, genomics, and epigenomics. *Kidney Int,* 95**:** 624-635, 2019.
28. Morris, AP, Le, TH, Wu, H, Akbarov, A, van der Most, PJ, Hemani, G, et al.: Trans-ethnic kidney function association study reveals putative causal genes and effects on kidney-specific disease aetiologies. *Nature communications,* 10**:** 29, 2019.
29. Williams, SA, Kivimaki, M, Langenberg, C, Hingorani, AD, Casas, JP, Bouchard, C, et al.: Plasma protein patterns as comprehensive indicators of health. *Nature medicine,* 25**:** 1851-1857, 2019.
30. Huang, DW, Sherman, BT, Lempicki, RA: Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols,* 4**:** 44-57, 2009.
31. Hill, NR, Fatoba, ST, Oke, JL, Hirst, JA, O'Callaghan, CA, Lasserson, DS, et al.: Global Prevalence of Chronic Kidney Disease - A Systematic Review and Meta-Analysis. *PloS one,* 11**:** e0158765, 2016.

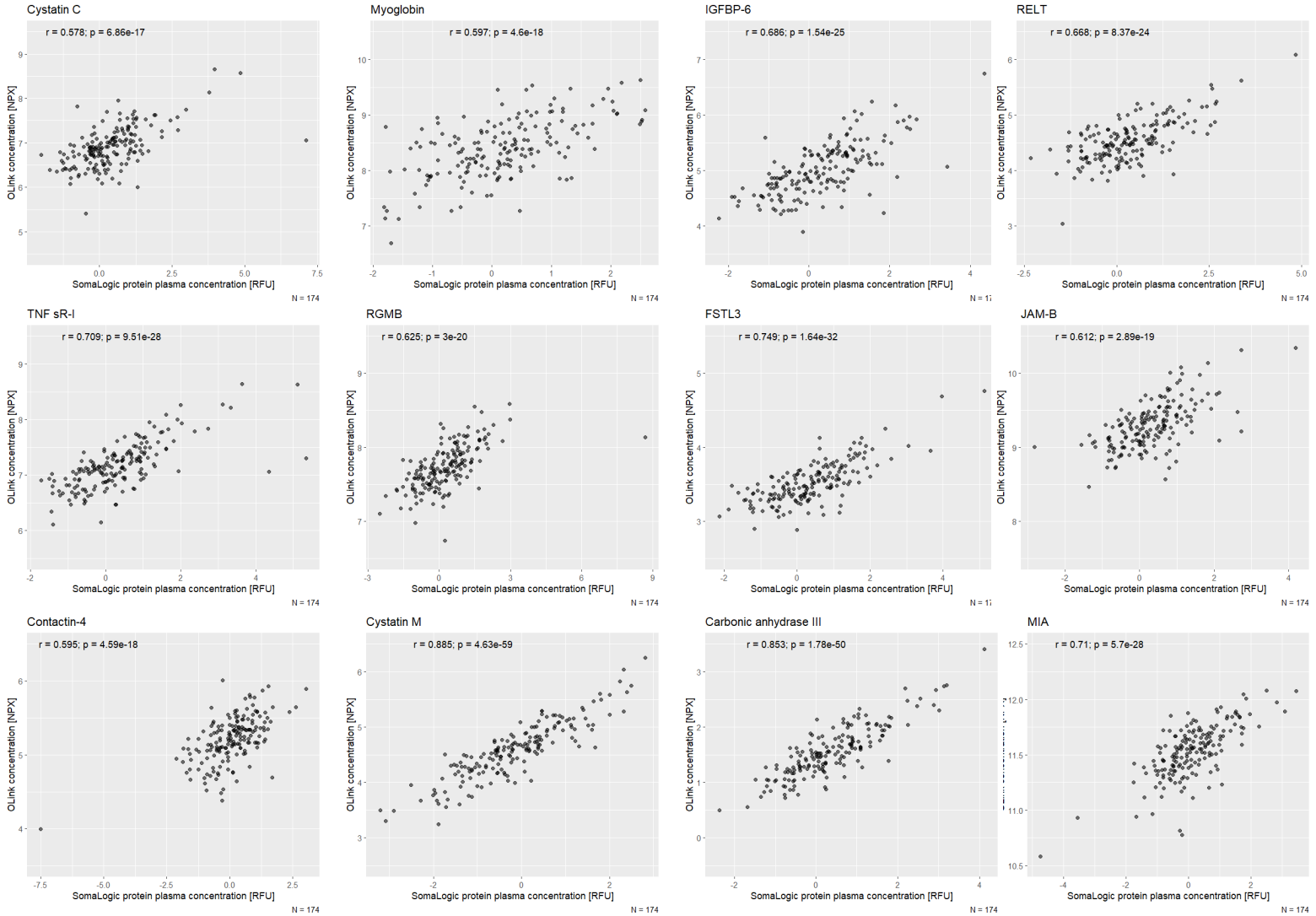
Supplementary Figures

Plasma proteomics of renal function: a trans-ethnic meta-analysis and Mendelian randomization study

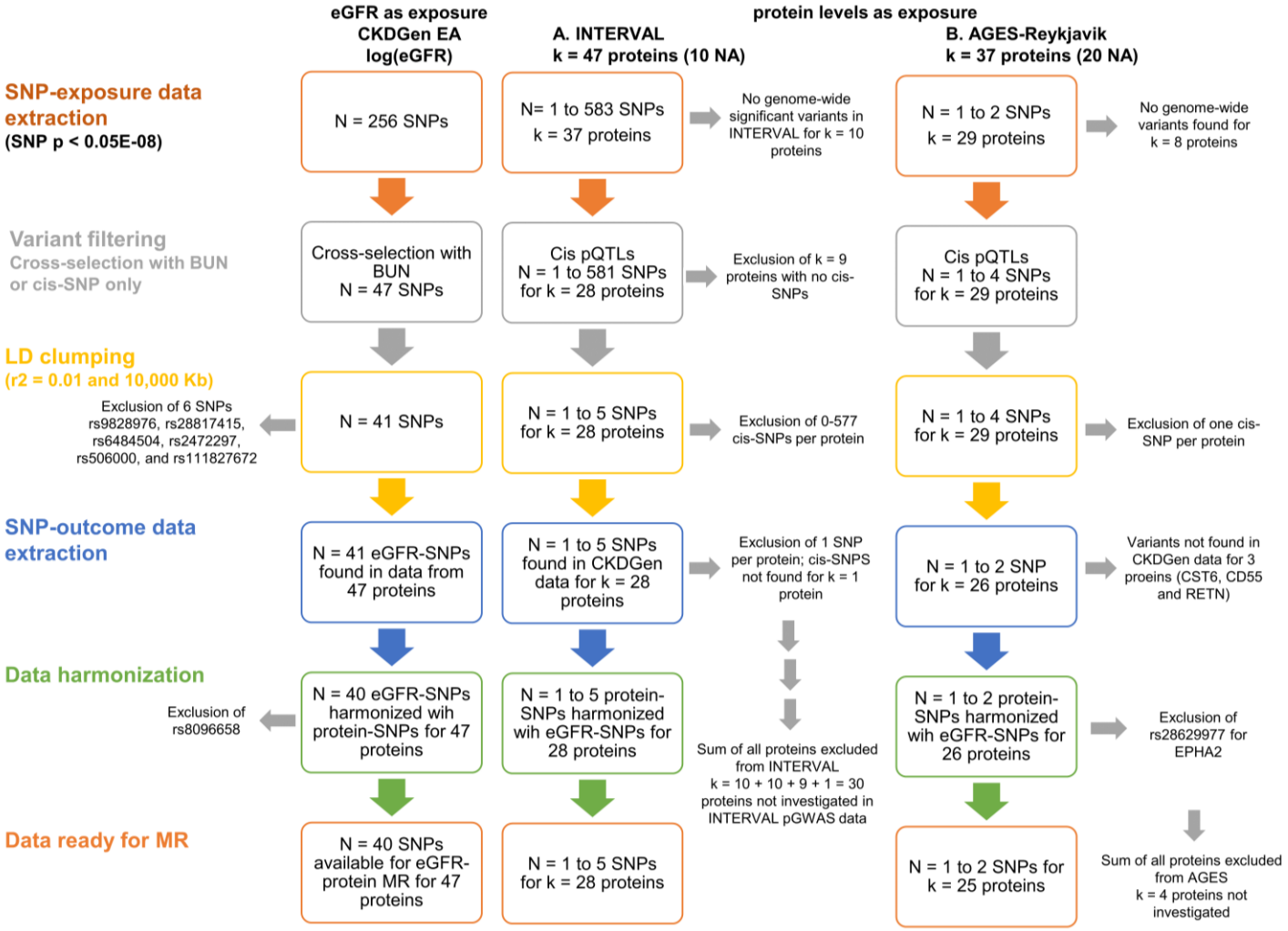
Suppl. Fig. 1. Correlation of serum creatinine variables in KORA F4



