University of Groningen

Observational causality from -omics

van der Graaf, Adriaan

DOI:
[10.33612/diss.198282903](https://doi.org/10.33612/diss.198282903)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2022

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):
van der Graaf, A. (2022). *Observational causality from -omics*. [Thesis fully internal (DIV), University of Groningen]. University of Groningen. <https://doi.org/10.33612/diss.198282903>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Observational causality from -omics

Adriaan van der Graaf

university of
groningen

Observational causality from -omics

PhD thesis

to obtain the degree of PhD at the
University of Groningen
on the authority of the
Rector Magnificus Prof. C. Wijmenga
and in accordance with
the decision by the College of Deans.

This thesis will be defended in public on

Wednesday 19 January 2022 at 16.15 hours

by

Aart Adriaan Cornelis van der Graaf

born on 21 January 1991
in Groningen

Supervisors

Prof. L.H. Franke
Prof. C. Wijmenga

Co-supervisor

Dr. S. Sanna

Assessment Committee

Prof. G.A. Lunter
Prof. H. Snieder
Prof. P.M. Visscher

table of contents

---

part 1 genetic loci for causal inference 2

---

chapter 1 Introduction 3

chapter 2 Lack of Association Between Genetic Variants at ACE2 and TMPRSS2 Genes Involved in SARS-CoV-2 Infection and Human Quantitative Phenotypes 27

Introduction	27
Materials and Methods	29
Results	31
Discussion	36
References	40

chapter 3 ImmunoChip meta-analysis in European and Argentinian populations identifies two novel genetic loci associated with celiac disease 43

Introduction	43
Material and methods	44
Results	47
Discussion	52

---

part 2 causal inference of molecular traits 58

---

chapter 4 Large-scale association analyses identify host factors influencing human gut microbiome composition 61

Introduction	61
Results	62
Discussion	70
Methods	72
References	79

chapter 5 Effect of host genetics on the gut microbiome in 7,738 participants of the Dutch Microbiome Project 85

Introduction	85
Results	87
Discussion	96
Methods	99
References	107

chapter 6 Systematic prioritization of candidate genes in disease loci identifies TRAFD1 as a master regulator of IFN γ signalling in celiac disease 113

Introduction	113
Materials and methods	114
Results	121
Discussion	129
References	132

chapter 7 Mendelian randomization while jointly modeling *cis* genetics identifies causal relationships between gene expression and lipids 139

Introduction	139
Discussion	150
Methods	152
References	161

chapter 8 Discussion 167

English summary	193
Nederlandse samenvatting	196
List of Publications	200
Acknowledgements	202

part one

genetic loci for causal inference

Adriaan van der Graaf

chapter 1

Introduction

Causal relationships provide valuable insights into how individuals work on the inside. In the last decade, we have seen immense progress in the development of large observational cohort studies of many individuals that measure nearly everything: an individual's DNA and other molecular measurements as well as clinical phenotypes, through to eating habits. These studies have provided us with a wealth of information on the correlations between measured factors. Unfortunately, observational studies have not been able to identify causal relationships because correlations are often distorted by confounding. This thesis attempts to identify causal relationships from observational studies through the use of instrumental variable methods like Mendelian randomization (MR). In this introduction, I provide background on the state of the art of human molecular measurements. I then discuss how observational studies of these measurements are often confounded, and how MR is able to remove this confounding. I explain the theory behind MR, show how it works, and discuss how it fails. Then I will outline all the steps that we have taken to identify causal relationships across a wide range of (molecular) human phenotypes.

Broadly speaking, we understand all the basics of human biology. For example, we know that a biological cell is made up of a combination of water, salts, small molecules, proteins, lipids, glycans, nucleotides and trace elements (1). We also understand how these components are metabolized and synthesized, how cells multiply through mitosis and how a new organism is created through embryogenesis. A newly formed organism divides into multiple cells, differentiates and eventually turns into a full-grown adult human. These broad strokes are the molecular basis of our lives, but many details are still lacking.

Researchers around the globe study different aspects of the human organism. And although progress has been made, there are still many aspects and processes we don't understand. For instance, it is still largely unknown how the most commonly used painkiller, paracetamol, reduces fever and pain (2). Fortunately, we don't need a full understanding of what the active compound in paracetamol does in the human body to benefit from its use¹. We also don't know the causes of most human diseases, in particular the more chronic diseases for which the burden is enormous on society, the health care system and our economy (3).

Understanding the causes of disease allows us to more accurately treat it. Take the recently approved COVID-19 vaccines as an example (4,5). These vaccines have been designed, developed and validated incredibly quickly. In part because researchers were aware of the infection mechanisms discovered from other corona virus epidemics such as SARS and MERS (6). Without an accurate understanding of the mechanisms of coronavirus viral infection, it is unlikely that we would have had such

¹ Although this knowledge would be needed to identify the individuals who will not benefit from paracetamol or those who will develop side effects.

effective COVID-19 vaccines so soon after the discovery of the virus (7).

It may be worthwhile now to consider what has brought us to the stage where we know a fair amount about the inner workings of the human cell, and what steps still need to be taken to help us find causal relationships that are important to understand the human organism.

In the last decades, very large cohorts of many different individuals have been set up. These cohorts have been instrumental in identifying risk factors for disease. For instance, a relationship between cholesterol levels and heart disease was initially identified in the Framingham heart study (8). Nowadays, we have combined these cohorts with immense amounts of phenotypic measurements. Measurements which help us identify which phenotypes may be important for the development of human disease (9–11).

In parallel, there has been enormous progress into measuring molecular phenotypes in high throughput, which allows us to measure genetics, DNA methylation, mRNA expression, proteins and even our microbes with which we have a synergistic relationship. I call these factors “-omics” layers as they measure the genome, epigenome, transcriptome and microbiome of an individual. These -omics layers can also be measured in a cohort specific manner, allowing for the identification of molecular biomarkers through correlation analyses. This has been very successful, and many correlations between different human traits have been found. These correlations are very interesting, but it can be difficult to discern a cause from consequence based on correlations alone. For instance, there have been many observational associations between the molecular measurements of polyunsaturated fatty acids. An increase of Omega-3 fatty acid intake has been correlated to health-related outcomes like cancer and cardiovascular disease, however, after clinical trials were performed, the health effects of Omega-3 intake seem negligible or non-existent (12,13).

In this introduction I will discuss the state of the art for molecular measurements, and how they are used to identify risk factors for human disease. Then I will discuss why we cannot rely on correlations between these risk factors to identify causal relationships, and finally, I will discuss how we can use a technique called Mendelian randomization to identify causal relationships from observational traits. After this introduction to the field, I will state the aim of my PhD and outline how I have tried to achieve this aim through the chapters of the rest of this book.

New molecular measurement techniques are opening Pandora’s box of human biology

High-throughput molecular measurements using -omics techniques

The availability of more and better ways to measure the human cells has

Chapter 1

aided in our knowledge of the inner workings of ourselves. If we consider how much work has been done to determine the basic building blocks of human physiology, we are truly standing on the shoulders of giants. Presently, we can measure most cellular molecules relatively inexpensively with high throughput. For instance, the most common genetic variation² of the human genome is easily measurable using genotyping arrays for tens of euros per sample (14). For a little more money (a few hundred to a thousand euros), it is possible to measure the full genome as well as all of an individual's mRNA gene expression³ (transcriptome) and epigenetic marks (epigenome) through sequencing techniques (15–17). On top of that, we are in the midst of a single cell revolution, where we are able to identify the genetics, expression and epigenetics of individual cells (18,19). High-throughput sequencing can be applied to any organism, allowing us to also characterize the microbiome, the community of bacteria, virus and fungi and their metabolites that inhabit body sites, such as gut or skin, all of which could be involved in disease onset and progression (20).

The availability of all these layers of information has opened up the study of so called -omics data⁴. Nowadays, most large population studies or cohorts at least include DNA measurements of their participants next to a variety of directly measured phenotypes and metadata information recorded through questionnaires (9,10). Many cohorts go even further, developing so-called multi-omics⁵ cohorts, in which multiple data layers are measured, from the genome through mRNA expression to the microbiome (11,21,22).

In this time of high-throughput measurements of tens of thousands of molecular phenotypes across thousands to hundreds of thousands of individuals, it would seem that we're in a perfect position to understand everything there is to know about human biology.

Genome wide association studies map phenotype variation to locations on the DNA

Unfortunately, measuring all molecules in humans will not directly help us in understanding how human disease develops. Just measuring a molecule will not be enough, these molecules need to be put into the context of other human traits.

In the last two decades, one popular approach was to start with an easily measurable (quantitative) or clinically relevant phenotype, such as cholesterol levels, or a clinical outcome like a disease, and identify which markers on the DNA are likely to be involved in the modulation of that quantitative phenotype or the development of the disease. For example, if we're interested in a disease like celiac disease⁶, we could gather a large group of individuals diagnosed with celiac disease, and a large control group who have not been diagnosed with celiac disease, and measure their DNA (**Figure 1**). We can then compare the prevalence or frequency of each variant on the DNA between the group of cases and the group of controls. If some alleles are much more prevalent in the disease group compared to the control group, we could conclude that this variant is implicated in disease onset. Technically speaking, we say that this variant

² Genetic variation describes all the places on the DNA with individual variation

³ Gene expression is the measurement of how much a gene is transcribed from the DNA: How much is a specific gene 'turned on' or 'turned off'

⁴ -omics is the study of molecular marks. It's often used to describe a specific class of molecule that is under investigation

⁵ Multi-omics is the study of multiple layers of -omics data together

⁶ Sometimes also known as gluten allergy. Individuals who suffer from celiac disease experience a strong immune response in the small intestine upon ingestion of gluten. Celiac disease is an autoimmune disease

Figure 1
A simple GWAS

Genome-wide association studies determine the genetic locations on the DNA that are associated with a trait. In this fictional example, a case and control group are recruited, their DNA is measured, and the allele frequencies (or allele counts) between the two groups are compared.

Chapter 1

is associated with the disease (**Figure 1**) (23) .

This method is called a genome-wide association study (GWAS), and hundreds of diseases or complex phenotypes have been studied with this approach (24). The locations on the DNA that are associated with the disease are informative, because we know which genes and regulatory elements reside in these DNA locations. Integrating the knowledge of these proximal DNA features can help us to better understand the causes of a disease.

Interpreting GWAS can be difficult

GWAS studies have been able to identify thousands of locations on the DNA that are associated with human phenotypes (GWAS locus or quantitative trait loci (QTL)). In most of these loci, often many (tens to hundreds) genetic variants are associated with the phenotypes with strong statistical evidence and so identifying the mechanism that causes the association can be difficult. In the easiest cases, explaining the causal role of a DNA location is trivial: if the most-associated variant is a protein-changing sequence, this leads to the strong claim that this is the causal variant and the disease or the quantitative trait is caused by a specific protein isoform⁷. Interestingly, however, most GWAS loci are located in non-coding locations of the genome (23,25,26), meaning that they do not alter the final sequence of any proteins. Therefore, figuring out which (molecular) biological factor is causal to the disease requires further analyses or experiments. In the last decade there has been a strong push to try to understand the causal variants and genes that GWAS loci arise from, and how they transfer onto the final phenotype of interest.

Several approaches exist to interpret GWAS loci. One popular class of methods is gene-set analysis (27,28), whereby the genes or other DNA features that are close to the GWAS hits are analyzed in an enrichment analysis. These aggregate analysis techniques are very suitable for describing biological function but will not be able to pin-point causal mechanisms at a single locus. To really describe the function of a single GWAS locus. We have to move to other techniques.

GWAS studies can be combined to interpret GWAS loci

Luckily, we can use GWAS to understand GWAS. An interesting feature of a GWAS analysis is that it is largely agnostic to the phenotype of interest. There is only a single requirement: the phenotype of interest needs to be heritable⁸. Given this minimal requirement, GWAS have been done on almost everything, from human height, to questionnaire data, to psychic ability⁹ (29,30).

