Leveraging genomics and proteomics to identify therapeutic targets for COVID-19 and cardiometabolic diseases

Satoshi Yoshiji

Department of Human Genetics, Faculty of Medicine and Health Sciences,
McGill University, Montréal

Kyoto-McGill International Collaborative Program in Genomic Medicine,
Graduate School of Medicine, Kyoto University, Kyoto

September 2023

A thesis submitted to McGill University and Kyoto University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

Table of Contents

Abstract

Obesity is a major risk factor for coronavirus disease 2019 (COVID-19) severity and cardiometabolic diseases, including coronary artery disease (CAD), stroke, and type 2 diabetes. However, the underlying mechanisms through which obesity influences these diseases are not fully understood. By leveraging genomics and proteomics, this thesis investigates how circulating proteins mediate the effect of obesity on COVID-19 severity and cardiometabolic diseases.

First, we evaluated the causal effect of obesity on COVID-19 severity using Mendelian randomization (MR), a causal inference method in genetic epidemiology. Given that body fat mass and fat-free mass are genetically interrelated, we used multivariable MR to discern their independent causal effects. Our findings showed that body fat mass is independently associated with an increased risk of COVID-19 severity.

Second, considering that obesity strongly influences the plasma proteome, we sought to identify circulating proteins that mediate the effect of obesity on COVID-19 severity. We used proteome-wide MR to estimate the causal effect of BMI on circulating protein levels and identified proteins whose plasma levels are influenced by BMI, termed “BMI-driven proteins”. Then, we evaluated the causal effects of the BMI-driven proteins on COVID-19 outcomes, again using MR with cis-acting protein quantitative trait loci (cis-pQTLs). This two-step MR approach found that increased circulating nephronectin (NPNT) levels were associated with an increased risk of critically ill and COVID-19 hospitalization. To ensure the robustness of our findings, we repeated the analyses using body fat percentage and cis-pQTLs from different cohorts, which consistently showed that NPNT partially mediates the effect of obesity on COVID-19 severity. In further follow-up analyses for NPNT, we showed that a specific NPNT isoform drives the effect. In single-cell RNA sequencing of the lung from individuals who died of COVID-19, NPNT was significantly expressed in fibroblasts and alveolar cells. Finally, multivariable MR revealed that decreasing body fat mass and increasing fat-free mass can lower NPNT levels and thus may improve COVID-19 severity—underscoring NPNT’s potential clinical relevance and actionability.

Finally, we expanded this framework to other major complications of obesity: CAD, stroke, and type 2 diabetes. Using the two-step MR approach, followed by colocalization and mediation analysis, we identified seven plasma protein mediators with eight protein-disease associations. Among them, circulating collagen type VI alpha-3 (COL6A3) was strongly increased by BMI and increased the risk of CAD. In follow-up analysis for COL6A3, we evaluated the causal effect of its C- and N-terminal effects on CAD. This domain-aware MR found that the C-terminal fragment of COL6A3, known as endotrophin, mediated the effect. In single-cell RNA sequencing of adipose tissues and coronary arteries, COL6A3 was highly expressed in cell types involved in metabolic dysfunction and fibrosis. Finally, multivariable MR revealed that body fat reduction can lower plasma levels of COL6A3-derived endotrophin and other protein mediators and reduce the risk of cardiometabolic diseases.

In summary, this thesis provides clinically relevant insights into how circulating proteins mediate the effect of obesity on COVID-19 severity and cardiometabolic diseases. This integrative proteogenomics approach prioritizes potential therapeutic targets, including NPNT for COVID-19 and endotrophin for CAD.

Résumé

L'obésité est un facteur de risque majeur pour la gravité de la maladie à coronavirus 2019 (COVID-19) et les maladies cardiométaboliques, notamment les maladies coronariennes (CAD), les accidents vasculaires cérébraux et le diabète de type 2. Toutefois, les mécanismes sous-jacents par lesquels l'obésité influence ces maladies ne sont pas entièrement compris. En exploitant la génomique et la protéomique, cette thèse étudie comment les protéines circulantes médient l'effet de l'obésité sur la gravité de la COVID-19 et les maladies cardiométaboliques.

D'abord, nous avons évalué l'effet causal de l'obésité sur la gravité de la COVID-19 en utilisant la randomisation mendélienne (RM), une méthode d'inférence causale en épidémiologie génétique. Étant donné que la masse grasse et la masse non grasse sont génétiquement liées, nous avons utilisé la RM multivariable pour discerner leurs effets causaux indépendants. Nos résultats montrent que la masse grasse est associée de manière indépendante à un risque accru de gravité de la COVID-19.

Ensuite, considérant que l'obésité influence fortement le protéome plasmatique, nous avons cherché à identifier les protéines circulantes qui médient l'effet de l'obésité sur la gravité de la COVID-19. Nous avons utilisé la RM à l'échelle du protéome pour estimer l'effet causal de l'IMC sur les niveaux de protéines circulantes et avons identifié les protéines influencées par l'IMC. Puis, nous avons évalué les effets causaux de ces protéines influencées par l'IMC sur les issues de la COVID-19, à nouveau avec la RM. Nous avons constaté qu'une augmentation des niveaux circulants de néphronectine (NPNT) était associée à un risque accru de gravité de la COVID-19. Nous avons répété les analyses en utilisant le pourcentage de graisse corporelle et les cis-pQTLs de différentes cohortes, montrant de manière cohérente que la NPNT médiait partiellement l'effet de l'obésité sur la gravité de la COVID-19. Lors d'analyses de suivi pour la NPNT, nous avons montré qu'une isoforme spécifique de la NPNT était à l'origine de cette association. Dans le séquençage de l'ARN d'une seule cellule du poumon de personnes décédées de la COVID-19, la NPNT était exprimée de manière significative dans les fibroblastes et les cellules alvéolaires. Finalement, la RM multivariable a révélé que la

réduction de la masse grasse et l'augmentation de la masse non grasse peuvent abaisser les niveaux de NPNT et ainsi améliorer la gravité de la COVID-19, soulignant la pertinence clinique potentielle de la NPNT.

Enfin, nous avons étendu ce cadre aux maladies cardiométaboliques. En utilisant l'approche RM en deux étapes, suivie d'une analyse de colocalisation et de médiation, nous avons identifié sept médiateurs protéiques plasmatiques avec huit associations protéine-maladie. Parmi eux, le collagène de type VI alpha-3 (COL6A3) était fortement augmenté par l'IMC et augmentait le risque de CAD. Dans l'analyse de suivi pour le COL6A3, nous avons évalué l'effet causal de ses effets C- et N-terminaux sur la CAD. Cette RM axée sur le domaine a montré que le fragment C-terminal de COL6A3, connu sous le nom d'endotrophine, médiait cet effet. Dans le séquençage de l'ARN d'une seule cellule des tissus adipeux et des artères coronaires, COL6A3 était fortement exprimé dans les types de cellules associés au dysfonctionnement métabolique et à la fibrose. Finalement, la RM multivariable a montré que la réduction de la masse grasse peut diminuer les niveaux plasmatiques d'endotrophine dérivée de COL6A3 et d'autres médiateurs protéiques et réduire le risque de maladies cardiométaboliques.

En résumé, cette thèse offre des informations cliniquement pertinentes sur la manière dont les protéines circulantes médient l'effet de l'obésité sur la gravité de la COVID-19 et les maladies cardiométaboliques. Cette approche protéogénomique intégrative priorise les cibles thérapeutiques potentielles, y compris la NPNT pour la COVID-19 et l'endotrophine pour la CAD.

List of Abbreviations

ACD: Acid citrate dextrose

BMI: Body mass index

BMP1: Bone morphogenetic protein 1

CADD: Combined Annotation Dependent Depletion

CAD: Coronary artery disease

cis-pQTL: cis-acting protein quantitative trait loci

COL6A3: Collagen type VI α3

COPD: Chronic obstructive lung disease

COVID-19: Coronavirus disease 2019

DXA: Dual-energy X-ray absorptiometry

eQTL: Expression quantitative trait loci

FEV1: Forced expiratory volume

FVC: Forced vital capacity

F11: coagulation factor XI

GWAS: Genome-wide association studies

HSD17B14: Hydroxysteroid 17-beta dehydrogenase 14

INFO: Imputation quality

InSIDE: Instrument Strength Independent of Direct Effect

LD: Linkage disequilibrium

MHC: The human major histocompatibility complex

MMP14: Matrix metallopeptidase 14

MR: Mendelian randomization

NPNT: Nephronectin

OR: Odds ratio

pQTL: Protein quantitative trait loci

PP: Posterior probability

RCTs: Randomized controlled trials

SARS-CoV-2: Severe acute respiratory syndrome coronavirus 2

SD: Standard deviation

SE: Standard error

SNPs: Single-nucleotide polymorphisms

sQTL: Splicing quantitative trait loci

trans-pQTL: Trans-acting protein quantitative trait loci

List of Figures

Chapter 2

Chapter 3

Chapter 4

List of Tables

Chapter 2

Format of the Thesis

This thesis adopts the manuscript-based format outlined in the Thesis Preparation Guidelines provided by the Department of Graduate and Postdoctoral Studies. It encompasses 6 chapters. Chapter 1 serves as an Introduction. Chapter 2 has been published in *Frontiers in Endocrinology*. Chapter 3 has been published in *Nature Metabolism*. Chapter 4 has been posted on *medRxiv*. Chapter 5 discusses the findings of Chapters 2–4. Chapter 6 summarizes the thesis work and discusses the future directions. The summary of the author's significant contributions is provided in the Appendices.

Acknowledgments

I would like to sincerely thank my mentors, collaborators, friends, and family for their unwavering support throughout my PhD. This doctoral thesis would not have been possible without them.

First, I would like to express my gratitude to my supervisor, Dr. Brent Richards. Your generous support and guidance have been invaluable. Your kindness, passion, charisma, and attitude to enjoy science have been truly inspirational. Despite the challenges posed by the pandemic, the onboarding process to your lab was smooth, thanks to the warmth and support from you and the team. Every weekly meeting with you has been an enlightening experience. Our shared values as clinician-scientists specializing in endocrinology have meant that what excites you invariably excites me. Your guidance has solidified my interest in using human genetics to elucidate the underlying mechanisms of human diseases and to identify drug targets—with the ultimate goal of transforming clinical care. Thanks to your unwavering support, I have had the opportunity to collaborate with exceptional scientists worldwide and be at the forefront of cutting-edge science. I am eager to continue collaborating to establish a world-class multi-omics biobank in Montréal, aiming to make a difference in science and clinical care.

I am also deeply grateful to my co-supervisor, Dr. Nobuya Inagaki. Your unwavering support, which began during my residency at Kyoto University Hospital and continued through my endocrinology fellowship and graduate school, has been instrumental. You have consistently encouraged my scientific endeavors and nurtured my early career as a clinician-scientist. Being a part of the Japan Diabetes Society's nationwide monogenic diabetes project under your leadership has been a great honor. The knowledge and insights I have gained from you will undoubtedly shape how I guide my future students. I sincerely thank you for your steadfast support and wisdom.

I would like to thank Dr. Fumihiko Matsuda, Director of the Center for Genomic Medicine at Kyoto University. Enrolling in the Kyoto-McGill International Collaborative Program in

Genomic Medicine, a joint PhD initiative you co-created with Dr. Mark Lathrop, was a transformative experience. This invaluable program was pivotal in starting my career as a human geneticist. I am deeply thankful for your guidance and mentorship.

I would like to express my gratitude to Dr. Vincent Mooser, the Canada Excellence Research Chair in Genomic Medicine at McGill University. Collaborating with you has been enlightening, and I eagerly anticipate our continued work on the BioPortal program.

I extend a heartfelt thanks to all members of the Richards lab. I appreciate unwavering and warm support from Tianyuan, Guillaume, Yiheng, Kevin, Chen-Yang, Yann, Julian, Tomoko, Vince, Yossi, Darin, Dave, Laetitia, Mariana, and Zaman. My time with the lab members has made Montreal a home away from home. I am also grateful to the lab members at Kyoto University and Kitano Hospital, especially Dr. Yorihiro Iwasaki and Dr. Akihiro Hamasaki, for their invaluable supervision and consistent support.

Special thanks to my supervisory committee members, Dr. Celia Greenwood and Dr. Sirui Zhou, and to my external examiner, Dr. Aaron Leong. Celia's insightful lectures have been enriching, and Sirui’s OAS1 paper influenced my interest in proteogenomics.

I extend my gratitude to my collaborators and peers from around the globe: Dr. Julia Carrasco-Zanini-Sanchez, Dr. Shidong Wang, Dr. Takayoshi Sasako, Dr. Hugo Zeberg, Dr. Richard Ågren, Dr. Michael Hultström, Dr. Mitchell Machiela, Dr. Nicholas J. Timpson, Dr. Claudia Langenberg, Dr. Hans Markus Münter, Dr. Marc Afilalo, Dr. Jonathan Afilalo, Dr. Nicholas J Timpson, and Dr. Jason Flannick.

To my cherished family, thank you. Your enduring support have been the foundation of my journey. The time spent with you have been both revitalizing and empowering.

I wish to acknowledge the institutions that have generously funded my research. The unwavering support from the Japan Society for the Promotion of Science, McGill

University, Kyoto University, and the Lady Davis Institute has been indispensable. The scholarships and awards I received from these esteemed institutions allowed me to immerse myself fully in my research.

Contribution to original knowledge

This thesis employs an integrative proteogenomics approach to understand how circulating proteins mediate the impact of obesity on coronavirus disease 2019 (COVID-19) severity and cardiometabolic diseases. By combining genetic epidemiology approaches such as Mendelian randomization (MR) and colocalization with large-scale genomics and proteomics data, we evaluate the clinical relevance of these protein mediators and prioritize potential therapeutic targets, showcasing the power of human genetics to support therapeutic target discovery.

Chapter 2 is titled “Causal associations between body fat accumulation and COVID-19 severity: A Mendelian randomization study”. Previous studies have reported associations between body mass index (BMI) and the severity of COVID-19. However, since BMI is solely a function of height and weight, it does not differentiate between body fat mass and lean mass. As a result, it remains uncertain whether body fat, fat-free mass, or both modulate the association between obesity and COVID-19 severity. Using MR, this chapter aims to dissect the independent causal associations of adipose tissue mass and lean mass with COVID-19 severity. Our analyses demonstrated that an increase in body fat is associated with an increased risk of severe COVID-19 outcomes. Given the genetic correlation between body fat and lean mass, we further refined our investigation using multivariable MR to delineate their independent causal effects. The results showed that body fat mass is independently associated with an increased risk of COVID-19 severity, suggesting that body fat accumulation is an independent risk factor.

Chapter 3 is titled “Proteome-wide Mendelian randomization implicates nephronectin as an actionable mediator of the effect of obesity on COVID-19 severity”. There have been recent efforts to harness proteogenomics in combination with MR to illuminate the causal biology and identify potential therapeutic targets. However, the potential of these methods to discern circulating mediators in a proteome-wide manner remains largely unexplored. Given the strong influence of obesity on the plasma proteome, we aimed to

pinpoint circulating proteins that mediate obesity's impact on COVID-19 severity. Identifying these proteins is valuable, as circulating proteins are easily measurable and, in some instances, modifiable, offering potential therapeutic targets. In this chapter, we first conducted a comprehensive screening of 4,907 plasma proteins, utilizing MR to identify those affected by BMI. This process revealed 1,216 proteins whose plasma levels were influenced by BMI. We subsequently assessed their influence on COVID-19 severity, again employing MR. Through this two-step MR methodology, we determined that a standard deviation increase in nephronectin (NPNT) was associated with an increased risk of COVID-19 severity outcomes. This effect was attributed to an NPNT splice isoform. Subsequent mediation analyses confirmed NPNT's role as a mediator. To ensure the robustness of the findings, we repeated the analysis using body fat percentage and cis-acting protein quantitative loci from different cohorts, further affirming NPNT's mediating role. Single-cell RNA sequencing revealed *NPNT* expression in the alveolar cells and lung fibroblasts of individuals who died of COVID-19. Lastly, we showed that reducing body fat mass and increasing fat-free mass can decrease plasma NPNT levels. These findings shed light on the underlying mechanism by which obesity influences the risk of COVID-19 severity and prioritize NPNT as a therapeutic target, especially in individuals with obesity.

Chapter 4 is titled “COL6A3-derived endotrophin mediates the effect of obesity on coronary artery disease: an integrative proteogenomics analysis”. This chapter explores whether the two-step MR approach can identify circulating protein mediators for other major complications of obesity, such as coronary artery disease, stroke, and type 2 diabetes. Through an integrated analysis that combined a two-step MR screening of 4,907 plasma proteins, colocalization, and mediation analyses, we identified seven plasma proteins linked with eight protein-disease associations. This includes collagen type VI α3 (COL6A3). The two-step MR approach revealed that an increase in BMI is associated with increased plasma levels of COL6A3, which, in turn, is associated with an elevated risk of CAD. Importantly, the C-terminus of COL6A3 is cleaved to produce endotrophin, which we identified as the mediating factor in CAD risk. Single-cell RNA sequencing of adipose tissues and coronary arteries indicated marked *COL6A3*

expression in cells linked with metabolic dysfunction and fibrosis. Additionally, our analysis suggested that reducing body fat can decrease plasma endotrophin levels derived from COL6A3. Overall, this chapter emphasizes the pivotal role of circulating proteins in the effects of obesity on cardiometabolic diseases and highlights the therapeutic potential of endotrophin.

In conclusion, this thesis provides clinically relevant insights into the role of circulating proteins in mediating the effects of obesity on COVID-19 severity and cardiometabolic diseases. Through an integrative proteogenomics approach, we highlight potential therapeutic candidates such as NPNT for COVID-19 and endotrophin for coronary artery disease. This emphasizes the transformative potential of human genomics, proteomics, and genetic epidemiology in identifying therapeutic targets.

Contributions of authors

Chapter 2 is a manuscript authored by Satoshi Yoshiji, Daisuke Tanaka, Hiroto Minamino, Tianyuan Lu, Guillaume Butler-Laporte, Takaaki Murakami, Yoshihito Fujita, J. Brent Richards, and Nobuya Inagaki. It was published in *Frontiers in Endocrinology* on August 3rd, 2022. Satoshi Yoshiji conceptualized and analyzed the data. Satoshi Yoshiji, Hiroto Minamino, and J. Brent Richards wrote the original draft of the manuscript. J. Brent Richards and Nobuya Inagaki supervised the study. All authors discussed the results and contributed to the final manuscript.

Chapter 3 is a manuscript authored by Satoshi Yoshiji, Guillaume Butler-Laporte, Tianyuan Lu, Julian Daniel Sunday Willett, Chen-Yang Su, Tomoko Nakanishi, David R. Morrison, Yiheng Chen, Kevin Liang, Michael Hultström, Yann Ilboudo, Zaman Afrasiabi, Shanshan Lan, Naomi Duggan, Chantal DeLuca, Mitra Vaezi, Chris Tselios, Xiaoqing Xue, Meriem Bouab, Fangyi Shi, Laetitia Laurent, Hans Markus Münter, Marc Afilalo, Jonathan Afilalo, Vincent Mooser, Nicholas J. Timpson, Hugo Zeberg, Sirui Zhou, Vincenzo Forgetta, Yossi Farjoun, and J. Brent Richards. It was published in *Nature Metabolism* on February 20th, 2023. Conception and design were the responsibility of Satoshi Yoshiji and J. Brent Richards. Methodology was overseen by Satoshi Yoshiji, Tianyuan Lu. and J. Brent Richards. Data analysis was performed by Satoshi Yoshiji, Tianyuan Lu, Chen-Yang Su, Julian Daniel Sunday Willett, and J. Brent Richards. Visualization was conducted by Satoshi Yoshiji and Tianyuan Lu. Writing of the original draft was carried out by Satoshi Yoshiji. All authors critically reviewed and edited the manuscript.

Chapter 4 is a manuscript authored by Satoshi Yoshiji, Tianyuan Lu, Guillaume Butler-Laporte, Julia Carrasco-Zanini-Sanchez, Yiheng Chen, Kevin Liang, Julian Daniel Sunday Willett, Chen-Yang Su, Shidong Wang, Darin Adra, Yann Ilboudo, Takayoshi Sasako, Vincenzo Forgetta, Yossi Farjoun, Hugo Zeberg, Sirui Zhou, Michael Hultström, Mitchell Machiela, Nicholas J. Wareham, Vincent Mooser, Nicholas J. Timpson, Claudia Langenberg, and J. Brent Richards. It was posted on *medRxiv* on April 25th, 2023. It is under review at *Nature Genetics* as of Sep 20th 2023. Conception

and design was the responsibility of Satoshi Yoshiji and J. Brent Richards. Methodology was overseen by Satoshi Yoshiji, Tianyuan Lu and J. Brent Richards. Data analysis was performed by Satoshi Yoshiji, Tianyuan Lu, Julia Carrasco-Zanini-Sanchez, and J. Brent Richards. Visualization was conducted by Satoshi Yoshiji and Tianyuan Lu. Writing of the original draft was carried out by Satoshi Yoshiji. All authors critically reviewed and edited the manuscript.

Chapter 1: Introduction

1.1 Obesity and its complications

Obesity affects more than 1.9 billion individuals worldwide and is now considered a global epidemic. There is mounting evidence that obesity is strongly linked to the risk of severe COVID-19 outcomes as well as various cardiometabolic diseases, including coronary artery disease (CAD), stroke, and type 2 diabetes[1,2]. Furthermore, obesity is associated with a decreased quality of life, reduced life expectancy, and elevated healthcare costs[3]. Thus, understanding the underlying mechanisms by which obesity increases the risk of these diseases and identifying potential therapeutic targets is urgently required to tackle the global obesity crisis.

Obesity cannot be only attributed to an energy imbalance between calorie intake and expenditure; its pathophysiological roots are complex and multifactorial[3]. Several biological mechanisms, such as metabolic abnormalities, oxidative stress, mitochondrial dysfunction, immune disturbances, and chronic low-grade inflammation, are implicated in its pathogenesis[3,4]. Yet, many of these insights are derived from rodent studies or observational analyses; findings in rodent studies do not necessarily translate into human biology, and observational analyses are prone to confounding and reverse causation, making it challenging to distinguish causation from consequences.

Human genetics is increasingly recognized as a valuable tool to uncover causal biology and support drug target discovery[5-7]. Since germline genetic variants are randomly allocated at conception, evidence derived from genetics is less susceptible to confounding and reverse causation[8]. The discovery of disease-causing variants has led to novel targets, such as PCSK9, whose loss-of-function variants were found to significantly lower LDL cholesterol[9], leading to the development of PCSK9 inhibitors[10,11]. Furthermore, there have been extraordinary breakthroughs in high-throughput

sequencing and computational genomics methodologies, facilitating the elucidation of human diseases' molecular genetic basis. Meanwhile, proteomics offers another valuable perspective. Obesity markedly impacts plasma protein levels[12,13], with these proteins playing a substantial role in the development and progression of diseases. Additionally, circulating proteins can be quantified and, in some cases, modulated[14], making them attractive therapeutic targets.

1.2 Obesity and COVID-19

COVID-19 severity varies markedly among patients[15]. This underscores the importance of identifying and understanding the modifiable risk factors, which may enhance public health strategies, optimize resource allocation, and facilitate clinical decisions[16]. Notably, BMI has emerged as an independent risk factor for severe COVID-19 outcomes, which consist of increased hospital admissions, the necessity for invasive mechanical ventilation, and higher mortality rates[17]. This has been corroborated in cohort studies from China, the US, and Europe, with a population-based observational study reporting a nearly two-fold increase in the risk of COVID-19-related death in individuals with a BMI of 40 kg/m^2 or higher[18].

Multiple theories have been suggested to explain the increased risk of severe COVID-19 in individuals with obesity[19]. Obesity has broad effects on pulmonary physiology, adipose tissue biology, metabolism, and immune system function[16,20]. Obesity has been associated with respiratory dysfunction, manifesting as altered respiratory mechanisms, increased airway resistance, decreased gas exchange efficiency, and reductions in lung volume and muscle strength. Such impairments elevate the risk of pneumonia in these individuals, a risk that is exacerbated by associated conditions like hypoventilation, pulmonary hypertension, and cardiac strain[21]. Beyond respiratory dysfunction, obesity is associated with metabolic derangements and immune dysfunction. Immuno-metabolically abnormal adipose tissue may play a role in an exaggerated inflammatory response to SARS-CoV-2 infection, which could underlie the convergence of obesity,

severe systemic inflammation, and poor outcomes[22]. Experimental mouse models and clinical studies have shown immune cell infiltration in adipose tissues, and obesity converges with low-grade, systemic inflammation in increased susceptibility to SARS-CoV-2[23]. Furthermore, obesity is associated with an increased risk of comorbidities such as diabetes mellitus, cardiovascular diseases, and kidney complications. These conditions can jointly contribute to poor outcomes[24]. Given the intertwined nature of these biological systems and their effects, distinguishing between causation and correlation—while accounting for potential confounders—is challenging.

1.3 Obesity and cardiometabolic diseases

The incidence of obesity has consistently increased since the 1980s, and it is now recognized as an obesity epidemic[1,16]. The rise in obesity rates has directly led to an increased incidence of diseases linked to obesity. Notably, since the 1980s, over four million deaths worldwide were attributed to excessive body weight, and cardiovascular disease was the primary cause for most of these deaths[25]. A previous large-scale observational study found that every 5 kg/m^2 increase in BMI over 25 kg/m^2 increased the risk of mortality from any cause by 30%[25]. Interestingly, centenarians, especially those without cardiovascular diseases, rarely exhibited obesity throughout their lifetimes, further suggesting that obesity can reduce lifespan[26]. While obesity is connected to various distinct diseases, such as digestive, respiratory, and neurological disorders, cardiometabolic diseases have been the leading cause of BMI-related disease burdens[27]. Epidemiological studies showed that the risk of developing cardiometabolic diseases such as coronary heart disease, stroke, or type 2 diabetes is nearly five times higher with obesity and up to 15 times higher for more severe obesity classifications[27].

Several proposed mechanisms may connect obesity to cardiometabolic diseases. Among these are systemic inflammation and histopathological remodeling[28,29]. It has been suggested that there is an interaction between adipocytes and macrophages[30,31]; long-chain saturated fatty acids from adipocytes can stimulate macrophages to produce

inflammatory cytokines. This, in turn, prompts these macrophages to incite pro-inflammatory reactions, causing an overarching stress response in the body. Another mechanism centers on the histopathological remodeling of adipose tissues and other related tissues[32,33]. As obesity progresses, adipocytes within the adipose tissues die due to metabolic stress. This type of adipocyte death is a distinctive feature of obesity and aligns with an increase in the size of fat cells observed in both mice and humans afflicted with obesity. Beyond the general adipose tissues, obesity also impacts other adipose tissues, such as the epicardial adipose tissues[34,35]. Although epicardial adipose tissue can offer cardioprotective benefits through its ability to handle free fatty acids, obesity can lead to its malfunctioning. Consequently, epicardial fat may release inflammatory substances that result in dysfunction and scarring of the nearby heart muscle. Additionally, the growth or inflammation of this tissue can further provoke issues in the surrounding heart muscle regions.

In short, obesity is intricately linked to systemic inflammation and histopathological remodeling. However, it should be noted that many of these studies predominantly rely on rodent models, cellular models, or observational methods. Elucidating the causal mechanisms connecting obesity to cardiometabolic diseases remains challenging, largely due to the chronic and complex nature of obesity. Often, the onset of obesity-induced cardiometabolic conditions spans years or even decades[36]. Consequently, executing interventional studies in humans to comprehensively understand these causal mechanisms is frequently impractical due to factors like high costs, logistical challenges, and ethical considerations.

1.4 Obesity and confounding

The relationship of obesity with COVID-19 and cardiometabolic diseases is confounded by numerous factors, including age, sex, smoking, alcohol consumption, and socioeconomic status[37-40]. A confounder is a variable that influences both the exposure (*X*) and the outcome (*Y*), potentially causing a spurious association and introducing bias into causal inference[41]. Formally, a confounder is defined by three criteria: (i) it is

associated with *X*; (ii) it is associated with *Y*, conditional on *X*; and (iii) it is not on the causal pathway between *X* and *Y*[42]. Traditional observational studies in epidemiology strive to account for confounders by adjusting for them. However, such adjustments are limited by the availability of data. While the majority of observational studies adjust for factors like age and sex, it is impractical to measure all potential confounders[43]. Moreover, some confounders, such as socioeconomic status, are particularly challenging to quantify and less often available[44,45]. Still, consistent evidence shows the influence of socioeconomic status on BMI and other obesity-related anthropometric traits. For example, in developing countries, a higher income correlates with an increased BMI. Conversely, in developed countries, BMI has shown an inverse relationship with median household income[46]. Furthermore, even after accounting for all known confounders, the risk of residual confounding persists[43]. Additionally, COVID-19 and cardiometabolic diseases might influence BMI and other metrics employed as obesity proxies, leading to issues of reverse causation. In such cases of reverse causation, *Y* influences *X* rather than the other way around, providing another potential source of biased estimations of the relationship between *X* and *Y*[47]. Given these observational analysis limitations, genetic epidemiology aims to uncover insights into causal biology and potential therapeutic targets by harnessing the principle that germline genetic variants are randomly assigned at conception and remain unaltered by disease. As such, genetic-derived evidence is less prone to confounding and reverse causation[48].

1.5 Genomics and Proteomics

Genome-wide association studies (GWAS) have significantly advanced our knowledge of obesity biology. The inaugural GWAS on obesity traits emerged in 2007, pinpointing a cluster of significant common variants within the intron of the *FTO* locus associated with BMI[50,51]. Since then, nearly 60 GWAS have revealed over 1,100 independent loci associated with obesity-related traits[52,53]. Yet, one key challenge is translating GWAS loci into candidate genes and shaping our understanding of biology and drug development. This process involves discerning the regulatory roles of non-coding

variants, identifying their potential effector transcripts, and determining where they function in the body. With the recent advent of cutting-edge genome-scale technologies that map regulatory elements, comprehensive multi-omics databases, sophisticated computational methods, and the latest genetic and molecular techniques, we are now better positioned to transform GWAS loci findings into actionable biological insights[54].

A viable approach to disentangle the intricacies of the relationship between obesity and its complications is to identify circulating proteins that act as mediators. Plasma proteins are involved in various biological functions, encompassing signaling, transportation, growth, repair, and protection against pathogens. Often, these proteins become dysregulated during diseases and serve as critical targets for medications. Thus, discerning the mechanisms that influence individual protein variations can provide valuable biological perspectives[55]. Furthermore, given that these proteins can be quantified and, in certain instances, modulated[14], understanding these mediatory proteins can shed light on the mechanisms through which obesity increases the risk of obesity-related complications, including COVID-19 and cardiometabolic diseases. This approach may present potential avenues for therapeutic measures. Advancements in large-scale proteomics have enabled the identification of genetic variants that influence plasma protein levels across the entire proteome. These genetic variants, known as protein quantitative trait loci (pQTLs), have been employed to identify causal proteins linked to diseases, their underlying mechanisms, and potential drug targets[55-57].

1.6 Mendelian randomization (MR)

Epidemiological studies often investigate the relationships between exposures and health outcomes. Yet, the associations found in these studies may not consistently provide accurate estimates of causal effects[58]. These discrepancies can arise due to confounding, wherein another variable influences both the outcome and the exposure. The gold standard method to evaluate such causation is randomized controlled trials

(RCTs), but they are not always possible due to cost, logistic, and ethical reasons. One of the effective ways to make causal inferences is through the use of Mendelian randomization (MR)[58]. MR is an effective genetic epidemiology approach to identify the causal relationship between modifiable risk factors or exposures and outcomes. MR can be described as a natural experiment somewhat analogous to RCTs because MR relies upon the random allocation of genetic variants at conception, similar to the randomization process in RCTs. Moreover, reverse causation does not affect genetic variations because genotype is always assigned prior to the onset of disease, and the disease does not change germline genotypes[59].

Nevertheless, MR is based on several instrumental variable assumptions: (I) the genetic variants used as instrumental variables are associated with the exposure; (II) they are not associated with factors that confound the relationship between the exposure and the outcome and (III) they influence the outcome only through the exposure (also known as exclusion restriction). Violation of the exclusion restriction is termed horizontal pleiotropy, wherein the variant used as an instrumental variable affects the disease independently of its effect on the exposure. Although careful assessment of directional horizontal pleiotropy is required, with proper selection of instrumental variables and sensitivity analyses, MR can serve as a powerful tool to help understand causal mechanisms for diseases in humans. In the case of COVID-19 severity and cardiometabolic diseases, MR can be used to rapidly screen thousands of proteins that may help to explain this relationship and identify potential therapeutic targets[60,61].

1.7 MR with proteomics

MR has been increasingly used in combination with proteomics, especially the plasma proteome, which facilitates the identification of circulating proteins that are causal for human diseases[56,62-65]. A straightforward way to classify protein-associated variants is by categorizing them into cis-acting pQTLs (cis-pQTLs) and trans-acting pQTLs (trans-pQTLs): cis-pQTLs are variants located in close proximity to the encoding gene (typically defined as either ≤500kb or ≤1Mb from the sentinel pQTL of the assessed

protein), while trans-pQTLs are variants located beyond this boundary[66]. The cis-pQTLs are considered more likely to directly influence the transcription and translation of proteins than trans-pQTLs due to their proximity to the protein-coding gene, making them less likely to be susceptible to horizontal pleiotropy. In contrast, trans-acting pQTLs might function through indirect mechanisms and are, thus, more prone to pleiotropy. Therefore, in the context of MR, employing cis-pQTLs as exposures can be beneficial in minimizing the risk of bias due to horizontal pleiotropy. However, because not all proteins have associated cis-pQTLs, this restriction may reduce the number of proteins available for causal inference testing[14,67].

1.8 Drug target discovery

There is increasing interest in using genomics and proteomics with genetic epidemiology methods to facilitate drug target discovery. Drug development is a lengthy, costly, and risky undertaking. The failure rate exceeds 96%[68], and the anticipated expenditure for launching a single drug to market hovers at approximately $1.8 billion CAD ($1.3 billion USD)[6]. Therefore, finding a strategy to increase the success rate is of paramount importance. Incorporating genomics data has demonstrably improved these success rates. Specifically, targets backed by genetic evidence have a more than twofold increase in success rate during clinical development. Further studies indicate that such genetically backed targets exhibit higher likelihoods of success during phase II and III trials. Notably, around two-thirds of the drugs that received FDA approval in 2021 possessed corroborating human genetic evidence[5].

The advent of broad-capture proteomics presents another promising avenue to identify potential drug targets. Efforts are underway to harness both proteomics and genomics in a comprehensive exploration and assessment of genetic signals, particularly those linking protein function and abundance to various diseases. Within this framework, the pQTLs—genetic variants that modulate protein abundance—serve as valuable tools in

MR to ascertain the influence of circulating protein levels on disease outcomes. Leveraging this methodology, researchers have pinpointed promising therapeutic targets[56,57,64,65,69-71]. Nonetheless, the potential of pQTL MR in discerning mediators across the entire proteome remains largely unexplored.

1.9 Rationale and structure of the thesis

The primary aim of this thesis is to utilize genomics and proteomics to gain clinically relevant insights into the causal relationship between obesity and COVID-19 severity, thereby identifying potential therapeutic targets. In Chapter 2, given that a high BMI has emerged as a critical risk factor for COVID-19, we employed multivariable MR to dissect the causal association between body fat mass and fat-free mass with the severity of COVID-19. In Chapter 3, we identified the mediators underpinning the relationship between obesity and COVID-19 severity, employing a combination of large-scale genomics, proteomics, and genetic epidemiology methods. The intent was to conduct a rapid, hypothesis-free proteome-wide scan to identify causal proteins, shedding light on the causal biology and suggesting potential therapeutic targets. In Chapter 4, we examined whether this methodology, termed "two-step MR," can be used to discern circulating proteins that mediate the effects of obesity on cardiometabolic diseases. Identifying such mediators can facilitate drug development for these diseases, which are the leading causes of obesity-related deaths.

Chapter 2: Causal associations between body fat accumulation and COVID-19 severity: A Mendelian randomization study

Satoshi Yoshiji[1–5], Daisuke Tanaka[1], Hiroto Minamino[1,5], Tianyuan Lu[3,6], Guillaume Butler-Laporte[3,7], Takaaki Murakami[1], Yoshihito Fujita[1], J. Brent Richards[2,3,7–9†], and Nobuya Inagaki[1†]

[1] Department of Diabetes, Endocrinology and Nutrition, Graduate School of Medicine, Kyoto University, Kyoto, Japan.
[2] Department of Human Genetics, McGill University, Montréal, Québec, Canada.
[3] Centre for Clinical Epidemiology, Department of Medicine, Lady Davis Institute, Jewish General Hospital, McGill University, Montréal, Québec, Canada.
[4] Kyoto-McGill International Collaborative Program in Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto, Japan.
[5] Japan Society for the Promotion of Science, Tokyo, Japan.
[6] Quantitative Life Sciences Program, McGill University, Montréal, Québec, Canada
[7] Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, Québec, Canada.
[8] Department of Twin Research, King's College London, London, United Kingdom.
[9] 5 Prime Sciences, Montréal, Québec, Canada.
[†] These authors contributed equally to this study.

Correspondence:

J. Brent Richards

Professor of Medicine, McGill University, Senior Lecturer, King's College London (Honorary), Pavilion H-413, Jewish General Hospital, 3755 Côte-Ste-Catherine Montréal, Québec, H3T 1E2, Canada.

Tel: +1-514-340-8222

Fax: +1-514-340-7529

E-mail: brent.richards@mcgill.ca

Nobuya Inagaki

Professor and Chairman, Department of Diabetes, Endocrinology and Nutrition, Graduate School of Medicine, Kyoto University, 54 Kawahara-cho, Shogoin, Sakyo-ku, Kyoto, 606-8507, Japan.

Tel: +81-075-751-3560

Fax: +81-075-751-4244

E-mail: inagaki@kuhp.kyoto-u.ac.jp

2.1 Abstract

Previous studies reported associations between obesity measured by body mass index (BMI) and coronavirus disease 2019 (COVID-19). However, BMI is calculated only with height and weight and cannot distinguish between body fat mass and fat-free mass. Thus, it is not clear if one or both of these measures are mediating the relationship between obesity and COVID-19. Here, we used Mendelian randomization (MR) to compare the independent causal relationships of body fat mass and fat-free mass with COVID-19 severity. We identified single nucleotide polymorphisms associated with body fat mass and fat-free mass in 454,137 and 454,850 individuals of European ancestry from the UK Biobank, respectively. We then performed two-sample MR to ascertain their effects on severe COVID-19 (cases: 4,792; controls: 1,054,664) from the COVID-19 Host Genetics Initiative. We found that an increase in body fat mass by one standard deviation was associated with severe COVID-19 (odds ratio $(OR)_{body\ fat\ mass}$ = 1.61, 95% confidence interval [CI]: 1.28–2.04, $P = 5.51\times10^{-5}$; $OR_{body\ fat\text{-}free\ mass}$ = 1.31, 95% CI: 0.99–1.74, $P = 5.77\times10^{-2}$). Considering that body fat mass and fat-free mass were genetically correlated with each other (r = 0.64), we further evaluated independent causal effects of body fat mass and fat-free mass using multivariable MR and revealed that only body fat mass was independently associated with severe COVID-19 ($OR_{body\ fat\ mass}$ = 2.91, 95%CI: 1.71–4.96, $P = 8.85\times10^{-5}$ and $OR_{body\ fat\text{-}free\ mass}$ = 1.02, 95%CI: 0.61–1.67, P = 0.945). In summary, this study demonstrates the causal effects of body fat accumulation on COVID-19 severity and indicates that the biological pathways influencing the relationship between COVID-19 and obesity are likely mediated through body fat mass.

2.2 Introduction

More than 500 million individuals have been infected by the coronavirus disease-19 (COVID-19) with 6 millions of deaths worldwide to date (1). The severity of COVID-19 varies considerably among individuals and identifying modifiable risk factors associated with COVID-19 severity is essential for optimizing public health policies, allocating resources, and assisting clinical decisions.

A major risk factor for COVID-19 appears to be obesity. A community-based cohort study involving 6.9 million individuals in England showed a positive association between body mass index (BMI) and COVID-19 severity (2), which was replicated in other independent observational studies (3-5). However, the key limitation of BMI is that it is a crude proxy of obesity because it is calculated only with height and weight and does not consider body composition (i.e., body fat mass and body fat-free mass) (6). Therefore, direct measures of body composition assessed by dual-energy X-ray absorptiometry or bioelectrical impedance analysis might better elucidate the association of body fat accumulation with COVID-19 outcomes. In this regard, two recent studies utilized the direct measures of body composition to evaluate the effect of obesity on COVID-19 (7, 8). However, individuals with increased body fat mass are also more likely to have increased body fat-free mass because there is a positive correlation between body fat mass and body fat-free mass (9). Thus, we have to specifically study the independent effects of body fat mass and body fat-free mass to disentangle the causal effects of obesity on COVID-19.

Regarding a means of exploring the associations between risk factors and outcomes of the interest, observational studies can evaluate correlations but not causations; in fact, interpreting the results of observational studies as a causal relationship relies on untestable and usually implausible assumptions, including the absence of unmeasured confounders and reverse causation (10). Given these limitations inherent to traditional observational epidemiology studies, Mendelian randomization (MR) has emerged as a way to mitigate against such shortcomings through its use of genetic variants as instrumental variables to infer a causal relationship between exposures and outcomes (11, 12). Using MR, we can estimate the causal effects of genetically predicted levels of

adiposity-related exposures on COVID-19 outcomes, in contrast to typical observational studies that evaluate only associations. Because genetic alleles are randomly assigned at conception, which is generally well before the onset of the disease, the risk of reverse causation is substantially decreased. Taking advantage of MR analysis, previous studies evaluated causal associations of anthropometric traits of obesity and some direct measures of body composition, such as body fat percentage (7, 13-15). However, none has taken into account the correlation of body fat and fat-free mass and evaluated the independent causal associations of body fat mass and body fat-free mass with COVID-19 outcomes.

In this study, we conducted a two-sample MR to assess independent causal associations of body fat mass and body fat-free mass with COVID-19 severity outcomes using data from the UK Biobank and the COVID-19 Host Genetics Initiative.

2.3 Methods

2.3.1 Instrumental Variables for Body Fat Mass, Body Fat-free Mass, Body Fat Percentage, and BMI

Instrumental variables were defined as independent genome-wide significant single-nucleotide polymorphisms (SNPs) ($P < 5\times10^{-8}$) for exposure traits. Independence of SNPs was defined as not in linkage disequilibrium with other SNPs ($r^2 < 0.001$ within a 10,000 kilobase [kb] window). The exposures used in this study were body fat mass, body fat-free mass, body fat percentage, and BMI. Body fat percentage and BMI were included as supplementary analyses. To select SNPs used as instrumental variables, we obtained the genome-wide association study (GWAS) results of body fat mass, body fat-free mass, body fat percentage, and BMI from individuals with European ancestry in the UK Biobank (**Figure 1**), using the OpenGWAS and MR-Base platform of the MRC Integrative Epidemiology Unit at the University of Bristol (16). Accession IDs were as follows: body fat mass (ukb-b-19393), body fat-free mass (ukb-b-13354), body fat percentage (ukb-b-8909), and BMI (ukb-b-19953). A full description of the study design, participants and

quality control procedures were described in detail previously (17). Briefly, GWAS was performed using 12,370,749 SNPs on 463,005 individuals by BOLT-LMM (18) with the following quality control criteria: Imputation quality (INFO) score > 0.3 for SNPs with a MAF > 3%; INFO score > 0.6 for SNPs with a MAF between 1–3%; INFO score > 0.8 for SNPs with a MAF between 0.5–1%; INFO score > 0.9 for SNPs with a MAF between 0.1–0.5%; SNPs with a MAF below 0.1% were excluded; individuals who were outliers in heterozygosity and missing rates, and individuals with sex-mismatch (i.e. different genetic sex and reported sex) or sex-chromosome aneuploidy were excluded. The fat mass and fat-free mass of the UK Biobank participants were evaluated by performing bioelectrical impedance analysis using the Tanita BC418MA body composition analyzer (Tanita, Tokyo, Japan). We restricted the analyses to individuals of European ancestry to maximize the statistical power, given that the majority of UK Biobank participants were of European ancestry. To select instrumental variables, SNPs were clumped using PLINK (v1.90) according to a linkage disequilibrium threshold of $r^2 < 0.001$ with a clumping window of 10,000 kb using the 1000G European reference panel (16, 19) in order to select an independent SNP with the lowest *P*-value in each linkage disequilibrium block. When a selected SNP was not present in the results of the GWAS of COVID-19 severity outcomes, we instead used a proxy SNP that was in linkage disequilibrium with the selected SNP, with an r^2 of ≥0.8 and minor allele frequency of ≤0.3 using 1000G European reference panel as described before (12). We calculated *F*-statistics for the exposure traits and a genetic correlation between body fat mass and body fat-free mass using LDAK (v5.1) (19).