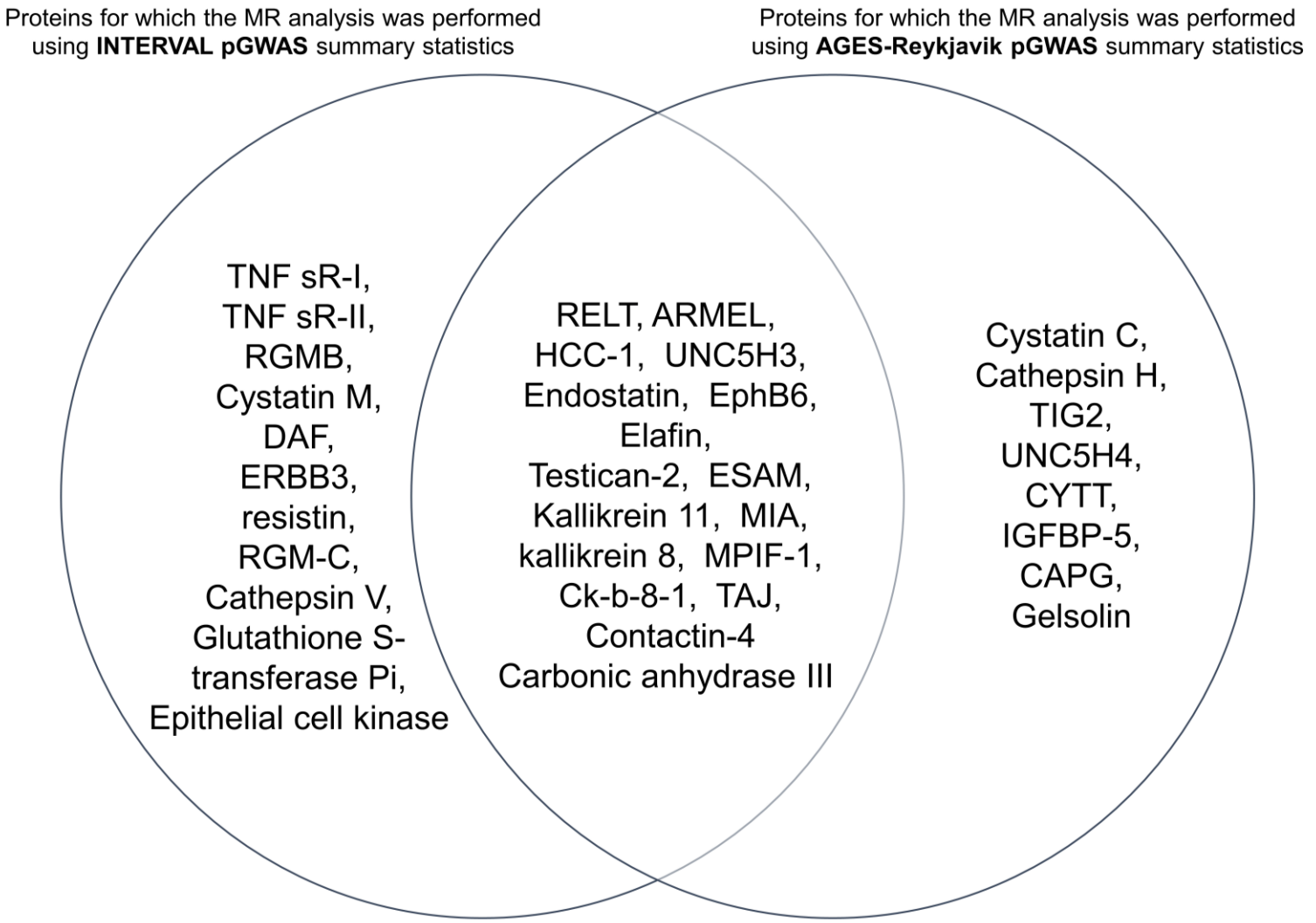
Suppl. Fig. 2. Correlation between plasma proteomic measurements in KORA F4



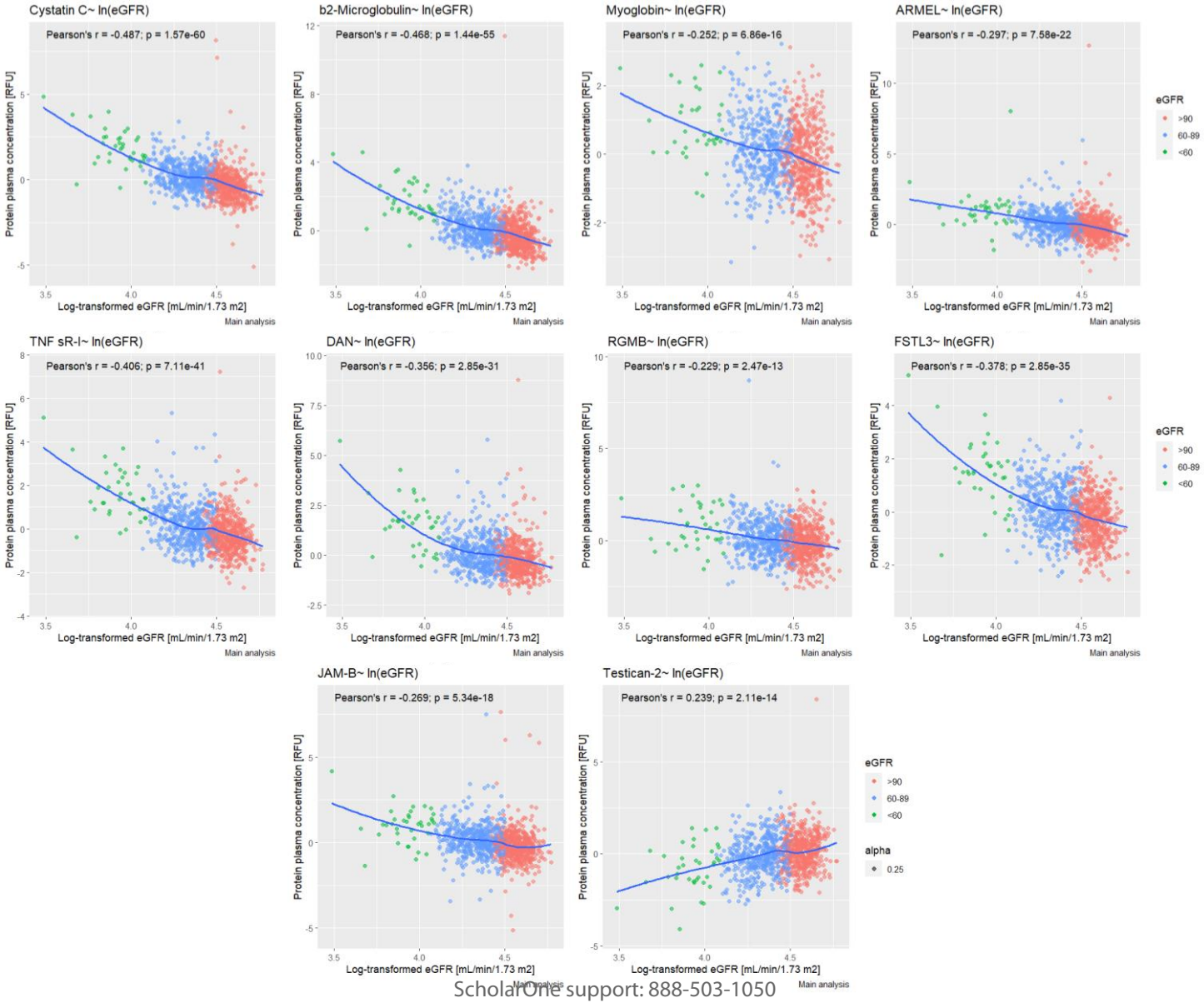
Suppl. Fig. 3. Genetic instrument selection and data harmonization for MR



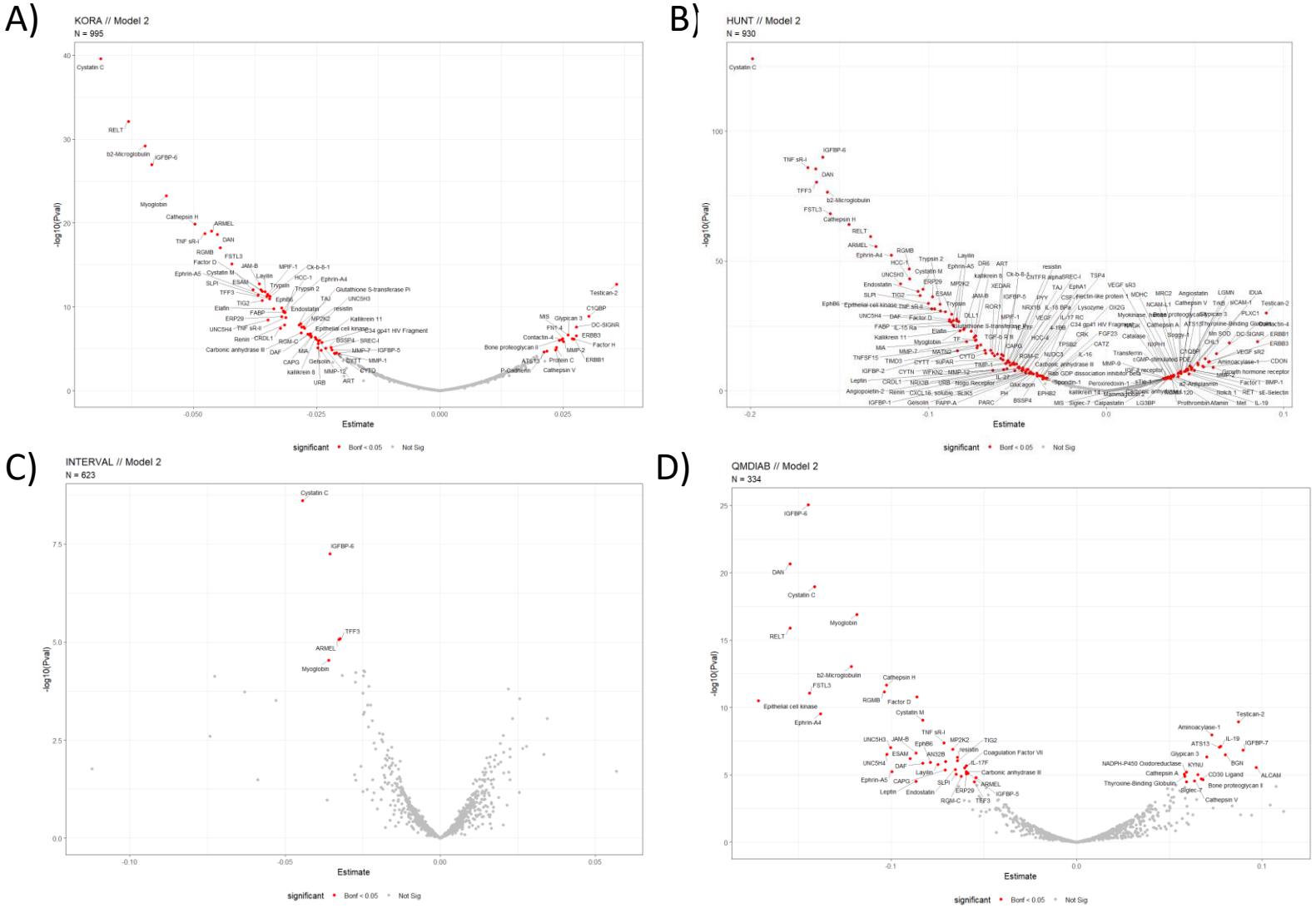
Suppl. Fig. 4. Protein overlap in pGWAS datasets used in reverse direction of MR