If GWAS results are available for two traits, we can compare the associated GWAS loci to see if they share the same causal genetic variants. This is called GWAS colocalization. Multiple methods exist to perform GWAS colocalization and they have been widely, and successfully used to identify which variants overlap on the genome (31,32).

Colocalization methods are particularly helpful when applied to -omics

⁷ A protein isoform is a specific version of a protein that is the result of a genetic difference

⁸ Heritability means that a certain phenotype runs in families. A more statistically sound explanation would be the amount of variance of a phenotype that can be explained by an individuals ancestry.

⁹ Which is really the same as filling out that you're psychic on a questionnaire.

Figure 2
Explanation of
colocalization methods

(A) Colocalization methods can be useful to identify causal hypothesis. (B) If many phenotypes colocalize in the same locus, it is difficult to interpret what the causal relationships could be. Any direction is, in principle, possible.

molecular measurements in conjunction with a quantitative phenotype or disease. We expect that some molecules are an intermediate causal step towards the disease (**Figure 2A**), which also means that the molecular trait should share the same causal variants with the disease. In this way colocalization methods can estimate causal relationships, although it is difficult to claim that one trait is causal for the other based solely on the sharing of a causal variant. For instance, both traits could share causal variants by chance (**Figure 2B**).

Interpreting GWAS remains difficult

Protein-coding variants, gene set analysis and colocalization approaches have been widely applied to analyze GWAS results. These analyses have been able to derive interesting hypotheses on the causal elements of some GWAS loci, but the majority of the causal mechanisms underlying these loci remain unexplained even after using these techniques.

It is striking that we have been so successful in figuring out which genetic locations are important for human traits, but have not been able to discern how these loci work. One potential explanation for why some GWAS loci remain unexplained is that we're not measuring the correct molecular markers or we're not measuring them in the correct context. For instance, when looking at lipid traits, the relevant tissue is more likely to be the liver than blood, because lipid metabolism mostly takes place in the liver (33). This can be an issue in difficult-to-access tissues like brain, where a biopsy can only be done post-mortem. Next to tissue accessibility, there is also the issue of when measurements are taken. With immune responses, tissues need to be stimulated before the immune response can be seen. Similarly, with phenotypes that develop and mature early, like human standing height, if we measure a molecular layer when growth has stopped, the dynamics of the phenotype are not taken into account in the analysis.

Another reason why GWAS loci remain unexplained is simply the sheer

Chapter 1

number of loci that have been found. Nearly every part of the human genome has been mapped to some trait or another. Single (non-protein-coding) genetic variants can be associated with tens or even hundreds of different other phenotypes (34,35). Clearly, applying colocalization methods on these hits will naturally produce results that are difficult to interpret, because there may be evidence for a shared causal variant associated with many phenotypes.

This is not to argue that GWAS has been an exercise in futility. GWAS analyses have led to profound insights into many different diseases, and have shown how phenotypically different traits relate to each other genetically and, by extension biologically. Furthermore, GWAS have paved the way toward -omics and high-throughput analyses of molecular systems. Still, GWAS remains an observational study, that is not designed to identify causal relationships on its own.

Causal relationships from experiments and observational studies

Experimentation is required to understand causal mechanisms in human biology

Strictly speaking, a GWAS cannot find a causal variant. Nor will colocalizing the GWAS loci with a molecular trait assure that we will find a true biological causal mechanism, given the issues described above. It is therefore crucial to implement systematic approaches that could be used to robustly identify causal relations and inspect several hypotheses in a limited amount of time.

Our current understanding of (human) biology is due mostly to experiments and long-term epidemiological studies (36,37). In these experiments, researchers aim to identify a causal relation between a factor and an outcome by comparing one group that has been exposed to the factor with a control group. In this scenario, the null hypothesis would state: if there is no causal relationship, there should be no difference in the outcome between the two groups. On the other hand, when there is a difference between the two groups the null hypothesis is rejected, and the difference can be attributed to the factor to which one of the groups has been exposed (**Figure 3**). An example of such an experiment is a study interested in the effect of a new drug proposed to reduce the chance of heart disease. The drug can be said to work if the untreated group has a higher prevalence of heart disease compared to the treated group (**Figure 3**).

We call these types of experiments ‘treatment - control’ experiments, and they are designed for the identification of so-called causal relationships (**Figure 3**). Causal relationships are extremely informative because they reveal how one factor affects another, for instance the effect of a new cholesterol-lowering medication on the prevalence of heart disease.

These treatment-control experiments don’t find all the causes that define

Figure 3

An example randomized control trial

A group of individuals (top) is randomly assigned into two groups. The treatment group receives a treatment as an exposure (right), whereas the control group does not or receives a placebo treatment (left). At the end of the trial, the outcome is compared between the groups. In this case there are different occurrences of the disease between the two groups

Chapter 1

a biological trait, rather they identify contributory causes. Evidence found in experiments does not provide proof that the new drug will always work against the disease (38). An experiment can only show that the chance of heart disease onset is smaller when the drug is taken, compared to when the drug is not taken. Likewise, contributory causes do not exclude other factors from also causing a reduction or increasing the prevalence of heart disease. You remain at higher risk for heart disease if you are, for instance, an avid smoker.

Causal relationships are critical for discovering potential treatments. If a relationship is causal, say LDL-cholesterol levels are causal for heart disease, the perturbation of LDL cholesterol will also have an effect on heart disease prevalence. For this reason, causal relationships are important in the identification of potential therapeutic targets, and thus for improving patient care and disease prevention. A causal relationship also automatically shows how one factor influences another, providing essential information for the identification of biological mechanisms and pathways (38). Pathway information can be used to understand human biology, but, perhaps more interestingly, broad integration of all known pathways will provide further understanding into disease course and therapeutic approaches for other diseases.

It is infeasible to find causal relationships between all combinations of traits

Treatment-control experiments perturb potential risk factors, then assess if an outcome is also changed. We could ask ourselves, how many experiments would we need to carry out to fully understand the causal factors to human disease?

If we stand in the shoes of a pharmaceutical company, perhaps we would be interested in treating a disease with a drug. Not knowing where to start, we could do everything in a high-throughput fashion. Let's design compounds for all the human protein-coding genes and perform experiments to see if the drug would work. This would mean that the company has to design a compound that targets each of the 20,000 proteins in the human genome and subsequently run clinical trials for each of them to find all the drugs that may provide clinical benefit for a single disease (39). This is, of course, impossible in terms of money and time, as a single clinical trial usually costs a million US dollars and can run upwards of a billion US dollars (39,40).

Unfortunately, even considering all 20,000 protein coding genes (and thus 20,000 clinical trials) for a single disease would not be at all representative of all human molecular factors that may be important for disease. It would take 20,000 trials if we're only interested in proteins and a single disease, but there are other biological factors that we may want to target with a drug. And if we're interested in another disease, we would have to repeat all these trials again.

It is therefore crucial to find a way to prioritize which targets to perform experiments on, or alternatively, to identify candidate causal relationships

from the results of observational and treatment-control experiments
Epidemiological studies help us choose the right experimental targets

Of course, we already discussed how to pick the right target to perturb, we can use GWAS! Indeed, GWAS have been helpful in picking the right pharmaceutical drug targets (41). Still, even with the help of GWAS results, identification of which target to perturb is non-trivial because we don't know the underlying biological mechanism. Even today, very costly clinical trials fail due to efficacy. Sometimes, the drug of interest can target the wrong biological factor, as has been shown in the failure of HDL-cholesterol treatments for heart disease.

One major observation in observational studies was that individuals who suffer from coronary artery disease have high LDL cholesterol levels, and low HDL cholesterol levels (8,42). Initially, this prompted pharmaceutical companies to create medication to reduce LDL cholesterol levels in patients. These trials were very successful in proving clinical benefit, and the resulting statin drugs are widely used to reduce the risk for heart disease (43).

Figure 4
Correlations between ice cream consumption and drowning, illustrating confounding.

(A) In this fictitious example, we show a correlation between ice cream consumption and rates of drowning. (B) This correlation indicates that there could be a causal relationship between ice-cream consumption and drowning. (C) However, there is no causal relationship between ice cream and drowning, but there is a single cause for both, namely how warm the weather is. The warm weather is the uncontrolled confounder in (A) and (B).

Chapter 1

The success of treating LDL cholesterol motivated pharmaceutical companies to create medication that increases HDL cholesterol levels, as an increase of HDL cholesterol has been associated to heart disease in similar epidemiological studies. Unfortunately, drugs that targeted HDL cholesterol level increases failed to reduce the risk for coronary artery disease (44). The trials that have shown the beneficial effect of LDL-cholesterol on heart disease were some of the largest successes in the pharmaceutical industry. While the trials testing the effect of HDL cholesterol were some of the most high-profile failures of the pharmaceutical industry. Why were observational studies not able to differentiate between causal effects of LDL cholesterol and HDL cholesterol changes?

Observational studies are plagued by confounders, limiting their use in causal inference

If we zoom out a little, we can ask ourselves the question: Why can we find causal relationships using experiments and why can't we do so with an observational study? After all, observational studies were the first to identify causal relationships like 'smoking causes cancer' and 'LDL cholesterol causes heart disease' (8).

The reason why observational studies are not always able to identify causal relationships, is because of what we call a confounder (45). Confounders influence both variables that are under analysis, causing a spurious correlation with your variables of interest¹⁰. It is important to note confounders are not necessarily limited to observational studies, but good experimental design can control for them during treatment-control studies¹¹.

Let's consider a mock example with a confounder. If we plotted the amount of ice cream sales per day against the rate of drowning on any particular day, we could find a correlation that is reproducible across observational studies (**Figure 4A, 4B**). Claiming a causal relationship from this evidence would result in the conclusion that ice cream causes drowning. Time to ban the sale of ice cream!

Of course, ice cream consumption is not the cause of drowning (except for the rare ice cream-induced stomach cramp tragedies). The cause of both ice cream sales and drowning is that the weather and location influences how many people are swimming and consuming a refreshing ice cream. In this mock example, warm sunny weather is the confounder. If we correct for the effect of weather when considering ice cream sales and drowning, the correlation would disappear (**Figure 4C**).

In this example, it is obvious that the weather is a confounder, making it easy to control for in a statistical analysis. However, due to the complexity of human biology, it's not always obvious which factor is a confounder. Some common confounders could be an individual's sex or age and other intrinsic factors like body mass index (BMI), physical activity, education and smoking, but there is no guarantee that measuring a confounder is sufficient to ensure correct causal inference. Interestingly, even when the confounder is measured and known, causality could still be incorrectly

¹⁰ Confounders are sometimes also called 'confusors', as they can show a causal relationship that is not there

¹¹ Confounding can also occur in treatment control experiments if the groups are imbalanced with respect to the confounder. For instance if the control group contains more smokers.

claimed. Spurious correlation can occur through measurement error alone, indicating that claims to causality can still be incorrect even when all confounders are known and corrected for (46).

Very often, we don't know what all the confounders are in an analysis of causality, making it difficult to claim a causal relationship in an observational study. This is especially important when understanding human biology. As there are so many internal factors, many of them unmeasured and unstudied, it remains very difficult to anticipate all confounders.

Treatment-control experiments remove confounding through randomization

Treatment-control experiments are able to remove all confounders because of an important first step that happens before any observations can be made from the data. This is called a randomization step. An experiment on a group of individuals is first randomized into two or more groups that receive a different treatment. Only after this step is the outcome of interest observed. The randomization step ensures that any confounding that is present should be equally present between groups. We can then make causal conclusions about the data by comparing one group to another (Figure 3).