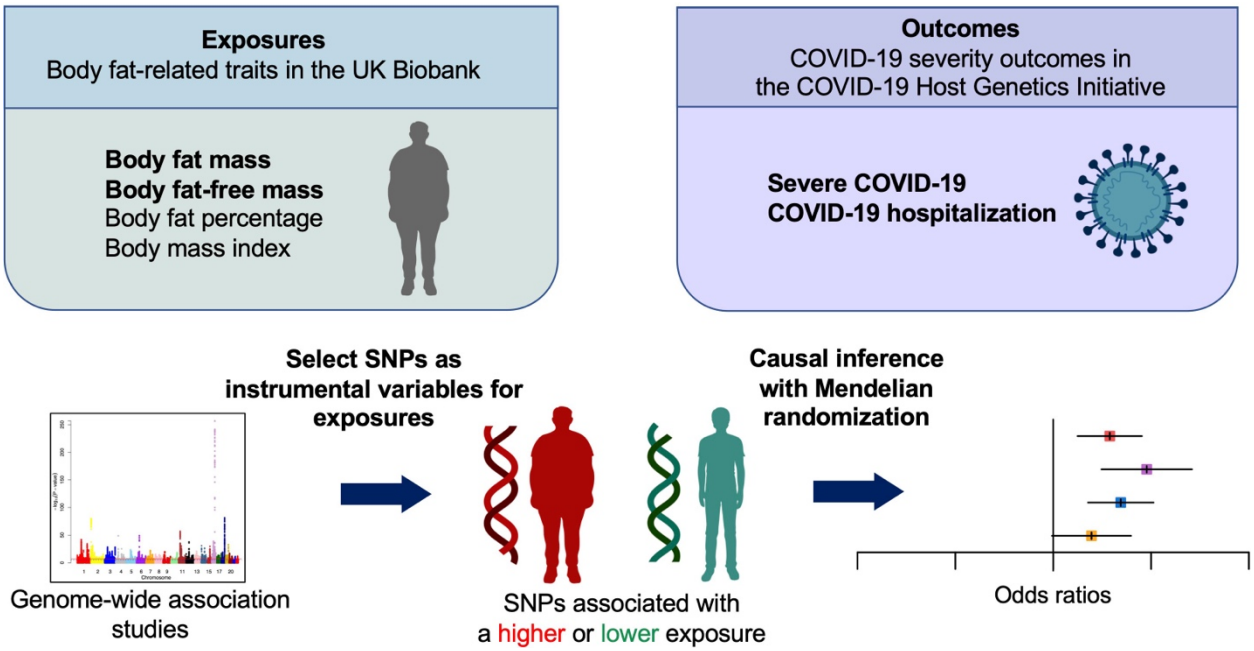


Figure 1. Schematic representation of the Mendelian randomization study. SNPs, single nucleotide polymorphisms.

2.3.2 Severe COVID-19 and COVID-19 Hospitalization Outcomes

For proxy outcomes of COVID-19 severity, we adopted the outcomes of the COVID-19 Host Genetics Initiative, an international consortium working collaboratively to share data and ideas, recruit patients, and disseminate scientific findings. The outcomes were severe COVID-19 and COVID-19 hospitalization (20). For definitions of COVID-19 outcomes, the severe COVID-19 group was defined as individuals whose death was due to COVID-19, or those requiring hospitalization and respiratory support due to symptoms related to laboratory-confirmed SARS-CoV-2 infection. The COVID-19 hospitalization group was defined as individuals requiring hospitalization due to symptoms associated with laboratory-confirmed severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection. For the definitions of controls in the GWAS data, ancestry-matched controls were sourced from participating population-based cohorts. Controls included individuals whose status of exposure to SARS-CoV-2 was either negative according to electronic health records/questionnaires or unknown (20). We used the largest GWAS summary statistics of the COVID-19 Host Genetics Initiative for severe COVID-19 and COVID-19 hospitalization outcomes in individuals of European-ancestry, excluding those from the UK Biobank. The datasets corresponding to each outcome were as follows: severe COVID-19 (cases: 4,792; controls: 1,054,664; dataset ID: COVID19_HGI_A2_ALL_eur_leave_ukbb_23andme_20210107 from data release 5) and COVID-19 hospitalization (cases: 14,652; controls: 1,114,836; and dataset ID: COVID19_HGI_B2_ALL_eur_leave_ukbb_23andme_20210622 from data release 6). We note that the COVID-19 Host Genetics Initiative's data release 6 did not include ancestry-specific GWAS for the severe COVID-19 outcome and also that the latest data release 7 did not include GWAS in European-ancestry individuals excluding those from the UK biobank. Hence, we used data release 5 for the severe COVID-19 outcome and data release 6 for the COVID-19 hospitalization outcome to minimize bias due to sample overlap or genetic confounding due to population stratification.

2.3.3 Mendelian Randomization

We performed univariable MR using the inverse variance weighted method (hereinafter referred to as univariable MR) to evaluate the relationship of body fat mass, body fat-free mass, body fat percentage, and BMI with severe COVID-19 and COVID-19 hospitalization. Univariable MR is a weighted linear regression model in which the effect of genetic variants i ($i = 1 \ldots n$) on an outcome $\hat{\beta}_{Y_i}$ is regressed on the effect of the same genetic variant i on the exposure $\hat{\beta}_{X_i}$ weighted by the inverse of the squared standard error ($se\left(\hat{\beta}_{Y_i}\right)^{-2}$). The estimated total effect (θ) of the exposure on the outcome can be formulated as follows:

$$\hat{\beta}_{Y_i} = \theta\hat{\beta}_{X_i} + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}\left(0, se\left(\hat{\beta}_{Y_i}\right)^{-2}\right)$$

The instrumental variable assumptions are as follows: (I) Relevance–genetic variant is associated with the exposure. (II) independence–genetic variant does not share the unmeasured cause or confounder with the outcome. (III) exclusion restriction–genetic variant does not influence the outcome except through the exposure (11, 12). These assumptions are illustrated by a canonical diagram in **Figure 2**.

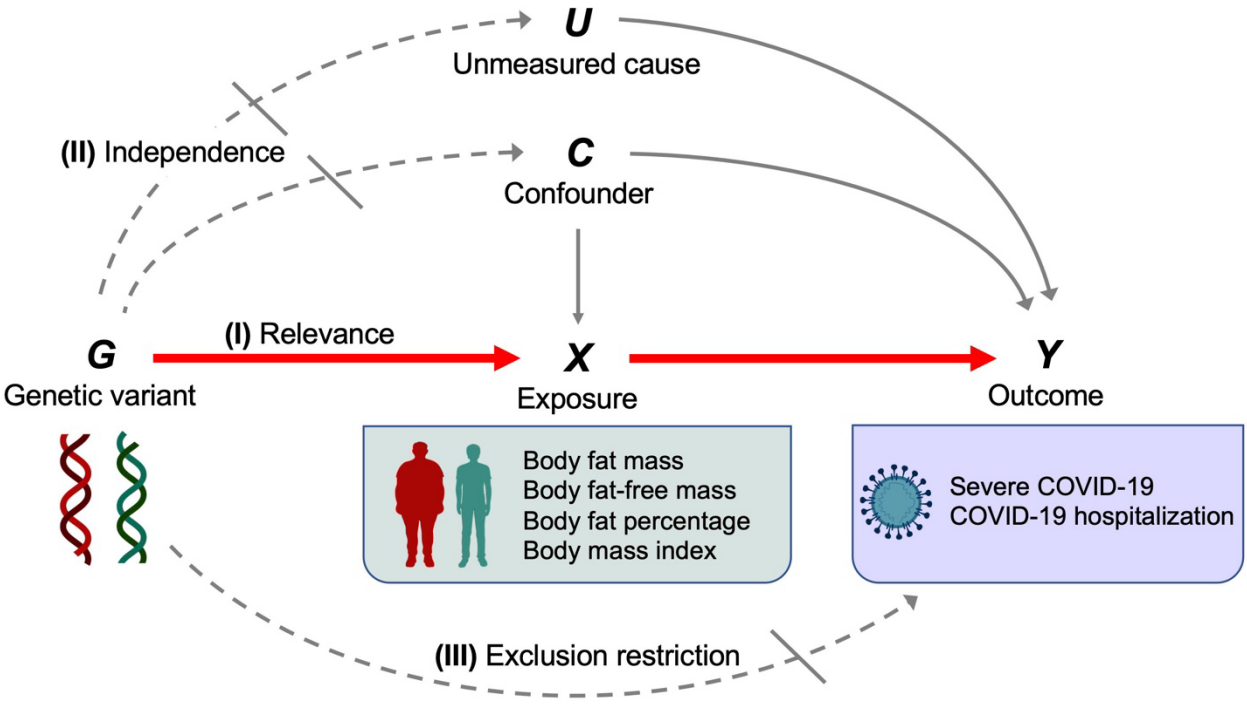


Figure 2. Canonical diagram illustrating the instrumental variable assumptions made in the Mendelian randomization analyses.

Genetic variant ***G*** is used as an instrumental variable for exposure ***X*** (body mass index, body fat percentage, body fat mass, or body fat-free mass) to evaluate the causal effect of ***X*** on the outcome ***Y*** (severe COVID-19 or COVID-19 hospitalization). Instrumental variable assumptions include the following: (I) Relevance–genetic variant ***G*** is associated with exposure ***X***. (II) independence–genetic variant ***G*** does not share the unmeasured cause or the confounder with the outcome ***Y***. (III) exclusion restriction–genetic variant ***G*** does not influence the outcome ***Y*** except through the exposure ***X***. Red solid arrows represent causal effects; gray solid arrows represent causal effects of the unmeasured cause or confounder that do not violate the instrumental variable assumptions; dashed arrows represent causal effects that are specifically prohibited by the instrumental variable assumptions.

Multivariable MR was performed using the inverse variance weighted method (hereinafter referred to as multivariable MR). This is an extension of univariable MR, in which the effects of genetic variant i ($i = 1 \ldots n$) on the outcome ($\hat{\beta}_{Y_i}$) are regressed on the effect of genetic variant i on two exposures of X_1 (fat mass) and X_2 (fat-free mass). In multivariable MR, genetic variants used as instrumental variables are associated with one or both of the exposures (21).

The causal associations were evaluated using odds ratios (ORs), which are expressed according to a standard deviation (SD) increase in genetically predicted body fat mass (kg), or body fat-free mass (kg), body fat percentage (%), and BMI (kg/m^2).

Results with a $P < 0.0125$ were considered statistically significant ($P = 0.05/4$; Bonferroni-corrected significance threshold according to the number of exposures). We note that such a correction is likely overly conservative, given that the exposures are non-independent. MR analyses were performed using TwoSampleMR (v0.5.6) in R (v4.02). This study was conducted in accordance with the STROBE-MR guideline (6, 7). STROBE-MR checklist is provided in **Supplementary Material** (22).

2.3.4 Sensitivity Analysis

We performed the MR-Egger intercept test, Cochran's Q test, and the MR-PRESSO global test (23, 24) to detect horizontal pleiotropy, which occurs when instrumental variables influence outcomes through pathways independent of the exposure. MR-Egger relaxes the exclusion restriction assumption and is valid under the Instrument Strength Independent of Direct Effect (InSIDE) assumption that associations of the genetic variants with the exposure trait are independent of direct effects of the genetic variants on the outcome. Deviation of the MR-Egger intercept from zero indicates horizontal pleiotropy. The results of Cochran's Q test were used to evaluate the heterogeneity of genetic variants used as instrumental variables. Results of Cochran's Q test were presented with I^2 index, based on which the heterogeneity of genetic variants was defined categorically with I^2 index as low (I^2 index ≤ 25%), moderate (I^2

index 26–50%), and high (I^2 index > 50%). Additionally, we performed the MR-PRESSO global test, which can detect horizontally pleiotropic outlier SNPs. A significant result indicates the presence of pleiotropic outlier SNPs and this method then generates ORs after removing and correcting for these outliers (outlier-corrected ORs). MR-PRESSO can also be used to evaluate the distortion of the causal estimates before and after the removal of pleiotropic outlier SNPs following the MR-PRESSO distortion test. MR-PRESSO requires at least 50% of the genetic variants to be valid instruments with no horizontal pleiotropy and also relies on the InSIDE assumption. We also performed leave-one-out analyses for all exposure-outcome associations, which repeated univariable weighted MR excluding each SNP to assess whether the overall estimate is driven by a single SNP. We also generated scatter plots and funnel plots to inspect for horizontal pleiotropy.

Results with a $P < 0.05$ were considered to indicate the presence of horizontal pleiotropy for the MR-Egger intercept test, Cochran's Q test, MR-PRESSO global test, and MR-PRESSO distortion test. Sensitivity analyses were performed with TwoSampleMR (v.0.5.6) and MR-PRESSO (v1.0).

2.3.5 Ethics Statements

The UK Biobank and COVID-19 Host Genetics Initiatives obtained ethics approval from the relevant institutional ethics committees. We used publicly available summary statistics of GWAS results of UK Biobank and COVID-19 Host Genetics Initiative and did not use individual-level data.

2.4 Results

2.4.1 Instrumental Variables or Exposure Traits

The characteristics of the exposure traits (body fat mass, body fat-free mass, body fat percentage, and BMI) are presented in **Table 1**. The mean ± SD of body fat mass was

24.9 ± 9.6 kg, body fat-free mass was 53.2 ± 11.5 kg, body fat percentage was 31.4 ± 8.5%, and BMI was 27.4 ± 4.8 kg/m^2 (**Table 1**). For body fat mass, body fat-free mass, body fat percentage, and BMI, 417, 530 377, and 439 independent genome-wide significant SNPs were identified as instrumental variables from the GWAS results of the UK Biobank, respectively. *F*-statistics for these exposure traits were 502.2, 607.4, 496.9, and 507.6, respectively. The SNPs used as instrumental variables are presented in **Supplementary Table 1 (22)**.

Table 1. Dataset descriptions.

Data source	Dataset details	Phenotype	Sample size of each dataset	Mean ± SD
UK Biobank	• GWAS in individuals of European ancestry. • Body fat and body fat–free mass were measured using bioelectrical impedance analysis.	Body fat mass	454,137	24.9 ± 9.6 kg
		Body fat–free mass	454,850	53.2 ± 11.5 kg
		Body fat percentage	454,633	31.4 ± 8.5%
		Body mass index	461,460	27.4 ± 4.8 kg/m^2
COVID–19 Host Genetics Initiative	• Meta–analysis of GWAS in individuals of European ancestry excluding those from UK biobank	Severe COVID–19	Cases: 4,792 Controls: 1,054,664	–
		COVID–19 hospitalization	Cases: 14,652 Controls: 1,114,836	–

2.4.2 Severe COVID-19 Outcome

For the severe COVID-19 outcome, univariable MR showed that the genetically predicted increase per SD in body fat mass, body fat percentage, and BMI was associated with an increased risk of severe COVID-19 ($OR_{body\ fat\ mass}$ = 1.61, 95%CI 1.28–2.04, $P = 5.51\times10^{-5}$; and $OR_{body\ fat\text{-}free\ mass}$ =1.31, 95%CI: 0.99–1.74, $P = 5.77\times10^{-2}$; $OR_{body\ fat\ percentage}$ = 1.94, 95% confidence interval [CI]: 1.41–2.67; $P = 5.07\times10^{-5}$; OR_{BMI} = 1.49, 95%CI: 1.19–1.87, $P = 5.57\times10^{-4}$) (**Figure 3**). Further, as instrumental variables for body fat mass and body fat-free mass were not independent of each other (r = 0.64 for the genetic correlation of the two traits) (**Figure 4**), we performed multivariable MR to elucidate the independent causal effects of body fat mass and body fat-free mass on the severe COVID-19 outcome, which showed that only body fat mass was independently associated with the severe COVID 19 outcome (body fat mass: $OR_{body\ fat\ mass}$ = 2.91, 95%CI: 1.71–4.96, $P = 8.85\times10^{-5}$, and $OR_{body\ fat\text{-}free\ mass}$ = 1.02, 95%CI: 0.61–1.67, P = 0.945) (**Figure 5**).

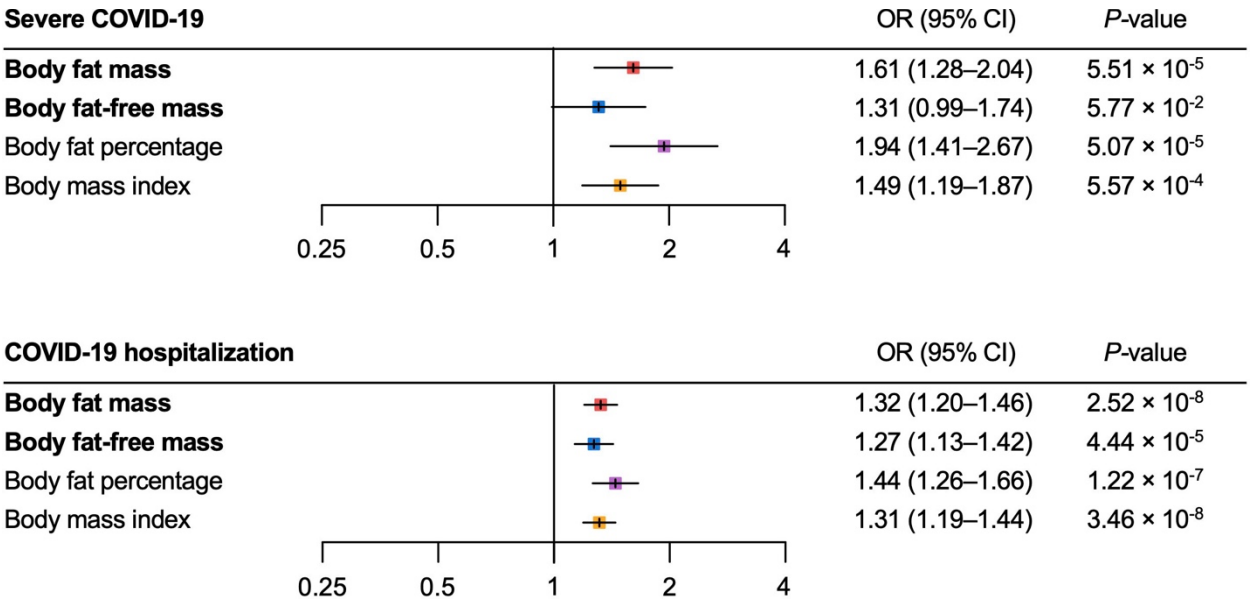


Figure 3. Univariable Mendelian randomization analysis for the severe COVID-19 and COVID-19 hospitalization outcomes.

MR, Mendelian randomization.

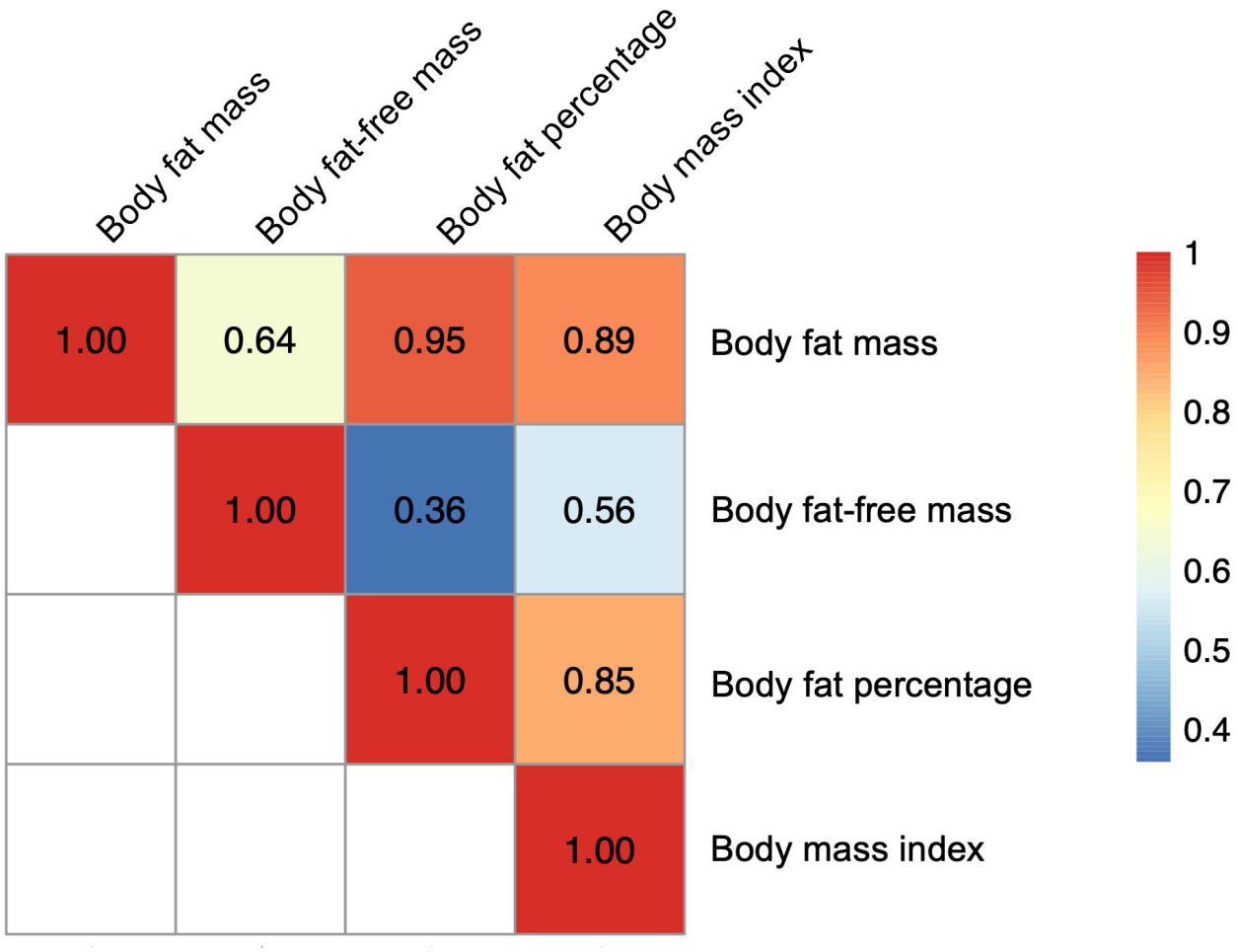


Figure 4. Heatmap for genetic correlation coefficients between the body fat-related traits.

Genetic correlations among the four exposures (body fat mass, body fat-free mass, body fat percentage, and body mass index) were analyzed with LDAK using the results of corresponding genome-wide association studies.

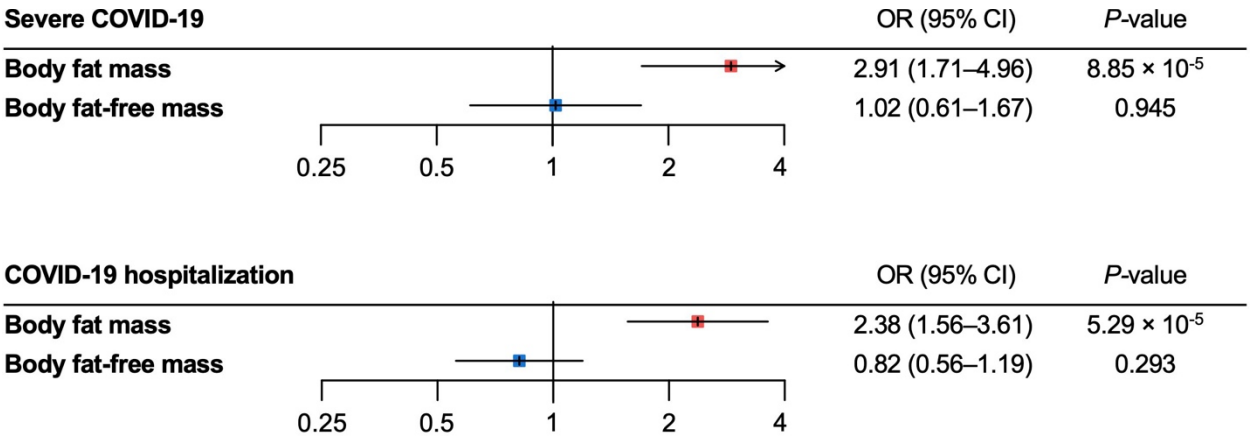


Figure 5. Multivariable Mendelian randomization analysis for the severe COVID-19 and COVID-19 hospitalization outcomes.

MR, Mendelian randomization.

2.4.3 COVID-19 Hospitalization Outcome

For the COVID-19 hospitalization outcome, univariable MR showed that a genetically predicted increase per SD in body fat mass, body fat-free mass, body fat percentage, and BMI and was associated with an increased risk of COVID-19 hospitalization ($OR_{body\ fat\ mass}$ = 1.32, 95%CI: 1.20–1.46, $P = 2.52 \times 10^{-8}$; $OR_{body\ fat\text{-}free\ mass}$ = 1.27 95%CI: 1.13–1.42, $P = 4.44 \times 10^{-5}$; $OR_{body\ fat\ percentage}$ = 1.44, 95%CI: 1.26–1.66, $P = 1.22 \times 10^{-7}$; OR_{BMI} = 1.31, 95%CI: 1.19–1.44, $P = 3.46 \times 10^{-8}$) (**Figure 3**). In multivariable MR, only body fat mass was independently associated with COVID-19 hospitalization ($OR_{body\ fat\ mass}$ = 2.38, 95%CI: 1.56–3.61, $P = 5.29 \times 10^{-5}$; $OR_{body\ fat\text{-}free\ mass}$ = 0.82, 95%CI: 0.56–1.19, P = 0.293), consistent with the findings for severe COVID-19 (**Figure 5**).

2.4.4 Sensitivity Analysis

We performed MR-Egger, Cochran's Q test and MR-PRESSO for sensitivity analysis (**Table 2**). In the MR-Egger, the 95%CI results of the MR-Egger intercept (Egger intercept) contained the null hypothesis value zero for all exposure-outcome relationships, suggesting no evidence of horizontal pleiotropy. Heterogeneity estimates of instrumental variables were low according to the I^2 index (I^2 index were ≤ 25% for all exposure traits). The leave-one-out analyses showed that causal estimates were robust to exclusion of single SNPs (**Supplementary Table 2–5**). Visual inspection of the scatter plots and funnel plots did not suggest biased estimates or pleiotropy (**Figure 6 and Supplementary Figure 1**). However, MR-PRESSO detected some pleiotropic outlier SNPs in instrumental variables body fat mass, body fat percentage, and BMI with the COVID-19 hospitalization outcome (P-value for global test < 0.05). Nevertheless, results with MR-PRESSO after removal and correction for these pleiotropic outlier SNPs were directionally consistent with those from univariable MR, supporting the robustness of the findings with univariable MR. In addition, the MR-PRESSO distortion test detected no significant distortion in the causal estimates before and after the removal of outlier pleiotropic SNPs (**Table 2**).

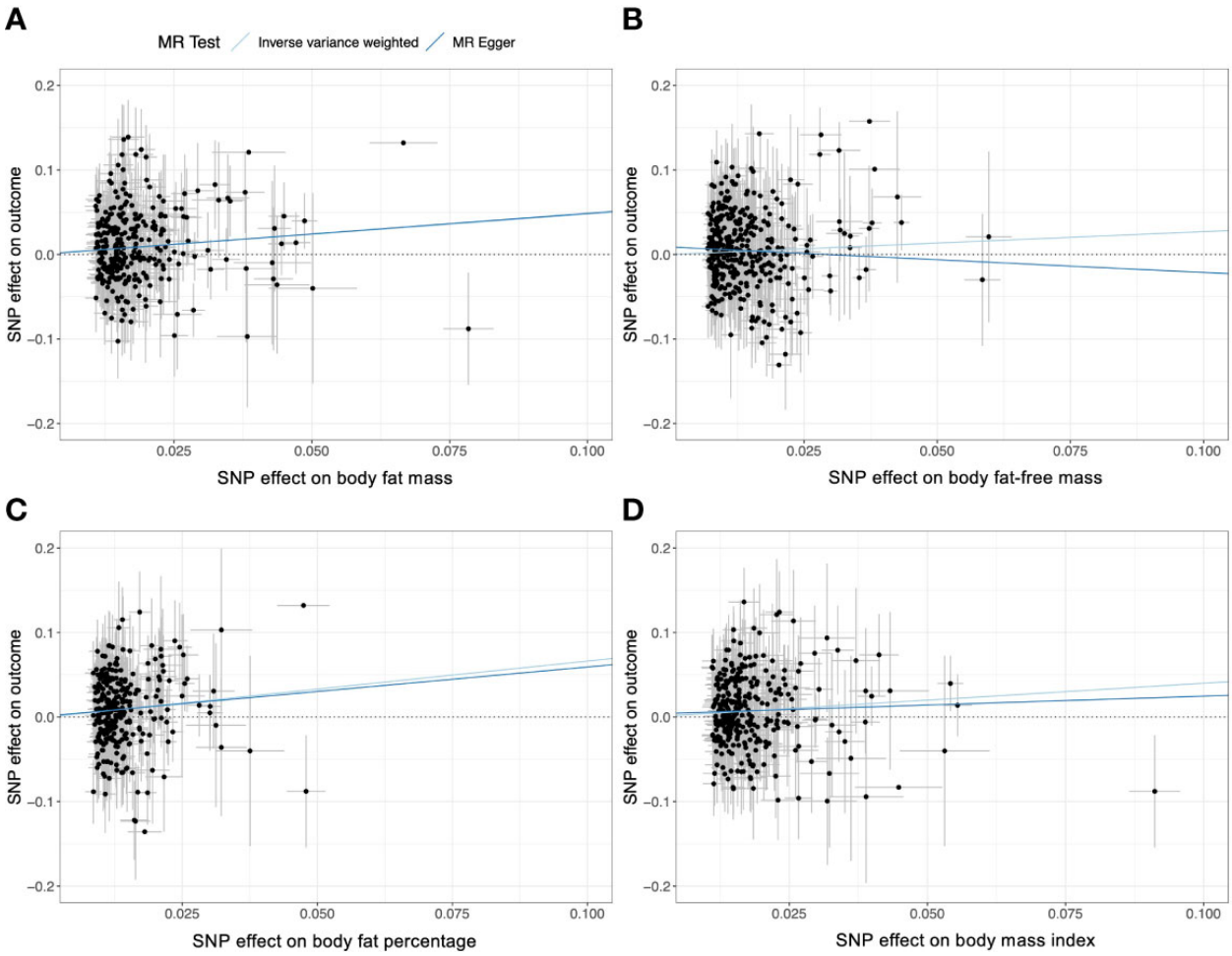


Figure 6. Scatter plots of the univariable weighted MR analyses for (a) body fat mass, (b) body fat-free mass, (c) body fat percentage, and (d) body fat mass. Each dot represents a genetic instrumental variable. Two lines represent causal estimate (β_{IV}) by the inverse variance weighted method (light blue) and the MR-Egger method (blue). Error bars represent 95%CIs.

MR, Mendelian randomization.

Table 2. Sensitivity analysis results.

Exposures	Outcomes	Sensitivity analysis methods								
		MR–Egger				Cochran's Q test I^2 index	MR–PRESSO			
		Egger slope (95% CI)	*P*–value (Egger slope)	Egger intercept (95% CI)	*P*–value (Egger intercept)		Global test *P*–value	Outlier–corrected OR (95% CI)	Outlier–corrected *P*–value	Distortion test *P*–value
Body fat mass	Severe COVID–19	1.63 (0.84–3.15)	0.147	−0.0002 (−0.012–0.012)	0.975	4.0	0.273	No outlier	—	—
Body fat–free mass		0.74 (0.38–1.45)	0.378	0.009 (−0.001–0.018)	0.066	5.0	0.551	No outlier	—	—
Body fat percentage		1.79 (0.62–5.14)	0.282	0.001 (−0.013–0.016)	0.874	9.7	0.108	No outlier	—	—
Body mass index		1.24 (0.65–2.37)	0.523	0.004 (−0.008–0.015)	0.542	1.8	0.417	No outlier	—	—
Body fat mass	COVID–19 hospitalization	1.53 (1.17–2.00)	2.31×10^{-3}	−0.003 (−0.008–0.002)	0.264	17.3	0.004	1.3351 (1.3348–1.3353)	3.98×10^{-3}	0.881
Body fat–free mass		1.12 (0.85–1.46)	0.433	0.002 (−0.002–0.006)	0.302	12.3	0.174	No outlier	—	—
Body fat percentage		1.80 (1.17–2.76)	8.04×10^{-3}	−0.003 (−0.009–0.003)	0.297	18.4	0.003	1.4498 (1.4493–1.4503)	1.60×10^{-7}	0.810
Body mass index		1.27 (0.99–1.64)	6.34×10^{-2}	0.001 (−0.004–0.005)	0.812	19.5	0.001	1.3085 (1.3082–1.3088)	7.09×10^{-9}	0.864

2.5 Discussion

In this study, we first used two-sample MR to disentangle the independent effects of body fat mass and body fat-free mass and showed that body fat mass, but not body fat-free mass, is independently associated with severe COVID-19 outcomes. First, we performed univariable weighted MR and found that increased body fat mass, along with BMI and body fat percentage, were associated with an increased risk of severe COVID-19 and COVID-19 hospitalization. We further used multivariable MR to disentangle the independent causal effects of body fat mass and body fat-free mass on these outcomes and revealed that only body fat mass was independently associated with the outcomes.

During the COVID-19 pandemic, obesity has emerged as a major risk factor for COVID-19 outcomes. Multiple observational and MR studies suggested that obese individuals present an increased risk of severe diseases, hospitalization, and death due to COVID-19 (2-4, 25). However, observational studies are prone to confounding bias and reverse causation and do not estimate the causal effects of exposures on outcomes. To tackle this problem, recent studies have used MR to estimate the causal effect of obesity on the risk of COVID-19. For instance, the landmark paper from the COVID-19 Host Genetics Initiative showed that BMI was causally associated with an increased risk of COVID-19 hospitalization (20). This was supported by multiple MR studies and our analysis, which included BMI as the supplementary exposure. Other studies also assessed multiple anthropometric traits, including waist circumference, hip circumference, waist-to-hip ratio, and trunk fat ratio as well as BMI to evaluate the effect of adiposity on the risk of COVID-19 (7, 8, 13, 14, 26-32). These MR studies consistently estimated that increases in BMI, waist circumference, and hip circumference are causal for COVID-19 severity (7, 13, 14, 27, 29). On the other hand, the waist-to-hip ratio was not associated with COVID-19 severity (7, 29), contradicting observational studies. These discrepancies may be explained by confounding factors involved in observational studies but also by the limited ability of anthropometric traits to act as proxies for body composition (i.e., body fat mass and fat-free mass). It should also be noted that BMI is a function only of weight and height and an indirect measurement of obesity. Thus, it may not necessarily reflect body composition, which can be directly measured with bioelectrical impedance analysis or

dual-energy X-ray absorptiometry (DXA). For example, individuals with similar BMI may have very different body composition, if there are large changes in lean body mass. This highlights the importance of directly measuring adiposity. In this regard, two recent MR studies used GWAS of direct measurements of obesity (i.e., body fat mass, fat-free mass, and body fat percentage) and found that they influence the risk of COVID-19, which was replicated by our univariable MR analyses (7, 8). However, analyses using body composition measurements still have limitations such as the high correlation between body fat mass and body fat-free mass, which was highlighted by our genetic correlation analysis (r = 0.64). To the best of our knowledge, the present study is the first to disentangle the independent causal effects of body fat mass and body fat-free mass on COVID-19 severity.

Our multivariable MR showed that one SD increase in body fat mass (9.6 kg) is causally associated with 2.91-fold and 2.38-fold increase in the risk of severe COVID-19 and COVID-19 hospitalization, respectively, highlighting the burden of body fat accumulation on COVID-19 severity. On the contrary, body fat-free mass was not independently associated with increased risk of severe COVID-19 or hospitalization. We used multivariable MR since most instrumental variables of adiposity affect both fat mass and fat-free mass, although some variants more strongly and proportionally influence fat mass, whereas others influence fat-free mass more strongly. Therefore, multivariable MR can test the differential causal effects of fat mass and fat-free mass. Using this approach, recent MR studies showed differential associations between body fat mass and body fat-free mass with various disorders (9, 33-35). The present findings extend this knowledge to COVID-19. Results from multivariable MR showed that body fat mass but not body fat-free mass was independently associated with severe COVID-19 and COVID-19 hospitalization. The association between body fat mass and COVID-19 severity was strengthened in multivariable MR relative to findings using univariable MR, whereas the effects of body fat-free mass on COVID-19 severity was markedly attenuated in multivariable MR, thereby illustrating the independent causal effects of body fat mass on COVID-19 severity.

The underlying mechanism of these associations remains to be clarified. Obesity is a metabolic disease characterized by systemic changes in metabolism, including insulin resistance, glucose intolerance, dyslipidemia, changes in adipokines (e.g., increased leptin and decreased adiponectin levels), chronic inflammation, and altered immune response, all of which could collectively increase the risk of COVID-19 severity (36-38). In addition, recent studies suggests that adipose tissue is a potential organ for direct infection with SARS-CoV2 in obese individuals (39). The infection of adipose tissue can cause systemic metabolic dysregulation including hyperglycemia, which is known as another risk factor for COVID-19 severity (40). Moreover, obesity causes respiratory dysfunction, including impaired respiratory physiology, increased airway resistance, impaired gas exchange, low lung volume, and low muscle strength, which can also increase the risk of COVID-19 severity. Furthermore, the physical characteristics of obese individuals render intubation and laryngoscopy difficult, which could also aggravate outcomes (41). Further studies are needed to explore the pathways linking adiposity to increased risk of COVID-19 severity.

This study has several strengths. We used an MR design, which minimized bias from reverse causation and confounders, thereby enabling us to test for causal effects, provided compliance with MR assumptions. In this MR study, we used the data from the UK Biobank for the exposure traits (*F*-statistics > 10 for all exposure traits) and COVID-19 Host Genetics Initiative for the outcomes, both of which have large sample sizes, thus increasing the statistical power of the analysis. Furthermore, as proxy measures of body composition, we not only considered BMI, which is a common indirect measure, but also direct measures, including body fat percentage, body fat mass, and body fat-free mass, and revealed associations of these traits with COVD-19 severity.

Our study also has important limitations. First, MR analysis relies on several key assumptions, the violation of which compromises causal inference: relevance, independence, and exclusion restriction (Figure 2). To test for possible violations of these assumptions, we performed multiple sensitivity analyses. The MR-Egger intercept test did not detect horizontal pleiotropy. Although heterogeneity of effects was detected

for certain SNPs when analyzing COVID-19 hospitalization, the removal of outlier SNPs via MR-PRESSO still showed results consistent with those from MR inverse variance weighted method. We believe that these sensitivity analyses demonstrate the robustness and validity of the present findings. However, we acknowledge that horizontal pleiotropy is difficult to exclude entirely. Second, regarding exposure traits, we used measures derived from the bioelectrical impedance analysis (i.e., body fat percentage, body fat mass, and body fat-free mass) instead of DXA-derived measures to maximize statistical power. Although the UK Biobank collected DXA-derived measures for body fat mass and body fat-free mass, the sample size was markedly smaller for these measurements (n = 5,170). Moreover, although DXA-derived measures are generally more accurate than impedance-derived measures, high correlations between the two were reported for fat mass (r = 0.96) and fat-free mass (r = 0.86) in the UK Biobank dataset (9). Hence, we believe impedance-derived measures can serve as clinically relevant exposure traits in the present analysis. Third, we only used summary-level data and did not use individual-level data. Therefore, we could not evaluate the nonlinear relationship between exposures and outcomes. However, it should be noted that MR using summary statistics can still test for the presence of causal effects of exposures on outcomes, even if the exposure-outcome relationship is nonlinear (42). Additionally, a recent prospective cohort study of 6.9 million individuals in the UK suggested that BMI and COVID-19 severity have a linear relationship within a BMI range ≥23 kg/m^2 (2). Notably, the BMI of a majority of the individuals in the UK Biobank population included in the present analysis fell within this range (≥23 kg/m^2). Fourth, we restricted our analysis to individuals of European ancestry given that majority of participants in the UK Biobank were of European ancestry. Future studies are warranted to evaluate the generalizability of our findings to other populations. Lastly, we did not evaluate other clinically established risk factors such as diabetes, respiratory, heart, kidney, liver, autoimmune disorders, older age, smoking, and lower socioeconomic status (43). When considering risk factors for COVID-19 severity, we have to take into account phenotypic and genetic correlations. This was highlighted by a recent study showing that the causal effect of diabetes on COVID-19 severity is mediated by BMI (44). Another study also showed that the effect of BMI on severe

COVID-19 is partially mediated by socioeconomic status measured by household income (27). Furthermore, obesity is associated with other risk factors for severe COVID-19, including, but not limited to, chronic obstructive lung disease, heart failure, chronic kidney disease, liver cirrhosis and autoimmune disorders (37, 45, 46). The interconnected nature of these risk factors highlights the importance of disentangling the independent causal effect of each risk factor, which requires further investigation.

In summary, the present MR study provides evidence that indicates a causal relationship between body fat accumulation and COVID-19 severity. Because excess fat can be reduced by following an appropriate diet and exercising, it might represent an important modifiable risk factor. Thus, body weight reduction considering direct measurements of body fat (i.e., body fat percentage and body fat mass) can be an effective strategy to reduce the risk of COVID-19 severity.

2.6 Conflict of Interest

JBR's institution has received investigator-initiated grant funding from Eli Lilly, GlaxoSmithKline and Biogen for projects unrelated to this research. He is the founder of 5 Prime Sciences (www.5primesciences.com), which provides research services for biotech, pharma and venture capital companies for projects unrelated to this research. NI received research funds from Terumo Corp., Drawbridge, Inc., and Asken Inc. NI received speaker honoraria from Kowa Co., Ltd., MSD K.K, Astellas Pharma Inc., Novo Nordisk Pharma Ltd., Ono Pharmaceutical Co., Ltd., Nippon Boehringer Ingelheim Co., Ltd., Takeda Pharmaceutical Co., Ltd., Mitsubishi Tanabe Pharma Corp., Sumitomo Dainippon Pharma Co., Ltd., Sanofi K.K., Eli Lilly Japan K.K.; received scholarship grant from Kissei Pharmaceutical Co., Ltd., Sanofi K.K., Daiichi-Sankyo Co., Ltd., Mitsubishi Tanabe Pharma Corp., Takeda Pharmaceutical Co., Ltd., Japan Tobacco Inc., Kyowa Kirin Co., Ltd., Sumitomo Dainippon Pharma Co., Ltd., Astellas Pharma Inc., MSD K.K., Ono Pharmaceutical Co., Ltd., Sanwa Kagaku Kenkyusho Co., Ltd., Nippon Boehringer Ingelheim Co., Ltd., Novo Nordisk Pharma Ltd., Novartis Pharma K.K., and Life Scan Japan K.K. NI is an advisory board member of Novo Nordisk. These agencies did not

play any role in study design; the collection, analysis, or interpretation of data; the writing of the report; or the decision to submit this paper for publication. The other authors declare no conflict of interests.

2.7 Author Contributions

SY conceptualized and analyzed the data. SY, HM, and JBR wrote the original draft of the manuscript. JBR and NY supervised the study. All authors discussed the results and contributed to the final manuscript.

2.8 Funding

The Richards research group is supported by the Canadian Institutes of Health Research (CIHR: 365825; 409511, 100558, 169303), the McGill Interdisciplinary Initiative in Infection and Immunity (MI4), the Lady Davis Institute of the Jewish General Hospital, the Jewish General Hospital Foundation, the Canadian Foundation for Innovation, the NIH Foundation, Cancer Research UK, Genome Québec, the Public Health Agency of Canada, McGill University, Cancer Research UK [grant umber C18281/A29019] and the Fonds de Recherche Québec Santé (FRQS). JBR is supported by an FRQS Mérite Clinical Research Scholarship. Support from Calcul Québec and Compute Canada is acknowledged. TwinsUK is funded by the Welcome Trust, Medical Research Council, European Union, the National Institute for Health Research (NIHR)-funded BioResource, Clinical Research Facility and Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust in partnership with King's College London. These funding agencies had no role in the design, implementation or interpretation of this study. SY and HM are supported by the Japan Society for the Promotion of Science.

2.9 Data Availability

All GWAS summary statistics used in this study are publicly available. All results are included in the present article.

2.10 References

1. World Health Organization. WHO Coronavirus (COVID-19) Dashboard. Accessed on 5 June, 2022. https://covid19.who.int.
2. Gao M, Piernas C, Astbury NM, Hippisley-Cox J, O'Rahilly S, Aveyard P, *et al.* Associations between body-mass index and COVID-19 severity in 6·9 million people in England: a prospective, community-based, cohort study. *Lancet Diabetes Endocrinol*. 2021;9(6):350-9.
3. Recalde M, Pistillo A, Fernandez-Bertolin S, Roel E, Aragon M, Freisling H, *et al.* Body Mass Index and Risk of COVID-19 Diagnosis, Hospitalization, and Death: A Cohort Study of 2 524 926 Catalans. *J Clin Endocrinol Metab*. 2021;106(12):e5040-e2.
4. Recalde M, Roel E, Pistillo A, Sena AG, Prats-Uribe A, Ahmed W-U-R, *et al.* Characteristics and outcomes of 627 044 COVID-19 patients living with and without obesity in the United States, Spain, and the United Kingdom. *Int J Obes*. 2021;45:2347-57.
5. Helvaci N, Eyupoglu ND, Karabulut E, Yildiz BO. Prevalence of Obesity and Its Impact on Outcome in Patients With COVID-19: A Systematic Review and Meta-Analysis. *Front Endocrinol (Lausanne)*. 2021;12:598249.
6. Rothman KJ. BMI-related errors in the measurement of obesity. *Int J Obes*. 2008;32(S3):S56-S9.
7. Gao M, Wang Q, Piernas C, Astbury NM, Jebb SA, Holmes MV, *et al*. Associations between body composition, fat distribution and metabolic consequences of excess adiposity with severe COVID-19 outcomes: observational study and Mendelian randomisation analysis. *Int J Obes (Lond)*. 2022(46):943-50.
8. Sun Y, Zhou J, Ye K. Extensive Mendelian randomization study identifies potential causal risk factors for severe COVID-19. *Commun Med*. 2021;1(1).
9. Tikkanen E, Gustafsson S, Knowles JW, Perez M, Burgess S, Ingelsson E. Body composition and atrial fibrillation: a Mendelian randomization study. *Eur Heart J*. 2019;40(16):1277-82.
10. Grimes DA, Schulz KF. Bias and causal associations in observational research. *Lancet*. 2002;359(9302):248-52.