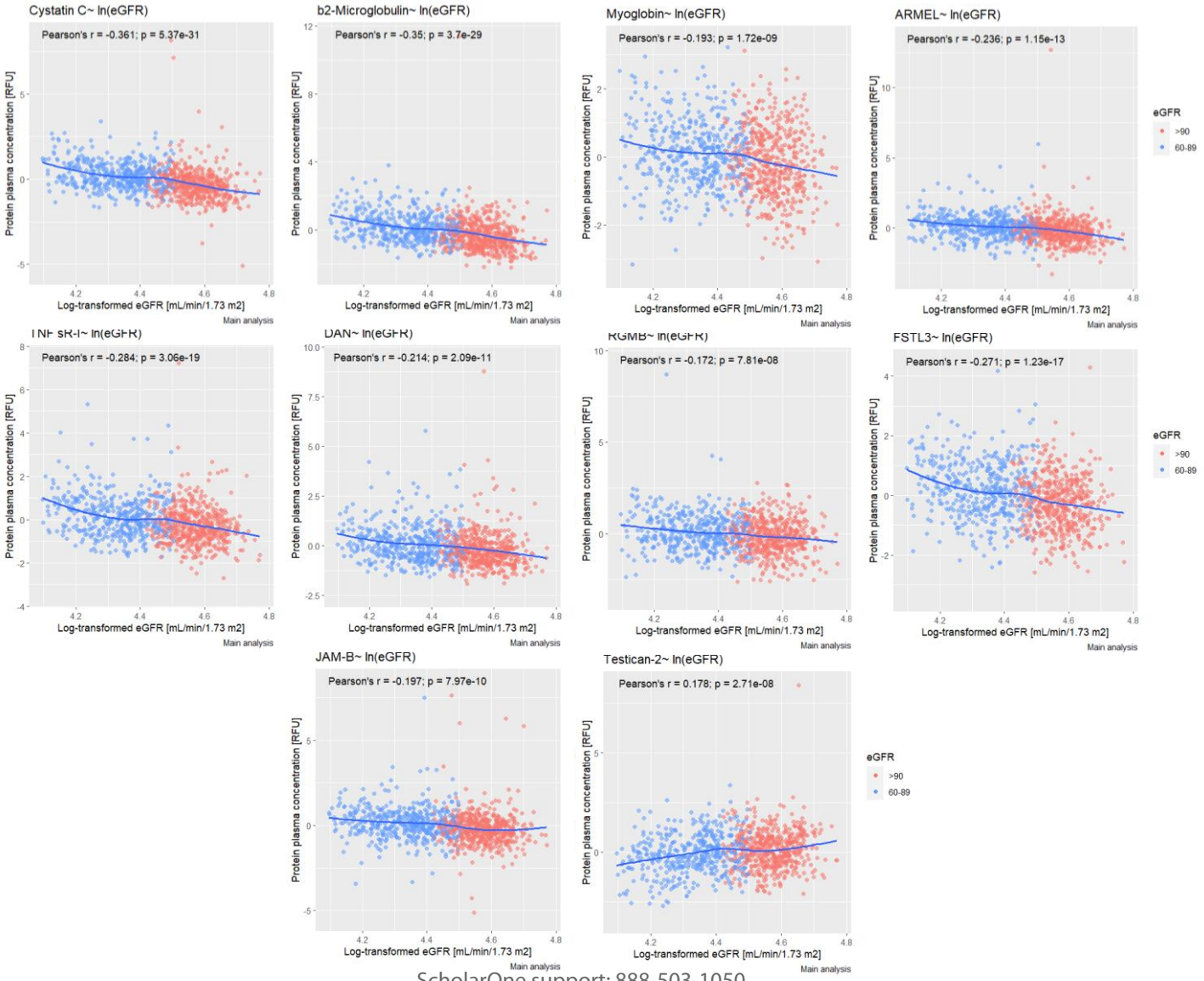
Suppl. Fig. 5. Proteins and log(eGFR) distribution in discovery dataset



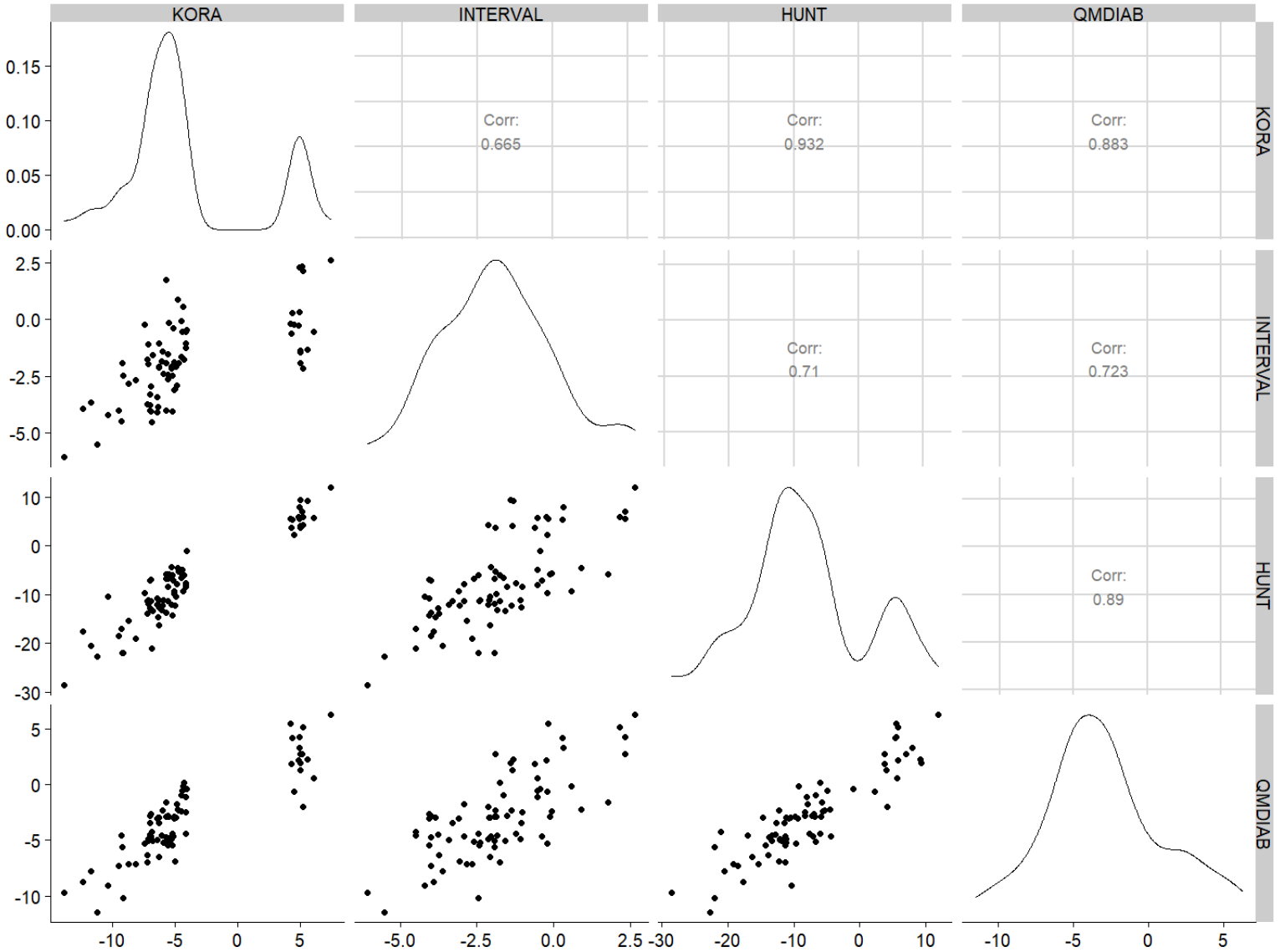
Suppl. Fig. 6. Cross sectional results for eGFR-protein associations across studies



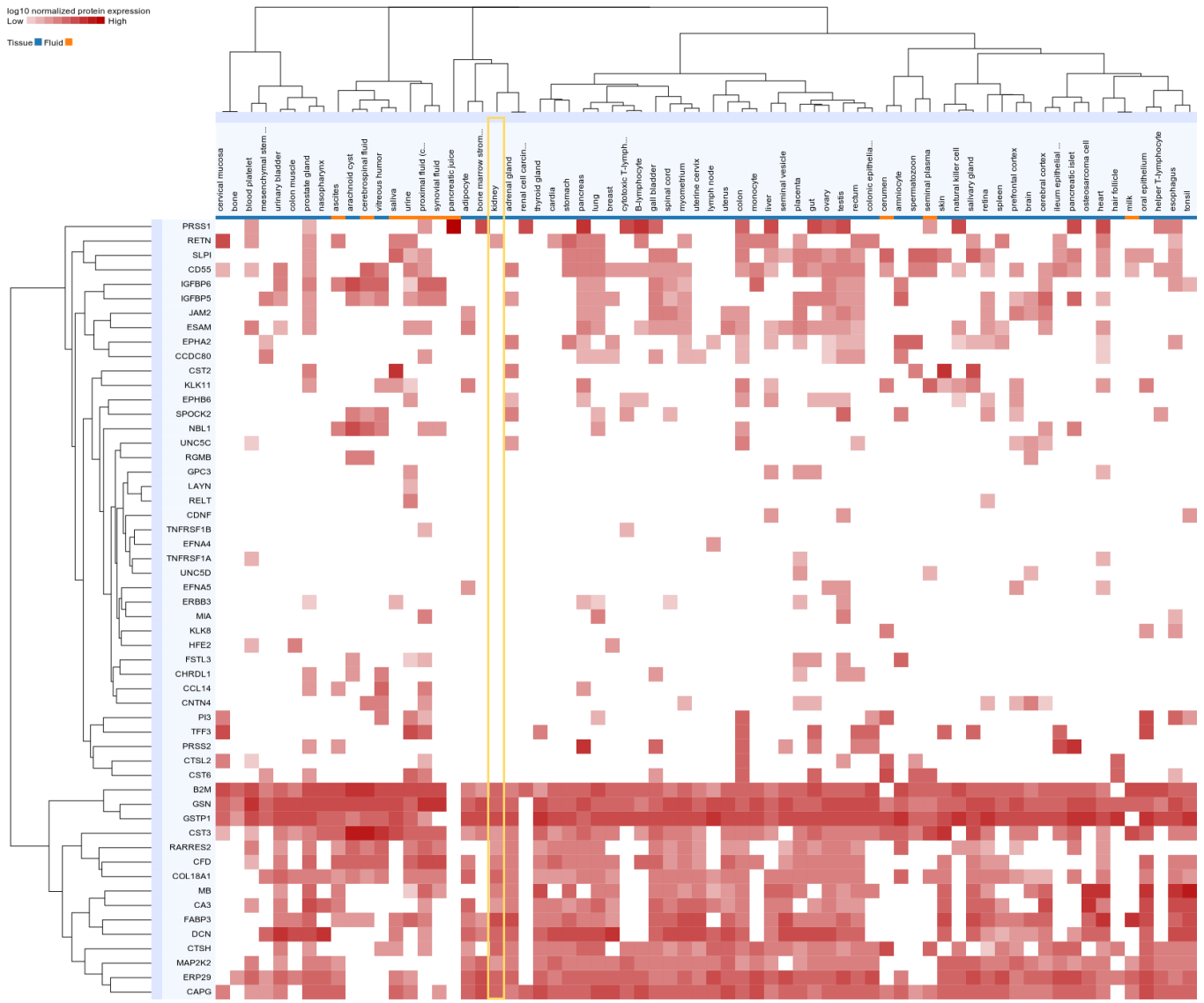
Suppl. Fig. 7. Proteins and log(eGFR) distribution in discovery dataset after CKD exclusion