¹² This is an example of a negative result. In finding causal relationships in biology, negative results are very important as the risk of false positives is so high. Therefore it's important to understand how negative results come about

As an example, let's construct an experiment where we put our mock hypothesis that ice cream consumption increases drowning to the test¹². We recruit a sufficiently large population close to a water source and divide it up into two groups. The treatment group receives an ice cream voucher from us, and the placebo group swears not to eat ice cream for the whole day. At the end of the day, we check back with each group and assess how many people have drowned. After the comparison, and ignoring the morbid nature of this outcome, we would not find a causal relationship because the consumption of ice cream does not cause individuals to drown. Note that we did not consider the confounder - the weather - in this experiment. The confounder has been accounted for through the randomization step, as the randomization makes ice cream consumption independent of the weather as a confounder.

Treatment-control experiments are not limited to experiments on humans but are done almost daily in model systems and model organisms that compare one treatment to another. In all cases, the experimental setup removes confounding from the observations and will only evaluate the effect of the treatment of interest.

Re-creating experimental designs in observational studies

Up to this point, I've discussed that experimental studies are expensive, a burden on the subjects and sometimes not feasible. Furthermore, observational studies do not find causality. So, it seems as if we are stuck in an endless cycle of finding potential targets to perturb in observational studies and testing these targets in expensive experiments.

Fortunately, there is a way out of this vicious circle of observational

Chapter 1

association and costly experimentation: we can mimic an experimental setup in observational data so that causal relationships can be claimed.

Let's again consider our mock example where ice cream is associated with drowning. We could use some variables that affect ice cream consumption in the same way that we did in the experimental setup. For instance, we know that sugar tax rates influence how much ice cream is consumed. Looking across jurisdictions with and without sugar taxes, we can check if the introduction of the tax decreased ice cream consumption and then look at the rates of drowning across these jurisdictions. In this case, the sugar tax has had the same effect as the treatment in our hypothetical experiment: the tax influenced ice cream consumption, and we can look at the causal downstream effects. The tax increase is independent of our confounder, the weather, and so only looking at tax-dependent ice cream consumption would be independent of the weather, making it possible to find a causal relationship using tax-controlled ice cream consumption differences. This analysis would reveal that tax-dependent ice cream consumption is not a cause of drowning.

This is an example of a so-called instrumental variable (IV) analysis. IV analyses are widely used in econometrics to identify potential causal relationships. IV analysis tests for the causal effect of an exposure (ice cream consumption) towards an outcome (the rate of drowning) through

Figure 5 Principles of Mendelian randomization

(A) Mendelian randomization can be used in observational cohorts with DNA information available. In the simplest example, Mendelian randomization would separate individuals into carriers and non-carriers to identify how many cases of a specific disease are in both groups. (B) Mendelian randomization will only identify causality when none of the assumptions are violated: IV-I requires that the genetic variant chosen as IV has an effect on the exposure (Strong instruments), IV-II requires that the genetic variants are not associated with the confounder (Independence) and IV-III requires that the genetic variant only has an effect on the outcome through the exposure (exclusion restriction). (C) For more power, we can provide a risk score to each individual based on multiple genetic variants and then perform MR.

some random factor (the sugar tax increase). It is worth noting that IV analysis is a statistical technique that assumes certain conditions to be true. Usually, a good IV analysis is done when the random factor is truly random with respect to a confounder. When these conditions are met, it can identify robust causal relationships. When these conditions are only partially (or not at all) satisfied, IV analysis finds, at best, nothing and, at worst, incorrect claims to causality.

Mendelian Randomization finds causal relationships in observational data

Mendelian randomization uses the randomness of genotypes to identify causality

So perhaps we can apply IV analysis to biological datasets to find causality? What sources of randomness are present in biological data that we can use? It may be easy to see that a human's DNA is randomly allocated from their parents. There is no control on which parental alleles are passed to their progeny. Genetic variation is, in principle, random, and thus it could be used in IV analysis to identify causality. Moreover, DNA influences the majority of human traits (47–49) as well as molecular mechanisms, such as gene expression or protein levels (35,50–52). So, if we have an observational cohort with DNA marker data information available, we could use its randomness to identify causal relationships from observational data. This method is called Mendelian randomization (MR), named after Gregor Mendel, who first observed genetic inheritance patterns (53,54).

To give an example on how this would work, let's construct a relatively simple MR experiment where we're interested in the causal effect of the LDL-cholesterol biomarker as an exposure (or risk factor) on the occurrence of heart disease as the outcome (the result of interest).

We take an observational study where we know the LDL cholesterol concentrations of individuals and if they have had heart disease. First, we determine which genetic variants are associated with variations in LDL cholesterol levels. Detection of these variants is done using GWAS. After we select a variant to use as an IV for our analysis, we separate the whole cohort into three groups of people depending on their genotype for the variant (i.e. homozygous for the reference allele, heterozygous and homozygous for the alternative allele). Then, we see if there are differences in heart disease occurrence between each group. If the experiment was performed correctly, a significant difference in heart disease occurrence between each group would indicate that there is a causal relationship (Figure 5A).

When is Mendelian randomization appropriate for causal inference?

If we want to rely on the inferences of MR, it is important to understand how and when MR fails. As with any statistical technique, there are

Chapter 1

assumptions that need to be met. The underlying theory behind MR (IV analysis) requires three assumptions (**Figure 5B**):

- IV-I:** The instrumental variable is robustly associated to the exposure variable (*Strong instrument*)
- IV-II:** The instrumental variable is not associated with any confounder (*Independence*)
- IV-III:** The instrumental variable is only associated to the outcome through the exposure (*Exclusion*)

I will approach these assumptions from a treatment-control experiment perspective and discuss how these assumptions may fail in the human genetics perspective to help understand cases when these assumptions fail, and how often this is likely to happen when genetic variants are used as IVs.

The IV-I assumption is violated when an IV does not have a sufficiently large effect on the exposure

The IV-I assumption requires that the instrumental variable is actually associated with changes in the exposure. If we consider the treatment-control experiment, we shouldn't be providing the treatment group too low a dose of our new treatment, i.e., there should be a sufficient difference between the treatment and control groups.

In the design of clinical trials, it is usually known what kind of effect the treatment has as these have been at least broadly tested in animal studies. This is not always the case in MR. The genetic variants we use are statistically associated with an exposure trait. Sometimes, these associations are not robust, or are even false positives. Using them as an instrument will result in invalid causal inference. When a genetic variant is selected as an instrument, MR will not question its validity. Fortunately, there are ways to reduce the bias from these weak instruments and these sources of bias can be eliminated by being very stringent about what variants to select for MR.

We can, for example, i) use very stringent cutoffs of association significance to select only robustly associated variants, ii) require that these associations were replicated in an independent cohort, and iii) potentially perform the selection of IV in a cohort independent from the one that will be used for MR, as confounding correlations with the outcome will be absent (55).

The IV-II assumption is violated through population stratification and selection in MR

The IV-II assumption requires that an IV is independent of any confounders. In other words the IV should be randomized with respect to any confounders. In the same way that we shouldn't put all the smokers in the treatment group in a treatment-control experiment, the IV-II assumption requires that the IVs are independent from any potential confounders (**Figure 5B**).

In experimental designs it is fairly obvious that incorrect randomization

would invalidate an experiment, but when and how do we violate the IV-I assumption when we use genetic variation as our IV? In general, the random segregation of alleles is ensured by evolutionary forces, and we assume this happens in human genetics (56). But we could think of selection processes that may influence the viability of individuals with certain variants, as we know that selection happens in the human genome (57,58). Fortunately, selection effects can be tested for in the genome, using, for instance, the Hardy-Weinberg equilibrium test (59). This makes it unlikely that selection or population stratification has large effects that will violate the assumptions of MR.

Next to selection processes, there is a chance that IV-II could be violated through a process called collider bias (60,61). Very often, we determine which variant to use in MR by using an already published GWAS study. These GWAS studies have pre-processed their phenotypes. If the phenotype was adjusted for a confounder in the analysis, this could make the resulting IV association also dependent on the confounder which violates IV-II. If, however, the phenotype is not corrected for covariates, collider bias is not an issue.

The IV-III assumption is violated when an IV affects multiple traits that are causal to the outcome

The IV-III assumption states that there should only be a single path from the IV to the exposure to the outcome (**Figure 5B**) a condition also known as the exclusion restriction and lack of (horizontal) pleiotropy assumption¹³. If there is any other path from the SNP to the outcome that does not pass through the exposure but rather through another phenotype, the IV-III assumption is violated.

¹³ In this thesis, we mostly refer to a violation of IV-III as the presence of pleiotropy.

In a treatment-control experiment, IV-III could be violated if the drug treatment has some sort of contamination in it in addition to the compound of interest. Thus, strictly speaking, any effect that is observed in the experiment cannot be purely attributable to the compound of interest since it could also be due to the contamination that is part of the treatment.

The IV-III assumption is perhaps the most difficult to account for when using genetic instruments. There is no control over the other effects of a specific IV for an exposure. It could thus be that the selected IV also has an effect on hundreds of other traits, which all could violate IV-III.

Ensuring that there are no known GWAS signals for other traits than the outcome on your IV of interest is important but not sufficient to control for IV-III because there could be other unobserved factors that are regulated by our genetic variant used as an IV. So, IV-III cannot really be ignored in an MR study, and it represents one of the major hurdles of this approach. Luckily, there are ways to reduce the effects of the IV-III when we apply MR on multiple variants.

Mendelian randomization is fairly easy to apply to genetic epidemiological studies

Chapter 1

Usually, the most difficult part of an IV analysis is identifying informative IVs. This is not necessarily the case for an MR analysis. Thanks to large-scale genomic studies, many genetic variants have been found that can be used for MR, across many different phenotypes. MR can use multiple GWAS-associated variants together to improve the causal estimate (**Figure 5C**). Meta-analysis techniques coupled with sensitivity analyses can be used to identify average causal effects across multiple IVs, to investigate if IV-III was violated by some of them and eventually to recalculate the causal relationship without these pleiotropic IVs (54,60,62–65).

Furthermore, MR can work in a ‘two-sample’ fashion, meaning that the exposure and outcome can be estimated in different datasets. This allows for causal inference across two independent observational studies. Moreover, it is not even necessary to perform MR on individual level data, as GWAS summary statistics can be used, allowing for causal inference without having access to individual-level data. Web tools that compile summary statistics and run MR are widely available (66,67).

Mendelian randomization has identified and disproved causal relationships. Easy applicability and the attractiveness of being able to claim causality has made MR a popular method for researchers in human biology and in the social sciences. Indeed, MR has been successfully applied to identify causal relationships in humans without having to perform costly experiments. MR confirmed the well-known causal relationship of LDL-cholesterol causes heart disease, and it was one of the first methods that disproved the causal relationship between HDL-cholesterol and heart disease (44). Another relationship that MR has disproved is the causal influence of vitamin D intake and the incidence of cancer (68,69).

These results showcase how MR has been able to identify causal relationships between easily observable complex traits. But, perhaps the main promise of the method is that there is no limitation to which traits can be assessed. MR is possible as long as there are variants that are suitable IVs. As sample sizes increase in large cohort-based GWAS and more traits are investigated, we are finding more and more genetic loci associated with all sorts of traits, with robust effects (to satisfy IV-I). The availability of more genetic variants to use together as IVs in a combined meta-analysis approach also make MR analyses more and more powerful.

Mendelian randomization on -omics phenotypes.

I find there is particular promise for performing MR in molecular phenotypes. We know that -omics phenotypes are difficult to interpret through association studies. Furthermore, treatment-control experiments that test the influence of one molecular factor on all other molecular factors are expensive and time-consuming. MR has the potential to identify the same causal effects as these experiments and is therefore an interesting candidate technique to understand how -omics phenotypes are related to one another as well as how -omics traits are related to complex human traits such as disease.