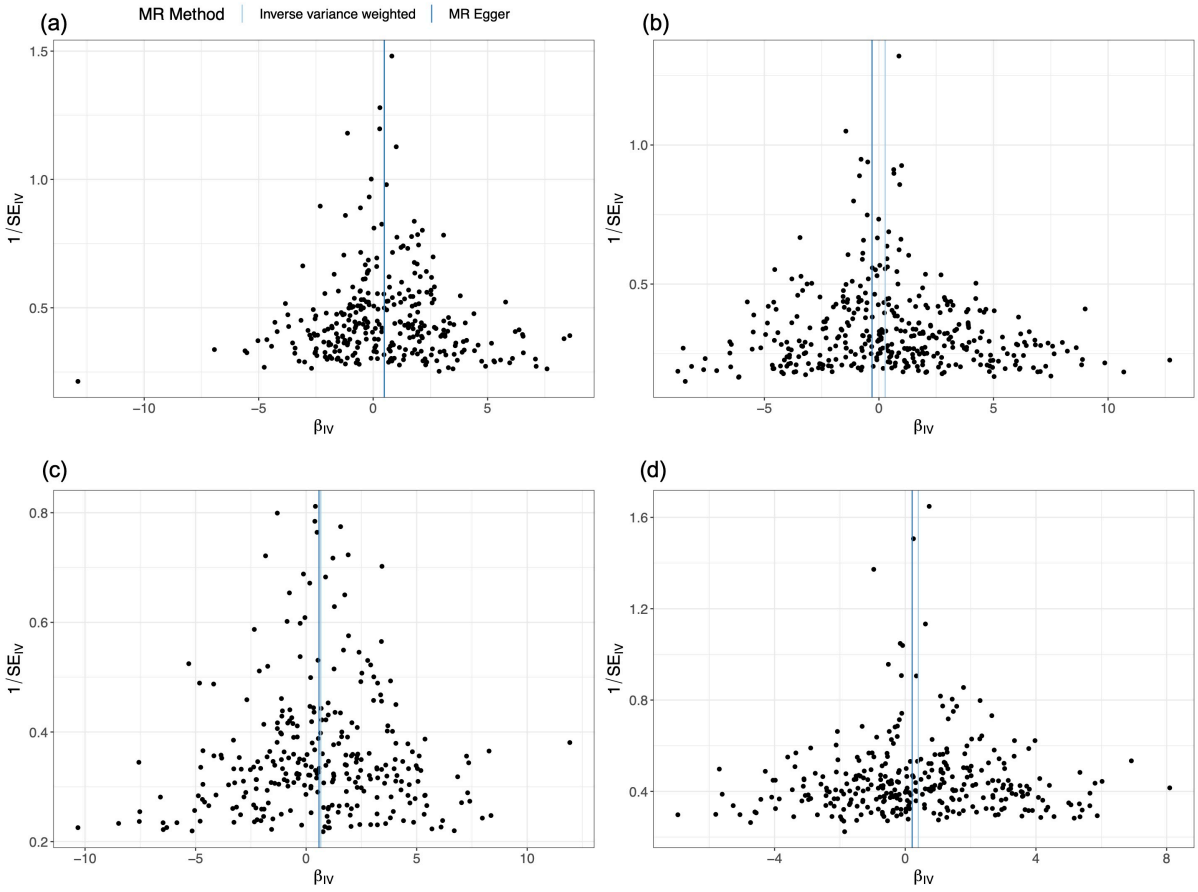
11. Skrivankova VW, Richmond RC, Woolf BAR, Yarmolinsky J, Davies NM, Swanson SA, *et al*. Strengthening the Reporting of Observational Studies in Epidemiology Using Mendelian Randomization. *JAMA*. 2021;326(16):1614.
12. Skrivankova VW, Richmond RC, Woolf BAR, Davies NM, Swanson SA, VanderWeele TJ, *et al*. Strengthening the reporting of observational studies in epidemiology using mendelian randomisation (STROBE-MR): explanation and elaboration. *BMJ*. 2021;375:n2233.
13. Ponsford MJ, Gkatzionis A, Walker VM, Grant AJ, Wootton RE, Moore LSP, *et al*. Cardiometabolic Traits, Sepsis, and Severe COVID-19: A Mendelian Randomization Investigation. *Circulation*. 2020;142(18):1791-3.
14. Lorincz-Comi N, Zhu X. Cardiometabolic risks of SARS-CoV-2 hospitalization using Mendelian Randomization. *Sci Rep*. 2021;11(1):7848.
15. Cecelja M, Lewis CM, Shah AM, Chowienczyk P. Cardiovascular health and risk of hospitalization with COVID-19: A Mendelian Randomization study. *JRSM Cardiovasc Dis*. 2021;10:20480040211059374.
16. 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*. 2018;7:e34408.
17. Mitchell. R, Elsworth B, Mitchell R, Raistrick C, Paternoster L, Hemani G, *et al*. MRC IEU UK Biobank GWAS pipeline version 2. 2019. https://doi.org/10.5523/bris.pnoat8cxo0u52p6ynfaekeigi.
18. Loh PR, Kichaev G, Gazal S, Schoech AP, Price AL. Mixed-model association for biobank-scale datasets. *Nat Genet*. 2018;50(7):906-8.
19. Speed D, Holmes J, Balding DJ. Evaluating and improving heritability models using summary statistics. *Nat Genet*. 2020;52(4):458-62.
20. Initiative C-HG. Mapping the human genetic architecture of COVID-19. *Nature*. 2021(600):472-7.
21. Lawlor DA, Harbord RM, Sterne JAC, Timpson N, Davey Smith G. Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Stat Med*. 2008;27(8):1133-63.

22. Yoshiji S, Tanaka D, Minamino H, Murakami T, Fujita Y, Richards JB, Inagaki N (2022). Supplementary material for “Causal association between body fat accumulation and COVID-19 severity: A Mendelian randomization study”. *Figshare.* Deposited May 12, 2022. https://figshare.com/s/c876361a354038ea988f?file=35063836.
23. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol*. 2015;44(2):512-25.
24. 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. *Nat Genet*. 2018;50(5):693-8.
25. Hippisley-Cox J, Coupland CA, Mehta N, Keogh RH, Diaz-Ordaz K, Khunti K, *et al*. Risk prediction of covid-19 related death and hospital admission in adults after covid-19 vaccination: national prospective cohort study. *BMJ*. 2021:n2244.
26. Aung N, Khanji MY, Munroe PB, Petersen SE. Causal Inference for Genetic Obesity, Cardiometabolic Profile and COVID-19 Susceptibility: A Mendelian Randomization Study. *Front Genet*. 2020;11:586308.
27. Cabrera-Mendoza B, Wendt FR, Pathak GA, De Angelis F, De Lillo A, Koller D, *et al*. The effect of obesity-related traits on COVID-19 severe respiratory symptoms is mediated by socioeconomic status: a multivariable Mendelian randomization study. *medRxiv* 2021.06.08.21258587. https://doi.org/10.1101/2021.06.08.21258587.
28. Freuer D, Linseisen J, Meisinger C. Impact of body composition on COVID-19 susceptibility and severity: A two-sample multivariable Mendelian randomization study. *Metabolism*. 2021;118:154732.
29. Leong A, Cole JB, Brenner LN, Meigs JB, Florez JC, Mercader JM. Cardiometabolic risk factors for COVID-19 susceptibility and severity: A Mendelian randomization analysis. *PLoS Med*. 2021;18(3):e1003553.
30. Li S, Hua X. Modifiable lifestyle factors and severe COVID-19 risk: a Mendelian randomisation study. *BMC Medical Genom*. 2021;14(1).
31. Richardson TG, Fang S, Mitchell RE, Holmes MV, Davey Smith G. Evaluating the effects of cardiometabolic exposures on circulating proteins which may contribute to severe SARS-CoV-2. *EBioMedicine*. 2021;64:103228.

32. Luo S, Liang Y, Wong THT, Schooling CM, Au Yeung SL. Identifying factors contributing to increased susceptibility to COVID-19 risk: a systematic review of Mendelian randomization studies. *Int J Epidemiol*. 2022: dyac076.
33. Larsson SC, Bäck M, Rees JMB, Mason AM, Burgess S. Body mass index and body composition in relation to 14 cardiovascular conditions in UK Biobank: a Mendelian randomization study. *Eur Heart J*. 2020;41(2):221-6.
34. Zeng H, Lin C, Wang S, Zheng Y, Gao X. Genetically predicted body composition in relation to cardiometabolic traits: a Mendelian randomization study. *Eur J Epidemiol*. 2021;36:1157–68.
35. Speed MS, Jefsen OH, Børglum AD, Speed D, Østergaard SD. Investigating the association between body fat and depression via Mendelian randomization. *Transl Psychiatry*. 2019;9(1).
36. Steenblock C, Schwarz PEH, Ludwig B, Linkermann A, Zimmet P, Kulebyakin K, *et al*. COVID-19 and metabolic disease: mechanisms and clinical management. *Lancet Diabetes Endocrinol*. 2021;9(11):786-98.
37. Gammone MA, D'Orazio N. Review: Obesity and COVID-19: A Detrimental Intersection. *Front Endocrinol (Lausanne)*. 2021;12:652639.
38. Foulkes AS, Selvaggi C, Shinnick D, Lumish H, Kim E, Cao T, *et al*. Understanding the link between obesity and severe COVID-19 outcomes: Causal mediation by systemic inflammatory response. *J Clin Endocrinol Metab*. 2021;107:e698-e707.
39. Zickler M, Stanelle-Bertram S, Ehret S, Heinrich F, Lange P, Schaumburg B, *et al.* Replication of SARS-CoV-2 in adipose tissue determines organ and systemic lipid metabolism in hamsters and humans. *Cell Metab*. 2022;34(1):1-2.
40. Reiterer M, Rajan M, Gomez-Banoy N, Lau JD, Gomez-Escobar LG, Ma L, *et al*. Hyperglycemia in acute COVID-19 is characterized by insulin resistance and adipose tissue infectivity by SARS-CoV-2. *Cell Metab*. 2021;33(11):2174-88 e5.
41. Stefan N, Birkenfeld AL, Schulze MB, Ludwig DS. Obesity and impaired metabolic health in patients with COVID-19. *Nat Rev Endocrinol*. 2020;16(7):341-2.
42. Burgess S, Davies NM, Thompson SG. Instrumental Variable Analysis with a Nonlinear Exposure–Outcome Relationship. *Epidemiology*. 2014;25(6):877-85.

43. Williamson EJ, Walker AJ, Bhaskaran K, Bacon S, Bates C, Morton CE, *et al*. Factors associated with COVID-19-related death using OpenSAFELY. *Nature*. 2020;584(7821):430-6.
44. COVID-19 Host Genetics Initiative, Ganna A. Mapping the human genetic architecture of COVID-19: an update. *medRxiv*. 2022:2021.11.08.21265944. https://doi.org/10.1101/2021.11.08.21265944.
45. Lin X, Li H. Obesity: Epidemiology, Pathophysiology, and Therapeutics. *Front Endocrinol (Lausanne)*. 2021;12:706978.
46. Kivimäki M, Strandberg T, Pentti J, Nyberg ST, Frank P, Jokela M, *et al.* Body-mass index and risk of obesity-related complex multimorbidity: an observational multicohort study. *Lancet Diab Endocrinol*. 2022;10(4):253-63.

2.11 Supplementary Figure



Supplementary Figure 1. Funnel plots of the univariable weighted MR for (a) body fat mass, (b) body fat-free mass, (c) body fat percentage, and (d) body fat mass.
Each dot represents a genetic instrumental variable. Two lines represent causal estimate (β_{IV}) by the inverse variance weighted method (light blue) and the MR-Egger method (blue). SE_{IV} represents standard error for each genetic instrumental variable. Error bars represent 95%CIs.
MR, Mendelian randomization; IV, genetic instrumental variable.

Transition from Chapter 2 to Chapter 3

In Chapter 2, using various proxies for obesity and MR, we demonstrated that obesity is associated with an increased risk of COVID-19 severity. Additionally, considering the genetic interrelation between body fat mass and fat-free mass, we employed multivariable MR to distinguish their independent causal effects. Our findings showed that body fat mass is independently associated with an increased risk of COVID-19 severity, underscoring the causal role of body fat accumulation. Yet, the specific mechanisms by which obesity affects COVID-19 severity remained elusive.

In Chapter 3, we sought to gain insights into this underlying mechanism. Considering that obesity substantially influences the plasma proteome, we aimed to pinpoint circulating proteins that may serve as mediators between obesity and COVID-19 severity. In this chapter, we integrated large-scale genomics and proteomics data with genetic epidemiology methods such as MR, colocalization, mediation analyses and single-cell RNA sequencing analysis. This integration enriched our understanding of the causal biology and aided in identifying potential therapeutic targets. Notably, during our analysis for Chapter 3, the COVID-19 Host Genetics Initiative released new sets of GWAS on COVID-19 severity. This updated GWAS more than doubled the sample size, allowing us to incorporate it into our Chapter 3 analyses, thereby enhancing our statistical power.

Chapter 3: Proteome-wide Mendelian randomization implicates nephronectin as an actionable mediator of the effect of obesity on COVID-19 severity

Satoshi Yoshiji[1,2,3,4], Guillaume Butler-Laporte[1,5], Tianyuan Lu[1,6,7], Julian Daniel Sunday Willett[1,6], Chen-Yang Su[1,8], Tomoko Nakanishi[1,2,3,4], David R. Morrison[1], Yiheng Chen[1,2], Kevin Liang[1,6], Michael Hultström[1,5,9,10], Yann Ilboudo[1], Zaman Afrasiabi[1], Shanshan Lan[1], Naomi Duggan[1], Chantal DeLuca[1], Mitra Vaezi[1], Chris Tselios[1], Xiaoqing Xue[1], Meriem Bouab[1], Fangyi Shi[1], Laetitia Laurent[1], Hans Markus Münter[11], Marc Afilalo[1,12], Jonathan Afilalo[1,5,13], Vincent Mooser[2,11], Nicholas J Timpson[14], Hugo Zeberg[15,16], Sirui Zhou[1,2,11], Vincenzo Forgetta[1,7], Yossi Farjoun[1], J. Brent Richards[1,2,5,7,17*]

[1] Lady Davis Institute, Jewish General Hospital, McGill University, Montréal, Québec, Canada
[2] Department of Human Genetics, McGill University, Montréal, Québec, Canada
[3] Kyoto-McGill International Collaborative Program in Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto, Japan
[4] Japan Society for the Promotion of Science, Japan
[5] Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, Québec, Canada
[6] Quantitative Life Sciences Program, McGill University, Montréal, Québec, Canada
[7] 5 Prime Sciences, Montréal, Québec, Canada
[8] Department of Computer Science, McGill University, Montréal, Québec, Canada
[9] Anaesthesiology and Intensive Care Medicine, Department of Surgical Sciences, Uppsala University, Sweden
[10] Integrative Physiology, Department of Medical Cell Biology, Uppsala University, Uppsala, Sweden
[11] McGill Genome Centre, McGill University, Montréal, Québec, Canada
[12] Department of Emergency Medicine, Jewish General Hospital, McGill University, Montréal, Québec, Canada

[13] Division of Cardiology, Jewish General Hospital, McGill University, Montréal, Québec, Canada
[14] MRC Integrative Epidemiology Unit, University of Bristol, Bristol, United Kingdom
[15] Department of Physiology and Pharmacology, Karolinska Institutet, Stockholm, Sweden
[16] Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
[17] Department of Twin Research, King's College London, London, United Kingdom

3.1 Abstract

Obesity is a major risk factor for COVID-19 severity; however, the mechanisms underlying this relationship are not fully understood. Since obesity influences the plasma proteome, we sought to identify circulating proteins mediating the effects of obesity on COVID-19 severity in humans. Here, we screened 4,907 plasma proteins to identify proteins influenced by body mass index (BMI) using Mendelian randomization (MR). This yielded 1,216 proteins, whose effect on COVID-19 severity was assessed, again using MR. We found that a standard deviation increase in nephronectin (NPNT) was associated with increased odds of critically ill COVID-19 (OR = 1.71, $P = 1.63 \times 10^{-10}$). The effect was driven by an NPNT splice isoform. Mediation analyses supported NPNT as a mediator. In single-cell RNA-sequencing, *NPNT* was expressed in alveolar cells and fibroblasts of the lung in individuals who died of COVID-19. Finally, decreasing body fat mass and increasing fat-free mass were found to lower NPNT levels. These findings provide actionable insights into how obesity influences COVID-19 severity.

3.2 Introduction

Coronavirus disease (COVID-19) has claimed more than 6 million lives globally since the beginning of the pandemic[1], and obesity increases the risk of severe COVID-19[2,3]. To date, multiple pathways have been explored as mechanisms linking obesity to COVID-19 severity, including metabolic abnormalities, systemic inflammation, and respiratory dysfunction (e.g., impaired gas exchange) [3-5]. However, the underlying mediators whereby obesity influences COVID-19 outcomes are not fully understood. One strategy to disentangle this relationship is to identify circulating proteins mediating the effects of obesity on COVID-19 outcomes. Since circulating proteins can be measured and, in some cases modulated, the identification of mediator proteins may provide insights into the pathways whereby obesity increases the risk of severe COVID-19 and offer potential targets for therapeutic interventions.

A previous large cross-sectional study showed that body mass index (BMI) is significantly associated with changes in the plasma levels of 1,576 proteins[6], and another study supported the considerable influence of obesity on the plasma proteome[7]. However, since proteins are intricately involved in complex biological processes, observational studies may be biased by unmeasured confounding factors and reverse causation. Considering that COVID-19 has been associated with substantial changes in the levels of circulating proteins[8], the effects of circulating proteins on COVID-19 severity are also subject to such biases.

Mendelian randomization (MR) can help to protect against such biases. MR is a method that uses genetic variants as instrumental variables to evaluate the causal effects of exposures (risk factors) on outcomes. Since genetic variants are randomly allocated at conception, they are largely independent confounders, thereby decreasing the risk of confounding. Additionally, they are not subject to reverse causation since the allocation of genetic variants always precedes the onset of diseases[9,10].

From the beginning of the pandemic, MR has played a critical role in providing evidence of modifiable risk factors for COVID-19[2,11,12]. For example, multiple MR

studies have identified potential therapeutic targets for COVID-19, including OAS1, ABO, IFNAR2, IL-6, ELF5, and FAS[12-18]. Indeed, some MR findings have been validated by randomized controlled trials, thereby demonstrating the utility of MR[19-24].

Nevertheless, MR is based on several instrumental variable assumptions: [9,10] (I) the genetic variants used as instrumental variables are associated with the exposure; (II) they are not associated with factors that confound the relationship between the exposure and the outcome and; (III) they influence the outcome only through exposure (also known as exclusion restriction). The most problematic of these assumptions is the last, since the violation of exclusion restriction can bias the causal estimate through directional horizontal pleiotropy; therefore, careful assessment of directional horizontal pleiotropy is required. However, with a proper selection of instrumental variables and sensitivity analyses, MR can serve as a powerful tool to help understand causal mechanisms for disease in humans[9,10], and in the case of obesity and COVID-19, can be used to screen thousands of proteins that may help to explain this relationship.

Here, we integrated proteome-wide MR using the large-scale aptamer-based plasma protein measurements, multiple sensitivity analyses, colocalization, fine-mapping, single-cell RNA-sequencing analysis, and mediation analysis to identify plasma proteins mediating the effect of obesity on COVID-19 severity.

3.3 Results

3.3.1 Study overview and summary

The study was conducted in the following manner (**Figure 1**):

1) **Step 1 MR**

   First, we estimated the effect of BMI on circulating protein levels using two-sample MR on a proteome-wide scale. Two-sample MR can estimate the causal effect of the exposure on the outcome using summary statistics of genome-wide association studies (GWAS). For this, we used a GWAS of BMI in 694,649

individuals from GIANT and UK Biobank by Yengo *et al*.[25], and plasma proteome GWAS from the deCODE study[26] that measured plasma protein abundances using 4,907 aptamers in 35,559 individuals. Hereafter, "protein" is used when referring to an aptamer targeting a protein[26]. After two-sample MR and sensitivity analyses, 1,216 proteins were estimated to be influenced by BMI, including nephronectin (NPNT) and hydroxysteroid 17-beta dehydrogenase 14 (HSD17B14).

2) **Step 2 MR**

Next, we performed two-sample MR to estimate the causal effects of the above-identified proteins (BMI-driven proteins) on critically ill COVID-19 and COVID-19 hospitalization outcomes (collectively referred to as COVID-19 severity outcomes). For this analysis, we used cis-acting protein quantitative loci (cis-pQTLs) from the deCODE study[26], thereby minimizing the risk of directional horizontal pleiotropy. For COVID-19 severity outcomes, we used GWAS data from the COVID-19 Host Genetics Initiative[2]. This step 2 MR identified NPNT and HSD17B14 as putatively causal proteins for the COVID-19 severity outcomes.

3) Validation analyses for NPNT and HSD17B14

We performed multiple validation analyses for step 1 and step 2 MR as follows:

Step 1 MR validation

To validate whether NPNT and HSD17B14 were influenced by obesity, we performed separate MR analyses using body fat percentage (another proxy measure for obesity) as the exposure and plasma protein levels as the outcomes. The MR analysis found that both NPNT and HSD17B14 were increased by body fat percentage, consistent with the step 1 MR using BMI.

We also checked whether NPNT and HSD17B14 were observationally associated with BMI using a published observational association study from INTERVAL (n = 2,729) and found that NPNT, but not HSD17B14, was positively associated with BMI.

These validation analyses collectively supported the causal effect of obesity on NPNT, but not on HSD17B14.

Step 2 MR validation

To ensure that the findings in step 2 MR were not biased by linkage disequilibrium (LD), which can reintroduce confounding, we used colocalization analyses to evaluate whether the cis-pQTLs of the identified proteins and COVID-19 severity outcomes shared a single causal variant. We found that NPNT, but not HSD17B14, shared a single causal variant with COVID-19 severity outcomes, indicating that the MR estimates for HSD17B14 in step 2 MR could have been biased by LD.

In addition, we repeated MR using cis-pQTLs from two independent studies (the FENLAND study[27] and the AGES Reykjavik study[28]), which showed that NPNT, but not HSD17B14, influenced the COVID-19 severity outcomes. Given the lack of colocalization and replication for HSD17B14, we excluded HSD17B14 from further analyses.

For NPNT, we further investigated the consistency between MR findings and observational associations between NPNT and COVID-19 severity outcomes using the BQC19 cohort, which showed the same direction of effect as the MR analyses.

Collectively, these findings showed a causal effect of plasma NPNT levels on COVID-19 severity outcomes.

4) **Follow-up analyses for NPNT**

Colocalization of NPNT's cis-pQTL with eQTL and sQTL:

The observed differences in aptamer-measured NPNT levels may be predominantly explained by a particular isoform, rather than total NPNT levels. Thus, we used colocalization to evaluate whether the cis-pQTL shared the same causal variant with either its expression QTL (eQTL; genetic variants that explain the total RNA expression levels of NPNT) or its splicing QTL (sQTL; genetic variants that explain a specific isoform level of NPNT). We found that the cis-

pQTL shared the same causal variant with the sQTL but not with the eQTL. These findings demonstrate that an NPNT splice isoform is likely measured by the aptamer targeting NPNT.

Single-cell RNA-sequencing data of SARS-CoV-2-infected lungs

Next, to gain insights into the biological role of NPNT in SARS-CoV-2-infected lungs, we analyzed single-cell RNA-sequencing data of the lung autopsy samples from patients who died due to COVID-19. We found that NPNT is significantly expressed in in alveolar cells and fibroblasts of the lung, indicating its role in air exchange and fibrosis.

Mediation analysis

Further, we performed an MR mediation analysis to quantify the extent to which the total effect of obesity on COVID-19 severity outcomes was mediated by plasma NPNT levels. We found that NPNT partially mediated the total effect of BMI, and consistent results were found for body fat percentage.

Multivariable MR analyses of body fat and fat-free mass

Finally, we estimated the independent causal effect of body fat and fat-free mass on plasma NPNT levels and COVID-19 severity outcomes using multivariable MR. We found that body fat mass increased plasma NPNT levels and the risk of COVID-19 severity outcomes, whereas fat-free mass decreased plasma NPNT levels and the risk of COVID-19 severity outcomes. These findings demonstrate that decreasing body fat and increasing body fat mass (e.g., through actions such as appropriate exercise and diet) can decrease plasma NPNT levels, and thus, may reduce COVID-19 severity outcomes, thereby suggesting NPNT as an actionable target.

Each of these steps is described in more detail below.

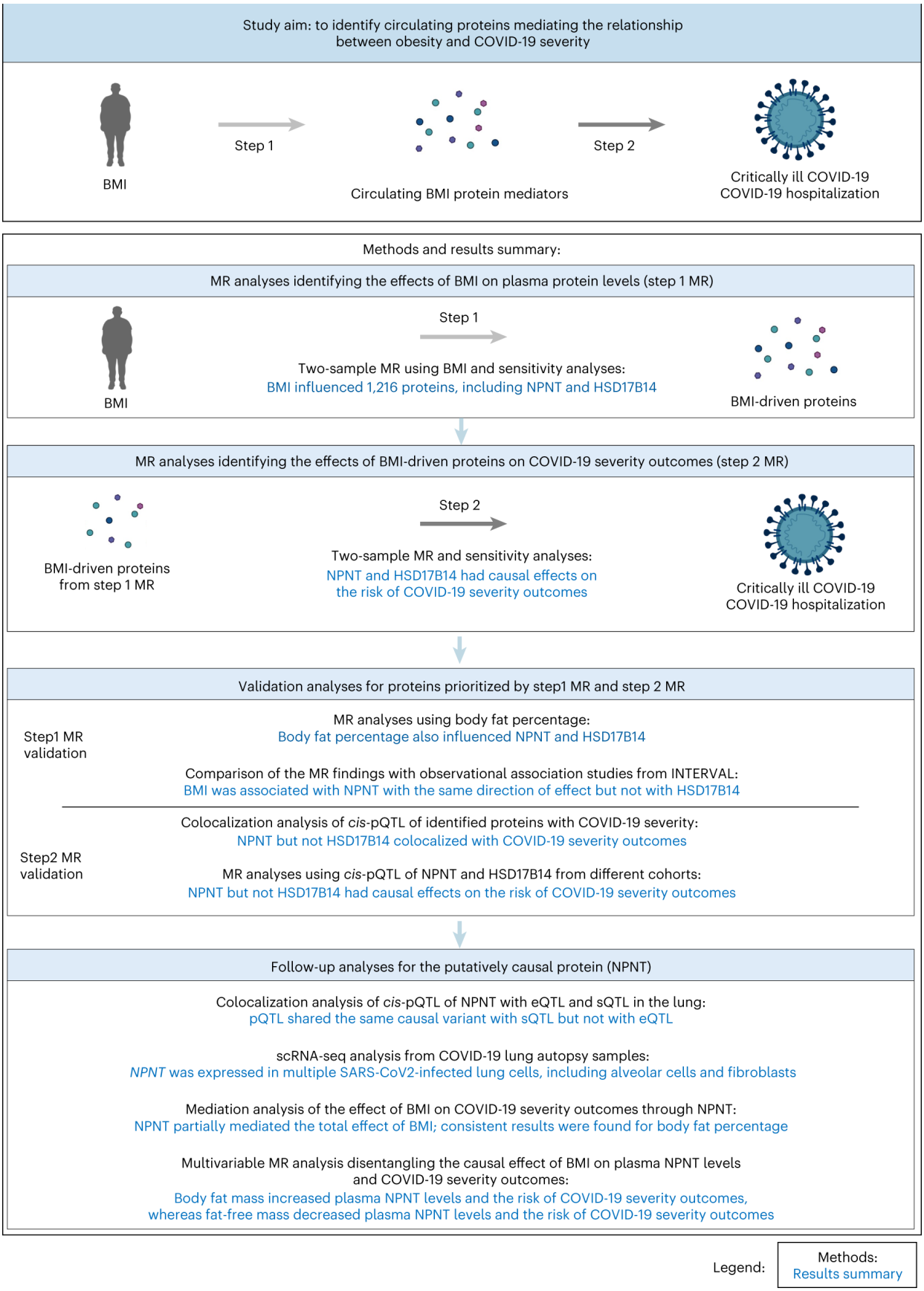


Figure 1. Study overview and summary.

We identified circulating proteins mediating the effect of obesity on COVID-19 severity using two-step MR approach: First, we estimated the effect of BMI on 4,907 plasma proteins using MR, which yielded 1,216 BMI-driven proteins (Step 1 MR). Second, we estimated the effect of the BMI-driven proteins on COVID-19 severity outcomes, again using MR (Step 2 MR). This was followed by multiple validity assessments and follow-up analyses.

MR: Mendelian randomization, BMI: body mass index, NPNT: nephronectin, HSD17B14: hydroxysteroid 17-beta dehydrogenase 14, cis-pQTL: cis-acting quantitative trait loci, e-QTL: expression quantitative trait loci, sQTL: splicing quantitative trait loci.

3.3.2 BMI to plasma proteins (Step 1 MR)

To estimate the causal effect of BMI on plasma protein levels on a proteome-wide scale, we performed a two-sample MR using BMI as the exposure and 4,907 plasma protein levels as the outcomes. For two-sample MR, we used an inverse variance weighted method with a random-effects model (see **Methods** and **Supplementary Table 1** for details). The F-statistic, which is a measure of the strength of the association between genetic variants and BMI, was 94.2 and thus did not indicate weak instrument bias (suspected when F-statistic < 10) [29] (**Supplementary Tables 2**). Out of the 4,907 proteins screened, 1,304 were estimated to be influenced by BMI, using a Bonferroni-adjusted threshold of $P < 1.0 \times 10^{-5}$ (0.05/4907), highlighting the substantial influence of BMI on plasma protein levels (**Figure 2. Supplementary Table 3 and 4**).

In sensitivity analyses, we tested the robustness of the MR findings with a heterogeneity test, directional pleiotropy test, and reverse causation test (see **Methods**) to only retain the proteins that were robustly influenced by BMI. We did not find significant heterogeneity for 1,304 Bonferroni-significant proteins ($I^2 < 50\%$ for all). Of 1,304 proteins, 1,229 showed no apparent sign of directional horizontal pleiotropy with the MR Egger intercept test[30] ($P_{\text{Egger intercept}} > 0.05$). MR analyses with the weighted median, weighted mode, and MR-Egger slope methods also showed directionally consistent results with the inverse variance weighted method for NPNT and HSD17B14 **(Figure 2** and **Supplementary Table 4)**. To assess potential reverse causation, whereby the proteins influenced BMI, we performed bidirectional MR that used protein levels as the exposures and BMI as the outcome (see **Methods**). Thirteen proteins exhibited bidirectional effects (**Supplementary Table 5**) and were excluded from further analyses.

Hence, a total of 1,216 protein levels were identified as BMI-driven proteins, which were estimated to be influenced by BMI with no apparent heterogeneity, directional pleiotropy, or reverse causation. We proceeded to the second step of our study (step 2 MR) with these 1,216 proteins, including NPNT and HSD17B14.

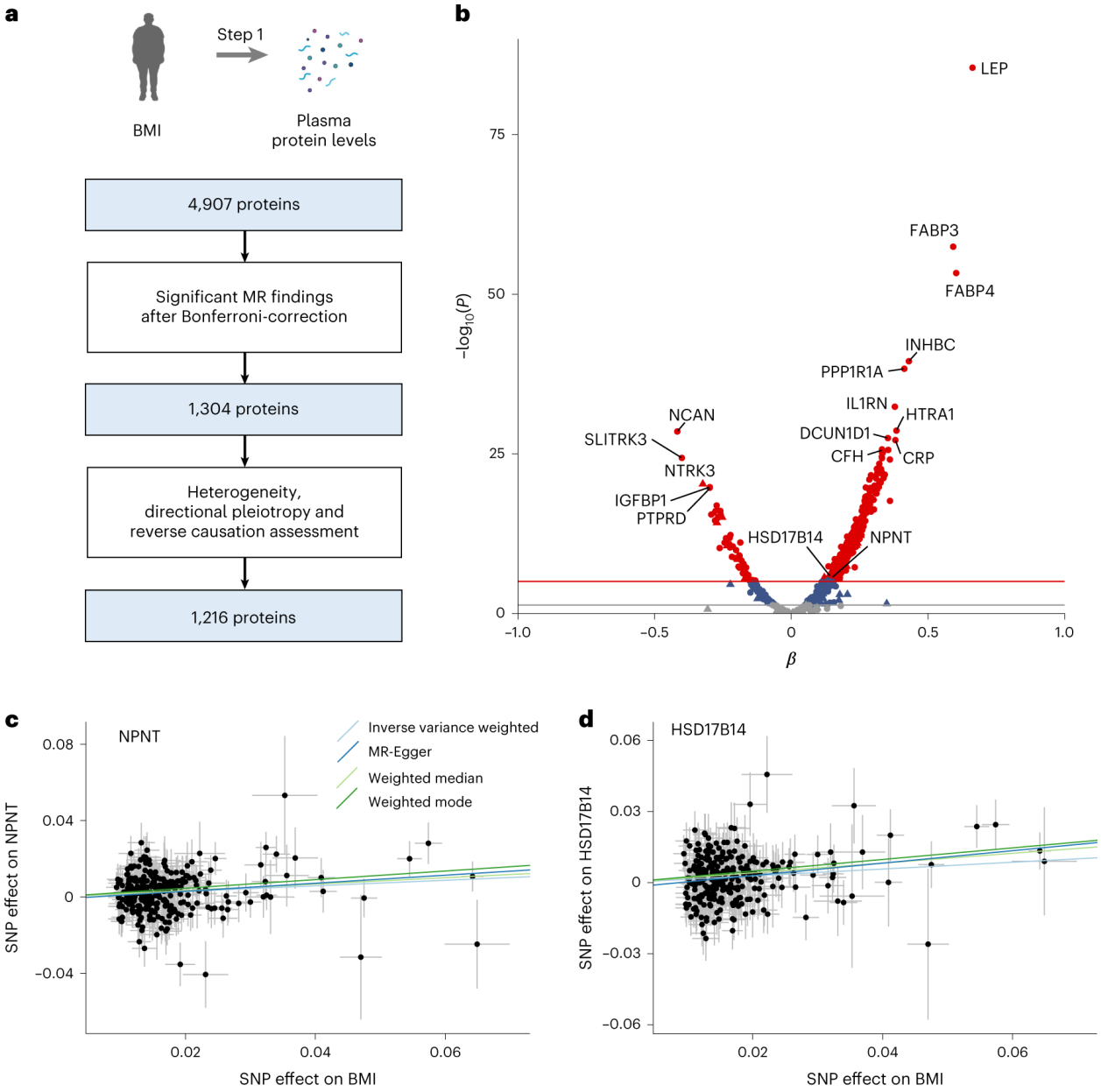


Figure 2. MR analyses for the effect of body mass index on plasma protein levels. (a) Flow diagram of the Step 1 MR analyses. (b) Volcano plot illustrating the effect of BMI on each plasma protein from the MR analyses using inverse variance weighted method. Red and blue horizontal lines represent $P = 1.0 \times 10^{-5}$ (Bonferroni correction for 4,907 proteins: 0.05/4,907) and 0.05, respectively. A proteins' shape denotes whether the protein passed all sensitivity tests (i.e., heterogeneity, directional pleiotropy, and reverse causation assessment) (circle) or failed any of them (triangle). (c) MR scatter plot for the effect of BMI on plasma NPNT levels. (d) MR scatter plot for the effect of

BMI on plasma HSD17B14 levels. A genetically predicted increase in BMI by one standard deviation was associated with increased levels of NPNT (beta = 0.145, 95% CI: 0.084–0.206, $P = 3.03 \times 10^{-6}$) and HSD17B14 (beta = 0.144, 95% CI: 0.085–0.202, $P = 1.71 \times 10^{-6}$) using the inverse variance weighted method. MR-Egger, weighted median, and weighted mode methods yielded directionally consistent results with the inverse variant weighted method.

MR: Mendelian randomization, BMI: body mass index, NPNT: nephronectin, HSD17B14: hydroxysteroid 17-beta dehydrogenase 14.

3.3.3 BMI-driven proteins to COVID-19 severity (Step 2 MR)

Next, we evaluated the causal effects of the above-obtained BMI-driven proteins on COVID-19 outcomes, again using two-sample MR. We used cis-pQTLs for these proteins as instrumental variables and GWASs from the COVID-19 Host Genetics Initiative (release 7) [2] as outcomes (**Supplementary Table 1)**. We used cis-pQTLs (pQTLs that reside within ± 1 Mb region around a transcription start site of a protein-coding gene) as the exposures to protect against bias from directional horizontal pleiotropy. This is because cis-pQTLs reside near the transcription start site of the protein-coding gene and are more likely to directly influence the protein levels than the trans-pQTLs[31,32]. Since cis-pQTLs are likely to directly influence the transcription or translation of their associated gene, the risk of directional horizontal pleiotropy would be greatly reduced.

We searched for the cis-pQTLs for 1,216 BMI-driven proteins using the deCODE study[26]. Following the cis-pQTL search and data harmonization, 358 and 352 proteins were tested in MR for their estimated causal effects on critically ill COVID-19 and COVID-19 hospitalization, respectively. The F-statistics for the tested proteins were all greater than 10, substantially reducing the risk of weak instrument bias[29]. F-statistics for NPNT and HSD17B14 were 252.5 and 66.3, respectively (**Supplementary Table 2**).

COVID-19 and COVID-19 hospitalization outcomes are collectively referred to as COVID-19 severity outcomes. Throughout the study, we focused on these two outcomes from the COVID-19 Host Genetics Initiative and did not include COVID-19 susceptibility outcome (reported COVID-19 infection). We did so because the determinants of COVID-19 susceptibility may reflect local testing strategy and resource allocation, which pose difficulties in the interpretation of genetic findings.

Based on a Bonferroni-adjusted threshold of $P < 1.40 \times 10^{-4}$, MR revealed that a one standard deviation (SD) increase in genetically predicted NPNT levels was associated with increased odds of critically ill COVID-19 (OR = 1.71, 95% CI: 1.45–2.02, $P = 1.63 \times 10^{-10}$) and COVID-19 hospitalization (OR = 1.36, 95% CI: 1.22–1.53, $P = 4.52 \times 10^{-8}$) (**Figure 3** and **Supplementary Table 6 and 7**). Similarly, a one SD

increase in genetically predicted HSD17B14 levels was associated with increased odds of critically ill COVID-19 (OR = 1.92, 95% CI: 1.39–2.65, $P = 6.85 \times 10^{-5}$).

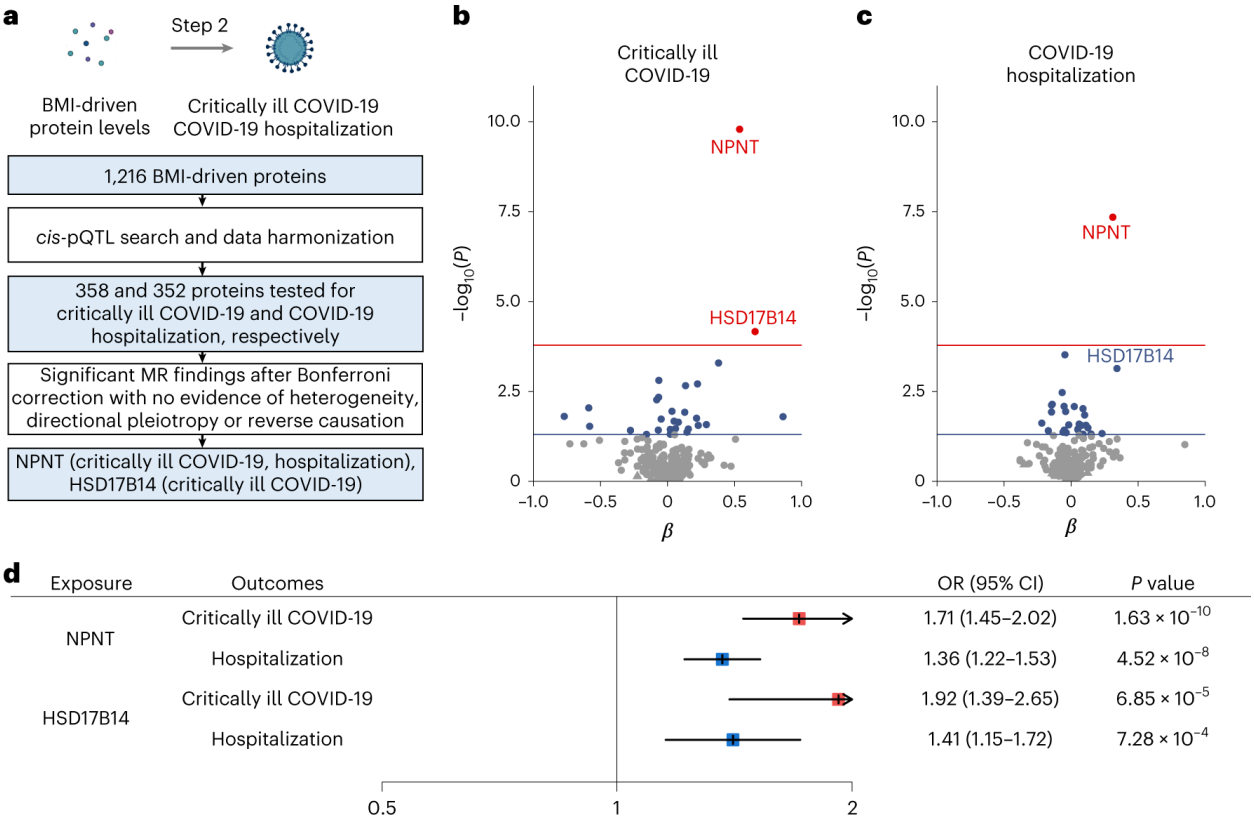


Figure 3. MR analyses of BMI-driven proteins on COVID-19 outcomes.

(a) Flow diagram of the Step 2 MR analyses. (b, c) Volcano plot illustrating the effect of BMI on critically ill COVID-19 and (b) COVID-19 hospitalization (c) from the MR analyses using the inverse variance weighted method or Wald ratio when only one SNP was available as an instrumental variable. Red and blue horizontal lines represent $P = 1.4 \times 10^{-4}$ (Bonferroni correction for 358 proteins: 0.05/358) and 0.05, respectively. A proteins' shape denotes whether the protein passed (circle) all sensitivity tests (i.e., heterogeneity, directional pleiotropy, and reverse causation assessment) or failed any of them (triangle). (d) Forest plot of the MR results for NPNT and HSD17B14, showing the odds ratio per one standard deviation increased in plasma levels of NPNT and HSD17B14 for critically ill COVID-19 and hospitalization outcomes.

MR: Mendelian randomization, BMI: body mass index, NPNT: nephronectin, HSD17B14: hydroxysteroid 17-beta dehydrogenase 14, OR: odds ratio, 95% CI: 95% confidence intervals.

To verify the assumption of a lack of directional pleiotropy which can reintroduce confounding, we checked whether the cis-pQTLs for NPNT and HSD17B14 were associated with any traits or diseases using the PhenoScanner (http://www.phenoscanner.medschl.cam.ac.uk/)[33] and Open Target Genetics (https://genetics.opentargets.org/) databases at the genome-wide significant threshold of $P = 5 \times 10^{-8}$. The lead cis-pQTL for NPNT (rs34712979) from the deCODE study was associated with lung-related traits (**Supplementary Table 8**). However, since NPNT has an established role as an extracellular matrix protein and fibrosis in the lungs[34-36], it is possible that the NPNT cis-pQTL affects such traits by altering NPNT levels, and thus should not violate the assumption of no directional pleiotropy. Indeed, MR showed that NPNT levels were estimated to influence the FEV1/FVC ratio (a key lung function index used for the definition of chronic obstructive lung disease (COPD) [37]) and asthma (**Supplementary Table 9**). This suggests that these findings may be a case of vertical pleiotropy, which does not bias MR interpretation[38, 39, 40]. Intriguingly, the NPNT-increasing rs1662979-G allele, which increases the risk of COVID-19 severity, was found to improve lung function (i.e., higher FEV1/FVC ratio) and decrease the risk of COPD (**Supplementary Table 8 and 9**). A similar phenomenon has been reported for ELF5 and MUCB5: the COVID-19 severity risk-increasing alleles of ELF5 and MUCB5 were observed to improve lung function and decrease the risk of idiopathic pulmonary fibrosis[18]. This suggests that COVID-19 has distinct underlying mechanisms that influence the severity risk. No other cis-pQTL for NPNT or HSD17B14 was associated with any trait or disease, thereby reducing the possibility of directional pleiotropy. Next, to assess potential bias from reverse causation, we performed the MR-Steiger test, which supported a causal direction of plasma NPNT and HSD17B14 levels influencing COVID-19 severity outcomes (**Supplementary Table 10**).

3.3.4 Validation analyses for NPNT and HSD17B14

3.3.4.1 Step 1 MR validation

MR analyses using body fat percentage

BMI is an easy-to-measure, widely used proxy for obesity, which can offer clinically relevant information. However, another proxy for obesity, body fat percentage, can more directly measures body fat accumulation. Thus, to evaluate whether circulating NPNT and HSD17B14 levels were influenced by body fat accumulation, we repeated step 1 MR using body fat percentage as the exposure instead of BMI. We used the same 4,907 plasma protein levels GWASs from the deCODE study[26] as the outcomes. The F-statistic for body fat percentage for these analyses was 61.1, which indicated no evidence of weak instrument bias.

We found that one SD increase in body fat percentage was associated with increased levels of NPNT (beta = 0.14, 95% CI: 0.07–0.22, $P = 1.23 \times 10^{-4,}$ and HSD17B14 (beta = 0.17, 95% CI: 0.10–0.24, $P = 3.61 \times 10^{-6}$), (see **Methods** and **Supplementary Table 11**), consistent with the step 1 MR findings with BMI.

Comparison with observational studies from INTERVAL

One way to assess potential biases is to test the same hypothesis using a different study design. Since each study design has its own inherent potential biases, similar results across designs can serve to strengthen causal inference through a triangulation of results[41]. Therefore, in supplementary analyses, we compared our MR findings for the effect of BMI on NPNT and HSD17B14 with published results from the INTERVAL study, which evaluated the associations between BMI and 3,622 plasma protein levels among 2,729 individuals[6].