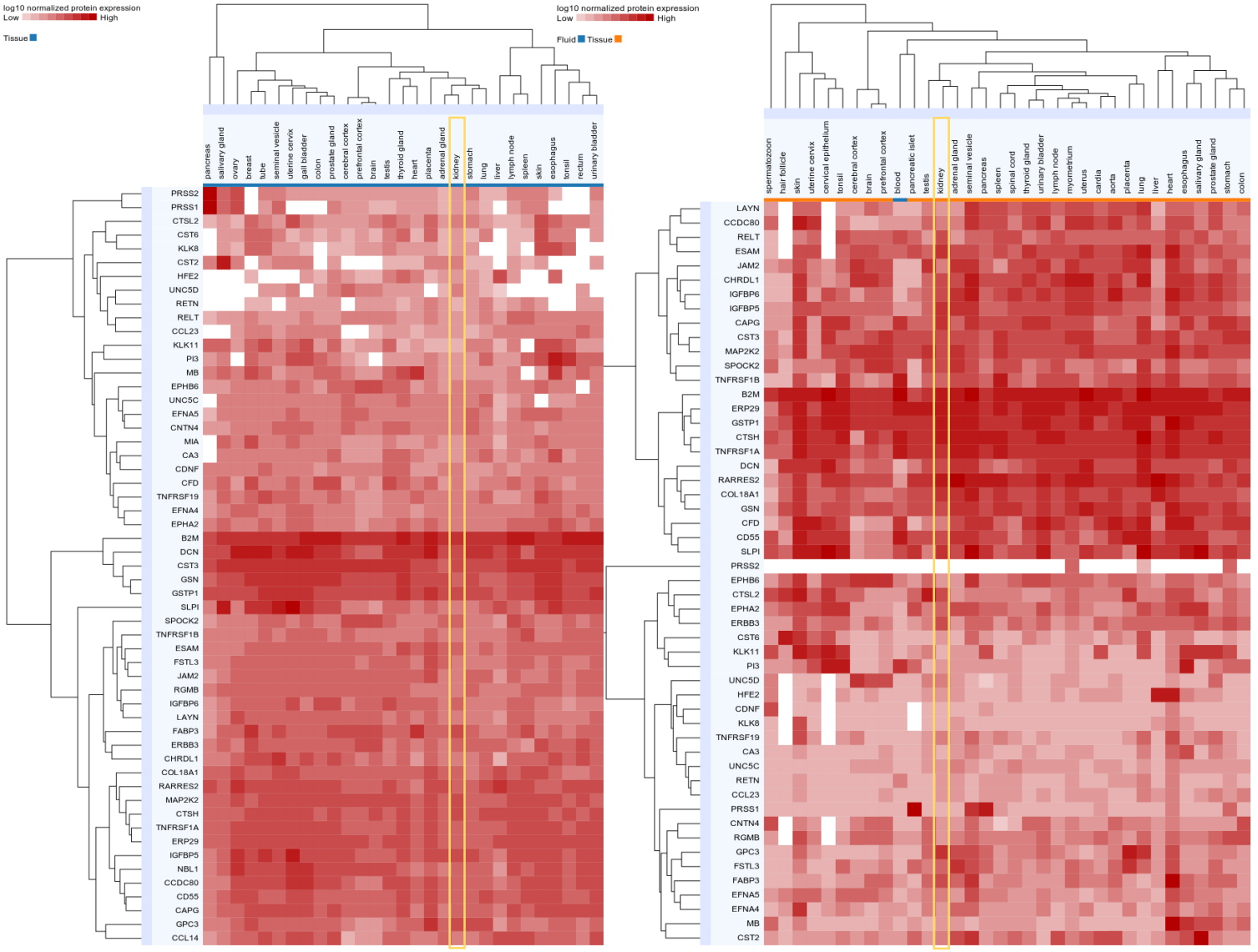
Suppl. Fig. 8. Correlation between Z-values for eGFR-protein associations across studies



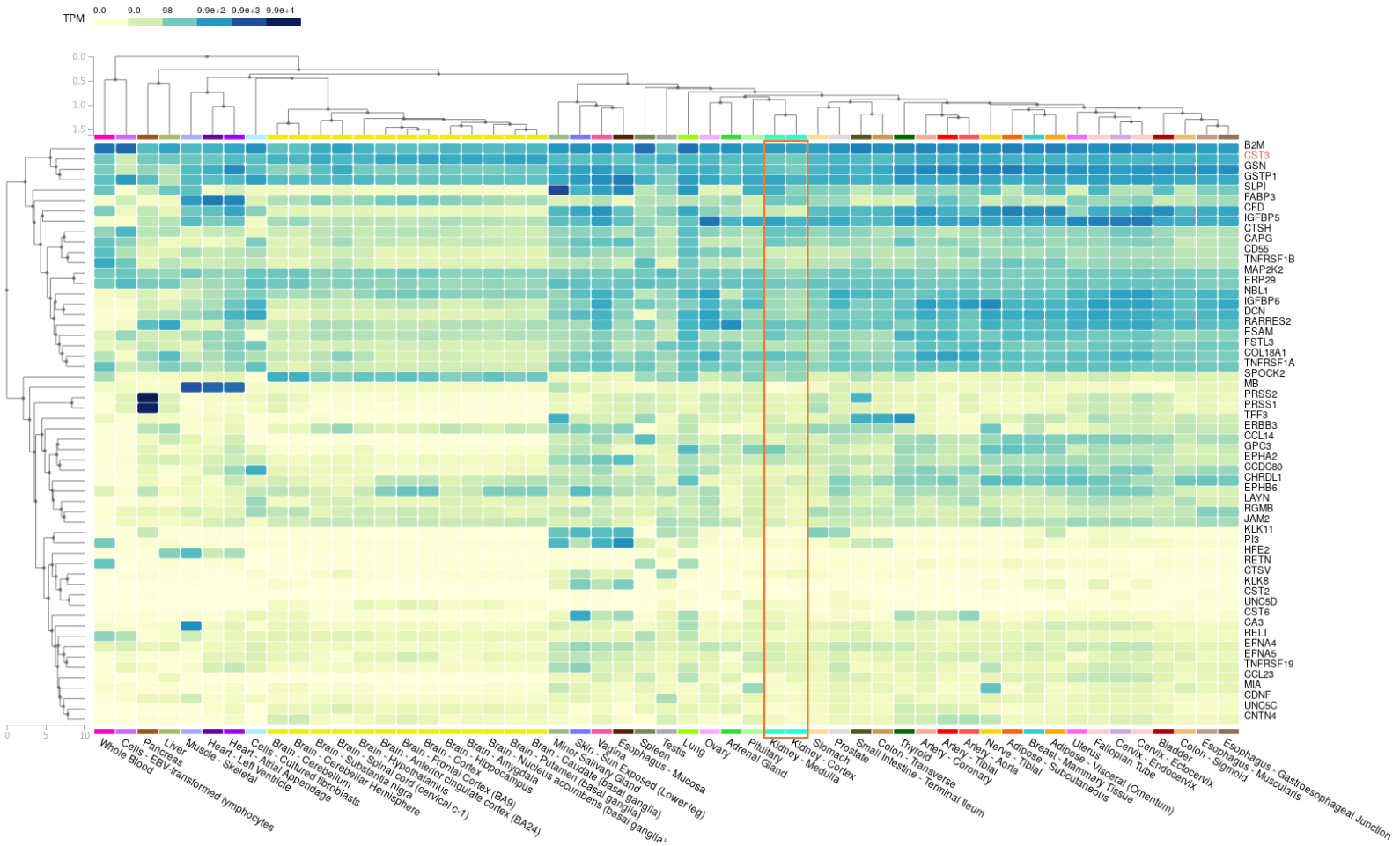
Suppl. Fig. 9. Tissue expression of 56 eGFR-associated proteins (ProteomeDB)



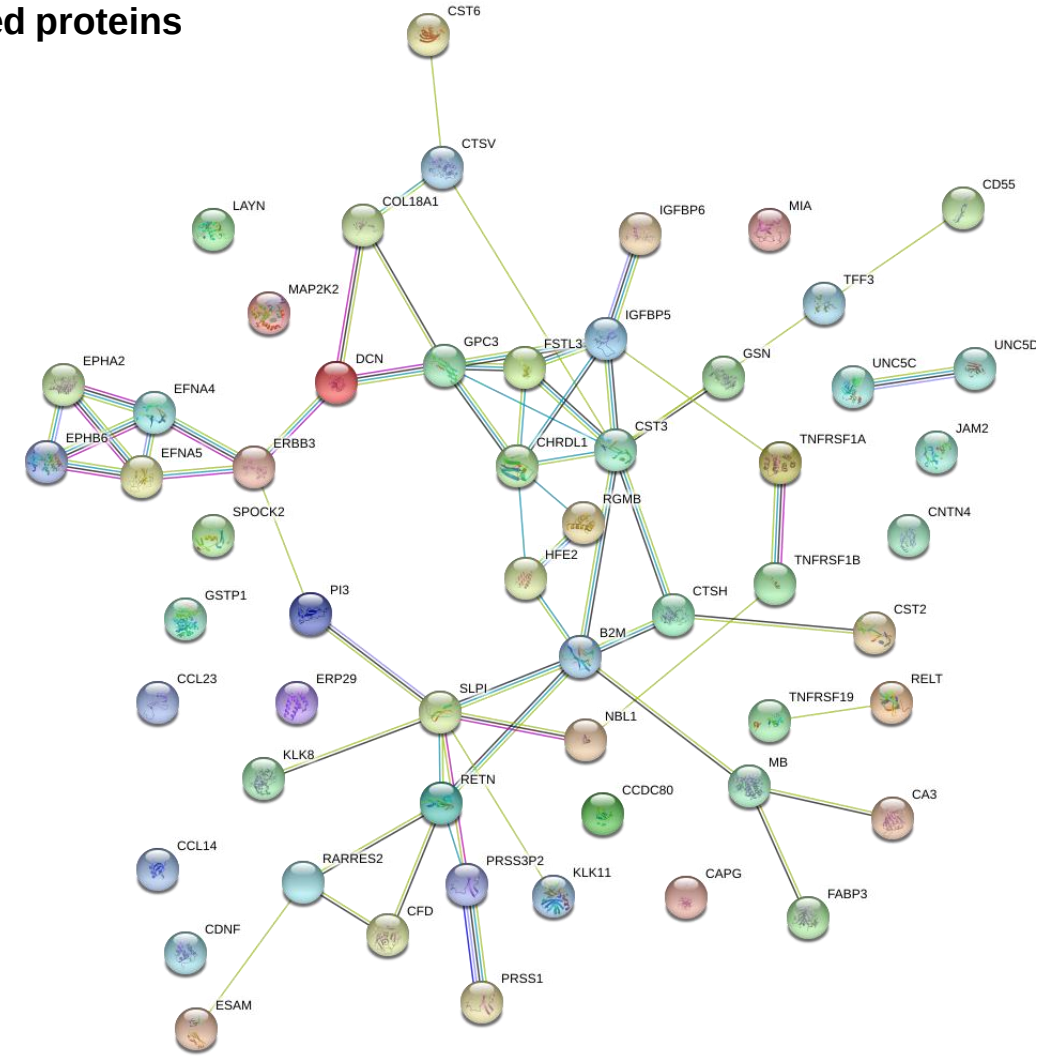
Suppl. Fig. 10. Tissue expression of 56 eGFR-associated protein coding genes (ProteomeDB)