However, MR on -omics traits is not as straightforward as MR on complex traits. For instance, the number of available IVs is more limited in -omics

analysis, making it difficult to know if the IV follows all the assumptions underlying MR. There is also the issue of extensive quality control that needs to occur in -omics traits, which may be a cause of collider bias in MR.

That is not to say that it is impossible to perform MR on -omics molecular phenotypes and the potential payoff for finding robust causality is very large. For example, we have benefited tremendously from the knowledge of the metabolic pathways such as the Krebs cycle (70). Having similar network structures for -omics data layers could help us in understanding even more about the human biological system.

The main goal and outline of this thesis

I have outlined that identification of causal relationships is very important in biology. We can identify causal relationships by using treatment-control experiments and IV based methods such as MR. This brings me to the main goal of this PhD thesis:

To identify causal relationships between human -omics and complex traits

In this introduction I alluded to the complexity of the human organism, and more broadly, how we hardly know what causes human traits such as disease. For instance, even when we know a GWAS locus for a certain disease, it remains very difficult to understand how this position on the DNA translates into biological mechanisms leading to disease onset. Thus, the goal of this thesis is to better understand which molecular markers are causal for human traits. In this book, I look at mRNA expression, protein levels and the microbiome as -omics traits.

Fortunately, because we have the GWAS results for many traits as well as -omics data, it becomes possible to perform MR in a systematic manner. Finding variants that are robustly associated with traits is the first step to performing successful MR. The strong instrument assumption (IV-I) would be violated if we didn't have good GWAS hits. Therefore, this thesis is subdivided into two sections. In the first section, I describe finding of genetic mapping studies that I performed with my colleagues to identify variants that are associated with a wide variety of traits, including celiac disease and gut microbiome composition and function. In the second section, I show how we can integrate this QTL information into causal conclusions using existing MR methods and also proposing a novel statistical approach to apply the MR concept in the context of -omics data.

Section 1: Finding robustly associated genetic locations

Before we can prioritize or identify genes that are potentially causal to

Chapter 1

certain traits, it is important to identify which genetic loci are associated with a trait. In [chapter 2](#), I performed genetic variant mapping across a wide range of phenotypes including drug usage in two genes that are necessary for coronavirus SARS-COV-2 infection. This paper was written in the early days of the coronavirus disease (COVID-19) pandemic, when many things were unknown about the disease. Of particular interest at that time was the fact that the *ACE2* and *TMPRSS2* genes are necessary for COVID-19 infection. We wondered if there were other human phenotypes that could have an association signal on *ACE2* and *TMPRSS2*, which could potentially inform us about which individuals are at risk of severe COVID-19. Furthermore, *ACE2* is also the target of a commonly used antihypertensive drug, and there were worries that the usage of *ACE2* inhibitors could increase COVID-19 infection severity. In our work, we were unable to identify an association for our genes of interest in drug usage or other clinical phenotypes, making it less likely that COVID-19 infection severity was due to variants on these gene.

One important part of this thesis is finding causal genes for celiac disease. For this reason, my colleagues and I performed the largest GWAS meta-analysis of celiac disease in [chapter 3](#). Here, we built upon an existing GWAS dataset and included an extra 4,125 individuals from an Argentinian population with European ancestry. We find two novel GWAS loci that we use later in this book to identify the most likely causal genes for celiac disease ([chapter 6](#)). While my contribution to this paper was limited, I discuss this meta-analysis in this book because of the reliance on this data in [chapter 6](#).

Section 2: Causality using -omics data

All the chapters in the second section perform follow-up of GWAS analyses across a wide range of -omics traits. In [chapter 4](#) and [chapter 5](#), we performed GWAS analysis and causal follow-up on microbiome measurements. In [chapter 4](#), these microbiome measurements were performed using the 16S microbiome measurement technique in 18,340 individuals across 24 cohorts. In [chapter 5](#), the measurements were performed with a more sensitive microbiome sequencing technique in a single cohort of 7,738 individuals. In both cases, we find overlapping genetics and draw new conclusions on causal links using MR. Most of my contribution is in [chapter 5](#), while my contribution for [chapter 4](#) was limited.

In [chapter 6](#), I show how we analyzed the GWAS results from [chapter 3](#) to identify the most likely causal genes in the associated genetic loci. Combining the GWAS results with RNA sequencing (RNA-seq) expression data across 3,503 individuals, I used several techniques, including MR, to prioritize genes. In this work, we have experimentally validated one gene (TRAFD1) to be a master regulator of 41 other genes.

In [chapter 6](#), we found that, based on gene expression data, gene prioritization techniques are quite variable in nature and there is no gold standard statistical method. Furthermore, to reduce the chances of

violating the IV-III assumption, MR methods are usually only applicable when three or more IVs are available. In [chapter 7](#), I describe a new MR method, which we named MR-link, that is robust to IV-III even when there is a single genetic variant as IV available. We tested MR-link in simulations and on different RNA-seq datasets and a protein-level dataset and identified genes that are causally related to variation in LDL cholesterol levels.

Finally, in [chapter 8](#), I discuss the findings of this book, combined with their drawbacks and provide an overview of how I expect causal relationships between traits will be investigated in the future.

References

1. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. *Molecular Biology of the Cell*. Taylor & Francis Group; 2007. 1732 p.
2. Aghababian. *Essentials of Emergency Medicine*. Jones & Bartlett Publishers; 2010. 1092 p.
3. Murray CJL, Abbafati C, Abbas KM, Abbasi M, Abbasi-Kangevari M, Abd-Allah F, et al. Five insights from the Global Burden of Disease Study 2019. *The Lancet*. 2020 Oct 17;396(10258):1135–59.
4. CDC. Different COVID-19 Vaccines [Internet]. Centers for Disease Control and Prevention. 2021 [cited 2021 Apr 22]. Available from: <https://www.cdc.gov/coronavirus/2019-ncov/vaccines/different-vaccines.html>
5. GLANVILLE D. COVID-19 vaccines: authorised [Internet]. European Medicines Agency. 2021 [cited 2021 Apr 22]. Available from: <https://www.ema.europa.eu/en/human-regulatory/overview/public-health-threats/coronavirus-disease-covid-19/treatments-vaccines/vaccines-covid-19/covid-19-vaccines-authorised>
6. Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, et al. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell*. 2020 Apr 16;181(2):271–280.e8.
7. Dai L, Gao GF. Viral targets for vaccines against COVID-19. *Nat Rev Immunol*. 2021 Feb;21(2):73–82.
8. Mahmood SS, Levy D, Vasan RS, Wang TJ. The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *The Lancet*. 2014 Mar 15;383(9921):999–1008.
9. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Med*. 2015 Mar 31;12(3):e1001779.
10. Scholtens S, Smidt N, Swertz MA, Bakker SJL, Dotinga A, Vonk JM, et al. Cohort Profile: LifeLines, a three-generation cohort study and biobank. *Int J Epidemiol*. 2015 Aug;44(4):1172–80.
11. Tigchelaar EF, Zernakova A, Dekens JAM, Hermes G, Baranska A, Mujagic Z, et al. Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. *BMJ Open*. 2015 Aug 1;5(8):e006772.
12. Cederholm T, Palmblad J. Are omega-3 fatty acids options for prevention and treatment of cognitive decline and dementia? *Curr Opin Clin Nutr Metab Care*. 2010 Mar;13(2):150–5.
13. Sala-Vila A, Calder PC. Update on the Relationship of Fish Intake with Prostate, Breast, and Colorectal Cancers. *Crit Rev Food Sci Nutr*. 2011 Oct 1;51(9):855–71.
14. Suratannon N, van Wijck RTA, Broer L, Xue L, van Meurs JB, Barendregt BH, et al. Rapid Low-Cost Microarray-Based Genotyping for Genetic Screening in Primary Immunodeficiency. *Front Immunol* [Internet]. 2020 [cited 2021 May 3];11. Available from: <https://www.frontiersin.org/articles/10.3389/>

- fimmu.2020.00614/full
15. Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PAC, Monlong J, Rivas MA, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*. 2013 Sep;501(7468):506–11.
16. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods*. 2013 Dec;10(12):1213–8.
17. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. *Nature*. 2015 Oct;526(7571):68–74.
18. Bonder MJ, Smail C, Gloudemans MJ, Frisard L, Jakubosky D, D'Antonio M, et al. Identification of rare and common regulatory variants in pluripotent cells using population-scale transcriptomics. *Nat Genet*. 2021 Mar;53(3):313–21.
19. van der Wijst MGP, Brugge H, de Vries DH, Deelen P, Swertz MA, Franke L. Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat Genet*. 2018 Apr;50(4):493–7.
20. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*. 2010 Mar;464(7285):59–65.
21. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. *Genome Biol*. 2017 May 5;18(1):83.
22. Bonder MJ, Kurilshikov A, Tigchelaar EF, Mujagic Z, Imhann F, Vila AV, et al. The effect of host genetics on the gut microbiome. *Nat Genet*. 2016 Nov;48(11):1407–12.
23. Gallagher MD, Chen-Plotkin AS. The Post-GWAS Era: From Association to Function. *Am J Hum Genet*. 2018 May 3;102(5):717–30.
24. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res*. 2019 Jan 8;47(D1):D1005–12.
25. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. Linking disease associations with regulatory information in the human genome. *Genome Res*. 2012 Jan 9;22(9):1748–59.
26. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science*. 2012 Sep 7;337(6099):1190–5.
27. Leeuw CA de, Mooij JM, Heskes T, Posthuma D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLOS Comput Biol*. 2015 Apr 17;11(4):e1004219.
28. Pers TH, Karjalainen JM, Chan Y, Westra H-J, Wood AR, Yang J, et al. Biological interpretation of genome-wide association studies using predicted gene functions. *Nat Commun*. 2015 Jan 19;6(1):5890.
29. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*. 2015 Feb 12;518(7538):197–206.
30. Wabbeh H, Radin D, Yount G, Woodley of Menie MA, Sarraf MA, Karpuj MV. Genetics of psychic ability - A pilot case-control exome sequencing study. *EXPLORE [Internet]*. 2021 Mar 1 [cited 2021 May 6]; Available from: <https://www.sciencedirect.com/science/article/pii/S1550830721000501>
31. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLOS Genet*. 2014 May 15;10(5):e1004383.
32. Hormozdizari F, van de Bunt M, Segri AV, Li X, Joo JWJ, Bilow M, et al. Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am J Hum Genet*. 2016 Dec 1;99(6):1245–60.
33. Ongen H, Brown AA, Delaneau O, Panousis NI, Nica AC, Dermitzakis ET. Estimating the causal tissues for complex traits and diseases. *Nat Genet*. 2017 Dec;49(12):1676–83.
34. Morris JA, Daniloski Z, Domingo J, Barry T, Ziosi M, Glinos DA, et al. Discovery of target genes and pathways of blood trait loci using pooled CRISPR screens and single cell RNA sequencing. *bioRxiv*. 2021 Apr 8;2021.04.07.438882.
35. Vösa U, Claringbould A, Westra H-J, Bonder MJ, Deelen P, Zeng B, et al.

Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. *bioRxiv*. 2018 Oct 19;447367.