In the INTERVAL study, a one SD increase in BMI was cross-sectionally associated with increased NPNT (beta = 0.13, 95% CI: 0.09–0.17, $P = 4.52 \times 10^{-10}$), which was directionally consistent with the MR findings. On the contrary, HSD17B14 showed inconsistent results, wherein a one SD increase in BMI was not associated with HSD17B14 (beta = -0.02, 95% CI: -0.05–0.02, $P = 0.415$).

These validation analyses collectively supported the causal effect of obesity on NPNT, but not on HSD17B14.

3.3.4.2 Step 2 MR validation

Colocalization of cis-pQTLs with COVID-19 severity outcomes

Considering that the MR analyses from Step 1 and Step 2 implicated NPNT and HSD17B14 as candidate proteins mediating the effect of BMI on COVID-19 severity, we performed colocalization to assess whether the cis-pQTLs for NPNT and HSD17B14 shared the same single causal variant with critically ill COVID-19 and hospitalization. This can test whether MR analyses were biased by LD.

The colocalization analyses revealed that NPNT had a high posterior probability of colocalization with critically ill COVID-19 and hospitalization (posterior probability (PP_{shared}) > 99.9% for both) (**Figure 4**). We confirmed the robustness of the colocalization results using different priors (**Methods and Supplementary Table 12**), which consistently showed that the cis-pQTL for NPNT colocalized with COVID-19 severity outcomes. Moreover, using Combined Annotation Dependent Depletion (CADD)-scores[42] as priors for fine-mapping, we confirmed that rs34712979, the lead cis-pQTL for NPNT, was a causal variant for critically ill COVID-19 and hospitalization with a posterior inclusion probability of > 0.99 (**Supplementary Table 13 and 14**). On the contrary, cis-pQTLs for HSD17B14 did not colocalize with either of the COVID-19 severity outcomes (**Figure 4**), indicating that the MR findings for HSD17B14 were likely to be biased by LD.

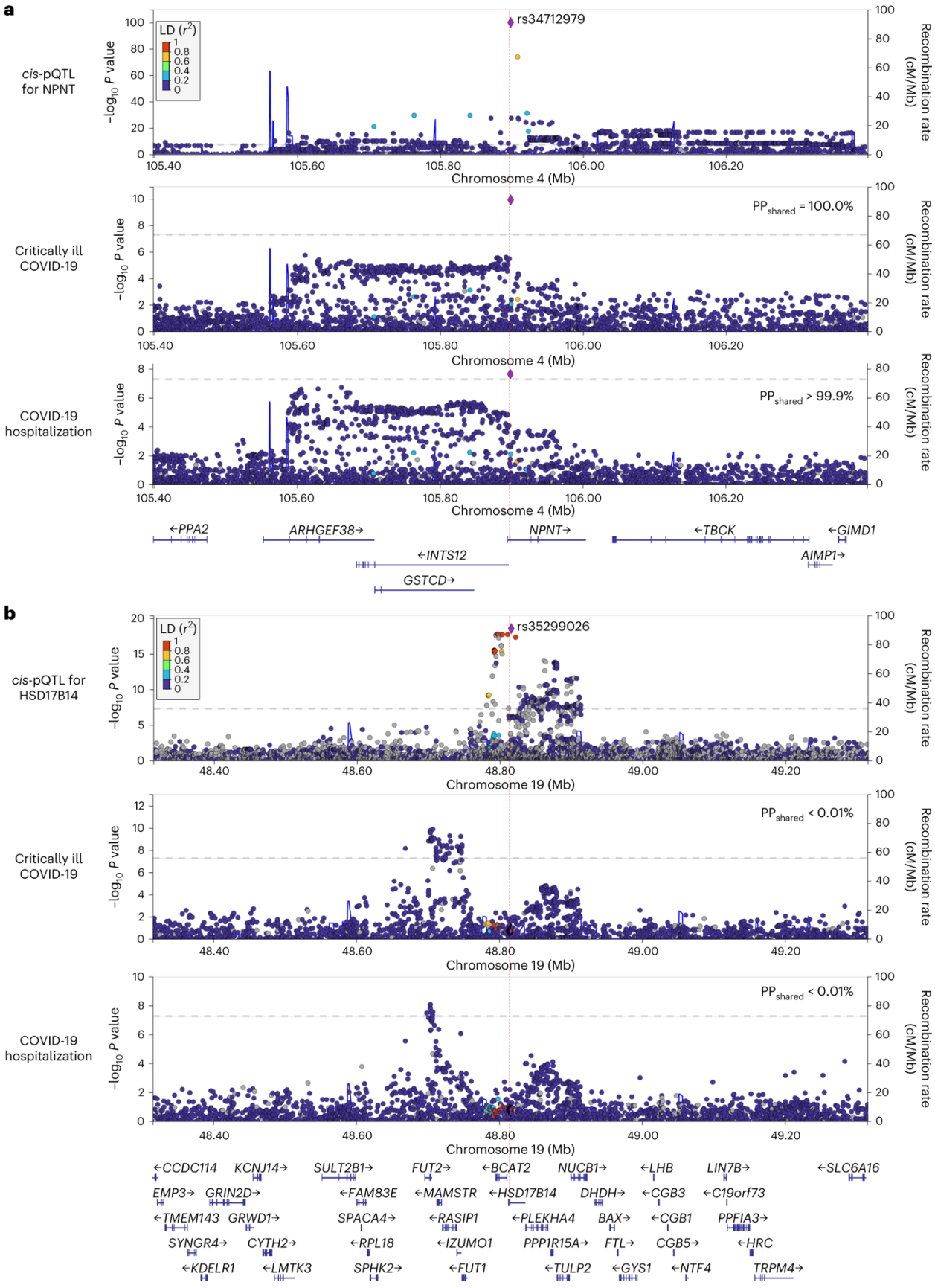

a
cis-pQTL for NPNT
Critically ill COVID-19
COVID-19 hospitalization
rs34712979
LD (r²)
1
0.8
0.6
0.4
0.2
0
−log10 P value
Recombination rate (cM/Mb)
PPshared = 100.0%
PPshared > 99.9%
105.40
105.60
105.80
106.00
106.20
Chromosome 4 (Mb)
←PPA2
ARHGEF38→
NPNT→
←TBCK
←GIMD1
←INTS12
AIMP1→
GSTCD→
b
cis-pQTL for HSD17B14
Critically ill COVID-19
COVID-19 hospitalization
rs35299026
PPshared < 0.01%
48.40
48.60
48.80
49.00
49.20
Chromosome 19 (Mb)
←CCDC114
KCNJ14→
SULT2B1→
FUT2→
←BCAT2
NUCB1→
←LHB
LIN7B→
←SLC6A16
EMP3→
GRIN2D→
←FAM83E
←MAMSTR
←HSD17B14
DHDH→
←CGB3
←C19orf73
←TMEM143
GRWD1→
SPACA4→
←RASIP1
←PLEKHA4
BAX→
←CGB1
PPFIA3→
SYNGR4→
CYTH2→
←RPL18
←IZUMO1
PPP1R15A→
FTL→
CGB5→
←HRC
←KDELR1
←LMTK3
SPHK2→
←FUT1
←TULP2
←GYS1
←NTF4
TRPM4→

Figure 4. Colocalization analyses of cis-pQTL for NPNT or HSD17B14 with COVID outcomes in the 1-Mb region around rs34712979.

We evaluated whether the cis-pQTL for NPNT (a) and HSD17B14 (b) shared the same causal variant with critically ill COVID-19 or COVID-19 hospitalization outcomes using colocalization.

PPshared: Posterior probability that cis-pQTL for NPNT shares a single causal signal with the COVID-19 outcome.

MR analyses using cis-pQTLs from different studies

To further test whether plasma levels of NPNT or HSD17B14 were causal for critically ill COVID-19, we performed additional sets of MR analyses using cis-pQTLs from different cohorts. The cis-pQTLs for NPNT and HSD17B14 were identified using data from the FENLAND study by Pietzner *et al*. (*n* = 10,708) [27] and the AGES Reykjavik study by Emilsson *et al*. (*n* = 3,200) [28]. We assessed the PhenoScanner and Open Target Genetics databases for additional associations for these cis-pQTLs with other phenotypes and found none.

MR analyses using these cis-pQTLs estimated a consistent causal effect of NPNT on critically ill COVID-19 (the FENLAND study: OR = 1.89, 95% CI: 1.56–2.29, $P = 1.21 \times 10^{-10}$; the AGES Reykjavik study: OR = 1.25, 95% CI: 1.06–1.48, $P = 8.26 \times 10^{-3}$) and COVID-19 hospitalization (the FENLAND study, OR =1.45, 95% CI: 1.28–1.66, $P = 2.17\times 10^{-8}$; the AGES Reykjavik study, OR = 1.17, 95% CI: 1.04–1.31, $P = 7.30 \times 10^{-3}$).

In contrast, the estimated causal effects of HSD17B14 on these COVID-19 outcomes were not supported (**Supplementary Table 15**). Given the lack of colocalization and replication, we concluded that the initial finding indicating a causal effect of HSD17B14 on COVID-19 severity outcomes using a cis-pQTL from the deCODE study was likely biased by a difference in LD structure. Therefore, we excluded HSD17B14 from further evaluation.

Comparing with observational associations using BQC19

Considering that the MR analyses (Step 2 MR) used plasma protein levels in a non-infectious state from the deCODE study as the exposures, in supplementary analyses, we observationally assessed whether the plasma levels of NPNT in a non-infectious state were associated with the risk of COVID-19 using logistic regression in 293 individuals from the BQC19 cohort. We restricted the analyses to individuals who met our criteria of critically ill COVID-19, COVID-19 hospitalization, or COVID-19 negative controls (see **Methods)**.

Logistic regression analysis adjusting for age, sex, and sample collection batch showed that increased non-infectious plasma levels of NPNT were associated with the increased risk of critically ill COVID-19 (OR = 1.47, 95% CI: 1.05–2.06, $P = 2.60 \times 10^{-2}$) and COVID-19 hospitalization (OR = 1.49, 95% CI: 1.11–2.02, $P = 9.87 \times 10^{-3}$), which were directionally consistent with the step 2 MR findings (**Supplementary Table 16**).

3.3.5 Follow-up analyses for the putatively causal protein (NPNT)

3.3.5.1 *Colocalization of NPNT's cis-pQTL with eQTL and sQTL*

The observed differences in NPNT levels measured by the SomaScan assay may be explained by levels of a particular isoform, rather than total NPNT levels. Thus, we performed colocalization analyses to evaluate whether the cis-pQTL share the same causal variant with either the total RNA expression level or a specific isoform level of NPNT. Given that the lung is a primary target organ in the context of COVID-19 severity, NPNT is highly expressed in the lung, NPNT splice isoforms have previously been implicated as a risk factor for COPD and other lung outcomes[43, 44,45], and SNPs influencing a specific isoform level would be sQTLs, we performed colocalization analysis of the cis-pQTL with eQTLs and sQTLs for NPNT in lung tissue from GTEx[46].

Colocalization analysis of NPNT pQTL and sQTL within a one-megabase (1-Mb) region (±500 kb) surrounding the lead cis-pQTL (rs34712979) revealed that there was a high probability of pQTL and sQTL sharing a single causal variant (posterior probability for a shared causal signal (PP_{shared}) = 100.0%). However, this was not true for the eQTL (PP_{shared} < 0.01%; **Figure 5**). We confirmed the robustness of the colocalization results using different priors (**Methods and Supplementary Table 17**).

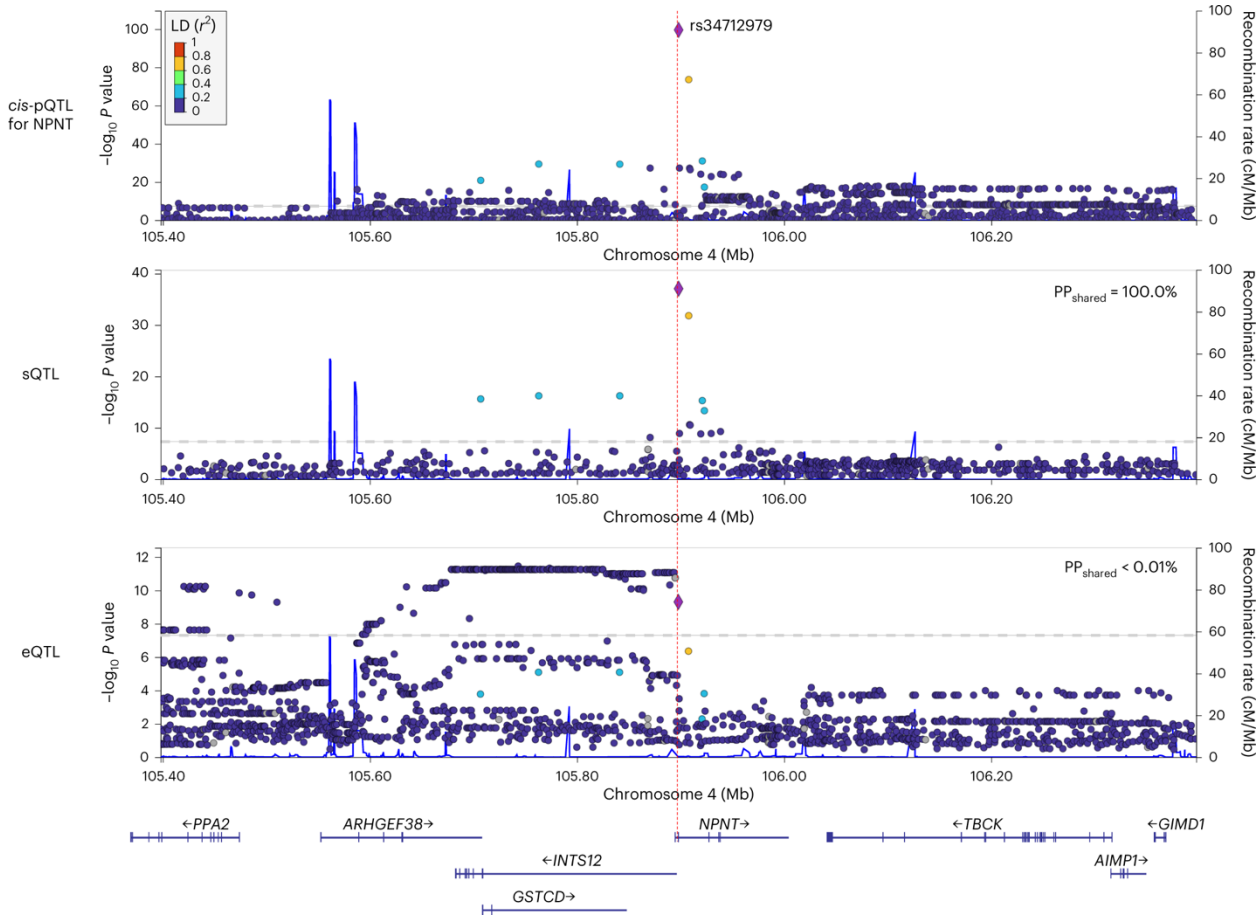


Figure 5. Colocalization analyses of cis-pQTL with sQTL and eQTL for NPNT.

PP_{shared}: Posterior probability that the cis-pQTL for NPNT shared a single causal signal with its sQTLs or eQTLs in the lung.

Given that the thyroid, lung, and arteries are the top three *NPNT*-expressing tissues according to the GTEx[46], we performed the same colocalization analyses in the thyroid and arteries (aorta, tibial, and coronary) and found that the pQTL of NPNT also colocalized with sQTL in these tissues (PP_{shared} = 100.0% for all). Furthermore, it colocalized with the eQTL in the aorta and tibial artery (PP_{shared} = 100.0% for both), but not in the thyroid or coronary artery (PP_{shared} = 11.9% and 5.8%, respectively).

Notably, the lead cis-pQTL from the deCODE study, i.e., rs34712979, was also identified by the FENLAND study[27] and has been reported to create a cryptic splice acceptor site, which inserts a three-nucleotide sequence coding a serine residue at the 5'-splice site of exon 2, resulting in perturbations of the alpha-helix motif[43]. Further, this lead cis-pQTL (rs34712979) and another cis-pQTL (rs78213340) from the AGES Reykjavik study—that was not in high LD with rs34712979 (r^2 = 0.234)—were both associated with the same exon-skipping splicing in the lung in GTEx (**Supplementary Table 18**). Hence, two different cis-pQTLs of NPNT from three different studies—all using SomaScan—were associated with the same splicing pattern, suggesting that the SomaScan assay measures a specific isoform of the NPNT protein.

Collectively, these findings indicate that the SomaScan NPNT-targeting aptamer measures specific isoform levels and the specific isoform of NPNT with a serine insertion at the N-terminus, influences the effect of NPNT on COVID-19 severity.

3.3.5.2 *Single-cell RNA-sequencing data of SARS-CoV-2-infected lungs*

To gain insights into the biological role of NPNT in SARS-CoV-2-infected lungs, we explored the lung cell types that significantly expressed the *NPNT* gene by analyzing single-cell RNA-sequencing data from lung autopsy samples (106,792 cells) of 16 patients who died of COVID-19[47] (Single Cell Portal of the Broad Institute (Accession ID: SCP1052)).

We found that *NPNT* is widely expressed in lung cell types, including epithelial cells (type 1 and type 2 alveolar cells) and fibroblasts, highlighting its role in air exchange and fibrosis. Furthermore, we conducted a subgroup analysis of SARS-CoV-

2-positive and -negative cells and found that *NPNT* was significantly expressed in SARS-CoV-2-positive alveolar cells and fibroblasts (permutation test $P < 0.001$, see **Methods** for details) (**Figure 6**).

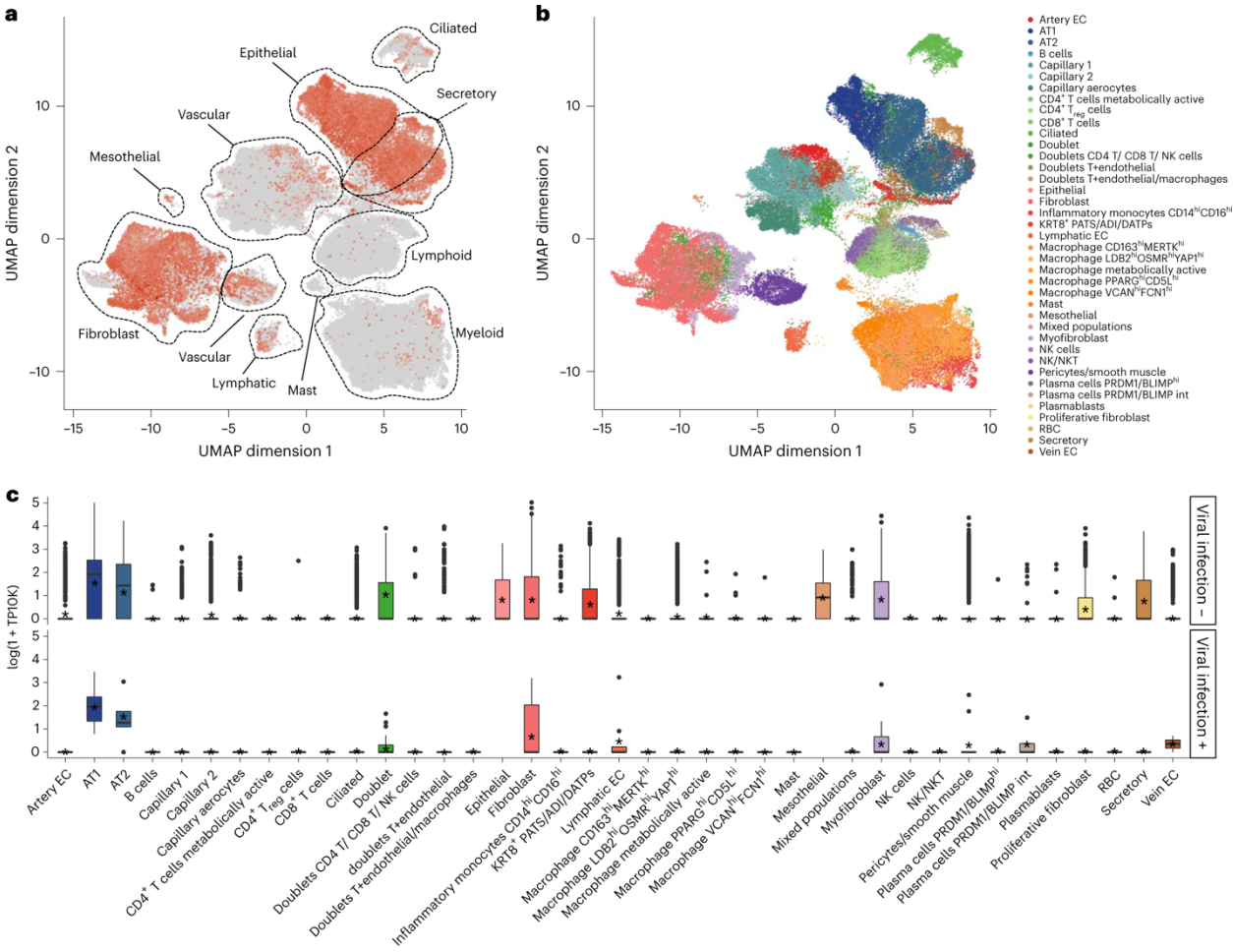


Figure 6. NPNT expression levels in lung cell types from COVID-19 lung autopsy samples at single-cell resolution.

(a) NPNT expression levels of each cell type at single-cell resolution in the 16 lung donors with COVID-19. (b) Twenty-eight annotated cell types of the lung. (c) NPNT expression status in 106,449 SARS-CoV-2 non-infected cells (viral infection −, top panel) or 343 SARS-CoV-2 infected cells (viral infection +, bottom panel) in 16 lung donors. NPNT expression levels of 28 cell types in the two groups are shown in a box plot. In each box, the horizontal line denotes a median value of the expression levels, and the asterisk inside each box denotes the mean value. Each box extends from the 25th to the 75th percentile of each group. Whiskers extend 1.5 times the interquartile range from the top and bottom of the box. Log (TP10K+1) was calculated by normalizing original gene counts by total unique molecular identifiers (UMI) counts, multiplying by 10,000 (TP10K), and then taking the natural logarithm

3.3.5.3 *Mediation analysis*

Since BMI has likely thousands of effects upon human physiology and creates an important perturbation of the proteome, we wanted to estimate the proportion of the effect of BMI that was mediated only through plasma NPNT levels. To do so, we performed a mediation analysis using network MR with the product of coefficients method to understand the extent to which plasma NPNT levels mediate the association between BMI and critically ill COVID-19 and COVID-19 hospitalization.

For critically ill COVID-19, we estimated the effect of BMI on critically ill COVID-19 mediated by plasma NPNT levels (**Figure 7**). We first estimated the effect of BMI on plasma NPNT levels and then multiplied this estimate by the effect of plasma NPNT levels on critically ill COVID-19 (see **Methods** for further details). The ratio of the effect mediated by NPNT was calculated by dividing the NPNT-mediated effect estimate by the total effect estimate of BMI on critically ill COVID-19. We repeated the same process for COVID-19 hospitalization.

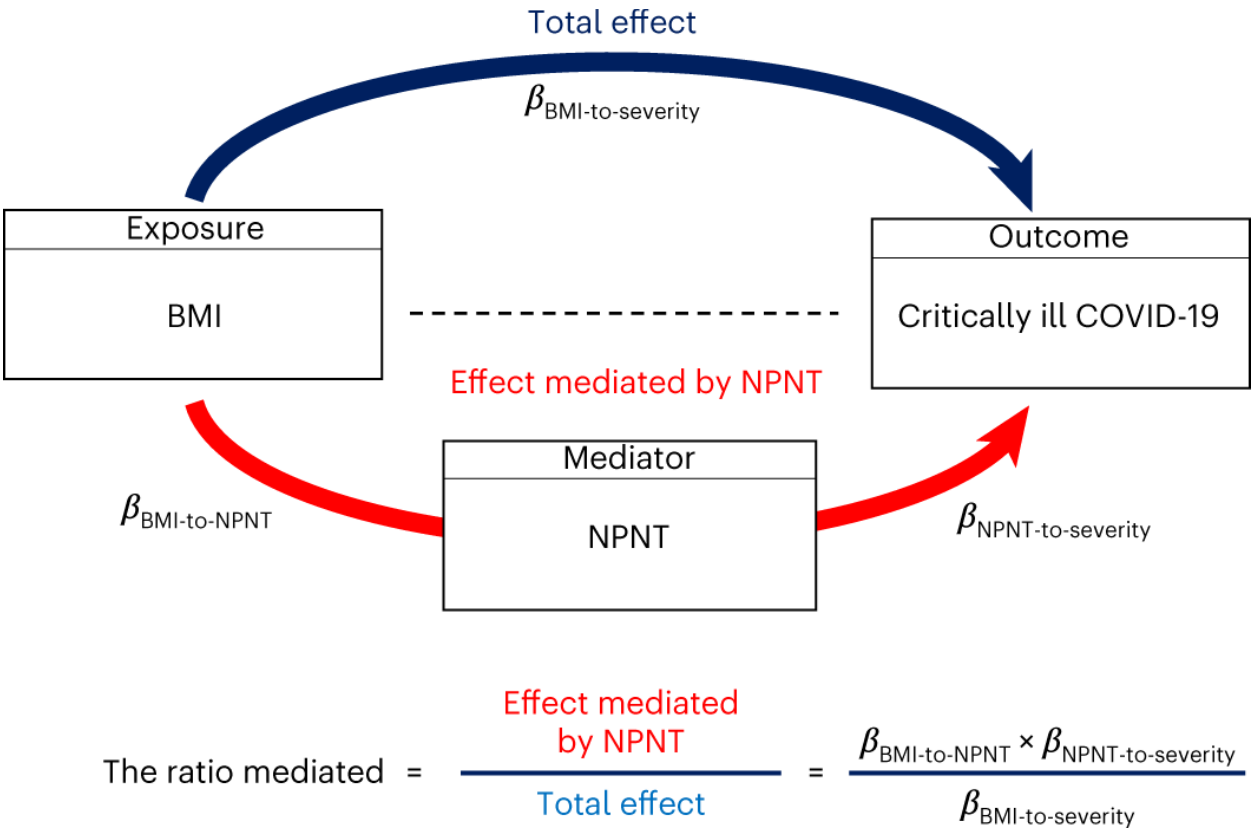


Figure 7. MR mediation analysis illustrated by the directed acyclic graph.

The dark blue arrow represents the total effect of BMI on critically ill COVID-19. The red arrow represents the effect of BMI on critically ill COVID-19 mediated by NPNT. For the effect of BMI on critically ill COVID-19 mediated by NPNT, the product of coefficients method calculates the proportion mediated by multiplying $\beta_{\text{BMI-to-NPNT}}$ and $\beta_{\text{NPNT-to-severity}}$, where $\beta_{\text{BMI-to-NPNT}}$ is the effect of BMI on NPNT and $\beta_{\text{NPNT-to-severity}}$ is the effect of NPNT on critically ill COVID-19. We evaluated the proportion mediated for the effect of obesity-related exposures (i.e., BMI, body fat percentage, and body fat mass) on COVID-19 severity outcomes (i.e., critically ill COVID-19 and COVID-19 hospitalization). BMI: Body mass index; MR: Mendelian randomization; NPNT: nephronectin.

We found that plasma NPNT levels partially mediated the total effect of BMI on critically ill COVID-19 (proportion mediated = 13.9%, 95% CI: 6.1–21.6%, $P = 4.52 \times 10^{-4}$) and COVID-19 hospitalization (proportion mediated = 10.6%, 95% CI: 4.4–16.7%, $P = 8.02 \times 10^{-4}$) (**Supplementary Table 19**).

In supplementary analyses, we evaluated whether NPNT mediates the total effect of body fat percentage on the COVID-19 severity outcomes. We found that plasma NPNT levels mediated the total effect of body fat percentage on critically ill COVID-19 (proportion mediated = 9.5%, 95% CI: 3.6–15.4%, $P = 1.59 \times 10^{-3}$) and COVID-19 hospitalization (proportion mediated = 7.7%, 95% CI: 2.7–12.7%, $P = 2.57 \times 10^{-3}$). We also found consistent results for body fat mass; plasma NPNT levels mediated the total effect of body fat mass on critically ill COVID-19 (proportion mediated = 13.4%, 95% CI: 6.1–20.6%, $P = 2.85 \times 10^{-4}$) and COVID-19 hospitalization (proportion mediated = 9.9%, 95% CI: 4.3–15.6%, $P = 5.61 \times 10^{-4}$) (**Supplementary Table 19**). All of these results consistently suggested that plasma NPNT levels partially mediated the effect of obesity, measured by BMI, body fat percentage, or body fat mass, on COVID-19 severity outcomes.

Throughout the above analyses, we did not adjust the exposure while estimating the effect of the mediator on the outcome ($\beta_{\text{NPNT-to-severity}}$ in **Figure 7**) to avoid weak instrument bias (see **Methods**). This approach was also used in previous studies[48, 49, 50]. In supplementary mediation analyses, we found that adjusting for the exposure when estimating $\beta_{\text{NPNT-to-severity}}$ (i.e., Sobel test) also support the role of NPNT as a mediator for the effect of obesity-related exposures (BMI, body fat percentage, and body fat mass) on COVID-19 severity outcomes; however, the estimated proportion mediated was modest, likely due to weak instrument bias (**Supplementary Table 20**).

3.3.6 Multivariable MR analyses of body fat and fat-free mass

Given the consistent evidence that BMI influenced NPNT levels, which in turn influence COVID-19 severity, we aimed to identify a way of modulating plasma NPNT levels by gaining a better understanding of how the estimated causal effect of BMI on plasma NPNT levels was influenced by fat or fat-free mass in humans. In the Step 1 MR, we

showed that increased BMI is estimated to increase plasma levels of NPNT. However, BMI is a function of height and weight and does not take into account body compositions, such as body fat and fat-free mass. Thus, we specifically assessed their independent effects on plasma NPNT levels. For this, we performed multivariable MR using body fat and fat-free mass as the exposures and plasma NPNT levels as the outcomes (see **Methods**). Conditional F-statistics for body fat mass and fat-free mass were 38.8 and 59.3, respectively, which did not indicate weak instrument bias (which is suspected when F-statistics is less than 10) (**Supplementary Table 21**) [51].

Multivariable MR using inverse variance weighted method found that a one SD increase in body fat mass was associated with increased plasma levels of NPNT (beta = 0.21, 95% CI: 0.14–0.28, $P = 2.74 \times 10^{-9}$), whereas a one SD increase in body fat-free mass was associated with decreased plasma levels of NPNT (beta = -0.13, 95% CI: -0.22, -0.05, $P = 2.98 \times 10^{-3}$) (**Figure 8**). In sensitivity analyses, Q-statistics for instrumental validity did not suggest evidence of pleiotropy (Q-statistics = 940.8, $P_{Q\text{-}statistics} = 0.09$). Multivariable MR-Egger also showed directionally consistent results, and no evidence of directional pleiotropy was observed with the MR-Egger intercept test (**Supplementary Table 21**).

Further, to test the influence of body fat and fat-free mass on COVID-19 severity outcomes, we conducted multivariable MR analyses using body fat mass and fat-free mass as the exposures and either critically ill COVID-19 or COVID-19 hospitalization as the outcome. We found that a one SD increase in body fat mass was associated with increased odds of critically ill COVID-19 (OR = 1.89, 95% CI: 1.65–2.16, $P = 4.83 \times 10^{-20}$) and COVID-19 hospitalization (OR = 1.59, 95% CI: 1.45–1.75, $P = 4.28 \times 10^{-22}$), whereas a one SD increase in body fat-free mass was associated with a decreased risk of critically ill COVID-19 (OR = 0.77, 95% CI: 0.65–0.91, $P = 1.94 \times 10^{-3}$) and COVID-19 hospitalization (OR = 0.87, 95% CI: 0.77–0.97, $P = 1.57 \times 10^{-2}$). Q-statistics for instrumental validity and the MR-Egger intercept test did not show evidence of pleiotropy (**Supplementary Table 21**).

These findings suggest that decreasing body fat mass and increasing fat-free mass (e.g., through actions such as appropriate diet and exercise) can reduce plasma

NPNT levels, and thus reduce the risk of critically ill COVID-19, thereby indicating NPNT as an actionable target.

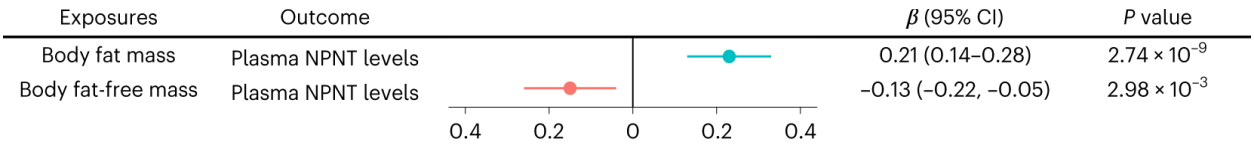


Figure 8. Multivariable MR analysis for evaluating independent effects of body fat and fat-free mass on plasma NPNT levels.

We performed multivariable MR with the inverse variance weighted method using body fat and fat-free mass as the exposures and plasma NPNT levels as the outcome.

3.4 Discussion

In the present study, we conducted MR analyses and found that NPNT likely mediates an important proportion of the effect of obesity on COVID-19 severity. Considering that BMI is a highly polygenic trait with more than 530 associated loci[25] and influences more than 1,200 circulating proteins, as shown in the present and previous studies[6,7], it is remarkable that a single protein explains a reasonably large proportion of the effect of BMI and other obesity-related traits on COVID-19 severity outcomes. Additionally, colocalization analyses provided evidence that the NPNT cis-pQTL is shared with the lung sQTL for NPNT, which leads to a splice isoform with a serine insertion at the N-terminus of NPNT[43]. These results suggest that NPNT mediates a proportion of obesity's effect on critically ill COVID-19 and that this effect may be conferred by alternative splicing of *NPNT* which introduces a serine residue at its N-terminus.

NPNT, or nephronectin, is an extracellular matrix protein that controls integrin binding activity. It is known as a functional ligand of integrin α8/β-1 in kidney development and is also associated with the development, remodeling, and survival of various tissues through the binding of integrins[52,53]. NPNT has also been implicated in inflammation and autoimmunity[54,55], which may align with the suggested role of obesity-induced inflammation in COVID-19 severity[56]. In addition, *NPNT* is expressed in lung alveolar cells and fibroblasts[57,58], and recent GWASs have identified the same NPNT splice variant, rs34712979, to be associated with lung function-related traits (e.g., forced expiratory volume (FEV1), forced vital capacity (FVC), and FEV1/FVC ratio) and COPD[44,45]. Our single-cell RNA-sequencing analysis on COVID-19 lung autopsy samples also found that *NPNT* was expressed in SARS-CoV-2-infected alveolar cells and fibroblasts, suggesting its role in air exchange and fibrosis in SARS-CoV-2-affected lungs. These findings collectively suggest that NPNT alternative splicing, which results in an isoform with a serine insertion in the N-terminus of the protein, increases the risk of deterioration of lung function and lung diseases such as COPD and COVID-19. Future studies are required to investigate the functional properties of this alternative splice isoform.

Our findings may have important clinical implications. The multivariable MR approach showed that decreasing body fat mass and increasing body fat-free mass, which can be achieved through non-pharmacological interventions such as exercise and appropriate diet[59], can reduce plasma NPNT levels and the risk of COVID-19 severity outcomes. Moreover, recent trials have shown GLP-1/GIP co-agonist tirzepatide[60,61] and GLP-1 receptor agonists including semaglutide[62,63] and liraglutide[64,65] can reduce body fat mass, while preserving body fat-free mass. We believe that these findings are important because they offer the possibility to potentially modulate plasma NPNT levels, using available therapies. Such hypotheses require further investigations in clinical trials.

This study has both strengths and weaknesses. MR and colocalization analyses robustly implicated NPNT as a causal mediator of the relationship between BMI and COVID-19 severity. The robustness of the MR findings was enhanced by the large sample sizes used to derive the findings. To the best of our knowledge, this is the first study to identify a mediator of obesity on COVID-19 employing a two-step MR approach, although a previous study using a similar framework with limited statistical power earlier in the pandemic could not identify a strong protein signal[66]. Furthermore, our MR findings withstood multiple sensitivity analyses and were supported by colocalization, fine-mapping, replication MR, observational evaluation, and RNA-sequencing studies in COVID-19 lung samples. Intriguingly, multivariable MR showed that plasma NPNT levels were increased by body fat mass but decreased by fat-free mass, indicating that the effect of BMI on plasma NPNT levels was driven by body fat mass and partially counteracted by fat-free mass. These findings suggest that NPNT could be a potential intervention target in individuals with obesity to prevent critically ill COVID-19 and highlight one of the possible mechanisms by which appropriate diet and exercise can confer risk reduction of COVID-19 severity through the reduction in plasma NPNT levels.

This study also has important limitations. First, MR and colocalization analyses were restricted to individuals of European ancestry to avoid confounding by population

stratification and heterogeneity of genetic associations across different ancestries. Whether NPNT mediates the effect of BMI on COVID-19 severity in populations of non-European ancestries requires further investigation. Second, we did not perform sex-stratified analysis due to the unavailability of sex-specific datasets. Third, there were no genome-wide significant genetic instrumental variables for BMI in the 1-Mb region around the cis-region of NPNT (1-Mb around the transcription start site), we could not pinpoint a single genetic variant that directly links BMI to NPNT. Hence, the effect of BMI on NPNT is likely mediated by a combination of multiple trans-effects, which was also indicated by the scatter plot in the MR analysis for NPNT. However, this is not surprising considering that BMI is a highly polygenic, complex trait, and multiple pathways could confer the effect of BMI on other diseases[67]. Fourth, we do not know the molecular mechanism by which BMI influences the splice isoform. Previous studies have suggested that obesity is associated with alternative splicing[68-70] and extracellular matrix protein remodeling[71], which may align with our findings. However, further investigation is required to clarify the specific molecular mechanisms involved. Lastly, we do not rule out the possibility that total levels of NPNT (all isoforms) measured by the NPNT-targeting aptamer mediate the effect of obesity on COVID-19 severity. However, given the evidence provided by the MR, colocalization, fine-mapping, and a biological understanding of lead cis-pQTL (i.e., the variant causes alternative splicing resulting in perturbations of the alpha-helix motif in the lung), it is likely that the specific isoform measured by the aptamer is driving the effect. Nevertheless, isoform-specific measurements (e.g., mass spectrometry) will be required to confirm these findings.

In conclusion, we integrated a two-step MR approach, sensitivity analyses, colocalization, fine-mapping, single-cell RNA-sequencing, and mediation analyses to identify NPNT as an important mediator of the effect of obesity on COVID-19 severity outcomes. We also showed that decreasing body fat mass and increasing fat-free mass (e.g., by actions such as exercise and appropriate diet) can lower NPNT levels, and thus may improve COVID-19 severity outcomes. These findings provide actionable insights into how obesity influences COVID-19 severity.

3.5 Methods

3.5.1 Step 1: BMI to plasma proteins

Two-sample MR

BMI GWAS: We used the BMI GWAS meta-analysis with the largest sample size, comprising 693,529 European ancestry individuals from the GIANT consortium and UK Biobank[25]. The consortium details are provided in **Supplementary Table 1,** and post-hoc power calculation is provided in **Supplementary Table 22**.

Proteomic GWAS: For GWAS of plasma protein levels, we used the largest proteomic GWAS available[26], which measured 4,907 proteins in 35,559 individuals of European ancestry using the SomaLogic SomaScan assay v4 (SOMAscan, SomaLogic, Boulder, Colorado, USA).

Two-sample MR: The effect of BMI on plasma protein levels was assessed using the inverse variance weighted method with a random-effects model in TwoSampleMR v.0.5.6[72] (https://mrcieu.github.io/TwoSampleMR/). The instrumental variables for the exposure were defined as genome-wide significant and independent single nucleotide polymorphisms (SNPs) ($P < 5 \times 10^{-8}$; $r^2 < 0.001$, with a clumping window of 10 Mb). SNPs in the human major histocompatibility complex (MHC) region at chromosome 6: 28,477,797–33,448,354 (GRCh37) were excluded considering its complex LD structure. We used PLINK v1.9[73] (http://pngu.mgh.harvard.edu/purcell/plink/) to obtain instrumental variables by clumping SNPs using the 1000 Genomes Project European reference panel[74] and applying an LD threshold of $r^2 < 0.001$. The genome-wide significant independent SNPs with the lowest *P*-value were selected from each LD block. If instrumental variable SNPs were not present in an outcome GWAS, we used proxy SNPs ($r^2 > 0.8$ with the original SNP). Proxy SNPs were identified using snappy v1.0 (https://gitlab.com/richards-lab/vince.forgetta/snappy) with 1000 Genomes Project's European reference panel[74]. Data harmonization and MR analyses were conducted using TwoSampleMR v0.5.6. Specifically, data harmonization was performed with the "harmonise_data()" function using default settings, including the removal of palindromic

SNPs that have minor allele frequency above 0.42. MR was performed using the “mr() “ function. A Bonferroni correction was used to set a statistical significance threshold by dividing 0.05 by the number of proteins (P = 1.0 × 10^{-5} (0.05/4907)). We note that this correction is overly-conservative since many proteins are non-independent. We did so to safeguard against false positive findings. In a heterogeneity test, we calculated I^2 statistics using “Isq()” function and heterogeneity P-value with “mr_heterogeneity()” function; results with an I^2 > 50% and heterogeneity P-value (Q_pval) < 0.05 were considered to be heterogeneous (substantial heterogeneity) [75]. For evaluating directional pleiotropy, we used the MR-Egger intercept test, which was performed using the “mr_pleiotropy_test()” function; where directional pleiotropy was considered to be present when the MR-Egger intercept differed from the null (P < 0.05). We note that even in the presence of moderate heterogeneity, balanced horizontal pleiotropic effects would not violate the MR assumption of a lack of directional pleiotropy[76,77]. For NPNT and HSD17B14, we used MR-Egger, weighted median, and weighted mode methods as additional sensitivity analyses to evaluate the directional consistency of beta coefficients with the inverse variance weighted method.

For reverse MR, whereby the effects of plasma protein levels on BMI were assessed, we used cis-pQTLs from the deCODE study as the exposures and BMI GWAS as the outcome, deploying the inverse variance weighted method or Wald ratio method when only one SNP was available as an instrumental variable. After the cis-pQTL search, proxy, and data harmonization, 357 proteins were tested in MR for their estimated reverse effects. Results with P < 1.4 × 10^{-4} (0.05/357; Bonferroni correction) were considered statistically significant.

We used BMI GWAS from the UK Biobank (not the meta-analysis GWAS) because multiple variants, including cis-pQTL for NPNT from the deCODE study (rs34712979), were dropped during the stringent quality control process of the meta-analysis (e.g., rs34712979 was not present in the GIANT GWAS and thus dropped during the meta-analysis process).

To assess statistical power, F-statistics were calculated as previously described[78] using the following formula: $F = \frac{R^2(n-2-k)}{(1-R^2)k}$ where: R^2 = proportion of variance in the

exposure trait and *k* = number of instrumental variables (**Supplementary Table 2).** We also performed post-hoc power calculations using an online power calculator for Mendelian randomization (https://sb452.shinyapps.io/power/) (**Supplementary Table 22**).

3.5.2 Step 2: BMI-driven proteins to COVID-19 severity outcomes

Two-sample MR

Cis-pQTL GWAS: We identified cis-pQTLs from the GWAS of 4,907 proteins in 35,559 individuals of European ancestry from the deCODE study[26]. Cis-pQTLs were defined as pQTLs located within 1-Mb around a transcription start site of a protein-coding gene. Further details of the dataset are provided in **Supplementary Table 1**. Genetic coordinates of transcription start sites of each gene used to define cis-pQTLs are provided in Supplementary Table 2 of the deCODE study[26].

COVID-19 severity outcome GWAS: For COVID-19 outcomes, we used a GWAS meta-analysis from the COVID-19 Host Genetics Initiatives data release 7 (https://www.covid19hg.org/). The outcomes included critically ill COVID-19 (13,769 cases and 1,072,442 controls) and COVID-19 hospitalization (32,519 cases and 2,062,805 controls). These two outcomes were collectively referred to as COVID-19 severity outcomes. Critically ill COVID-19 was defined as a requirement for respiratory suport among hospitalized individuals with laboratory-confirmed SARS-CoV-2 infection or death due to COVID-19. COVID-19 hospitalization was defined as a laboratory-confirmed SARS-CoV-2 infection that required hospitalization.

Two-sample MR: Using the data above, we carried out two-sample MR for the effects of plasma protein levels on COVID-19 severity outcomes. We used the inverse variance weighted method for proteins with ≥2 instrumental variables and the Wald ratio method for proteins with a single instrumental variable. When an instrumental variable SNP was not found in an outcome GWAS, we searched and used a proxy SNP ($r^2 < 0.8$ with the original SNP) using snappy v1.0 (https://gitlab.com/richards-lab/vince.forgetta/snappy) with 1000 Genomes Project European reference panel[74]. Results with a $P < 1.4 \times 10^{-4}$

(0.05/358; Bonferroni correction with the number of proteins tested in the Step 2 MR) were considered statistically significant. We used the MR-Egger intercept test to assess lack of directional pleiotropy for proteins with three or more genetic instrumental variables. However, we could not use this test for NPNT because it had fewer than three genetic instrumental variables. Therefore, we used the PhenoScanner (http://www.phenoscanner.medschl.cam.ac.uk/) and Open Targets Genetics (https://genetics.opentargets.org/) databases to test whether variants had potential pleiotropic associations with other diseases or traits. Associations with $P < 5 \times 10^{-8}$ were considered statistically significant. We also assessed reverse causation, wherein the effect of COVID-19 severity may influence plasma protein levels with the MR-Steiger test using "directionality_test()" function from TwoSampleMR v.0.5.6[72].