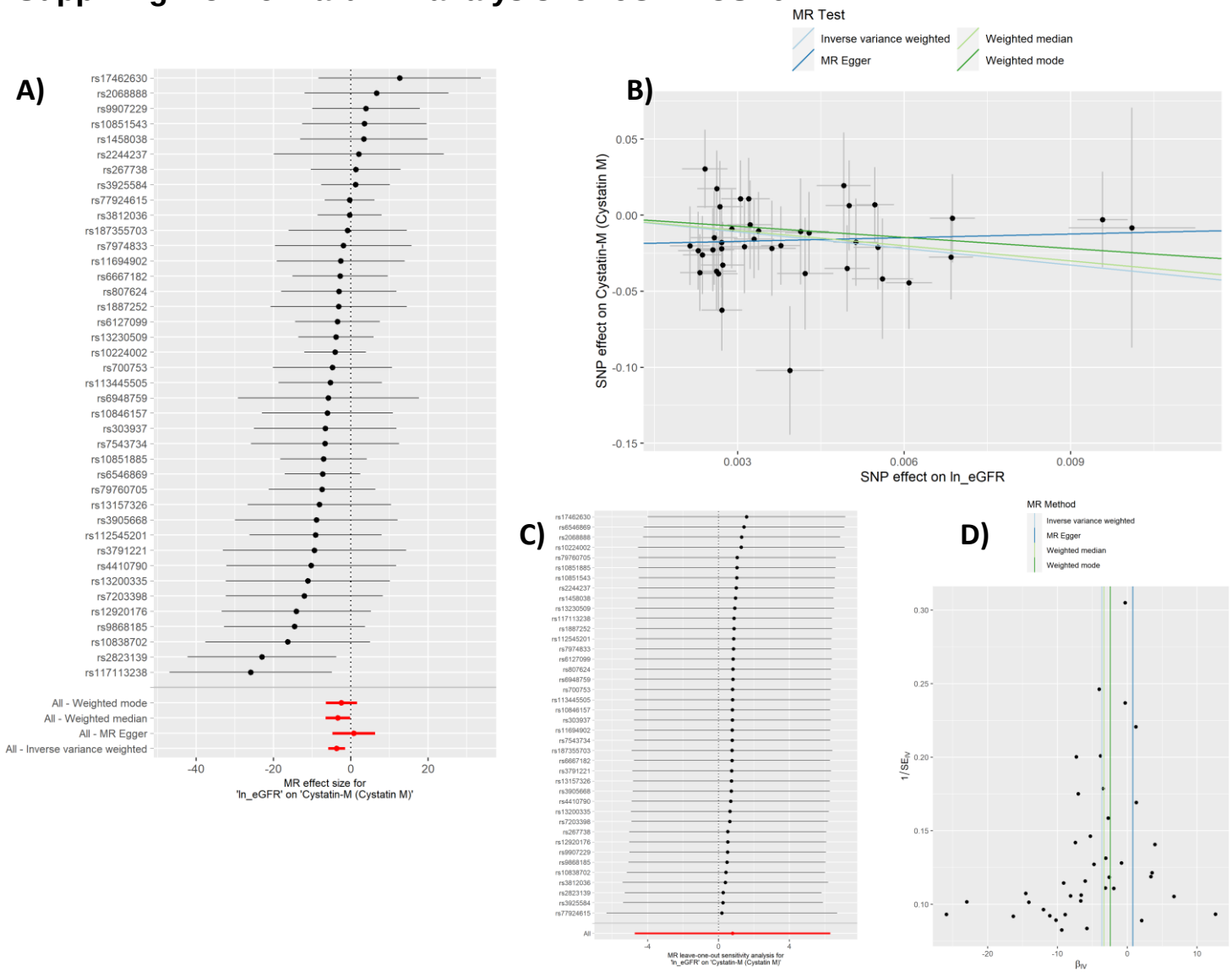
Suppl. Fig. 11. Tissue expression of 56 eGFR-associated proteins (GTEx, transcriptomics)



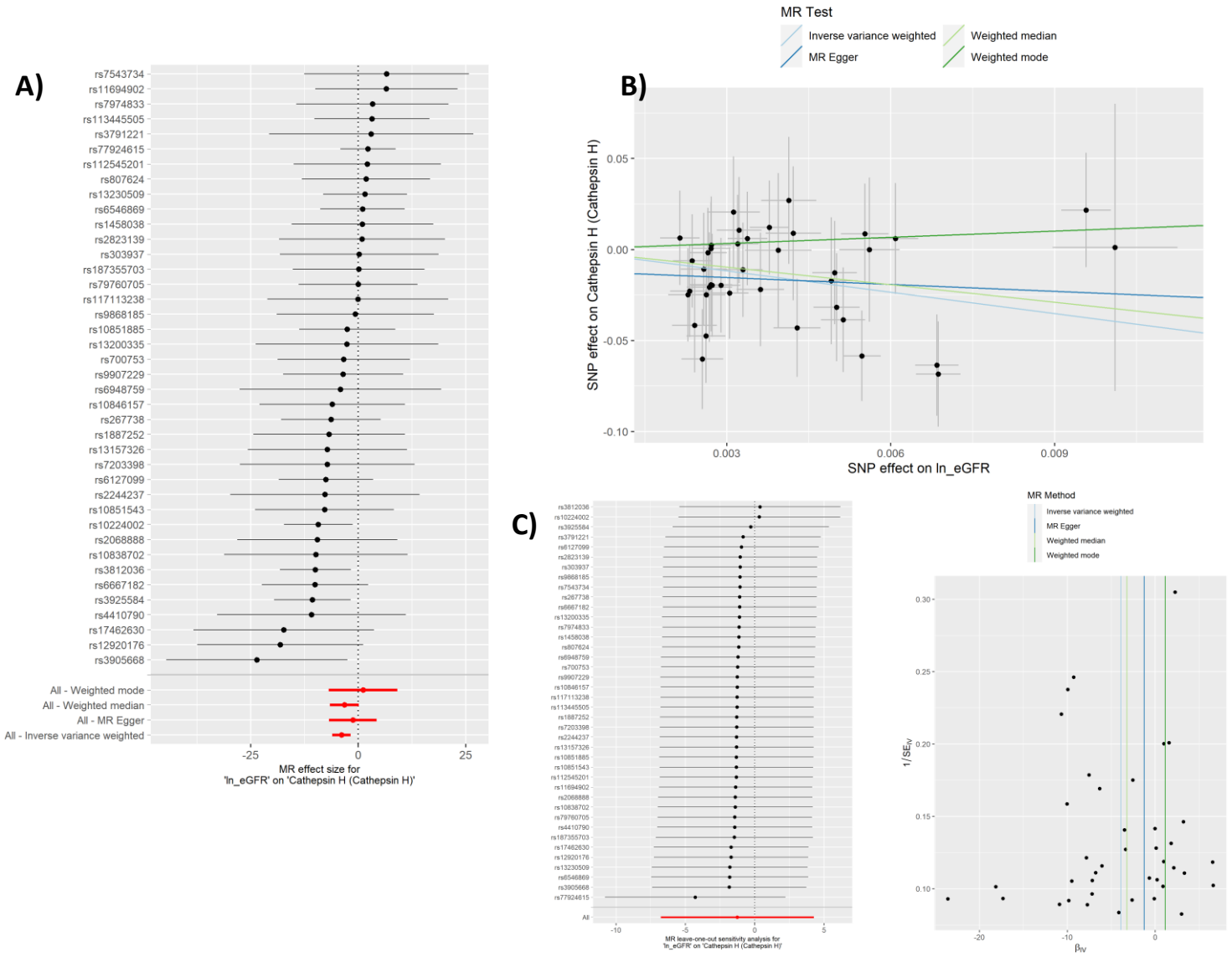
Suppl. Fig. 12. Protein-protein interaction network of 57 replicated eGFR-associated proteins