36. Montgomery DC. Design and Analysis of Experiments. John Wiley & Sons Incorporated; 2013. book.
37. Greenland S. Randomization, Statistics, and Causal Inference. *Epidemiology*. 1990;1(6):421–9.
38. Riegelman R. Contributory cause: Unnecessary and insufficient. *Postgrad Med*. 1979 Aug 1;66(2):177–9.
39. Sertkaya A, Wong H-H, Jessup A, Beleche T. Key cost drivers of pharmaceutical clinical trials in the United States. *Clin Trials*. 2016 Apr 1;13(2):117–26.
40. DiMasi JA, Hansen RW, Grabowski HG. The price of innovation: new estimates of drug development costs. *J Health Econ*. 2003 Mar 1;22(2):151–85.
41. Plenge RM, Scolnick EM, Altshuler D. Validating therapeutic targets through human genetics. *Nat Rev Drug Discov*. 2013 Aug;12(8):581–94.
42. Lloyd-Jones DM, O'Donnell CJ, D'Agostino RB, Massaro J, Silbershatz H, Wilson PWF. Applicability of Cholesterol-Lowering Primary Prevention Trials to a General Population: The Framingham Heart Study. *Arch Intern Med*. 2001 Apr 9;161(7):949.
43. Adhyaru BB, Jacobson TA. Safety and efficacy of statin therapy. *Nat Rev Cardiol*. 2018 Dec;15(12):757–69.
44. Rader DJ, Hovingh GK. HDL and cardiovascular disease. *The Lancet*. 2014 Aug 16;384(9943):618–25.
45. Greenland S, Pearl J, Robins JM. Confounding and Collapsibility in Causal Inference. *Stat Sci*. 1999 Feb;14(1):29–46.
46. Fewell Z, Davey Smith G, Sterne JAC. The Impact of Residual and Unmeasured Confounding in Epidemiologic Studies: A Simulation Study. *Am J Epidemiol*. 2007 Sep 15;166(6):646–55.
47. Shin S-Y, Fauman EB, Petersen A-K, Krumsiek J, Santos R, Huang J, et al. An atlas of genetic influences on human blood metabolites. *Nat Genet*. 2014 Jun;46(6):543–50.
48. Ge T, Chen C-Y, Neale BM, Sabuncu MR, Smoller JW. Phenome-wide heritability analysis of the UK Biobank. *PLOS Genet*. 2017 Apr 7;13(4):e1006711.
49. Teegene Balew Gizie S., Man Tengfei, van Roon Arie M., Asefa Nigus G., Riese Harriette, Nolte Ilja, et al. Heritability and the Genetic Correlation of Heart Rate Variability and Blood Pressure in >29000 Families. *Hypertension*. 2020 Oct 1;76(4):1256–62.
50. Sinnott-Armstrong N, Tanigawa Y, Amar D, Mars N, Benner C, Aguirre M, et al. Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nat Genet*. 2021 Feb;53(2):185–94.
51. Sun BB, Maranville JC, Peters JE, Stacey D, Staley JR, Blackshaw J, et al. Genomic atlas of the human plasma proteome. *Nature*. 2018 Jun;558(7708):73–9.
52. Consortium TGte. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*. 2020 Sep 11;369(6509):1318–30.
53. Katan M. Apolipoprotein E isoforms, serum cholesterol, and cancer. *The Lancet*. 1986 Mar 1;327(8479):507–8.
54. Smith GD. Mendelian Randomization for Strengthening Causal Inference in Observational Studies: Application to Gene × Environment Interactions. *Perspect Psychol Sci*. 2010 Sep 1;5(5):527–45.
55. Burgess S, Thompson SG, CRP CHD Genetics Collaboration. Avoiding bias from weak instruments in Mendelian randomization studies. *Int J Epidemiol*. 2011 Jun;40(3):755–64.
56. Edwards AWF. Natural Selection and the Sex Ratio: Fisher's Sources. *Am Nat*. 1998 Jun 1;151(6):564–9.
57. Bamshad M, Wooding SP. Signatures of natural selection in the human genome. *Nat Rev Genet*. 2003 Feb;4(2):99–110.
58. Beauchamp JP. Genetic evidence for natural selection in humans in the contemporary United States. *Proc Natl Acad Sci*. 2016 Jul 12;113(28):7774–9.
59. Weir BS. Genetic data analysis. Methods for discrete population genetic data. *Genet Data Anal Methods Discrete Popul Genet Data* [Internet]. 1990 [cited 2021 May 18]; Available from: <https://www.cabdirect.org/cabdirect/abstract/19900180990>
60. Burgess S, Thompson SG. Mendelian Randomization: Methods for Using Genetic

- Variants in Causal Estimation. Taylor & Francis; 2015. 224 p.
61. Barry C, Liu J, Richmond R, Rutter MK, Lawlor DA, Dudbridge F, et al. Exploiting collider bias to apply two-sample summary data Mendelian randomization methods to one-sample individual level data. medRxiv. 2020 Oct 23;2020.10.20.20216358.
62. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol*. 2015 Apr 1;44(2):512–25.
63. 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 May;50(5):693.
64. Burgess S, Thompson SG. Multivariable Mendelian Randomization: The Use of Pleiotropic Genetic Variants to Estimate Causal Effects. *Am J Epidemiol*. 2015 Feb 15;181(4):251–60.
65. Rees JMB, Wood AM, Burgess S. Extending the MR-Egger method for multivariable Mendelian randomization to correct for both measured and unmeasured pleiotropy. *Stat Med*. 36(29):4705–18.
66. 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. Loos R, editor. *eLife*. 2018 May 30;7:e34408.
67. Kamat MA, Blackshaw JA, Young R, Surendran P, Burgess S, Danesh J, et al. PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations. *Bioinforma Oxf Engl*. 2019 Nov 1;35(22):4851–3.
68. Ong J-S, Gharahkhani P, An J, Law MH, Whiteman DC, Neale RE, et al. Vitamin D and overall cancer risk and cancer mortality: a Mendelian randomization study. *Hum Mol Genet*. 2018 Dec 15;27(24):4315–22.
69. Chandler PD, Chen WY, Ajala ON, Hazra A, Cook N, Bubes V, et al. Effect of Vitamin D 3 Supplements on Development of Advanced Cancer: A Secondary Analysis of the VITAL Randomized Clinical Trial. *JAMA Netw Open*. 2020 Nov 18;3(11):e2025850.
70. Krebs HA, Johnson WA. Metabolism of ketonic acids in animal tissues. *Biochem J*. 1937 Apr 1;31(4):645–60.

Esteban A. Lopera-Maya†, Adriaan van der Graaft, Pauline Lanting, Marije van der Geest, Jingyuan Fu, Morris Swertz, Lude Franke, Cisca Wijmenga, Patrick Deelen, Alexandra Zhernakova, Serena Sanna,* and Lifelines Cohort Study

chapter 2

Lack of Association Between Genetic Variants at ACE2 and TMPRSS2 Genes Involved in SARS-CoV-2 Infection and Human Quantitative Phenotypes

Coronavirus disease 2019 (COVID-19) shows a wide variation in expression and severity of symptoms, from very mild or no symptoms, to flu-like symptoms, and in more severe cases, to pneumonia, acute respiratory distress syndrome, and even death. Large differences in outcome have also been observed between males and females. The causes for this variability are likely to be multifactorial, and to include genetics. The SARS-CoV-2 virus responsible for the infection depends on two human genes: the human receptor angiotensin converting enzyme 2 (*ACE2*) for cell invasion, and the serine protease *TMPRSS2* for S protein priming. Genetic variation in these two genes may thus modulate an individual's genetic predisposition to infection and virus clearance. While genetic data on COVID-19 patients is being gathered, we carried out a phenome-wide association scan (PheWAS) to investigate the role of these genes in other human phenotypes in the general population. We examined 178 quantitative phenotypes including cytokines and cardio-metabolic biomarkers, as well as usage of 58 medications in 36,339 volunteers from the Lifelines population cohort, in relation to 1,273 genetic variants located in or near *ACE2* and *TMPRSS2*. While none reached our threshold for significance, we observed several interesting suggestive associations. For example, single nucleotide polymorphisms (SNPs) near the *TMPRSS2* genes were associated with thrombocytes count ($p = 1.8 \times 10^{-5}$). SNPs within the *ACE2* gene were associated with (1) the use of angiotensin II receptor blockers (ARBs) combination therapies ($p = 5.7 \times 10^{-4}$), an association that is significantly stronger in females ($p_{\text{diff}} = 0.01$), and (2) with the use of non-steroid anti-inflammatory and antirheumatic products ($p = 5.5 \times 10^{-4}$). While these associations need to be confirmed in larger sample sizes, they suggest that these variants could play a role in diseases such as thrombocytopenia, hypertension, and chronic inflammation that are often observed in the more severe COVID-19 cases. Further investigation of these genetic variants in the context of COVID-19 is thus promising for better understanding of disease variability. Full results are available at <https://covid19research.nl>.

Introduction

The recent outbreak of the coronavirus disease 2019 (COVID-19) caused by the SARS-CoV-2 virus has quickly become a pandemic and poses a great threat to public health. COVID-19 has a wide range of clinical manifestations: infected people can be asymptomatic, symptomatic with mild respiratory symptoms, or have severe pneumonia (1–4). Estimates based on reported cases from February 2020 in China indicated that approximately 20% of patients develop severe respiratory illness requiring hospitalization, and that overall mortality estimates are around 2.3% (3). These estimates are not fixed and are becoming more precise as more

*Adapted from: Lopera-Maya, et al.
Lack of Association Between Genetic
Variants at ACE2 and TMPRSS2 Genes
Involved in SARS-CoV-2 Infection and
Human Quantitative Phenotypes.
Front. Genet. 11, (2020).*

†: equal contribution
*: Corresponding author

cases are reported, screened and analyzed. Interestingly, there is high variability in these estimates when comparing countries and continents, as well as differences in COVID-19 severity between males and females and between different age groups (1,2,5) [WHO Situation Report 70, from March 30, 2020]. Differences in response to SARS-CoV-2 infection between individuals and countries may be explained by diminished immune response in the elderly, comorbidities or smoking habits (6), but severe COVID-19 cases have also been observed in young individuals, seemingly without risk factors. This indicates that most factors explaining COVID-19 severity are still unknown. It is therefore critical to understand the mechanisms behind COVID-19 severity in order to provide appropriate prevention measures and adequate triage strategies, guide the drug discovery process and ultimately combat the SARS-CoV-2 pandemic.

The large variation in SARS-CoV-2 infection rates and COVID-19 severity could potentially be explained by genetic differences between hosts. While large-scale genetic studies of COVID-19 patients are being assembled, such as those coordinated by the COVID host genetics consortium (<https://www.covid19hg.com/>), it is worthwhile to evaluate the effects of genetic variants in genes involved in SARS-CoV-2 infection on human phenotypes, including quantitative traits, taking advantage of already existing cohorts. In fact, while quantitative phenotypes are not always directly associated with a disease, knowledge on the genetic variants that modulate these traits can improve our understanding of disease onset and the variability in symptoms. In one example of how this can work, genetic variants in the *BCL11A* gene were associated by genome-wide association studies (GWAS) to fetal hemoglobin (HbF) production in the general population (7), and these genetic variants were subsequently found to modulate the severity of beta-thalassemia and sickle cell diseases (8,9). This observation explained why certain individuals were naturally predisposed to mild symptoms of these diseases, while others had very severe clinical outcomes and benefitted from HbF increasing drugs. Therefore, understanding the role of genetic variants at genes essential for SARS-CoV-2 infection in human quantitative phenotypes is important to explain the observed variability in infection susceptibility and severity of COVID-19 and this understanding may suggest potential treatments.

Some factors that are necessary for SARS-CoV-2 infection are known (10,11). Angiotensin converting enzyme 2 (*ACE2*) is necessary for the invasion of the virus into the host cell through viral spike proteins, and the transmembrane Serine Protease 2 (*TMPRSS2*) is necessary for the correct maturation of these same viral spike proteins that enter the cell through *ACE2* (11). According to the GWAS Catalogue¹, genetic variants in or near *TMPRSS2*, located on chromosome 21, are associated with susceptibility of prostate cancer and mortality rate in the population, while no associations have been reported for variants in or near *ACE2*. This can be partly explained by the fact that the *ACE2* gene is located on the X chromosome, a part of the genome that is often not analysed by large scale genome wide association studies (GWAS) due to differences in analysis workflow with the autosomal chromosomes. Potential associations with human phenotypes near *ACE2* could have therefore been missed.