3.5.3 Validation analyses for proteins prioritized by step 1 and step 2 MR (NPNT and HSD17B14)

3.5.3.1 Step 1 MR validation

MR analysis using body fat percentage

We repeated the step 1 MR (described above) for NPNT and HSD17B14 using body fat percentage as the exposure instead of BMI. For this, we used body fat percentage GWAS in 454,633 individuals of European ancestry from the UK Biobank, obtained from the IEU OpenGWAS project (https://gwas.mrcieu.ac.uk/). The accession ID was ukb-b-8909.

Comparison of the MR findings with the published observational association study from INTERVAL

The INTERVAL study was a prospective cohort study conducted in England. The study recruited ~50,000 participants of primarily European ancestry without a self-reported history of any major disease. The recruitment took place between 2012 and 2014,

before the COVID-19 pandemic. The study measured 3,622 plasma protein levels in 2,737 individuals using the SomaScan assay. The study performed a linear regression to evaluate associations between BMI and plasma protein levels, adjusting for age and sex[6]. The derived beta estimates represented normalized SD-unit difference in each protein level per one SD (4.8 kg/m^2) increase in BMI. Further details of the study have been described in depth previously[6].

3.5.3.2 Step 2 MR validation

Colocalization of cis-pQTLs with COVID-19 severity outcomes

To identify potential bias caused by confounding owing to LD, we performed colocalization to evaluate whether cis-pQTLs shared a single causal variant between the three COVID-19 outcomes for the causal proteins and NPNT or HSD17B14. We used the coloc R package v5.1.0[79] (https://chr1swallace.github.io/coloc/) to evaluate all SNPs within the 1-Mb region around the top cis-pQTL for the NPNT (defined as the cis-pQTL with the smallest *P*-value). The posterior probability of hypothesis 4 (H_4) or PP_{shared} (two traits sharing a single causal variant) > 0.8 was considered to indicate strong evidence of colocalization. Colocalization analyses were conducted using the coloc R package v5.1.0[79], using default priors of $p1 = p2 = 10^{-4}$ and $p12 = 10^{-5}$, where $p1$ is a prior probability that only trait 1 has a genetic association in the region, $p2$ is a prior probability that only trait 2 has a genetic association in the region, and $p12$ is a prior probability that both trait 1 and trait 2 share the same genetic association in the region. To test the robustness of the results, we evaluated different combinations of priors: $p1$ = c(10^{-4}, 10^{-5}, 10^{-6}), $p2$ = c(10^{-4}, 10^{-5}, 10^{-6}), $p12$ = c(10^{-5}, 5×10^{-6}, 10^{-6}), as performed previously[18].

Fine-mapping of the NPNT region in the COVID-19 severity GWAS

The COVID-19 Host Genetics Initiatives release 7 data for those of European ancestry with critically ill COVID-19 and hospitalization was inputted into FINEMAP v1.4[80] (http://www.christianbenner.com/) using default options. We assumed a maximum of one causal SNP, given a single GWAS significant SNP in the NPNT region. Summary statistics were filtered for SNPs present in the European ancestry, as done by Huffman *et al*[81]. We set prior probabilities using scaled CADD-PHRED scores[42] (https://cadd.gs.washington.edu/). The scaled CADD-PHRED scores were normalized so that their sum equaled one. The variance of effect size was set such that the maximum odds ratio a variant can have with 95% probability is two, then scaled using the case-control ratio as described previously[81].

MR analyses using cis-pQTLs for NPNT and HSD17B14 from different cohorts

We also obtained cis-pQTLs and their beta estimates for plasma protein levels of NPNT and HSD17B14 from the FENLAND study and AGES Reykjavik study (**Supplementary Table 10**). We repeated the two-sample MR using these cis-pQTLs as instrumental variables and the COVID-19 severity outcomes as described above.

Comparing with observational associations using BQC19

The BQC-19 cohort: BQC19 (Biobanque Québécoise de la COVID-19) is a province-wide biobank that provides global access to important biological and clinical data from patients with COVID-19 and control subjects in Québec, Canada (https://www.bqc19.ca/). Blood samples were collected in acid citrate dextrose (ACD) tubes from 264 SARS-CoV-2 infectious and 463 non-infectious patients of European ancestry (see **Supplementary Information** for the definition of infectious or non-infectious state). Detailed description of sample processing can be found in the **Supplementary Information.** Briefly, the samples underwent proteomic profiling on the SomaScan v4 assay, and 4,907 aptamers were used for analysis, consistent with the

deCODE study[26]. For quality control, protein levels were natural log-transformed and then batch-corrected using the ComBat function implemented in the sva R package v3.44.0[82].

Logistic regression analysis in BQC19

Given that the MR analyses used plasma protein levels in a non-infectious state as the exposures, we observationally assessed an association between plasma NPNT in a non-infectious state and the risk of COVID-19 severity outcomes in the BQC19 cohort. We defined the COVID-19 severity outcomes in accordance with those for the GWAS from COVID-19 Host Genetics Initiative (**Supplementary Information**).

We performed logistic regression analysis using one of the COVID-19 outcomes as the dependent variable and standardized plasma NPNT levels as the independent variable while adjusting for sex, age, and batch number. We did not adjust for clinical risk factors such as smoking and socioeconomic status.

3.5.4 Follow-up analyses for the putatively causal protein (NPNT)

3.5.4.1 Colocalization of NPNT's cis-pQTL with eQTL, and sQTL

sQTL and eQTL GWAS: We used sQTL and eQTL GWASs derived from the lung and those of thyroid and arteries (i.e., the top three *NPNT*-expressing tissues) in the European-ancestry individuals from the GTEx Portal V8 dataset[46] (https://gtexportal.org/).

Colocalization analyses: To evaluate whether cis-pQTL is more affected by sQTL, rather than total expression, we also carried out colocalization of the cis-pQTL with eQTLs and sQTLs of NPNT using the GTEx V8 dataset. Colocalization analyses were performed in the 1 Mb region around the cis-pQTL for NPNT (rs34712979) using "coloc.abf()" function from the coloc v5.1.0[79] with default priors of $p1 = p2 = 10^{-4}$ and $p12 = 10^{-5}$ (the definitions of $p1$, $p2$, and $p12$ can be found above). $PP_{shared} > 0.8$ was considered to indicate strong evidence of colocalization. To test the robustness of the

results, we evaluated 27 different combinations of priors: $p1$ = c(10^{-4}, 10^{-5}, 10^{-6}), $p2$ = c(10^{-4}, 10^{-5}, 10^{-6}), $p12$ = c(10^{-5}, 5×10^{-6}, 10^{-6}), as performed previously[18]. We also evaluated whether the lead cis-pQTL from the deCODE study (rs34712979), the FENLAND study (rs34712979), and the AGES Reykjavik study (rs78213340) were associated with the same splice pattern in the lung using GTEx. For eQTL and sQTL, we considered associations with $P < 1 \times 10^{-7}$ statistically significant, as did the deCODE study[26].

3.5.4.2 Single-cell RNA-sequencing data of SARS-CoV-2-infected lungs

To understand the *NPNT* expression pattern in the lungs of SARS-CoV-2 infected individuals, we obtained single-cell transcriptomic data of SARS-CoV-2-infected lungs by Delorey *et al*[47] from the Single Cell Portal of the Broad Institute (https://singlecell.broadinstitute.org/single_cell/) (Accession ID: SCP1052). The data contained 106,792 single cells from the lungs of 16 autopsy donors aged 30 to older than 89 years who died due to COVID-19.

We reanalyzed the data focusing on NPNT expression status. We analyzed the gene expression matrix and the associated metadata using R v4.1.2 and Seurat R package v4.1.1[83] (https://satijalab.org/seurat/). For visualization, NPNT expression levels were represented on a log-transcript per 10 thousand + 1, i.e., log (TP10K+1) scale. To cluster the cells, we adopted the clustering annotation from the original study[47]. To test whether *NPNT* expression was enriched in a cell type, we calculated the proportion of *NPNT*-expression cells in this cell type. Subsequently, we permuted the cell type labels 1,000 times and obtained the frequency (permutation p-value) of the same cell type containing the same or a larger proportion of *NPNT*-expression cells. Additionally, we compared the *NPNT* expression level between a target cell type and the other cell types using Wilcoxon rank-sum test. This enrichment analysis was performed separately in SARS-CoV-2-infected and uninfected cells.

3.5.4.3 Mediation analysis

We undertook mediation analysis to calculate the proportion of the effect of BMI on critically ill COVID-19 mediated by NPNT using network MR (**Figure 7**). We used the product of coefficient method to estimate the NPNT-mediated effect (i.e., the effect of BMI on critically ill COVID-19 that was accounted for by NPNT) and the same instrumental variables and outcome GWASs from step 1 and step 2 MR.
First, we estimated the effect of BMI on NPNT, then multiplied this by the effect of NPNT on critically ill COVID-19. Subsequently, the proportion of the total effect of BMI on COVID-19 mediated by NPNT was estimated by dividing the NPNT-mediated effect ($\beta_{\text{NPNT-to-severity}}$) by the total effect ($\beta_{\text{BMI-to-severity}}$), as described previously[49,84]. Additionally, we evaluated whether NPNT mediated the effect of body fat percentage on COVID-19 hospitalization and the effect of body fat mass on critically ill COVID-19 and COVID-19 hospitalization.

We used the product of coefficients method without adjusting for the exposure (BMI or body fat percentage) when estimating the effect of the mediator on the outcome ($\beta_{\text{NPNT-to-severity}}$) to avoid weak instrument bias. This approach was also used in previous studies[48, 49, 50]. We did so because the above-mentioned exposure adjustment requires multivariable MR using NPNT and BMI (or body fat percentage; fat mass) as exposures; however, there are only two instrumental variables for NPNT (cis-pQTL), but hundreds of instrumental variables for BMI. Thus, when using multivariable MR, which includes instrumental variables from both exposures in the model, the association between plasma NPNT levels and instrumental variables would be substantially weakened (i.e., the large number of instrumental variables of BMI would decrease the strength of the association between plasma NPNT and the genetic variants). Nevertheless, in sensitivity analyses, we performed mediation analyses with adjustment for the exposure when estimating $\beta_{\text{NPNT-to-severity}}$ (i.e., Sobel test) (**Supplementary Table 20**). For multivariable MR, we performed data harmonization using the “mv_harmonise_data()” function from TwoSampleMR v.0.5.6[72], followed by multivariable MR causal estimation using the “mv_multiple()” function from MVMR v0.3 (https://github.com/WSpiller/MVMR)[51]. We calculated conditional F-statistics and heterogeneity Q-statistics using the “strength_mvmr()” and “pleiotropy_mvmr()”

functions, respectively, again from MVMR v0.3. To the best of our knowledge, the sample dataset of the exposure did not overlap (one was the plasma NPNT levels from the deCODE study, which is an Icelandic cohort, and another was from the obesity-related traits from UK Biobank); thus, we set "gencov" to be zero when calculating these measures following the instruction by the package[85], as performed previously[86]. We also quantified directional pleiotropy with the MR-Egger intercept test with the "mr_mvegger()" function from MendelianRandomization v0.6.0[85] (https://github.com/cran/MendelianRandomization).

3.5.5 Multivariable MR of body fat and fat-free mass

Body fat and fat-free mass GWAS: We obtained GWAS of body fat and fat-free mass from UK Biobank using the IEU OpenGWAS project (https://gwas.mrcieu.ac.uk/). Accession ID for each GWAS was "ukb-b-19393" and "ukb-b-13354", respectively.

NPNT GWAS: For GWAS of plasma NPNT levels, we used the same GWAS from the deCODE study[26] as the one used in Step 1 MR, which measured plasma NPNT levels in 35,559 individuals of European ancestry by using the SomaLogic SomaScan assay v4.

COVID-19 severity outcomes: We used the same GWASs of critically ill COVID-19 and COVID-19 hospitalization from the COVID-19 Host Genetics Initiatives data release 7 (https://www.covid19hg.org/).

Multivariable MR: Since body fat mass and fat-free mass are genetically correlated with each other (r = 0.64) [87], we performed multivariable MR to estimate the independent effect of body fat and fat-free mass on plasma NPNT levels or COVID-19 severity outcomes. For instrumental variables, we identified genome-wide significant and independent SNPs for body fat and fat-free mass using the same criteria as in step 1 MR (i.e., $P < 5 \times 10^{-8}$ for significance and $r^2 < 0.001$ with a clumping window of 10 Mb for independence).

SNPs in the MHC region were excluded. After data harmonization, we undertook multivariable MR with the inverse variance weighted method using a random-effects model. We used body fat and fat-free mass as the exposures and plasma NPNT levels as the outcome. Results with $P < 0.025$ (0.05/2; Bonferroni correction) were considered statistically significant. LD clumping was performed using PLINK v1.9[73]. We performed data harmonization using the "mv_harmonise_data()" function from TwoSampleMR v.0.5.6[72], followed by multivariable MR causal estimation using the "mv_multiple()" function from MVMR v0.3 (https://github.com/WSpiller/MVMR)[51]. For sensitivity analyses, we first calculated genetic covariance matrix for exposures (i.e., body fat mass and fat-free mass) using the "phenocov_mvmr()" function, and then used "strength_mvmr()" and "pleiotropy_mvmr()" functions from MVMR v0.3[51] to calculate conditional F-statistics and Q-statistics, respectively. To calculate a phenotypic correlation matrix used in the "phenocov_mvmr()" function, we used metaCCA v1.22.0 (https://github.com/acichonska/metaCCA)[88], as previously performed by Vabistsevits *et al*[86]. Lastly, to perform multivariable MR-Egger analysis, we used the "mr_mvegger()" function from MendelianRandomization v0.6.0[85] (https://github.com/cran/MendelianRandomization).

3.6 Ethical approval

For summary-level data, all contributing cohorts obtained ethical approval from their intuitional ethics review boards. The contributing cohorts include: UK Biobank, GIANT consortium, deCODE study, FENLAND study, AGES Reykjavik study, INTERVAL study, COVID-19 Host Genetics Initiative, and BQC19. For individual-level data in BQC19, BQC19 received ethical approval from the Jewish General Hospital research ethics board (2020-2137) and the Centre Hospitalier de l'Université de Montréal institutional ethics board (MP-02-2020-8929, 19.389). All participants provided informed consent.

3.7 Data availability

· GWAS summary statistics for each trait are available as follows:

BMI (https://portals.broadinstitute.org/collaboration/giant/),

Plasma proteome from the deCODE study (https://www.decode.com/summarydata/),

COVID-19 outcomes (https://www.covid19hg.org/results/r7/),

GTEx Portal V8 (https://gtexportal.org/home/datasets/).

Plasma proteome from BQC19 (https://www.mcgill.ca/genepi/mcg-covid-19-biobank) Access to the data of BQC19 can be obtained upon approval of requests via bqc19.ca.

· Cis-pQTLs of each study are available in the corresponding publications' supplementary materials[26-28]

· Body fat percentage, body fat mass, and fat-free mass GWASs are available at IEU OpenGWAS project with Accession ID of ukb-b-8909, ukb-b-19393, and ukb-b-13354, respectively (https://gwas.mrcieu.ac.uk/)

· Single-cell RNA-sequencing data of COVID-19 lung autopsy samples are available at the Single Cell Portal under the Accession ID of SCP1052 (https://singlecell.broadinstitute.org/single_cell/)

· CADD-scores v.1.6 can be accessed at https://cadd.gs.washington.edu/score

· Genotype data from 1000G genomes project is available at https://www.internationalgenome.org/data

3.8 Code availability

We used R v4.1.2 (https://www.r-project.org/), TwoSampleMR v.0.5.6 (https://mrcieu.github.io/TwoSampleMR/), snappy v1.0 (https://gitlab.com/richards-lab/vince.forgetta/snappy), coloc v5.1.0 (https://chr1swallace.github.io/coloc/), FINEMAP R package v1.4, Seurat v4.0.6 (https://satijalab.org/seurat/), PLINK v1.9 (http://pngu.mgh.harvard.edu/purcell/plink/), and GCTA fastGWA v1.93.3

(https://yanglab.westlake.edu.cn/software/gcta/). Custom codes are available on GitHub (https://github.com/satoshi-yoshiji/TwostepMR_obesity_COVID/).

3.9 Acknowledgments

We thank the COVID-19 Host Genetics Initiative for providing the latest summary statistics for COVID-19 outcomes. We acknowledge Shidong Wang, Tala Khosroheidari, Lena Cuddeback, Will Schwarzmann, and DeAunne Denmark at SomaLogic, Inc. for constructive discussions. We acknowledge Biorender (biorender.com) for providing materials used to create the illustrative diagram. The Richards research group is supported by the Canadian Institutes of Health Research (CIHR: 365825, 409511, 100558, 169303), the McGill Interdisciplinary Initiative in Infection and Immunity (MI4), the Lady Davis Institute of the Jewish General Hospital, the Jewish General Hospital Foundation, the Canadian Foundation for Innovation, the NIH Foundation, Cancer Research UK, Genome Québec, the Public Health Agency of Canada, McGill University, Cancer Research UK [grant number C18281/A29019] and the Fonds de Recherche Québec Santé (FRQS). J.B.R. is supported by an FRQS Mérite Clinical Research Scholarship. The support from Calcul Québec and Compute Canada is acknowledged. TwinsUK is funded by the Welcome Trust, Medical Research Council, European Union, the National Institute for Health Research (NIHR)-funded BioResource, Clinical Research Facility and Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust in partnership with King's College London. S.Y. is supported by the Japan Society for the Promotion of Science. T.L. is supported by a Vanier Canada Graduate Scholarship, an FRQS doctoral training fellowship, and a McGill University Faculty of Medicine Studentship. G.B.L. is supported by scholarships from the FRQS, the CIHR, and Québec's ministry of health and social services. Y.C. is supported by an FRQS doctoral training fellowship and the Lady Davis Institute/TD Bank Studentship Award. M.H. is supported by grants from the SciLifeLab/Knut and Alice Wallenberg national COVID-19 research program (M.H.: KAW 2020.0182, KAW 2020.0241), the Swedish Heart-Lung Foundation (M.H.: 20210089, 20190639, 20190637), and the Swedish Society of Medicine (M.H.:SLS-

938101). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

3.10 Author contributions

Conception and design: S.Y. and J.B.R. Methodology: S.Y., T.L., and J.B.R. Data analysis: S.Y., T.L., J.D.S.W., C.Y.S., and J.B.R. Visualization: S.Y. and T.L. Writing – original draft: S.Y. Writing – Review and editing: S.Y., G.B.L., T.L., J.D.S.W., C.Y.S., T.N., D.M., Y.C., K.L., M.H., Y.I., Z.A., S.L., N.D., C.D., M.V., C.T., X.X., M.B., F.S., L.L., H.M.M., M.A., J.A., V.M., N.J.T., H.Z., S.Z., V.F., Y.F., and J.B.R.

3.11 Competing Interests

J.B.R. has served as an advisor to GlaxoSmithKline and Deerfield Capital. J.B.R.'s institution has received investigator-initiated grant funding from Eli Lilly, GlaxoSmithKline, and Biogen for projects unrelated to this research. He is the CEO of 5 Prime Sciences (www.5primesciences.com), which provides research services for biotech, pharma, and venture capital companies for projects unrelated to this research. T.L. and V.F. are employees of 5 Prime Sciences. T.N. has received speaking fees from Boehringer Ingelheim and AstraZeneca regarding the projects unrelated to this research. The remaining authors declare no competing interests.

3.12 References

1 Johns Hopkins University. COVID-19 Global Map, Available from: https://coronavirus.jhu.edu/map.html. Last accessed on Jan. 4, 2022.

2 COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of COVID-19. *Nature*, 472–477. https://doi.org:10.1038/s41586-021-03767-x (2021).

3 Stefan, N., Birkenfeld, A. L., Schulze, M. B. & Ludwig, D. S. Obesity and impaired metabolic health in patients with COVID-19. *Nat. Rev. Endocrinol.* **16**, 341–342. https://doi.org:10.1038/s41574-020-0364-6 (2020).

4 Foulkes, A. S. *et al.* Understanding the link between obesity and severe COVID-19 outcomes: Causal mediation by systemic inflammatory response. *J. Clin. Endocrinol. Metab*. https://doi.org:10.1210/clinem/dgab629 (2021)

5 Zickler, M. *et al.* Replication of SARS-CoV-2 in adipose tissue determines organ and systemic lipid metabolism in hamsters and humans. *Cell. Metab.* **34**, 1–2. https://doi.org:10.1016/j.cmet.2021.12.002 (2022).

6 Goudswaard, L. J. *et al.* Effects of adiposity on the human plasma proteome: observational and Mendelian randomisation estimates. *Int. J. Obes. (Lond.)* **45**, 2221–2229. https://doi.org:10.1038/s41366-021-00896-1 (2021).

7 Zaghlool, S. B. *et al.* Revealing the role of the human blood plasma proteome in obesity using genetic drivers. *Nat. Commun.* **12**, 1279. https://doi.org:10.1038/s41467-021-21542-4 (2021).

8 Filbin, M. R. *et al.* Longitudinal proteomic analysis of severe COVID-19 reveals survival-associated signatures, tissue-specific cell death, and cell-cell interactions. *Cell. Rep. Med.* **2**, 100287. https://doi.org:10.1016/j.xcrm.2021.100287 (2021).

9 Skrivankova, V. W. *et al.* Strengthening the Reporting of Observational Studies in Epidemiology Using Mendelian Randomization. *JAMA* **326**, 1614. https://doi.org:10.1001/jama.2021.18236 (2021).

10 Skrivankova, V. W. *et al.* Strengthening the reporting of observational studies in epidemiology using mendelian randomisation (STROBE-MR): explanation and elaboration. *BMJ* **375**, n2233. https://doi.org:10.1136/bmj.n2233 (2021).

11 Ponsford, M. J. *et al.* Cardiometabolic Traits, Sepsis, and Severe COVID-19: A Mendelian Randomization Investigation. *Circulation* **142**, 1791–1793. https://doi.org:10.1161/CIRCULATIONAHA.120.050753 (2020).

12 Luo, S., Liang, Y., Wong, T. H. T., Schooling, C. M. & Au Yeung, S. L. Identifying factors contributing to increased susceptibility to COVID-19 risk: a systematic review of Mendelian randomization studies. *Int. J. Epidemiol.*, dyac076. https://doi.org:10.1093/ije/dyac076 (2022).

13 Zhou, S. *et al.* A Neanderthal OAS1 isoform protects individuals of European ancestry against COVID-19 susceptibility and severity. *Nat. Med.* **27**, 659–667. https://doi.org:10.1038/s41591-021-01281-1 (2021).

14 Gaziano, L. *et al.* Actionable druggable genome-wide Mendelian randomization identifies repurposing opportunities for COVID-19. *Nat. Med.* **27**, 668–676. https://doi.org:10.1038/s41591-021-01310-z (2021).

15 Bovijn, J., Lindgren, C. M. & Holmes, M. V. Genetic variants mimicking therapeutic inhibition of IL-6 receptor signaling and risk of COVID-19. *Lancet Rheumatol.* **2**, e658–e659. https://doi.org:10.1016/s2665-9913(20)30345-3 (2020).

16 Klaric, L. *et al.* Mendelian randomisation identifies alternative splicing of the FAS death receptor as a mediator of severe COVID-19. *medRxiv*. https://doi.org:10.1101/2021.04.01.21254789 (2021).

17 Niemi, M. E. K., Daly, M. J. & Ganna, A. The human genetic epidemiology of COVID-19. *Nat. Rev. Genet*. https://doi.org:10.1038/s41576-022-00478-5 (2022).

18 Pietzner, M. *et al.* ELF5 is a potential respiratory epithelial cell-specific risk gene for severe COVID-19. *Nat. Commun.* **13**, 4484. https://doi.org:10.1038/s41467-022-31999-6 (2022).

19 Recovery Collaborative Group. Tocilizumab in patients admitted to hospital with COVID-19 (RECOVERY): a randomised, controlled, open-label, platform trial. *Lancet* **397**, 1637–1645. https://doi.org:10.1016/S0140-6736(21)00676-0 (2021).

20 Holmes, M. V., Richardson, T. G., Ference, B. A., Davies, N. M. & Davey Smith, G. Integrating genomics with biomarkers and therapeutic targets to invigorate

cardiovascular drug development. *Nat. Rev. Cardiol.* **18**, 435–453. https://doi.org:10.1038/s41569-020-00493-1 (2021).

21 Manousaki, D., Mokry, L. E., Ross, S., Goltzman, D. & Richards, J. B. Mendelian Randomization Studies Do Not Support a Role for Vitamin D in Coronary Artery Disease. *Circ. Cardiovasc. Genet.* **9**, 349–356. https://doi.org:10.1161/CIRCGENETICS.116.001396 (2016).

22 Jiang, X. *et al.* Circulating vitamin D concentrations and risk of breast and prostate cancer: a Mendelian randomization study. *Int. J. Epidemiol.* **48**, 1416–1424. https://doi.org:10.1093/ije/dyy284 (2019).

23 Meng, X. *et al.* Phenome-wide Mendelian-randomization study of genetically determined vitamin D on multiple health outcomes using the UK Biobank study. *Int. J. Epidemiol.* **48**, 1425–1434. https://doi.org:10.1093/ije/dyz182 (2019).

24 Manson, J. E. *et al.* Vitamin D Supplements and Prevention of Cancer and Cardiovascular Disease. *N. Engl. J. Med.* **380**, 33–44. https://doi.org:10.1056/NEJMoa1809944 (2019).

25 Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. *Human Mol. Genet.* **27**, 3641–3649. https://doi.org:10.1093/hmg/ddy271 (2018).

26 Ferkingstad, E. *et al.* DECODE: Large-scale integration of the plasma proteome with genetics and disease. *Nat. Genet.* **53**, 1712–1721. https://doi.org:10.1038/s41588-021-00978-w (2021).

27 Pietzner, M. *et al.* Synergistic insights into human health from aptamer- and antibody-based proteomic profiling. *Nat. Commun.* **12**, 6822. https://doi.org:10.1038/s41467-021-27164-0 (2021).

28 Emilsson, V. *et al.* Co-regulatory networks of human serum proteins link genetics to disease. *Science* **361**, 769–773 (2018).

29 Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N. & Davey Smith, G. Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Stat. Med.* **27**, 1133–1163. https://doi.org:10.1002/sim.3034 (2008).

30 Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **44**, 512–525. https://doi.org:10.1093/ije/dyv080 (2015).

31 Swerdlow, D. I. *et al.* Selecting instruments for Mendelian randomization in the wake of genome-wide association studies. *Int. J. Epidemiol.* **45**, 1600–1616. https://doi.org:10.1093/ije/dyw088 (2016).

32 Zhang, J. *et al.* Plasma proteome analyses in individuals of European and African ancestry identify cis-pQTLs and models for proteome-wide association studies. *Nat. Genet*. https://doi.org:10.1038/s41588-022-01051-w (2022).

33 Staley, J. R. *et al.* PhenoScanner: a database of human genotype-phenotype associations. *Bioinformatics* **32**, 3207–3209. https://doi.org:10.1093/bioinformatics/btw373 (2016).

34 Tsukui, T. *et al.* Collagen-producing lung cell atlas identifies multiple subsets with distinct localization and relevance to fibrosis. *Nat. Commun.* **11**, 1920. https://doi.org:10.1038/s41467-020-15647-5 (2020).

35 Xie, T. *et al.* Single-Cell Deconvolution of Fibroblast Heterogeneity in Mouse Pulmonary Fibrosis. *Cell. Rep.* **22**, 3625–3640. https://doi.org:10.1016/j.celrep.2018.03.010 (2018).

36 Obeidat, M. e. *et al.* Molecular mechanisms underlying variations in lung function: a systems genetics analysis. *Lancet Respir. Med.* **3**, 782–795. https://doi.org:10.1016/s2213-2600(15)00380-x (2015).

37 Bhatt, S. P. *et al.* Discriminative Accuracy of FEV1:FVC Thresholds for COPD-Related Hospitalization and Mortality. *JAMA* **321**, 2438–2447. https://doi.org:10.1001/jama.2019.7233 (2019).

38 Davies, N. M., Holmes, M. V. & Davey Smith, G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *BMJ* **362**, k601. https://doi.org:10.1136/bmj.k601 (2018).

39 Holmes, M. V., Ala-Korpela, M. & Smith, G. D. Mendelian randomization in cardiometabolic disease: challenges in evaluating causality. *Nat. Rev. Cardiol.* **14**, 577–590. https://doi.org:10.1038/nrcardio.2017.78 (2017).

40 Au Yeung, S. L., Li, A. M., He, B., Kwok, K. O. & Schooling, C. M. Association of smoking, lung function and COPD in COVID-19 risk: a two-step Mendelian randomization study. *Addiction* **117**, 2027–2036. https://doi.org:10.1111/add.15852 (2022).

41 Lawlor, D. A., Tilling, K. & Davey Smith, G. Triangulation in aetiological epidemiology. *Int. J. Epidemiol.* **45**, 1866-1886. https://doi.org:10.1093/ije/dyw314 (2016)

42 Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315. https://doi.org:10.1038/ng.2892 (2014).

43 Saferali, A. *et al.* Characterization of a COPD-associated NPNT functional splicing genetic variant in human lung tissue via long-read sequencing. *medRxiv* https://doi.org:10.1101/2020.10.20.20203927 (2020).

44 Shrine, N. *et al.* New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. *Nat. Genet.* **51**, 481–493. https://doi.org:10.1038/s41588-018-0321-7 (2019).

45 Sakornsakolpat, P. *et al.* Genetic landscape of chronic obstructive pulmonary disease identifies heterogeneous cell-type and phenotype associations. *Nat. Genet.* **51**, 494–505. https://doi.org:10.1038/s41588-018-0342-2 (2019).

46 GTEx Consortium. The GTEx Consortium atlas of genetic regulatoryeffects across human tissues. *Science* **369**, 1318–1330 (2020).

47 Delorey, T. M. *et al.* COVID-19 tissue atlases reveal SARS-CoV-2 pathology and cellular targets. *Nature* **595**, 107–113. https://doi.org:10.1038/s41586-021-03570-8 (2021).

48 Woolf, B., Zagkos, L. & Gill, D. TwoStepCisMR: A Novel Method and R Package for Attenuating Bias in cis-Mendelian Randomization Analyses. *Genes* **13**. https://doi.org:10.3390/genes13091541 (2022).

49 Burgess, S., Daniel, R. M., Butterworth, A. S., Thompson, S. G. & Consortium, E. P.-I. Network Mendelian randomization: using genetic variants as instrumental variables to investigate mediation in causal pathways. *Int. J. Epidemiol.* **44**, 484–495. https://doi.org:10.1093/ije/dyu176 (2015).

50 Relton, C. L. & Davey Smith, G. Two-step epigenetic Mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease. *Int. J. Epidemiol.* **41**, 161–176. https://doi.org:10.1093/ije/dyr233 (2012).

51 Sanderson, E., Spiller, W. & Bowden, J. Testing and correcting for weak and pleiotropic instruments in two-sample multivariable Mendelian randomization. *Stat. Med.* **40**, 5434–5452. https://doi.org:10.1002/sim.9133 (2021).

52 Brandenberger, R. *et al.* Identification and characterization of a novel extracellular matrix protein nephronectin that is associated with integrin alpha8beta1 in the embryonic kidney. *J. Cell Biol.* **154**, 447–458. https://doi.org:10.1083/jcb.200103069 (2001).

53 Morimura, N. *et al.* Molecular cloning of POEM: a novel adhesion molecule that interacts with alpha8beta1 integrin. *J. Biol. Chem.* **276**, 42172–42181. https://doi.org:10.1074/jbc.M103216200 (2001).

54 Inagaki, F. F. *et al.* Nephronectin is upregulated in acute and chronic hepatitis and aggravates liver injury by recruiting CD4 positive cells. *Biochem. Biophys. Res. Commun.* **430**, 751–756. https://doi.org:10.1016/j.bbrc.2012.11.076 (2013).

55 Kon, S., Honda, M., Ishikawa, K., Maeda, M. & Segawa, T. Antibodies against nephronectin ameliorate anti-type II collagen-induced arthritis in mice. *FEBS Open Bio.* **10**, 107–117. https://doi.org:10.1002/2211-5463.12758 (2020).

56 O'Rourke, R. W. & Lumeng, C. N. Pathways to Severe COVID-19 for People with Obesity. *Obesity (Silver Spring)* **29**, 645–653. https://doi.org:10.1002/oby.23099 (2021).

57 Strunz, M. *et al.* Alveolar regeneration through a Krt8+ transitional stem cell state that persists in human lung fibrosis. *Nat. Commun.* **11**, 3559. https://doi.org:10.1038/s41467-020-17358-3 (2020).

58 Xie, T. *et al.* Mesenchymal growth hormone receptor deficiency leads to failure of alveolar progenitor cell function and severe pulmonary fibrosis. *Sci. Adv.* **7**, eabg6005. https://doi.org:doi:10.1126/sciadv.abg6005 (2021).

59 Stiegler, P. & Cunliffe, A. The Role of Diet and Exercise for the Maintenance of Fat-Free Mass and Resting Metabolic Rate During Weight Loss. *Sports Med.* **36**, 239–262. https://doi.org:10.2165/00007256-200636030-00005 (2006).

60 Jastreboff, A. M. *et al.* Tirzepatide Once Weekly for the Treatment of Obesity. *N. Engl. J. Med.* **387**, 205–216. https://doi.org:10.1056/NEJMoa2206038 (2022).

61 Heise, T. *et al.* Effects of subcutaneous tirzepatide versus placebo or semaglutide on pancreatic islet function and insulin sensitivity in adults with type 2 diabetes: a multicentre, randomised, double-blind, parallel-arm, phase 1 clinical trial. *Lancet Diabetes Endocrinol.* **10**, 418–429. https://doi.org:10.1016/s2213-8587(22)00085-7 (2022).

62 Garvey, W. T. *et al.* Two-year effects of semaglutide in adults with overweight or obesity: the STEP 5 trial. *Nat. Med.* **28**, 2083–2091. https://doi.org:10.1038/s41591-022-02026-4 (2022).

63 Blundell, J. *et al.* Effects of once-weekly semaglutide on appetite, energy intake, control of eating, food preference and body weight in subjects with obesity. *Diabetes Obes. Metab.* **19**, 1242–1251. https://doi.org:10.1111/dom.12932 (2017).

64 Pi-Sunyer, X. *et al.* A Randomized, Controlled Trial of 3.0 mg of Liraglutide in Weight Management. *N. Engl. J. Med.* **373**, 11–22. https://doi.org:10.1056/NEJMoa1411892 (2015).

65 Grannell, A. *et al.* Liraglutide Does Not Adversely Impact Fat-Free Mass Loss. *Obesity (Silver Spring)* **29**, 529–534. https://doi.org:10.1002/oby.23098 (2021).

66 Richardson, T. G., Fang, S., Mitchell, R. E., Holmes, M. V. & Davey Smith, G. Evaluating the effects of cardiometabolic exposures on circulating proteins which may contribute to severe SARS-CoV-2. *EBioMedicine* **64**, 103228. https://doi.org:10.1016/j.ebiom.2021.103228 (2021).

67 Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206. https://doi.org:10.1038/nature14177 (2015).

68 Pihlajamaki, J. *et al.* Expression of the splicing factor gene SFRS10 is reduced in human obesity and contributes to enhanced lipogenesis. *Cell Metab.* **14**, 208–218. https://doi.org:10.1016/j.cmet.2011.06.007 (2011).

69 Zhao, X. *et al.* FTO-dependent demethylation of N6-methyladenosine regulates mRNA splicing and is required for adipogenesis. *Cell Res.* **24**, 1403–1419. https://doi.org:10.1038/cr.2014.151 (2014).

70 Roundtree, I. A., Evans, M. E., Pan, T. & He, C. Dynamic RNA modifications in gene expression regulation. *Cell* **169**, 1187–1200. https://doi.org:10.1016/j.cell.2017.05.045 (2017).

71 Kim, M., Lee, C. & Park, J. Extracellular matrix remodeling facilitates obesity-associated cancer progression. *Trends Cell Biol.* https://doi.org:10.1016/j.tcb.2022.02.008 (2022).

72 Hemani, G. *et al.* The MR-Base platform supports systematic causal inference across the human phenome. *eLife* **7**, e34408. https://doi.org:10.7554/elife.34408 (2018).

73 Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575. https://doi.org:10.1086/519795 (2007).

74 The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74. https://doi.org:10.1038/nature15393 (2015).

75 Deeks, J.J., *et al*. Analysing data and undertaking meta-analyses. *Cochrane Handbook for Systematic Reviews of Interventions.* 241- 284. https://doi.org/10.1002/9781119536604.ch10 (2019).

76 Hemani, G., Bowden, J. & Davey Smith, G. Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Hum. Mol. Genet.* **27**, R195–R208. https://doi.org:10.1093/hmg/ddy163 (2018).

77 Burgess, S., Bowden, J., Fall, T., Ingelsson, E. & Thompson, S. G. Sensitivity Analyses for Robust Causal Inference from Mendelian Randomization Analyses with Multiple Genetic Variants. *Epidemiology* **28**, 30–42. https://doi.org:10.1097/ede.0000000000000559 (2017).

78 Pierce, B. L., Ahsan, H. & Vanderweele, T. J. Power and instrument strength requirements for Mendelian randomization studies using multiple genetic variants. *Int. J. Epidemiol.* **40**, 740–752. https://doi.org:10.1093/ije/dyq151 (2011).

79 Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383. https://doi.org:10.1371/journal.pgen.1004383 (2014).

80 Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501. https://doi.org:10.1093/bioinformatics/btw018 (2016).

81 Huffman, J. E. *et al.* Multi-ancestry fine mapping implicates OAS1 splicing in risk of severe COVID-19. *Nat. Genet.* **54**, 125–127. https://doi.org:10.1038/s41588-021-00996-8 (2022).

82 Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882–883. https://doi.org:10.1093/bioinformatics/bts034 (2012).

83 Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587 e3529. https://doi.org:10.1016/j.cell.2021.04.048 (2021).

84 Carter, A. R. *et al.* Understanding the consequences of education inequality on cardiovascular disease: mendelian randomisation study. *BMJ* **365**, l1855. https://doi.org:10.1136/bmj.l1855 (2019).

85 Grant, A. J. & Burgess, S. Pleiotropy robust methods for multivariable Mendelian randomization. *Stat. Med.* **40**, 5813-5830. https://doi.org:10.1002/sim.9156 (2021).

86 Vabistsevits, M. *et al.* Deciphering how early life adiposity influences breast cancer risk using Mendelian randomization. *Commun. Biol.* **5**, 337. https://doi.org:10.1038/s42003-022-03272-5 (2022).

87 Yoshiji, S. *et al.* Causal associations between body fat accumulation and COVID-19 severity: A Mendelian randomization study. *Front. Endocrinol. (Lausanne)* **13**, 899625. https://doi.org:10.3389/fendo.2022.899625 (2022).

88 Cichonska, A. *et al.* metaCCA: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis. *Bioinformatics* **32**, 1981–1989. https://doi.org:10.1093/bioinformatics/btw052 (2016).

3.13 Supplementary Information

Sample processing in the BQC19 cohort The samples underwent proteomic profiling using the SomaScan v4 assay. Up to 5,284 aptamers were measured for these samples. After removing any aptamers that represented non-human proteins or controls, we retained 4,907 aptamers for analysis, consistent with the deCODE study26. SomaLogic performed normalization and calibration steps to remove systematic biases, which are detailed in their technical note (https://www.mcgill.ca/genepi/files/genepi/bqc19_jgh_prt_tech_note_0.pdf). Blood samples were sent to SomaLogic at two-time points during the pandemic, providing two batches of proteomic measurements. For quality control, the protein levels were natural log-transformed and subsequently batch-corrected using the ComBat function implemented in the sva R package v3.44.081, which uses an empirical Bayesian framework to perform batch effect removal. Definition of SARS-CoV-2 infectious or non-infectious state in the BQC-19 cohort We defined infectious and non-infectious states as follows (all date ranges are inclusive): infectious samples were defined as blood samples collected from individuals who tested positive for SARS-CoV-2 test from 7 days before and up to, and including 14 days after the first date of SARS-CoV-2-associated symptoms. Non-infectious samples were defined as those meeting either of the following three criteria: (1) samples were collected from individuals who tested negative for SARS-CoV-2; (2) samples were collected within 31 days after, but 90 days before the first date of symptoms from patients whose positive SARS-CoV-2 test was confirmed within 7 days from the first symptom onset; (3) samples were collected at least 15 days before or 31 days after the positive SARS-CoV-2 test from patients whose positive SARS-CoV-2 test was confirmed at least 7 days before or after the first symptom onset. Further details can be found at https://github.com/richardslab/BQC19_phenotypeQC/blob/main/src/COVID-19%20Omics.svg. Definition of the COVID-19 severity outcomes in the BQC-19 cohort We defined the COVID-19 outcomes in accordance with those for the GWAS from COVID-19 Host Genetics Initiative. (I) Critically ill COVID-19: cases were defined as laboratory-confirmed SARS-CoV-2 infection by PCR or serology testing along with the requirement for respiratory support or death. (II) Hospitalization: cases were defined as

those who were hospitalized due to COVID-19-related symptoms. Controls were defined as individuals who tested negative for SARS-CoV-2.

3.14 Supplementary Tables

All supplementary tables can be found at https://doi.org/10.1038/s42255-023-00742-w

Transition from Chapter 3 to Chapter 4

In Chapter 3, we employed two-step proteome-wide MR, which consists of two sets of MR followed by mediation analysis and replication analysis to identify nephronectin as the circulating protein that mediates the effect of obesity on COVID-19 severity. This strategy enabled a deeper exploration into the causal biology and identified a potential therapeutic target while minimizing the risk of confounding and reverse causation. Our findings underscored the power of integrating MR with large-scale proteogenomics data, enabling causal inferences, dissecting complex biological systems, and supporting therapeutic target discovery.

Chapter 4 broadens the scope of our investigation. While Chapter 3 elucidated the role of the circulating protein nephronectin in mediating the effects of obesity on COVID-19 severity, the implications of obesity are multifaceted and extend beyond infectious diseases. Since the 1980s, obesity has been implicated in over 4 million deaths worldwide, with the leading cause of death being cardiovascular diseases. Moreover, obesity heightens the risk for conditions such as stroke and type 2 diabetes, significantly impacting global health. Thus, Chapter 4 employs the two-step MR methodology to explore the associations between obesity and a trio of cardiometabolic diseases: coronary artery disease, stroke, and type 2 diabetes. The overarching objective of Chapter 4 is to decipher the mechanistic pathways through which obesity amplifies the risk of these cardiometabolic diseases. By pinpointing key circulating proteins implicated in this relationship, we aim to spotlight potential therapeutic targets that could be prioritized for future drug development, as well as intervention and prevention strategies.

Chapter 4: COL6A3-derived endotrophin mediates the effect of obesity on coronary artery disease: an integrative proteogenomics analysis

Satoshi Yoshiji[1,2,3,4], Tianyuan Lu[1,5,6], Guillaume Butler-Laporte[1,7,8] Julia Carrasco-Zanini-Sanchez[9], Yiheng Chen[1,2], Kevin Liang[1,10], Julian Daniel Sunday Willett[1,10], Chen-Yang Su[10], Shidong Wang[11], Darin Adra[1], Yann Ilboudo[1], Takayoshi Sasako[1], Vincenzo Forgetta[1,6], Yossi Farjoun[1,12], Hugo Zeberg[13,14], Sirui Zhou[2,15], Michael Hultström[1,6,16,17], Mitchell Machiela[18], Nicholas J. Wareham[9], Vincent Mooser[2,15], Nicholas J. Timpson[19,20], Claudia Langenberg[9,21,22], J. Brent Richards[1,2,6,7,23*]

[1]Lady Davis Institute, Jewish General Hospital, McGill University, Montréal, Québec, Canada
[2]Department of Human Genetics, McGill University, Montréal, Québec, Canada
[3]Kyoto-McGill International Collaborative Program in Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto, Japan
[4]Japan Society for the Promotion of Science, Japan
[5]Department of Statistical Sciences, University of Toronto, Toronto, Ontario, Canada
[6]Prime Sciences, Montréal, Québec, Canada
[7]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, Québec, Canada
[8]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK
[9]MRC Epidemiology Unit, Institute of Metabolic Science, University of Cambridge, Cambridge, UK
[10]Quantitative Life Sciences Program, McGill University, Montréal, Québec, Canada
[11]SomaLogic, Boulder, Colorado, USA
[12]Fulcrum Genomics, Colorado, USA
[13]Department of Physiology and Pharmacology, Karolinska Institutet, Stockholm, Sweden
[14]Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
[15]McGill Genome Centre, McGill University, Montréal, Québec, Canada
[16]Anaesthesiology and Intensive Care Medicine, Department of Surgical Sciences, Uppsala University, Uppsala, Sweden

[17]Integrative Physiology, Department of Medical Cell Biology, Uppsala University, Uppsala, Sweden
[18]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, USA
[19]Integrative Epidemiology Unit, University of Bristol, Bristol, UK
[20]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK
[21]Computational Medicine, Berlin Institute of Health (BIH) at Charité – Universitätsmedizin Berlin, Germany
[22]Precision Healthcare University Research Institute, Queen Mary University of London, London, UK
[23]Department of Twin Research, King's College London, London, UK

***Correspondence:**
J. Brent Richards
Professor of Medicine, McGill University, Senior Lecturer, King's College London (Honorary), Pavilion H-413, Jewish General Hospital, 3755 Côte-Ste-Catherine Montréal, Québec, H3T 1E2, Canada.
CEO, 5 Prime Sciences, Montreal, Québec, Canada
Email: brent.richards@mcgill.ca
Tel: +1-514-340-8222 Fax: +1-514-340-7529

4.1 Abstract

Obesity strongly increases the risk of cardiometabolic diseases, yet the underlying mediators of this relationship are not fully understood. Given that obesity has broad effects on circulating protein levels, we investigated circulating proteins that mediate the effects of obesity on coronary artery disease (CAD), stroke, and type 2 diabetes—since doing so may prioritize targets for therapeutic intervention. By integrating proteome-wide Mendelian randomization (MR) screening 4,907 plasma proteins, colocalization, and mediation analyses, we identified seven plasma proteins, including collagen type VI α3 (COL6A3). COL6A3 was strongly increased by body mass index (BMI) (β = 0.32, 95% CI: 0.26–0.38, $P = 3.7 \times 10^{-8}$ per s.d. increase in BMI) and increased the risk of CAD (OR = 1.47, 95% CI:1.26–1.70, $P = 4.5 \times 10^{-7}$ per s.d. increase in COL6A3). Notably, COL6A3 is cleaved at its C-terminus to produce endotrophin, which was found to mediate this effect on CAD. In single-cell RNA sequencing of adipose tissues and coronary arteries, *COL6A3* was highly expressed in cell types involved in metabolic dysfunction and fibrosis. Finally, we found that body fat reduction can reduce plasma levels of COL6A3-derived endotrophin, thereby highlighting a tractable way to modify endotrophin levels. In summary, we provide actionable insights into how circulating proteins mediate the effect of obesity on cardiometabolic diseases and prioritize endotrophin as a potential therapeutic target.