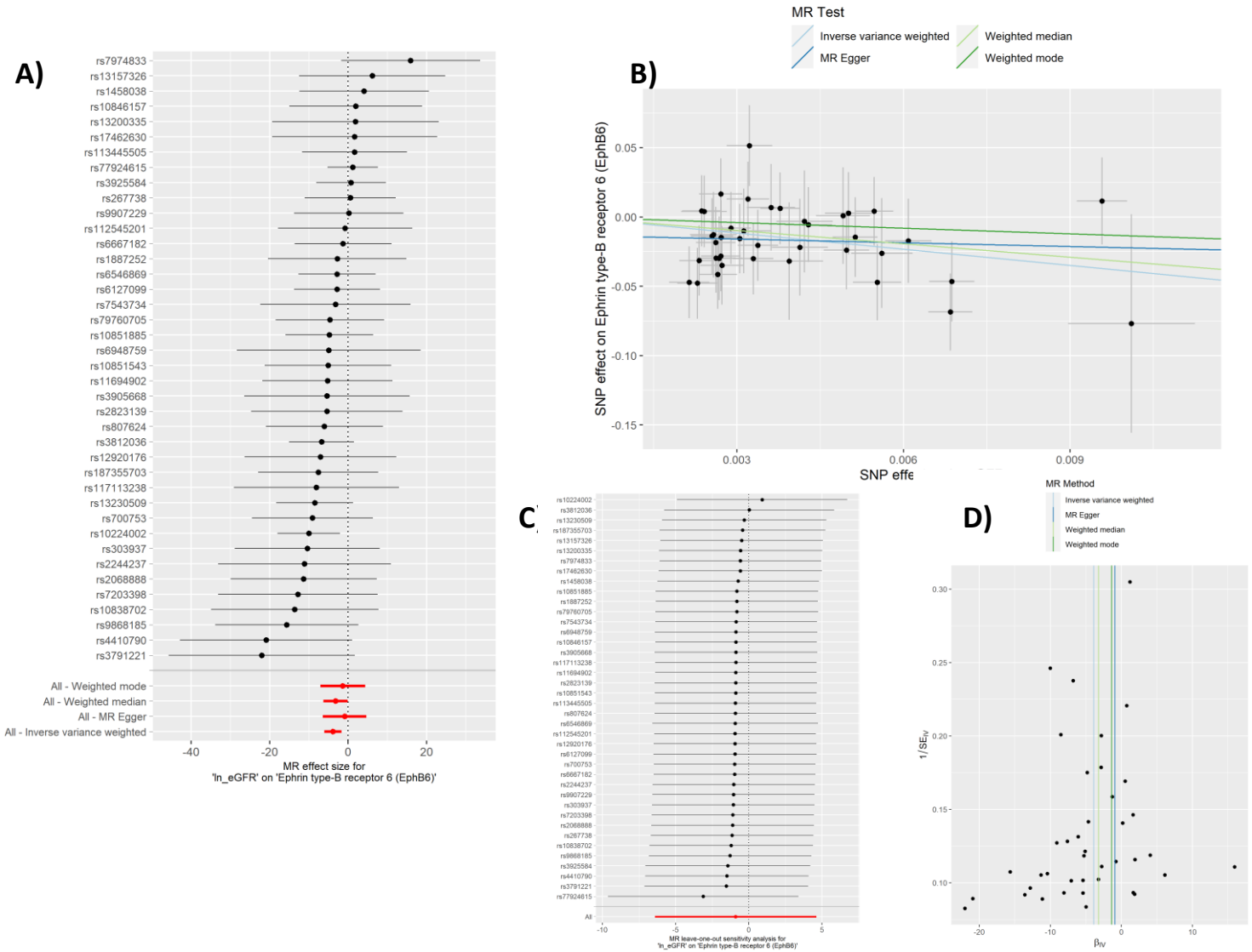
Suppl. Fig. 13. Forward MR analysis for eGFR-CST6



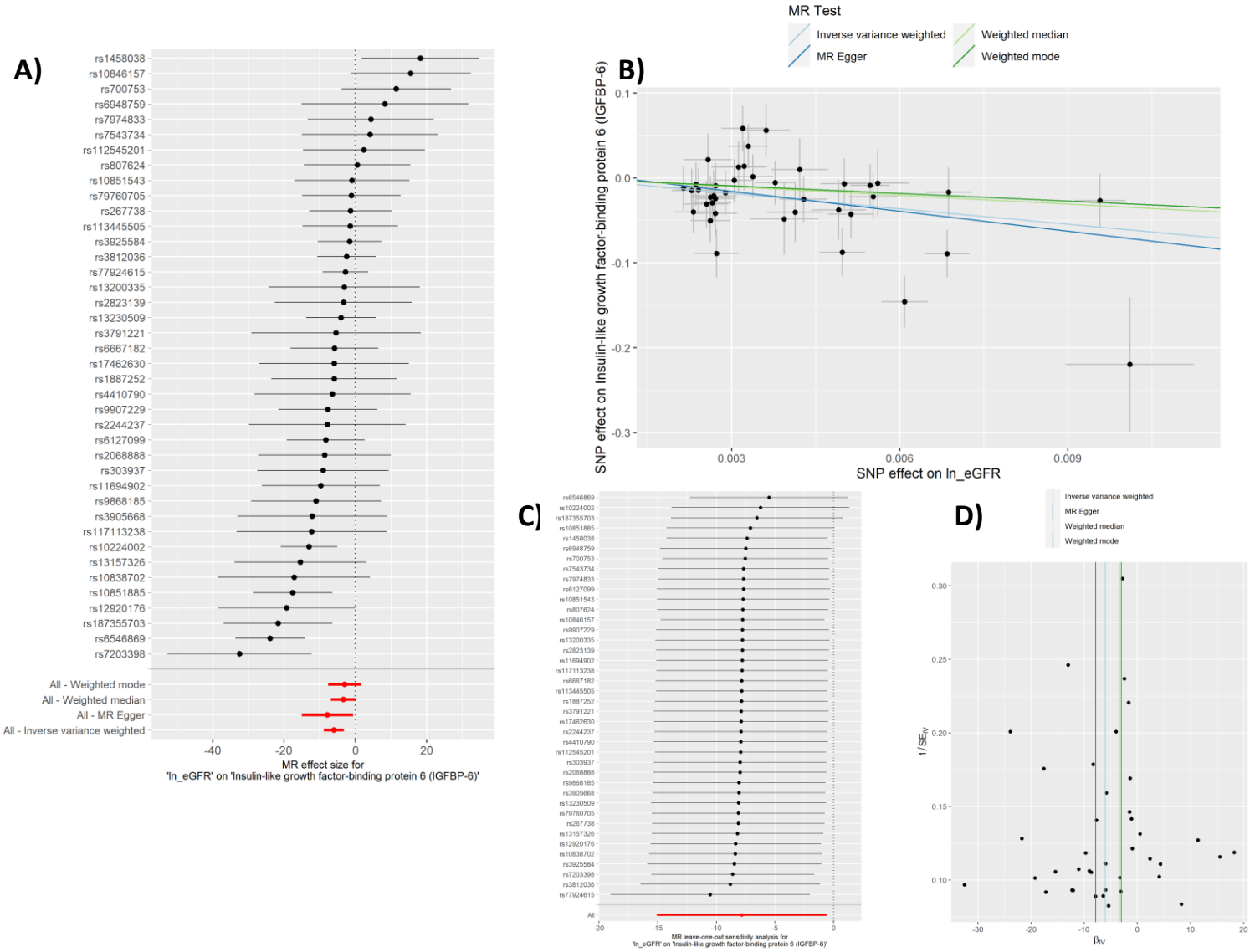
Suppl. Fig. 14. Forward MR analysis for eGFR-Cathepsin H



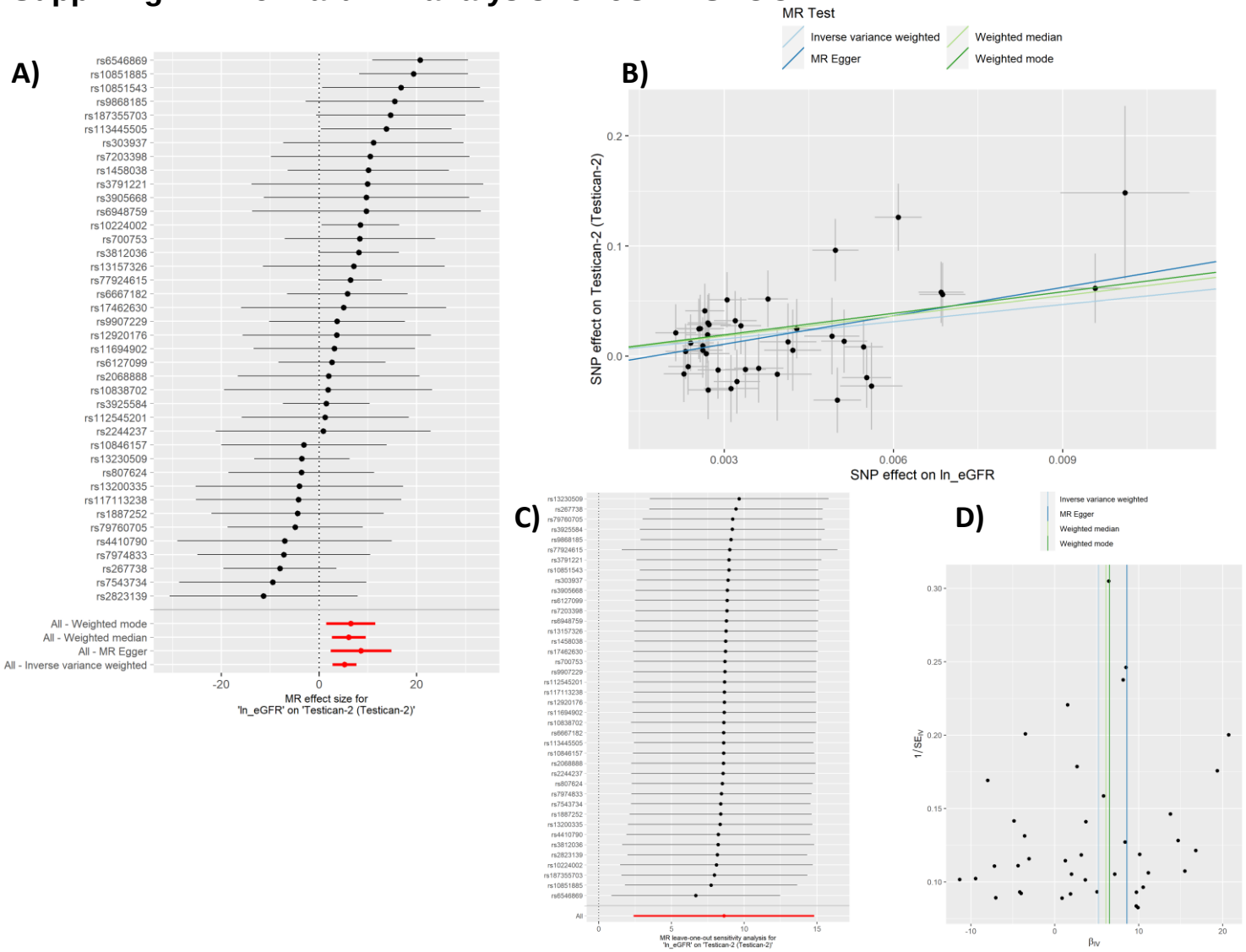
Suppl. Fig. 15. Forward MR analysis for eGFR-EphB6