¹ GWAS catalogue:
<https://www.ebi.ac.uk/gwas/home>

Here we investigated the association of genetic variants within or near ($\pm 100\text{Kb}$) *ACE2* and *TMPRSS2* transcripts through a phenome-wide association scan (PheWAS) in 36,339 volunteers from the Lifelines population cohort. We analysed 72 quantitative phenotypes and the medication usage of 58 different drug categories in the entire cohort, and 92 protein levels in plasma and 14 cytokines in a subset of ~600 individuals. The quantitative phenotypes selected are anthropometric traits and measurable parameters of lung, hearth, kidney, haematological, immune and cardio-metabolic functions. Finally, in a sex-stratified analysis we evaluated whether these variants were sex-specific or differed in their association between males and females to explore potential differences between sexes that could modulate SARS-CoV-2 infection.

Materials and Methods

Lifelines cohort

The Lifelines cohort (12) is a multi-disciplinary prospective population-based cohort study, with a unique three generation design, that is examining the health and health-related behaviours of 167,729 individuals living in the North of the Netherlands. It was approved by the medical ethics committee of the University Medical Center Groningen and conducted in accordance with Helsinki Declaration Guidelines. All participants signed an informed consent form prior to enrollment. Lifelines employs a broad range of investigative procedures to assess the impact of biomedical, socio-demographic, behavioural, physical and psychological factors on multi-morbidity and complex genetics.

Genotyping data

A subset of 38,030 volunteers were genotyped using the Infinium Global Screening Array® (GSA) MultiEthnic Disease Version, according to manufacturer's instructions, at the Rotterdam genotyping center and the Department of Genetics, University Medical Center Groningen. We performed standard quality controls on both samples and markers, including removal of samples and variants with a low genotyping call rate ($< 99\%$), variants showing deviation from Hardy-Weinberg equilibrium ($p < 1 \times 10^{-6}$) or excess of Mendelian errors in families ($> 1\%$ of the parent-offspring pairs), and samples with very high or low heterozygosity. We further checked and removed samples that did not show consistent information between reported sex and genotypes on the X chromosome, between reported familial information and observed identity-by-descent sharing with family members, and between genotypes available from previous studies (13,14). A detailed description of the process can be found at the following link: <https://covid19research.nl> (15). After quality checks, a total of 36,339 samples and 571,420 autosomal and X-chromosome markers were available for analysis.

The genotyping dataset was then imputed using the Haplotype Reference Consortium (HRC) panel v1.1 at the Sanger imputation server² (16), and variants with an imputation quality score higher than 0.4 for variants

²Sanger imputation server:
<https://imputation.sanger.ac.uk>

with a MAF > 0.01 and higher than 0.8 for rare variants (MAF < 0.01) were retained. 58.40% of the 21,241 individuals whose genotype passed quality control were female, and the average age at phenotype collection was 39.9 years (± 16.3 years).

Phenotypes

Quantitative phenotypes were measured as previously described (11). We removed illegal zero or negative values for the 'QRS', 'QT', 'HALB', 'MAP', 'MOP', 'EOP', 'BAP', 'U24HVOL', 'ALT', 'HR', 'EO', 'PQ', 'MO' and 'BA' phenotypes, and removed -999 values from the electrocardiogram phenotypes 'P_AXIS', 'T_AXIS' and 'QRS_AXIS' (**Supplementary Table 1**). Protein levels in plasma for 92 cardiovascular-related proteins were determined using Olink Proseek Multiplex CVD III panel (OLINK, Uppsala, Sweden), and concentrations of plasma citrulline and cytokines were measured by ProcartaPlex^{TM} multiplex immunoassay (eBioscience, USA) as described previously (17,18). Medication use was recorded based on drug packaging brought in by the participant's on their first visit to the Lifelines inclusion center. Registration of medication use in this way has been shown to be fairly to highly concordant with health record information (19). After conversion to anatomical therapeutic chemical classification (ATC) codes, the first four letters (level 3) were used to define drug categories for association analyses. ATC codes with less than 100 observations were not considered for analysis, leaving 58 drug categories for analysis (**Supplementary Table 2**).

Statistical Analyses

We analyzed quantitative phenotypes using linear-mixed models implemented in SAIGEgds v1.0.0 so as to correct for familial relationships and cryptic population structure (20,21). For the X chromosome, genotypes in males were considered diploid. We tested the additive effect of 1,273 genetic variants within and near (± 100 Kb) *ACE2* (chrX:15,579,156-15,620,271, GRCh37) and *TMPRSS2* (chr21:42,836,478-42,903,043, GRCh37) transcripts. These are all single-nucleotide polymorphisms (SNPs) with minor allele frequency (MAF) > 0.005 that were genotyped or imputed and that passed our quality controls as described above. Analysis through SAIGEgds was carried out for 72 quantitative phenotypes available for all, or a subset of the 36,339 samples (**Table 1**). Drug categories were analysed as binary traits (1 = if medication currently in use, 0 otherwise) and restricted only to 1,240 genetic variants with MAF > 0.01. In both analyses age and sex were used as covariates. Inverse-normal transformation was applied to all quantitative traits prior to model fit. We searched for sex-specific effects by analyzing males and females separately (sex-stratified analyses), using only age as covariate and the same transformations as used for the analysis on the entire cohort. We also used the sex-stratified analysis results to investigate differential genetic effects between sexes at suggestive associations identified in the combined analysis. This approach is typically used in small to moderate studies as an alternative to an analysis with an interaction term (22)

The 92 circulating plasma proteins and 14 cytokines were measured in a

small subset of unrelated individuals and thus did not require correction for familial relationships. These were analysed using PLINK v2.00a3LM. We performed the association mapping with both sexes jointly, or separately as described above, and using inverse-normal transformation on the trait. We analyzed each variant and trait combination with or without the inclusion of age and sex covariates, as some genetic variants were too highly correlated due to the small sample size, and thus an estimate with covariates included in the model was not possible. To evaluate the statistical power of our study we used the package GeneticsDesign in R (23). For quantitative variable analyses, we used the function GeneticPower. Quantitative.Numeric() and calculated the minimum detectable additive effect (variance explained) with 80% power and at a significance threshold of 5×10^{-8} , for an increasing number of samples up to 36,339 (our study size). For binary variables analysis, we used the function GPC.default() and calculated the minimum detectable additive effect size (genotype relative risk) with 80% power and at a significance threshold of 5×10^{-8} , for an increasing number of cases in a cohort of 36,339 and for a risk allele frequency varying from 0.05 to 0.5. We set the number of cases up to 4000 to reflect the maximum number of users for the analyzed drug categories in our study. We also assumed that the causal variant was included in our genotyping data set, therefore we constrained full linkage disequilibrium ($D_{\text{prime}} = 1$) with the tag marker. Since disease prevalence (pD) could also impact power, we calculated the minimum detectable effect for pD varying from 1% to 20%.

Results

Quantitative phenotypes

Using a linear-mixed model, we analyzed 1,273 common and low frequency (minor allele frequency - MAF > 0.005) genetic variants in and near ($\pm 100\text{Kb}$) *ACE2* and *TMPRSS2* transcripts for association with 178 quantitative traits (**Supplementary Table 1**). None were found to be significant at the standard genome-wide level ($p = 5 \times 10^{-8}$) or at the Benjamin-Hochberg false discovery rate ($\text{FDR} < 0.1$). The most significant associations found with quantitative traits at the *ACE2* locus were with triglycerides (rs5980163, $p = 1.6 \times 10^{-4}$) and with the eosinophil counts (rs17264937, $p = 1.5 \times 10^{-4}$) (**Figure 1**). The strongest associations at the *TMPRSS2* locus were with plasma levels of CHIT1 (rs150965978, $p = 2.1 \times 10^{-6}$) and thrombocytes (rs28401567, $p = 1.7 \times 10^{-5}$) (**Table 1**) (**Figure 2**). Only the association at rs5980163 with triglycerides at *ACE2* showed a differential effect between males and females (Cochran Q test $p_{\text{diff}} = 0.01$), with most of the signal being attributable to males, although the association remains only suggestive ($p = 4.12 \times 10^{-4}$). We did not find any signal that was restricted to either males or females ($p > 1 \times 10^{-6}$ for all associations in the sex-stratified analyses). The SNP-trait associations reported in **Table 1** were not replicated in the UK Biobank, based on summary statistics from an analysis that included at least 343,992 samples (see URLs) (all $p > 0.05$). No replication was observed also for the association with CHIT1 plasma levels using results from the INTERVAL study (20).

Trait	Gene	rs.id	Chr:position	Ref/Alt	AF.Alt	Analysis	N	beta (SD)	p
EO	ACE2	rs17264937	X:15647332	T/C	0.312	All	35494	0.416 (0.10)	1.49x10 ⁻⁴
						Males only	14,751	0.357 (0.146)	0.015
						Females only	20,743	0.531 (0.159)	8.17x10 ⁻⁴
TGL	ACE2	rs5980163	X:15521666	C/G	0.016	All	36,112	0.071 (0.019)	1.63x10 ⁻⁴
						Males only	15,004	0.108 (0.031)	4.12x10 ⁻⁴
						Females only	21,108	0.014 (0.021)	0.488
CHIT1	TMPRSS2	rs150965978	21:42942652	C/A	0.063	All	526	-0.630 (0.131)	2.13x10 ⁻⁶
						Males only	241	-0.502 (0.209)	0.017
						Females only	285	-0.555 (0.175)	0.002
TR	TMPRSS2	rs28401567	21:42951813	C/T	0.166	All	36,049	-2.50 (5.83)	1.77x10 ⁻⁵
						Males only	14,975	-1.82 (0.788)	0.021
						Females only	21,074	-2.74 (0.775)	4.04x10 ⁻⁴

Table 1:
Most-significant associations
with phenotypes at *ACE2* and
***TMPRSS2* loci**

The table reports summary statistics for the two most-associated phenotypes and genetic variants at *ACE2* and *TMPRSS2* loci. Positions refer to genome build GRCh 37. Beta indicates the effect for each copy of the alternative allele, in standard deviation units. Ref: reference allele; Alt: alternative allele; AF.Alt: alternate allele frequency; SD: Standard Deviation; EO: Eosinophils; TGL: triglycerides; CHIT1: plasma levels of CHIT1 protein; TR: thrombocytes.

Medication use

For this analysis, we focused on 1,240 variants with minor allele frequency - MAF > 0.01. As with the quantitative phenotypes, none of the genetic variants showed genome-wide significant association with medication use (**Supplementary Table 2**). The strongest associations at the *ACE2* locus were observed for the group of drugs that contains non-steroid anti-inflammatory and antirheumatic products (NSAIDs) (ATC = M01A) (odds ratio (OR) = 1.34, 95% C.I. = 1.14-1.58, $p = 5.5 \times 10^{-4}$ for the G allele of rs4646190) (**Table 2**) (**Figure 1**), and for the group that contains angiotensin II receptor blockers (ARBs) in combination with other antihypertensive drugs (ATC = C09D) (OR = 1.35, 95% C.I. = 1.14-1.62 $p = 5.7 \times 10^{-4}$ for the T allele of rs4646156) (**Table 2**) (**Figure 2**). These SNPs are both located in intron 8 of the *ACE2* transcript and only 525bp apart, but they are not in linkage disequilibrium ($r^2 = 0.05$ in 1000 Genomes Europeans).

NSAIDs are used for treating pain, fever and inflammation, and include ibuprofen. The significance of rs4646190 was stronger in males ($p = 3.7 \times 10^{-4}$) than in females ($p = 0.08$), but the effect sizes were not statistically different ($p_{\text{diff}} = 0.054$).