4.2 Background

Over 1.9 billion people worldwide have obesity, which is strongly linked to the risk of many cardiometabolic diseases, including coronary artery disease (CAD), stroke, and type 2 diabetes[1,2]. There are many biological mechanisms whereby obesity causes disease, including metabolic dysfunction, inflammation, and endothelial damage[3]. However, most of the factors mediating this relationship are not yet fully understood. Therefore, identifying modifiable mediators of this relationship could yield potential therapeutic targets, which may be targeted pharmaceutically or non-pharmaceutically, for example with lifestyle interventions. Circulating proteins are potential candidates because obesity strongly influences the level of plasma proteins[4,5], and they play a critical role in disease development and progression. Moreover, circulating proteins can be measured and sometimes modulated[6], and their levels can be used as a surrogate measure of target engagement in drug development programs. Therefore, understanding their role in disease could provide multiple avenues to lessen the impact of obesity on cardiometabolic disease.

One way to understand the role of circulating proteins in disease has been through observational epidemiology studies. However, such studies are not ideal for identifying causal mediators of disease because they are prone to bias from unmeasured confounders and reverse causation[7,8], wherein the disease itself influences the protein level. What is therefore needed is a method to understand mechanisms of disease, while reducing such biases.

Mendelian randomization (MR) is a genetic epidemiology approach that can contribute to the understanding of the causal relationship between exposures and outcomes while minimizing the bias from confounding and avoiding reverse causation [6-12]. MR can be described as a natural experiment somewhat analogous to randomized controlled trials (RCTs) [13] because both rely upon randomization to reduce bias from confounding. In MR studies randomization is achieved through the random allocation of alleles at conception. Moreover, reverse causation can be theoretically avoided because genotype is always assigned prior to the onset of disease.

Despite these advantages, MR relies on three key assumptions[7,8]: there exist genetic variants that: (I) are associated with the risk factor of interest; (II) are not correlated with confounders of the exposure-outcome relationship; (III) affect the outcome only through the exposure (also known as lack of horizontal pleiotropy). Of these, the third assumption is the most problematic and can be a source of potential bias in MR. Nevertheless, when these main assumptions are met, MR can be a powerful tool to describe causal relationships in humans—free of model systems.

Advancements in large-scale proteomics have facilitated the discovery of genetic variants that influence plasma protein levels on a proteome-wide scale[14-16]. These genetic variants, referred to as protein quantitative trait loci (pQTLs), can be utilized in MR to estimate the causal effect of circulating protein levels on disease. Such methods have been successfully leveraged to prioritize therapeutic targets, including OAS1 for COVID-19[9,17] and IL6R for both COVID-19[18,19] and CAD[20], and ANGPTL3 for CAD[21]. As drug discovery is costly and prone to failure[22], proteo-genomics-based MR could play an important role since such studies could provide causal targets, which can be measured, thereby providing proximal read-out of drug target engagement, but also providing biomarkers for recruitment into clinical trials. Indeed, drugs with human genetics evidence are more likely to be successful in Phase II and III trials, and two-thirds of FDA-approved drugs in 2021 were supported by human genetics evidence[23,24].

Furthermore, MR methods can be leveraged to understand mediators of the biological pathways connecting obesity with cardiometabolic disease when deployed in a two-step study design[25,26]. Step 1 begins by estimating the effect of BMI on protein mediators. Step 2 estimates the effect of the identified mediators on the outcome of interest (in this case, cardiometabolic diseases). Previously, we have successfully used this approach to identify a circulating protein, nephronectin, that mediates the impact of obesity on COVID-19 severity[27].

In the present study, we conducted an integrative analysis of proteome-wide MR screening 4,907 proteins, statistical colocalization, and mediation analysis to identify circulating proteins that mediate the effects of obesity on CAD, ischemic stroke, cardioembolic stroke, and type 2 diabetes. We then focused on collagen type VI α3 (COL6A3) as a potential target, performing multiple follow-up analyses, including replication and single-cell sequencing analysis. Additionally, we evaluated the actionability of COL6A3 by assessing the effect of reducing body fat on its circulating protein level in multivariable MR and also assessed the implication of reducing the identified proteins on a phenome-wide association study.

4.3 Results

The overall study design and a summary of the results are illustrated in **Fig. 1**.
The study consisted of four main sections:
1) Step 1 MR, which evaluated the causal effect of body mass index (BMI) on the levels of circulating plasma proteins. We also evaluated the consistency of MR findings when BMI and body fat percentage were used as the exposures.
2) Step 2 MR, which assessed the causal effects of BMI-driven proteins on four cardiometabolic outcomes (CAD, ischemic stroke, cardioembolic stroke, and type 2 diabetes).
3) Follow-up analyses for COL6A3 and its cleavage product, known as endotrophin, which assessed its role in CAD.
4) Assessment of clinical actionability for COL6A3-derived endotrophin and other protein mediators by reducing body fat mass.
Each of these four steps and their results is described in detail below.

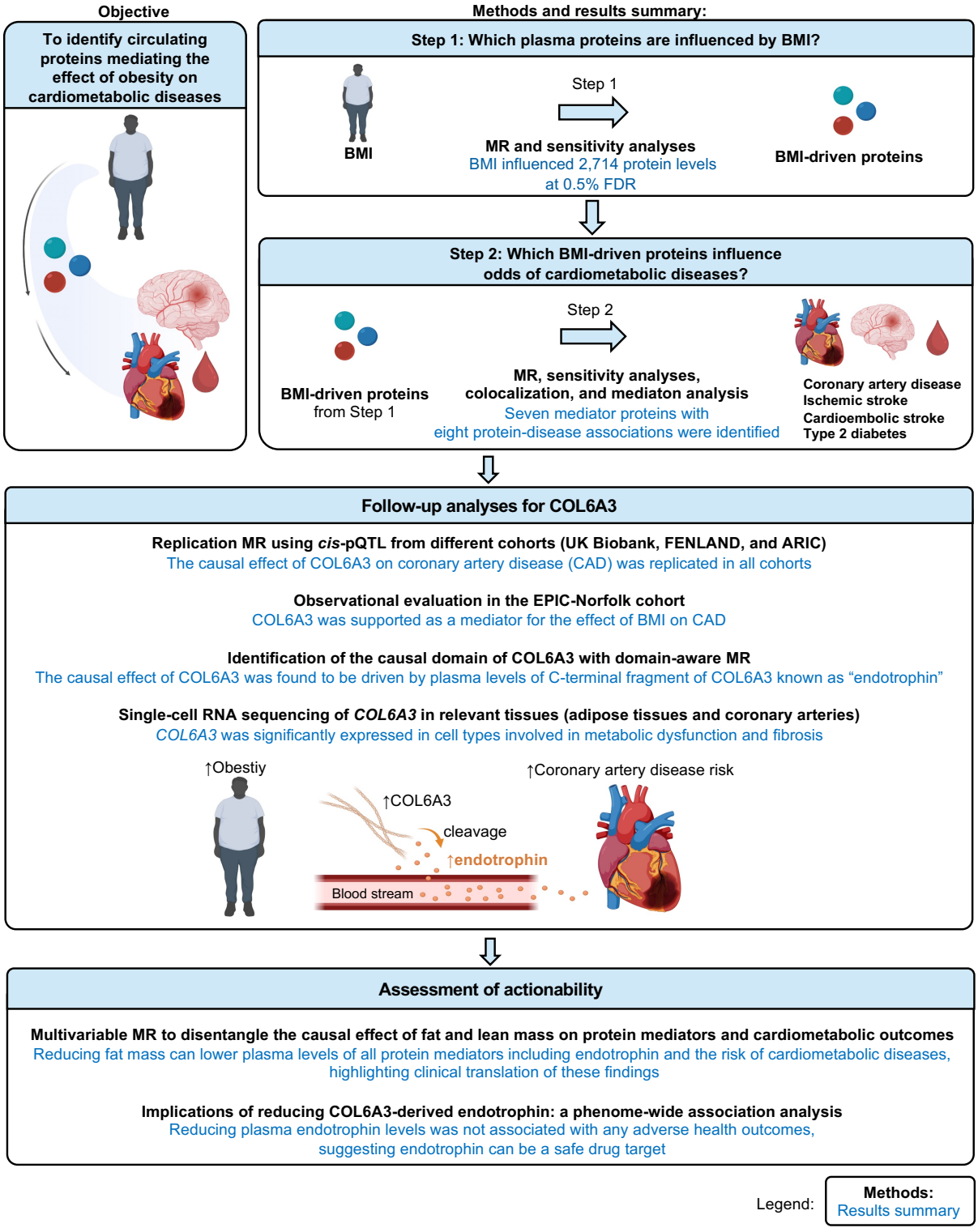


Figure 1. Study design.

To identify proteins that mediates the effect of obesity on cardiometabolic diseases, we used a two-step approach. In Step 1 Mendelian randomization (MR), we assessed the

effect of body mass index (BMI) on 4,907 plasma proteins, which led to the identification of 2,714 proteins influenced by BMI (referred to as “BMI-driven proteins”) using two-sample MR.

In Step 2 MR, we assessed the effect of these BMI-driven proteins on cardiometabolic diseases, again using two-sample MR.

In the subsequent sections, we conducted follow-up analyses of COL6A3 and evaluated the potential for actionability of this protein and other mediators we identified.

BMI: body mass index, *cis*-pQTL: *cis*-acting quantitative trait loci.

4.3.1 Step 1 MR: Identification of the causal effect of BMI on plasma protein levels

We evaluated the causal effect of BMI on 4,907 circulating proteins using the SomaScan v4 aptamer binding assay (SomaLogic, Boulder, CO). For clarity, we will refer to protein-targeting aptamers as "proteins" unless otherwise specified. We performed causal inference using two-sample MR, to estimate the effect of an exposure on an outcome of interest using two separate genome-wide association studies (GWAS); one for the BMI and the second for circulating proteins [13] (**Methods**). Specifically, we used the GWAS of BMI from the GIANT and UK Biobank consortia[28] (*n* = 681,275 individuals) and circulating protein levels from the deCODE study[15] (*n* = 35,559 individuals). In both studies we included only participants of European genetic ancestry (**Supplementary Table 1**). We performed two-sample MR, using the inverse variance weighted method as the primary analysis and then filtered these results dependent upon sensitivity analyses, including tests for heterogeneity, directional horizontal pleiotropy, and reverse causation. We used false discovery rate (FDR) correction with 0.5% as a strigent threshold for significance, given that many protein levels are correlated with each other and therefore a Bonferroni correction would be overly conservative (see **Methods**). No evidence of weak instrumental variables (suspected when F-statistics < 10) were found (**Supplementary Table 2**).

We found that BMI influenced 2,728 proteins, passing tests of significance, heterogeneity, and directional pleiotropy (**Supplementary Table 3**). However, among them, 14 showed evidence of reverse causation, wherein the protein influenced BMI (**Supplementary Table 4**), and these 14 proteins were removed from further analyses. Thus, we identified a total of 2,714 plasma proteins that are influenced by BMI. Hereafter, these 2,714 proteins are referred to as BMI-driven proteins (**Fig. 2a, 2b, and 2c**).

Additionally, we performed MR to evaluate the effect of body fat percentage on the same 4,907 plasma proteins (**Methods**). We did this because body fat percentage is considered to be a more direct proxy of obesity, whereas BMI is an easy-to-measure, clinically relevant proxy. [29] However, the sample size available to assess the genetic determinants of BMI is larger than that of body fat percentage, provide more precise estimates. We

found that body fat percentage influenced 94.7% of all BMI-driven proteins with the same direction of effect as BMI **(Fig. 2d**), illustrating a high concordance of results between the two different measures of obesity (r = 0.93; $P < 2.2 \times 10^{-16}$). Given the high concordance between MR results from BMI and body fat percentage, we proceed to Step 2 MR with BMI-driven protein results.

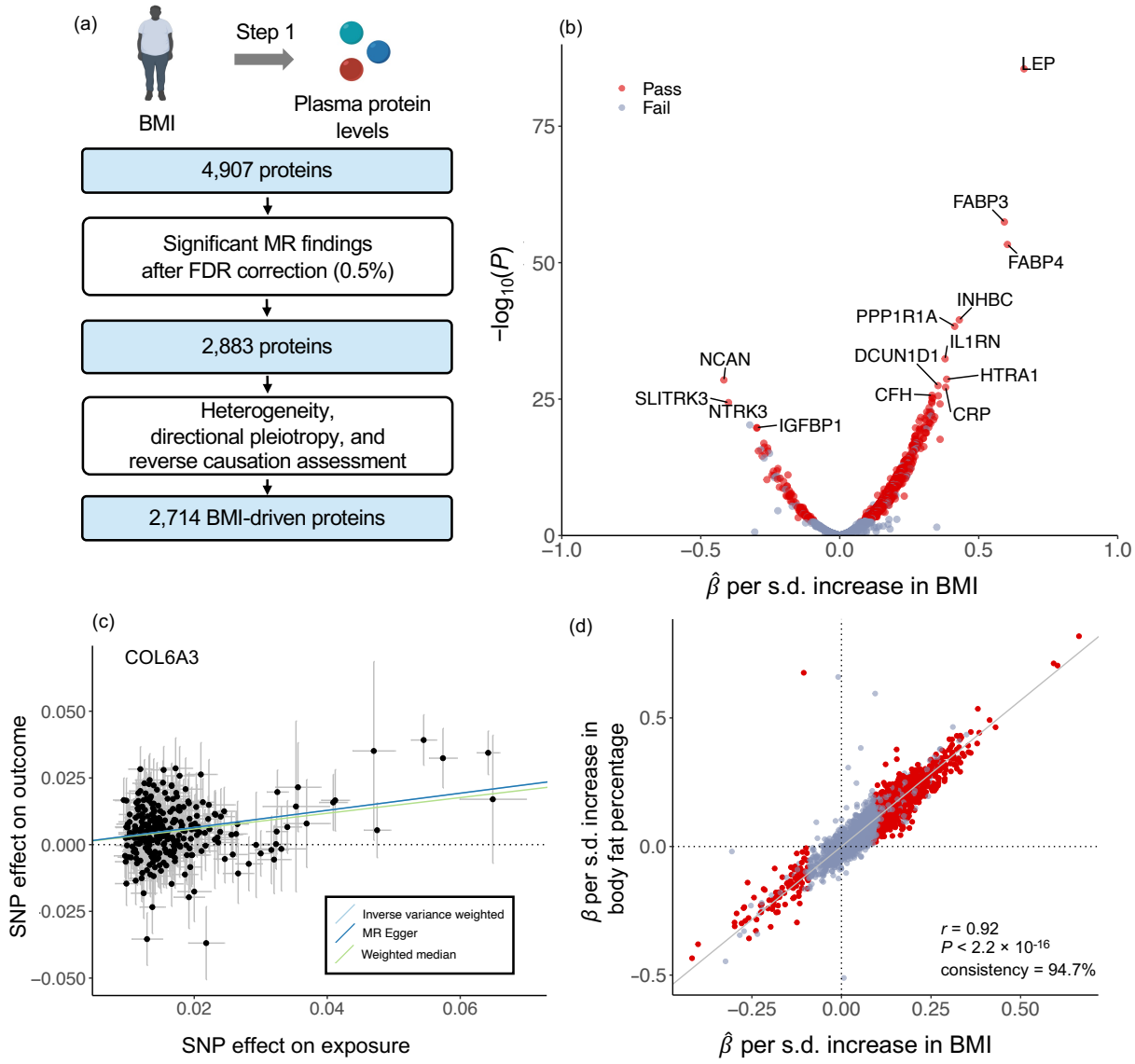


Figure 2. MR analyses for the effect of BMI on plasma protein levels.

(a) Flow diagram outlines Step 1 Mendelian randomization (MR).

(b) A volcano plot illustrates the effect of BMI on each plasma protein from MR analyses using the inverse variance weighted method. The x-axis represents beta estimates, and the y-axis represents -$\log_{10}(P)$ values from MR results. Red dots represent proteins that passed all tests, including significance with a false discovery rate (FDR) < 0.5%, as well as tests for heterogeneity, directional pleiotropy, and reverse causation. Grey dots represent proteins that failed any of these tests.

(c) MR scatter plot shows the effect of BMI on plasma levels of COL6A3 using the inverse-variance weighted method (primary analysis), weighted median, or MR-Egger slope methods.
(d) Directional consistency between MR results for the effect of BMI on plasma proteins and MR results for the effect of body fat percentage on plasma protein levels using the inverse variance weighted method.
The x-axis denotes beta estimates from MR results, and *r* denotes Pearson's correlation.

4.3.2 Step 2 MR: Identification of the causal effect of BMI-driven proteins on cardiometabolic diseases

Next, we estimated the causal effect of these BMI-driven proteins on CAD, ischemic stroke, cardioembolic stroke, and type 2 diabetes, again using two-sample MR (**Fig. 3a**). We used the BMI-driven protein levels identified in Step 1 MR as exposures. The outcomes were CAD, ischemic stroke, cardioembolic stroke, and type 2 diabetes (see **Methods**). To minimize the risk of bias from horizontal pleiotropy, we used *cis*-acting protein quantitative trait loci (*cis*-pQTLs) identified from 35,559 individuals from the deCODE study[15] as instrumental variables. In this context, instrumental variables are genetic variants that influence the exposure (i.e., circulating protein levels). We have defined *cis*-pQTLs as pQTLs that reside within a ± 1 Mb region around a transcription start site of a protein-coding gene. Since such *cis*-pQTLs would be likely to directly influence the circulating protein level by influencing the transcription or translation of mRNA from the gene that encodes the protein, they are less prone to bias from horizontal pleiotropy. Horizontal pleiotropy produces bias from the genetic variant influences the outcome independently of the circulating protein level.

To further reduce the risk of horizontal pleiotropy, we restricted instrumental variables to genetic variants that were *cis*-pQTLs to only one protein. To do so, we removed variants associated with more than two proteins in a *cis*-acting manner (**Fig. 3a; see Methods**). For the outcomes, we used the largest available GWAS for CAD[30] (181,522 cases and 1,165,690 controls), ischemic stroke, and cardioembolic stroke[31] (34,217 ischemic stroke cases, 7,193 cardioembolic stroke cases, and up to 2,703,029 controls), and type 2 diabetes[32] (80,154 cases and 853,816 controls).

Following MR with *cis*-pQTLs and sensitivity analyses (heterogeneity, pleiotropy, and reverse causation assessment), we performed colocalization to evaluate whether the pQTL of the protein of interest and the disease outcome shared a single causal variant around a 1-Mb (± 500 kb) region surrounding the lead *cis*-pQTL. As different linkage disequilibrium (LD) structures across different study populations may lead to bias in the

MR estimates, the presence of a shared single causal variant between the pQTL and the disease outcome can increase the robustness of MR findings (see **Methods**).

After MR with *cis*-pQTLs, sensitivity analyses, and colocalization, we identified 21 protein-disease associations that passed both step 1 and step 2 MR (**Supplementary Table 5**), including collagen type VI α3 (COL6A3) and PCSK9 for CAD, F11 for ischemic and cardioembolic stroke, and SF3B4 for type 2 diabetes. Among these proteins, COL6A3 was associated with the highest odds of CAD per one standard deviation (s.d.) increase in the protein levels (odds ratio (OR) = 1.47, 95% CI: 1.26–1.70, $P = 4.7 \times 10^{-7}$). We note that the finding of PCSK9 serves as a "positive control" and illustrates the utility of this method as PCSK9 is a well-known drug target, and its inhibition has been shown to reduced cardiovascular outcomes in multiple clinical trials[33-35]. Full results for CAD, ischemic stroke, cardioembolic stroke, and type 2 diabetes are provided in **Supplementary Table 6–9.**

As an additional filtering step, we performed mediation analyses for the identified protein-disease associations. To do this, we used the product of coefficients method[27,36-38] (**Methods**). Given that BMI increases the risk of cardiometabolic diseases ($\beta_{\text{BMI-to-cardiometabolic}} > 0$ in **Extended Fig. 1**; **Supplementary Table 10)**, we restricted the analysis to proteins that increased the risk of cardiometabolic diseases through their mediation pathway ($\beta_{\text{BMI-to-protein}} \times \beta_{\text{protein-to-cardiometabolic}} > 0$ in **Extended Fig. 1; Supplementary Fig. 10**). Among the 21 protein-disease associations, 8 met this condition. Notably, all eight protein-disease associations were supported by mediation analyses, suggesting that the effect of BMI on the cardiometabolic outcome was mediated, at least partially, by the circulating proteins (**Figure 3b; Supplementary Table 10**).

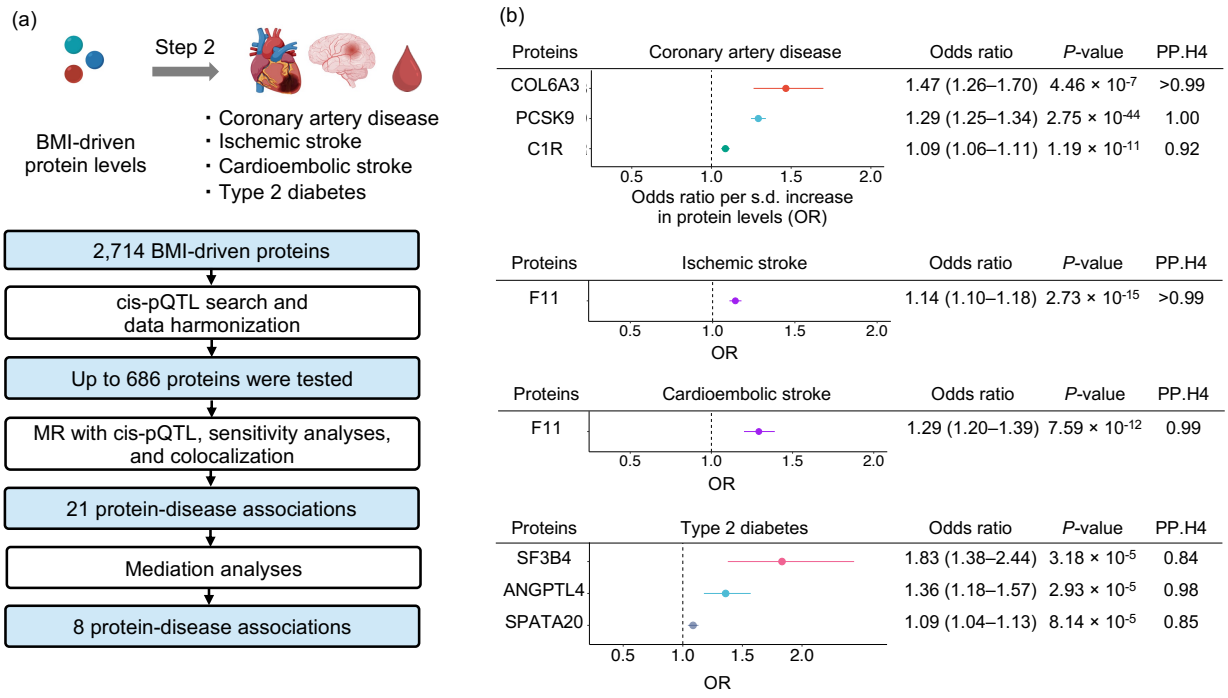


Figure 3. MR analyses for the effect of BMI-driven proteins on cardiometabolic diseases.

(a) Flow diagram of the Step 2 Mendelian randomization (MR) analyses.

(b) Forest plots for the effect of body mass index (BMI)-driven proteins on four cardiometabolic diseases (coronary artery disease, ischemic stroke, cardioembolic stroke, type 2 diabetes). The MR analyses were conducted using the largest available GWAS of coronary artery disease[30] (181,522 cases and 1,165,690 controls), ischemic stroke (34,217 cases and 2,703,029 controls), cardioembolic stroke[31] (7,193 cases and 2,703,029 controls), and type 2 diabetes[32] (80,154 cases and 853,816 controls).

PP.H4 = Posterior probability of having the shared causal variant (hypothesis H4 in colocalization).

4.3.3 Follow-up analyses of COL6A3 (collagen type VI α3)

Circulating COL6A3 levels had the strongest effects on CAD across all the mediators of the relationship between BMI and this outcome. We therefore sought to further test the hypothesis that COL6A3 mediates the relationship between obesity and cardiometabolic disease using analyses from orthogonal resources.

4.3.3.1 Replication MR using cis-pQTL from different cohorts

We evaluated whether the causal relationship between COL6A3 and CAD could be replicated using different sources of *cis*-pQTLs from other cohorts**.** For this, we conducted two-sample MR using *cis*-pQTLs from three additional cohorts: UK Biobank[39] (*n* = 35,571 individuals), Fenland[14] (*n* = 10,708 individuals), and ARIC[16] (*n* = 7,213 individuals). MR in all cohorts supported the causal effect of COL6A3 levels on CAD, in the same direction **(Supplementary Table 11**). Specifically, each s.d. increase in COL6A3 was associated with increased odds of CAD in UK Biobank[39] (OR = 1.30, 95% CI: 1.17–1.45, $P = 2.4 \times 10^{-6}$), Fenland (OR = 1.23, 95%CI: 1.12–1.35, $P = 8.9 \times 10^{-6}$), and ARIC (OR = 1.09, 95%CI: 1.05–1.13, $P = 1.6 \times 10^{-5}$). Notably, UK Biobank used Olink Explore 3072 assay[39], whereas deCODE[15], Fenland[14], and ARIC[16] used SomaScan v4 assay. Hence, concordant MR results using *cis*-pQTLs from the different studies from two different proteomic platforms further strengthened the evidence that COL6A3 partially mediates the relationship between obesity and CAD.

4.3.3.2 Observational epidemiological evaluation in the EPIC-Norfolk cohort

If testing a hypothesis using different designs yields similar results, it is less likely that the results are due to bias specific to one of the study designs. This is because different study designs have different bias architectures and concordant results across study designs strengthens causal inference because it is less likely that a single source of bias generated the results. Such testing has been referred to as a triangulation of evidence[40]. We therefore performed observational association analysis with a randomly selected sub-

cohort of the EPIC-Norfolk study (n = 872), which included 207 prevalent or incident cases of CAD (see **Methods**). EPIC-Norfolk is a population-based cohort from the United Kingdom. We found that increased BMI was associated with increased plasma levels of COL6A3 (β = 0.06, 95% CI: 0.04–0.08, $P = 8.5 \times 10^{-12}$), and a s.d. increase in plasma COL6A3 levels was associated with increased odds of CAD (OR = 1.34, 95% CI: 1.12–1.59, $P = 1.1 \times 10^{-3}$). The mediation analysis supported that plasma COL6A3 levels partially mediated the effect of BMI on CAD (**Supplementary Table 12**).

Given the robustness of these findings, we then explored the potential mechanism whereby COL6A3 may influence CAD.

4.3.3.3 Identification of the causal domain of COL6A3

Cleavage of proteins can influence their biological mechanism[41]. Previous studies have shown that the C-terminal domain, also known the Kunitz domain, of COL6A3 is proteolytically cleaved to form a biologically active fragment known as “endotrophin”. Endotrophin is produced in multiple tissues, including adipose tissue[41,42]. Endotrophin strongly induces fibrosis and inflammation, and recent evidence suggests that it is involved in obesity-induced metabolic dysfunction[41-46] (**Fig 4a**). Therefore, we evaluated whether this particular domain of COL6A3 is driving its effect on CAD.

The SomaScan v4 assay measures target protein levels using aptamers, which are short, single-stranded DNA or RNA molecules that can selectively bind to the target protein[47]. SomaScan v4 assay has two separate aptamers targeting two domains of COL6A3, the N-terminal and C-terminal (Kunitz domain) (**Methods).** These two separate aptamers thus allowed us to disentangle the effects of the N-terminal and C-terminal containing fragments of COL6A3.

Intriguingly, we found that the aptamer binding the C-terminal of COL6A3 (**Fig. 4b)** was associated with an increased risk of CAD (OR = 1.46 per s.d. increase in the protein level, 95% CI: 1.37–1.93, $P = 2.7 \times 10^{-8}$), whereas the aptamer binding the N-terminal (i.e., the

non-cleaved portion of COL6A3) was not associated with the risk of CAD (OR = 1.06, 95% CI: 0.96–1.18, P = 0.22) in domain-aware MR (**Supplementary Table 13**). These findings suggest that the C-terminal of COL6A3, which is cleaved into endotrophin, explains the effect of COL6A3 on CAD and the aptamer binding to the C-terminal of COL6A3 may be capturing the plasma levels of endotrophin or endotrophin-containing fragments. In the remainder of the manuscript, we refer to such fragments as endotrophin for clarity.

To further test the hypothesis that endotrophin is responsible for COL6A3's effect upon CAD, we tested whether *cis*-pQTLs from the Olink Explore 3072 assay[39,48] for COL6A3 were associated with CAD. The Olink Explore 3072 assay uses a polyclonal antibody to target the C-terminal (Kuniz domain) of COL6A3. The *cis*-pQTL (rs1050785) from UK-Biobank, which uses the Olink platform, was in high linkage disequilibrium (R^2 = 0.73) with *cis*-pQTL (rs11677932) of the C-terminal-targeting aptamer from the deCODE study but not in LD (R^2 = 0.0) with the *cis*-pQTL of the N-terminal-targeting aptamer of COL6A3 (rs2646260). We found that the *cis*-pQTL from the Olink platform was strongly associated with increased odds of CAD (OR = 1.32, 95%CI: 1.16–1.50, $P = 1.75 \times 10^{-5}$) (**Supplementary Table 13**), which was consistent with the finding using SomaScan v4 assay's aptamer binding the C-terminal of COL6A3. Taken together, these results provide evidence from orthogonal proteomic assays that circulating levels of C-terminus COL6A3-derived endotrophin likely explain the effect of COL6A3 levels on CAD.

Moreover, domain-aware MR analysis revealed that the aptamer targeting the C-terminal of COL6A3 (cleaved portion) was more strongly increased by an increase in BMI (β = 0.32, 95% CI: 0.26–0.38, $P = 3.7 \times 10^{-24}$) than the aptamer targeting N-terminal (uncleaved portion) (β = 0.10, 95% CI: 0.04–0.16, $P = 2.1 \times 10^{-3}$), as shown by non-overlapping confidence intervals. These findings indicate that an increase in BMI could increase both the expression of *COL6A3* and its cleavage, but has a preferential effect on the cleavage of COL6A3 into endotrophin.

4.3.3.4 COL6A3 expression analyses

We next explored the tissues in which COL6A3 is expressed using GTEx v8, which is a compendium of expression data from 49 tissues across 838 individuals[49]. In GTEx v8 (https://gtexportal.org/), *COL6A3* was significantly expressed in multiple tissues, including adipose tissue and coronary arteries when compared to the whole blood ($P < 0.001$) (**Fig. 4c**). Therefore, it is possible that these tissues may locally produce COL6A3 and consequently its cleavage product, endotrophin. While tissue-level examination of expression is helpful, such methods do not permit resolution to the cellular level. Considering that the adipose tissue is reported to be the primary source of COL6A3[46] and that the coronary artery is the location of primary lesions in CAD[50], to better understand the cell type of origin of COL6A3 we analyzed single-cell *COL6A3* expression in human white adipose tissues[51] (SCP1376 at https://singlecell.broadinstitute.org/) and coronary arteries in patients with CAD[50] (GSE131780 at https://www.ncbi.nlm.nih.gov/geo/).

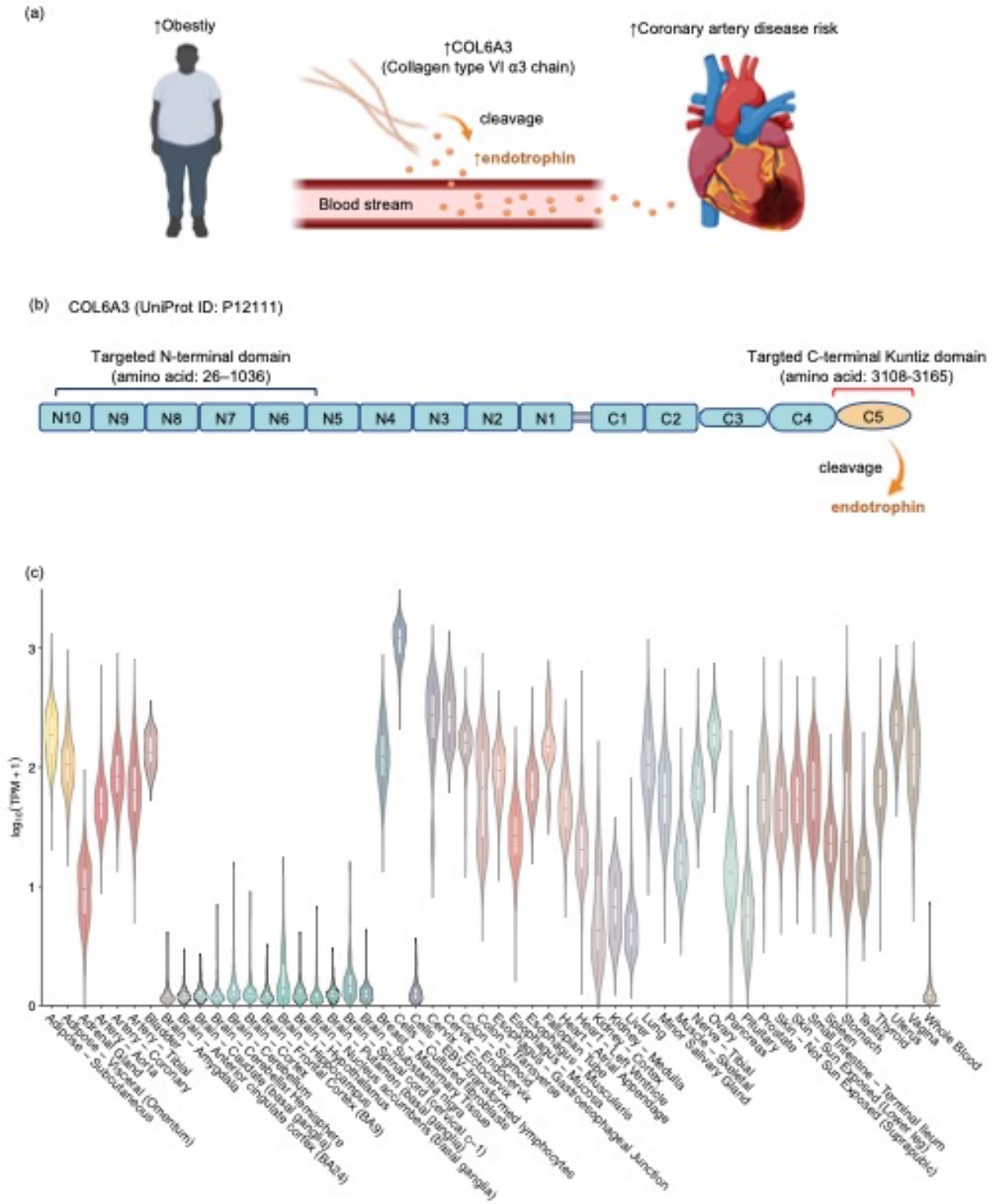


Figure 4. Follow-up analyses for collagen type VI α3 (COL6A3).

(a) Schematic illustration of proposed relationship between obesity, COL6A3 (Collagen type VI α3 chain), endotrophin, and coronary artery disease. Obesity leads to increased

production of COL6A3, whose C-terminal is cleaved into an active form termed endotrophin, which increases the risk of coronary artery disease.

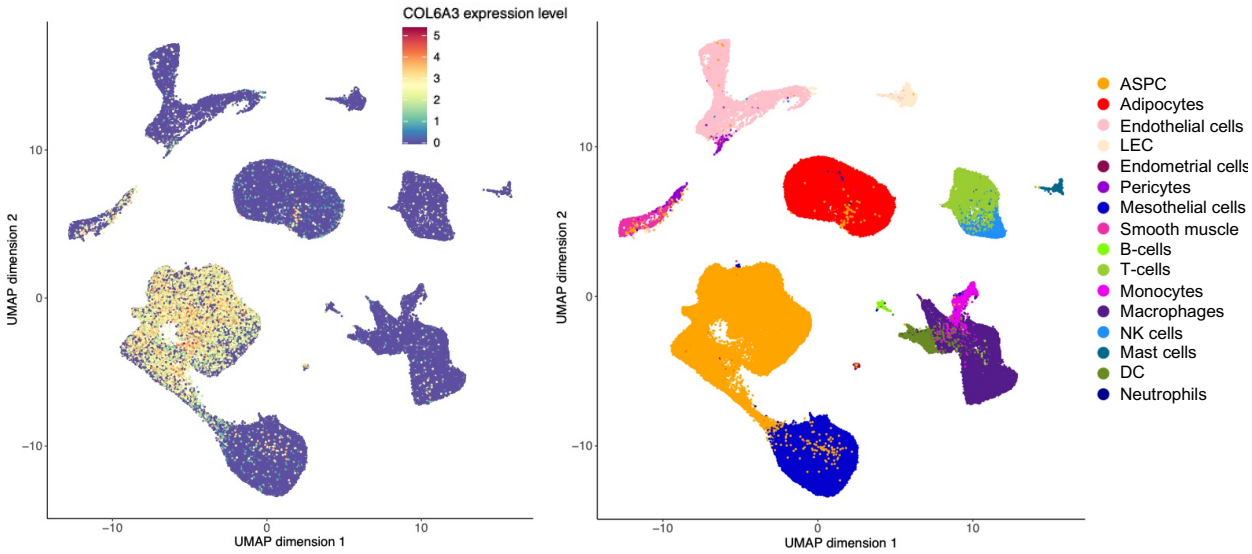
(b) Schematic diagram of COL6A3 (UniProt ID: P12111). COLA3 consists of a short collagenous region flanked by multiple von Willebrand factor type A (vWF-A) modules (N1–N10 in the N-terminal and C1,2 in the C-terminal). There are three additional C-terminal domains unique to COL6A3 (C3–C5), which are not present in other collagen type VI families. The most C-terminal domain (C5) is cleaved into a soluble protein termed endotrophin.

The two amino acid sequences targeted by the aptamers are as follows: the N-terminal-binding aptamer targets the amino acid sequence 26–1036 (uncleaved section), while the C-terminal aptamer targets the amino acid sequence 3108–3165 (cleaved section). The figure has been modified from ref[52,53].

(c) *COL6A3* expression profile in human tissues in GTEx v8[49]. *COL6A3* expression levels were represented on a log transcript per 10 thousand plus one (TPM + 1) scale.

In single-cell sequencing, *COL6A3* was significantly enriched in adipose progenitor/stem cells of adipose tissues when compared to other cell types in adipose tissues (permutation $P < 0.001$; see **Methods**) (**Fig. 5a)**. Given that these cell populations play critical roles in maintaining adipose tissue and metabolic function[54,55], the findings indicate that metabolic dysfunction may be an underlying biological mechanism whereby COL6A3 influences CAD. Additionally, we found that COL6A3 was significantly expressed in fibroblasts, which plays a key role in the atherosclerosis of the coronary artery[56], when compared to other cell types in the coronary artery (permutation $P < 0.001$; see **Methods**) (**Fig. 5b)**. Taken together, these findings suggested that these cell types may be responsible for the local production of COL6A3 in these tissues.

(a) *COL6A3* expression in adipose tissues



(b) *COL6A3* expression in coronary arteries

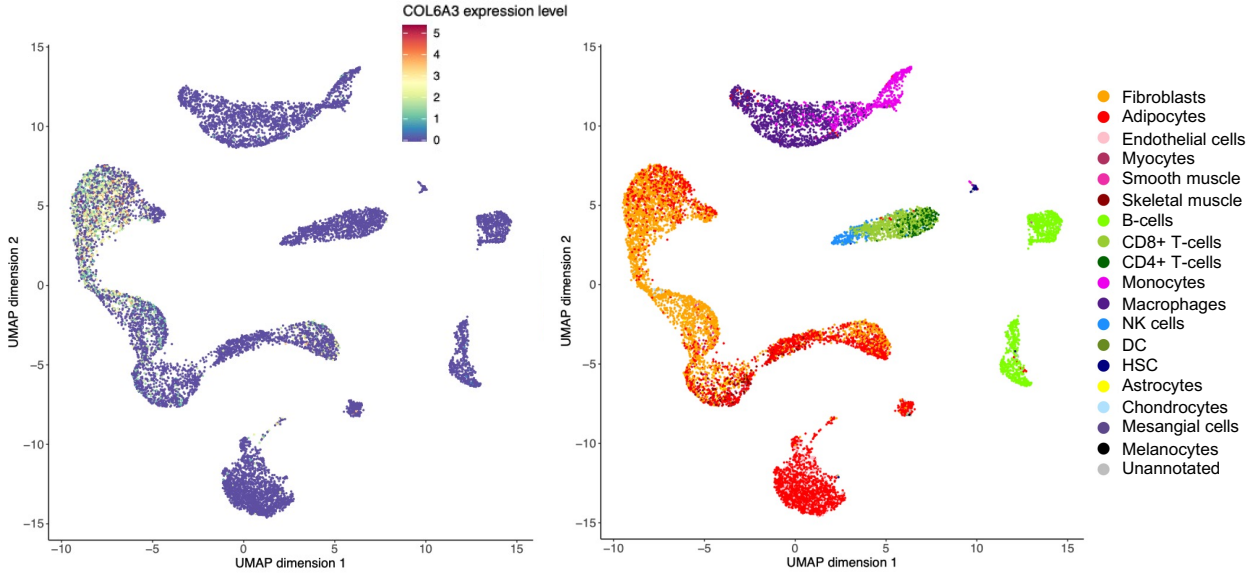


Figure 5. Single-cell sequencing analyses of COL6A3.

COL6A3 expression patterns in the adipose tissues (a) and coronary arteries (b). We obtained single-cell transcriptomic data of human adipose tissue from Emont et al. [51] (SCP1376 at https://singlecell.broadinstitute.org/) and the data of coronary arteries from Wirka et al. [50] (GSE131780 at the Gene Expression Omnibus database https://www.ncbi.nlm.nih.gov/geo/). ASPC: adipose stem and progenitor cells, LEC: lymphatic endothelial cells, NK: natural killer cells, DC: dendritic cells.

4.3.4 Assessment of clinical actionability

While identifying mediators of the effect of obesity on cardio-metabolic disease is relevant, such targets could become clinically relevant if their modification through weight loss or other methods influenced disease outcomes. We therefore explored whether reducing fat mass and/or increasing lean mass could improve plasma COL6A3-derived endotrophin and other protein levels, thereby reducing the risk of cardiometabolic diseases. For this, we used multivariable MR to evaluate the independent effects of body fat and lean mass (i.e., body fat-free mass) on the protein mediators and cardiometabolic disease outcomes (**Methods**).

We found that an s.d. increase in fat mass was independently associated with increased plasma levels of all protein mediators (COL6A3-derived endotrophin, F11, PCSK9, C1R, SPATA20, SF3B4, and ANGPTL4) (**Fig. 6b and Supplementary Table 14**) and increased odds of type 2 diabetes, CAD, and ischemic stroke. On the contrary, an s.d. increase in lean mass was independently associated with decreased plasma levels of some protein mediators including F11 and PCSK9 (**Fig. 6b and Supplementary Table 15**).

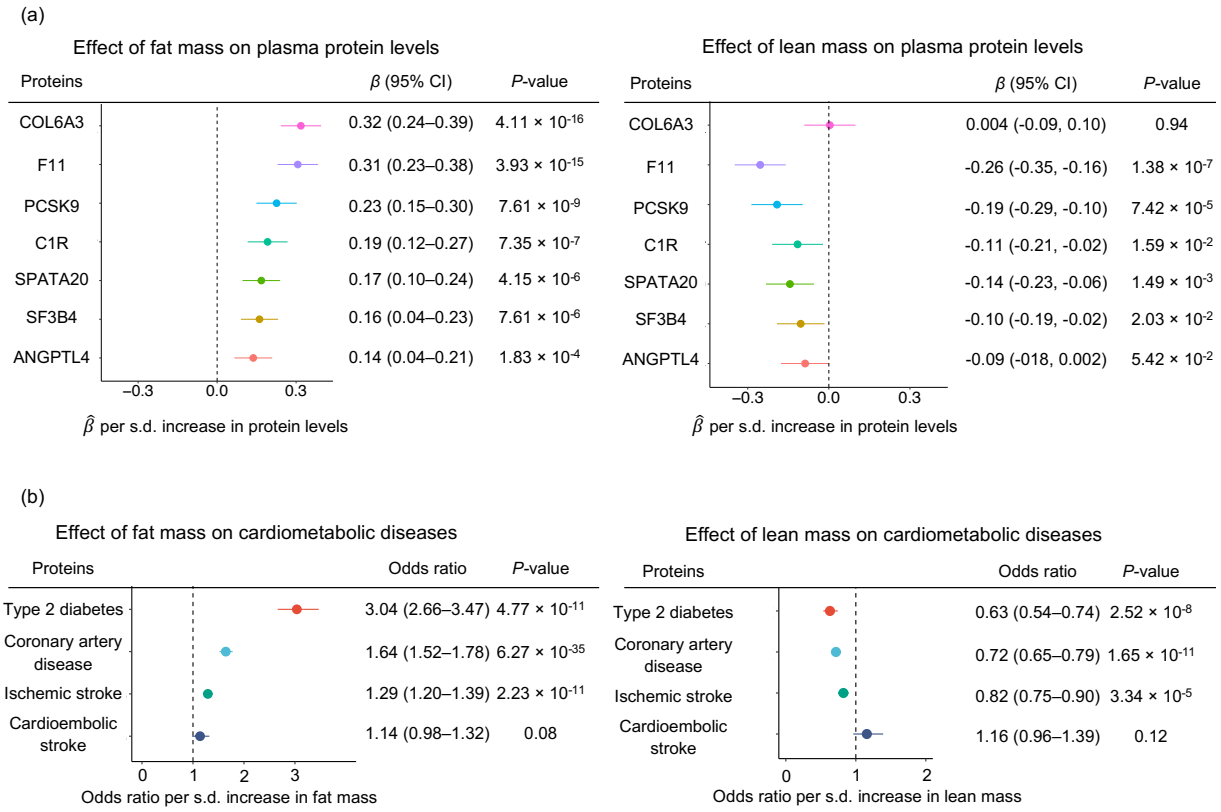


Figure 6. Multivariable MR analysis for evaluating the independent effects of fat mass and lean mass on plasma protein levels (a) and cardiometabolic diseases (b).