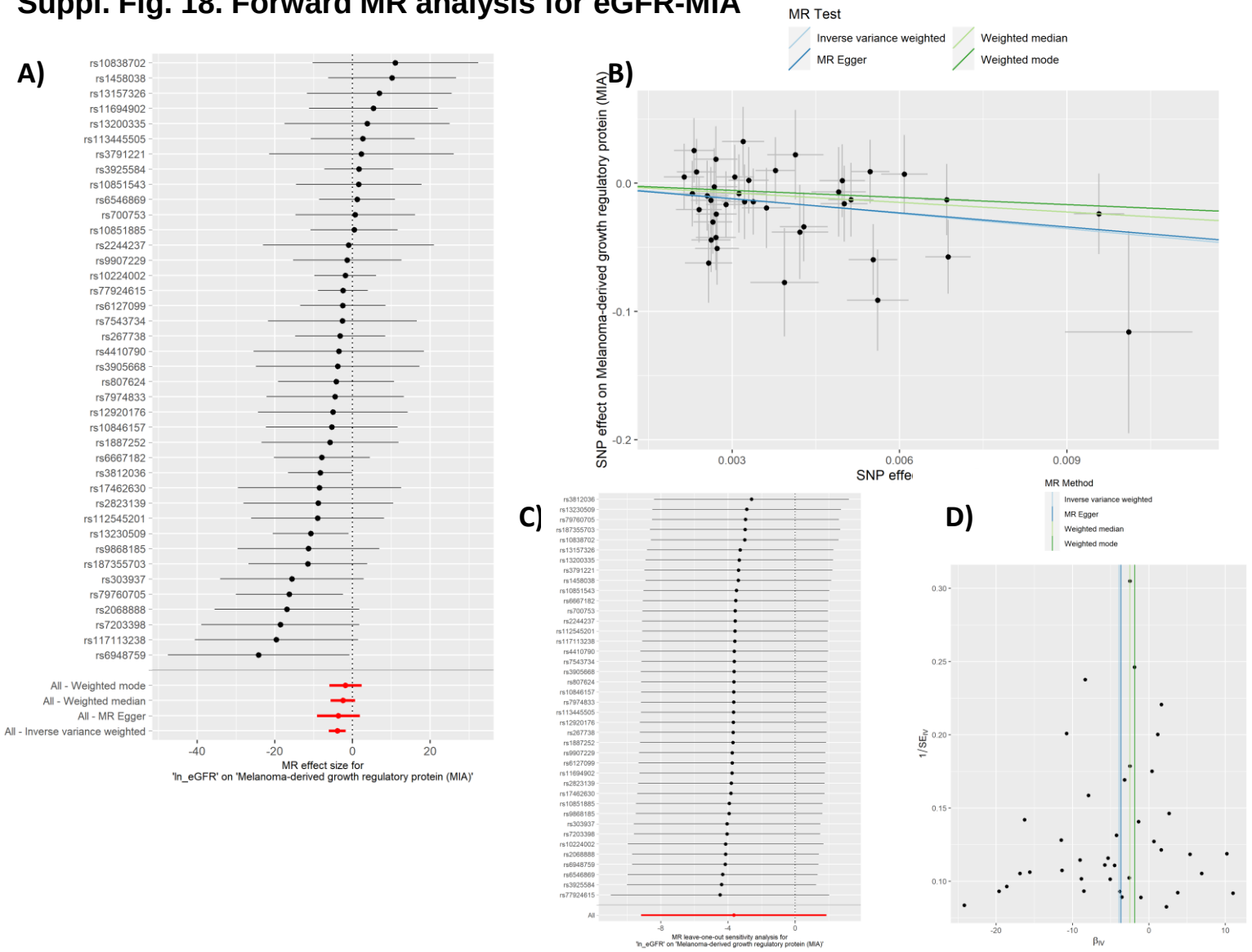
Suppl. Fig. 16. Forward MR analysis for eGFR-IGFBP-6



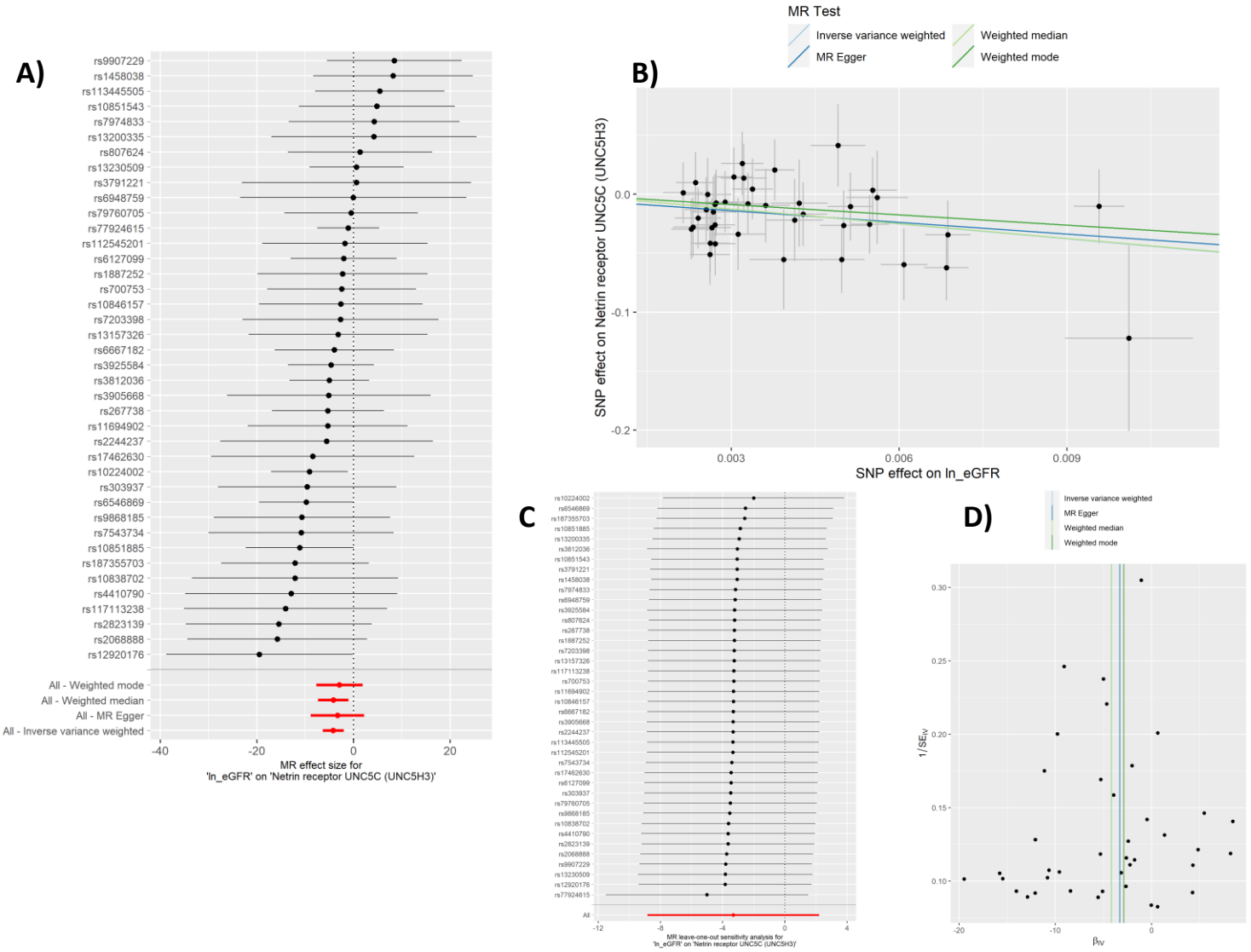
Suppl. Fig. 17. Forward MR analysis for eGFR-SPOCK2



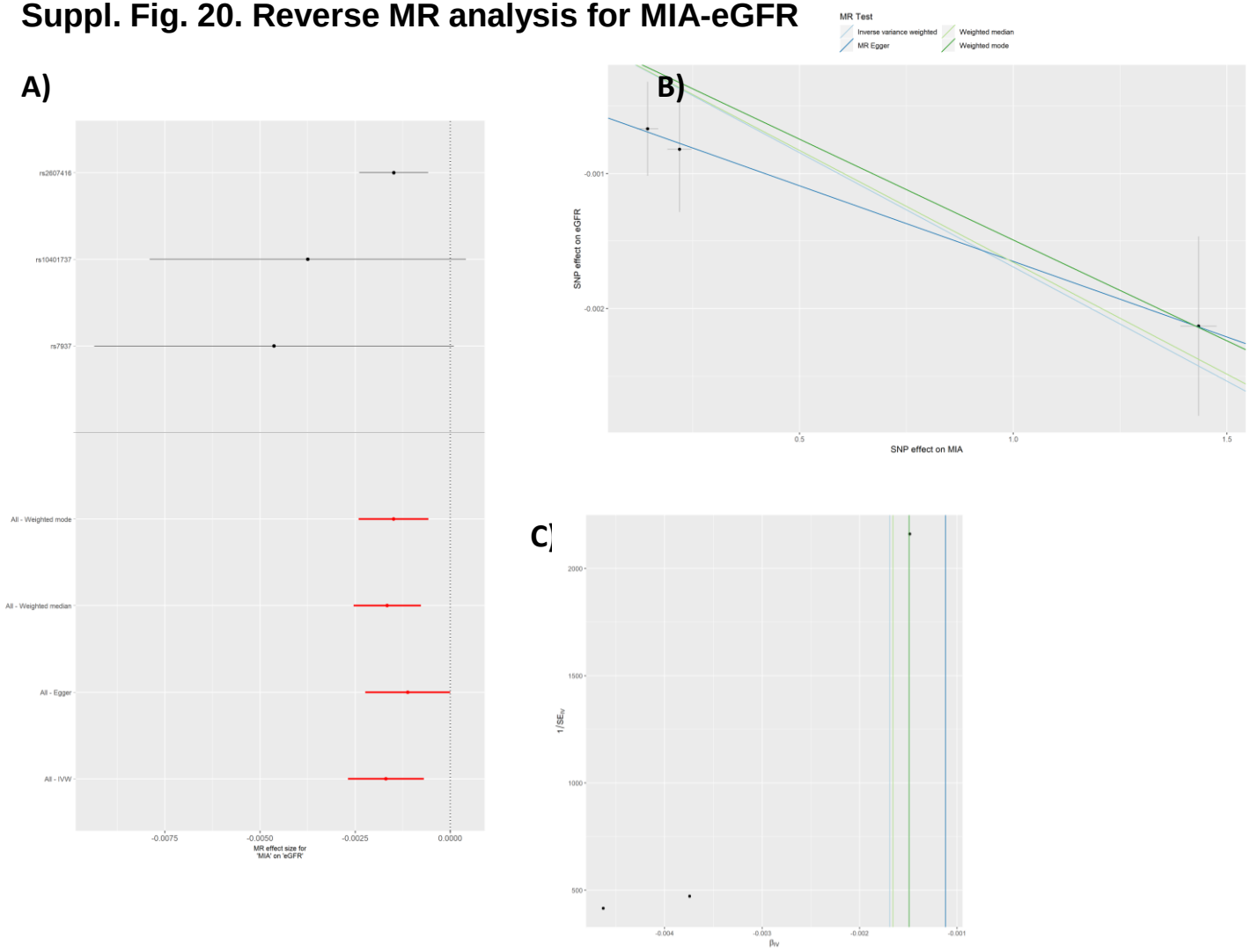
Suppl. Fig. 18. Forward MR analysis for eGFR-MIA