The second group of drugs encodes for a combined therapy used to treat hypertension. Combination therapy of ARBs with other hypertensive drugs is usually initiated as a second option when the antihypertensive effect of an ARB alone is not sufficient (25,26). Our results indicate that individuals carrying at least one T allele at the rs4646156 polymorphism were more likely to take this combined therapy compared to individuals with the other allele. The effect of this SNP was also not significant when considering only ARBs intake (ATC = C09C, $p = 0.66$). Thus, the association with ARB combination therapies could indicate that individuals in whom it is difficult to manage hypertension may be genetically predisposed to this state by rs4646156. Interestingly, when analysing males and females

Drug ATC code	Gene	rs.id	Chr:position	Ref/Alt	AF.Alt	Analysis	N users	OR (95% C.I.)	p
M01A	ACE2	rs4646190	X:15597568	A/G	0.046	All	1274	1.34 (1.14-1.58)	5.5x10 ⁻⁴
						Males only	412	1.54 (1.22-1.97)	3.7x10 ⁻⁴
						Females only	862	1.16 (0.92-1.46)	0.213
						V			
C09D	ACE2	rs4646156	X:15597043	A/T	0.646	All	202	1.36 (1.14-1.62)	5.71x10 ⁻⁴
						Males only	88	1.14 (0.92-1.42)	0.231
						Females only	114	1.78 (1.35-2.34)	4.69x10 ⁻⁵
J02A	TMPRSS2	rs457274	21:42792485	C/G	0.406	All	523	1.33 (1.16-1.51)	3.36x10 ⁻⁴
						Males only	234	1.23 (1.01-1.50)	0.041
						Females only	289	1.41 (1.25-1.60)	1.46x10 ⁻⁴
D07A	TMPRSS2	rs9975623	21:42920296	A/G	0.304	All	897	1.23 (1.07-1.36)	1.04x10 ⁻⁴
						Males only	348	1.29 (1.09-1.51)	2.52x10 ⁻³
						Females only	549	1.19 (1.04-1.35)	0.011

separately, we found that the signal of rs4646156 on ARB combination therapy was mostly attributable to females, even when accounting for differences in number of users (OR = 1.78, 95% C.I.= 1.35-2.34 , p = 4.7x10⁻⁵ in females vs OR = 1.14, 95% C.I.= 0.92-1.42, p = 0.23 in males, p_{diff} = 0.01).

We reiterate that none of these associations (in the combined and in the sex-specific analyses) meet either the genome-wide or FDR thresholds for significance. To confirm these findings larger sample sizes are necessary. The strongest associations at the *TMPRSS2* locus were observed for the group of drugs containing antimycotics prescriptions (ATC = J02A) (p = 3.65x10⁻⁵) and for corticosteroids (ATC = D07A) (p = 1.0x10⁻⁴) (**Figure 2**). No significant difference in effect size between sexes was observed for these two associations (p_{diff} > 0.2). These SNPs were independent from each other and from the top associations with quantitative traits described in **Table 2**.

We attempted to validate our findings on medication use using the UK Biobank public GWAS summary statistics³, although their data refers to the use of individual medications rather than drug categories. When considering the medications most commonly used (> 1,000 users in the UK Biobank cohort) in the categories of interest (M01A, C09D, J02A and D07A), we found nominal association with the same direction of effects only for glucosamine use (ATC = M01A, p = 0.002 in the combined analysis and p = 0.008 in males only) and with candesartan_cilexetil (ATC = C09D, p = 0.008 in females only) (**Supplementary Table 3**). A similar detailed analysis in our cohort was underpowered to detect an association signal for single medications of C09D and M01A categories (**Supplementary Table 4**). This lack of replication could be attributable to differences in medication usage reporting between studies. While both are based on self-reported information, in the Lifelines study records

Table 2
Top associations with drug prescriptions at *ACE2* and *TMPRSS2* loci

The table reports summary statistics for the two most associated drug categories (given in ATC codes) for genetic variants at *ACE2* and *TMPRSS2* loci. Positions refer to genome build GRCh 37. The Odds Ratio refers to the alternative allele. Ref: reference allele; Alt: alternative allele; OR: Odds ratio; M01A: Anti-inflammatory and antirheumatic products, non-steroids; C09D: Angiotensin II receptor blockers (ARBs), combinations; J02A: Antimycotics for systemic use; D07A: Corticosteroids, plain.

³ UK biobank all phenotype associations: <http://www.nealelab.is/uk-biobank>.

Figure 1

Regional associations plot at the *ACE2* locus. Graphical representation of the association results at the *ACE2* locus for the SNP-trait associations reported in **Table 1 (A,B)** and **Table 2 (C,D)**. In each panel, each dot represents a genetic variant, and shown is the association strength (expressed as negative log₁₀ p-values, Y-axis) vs. the genomic position (on the hg19/GRCh37 genomic build, X-axis). The strongest associated variant is depicted with a purple diamond, while other variants are color-coded to reflect their linkage disequilibrium with it (taken from pairwise r^2 values calculated from the 1,000 Genomes Europeans). A legend for color-coding is provided in (A). In (D), an additional box shows the location of associations reported in the GWAS catalog (no associations were reported at this locus) and below this box the location of genes is shown with specification of exons and direction of transcription. This figure was drawn using LocusZoom web tool (31).

Figure 2

Regional associations plot at the *TMPRSS2* locus. Graphical representation of the association results at the *TMPRSS2* locus for the SNP-trait associations reported in **Table 1 (A,B)** and **Table 2 (C,D)**. In each panel, each dot represents a genetic variant and shown is the association strength (expressed as negative log₁₀ p-values, Y-axis) vs. the genomic position (on the hg19/GRCh37 genomic build, X-axis). The strongest associated variant is depicted with a purple diamond, while other variants are color-coded to reflect their linkage disequilibrium with it (taken from pairwise r^2 values calculated from the 1,000 Genomes Europeans). A legend for color-coding is provided in (A). In (D), an additional box shows the location of associations reported in the GWAS catalog (associations here reported from left to right are: melanoma, age-related diseases and mortality, and prostate cancer) and below this box the location of genes is shown with specification of exons and direction of transcription. This figure was drawn using LocusZoom web tool (31).

are confirmed by medication packaging collected by a nurse during the recruitment.

Discussion

Recent studies have demonstrated that SARS-CoV-2 uses *ACE2* as the key receptor to invade cells (11) and that *ACE2*-mediated cell invasion is enhanced by *TMPRSS2* expression (10). Genetic variations in these two genes that interfere with the gene function may thus be involved in the observed variability of SARS-CoV-2 susceptibility and COVID-19 severity. The association of these genetic variants with human phenotypes in the general population may suggest potential treatments and help to better identify at-risk individuals. Here we used a cohort of 36,339 individuals from the Lifelines general population cohort to investigate the impact of variants near and within these two genes on 178 quantitative traits including measurable parameters of lung, hearth, kidney, hematological, immune, and cardio-metabolic functions.

We found no significant evidence that common and low frequency variants in these loci were associated with the measured quantitative traits in the general population. We did observe suggestive signals for phenotypes (triglycerides and thrombocytes) that are involved in cardiovascular diseases, which are considered risk factors for COVID-19 diseases (3), but none of the genetic variants reached statistical significance despite our large sample size. Nevertheless, we cannot exclude a role of these variants in the regulation of COVID-19 severity through other relevant phenotypes such as specific immune cell types or cytokine levels that were not measured in our cohort.

To evaluate the effect of genetic variation in clinically relevant phenotypes, we investigated the association of genetics with medication use. We observed a marginal association of variants within *ACE2* with use of ARBs combination therapy (ATC = C09D, rs4646156) and with use of non-steroidal anti-inflammatory and antirheumatic drugs (NSAIDs, rs4646190). Interestingly, a marginal association with ARBs (C09C category) was also observed at the *TMPRSS2* locus (rs75833467, $p = 3.5 \times 10^{-4}$). These results are intriguing considering the current debate about whether the use of ARBs and NSAIDs could worsen COVID-19 severity (27-29), and their potential effect on increasing *ACE2* expression. No significant associations were found for these variants with blood pressure measurements or inflammatory markers in our cohort ($p < 0.05$), not even when the use of such drugs were added as covariates (data not shown). Association with diastolic and systolic blood pressure was also not observed in the large UK Biobank cohort. Thus, these variants are likely to be associated only with clinical conditions such as hypertension and chronic inflammation or with a better drug response. It has to be noted that our sample size allowed sufficient statistical power to detect genetic variants with small effects (down to 0.001 of variance explained), and thus we are confident in claiming lack of association at *TMPRSS2* and *ACE2* loci with the quantitative phenotypes assessed. For analyses on medication usage we were instead sufficiently powered to find small effects (genotype relative risk ~ 1.1) only for very common SNPs

(frequency > 0.3), but we are underpowered for smaller effects and, in general, at less common variants (**Figure S1**). Therefore, our suggestive results for medication usage could indicate a real effect for which we were underpowered to find a genome-wide significance evidence. Further exploration of these associations is needed.

ARBs are the preferred alternative for patients who experience ACE-inhibitor induced coughing. However, as rs4646156 is not associated with this adverse drug reaction (ADR), our results are likely independent of the switch to ARBs due to ACE-inhibitor induced coughing (30). Interestingly, the association of this SNP with ARBs was specific to ARBs combination therapy, thus pointing to individuals with difficult-to-manage hypertension. The major allele (T) of rs4646156 has different frequencies across populations: 0.653 in Europeans, 0.997 in East Asians and 0.797 South Asians, according to 1,000 Genomes. Likewise, the G allele at the rs4646190 SNP, associated with a higher probability of NSAIDs use, shows substantial different frequencies among populations. It is mostly absent in Asians but not in Europeans: 0.03 in Europeans, 0 in East Asians and 0.003 in South Asians, according to 1,000 Genomes.

The suggestive genetics associations we find for NSAIDs and ARBs combination therapy indicate that, depending on their genotype, certain individuals are predisposed to take these drugs, and thus to suffer from hypertension and chronic inflammation, diseases often described among COVID-19 comorbidities. This, together with the observed different allele frequencies across continents and the sex-related differential effects could explain the observed variation in COVID-19 severity between countries and sexes. Unfortunately, we could only speculate around this hypothesis as this study is not suited to prove that these genetic associations are directly related to SARS-CoV-2 susceptibility or COVID-19 severity, nor we can determine if ARBs or NSAIDs improve or worsen COVID-19 severity. A role of ARBs in worsening severity seems however unlikely (30-33).

We acknowledge the following limitations in our study. First, only age and sex were used as covariates in our analyses, which may not be sufficient to correct for confounders for all traits, such as drug usage or diseases, although the effect of these confounders should be mitigated by our sample size. Secondly, our analyses on medication use are underpowered given the limited number of individuals in the general population who use the medications that we tested, and thus none of the associations found here met the multiple-testing adjusted significance. Third, our results for medication use did not include low frequency and none of the analysis include rare variants (MAF < 0.005) which could still be relevant. Fourth, while we can speculate about potential connections of our results with current knowledge of COVID-19, longitudinal and well-characterized data on patients is needed to further explore our hypothesis.

In conclusion we carried out an extensive screening of potential genetic associations at common and low frequency variants in the *ACE2* and *TMPRSS2* genes, and found a lack of substantial effect in human quantitative phenotype variation in the general population. Genetic

analyses in more phenotypes are needed to evaluate their functional role in other physiological processes.

Finally, since genetic variation in other genes, for example those involved in regulating the immune system, could also be important in determining SARS-CoV-2 susceptibility and disease severity, large scale genetic initiatives like the COVID-19 host genetics consortium (The COVID-19 Host Genetics Initiative, 2020; <https://www.covid19hg.com/>) that directly involve patients with COVID-19 and deeply characterization of genomes and phenotypes are urgently needed.