We performed multivariable Mendelian randomization (MR) using fat mass and lean mass as exposures and plasma protein levels of the seven protein mediators or cardiometabolic diseases as outcomes.

This has important clinical implications for actionability because interventions such as exercise, appropriate diet, or weight loss drugs such as the GLP-1 receptor agonist semaglutide and GLP-1/GIP co-agonist tirzepatide, which reduces body fat mass more than lean mass[57,58], could be effective in improving these protein levels and subsequently decreasing the risk of cardiometabolic diseases. However, future clinical trials are needed to confirm this hypothesis.

Lastly, we evaluated whether reducing COL6A3-derived endotrophin is associated with any adverse health outcomes using a phenome-wide association analysis in the UK Biobank, FinnGen, and the GWAS catalog. We did this because clinical trials for some drug candidates have been terminated due to unexpected adverse events in later stages of the trials[22,59]; thus, understanding the potential effects of perturbing the target on a phenome-wide level may to anticipate possible adverse events. Therefore, we assessed whether reducing COL6A3-derived endotrophin levels may have any implications on other traits. For this, we queried traits associated with the lead *cis*-pQTL of COL6A3 (rs11677932) from the deCODE study (a proxy for the COL6A3's C-terminal-derived endotrophin) in data from UK Biobank, FinnGen, and GWAS catalog using the Open Target Genetics (https://genetics.opentargets.org/) at $P < 1.0 \times 10^{-5}$. The phenome-wide association analysis revealed that decreased plasma levels of COL6A3-derived endotrophin (A-allele of rs11677932; β = -0.07, $P = 1.5 \times 10^{-14}$) was associated with decreased risk of coronary atherosclerosis (β = -0.05, $P = 1.0 \times 10^{-5}$), increased heel bone mineral density (β = 0.02, $P = 2.9 \times 10^{-12}$), and increased lung function (FEV1/FVC) (β = 0.02, $P = 5.2 \times 10^{-13}$) in addition to reduced risk of CAD (β = -0.03, $P = 2.9 \times 10^{-12}$) (**Supplementary Table 16**). This suggests that decreasing COL6A3-derived endotrophin may decrease the risk of multiple morbidities, including coronary atherosclerosis and CAD, and may also improve risk factors such as bone mineral density and lung function, offering COL6A3-derived endotrophin as an attractive therapeutic target.

4.4 Discussion

Obesity is a major risk factor of multiple diseases, and therapies are required that reduce its clinical consequences. Here, we identified seven protein mediators (from eight protein-disease associations) that partially mediate the effect of obesity on cardiometabolic diseases in humans. All of these protein levels, including COL6A3, could potentially be improved through body fat reduction, illustrating their possible clinical actionability. Furthermore, triangulation of evidence with multiple follow-up analyses indicated that endotrophin, which is derived from the cleavage of COL6A3, drives a part of the effect of obesity on CAD. These findings provide insights into how obesity causes cardiometabolic disease and provide circulating proteins that could be investigated as potential drug targets to lessen the public health burden of obesity.

The major finding of this study is the mediating role of endotrophin in the effect of obesity on CAD in humans. Previous studies reported endotrophin as an important hormone that induces metabolic dysfunction, fibrosis, and inflammation in rodent models[41-43,60], and cross-sectional studies in humans have found that increased circulating endotrophin level was observationally associated with cardiovascular events and all-cause mortality[44-46,61]. However, cross-sectional observational studies cannot disentangle cause and consequence. Therefore, our study, which utilized MR to make causal inferences, provides evidence that endotrophin acts as a causal mediator for the relationship between obesity and CAD in humans. Considering our findings that reducing COL6A3 and its cleaved product, endotrophin, can reduce the risk of CAD without apparent adverse health outcomes, directly targeting endotrophin can be an attractive therapeutic approach, and it may be particularly effective in individuals with obesity.

Notably, we found that the aptamer targeting C-terminal of COL6A3 (also called the Kunitz domain, which is cleaved into endotrophin) was more strongly affected by an increase in BMI than the aptamer targeting the N-terminal. This indicates that obesity may increase both *COL6A3* expression and the cleavage of COL6A3, but with a preferential influence on the cleavage of COL6A3 into endotrophin, leading to an increase in endotrophin levels. Several studies using mice models have shown that the

bone morphogenetic protein 1 (BMP1) [42], matrix metallopeptidase 14 (MMP14) [62], and other MMPs[63] can release the C-terminal of COL6A3 as endotrophin after proteolytic cleavage. However, as these studies were conducted using rodent models, further research is needed to establish whether the same applies to humans. Despite this, inhibition of BMP1 reduces scar formation and supports the survival of cardiomyocytes[64], which may be partly due to lower levels of endotrophin. Nevertheless, BMP1 also cleaves other procollagens into mature collagens, which introduces pleiotropy. Therefore, more research is necessary to determine how to selectively inhibit the cleavage of the C-terminal of COL6A3 to reduce endotrophin levels.

Our study also illuminated other proteins, such as ANGPTL4, which mediate the relationship between obesity and type 2 diabetes. Previous studies have shown that ANGPTL4 inhibits lipoprotein lipase[65], thereby reducing triglyceride levels[66]. Additionally, ANGPTL4 has also been implicated as an important player in obesity-induced glucose intolerance[65-69], consistent with our findings. Currently, an ANGPTL4 inhibitor, which is hepatocyte-targeting GalNAc-conjugated antisense oligonucleotides that downregulate ANGPTL4 levels in liver and adipose tissue, is in phase 1 clinical trial for hypertriglycedemia[70]. Our research indicates that this drug may be tested for the prevention of type 2 diabetes, and further clinical trials are required to evaluate the safety and efficacy of ANGPTL4 inhibition in humans. Another notable finding is F11 (coagulation factor XI) as a mediator of the effect of obesity on cardiometabolic disease. F11 is a critical player in the coagulation pathway and has been identified as causal for stroke by multiple studies[6,71]. However, few studies highlighted its role as a mediator. Currently, the F11 inhibitor, abelacimab[72], is in phase III clinical trial for venous thromboembolism (NCT05171049 at https://www.clinicaltrials.gov/). Our findings suggest that this drug may be effective for reducing the risk of ischemic stroke, especially for individuals with obesity.

This study has important limitations. First, we focused on analyzing data solely from European-ancestry individuals to prevent confounding by population stratification. While the ARIC cohort reported *cis*-pQTL for individuals of African ancestry[16], the sample size

(*n* = 1,871) is still limited when compared to data for those of European ancestry (deCODE study; *n* = 35,559). The same applies to CAD GWAS, with 181,522 CAD cases in European ancestry individuals[73] compared to only 17,247 cases in African ancestry individuals[74]. This limited sample size in African ancestry individuals reduces the statistical power of MR analysis. Therefore, further efforts are needed to increase the sample size of non-European-ancestry data. Second, we did not perform sex-stratified analysis due to the unavailability of sex-specific datasets. Third, while the mediation analyses results with both MR and observational evaluation in EPIC-Norfolk provided additional evidence supporting COL6A3-derived endotrophin as a causal mediator, it should be noted that the mediation analyses are based on additional assumptions[75]. Therefore, we used them as one of several orthogonal validation methods. Fourth, we did not explore the molecular mechanism whereby these proteins mediated the effect. Finally, although we triangulated multiple lines of evidence to propose several promising therapeutic targets that mediate an important proportion of the effect of obesity on cardiometabolic diseases (e.g., COL6A-derived endotrophin and ANGPTL4), future clinical trials are required to explore the effect of pharmacologically influencing these protein levels.

4.5 Conclusions

Our integrative proteogenomics analysis provide actionable insights into how circulating proteins mediate the effect of obesity on cardiometabolic diseases. We highlight the importance of body fat reduction to reduce the risk of cardiometabolic diseases and offers potential therapeutic targets, including COL6A3-derived endotrophin, which may be prioritized for drug development.

4.6 Methods

4.6.1 Step 1 MR

MR to evaluate the effect of BMI on plasma protein levels

We performed two-sample MR using BMI as exposure and circulating protein levels as outcomes. The BMI exposure data came from a meta-analysis GWAS of UK Biobank and GIANT involving 693,529 European-ancestry individuals[28] (**Supplementary Table 1**). For the outcomes, we used a GWAS of protein levels from the deCODE study[15], measuring 4,907 proteins in 35,559 individuals of European ancestry using the SomaScan assay v4 from SomaLogic (Boulder, Colorado, USA).

We performed two-sample MR using genome-wide significant and independent single nucleotide polymorphisms (SNPs) with $P < 5 \times 10^{-8}$ and $r^2 < 0.001$ as instrumental variables. We excluded SNPs in the human major histocompatibility complex region because of their complex linkage disequilibrium structures. Clumping was performed using PLINK v1.9 (https://www.cog-genomics.org/plink/) with 10-Mb window. When the instrumental variable SNPs were not present in the outcome GWAS, we identified proxy SNPs with $r^2 \geq 0.8$ using snappy v1.0 (https://gitlab.com/richards-lab/vince.forgetta/snappy/). To reduce the risk of weak instrument bias, we calculated F-statistics and evaluated whether they were above ten[76,77] (**Supplementary Table 2)**.

After harmonizing the exposure and outcome GWAS, we performed two-sample MR analysis using the inverse variance weighted method with a random-effects model as the primary analysis, implemented using TwoSampleMR v0.5.6. We set FDR < 0.005 (0.5%) as a stringent threshold for significance. We used FDR correction, given that many proteins are correlated with each other and that a Bonferroni correction can be overly conservative in such situations. However, we used a strict threshold of 0.5% instead of a conventional threshold of 5% to reduce false positive findings, as our intention was not to generate a complete list of potential associations, but rather to generate a smaller set of high-confidence findings. We used weighted median, weighted mode, and MR-Egger

slope as supplementary analyses to evaluate the directional concordance of the effect. Heterogeneity was tested using the I^2 statistic with results of $I^2 > 50\%$ and heterogeneity $P < 0.05$ considered as substantial heterogeneity. Directional horizontal pleiotropy was tested using the MR-Egger intercept test, and results with $P < 0.05$ were considered to indicate the presence of directional horizontal pleiotropy.

For reverse MR, wherein we examined the effect of plasma protein levels on BMI, we performed two-sample MR using *cis*-pQTLs variants from the deCODE study as exposures and BMI GWAS from UK Biobank as an outcome. We used the inverse variance weighted method or the Wald ratio method when only one SNP was available. We used FDR < 0.5% as a threshold for significance. For BMI GWAS, we used data from the UK Biobank instead of the meta-analysis GWAS of UK Biobank and GIANT because a number of *cis*-pQTL SNPs were not available in the latter due to the stringent quality control process of the meta-analysis.

MR to evaluate the effect of body fat percentage on plasma protein levels

While BMI is an easily measurable, clinically relevant proxy of obesity with the largest GWAS, body fat percentage is considered a more direct measurement of body fat accumulation. Thus, a high concordance between the BMI and body fat accumulation MR results may strengthen the inference from the findings of Step 1 MR for BMI.

Therefore, we performed two-sample MR using body fat percentage as exposure and plasma protein levels as outcomes. We used GWAS of body fat percentage in 454,633 European-ancestry individuals from UK Biobank (Accession ID: ukb-b-8909 at IEU OpenGWAS project) and the same protein levels for GWAS from the deCODE study as used in Step 1 MR.

4.6.2 Step 2 MR

MR with cis-pQTL to evaluate the effect of BMI-driven proteins on disease outcomes

Next, we performed two-sample MR using circulating protein levels as exposures and cardiometabolic diseases as outcomes, separately for each disease outcome. We used *cis*-pQTL variants from the deCODE study in 35,559 European-ancestry individuals[15] as the instrumental variables. The *cis*-pQTL was defined as pQTL located within 1 Mb (± 1Mb) from the transcription start site of the corresponding protein-coding gene. For the outcome, we used the largest available GWAS of CAD[30] (181,522 CAD cases and 1,165,690 controls), ischemic stroke, and cardioembolic stroke[31] (34,217 ischemic stroke cases, 7,193 cardioembolic stroke cases, and up to 2,703,029 controls), and type 2 diabetes[32] (80,154 type 2 diabetes cases and 853,816 controls). After data harmonization, we estimated the effect of each of the BMI-driven proteins on these outcomes. Two-sample MR was performed using TwoSampleMR v0.5.6 with an inverse variance weighted method and a random-effects model or Wald ratio when only one SNP was available as an instrumental variable. FDR < 0.5% was set as the threshold for significance. To minimize the risk of horizontal pleiotropy, we removed the variants associated with more than one protein in a *cis*-acting manner; therefore, we only retained the variants that were *cis*-pQTL for one protein (7008 out of 7572 variants are associated with only one protein in a *cis*-acting manner, and these 7008 variants are used as instrumental variables). To further test the absence of directional horizontal pleiotropy, we used the MR-Egger intercept test when applicable (i.e., if there are at least three instrumental variables). Additionally, we used the MR-Steiger test from TwoSampleMR v.0.5.6 to assess reverse causation, whereby cardiometabolic diseases influence plasma levels of proteins.

Colocalization

To ensure that the proteins and cardiometabolic diseases share the same causal genetic signal and avoid false-positive findings, We also performed colocalization using coloc R

package v5.1.0[78]. We evaluated whether *cis*-pQTL of the protein shared the same causal variant with cardiometabolic diseases within 1 Mb (± 500 kb). We used default prior of $p_1 = 10^{-4}$, $p2 = 10^{-4}$, and $p_{12} = 10^{-5}$ for coloc, where p_1 is a prior probability of trait 1 having a genetic association in the region, p_2 is a prior probability of trait 2 having a genetic association in the region, and p_{12} is a prior probability of the two traits having a shared genetic association. We considered the posterior probability of a shared causal variant (PP_{shared}) > 0.8 as evidence of colocalization.

Mediation analyses

As a validation analysis, we performed mediation analyses using network MR with a product of coefficients method. We did not adjust for the exposure (BMI) when estimating the effect of the mediator on the outcome ($\beta_{mediator\text{-}to\text{-}cardiometabolic}$) to avoid weak instrument bias. This approach has been adopted in multiple studies[26, 36-38].
Considering that the proportion mediated can be only estimated when the direction of effects is consistent between total causal effect and causal mediation effect, we restricted the analyses to proteins that meet the following criteria: $\beta_{total} \times \beta_{mediated} > 0$
where: β_{total} denotes the total effect (i.e., the effect of BMI on cardiometabolic diseases), and $\beta_{mediated}$ denotes the causal mediation effect (i.e., the effect mediated by the circulating proteins).

To estimate the causal mediation effects ($\beta_{mediated}$), we estimated the effect of BMI on the plasma protein levels ($\beta_{BMI\text{-}to\text{-}protein}$) and the effect of the plasma proteins on cardiometabolic diseases ($\beta_{protein\text{-}to\text{-}cardiometabolic}$), and then multiplied these values ($\beta_{mediated} = \beta_{BMI\text{-}to\text{-}protein} \times \beta_{protein\text{-}to\text{-}cardiometabolic}$). For this, we performed MR using the same instrumental variables as in Steps 1 and 2 of MR. Subsequently, we divided $\beta_{mediated}$ by β_{total} to estimate the proportion mediated and calculated the *P*-value under the null hypothesis that the protein of interest did not mediate the effect of BMI on the outcome of interest. We considered results with $P < 0.05$ to be significant. Since proteins can be correlated (e.g., in the same biological pathways), we did not apply Bonferroni correction.

4.6.3 Follow-up analyses

4.6.3.1 Replication MR using cis-pQTL from different cohorts

To replicate the causal estimates for the effect of COL6A3 on coronary artery disease, we conducted two-sample MR using *cis*-pQTLs from different cohorts: UK Biobank[39] (*n* = 35,571 individuals), Fenland[14] (*n* = 10,708 individuals), and ARIC (*n* = 7,213 individuals), using the same method as described in Step 2 MR.

4.6.3.2 Mediation analysis with individual-level data in the EPIC-Norfolk cohort

The EPIC-Norfolk study, a component of the pan-European EPIC Study, is a cohort of 25,639 middle-aged individuals from the general population of Norfolk, a county in Eastern England[79], who attended the baseline assessment between 1993–1998. We performed mediation analysis in a randomly selected subcohort (*n* = 872) of the EPIC-Norfolk study, in which proteomic profiling was performed using the SomaScan v4 assay. Death certificates and hospitalisation data were obtained using National Health Service numbers through linkage with the NHS digital database. Electronic health records were coded by trained nosologists according to the International Statistical Classification of Diseases and Related Health Problems, 9th (ICD-9) or 10th Revision (ICD-10). Participants were identified as CAD cases if the corresponding ICD-codes (ICD-9: 410-414, ICD-10:I20-I25) were registered on the death certificate (as the underlying cause of death or as a contributing factor), or as the cause of hospitalization. The current study is based on follow-up to the 31st March 2018. The case definition included all individuals identified as prevalent (at the baseline study assessment) or incident CAD cases over the follow-up period of over 20-years.

The plasma protein levels were normalized with rank-based inverse normal transformation using R package RNOmni v1.01. We used the product of coefficients methods to calculate the proportion mediated, as described above, using the R package mediation v4.5.0. We used linear regression adjusting for age and sex to estimate the effect of BMI on plasma COL6A3 levels and the effect of BMI, and logistic regression adjusting for age and sex to estimate the effect of BMI on the risk of CAD and the effect

of plasma COL6A3 levels on the risk of CAD. Significance of the indirect effect and the proportion mediated was estimated by computing unstandardized effects in 1000 bootstrapped samples, and calculating the corresponding 95% confidence intervals.

4.6.3.3 Identification of the causal domain of COL6A3

Target region of the SomaScan v4 assay and the Olink Explore 3072 assay

We used SomaScan Menu 7K (https://menu.somalogic.com/) to determine the target amino acid sequence of two aptamers for COL6A3 from on SomaScan v4 assay with additional support from SomaLogic (Boulder, Colorado, USA). We also obtained data on the target region of Olink Explore 3072 assay from Olink (Uppsala, Sweden). In SomaScan v4 assay, two aptamers target COL6A3: one for the C-terminal of COL6A3, also known as Kunitz domain (UniProt ID: P12111, target amino acid sequence: 3108-3165) and another for the N-terminal (UniProt ID: P12111, target amino acid sequence: 26-1036). In Olink Explore 3072 assay, the assay targets the C-terminal Kunitz domain of COL6A3 with polyclonal antibody (OID20292:v1).

Linkage disequilibrium of COPL6A3's cis-pQTL from the deCODE study and UK Biobank

We used the LDmatrix tool available at LDlink (https://ldlink.nci.nih.gov) with the 1000 genomes European samples as the reference panel[80] to calculate R^2 values between three SNPs: the *cis*-pQTL for COL6A3 from UK Biobank (rs1050785), the *cis*-pQTL of the C-terminal-targeting aptamer (rs11677932) from the deCODE study, and the *cis*-pQTL of the N-terminal-targeting aptamer of COL6A3 (rs2646260) from the deCODE study.

4.6.3.4 COL6A3 expression analyses

We downloaded bulk gene expression data in human tissues (GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_tpm.gct.gz) from GTEx portal (https://gtexportal.org/).

We generated the violin plots of *COL6A3* expression levels in each tissue using R v4.1.2. We used a two-sided Wilcoxon rank sum test to compare *COL6A3* expression in each tissue with its expression in the whole blood.

4.6.3.5 Single-cell RNA sequencing analysis

To investigate *COL6A3* expression at single-cell resolution in adipose tissues and coronary arteries, we reanalyzed the published expression matrix data from Emont et al. [51] (SCP1376 at https://singlecell.broadinstitute.org/) and Wirka et al. [50] (GSE131780 at Gene Expression Omnibus database https://www.ncbi.nlm.nih.gov/geo/), focusing on *COL6A3* expression. Following Wirka et al. [50], we removed low-quality cells that expressed < 500 genes or had a mitochondrial content > 7.5%, and genes expressed in < 5 cells. Cells expressing > 3,500 genes were also removed to avoid bias due to doublets. The retained gene expression profiles were normalized to library size. The top 2,000 most variable genes were selected after variance-stabilizing transformation using the FindVariableFeatures function in Seurat v4.0.6. Principal component analysis was performed based on these 2,000 most variable genes after scaling and centering. Nearest-neighbor graph construction was conducted based on the first 10 principal components using the FindNeighbors function in Seurat v4.0.6 with default settings. Cell clusters were identified using the FindClusters function in Seurat v4.0.6 with default settings. Uniform Manifold Approximation and Projection (UMAP) was also performed on the first 10 principal components. Two-dimensional visualization of the cell clusters was based on the first two UMAP dimensions. We used SingleR v2.0.0 to annotate the cell clusters with the Blueprint/ENCODE dataset as the reference using default settings.

To assess whether certain cell types express *COL6A3* more significantly than others, we performed 1,000 permutations of the cell type labels and calculated the frequency (permutation p-value) of the same cell type containing the same or a larger proportion of cells expressing *COL6A3* compared to all cells.

4.6.4 Follow-up analyses for the identified proteins

Assessment of actionability

To estimate the independent effects of fat mass and lean mass on plasma protein levels, we performed multivariable MR using fat mass and lean mass as exposures and protein levels as outcomes.

GWAS of fat mass and lean mass

We retrieved the GWAS data for fat mass and lean mass (i.e., fat-free mass) from UK Biobank through the OpenGWAS portal (https://gwas.mrcieu.ac.uk/). The data included 454,137 individuals of European ancestry for fat mass and 454,850 individuals for lean mass. The accession codes for the datasets were ukb-b-19393 for fat mass and ukb-b-13354 for lean mass. The fat mass and fat-free mass of the UK Biobank participants (second release, 2017) were evaluated by UK biobank with bioelectrical impedance analysis using the Tanita BC418MA body composition analyzer (Tanita, Tokyo, Japan).

Multivariable MR to evaluate the independent effect of fat mass and lean mass on protein levels and cardiometabolic diseases

To obtain instrumental variables, we applied the same selection criteria as in Steps 1 and 2 of MR ($P < 5 \times 10^{-8}$ and $r^2 < 0.001$), excluding those in the MHC region (GRCh37; chr6: 28,477,797–33,448,354). We performed data harmonization in TwoSampleMR v0.56 and multivariable MR with the inverse variance weighted method and a random-effect model in MVMR v0.3[77]. We calculated conditional F-statistics using MVMR v0.3[77] and evaluated whether they were above ten[76,77] (**Supplementary Table 14 and 15**). The phenotypic correlation matrix was calculated using metaCCA v1.22.0[81]. As additional sensitivity analyses, we performed multivariable MR-Egger analysis using MendelianRandomization v0.6.085[82].

Phenome-wide association study for rs11677932

We queried traits associated with the lead *cis*-pQTL of COL6A3 (rs11677932) from the deCODE study in the UK Biobank, FinGen, and GWAS catalog using the Open Target Genetics (https://genetics.opentargets.org/)

4.7 Ethical approval

All contributing cohorts obtained ethical approval from their intuitional ethics review boards. The contributing cohorts include UK Biobank, GIANT consortium, deCODEstudy, Fenland study, AGES Reykjavik study, INTERVAL study, CARDIoGRAMplusC4D, GIGASTROKE, and MAGIC consortium. The study was approved by the Norfolk Research Ethics Committee (no. 05/ Q0101/191), and all participants gave their informed written consent.

4.8 Data availability

We used GWAS summary statistics from the following source:
BMI GWAS from GIANT and UK Biobank (https://portals.broadinstitute.org/collaboration/giant/).
Plasma proteome GWAS from the deCODEstudy (https://www.deCODE.com/summarydata/), UK Biobank (https://doi.org/10.1101/2022.06.17.496443), Fenland (https://omicscience.org/apps/pgwas/), and the AGES Reykjavik study (https://doi.org/.1126/science.aaq1327).
We also used coronary artery disease GWAS from CARDIoGRAMplusC4D (http://www.cardiogramplusc4d.org/), stroke GWAS from GIGASTROKE (GCST90104534 and GCST90104535, at https://www.ebi.ac.uk/gwas/studies/), and type 2 diabetes GWAS from Mahajan *et al*. (https://doi.org/10.1038/s41588-022-01058-3).

For gene expression data, we used data from Nathan et al. (SCP498 at Single Cell Portal https://singlecell.broadinstitute.org/) and Wirka et al (GSE131780 at Gene Expression Omnibus database https://www.ncbi.nlm.nih.gov/geo/).

4.9 Code availability

We used R v4.1.2 (https://www.r-project.org/), TwoSampleMR v.0.5.6 (https://mrcieu.github.io/TwoSampleMR/), snappy v1.0 (https://gitlab.com/richards-lab/vince.forgetta/snappy), coloc v5.1.0 (https://chr1swallace.github.io/coloc/), PLINK v1.9 (http://pngu.mgh.harvard.edu/purcell/plink/), GCTA fastGWA v1.93.3 (https://yanglab.westlake.edu.cn/software/gcta/), and Seurat v4.0.6 (https://satijalab.org/seurat/). Custom codes will be made available on GitHub (https://github.com/satoshi-yoshiji/cm_proteogenomics/) upon publication of the manuscript.

4.10 Acknowledgments

The Richards research group is supported by the Canadian Institutes of Health Research (CIHR: 365825, 409511, 100558, 169303), the McGill Interdisciplinary Initiative in Infection and Immunity (MI4), the Lady Davis Institute of the Jewish General Hospital, the Jewish General Hospital Foundation, the Canadian Foundation for Innovation, the NIH Foundation, Cancer Research UK, Genome Québec, the Public Health Agency of Canada, McGill University, Cancer Research UK [grant number C18281/A29019] and the Fonds de Recherche Québec Santé (FRQS). J.B.R. is supported by an FRQS Mérite Clinical Research Scholarship. Support from Calcul Québec and Compute Canada is acknowledged. TwinsUK is funded by the Welcome Trust, Medical Research Council, European Union, the National Institute for Health Research (NIHR)-funded BioResource, Clinical Research Facility and Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust in partnership with King's College London. NJT is a Wellcome Trust Investigator (202802/Z/16/Z), is the PI of the Avon Longitudinal Study of Parents and Children (MRC & WT 217065/Z/19/Z), is supported by the

University of Bristol NIHR Biomedical Research Centre (BRC-1215-2001), the MRC Integrative Epidemiology Unit (MC_UU_00011/1) and works within the CRUK Integrative Cancer Epidemiology Programme (C18281/A29019).
The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from: the GTEx Portal on March 26, 2023.
S.Y. is supported by the Japan Society for the Promotion of Science. T.L. is supported by a Schmidt AI in Science Postdoctoral Fellowship, a Vanier Canada Graduate Scholarship, an FRQS doctoral training fellowship, and a McGill University Faculty of Medicine Studentship. G.B.L. is supported by scholarships from the FRQS, the CIHR, and Québec's ministry of health and social services. Y.C. is supported by an FRQS doctoral training fellowship and the Lady Davis Institute/TD Bank Studentship Award. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We acknowledge Biorender (biorender.com) for providing materials used to create the illustrative diagram.

4.11 Competing Interests

J.B.R. has served as an advisor to GlaxoSmithKline and Deerfield Capital. J.B.R.'s institution has received investigator-initiated grant funding from Eli Lilly, GlaxoSmithKline, and Biogen for projects unrelated to this research. J.B.R is the CEO of 5 Prime Sciences (www.5primesciences.com), which provides research services for biotech, pharma, and venture capital companies for projects unrelated to this research. T.L. and V.F. are employees of 5 Prime Sciences. The remaining authors declare no competing interests.

4.12 References

1 Powell-Wiley, T. M. *et al.* Obesity and Cardiovascular Disease: A Scientific Statement From the American Heart Association. *Circulation* **143**, e984-e1010 (2021). https://doi.org:doi:10.1161/CIR.0000000000000973

2 Czech, M. P. Insulin action and resistance in obesity and type 2 diabetes. *Nat Med* **23**, 804-814 (2017). https://doi.org:10.1038/nm.4350

3 Koenen, M., Hill, M. A., Cohen, P. & Sowers, J. R. Obesity, Adipose Tissue and Vascular Dysfunction. *Circ. Res.* **128**, 951-968 (2021). https://doi.org:10.1161/CIRCRESAHA.121.318093

4 Zaghlool, S. B. *et al.* Revealing the role of the human blood plasma proteome in obesity using genetic drivers. *Nat. Commun.* **12**, 1279 (2021). https://doi.org:10.1038/s41467-021-21542-4

5 Goudswaard, L. J. *et al.* Effects of adiposity on the human plasma proteome: observational and Mendelian randomisation estimates. *Int. J. Obes. (Lond.)* **45**, 2221-2229 (2021). https://doi.org:10.1038/s41366-021-00896-1

6 Zheng, J. *et al.* Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. *Nat. Genet.* **52**, 1122-1131 (2020). https://doi.org:10.1038/s41588-020-0682-6

7 Skrivankova, V. W. *et al.* Strengthening the reporting of observational studies in epidemiology using mendelian randomisation (STROBE-MR): explanation and elaboration. *BMJ* **375**, n2233 (2021). https://doi.org:10.1136/bmj.n2233

8 Skrivankova, V. W. *et al.* Strengthening the Reporting of Observational Studies in Epidemiology Using Mendelian Randomization. *JAMA* **326**, 1614 (2021). https://doi.org:10.1001/jama.2021.18236

9 Zhou, S. *et al.* A Neanderthal OAS1 isoform protects individuals of European ancestry against COVID-19 susceptibility and severity. *Nat. Med.* **27**, 659-667 (2021). https://doi.org:10.1038/s41591-021-01281-1

10 Yao, C. *et al.* Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nat. Commun.* **9**, 3268 (2018). https://doi.org:10.1038/s41467-018-05512-x

11 Miller, C. L. *et al.* Integrative functional genomics identifies regulatory mechanisms at coronary artery disease loci. *Nat. Commun.* **7**, 12092 (2016). https://doi.org:10.1038/ncomms12092

12 Lu, T., Forgetta, V., Greenwood, C. M. T., Zhou, S. & Richards, J. B. Circulating Proteins Influencing Psychiatric Disease: A Mendelian Randomization Study. *Biol. Psychiatry* **93**, 82-91 (2023). https://doi.org:10.1016/j.biopsych.2022.08.015

13 Burgess, S. *et al.* Using genetic association data to guide drug discovery and development: Review of methods and applications. *Am. J. Hum. Genet.* **110**, 195-214 (2023). https://doi.org:10.1016/j.ajhg.2022.12.017

14 Pietzner, M. *et al.* Mapping the proteo-genomic convergence of human diseases. *Science* **374**, eabj1541 (2021). https://doi.org:10.1126/science.abj1541

15 Ferkingstad, E. *et al.* DECODE: Large-scale integration of the plasma proteome with genetics and disease. *Nat. Genet.* **53**, 1712-1721 (2021). https://doi.org:10.1038/s41588-021-00978-w

16 Zhang, J. *et al.* Plasma proteome analyses in individuals of European and African ancestry identify cis-pQTLs and models for proteome-wide association studies. *Nat. Genet.* (2022). https://doi.org:10.1038/s41588-022-01051-w

17 Reis, G. *et al.* Early Treatment with Pegylated Interferon Lambda for Covid-19. *N. Engl. J. Med.* **388**, 518-528 (2023). https://doi.org:10.1056/NEJMoa2209760

18 Bovijn, J., Lindgren, C. M. & Holmes, M. V. Genetic variants mimicking therapeutic inhibition of IL-6 receptor signaling and risk of COVID-19. *Lancet Rheumatol.* **2**, e658-e659 (2020). https://doi.org:10.1016/s2665-9913(20)30345-3

19 Group, R. C. Tocilizumab in patients admitted to hospital with COVID-19 (RECOVERY): a randomised, controlled, open-label, platform trial. *Lancet* **397**, 1637-1645 (2021). https://doi.org:10.1016/S0140-6736(21)00676-0

20 Georgakis, M. K. *et al.* Interleukin-6 Signaling Effects on Ischemic Stroke and Other Cardiovascular Outcomes: A Mendelian Randomization Study. *Circ. Genom. Precis. Med.* **13**, e002872 (2020). https://doi.org:10.1161/CIRCGEN.119.002872

21 Dewey, F. E. *et al.* Genetic and Pharmacologic Inactivation of ANGPTL3 and Cardiovascular Disease. *N. Engl. J. Med.* **377**, 211-221 (2017). https://doi.org:10.1056/NEJMoa1612790

22 Pirmohamed, M. Pharmacogenomics: current status and future perspectives. *Nat. Rev. Genet.* (2023). https://doi.org:10.1038/s41576-022-00572-8

23 Ochoa, D. *et al.* Human genetics evidence supports two-thirds of the 2021 FDA-approved drugs. *Nat. Rev. Drug Discov.* **21**, 551 (2022). https://doi.org:10.1038/d41573-022-00120-3

24 King, E. A., Davis, J. W. & Degner, J. F. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. *PLoS Genet.* **15**, e1008489 (2019). https://doi.org:10.1371/journal.pgen.1008489

25 Dastani, Z. *et al.* Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: a multi-ethnic meta-analysis of 45,891 individuals. *PLoS Genet.* **8**, e1002607 (2012). https://doi.org:10.1371/journal.pgen.1002607

26 Richardson, T. G., Fang, S., Mitchell, R. E., Holmes, M. V. & Davey Smith, G. Evaluating the effects of cardiometabolic exposures on circulating proteins which may contribute to severe SARS-CoV-2. *EBioMedicine* **64**, 103228 (2021). https://doi.org:10.1016/j.ebiom.2021.103228

27 Yoshiji, S. *et al.* Proteome-wide Mendelian randomization implicates nephronectin as an actionable mediator of the effect of obesity on COVID-19 severity. *Nat. Metab.* **5**, 248-264 (2023). https://doi.org:10.1038/s42255-023-00742-w

28 Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. *Human Mol. Genet.* **27**, 3641-3649 (2018). https://doi.org:10.1093/hmg/ddy271

29 Peltz, G., Aguirre, M. T., Sanderson, M. & Fadden, M. K. The role of fat mass index in determining obesity. *Am. J. Hum. Biol.* **22**, 639-647 (2010). https://doi.org:10.1002/ajhb.21056

30 van der Harst, P. & Verweij, N. Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. *Circ. Res.* **122**, 433-443 (2018). https://doi.org:10.1161/CIRCRESAHA.117.312086

31 Mishra, A. *et al.* Stroke genetics informs drug discovery and risk prediction across ancestries. *Nature* **611**, 115-123 (2022). https://doi.org:10.1038/s41586-022-05165-3

32 Mahajan, A. *et al.* Multi-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation. *Nat. Genet.* **54**, 560-572 (2022). https://doi.org:10.1038/s41588-022-01058-3

33 Sabatine, M. S. *et al.* Evolocumab and Clinical Outcomes in Patients with Cardiovascular Disease. *N. Engl. J. Med.* **376**, 1713-1722 (2017). https://doi.org:10.1056/NEJMoa1615664

34 Schwartz, G. G. *et al.* Alirocumab and Cardiovascular Outcomes after Acute Coronary Syndrome. *N Engl J Med* **379**, 2097-2107 (2018). https://doi.org:10.1056/NEJMoa1801174

35 Ray, K. K. *et al.* Two Phase 3 Trials of Inclisiran in Patients with Elevated LDL Cholesterol. *N. Engl. J. Med.* **382**, 1507-1519 (2020). https://doi.org:10.1056/NEJMoa1912387

36 Woolf, B., Zagkos, L. & Gill, D. TwoStepCisMR: A Novel Method and R Package for Attenuating Bias in cis-Mendelian Randomization Analyses. *Genes* **13** (2022). https://doi.org:10.3390/genes13091541

37 Burgess, S., Daniel, R. M., Butterworth, A. S., Thompson, S. G. & Consortium, E. P.-I. Network Mendelian randomization: using genetic variants as instrumental variables to investigate mediation in causal pathways. *Int. J. Epidemiol.* **44**, 484-495 (2015). https://doi.org:10.1093/ije/dyu176

38 Relton, C. L. & Davey Smith, G. Two-step epigenetic Mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease. *Int. J. Epidemiol.* **41**, 161-176 (2012). https://doi.org:10.1093/ije/dyr233

39 Sun, B. B. *et al.* Genetic regulation of the human plasma proteome in 54,306 UK Biobank participants. *bioRxiv*, 2022.2006.2017.496443 (2022). https://doi.org:10.1101/2022.06.17.496443

40 Lawlor, D. A., Tilling, K. & Davey Smith, G. Triangulation in aetiological epidemiology. *Int. J. Epidemiol.* **45**, 1866-1886 (2016). https://doi.org:10.1093/ije/dyw314

41 Williams, L., Layton, T., Yang, N., Feldmann, M. & Nanchahal, J. Collagen VI as a driver and disease biomarker in human fibrosis. *FEBS J.* **289**, 3603-3629 (2022). https://doi.org:10.1111/febs.16039

42 Heumuller, S. E. *et al.* C-terminal proteolysis of the collagen VI alpha3 chain by BMP-1 and proprotein convertase(s) releases endotrophin in fragments of different sizes. *J. Biol. Chem.* **294**, 13769-13780 (2019). https://doi.org:10.1074/jbc.RA119.008641

43 Przyklenk, M. *et al.* Lack of evidence for a role of anthrax toxin receptors as surface receptors for collagen VI and for its cleaved-off C5 domain/endotrophin. *iScience* **25**, 105116 (2022). https://doi.org:10.1016/j.isci.2022.105116

44 Staunstrup, L. M. *et al.* Endotrophin is associated with chronic multimorbidity and all-cause mortality in a cohort of elderly women. *EBioMedicine* **68**, 103391 (2021). https://doi.org:10.1016/j.ebiom.2021.103391

45 Holm Nielsen, S. *et al.* The novel collagen matrikine, endotrophin, is associated with mortality and cardiovascular events in patients with atherosclerosis. *J. Intern. Med.* **290**, 179-189 (2021). https://doi.org:10.1111/joim.13253

46 Sun, K., Park, J., Kim, M. & Scherer, P. E. Endotrophin, a multifaceted player in metabolic dysregulation and cancer progression, is a predictive biomarker for the response to PPARgamma agonist treatment. *Diabetologia* **60**, 24-29 (2017). https://doi.org:10.1007/s00125-016-4130-1

47 Joshi, A. & Mayr, M. In Aptamers They Trust: The Caveats of the SOMAscan Biomarker Discovery Platform from SomaLogic. *Circulation* **138**, 2482-2485 (2018). https://doi.org:10.1161/CIRCULATIONAHA.118.036823

48 He, B., Huang, Z., Huang, C. & Nice, E. C. Clinical applications of plasma proteomics and peptidomics: Towards precision medicine. *Proteomics Clin. Appl.* **16**, e2100097 (2022). https://doi.org:10.1002/prca.202100097

49 Consortium, T. G. The GTEx Consortium atlas of genetic regulatoryeffects across human tissues. *Science* **369**, 1318-1330 (2020).

50 Wirka, R. C. *et al.* Atheroprotective roles of smooth muscle cell phenotypic modulation and the TCF21 disease gene as revealed by single-cell analysis. *Nat. Med.* **25**, 1280-1289 (2019). https://doi.org:10.1038/s41591-019-0512-5

51 Emont, M. P. *et al.* A single-cell atlas of human and mouse white adipose tissue. *Nature* **603**, 926-933 (2022). https://doi.org:10.1038/s41586-022-04518-2

52 Wang, J. & Pan, W. The Biological Role of the Collagen Alpha-3 (VI) Chain and Its Cleaved C5 Domain Fragment Endotrophin in Cancer. *Onco Targets Ther.* **13**, 5779-5793 (2020). https://doi.org:10.2147/OTT.S256654

53 Chen, P., Cescon, M. & Bonaldo, P. Collagen VI in cancer and its biological mechanisms. *Trends Mol. Med.* **19**, 410-417 (2013). https://doi.org:10.1016/j.molmed.2013.04.001

54 Miklosz, A., Nikitiuk, B. E. & Chabowski, A. Using adipose-derived mesenchymal stem cells to fight the metabolic complications of obesity: Where do we stand? *Obes. Rev.* **23**, e13413 (2022). https://doi.org:10.1111/obr.13413

55 Liao, X., Zhou, H. & Deng, T. The composition, function, and regulation of adipose stem and progenitor cells. *J. Genet. Genomics* **49**, 308-315 (2022). https://doi.org:10.1016/j.jgg.2022.02.014

56 Liberale, L. *et al.* The Role of Adipocytokines in Coronary Atherosclerosis. *Curr Atheroscler Rep* **19**, 10 (2017). https://doi.org:10.1007/s11883-017-0644-3

57 Blundell, J. *et al.* Effects of once-weekly semaglutide on appetite, energy intake, control of eating, food preference and body weight in subjects with obesity. *Diabetes. Obes. Metab.* **19**, 1242-1251 (2017). https://doi.org:10.1111/dom.12932

58 Jastreboff, A. M. *et al.* Tirzepatide Once Weekly for the Treatment of Obesity. *New Engl. J. Med.* **387**, 205-216 (2022). https://doi.org:10.1056/NEJMoa2206038

59 Fogel, D. B. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: A review. *Contemp. Clin. Trials Commun.* **11**, 156-164 (2018). https://doi.org:10.1016/j.conctc.2018.08.001

60 Sun, K. *et al.* Endotrophin triggers adipose tissue fibrosis and metabolic dysfunction. *Nat. Commun.* **5**, 3485 (2014). https://doi.org:10.1038/ncomms4485

61 Chirinos, J. A. *et al.* Endotrophin, a Collagen VI Formation–Derived Peptide, in Heart Failure. *NEJM Evidence* **1**, EVIDoa2200091 (2022). https://doi.org:doi:10.1056/EVIDoa2200091

62 Li, X. *et al.* Critical Role of Matrix Metalloproteinase 14 in Adipose Tissue Remodeling during Obesity. *Mol. Cell. Biol.* **40**, e00564-00519 (2020). https://doi.org:doi:10.1128/MCB.00564-19

63 Jo, W. *et al.* MicroRNA-29 Ameliorates Fibro-Inflammation and Insulin Resistance in HIF1alpha-Deficient Obese Adipose Tissue by Inhibiting Endotrophin Generation. *Diabetes* **71**, 1746-1762 (2022). https://doi.org:10.2337/db21-0801

64 Vukicevic, S. *et al.* Bone morphogenetic protein 1.3 inhibition decreases scar formation and supports cardiomyocyte survival after myocardial infarction. *Nat. Commun.* **13**, 81 (2022). https://doi.org:10.1038/s41467-021-27622-9

65 Lim, G. B. Genetics: Polymorphisms in ANGPTL4 link triglycerides with CAD. *Nat. Rev. Cardiol.* **13**, 245 (2016). https://doi.org:10.1038/nrcardio.2016.46

66 Dewey, F. E. *et al.* Inactivating Variants in ANGPTL4 and Risk of Coronary Artery Disease. *N. Engl. J. Med.* **374**, 1123-1133 (2016). https://doi.org:10.1056/NEJMoa1510926

67 Morris, A. Obesity: ANGPTL4 - the link binding obesity and glucose intolerance. *Nat. Rev. Endocrinol.* **14**, 251 (2018). https://doi.org:10.1038/nrendo.2018.35

68 Janssen, A. W. F. *et al.* Loss of angiopoietin-like 4 (ANGPTL4) in mice with diet-induced obesity uncouples visceral obesity from glucose intolerance partly via the gut microbiota. *Diabetologia* **61**, 1447-1458 (2018). https://doi.org:10.1007/s00125-018-4583-5

69 Gusarova, V. *et al.* Genetic inactivation of ANGPTL4 improves glucose homeostasis and is associated with reduced risk of diabetes. *Nat Commun* **9**, 2252 (2018). https://doi.org:10.1038/s41467-018-04611-z

70 Deng, M. *et al.* ANGPTL4 silencing via antisense oligonucleotides reduces plasma triglycerides and glucose in mice without causing lymphadenopathy. *J. Lipid Res.* **63**, 100237 (2022). https://doi.org:10.1016/j.jlr.2022.100237

71 Georgakis, M. K. & Gill, D. Mendelian Randomization Studies in Stroke: Exploration of Risk Factors and Drug Targets With Human Genetic Data. *Stroke* **52**, 2992-3003 (2021). https://doi.org:10.1161/STROKEAHA.120.032617

72 Verhamme, P. *et al.* Abelacimab for Prevention of Venous Thromboembolism. *N. Engl. J. Med.* **385**, 609-617 (2021). https://doi.org:10.1056/NEJMoa2105872

73 Aragam, K. G. *et al.* Discovery and systematic characterization of risk variants and genes for coronary artery disease in over a million participants. *Nat. Genet.* (2022). https://doi.org:10.1038/s41588-022-01233-6

74 Tcheandjieu, C. *et al.* Large-scale genome-wide association study of coronary artery disease in genetically diverse populations. *Nat. Med.* **28**, 1679-1692 (2022). https://doi.org:10.1038/s41591-022-01891-3

75 Carter, A. R. *et al.* Mendelian randomisation for mediation analysis: current methods and challenges for implementation. *Eur. J. Epidemiol.* **36**, 465-478 (2021). https://doi.org:10.1007/s10654-021-00757-1

76 Pierce, B. L., Ahsan, H. & Vanderweele, T. J. Power and instrument strength requirements for Mendelian randomization studies using multiple genetic variants. *Int. J. Epidemiol.* **40**, 740-752 (2011). https://doi.org:10.1093/ije/dyq151

77 Sanderson, E., Spiller, W. & Bowden, J. Testing and correcting for weak and pleiotropic instruments in two-sample multivariable Mendelian randomization. *Stat. Med.* **40**, 5434-5452 (2021). https://doi.org:10.1002/sim.9133

78 Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014). https://doi.org:10.1371/journal.pgen.1004383

79 Day, N. *et al.* EPIC-Norfolk: study design and characteristics of the cohort. European Prospective Investigation of Cancer. *Br. J. Cancer* **80 Suppl 1**, 95-103 (1999).