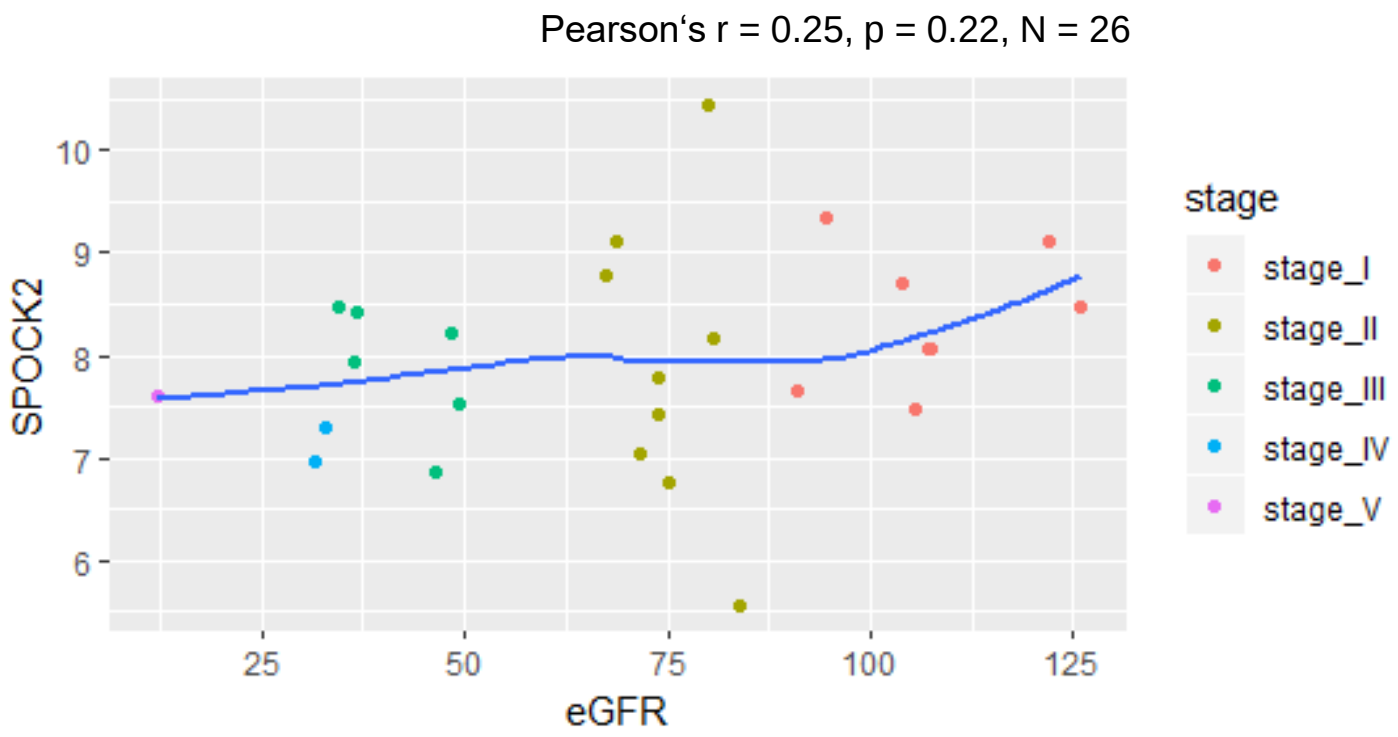
Suppl. Fig. 19. Forward MR analysis for eGFR-UNC5H3



Suppl. Fig. 20. Reverse MR analysis for MIA-eGFR



Suppl. Fig. 21. *SPOCK2* gene expression in renal tissue from 26 CKD patients



Journal of the American Society of NEPHROLOGY

CJASN Clinical Journal of American Society of Nephrology

JASN JOURNAL OF THE AMERICAN SOCIETY OF NEPHROLOGY

Manuscript ID: JASN-2020- 07-1070

Adapted* STROBE Checklist for Observational Studies

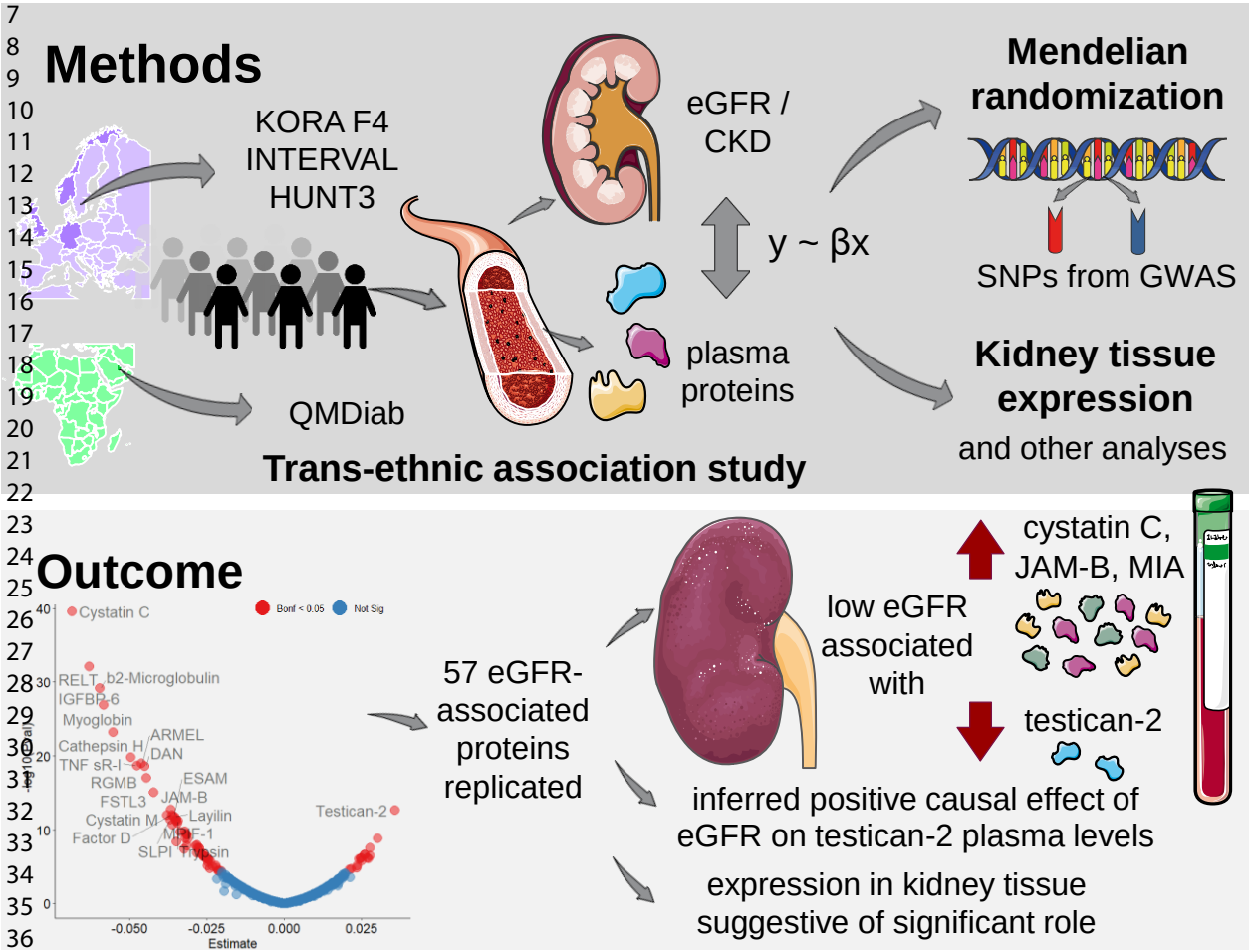
Section/Topic	Checklist Item	Response
Background	Specific objectives clearly stated, including any pre-specified hypotheses.	Yes
METHODS		
Study Design	Key elements of study design presented early in the paper.	Yes
Setting	Setting and location from where study participants drawn clearly described.	Yes
	Relevant dates for periods of recruitment, exposure, follow-up, and data collection.	Yes
Participants	Eligibility criteria or matching criteria, as appropriate, clearly described.	NA
	Sources and methods of selection of participant (or cases and controls) selection.	NA
	Number of exposed/unexposed OR number of controls per case, as appropriate.	Yes
	Methods of follow-up clearly described.	Yes
Variables	Exposures, outcomes, predictors, potential confounders, effect modifiers clearly defined.	Yes
Data Sources/ Measurement	Sources of data and details of methods of assessment.	Yes
	Comparability of assessment methods, if more than one group.	Yes
Bias	Clear description of efforts to address potential sources of bias.	Yes
Statistical Methods	All statistical methods clearly described, including those used to control for confounding.	Yes
	Methods used to examine subgroups and interactions clearly described.	NA
	Extent of missing data, and how it was handled clearly described.	NA
	Extent of loss to follow-up and how it was addressed clearly described.	NA
	Sensitivity Analyses appropriate and clearly described.	Yes
RESULTS		
Participants	Number of individuals at each stage – e.g., number eligible, examined for eligibility, confirmed eligible, included in study, completing follow-up and analyzed.	Yes
	Reasons for non-participation at each stage clearly described.	NA
Descriptive Data	Characteristics of study participants presented .	Yes
	Information on exposures and potential confounders available.	Yes
	Number of participants with missing data for each variable of interest provided.	Yes
	Follow-up time summarized (average and total, as appropriate).	NA
Outcomes	Number of outcome events or summary measures over time available.	Yes
	For case-control study, numbers in each exposure category, or summary measures of exposure provided.	NA
	For case-control study, numbers of outcomes events or summary measures.	NA
Main Results	Unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (95% confidence interval) provided.	Yes
	Category boundaries, when continuous variables were categorized, available.	NA
	Estimates of relative risk translated into absolute risk for a meaningful period.	NA
Other Analyses	Sub-group analyses, interactions, sensitivity analyses adequately presented.	Yes

Additional details:

*This checklist has been adapted by CJASN and JASN for author and editor usage.

Plasma proteomics of renal function: a trans-ethnic meta-analysis and Mendelian randomization study

JASN
JOURNAL OF THE AMERICAN SOCIETY OF NEPHROLOGY



Conclusion

In the largest multiplex plasma proteomics study to date, we identify 57 proteins as trans-ethnically associated with eGFR and/or CKD. Mendelian randomization analysis suggests that several proteins have a causal relationship with kidney function, results which highlight in particular testican-2. This work represents an early milestone in the identification and establishment of sets of proteins that could act as physiological markers of kidney disease progression, biomarkers of potential clinical relevance.

doi: 10.1681/ASN.