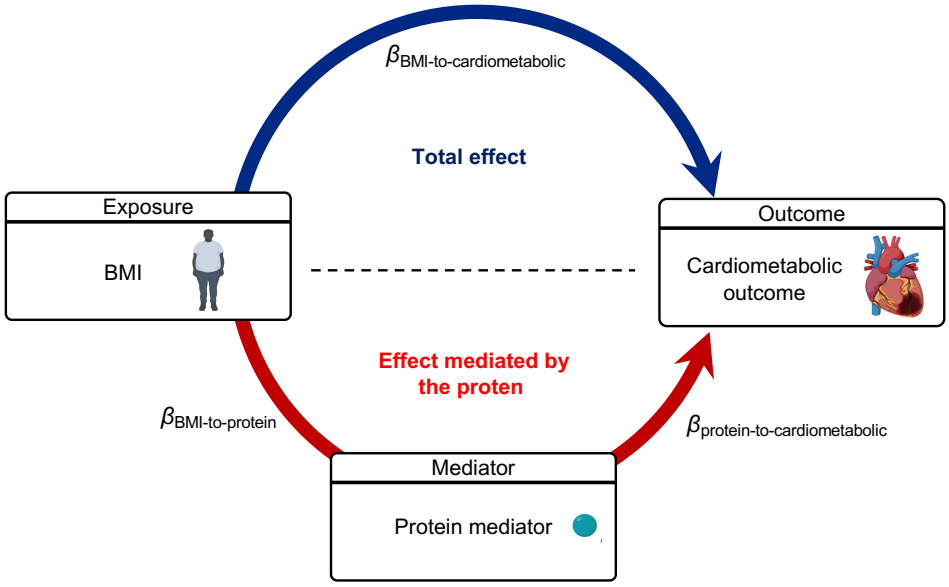
80 Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74 (2015). https://doi.org:10.1038/nature15393

81 Cichonska, A. *et al.* metaCCA: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation

analysis. *Bioinformatics* **32**, 1981-1989 (2016). https://doi.org:10.1093/bioinformatics/btw052

82 Grant, A. J. & Burgess, S. Pleiotropy robust methods for multivariable Mendelian randomization. *Stat. Med.* **40**, 5813-5830 (2021). https://doi.org:10.1002/sim.9156

4.13. Supplementary Figure.



Extended Figure 1. Schematic illustration of the mediation analysis.

The figure demonstrates the causal relationship between BMI, the protein mediator, and cardiometabolic diseases using directed acyclic graphs. The dark blue arrow represents the total effect of BMI on cardiometabolic diseases ($\beta_{BMI\text{-}to\text{-}cardiometabolic}$), while the red arrow represents the effect of BMI on cardiometabolic diseases mediated by the protein mediator. To calculate the ratio mediated, we used the product of coefficients method. This involved multiplying the effect of BMI on the protein mediator ($\beta_{BMI\text{-}to\text{-}protein}$) by the effect of the protein mediator on cardiometabolic diseases ($\beta_{protein\text{-}to\text{-}cardiometabolic}$) to estimate the effect mediated by the protein ($\beta_{mediated} = \beta_{BMI\text{-}to\text{-}protein} \times \beta_{protein\text{-}to\text{-}cardiometabolic}$). Subsequently, we divided $\beta_{mediated}$ by β_{total} to estimate the proportion mediated and calculated the *P*-value under the null hypothesis that the protein of interest did not mediate the effect of BMI on the outcome of interest.

BMI: body mass index, MR: Mendelian randomization.

4.14 Supplementary tables

All supplementary tables can be found at https://doi.org/10.1101/2023.04.19.23288706

Chapter 5. General discussions

The primary objective of this thesis is to utilize genetic epidemiology techniques in combination with large-scale genomics and proteomics datasets to elucidate the causal biology underlying obesity and its associated complications, thereby proposing potential therapeutic targets. Employing an integrative methodology, we offered clinically relevant insights into the role of circulating proteins in mediating the effects of obesity on COVID-19 severity and cardiometabolic diseases. Furthermore, we underscored potentially actionable therapeutic targets, including nephronectin for COVID-19 and endotrophin for CAD, highlighting the transformative potential of human genetics in therapeutic target identification.

This thesis program was initiated in response to the emerging COVID-19 pandemic, which started in late 2019. It was of critical importance to correctly understand the risk factors of severe COVID-19 and how to address them. As elevated BMI has emerged as one of the key risk factors for COVID-19 severity, we utilized MR to better understand and dissect the association between BMI and COVID-19 severity, which was presented in Chapter 2. Then, we moved on to identify the underlying mediators of this relationship using MR and large-scale genomics and proteomics. The motivation was to perform rapid and hypothesis-free proteome-wide scans to identify causal proteins and inform the causal biology and potential therapeutic targets. After the successful identification of nephronectin as a causal mediator for the effect of obesity on COVID-19 severity, we explored whether this framework, which we call “two-step MR”, can be leveraged to identify circulating proteins that mediate the effect of obesity on cardiometabolic diseases—another set of critical complications of obesity—since doing so may support drug target discovery for these conditions, which are the leading cause of obesity-related death. Below, we discuss each chapter in these contexts.

Chapter 2 assesses the causal relationship between body fat accumulation and COVID-19 severity. Previous research predominantly used BMI to study associations between

obesity and COVID-19; this is because BMI is a clinically relevant yet simplified measure, chosen primarily for its ease of measurement compared to other indices of obesity, such as waist circumference, waist-hip ratio, body fat percentage, and body fat mass. However, BMI, being derived only from height and weight, cannot differentiate between body fat mass and fat-free mass. Therefore, this raises questions concerning which of these components mediates the relationship between obesity and COVID-19 severity.

In this chapter, we employed MR to discern the independent causal relationships between body fat mass and fat-free mass on COVID-19 severity. Initial univariable two-sample MR analyses showed that an increase in body fat mass increased the risk of severe COVID-19. Interestingly, fat-free mass also appeared to be associated with a higher risk of COVID-19 severity, a finding that was somewhat unexpected given the general association of increased muscle mass with elevated metabolism.

Considering that body fat mass and fat-free mass are genetically intercorrelated, we estimated their independent causal effects using multivariable MR. This analysis confirmed that body fat mass independently contributed to the severity of COVID-19. These findings substantiated the causal impact of body fat accumulation on COVID-19 severity. Nonetheless, the specific biological mechanisms through which obesity increases the risk of severe COVID-19 remained to be elucidated.

In Chapter 3, our focus shifted towards understanding the causal biology underlying obesity and COVID-19 severity. Given the previous reports of the strong influence of obesity on plasma protein levels, we postulated that circulating proteins might serve as mediators in the relationship between obesity and COVID-19 severity. Another guiding rationale was the feasibility of measuring and, in some instances, modulating circulating proteins. Identifying these causal protein mediators could streamline drug target discovery.

To this end, we analyzed 4,907 plasma proteins to determine which were influenced by BMI. We used BMI in the initial analysis to maximize the statistical power given it being the largest GWAS to date among any proxies of obesity. Out of these, 1,216 proteins were found to be influenced by BMI. Continuing with MR, we evaluated their potential roles in increasing the risk of COVID-19 severity. In this analysis, we utilized a new set of GWAS of COVID-19 severity (i.e., critically ill COVID-19 and COVID-19 hospitalization) released by the COVID-19 Host Genetics Initiative. This GWAS doubled the sample size from the previous release, maximizing statistical power. Our MR analysis revealed that a standard deviation increase in nephronectin (NPNT) was associated with a heightened risk of severe COVID-19 outcomes (OR = 1.71, $P = 1.63 \times 10^{-10}$). To ensure the robustness of the findings, we also performed analyses using body fat percentage. Even though body fat percentage has a smaller sample size than BMI, it is considered a more direct proxy of body fat accumulation. The analyses using body fat percentage supported NPNT as a causal mediator.

As a further follow-up analysis for NPNT, we performed colocalization analyses of cis-pQTL of NPNT with eQTL and sQTL of NPNT in the lung. This showed that a specific splice variant of NPNT drives the association between NPNT and COVID-19. Subsequent mediation analyses validated the role of NPNT as a mediator in this relationship. Moreover, single-cell RNA sequencing analysis indicated NPNT expression in alveolar cells and lung fibroblasts of individuals who died of COVID-19.

Finally, using multivariable MR, we elucidated the independent causal effects of body fat mass and fat-free mass on COVID-19 severity. With the acquisition of a new set of GWAS of COVID-19 severity, we also repeated the multivariable MR, evaluating the causal effect of body fat mass and fat-free mass on COVID-19 severity, as performed in Chapter 2's study. We found that increased body fat mass and decreased fat-free mass were associated with heightened plasma levels of NPNT and an increased risk of

COVID-19 severity. This suggests that reducing body fat mass and increasing fat-free mass, predominantly muscle, can decrease plasma NPNT levels, reducing the risk of COVID-19 severity—highlighting the clinical relevance of NPNT.

In Chapter 4, we investigated whether the two-step approach, as employed in Chapter 3, could identify circulating proteins that mediate the effects of obesity on cardiometabolic diseases, specifically CAD, stroke, and type 2 diabetes. We integrated a two-step MR screening of 4,907 plasma proteins with colocalization and mediation analyses. This approach identified seven plasma proteins, placing a particular emphasis on collagen type VI α3 (COL6A3) in relation to CAD. Notably, for step 1 MR, we used FDR correction instead of the Bonferroni correction. We chose this approach because the Bonferroni correction, which assumes independence between tests, can be overly stringent for plasma proteins that often correlate with each other in complex biological systems. Using an FDR threshold of 0.5%, we identified 2,714 proteins influenced by BMI. This finding aligns with prior research that consistently demonstrated the substantial impact of obesity on the plasma proteome. The MR analysis using body fat percentage showed that body fat percentage influenced 94.7% of all BMI-driven proteins, with its direction of effect mirroring that of BMI. The high congruence between these two obesity proxies was compelling.

In step 2 MR, we found that COL6A3 was associated with an elevated risk of CAD (OR = 1.47, $P = 4.5 \times 10^{-7}$). Intriguingly, COL6A3 undergoes cleavage at its C-terminus, producing endotrophin. The domain-aware MR, which assessed the causal effects of both the N- and C-terminals of COL6A3 on CAD, indicated that only the C-terminal has a causal effect on CAD, while the N-terminal does not. We also observed that BMI preferentially elevates plasma levels of the C-terminal COL6A3 (cleaved site) compared to the N-terminal (non-cleaved site). This suggests that obesity increases plasma COL6A3 levels, leading to an increased cleavage of COL6A3, thereby elevating the risk of CAD.

To gain insights into the biological role of COL6A3, we conducted single-cell RNA sequencing analyses of adipose tissues and coronary arteries, revealing significant COL6A3 expression in cells associated with metabolic dysfunction and fibrosis. Furthermore, multivariable MR demonstrated that decreasing body fat mass can lower COL6A3 levels and reduce the risk of CAD. The phenome-wide association analysis for the proxy variant representing reduced plasma COL6A3 levels (cis-pQTL for COL6A3) indicated that reducing plasma COL6A3 levels was not associated with any adverse health outcomes.

Overall, by using a genetic epidemiology approach combined with extensive genomics, proteomics, and other omics datasets, we determined that endotrophin acts as a mediator for the effect of obesity on CAD in humans, positioning it as an attractive therapeutic target.

Chapter 6. Concluding remarks and future directions

This thesis utilized genetic epidemiology methods, combined with extensive genomics, proteomics, and other omics datasets, to provide clinically relevant insights into the role of circulating proteins in mediating the effects of obesity on COVID-19 severity and cardiometabolic diseases. We highlighted potential therapeutic candidates, including nephronectin for COVID-19 and endotrophin for coronary artery disease. These findings emphasize the transformative potential of human genetics in guiding therapeutic target identification.

However, there is much work to be done. First and foremost, not all proteins have pQTLs, and, therefore, cannot be evaluated in the pQTL MR setting. The size of the human proteome is a matter of debate, and numbers in the literature range from as few as 20,000 to several million[72]. Advancements in large-scale proteomics have facilitated the discovery of genetic variants that influence plasma protein levels on a proteome-wide scale[56,57,70]. The new aptamer-based assay, SomaScan, can measure up to 11,000 analytes, and the antibody-based assay, Olink, can measure up to 5,000 analytes. These platforms enabled large-scale measurements of plasma proteins, which was difficult with conventional high-throughput proteomics platforms[73-75]. Nevertheless, they are still far from covering the entirety of human proteins.

Just as importantly, measuring proteins in tissues other than plasma, such as adipose tissues and coronary arteries, has the potential to offer deeper insights into the biology with better resolution. As we are limited by what we can measure, further advancements in technology and the application of these methods to non-blood samples are essential to provide a more comprehensive understanding of the proteome and its role in the causal biology of human diseases.

Additionally, an increase in the statistical power of pQTLs is required. Currently, the largest pQTL study available is based on 54,306 individuals in the UK Biobank[65] using Olink, followed by the deCODE study using SomaScan, based on 35,559 Icelanders[56]. A major limitation of the UK Biobank study is the number of proteins, which is limited to 1,463 in the phase 1 release. Although it has been increased to 3,072 in the phase 2 release, it still only captures a minority of the human plasma proteome. Considering this limitation, we are currently working on a pQTL meta-analysis project of SomaScan-based studies, and the preliminary results look promising. We expect these pQTLs to be valuable resources for the community and facilitate a deeper understanding of the biology underpinned by human genomics, proteomics, and drug target discovery.

Furthermore, more diversity in human genomics and proteomics research is required. A majority of studies in human genetics have focused on individuals of European ancestry. Although there is an increasing effort to diversify the population, particularly in the GWAS field, pQTL data is still based on predominantly European-ancestry individuals. There are a few relatively large-scale pQTL datasets in African-American individuals [70,76] and East Asian-ancestry individuals[77], but these sample sizes are modest compared to those of European-ancestry.Genetic diversity is the largest in Africans[78] which highlights the importance of developing pQTL datasets specifically for African populations. Such efforts will enhance our understanding of the genetic architecture of the plasma proteome in non-European ancestry individuals, which are imperative for ensuring equitable, inclusive, and adequate representation of underrepresented populations in precision medicine.

Importantly, promoting the diversity of these datasets can benefit not only underrepresented populations but everyone. This is achieved by revealing novel mechanisms through which human genetics and proteomics influence disease risk and by highlighting potential therapeutic targets. A prime example of this is PCSK9. PCSK9 was first identified in French families with gain-of-function mutations in the PCSK9 gene (S127R and F216L), both of which are associated with autosomal dominant

hypercholesterolemia[79]. Subsequently, a loss-of-function variant (Q152H) was discovered in a French-Canadian family in Québec[80]. This variant was found to be linked with a substantial decrease in LDL cholesterol level. Additionally, the discovery of two nonsense mutations (Y142X and C679X), common in African Americans (with a combined allele frequency of 2%) but rare in European Americans (<0.1%), greatly enhanced our understanding of PCSK9. These two nonsense variants were associated with approximately 40% reduction in plasma LDL cholesterol levels, underscoring PCSK9's pivotal role in LDL metabolism and pinpointing it as a potential therapeutic target[9]. Large differences in allele frequencies across populations can provide such insights.

Another notable example is FinnGen, a large-scale biobank based in Finland. Their flagship paper demonstrated that many variants associated with common diseases in the Finnish population have an allele frequency of less than 5% in non-Finnish European individuals. This underscores the advantage of analyzing a well-phenotyped, isolated, and/or bottlenecked population[81]. In line with this, the author is engaged in creating a new biobank named BioPortal, a deeply phenotyped biobank with omics data from the Montreal population. This biobank aims to harness the cosmopolitan nature of Montreal[82], where approximately 40% are visible minorities (non-white), as well as the uniquely isolated French Canadian population, known for its richness in unique genetic variants[83-85]. These characteristics present a valuable opportunity to gain insights into disease biology and to identify potential drug targets.

In summary, this thesis provided novel insights into the mechanisms by which obesity influences the risk of COVID-19 and cardiometabolic diseases, underscoring the role of circulating proteins as mediators of the causal relationship of obesity with these diseases. By integrating genetic epidemiology methods with large-scale genomics and proteomics data, we have identified promising therapeutic targets, including nephronectin for COVID-19 and endotrophin for CAD. Additionally, we spotlight potential avenues for further exploration, harnessing the power of human genomics and

proteomics, such as the pQTL meta-analysis and the deeply phenotyped multi-omics biobank in the cosmopolitan city of Montreal. These strategies should be maximized to expedite drug discovery with the ultimate aim of transforming clinical care.

Master references

1 Powell-Wiley, T. M. *et al.* Obesity and Cardiovascular Disease: A Scientific Statement From the American Heart Association. *Circulation* **143**, e984-e1010 (2021). https://doi.org:doi:10.1161/CIR.0000000000000973

2 Czech, M. P. Insulin action and resistance in obesity and type 2 diabetes. *Nat. Med.* **23**, 804-814 (2017). https://doi.org:10.1038/nm.4350

3 Bluher, M. Obesity: global epidemiology and pathogenesis. *Nat. Rev. Endocrinol.* **15**, 288-298 (2019). https://doi.org:10.1038/s41574-019-0176-8

4 Saltiel, A. R. & Olefsky, J. M. Inflammatory mechanisms linking obesity and metabolic disease. *J. Clin. Invest.* **127**, 1-4 (2017). https://doi.org:10.1172/JCI92035

5 Ochoa, D. *et al.* Human genetics evidence supports two-thirds of the 2021 FDA-approved drugs. *Nat. Rev. Drug Discov.* **21**, 551 (2022). https://doi.org:10.1038/d41573-022-00120-3

6 Pirmohamed, M. Pharmacogenomics: current status and future perspectives. *Nat. Rev. Genet.* (2023). https://doi.org:10.1038/s41576-022-00572-8

7 Trajanoska, K. *et al.* From target discovery to clinical drug development with human genetics. *Nature* **620**, 737-745 (2023). https://doi.org:10.1038/s41586-023-06388-8

8 Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N. & Davey Smith, G. Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Stat. Med.* **27**, 1133-1163 (2008). https://doi.org:10.1002/sim.3034

9 Cohen, J. *et al.* Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. *Nat. Genet.* **37**, 161-165 (2005). https://doi.org:10.1038/ng1509

10 Sabatine, M. S. *et al.* Evolocumab and Clinical Outcomes in Patients with Cardiovascular Disease. *N. Engl. J. Med.* **376**, 1713-1722 (2017). https://doi.org:10.1056/NEJMoa1615664

11 Schwartz, G. G. *et al.* Alirocumab and Cardiovascular Outcomes after Acute Coronary Syndrome. *N. Engl. J. Med.* **379**, 2097-2107 (2018). https://doi.org:10.1056/NEJMoa1801174

12 Zaghlool, S. B. *et al.* Revealing the role of the human blood plasma proteome in obesity using genetic drivers. *Nat. Commun.* **12**, 1279 (2021). https://doi.org:10.1038/s41467-021-21542-4

13 Goudswaard, L. J. *et al.* Effects of adiposity on the human plasma proteome: observational and Mendelian randomisation estimates. *Int. J. Obes. (Lond.)* **45**, 2221-2229 (2021). https://doi.org:10.1038/s41366-021-00896-1

14 Zheng, J. *et al.* Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. *Nat. Genet.* **52**, 1122-1131 (2020). https://doi.org:10.1038/s41588-020-0682-6

15 Popkin, B. M. *et al.* Individuals with obesity and COVID-19: A global perspective on the epidemiology and biological relationships. *Obes. Rev.* **21**, e13128 (2020). https://doi.org:10.1111/obr.13128

16 Stefan, N., Birkenfeld, A. L., Schulze, M. B. & Ludwig, D. S. Obesity and impaired metabolic health in patients with COVID-19. *Nat. Rev. Endocrinol.* **16**, 341-342 (2020). https://doi.org:10.1038/s41574-020-0364-6

17 Kuehn, B. M. More Severe Obesity Leads to More Severe COVID-19 in Study. *JAMA* **325**, 1603-1603 (2021). https://doi.org:10.1001/jama.2021.4853

18 Recalde, M. *et al.* Body Mass Index and Risk of COVID-19 Diagnosis, Hospitalization, and Death: A Cohort Study of 2 524 926 Catalans. *J. Clin. Endocrinol. Metab.* **106**, e5040-e5042 (2021). https://doi.org:10.1210/clinem/dgab546

19 O'Rourke, R. W. & Lumeng, C. N. Pathways to Severe COVID-19 for People with Obesity. *Obesity (Silver Spring)* **29**, 645-653 (2021). https://doi.org:10.1002/oby.23099

20 Bonaventura, A. *et al.* Endothelial dysfunction and immunothrombosis as key pathogenic mechanisms in COVID-19. *Nat. Rev. Immunol.* **21**, 319-329 (2021). https://doi.org:10.1038/s41577-021-00536-9

21 Sattar, N., McInnes, I. B. & McMurray, J. J. V. Obesity Is a Risk Factor for Severe COVID-19 Infection. *Circulation* **142**, 4-6 (2020). https://doi.org:10.1161/circulationaha.120.047659

22 Korakas, E. *et al.* Obesity and COVID-19: immune and metabolic derangement as a possible link to adverse clinical outcomes. *Am. J .Physiol. Endocrinol. Metab.* **319**, E105-E109 (2020). https://doi.org:10.1152/ajpendo.00198.2020

23 Florindo, H. F. *et al.* Immune-mediated approaches against COVID-19. *Nat. Nanotechnol.* **15**, 630-645 (2020). https://doi.org:10.1038/s41565-020-0732-3

24 Boutari, C. & Mantzoros, C. S. A 2022 update on the epidemiology of obesity and a call to action: as its twin COVID-19 pandemic appears to be receding, the obesity and dysmetabolism pandemic continues to rage on. *Metabolism* **133**, 155217 (2022). https://doi.org:10.1016/j.metabol.2022.155217

25 Collaborators, G. B. D. O. *et al.* Health Effects of Overweight and Obesity in 195 Countries over 25 Years. *N. Engl. J. Med.* **377**, 13-27 (2017). https://doi.org:10.1056/NEJMoa1614362

26 Santos-Lozano, A. *et al.* Implications of obesity in exceptional longevity. *Ann. Transl. Med.* **4**, 416 (2016). https://doi.org:10.21037/atm.2016.10.35

27 Kivimäki, M. *et al.* Body-mass index and risk of obesity-related complex multimorbidity: an observational multicohort study. *Lancet Diabetes Endocrinol.* **10**, 253-263 (2022). https://doi.org:10.1016/s2213-8587(22)00033-x

28 Goossens, G. H. *et al.* Increased adipose tissue oxygen tension in obese compared with lean men is accompanied by insulin resistance, impaired adipose tissue capillarization, and inflammation. *Circulation* **124**, 67-76 (2011). https://doi.org:10.1161/CIRCULATIONAHA.111.027813

29 Cifarelli, V. *et al.* Decreased adipose tissue oxygenation associates with insulin resistance in individuals with obesity. *J. Clin. Invest.* **130**, 6688-6699 (2020). https://doi.org:10.1172/JCI141828

30 Matacchione, G. *et al.* Senescent macrophages in the human adipose tissue as a source of inflammaging. *Geroscience* **44**, 1941-1960 (2022). https://doi.org:10.1007/s11357-022-00536-0

31 Fruhbeck, G. *et al.* Increased Levels of Interleukin-36 in Obesity and Type 2 Diabetes Fuel Adipose Tissue Inflammation by Inducing Its Own Expression and Release by Adipocytes and Macrophages. *Front Immunol.* **13**, 832185 (2022). https://doi.org:10.3389/fimmu.2022.832185

32 Cinti, S. *et al.* Adipocyte death defines macrophage localization and function in adipose tissue of obese mice and humans. *J. Lipid Res.* **46**, 2347-2355 (2005). https://doi.org:10.1194/jlr.M500294-JLR200

33 Murano, I. *et al.* Dead adipocytes, detected as crown-like structures, are prevalent in visceral fat depots of genetically obese mice. *J. Lipid Res.* **49**, 1562-1568 (2008). https://doi.org:10.1194/jlr.M800019-JLR200

34 McKenney-Drake, M. L. *et al.* Epicardial Adipose Tissue Removal Potentiates Outward Remodeling and Arrests Coronary Atherogenesis. *Ann. Thorac. Surg.* **103**, 1622-1630 (2017). https://doi.org:10.1016/j.athoracsur.2016.11.034

35 Iacobellis, G. Epicardial adipose tissue in contemporary cardiology. *Nat. Rev. Cardiol.* **19**, 593-606 (2022). https://doi.org:10.1038/s41569-022-00679-9

36 Norris, T. *et al.* Duration of obesity exposure between ages 10 and 40 years and its relationship with cardiometabolic disease risk factors: A cohort study. *PLoS Med.* **17**, e1003387 (2020). https://doi.org:10.1371/journal.pmed.1003387

37 Hamer, M., Gale, C. R., Kivimaki, M. & Batty, G. D. Overweight, obesity, and risk of hospitalization for COVID-19: A community-based cohort study of adults in the United Kingdom. *Proc Natl Acad Sci U S A* **117**, 21011-21013 (2020). https://doi.org:10.1073/pnas.2011086117

38 Caussy, C. *et al.* Prevalence of obesity among adult inpatients with COVID-19 in France. *Lancet Diabetes Endocrinol* **8**, 562-564 (2020). https://doi.org:10.1016/S2213-8587(20)30160-1

39 Traversy, G. & Chaput, J. P. Alcohol Consumption and Obesity: An Update. *Curr Obes Rep* **4**, 122-130 (2015). https://doi.org:10.1007/s13679-014-0129-4

40 Iliodromiti, S. *et al.* The impact of confounding on the associations of different adiposity measures with the incidence of cardiovascular disease: a cohort study of 296 535 adults of white European descent. *Eur. Heart J.* **39**, 1514-1520 (2018). https://doi.org:10.1093/eurheartj/ehy057

41 Pearce, N. & Lawlor, D. A. Causal inference-so much more than statistics. *Int. J. Epidemiol.* **45**, 1895-1903 (2016). https://doi.org:10.1093/ije/dyw328

42 Sheehan, N. A., Didelez, V., Burton, P. R. & Tobin, M. D. Mendelian randomisation and causal inference in observational epidemiology. *PLoS Med* **5**, e177 (2008). https://doi.org:10.1371/journal.pmed.0050177

43 Fewell, Z., Davey Smith, G. & Sterne, J. A. The impact of residual and unmeasured confounding in epidemiologic studies: a simulation study. *Am. J. Epidemiol.* **166**, 646-655 (2007). https://doi.org:10.1093/aje/kwm165

44 Braveman, P. A. *et al.* Socioeconomic Status in Health ResearchOne Size Does Not Fit All. *JAMA* **294**, 2879-2888 (2005). https://doi.org:10.1001/jama.294.22.2879

45 Grundy, E. & Holt, G. The socioeconomic status of older adults: How should we measure it in studies of health inequalities? *J. Epidemiol. Community Health* **55**, 895-904 (2001). https://doi.org:10.1136/jech.55.12.895

46 McLaren, L. Socioeconomic status and obesity. *Epidemiol. Rev.* **29**, 29-48 (2007). https://doi.org:10.1093/epirev/mxm001

47 Hemani, G., Bowden, J. & Davey Smith, G. Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Hum. Mol. Genet.* **27**, R195-R208 (2018). https://doi.org:10.1093/hmg/ddy163

48 Ebrahim, S. & Davey Smith, G. Mendelian randomization: can genetic epidemiology help redress the failures of observational epidemiology? *Hum. Genet.* **123**, 15-33 (2008). https://doi.org:10.1007/s00439-007-0448-6

49 Loos, R. J. F. & Yeo, G. S. H. The genetics of obesity: from discovery to biology. *Nat. Rev. Genet.* (2021). https://doi.org:10.1038/s41576-021-00414-z

50 Frayling, T. M. *et al.* A Common Variant in the FTO Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity. *Science* **316**, 889-894 (2007). https://doi.org:doi:10.1126/science.1141634

51 Scuteri, A. *et al.* Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. *PLoS Genet.* **3**, e115 (2007). https://doi.org:10.1371/journal.pgen.0030115

52 Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. *Human Mol. Genet.* **27**, 3641-3649 (2018). https://doi.org:10.1093/hmg/ddy271

53 Valenzuela, P. L. *et al.* Obesity and the risk of cardiometabolic diseases. *Nat. Rev. Cardiol.* **20**, 475-494 (2023). https://doi.org:10.1038/s41569-023-00847-5

54 Cano-Gamez, E. & Trynka, G. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Front. Genet.* **11**, 424 (2020). https://doi.org:10.3389/fgene.2020.00424

55 Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73-79 (2018). https://doi.org:10.1038/s41586-018-0175-2

56 Ferkingstad, E. *et al.* Large-scale integration of the plasma proteome with genetics and disease. *Nat. Genet.* **53**, 1712-1721 (2021). https://doi.org:10.1038/s41588-021-00978-w

57 Pietzner, M. *et al.* Mapping the proteo-genomic convergence of human diseases. *Science* **374**, eabj1541 (2021). https://doi.org:10.1126/science.abj1541

58 Skrivankova, V. W. *et al.* Strengthening the reporting of observational studies in epidemiology using mendelian randomisation (STROBE-MR): explanation and elaboration. *BMJ* **375**, n2233 (2021). https://doi.org:10.1136/bmj.n2233

59 Skrivankova, V. W. *et al.* Strengthening the Reporting of Observational Studies in Epidemiology Using Mendelian Randomization. *JAMA* **326**, 1614 (2021). https://doi.org:10.1001/jama.2021.18236

60 Burgess, S. *et al.* Guidelines for performing Mendelian randomization investigations: update for summer 2023. *Wellcome Open Res.* **4**, 186 (2019). https://doi.org:10.12688/wellcomeopenres.15555.3

61 Davies, N. M., Holmes, M. V. & Davey Smith, G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *BMJ* **362**, k601 (2018). https://doi.org:10.1136/bmj.k601

62 Yao, C. *et al.* Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nat. Commun.* **9**, 3268 (2018). https://doi.org:10.1038/s41467-018-05512-x

63 Png, G. *et al.* Mapping the serum proteome to neurological diseases using whole genome sequencing. *Nat Commun* **12**, 7042 (2021). https://doi.org:10.1038/s41467-021-27387-1

64 Zhao, J. H. *et al.* Genetics of circulating inflammatory proteins identifies drivers of immune-mediated disease risk and therapeutic targets. *Nat. Immunol.* (2023). https://doi.org:10.1038/s41590-023-01588-w

65 Sun, B. B. *et al.* Genetic regulation of the human plasma proteome in 54,306 UK Biobank participants. *bioRxiv*, 2022.2006.2017.496443 (2022). https://doi.org:10.1101/2022.06.17.496443

66 Fauman, E. B. & Hyde, C. An optimal variant to gene distance window derived from an empirical definition of cis and trans protein QTLs. *BMC Bioinformatics* **23**, 169 (2022). https://doi.org:10.1186/s12859-022-04706-x

67 Karim, M. A. *et al.* Systematic disease-agnostic identification of therapeutically actionable targets using the genetics of human plasma proteins. *medRxiv*, 2023.2006.2001.23290252 (2023). https://doi.org:10.1101/2023.06.01.23290252

68 Hingorani, A. D. *et al.* Improving the odds of drug development success through human genomics: modelling study. *Sci. Rep.* **9**, 18911 (2019). https://doi.org:10.1038/s41598-019-54849-w

69 Suhre, K. Genetic associations with ratios between protein levels detect new pQTLs and reveal protein-protein interactions. *bioRxiv*, 2023.2007.2019.549734 (2023). https://doi.org:10.1101/2023.07.19.549734

70 Zhang, J. *et al.* Plasma proteome analyses in individuals of European and African ancestry identify cis-pQTLs and models for proteome-wide association studies. *Nat. Genet.* (2022). https://doi.org:10.1038/s41588-022-01051-w

71 Emilsson, V. *et al.* Co-regulatory networks of human serum proteins link genetics to disease. *Science* **361**, 769-773 (2018).

72 Aebersold, R. *et al.* How many human proteoforms are there? *Nat. Chem. Biol.* **14**, 206-214 (2018). https://doi.org:10.1038/nchembio.2576

73 Pietzner, M. *et al.* Synergistic insights into human health from aptamer- and antibody-based proteomic profiling. *Nat. Commun.* **12**, 6822 (2021). https://doi.org:10.1038/s41467-021-27164-0

74 Petrera, A. *et al.* Multiplatform Approach for Plasma Proteomics: Complementarity of Olink Proximity Extension Assay Technology to Mass Spectrometry-Based Protein Profiling. *J. Proteome Res.* **20**, 751-762 (2021). https://doi.org:10.1021/acs.jproteome.0c00641

75 Koprulu, M. *et al.* Proteogenomic links to human metabolic diseases. *Nat. Metab.* (2023). https://doi.org:10.1038/s42255-023-00753-7

76 Katz, D. H. *et al.* Whole Genome Sequence Analysis of the Plasma Proteome in Black Adults Provides Novel Insights Into Cardiovascular Disease. *Circulation* **145**, 357-370 (2022). https://doi.org:10.1161/CIRCULATIONAHA.121.055117

77 Xu, F. *et al.* Genome-wide genotype-serum proteome mapping provides insights into the cross-ancestry differences in cardiometabolic disease susceptibility. *Nat. Commun.* **14**, 896 (2023). https://doi.org:10.1038/s41467-023-36491-3

78 Fatumo, S. *et al.* A roadmap to increase diversity in genomic studies. *Nat. Med.* **28**, 243-250 (2022). https://doi.org:10.1038/s41591-021-01672-4

79 Abifadel, M. *et al.* Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. *Nature Genetics* **34**, 154-156 (2003). https://doi.org:10.1038/ng1161

80 Mayne, J. *et al.* Novel loss-of-function PCSK9 variant is associated with low plasma LDL cholesterol in a French-Canadian family and with impaired processing and secretion in cell culture. *Clin. Chem.* **57**, 1415-1423 (2011). https://doi.org:10.1373/clinchem.2011.165191

81 Kurki, M. I. *et al.* FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508-518 (2023). https://doi.org:10.1038/s41586-022-05473-8

82 Tremblay, K. *et al.* The Biobanque quebecoise de la COVID-19 (BQC19)-A cohort to prospectively study the clinical and biological determinants of COVID-19 clinical trajectories. *PLoS One* **16**, e0245031 (2021). https://doi.org:10.1371/journal.pone.0245031

83 Heyer, E. & Tremblay, M. Variability of the genetic contribution of Quebec population founders associated to some deleterious genes. *Am. J. Hum. Genet.* **56**, 970-978 (1995).

84 Hobbs, H. H., Brown, M. S., Russell, D. W., Davignon, J. & Goldstein, J. L. Deletion in the Gene for the Low-Density-Lipoprotein Receptor in a Majority of French Canadians with Familial Hypercholesterolemia. *N. Engl. J. Med.* **317**, 734-737 (1987). https://doi.org:10.1056/nejm198709173171204

85 Do, R. *et al.* Genetic Variants of FTO Influence Adiposity, Insulin Sensitivity, Leptin Levels, and Resting Metabolic Rate in the Quebec Family Study. *Diabetes* **57**, 1147-1150 (2008). https://doi.org:10.2337/db07-1267

Appendices

Copyright permissions

Ethical approval

For Chapters 3–5, all participants provided informed consent for the corresponding studies. All cohort studies were approved by the institutional review boards of the participating institutions.

Summary of the author's significant scientific contributions

Only manuscripts published during the PhD program are listed.

Publications (peer-reviewed, first author or co-first author)

***denotes co-first author and †denotes co-corresponding author**

1. Yoshiji, S., Butler-Laporte, G., Lu, T., Willett, JDS., Su, C-Y., Nakanishi, T., Morrison, DR., Chen, Y., Liang, K., Hultström, M., Ilboudo Y., Afrasiabi, Z., Lan, S., Duggan, N., DeLuca, C., Vaezi, M., Tselios, C., Xue, X., Bouab, M., Shi, F., Laurent, L., Münter, HM., Afilalo, M., Afilalo, J., Mooser, V., Timpson, NJ., Zeberg, H., Zhou, S., Forgetta, V., Farjoun, Y., and Richards, J.B. Proteome-wide Mendelian randomization implicates nephronectin as an actionable mediator of the effect of obesity on COVID-19 severity. *Nat. Metab.* **5**, 248-264 (2023). https://doi.org:10.1038/s42255-023-00742-w

   This is the published manuscript corresponding to Chapter 3.

2. Hasebe, M*., Yoshiji, S.*†, Keidai, Y.*, Minamino, H., Murakami, T., Tanaka, D., Fujita, Y., Harada, N., Hamasaki, A., Inagaki, N.† Efficacy of antihyperglycemic therapies on cardiovascular and heart failure outcomes: an updated meta-analysis and meta-regression analysis of 35 randomized cardiovascular outcome trials. *Cardiovasc. Diabetol.* **22**, 62 (2023). https://doi.org:10.1186/s12933-023-01773-z

   The author co-supervised the study.

3. Keidai, Y., Yoshiji, S.†, Hasebe, M., Minamino, H., Murakami, T., Tanaka, D., Fujita, Y., Inagaki, N.†. Stabilization of kidney function and reduction in heart failure events with SGLT2 inhibitors: a meta-analysis and meta-regression analysis. *Diabetes, Obes. and Metab*. **25**, 2505–2513 **(**2023). https://doi.org/10.1111/dom.15122

   The author co-supervised the study.

4. Yoshiji, S., Tanaka, D., Minamino, H., Murakami, T., Fujita, Y., Richards, J.B., Inagaki, N. Causal associations between body fat accumulation and COVID-19 severity: A Mendelian randomization study. *Front. Endocrinol.* (2022). https://doi.org/10.3389/fendo.2022.899625

   This is the published manuscript corresponding to Chapter 2.

5. Yoshiji, S., Minamino, H., Tanaka, D., Yamane, S., Harada, N., Inagaki, N. Effects of

glucagon-like peptide-1 receptor agonists on cardiovascular and renal outcomes: A meta-analysis and meta-regression analysis. *Diabetes, Obes. and Metab*. **24**, 1029-1037 (2022). https://doi.org/10.1111/dom.14666

6. Yoshiji, S*., Hasebe, M*., Iwasaki, Y., Shibue, K., Keidai, Y., Seno, Y., Iwasaki, K., Honjo, S., Fujikawa, J., Hamasaki, A. Exploring a Suitable Marker of Glycemic Response to Dulaglutide in Patients with Type 2 Diabetes: A Retrospective Study. *Diabetes Ther*. **13**, 733–746 (2022) https://doi.org/10.1007/s13300-022-01231-1
7. Yoshiji, S., Horikawa, Y., Kubota, S., Enya, M., Iwasaki, Y., Keidai, Y., Aizawa-Abe, M., Iwasaki, K., Honjo, S., Hosomichi, K., Yabe, D., Hamasaki, A. First Japanese Family with *PDX1*-MODY (MODY4): A Novel *PDX1* Frameshift Mutation, Clinical Characteristics, and Implications. *J. Endocr. Soc*. 2021. bvab159. https://doi.org/10.1210/jendso/bvab159
8. Yoshiji, S*, Iwasaki, Y., Iwasaki, K., Honjo, S., Hirano, K., Ono, K., Yamazaki, Y., Sasano, H., Hamasaki, A. *Alu*-Mediated *MEN1* Gene Deletion and Loss of Heterozygosity in a Patient with Multiple Endocrine Neoplasia Type 1. *J. Endocr. Soc*. bvaa051 (2020) https://doi.org/10.1210/jendso/bvaa051

Publications (peer-reviewed, second author or more)

1. Chen, Y., Lu, T., Pettersson-Kymmer, U., Stewart, I., Butler-Laporte, G., Nakanishi, T., Cerani, Agustin, Liang, K., Yoshiji, S., Willet, J., Su, C-Y., Raina, P., Greenwood, C., Fajoun, Y., Forgetta, V., Langenberg, C., Zhou, S., Ohlsson, C., Richards, J.B. Genomic atlas of the plasma metabolome prioritizes metabolites implicated in human diseases. *Nat. Genet.* **55**, 44–53 (2023). https://doi.org/10.1038/s41588-022-01270-1
2. Lu, T., Nakanishi, T., Yoshiji, S., Butler-Laporte, G., Greenwood, C.M.T., Richards, J.B. Dose-dependent Association of Alcohol Consumption With Obesity and Type 2 Diabetes: Mendelian Randomization Analyses. *J. Clin. Endocrinol. Metab* (2023). dgad324, https://doi.org/10.1210/clinem/dgad324
3. Willett, JDS., Lu, T, Nakanishi, Y., Yoshiji, S., Butler-Laporte, G., Zhou, S., Farjoun, Y., Richards, J.B. Colocalization of expression transcripts with COVID-19 outcomes is rare across cell states, cell types and organs. *Hum. Genet.* (2023).

https://doi.org/10.1007/s00439-023-02590-w

4. Liang, K.Y.H., Farjoun, Y., Forgetta, V., Chen, Y., Yoshiji, S., Lu, T., Richards, J.B. Predicting ExWAS findings from GWAS data: a shorter path to causal genes. *Hum. Genet.* (2023). https://doi.org:10.1007/s00439-023-02548-y
5. Butler-Laporte, G., Farjoun, Y., Chen, Y., Hultstrom, M., Liang, K., Nakanishi, T., Su, C-Y., Yoshiji, S., Forgetta, V., Richards, J.B. Increasing serum iron levels and their role in the risk of infectious diseases: a Mendelian randomization approach. *Int. J. Epidemiol*. (2023). https://doi.org:10.1093/ije/dyad010
6. Hultstrom, M., Lipcsey, M., Morrison, D.R., Nakanishi, T., Butler-Laporte, G., Chen, Y., Yoshiji, S., Forgetta, V., Fajoun, Y., Wallin, E., Larsson, I., Larsson, A., Marton, A., Titze, J.M., Nihlen, S., Richards, J.B., Frithiof, R. Dehydration is associated with production of organic osmolytes and predicts physical long-term symptoms after COVID-19: a multicenter cohort study. *Crit. Care* 26, **322** (2022). https://doi.org/10.1186/s13054-022-04203-w
7. Tsukaguchi, R., Murakami, M., Yoshiji, S., Shide, K., Fujita, Y., Ogura, M., Inagaki, N. Year-long effects of COVID-19 restrictions on glycemic control and body composition in patients with glucose intolerance in Japan: A single-center retrospective study. *J. Diabetes Investig*. (2022). https://doi.org/10.1111/jdi.13893

Preprints

1. Yoshiji, S., Lu, T., Butler-Laporte, G., Carrasco-Zanini-Sanchez, J., Chen, Y., Liang, K., Willett, J.D.S., Su, C.-Y., Wang, S., Adra, D., Ilboudo, Y., Sasako, T., Forgetta, V., Farjoun, Y., Zeberg, H., Zhou, S., Hultström, M., Machiela, M., Wareham, N.J., Mooser, V., Timpson, N.J., Langenberg, C., Richards, J.B. COL6A3-derived endotrophin mediates the effect of obesity on coronary artery disease: an integrative proteogenomics analysis. *medRxiv* (2023). https://doi.org/10.1101/2023.04.19.23288706

   This is the manuscript corresponding to Chapter 5, which is under review at the time of thesis submission.
2. Marks, A., Butler-Laporte, G., Yoshiji, S., Lu, T., Morrison, D., Nakanishi, T., Chen, Y., Forgetta, V., Farjoun, J., Frithiof, R., Lipcsey, M., Zeberg, H., Richards, J.B., Hultstrom,

M. Aquaporin 3 modulates the risk of death conferred by dehydration in COVID-19. *Research Square* (2023). https://doi.org/10.21203/rs.3.rs-3011474/v1

3. Butler-Laporte, G., Farjoun, Y., Nakanishi, T., Chen, Y., Hultström, M., Lu, T., Yoshiji, S., Ilboudo, Y., Liang, K., Su, C-Y., Willett, J., Zhou, S., Forgetta, V., Taliun, D., Richards, J.B. HLA allele-calling using whole-exome sequencing identifies 129 novel associations in 11 autoimmune diseases: a multi-ancestry analysis in the UK Biobank. *medRxiv* 2023.01.15.23284570 (2023). https://doi.org/10.1101/2023.01.15.23284570
4. Ilboudo, Y., Yoshiji, S., Lu, T., Butler-Laporte, G., Zhou, S., Richards, J.B. Vitamin D and the preclinical Alzheimer cognitive composite cognition (PACC) score: a two-sample mendelian randomization study. *medRxiv* 2022.11.23.22282674 (2022) https://doi.org/10.1101/2022.11.23.22282674