Data availability statement

The data analyzed in this study was obtained from the Lifelines biobank, under project application number OV18_0463. Requests to access this dataset should be directed to Lifelines Research Office (research@lifelines.nl). Full summary statistics of the results are available at <https://covid19research.nl>.

Ethics statement

The studies involving human participants were reviewed and approved by the medical ethics committee of the University Medical Center Groningen and conducted in accordance with Helsinki Declaration Guidelines. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

Author Contributions

EL and AG performed statistical analyses. EL, AG, PL, PD, AZ, and SS interpreted results. MG and MS provided computing infrastructure and web portal. Lifelines Cohort Study, LF, CW, JF, and AZ provided access to the data. EL, AG, and SS wrote the manuscript draft with critical input from PL, LF, CW, PD, JF, and AZ. All authors read and approved the manuscript.

Lifelines cohort study - group authors genetics

Raul Aguirre-Gamboa (1), PD (1), LF (1), Jan A Kuivenhoven (2), EL (1), Ilja M Nolte (3), SS (1), Harold Snieder (3), MS (1), Judith M Vonk (3), CW(1)

(1) Department of Genetics, University of Groningen, University Medical Center Groningen, Netherlands.

(2) Department of Pediatrics, University of Groningen, University Medical Center Groningen, Netherlands.

(3) Department of Epidemiology, University of Groningen, University Medical Center Groningen, Netherlands.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. The reviewer JK declared a shared affiliation with no collaboration with one of the authors PD to the handling editor at time of review.

Acknowledgements

The authors wish to acknowledge the services of the Lifelines Cohort Study, the contributing research centers delivering data to Lifelines, and all the study participants. We also thank K. McIntyre for editorial assistance, and the UMCG Genomics Coordination center, the UG Center for Information Technology and their sponsors BBMRI-NL & TarGet for storage and computational infrastructure. This manuscript has been released as a pre-print at <https://www.medrxiv.org/>.

Funding

This work was supported by the Netherlands Organization for Scientific Research (NWO): NWO Spinoza Prize SPI 92-266 (to CW). The Lifelines Biobank initiative has been made possible by funding from the Dutch Ministry of Health, Welfare and Sport; the Dutch Ministry of Economic Affairs; the University Medical Center Groningen (UMCG the Netherlands); the University of Groningen and the Northern Provinces of the Netherlands. The generation and management of GWAS genotype data for the Lifelines Cohort Study was supported by the UMCG Genetics Lifelines Initiative (UGLI). JF was supported by NWO Gravitation Netherlands Organ-on-Chip Initiative (024.003.001) and the Netherlands Heart Foundation CVON grant 2018-27. MS was supported by EOSC-Life, funded by European Union's Horizon 2020 programme under grant agreement 824087 and NWO-VIDI grant VIDI 917.164.455. AZ was supported by ERC Starting Grant 715772, NWO-VIDI grant 016.178.056, the Netherlands Heart Foundation CVON grant 2018-27, and NWO Gravitation grant ExposomeNL (024.004.017).

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2020.00613/full#supplementary-material>

References

1. Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *The Lancet*. 2020 Feb 15;395(10223):507–13.
2. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*. 2020 Feb 15;395(10223):497–506.
3. Wu Z, McGoogan JM. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention. *JAMA* [Internet]. 2020 Feb 24 [cited 2020 Apr 6]; Available from: <https://jamanetwork.com/journals/jama/fullarticle/2762130>
4. Xu X-W, Wu X-X, Jiang X-G, Xu K-J, Ying L-J, Ma C-L, et al. Clinical findings in a group of patients infected with the 2019 novel coronavirus (SARS-Cov-2) outside of Wuhan, China: retrospective case series. *BMJ* [Internet]. 2020 Feb 19 [cited 2020 Apr 6];368. Available from: <https://www.bmj.com/content/368/bmj.m606>
5. Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *The Lancet*. 2020 Mar 28;395(10229):1054–62.
6. Guan W, Liang W, Zhao Y, Liang H, Chen Z, Li Y, et al. Comorbidity and its impact on 1590 patients with Covid-19 in China: A Nationwide Analysis. *Eur Respir J* [Internet]. 2020 Jan 1 [cited 2020 Apr 3]; Available from: <https://erj.ersjournals.com/content/early/2020/03/17/13993003.00547-2020>
7. Menzel S, Garner C, Gut I, Matsuda F, Yamaguchi M, Heath S, et al. A QTL influencing F cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15. *Nat Genet*. 2007 Oct;39(10):1197–9.
8. Lettre G, Sankaran VG, Bezerra MAC, Aгаџо AS, Uda M, Sanna S, et al. DNA polymorphisms at the BCL11A, HBS1L-MYB, and beta-globin loci associate with fetal hemoglobin levels and pain crises in sickle cell disease. *Proc Natl Acad Sci U S A*. 2008 Aug 19;105(33):11869–74.
9. Uda M, Galanello R, Sanna S, Lettre G, Sankaran VG, Chen W, et al. Genome-wide association study shows BCL11A associated with persistent fetal hemoglobin and amelioration of the phenotype of β -thalassemia. *Proc Natl Acad Sci*. 2008 Feb 5;105(5):1620–5.
10. Hoffmann M, Kleine-Weber H, Schroeder S, Krѳger N, Herrler T, Erichsen S, et al. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* [Internet]. 2020 Mar 5 [cited 2020 Apr 6]; Available from: <http://www.sciencedirect.com/science/article/pii/S0092867420302294>
11. Yan R, Zhang Y, Li Y, Xia L, Guo Y, Zhou Q. Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science*. 2020 Mar 27;367(6485):1444–8.
12. Scholtens S, Smidt N, Swertz MA, Bakker SJ, Dotinga A, Vonk JM, et al. Cohort Profile: LifeLines, a three-generation cohort study and biobank. *Int J Epidemiol*. 2015 Aug 1;44(4):1172–80.
13. Francioli LC, Polak PP, Koren A, Menelaou A, Chun S, Renkens I, et al. Genome-wide patterns and properties of de novo mutations in humans. *Nat Genet*. 2015 Jul;47(7):822–6.
14. Tigchelaar EF, Zhernakova A, Dekens JAM, Hermes G, Baranska A, Mujagic Z, et al. Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. *BMJ Open*. 2015 Aug 1;5(8):e006772.
15. van der Velde KJ, Imhann F, Charbon B, Pang C, van Enckevort D, Slofstra M, et al. MOLGENIS research: advanced bioinformatics data software for non-bioinformaticians. *Bioinformatics*. 2019 Mar 15;35(6):1076–8.
16. Consortium T 1000 GP. A global reference for human genetic variation. *Nature*. 2015 Oct;526(7571):68–74.
17. Zhernakova A, Kurilshikov A, Bonder MJ, Tigchelaar EF, Schirmer M, Vatanen

- T, et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science*. 2016 Apr 29;352(6285):565–9.
18. Zhernakova DV, Le TH, Kurilshikov A, Atanasovska B, Bonder MJ, Sanna S, et al. Individual variations in cardiovascular-disease-related protein levels are driven by genetics and gut microbiome. *Nat Genet*. 2018 Nov;50(11):1524–32.
19. Sediq R, van der Schans J, Dotinga A, Alingh RA, Wilffert B, Bos JH, et al. Concordance assessment of self-reported medication use in the Netherlands three-generation Lifelines Cohort study with the pharmacy database iAdbNl: The PharmLines initiative. *Clin Epidemiol*. 2018 Aug 16;10:981–9.
20. Zheng X, Gogarten SM, Lawrence M, Stilp A, Conomos MP, Weir BS, et al. SeqArray-a storage-efficient high-performance data format for WGS variant calls. *Bioinforma Oxf Engl*. 2017 Aug 1;33(15):2251–7.
21. Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME, Wolford BN, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet*. 2018 Sep;50(9):1335–41.
22. Winkler TW, Justice AE, Graff M, Barata L, Feitosa MF, Chu S, et al. The Influence of Age and Sex on Genetic Associations with Adult Body Size and Shape: A Large-Scale Genome-Wide Interaction Study. *PLOS Genet*. 2015 Oct 1;11(10):e1005378.
23. Weiland Q, Lazarus R, Warnes G, Duffy D, Man M. Power Calculation for Testing If Disease is Associated with Marker in a Case-Control Study Using the GeneticsDesign Package [Internet]. 2019 [cited 2020 May 14]. Available from: <https://www.bioconductor.org/packages/release/bioc/html/GeneticsDesign.html>
24. Sun BB, Maranville JC, Peters JE, Stacey D, Staley JR, Blackshaw J, et al. Genomic atlas of the human plasma proteome. *Nature*. 2018 Jun;558(7708):73–9.
25. Ram CVS. Angiotensin receptor blockers and diuretics as combination therapy: Clinical implications. *Am J Hypertens*. 2004 Mar 1;17(3):277–80.
26. Flack JM. Maximising antihypertensive effects of angiotensin II receptor blockers with thiazide diuretic combination therapy: focus on irbesartan/hydrochlorothiazide. *Int J Clin Pract*. 2007 Dec;61(12):2093–102.
27. Kuster GM, Pfister O, Burkard T, Zhou Q, Twerenbold R, Haaf P, et al. SARS-CoV2: should inhibitors of the renin–angiotensin system be withdrawn in patients with COVID-19? *Eur Heart J* [Internet]. 2020 [cited 2020 Apr 17]; Available from: <https://academic.oup.com/eurheartj/advance-article/doi/10.1093/eurheartj/ehaa235/5810479>
28. Little P. Non-steroidal anti-inflammatory drugs and covid-19. *BMJ* [Internet]. 2020 Mar 27 [cited 2020 Apr 6];368. Available from: <https://www.bmj.com/content/368/bmj.m1185>
29. Russell B, Moss C, Rigg A, Van Hemelrijck M. COVID-19 and treatment with NSAIDs and corticosteroids: should we be limiting their use in the clinical setting? *ecancermedicallscience* [Internet]. 2020 Mar 30 [cited 2020 Apr 8];14. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7105332/>
30. Mas S, Gassot P, Alvarez S, Ortiz J, Sotoca JM, Francino A, et al. Pharmacogenetic predictors of angiotensin-converting enzyme inhibitor-induced cough: the role of ACE, ABO, and BDKRB2 genes. *Pharmacogenet Genomics*. 2011 Sep;21(9):531–8.
31. Gill D, Arvanitis M, Carter P, Cordero AIH, Jo B, Karhunen V, et al. ACE inhibition and cardiometabolic risk factors, lung ACE2 and TMPRSS2 gene expression, and plasma ACE2 levels: a Mendelian randomization study. *medRxiv*. 2020 Apr 14;2020.04.10.20059121.
32. Mancia G, Rea F, Ludergrani M, Apolone G, Corrao G. Renin–Angiotensin–Aldosterone System Blockers and the Risk of Covid-19. *N Engl J Med* [Internet]. 2020 May 1 [cited 2020 May 14]; Available from: <https://www.nejm.org/doi/10.1056/NEJMoa2006923>
33. Mehra MR, Desai SS, Kuy S, Henry TD, Patel AN. Cardiovascular Disease, Drug Therapy, and Mortality in Covid-19. *N Engl J Med*. 2020 Jun 18;382(25):e102.

Isis Ricaño-Ponce, Javier Gutierrez-Achury, Ana Florencia Costa, Patrick Deelen, Alexander Kurilshikov, Maria Magdalena Zorro, Mathieu Platteel, Adriaan van der Graaf, Consortium for the study of genetic associations of celiac disease in Latin-America, Serena Sanna, Oscar Daffra, Alexandra Zhernakova, Jingyuan Fu, Gosia Trynka, Edgardo Smecuol, Sonia Isabel Niveloni, Julio Cesar Bai, Vinod Kumar & Cisca Wijmenga*

chapter 3

ImmunoChip meta-analysis in European and Argentinian populations identifies two novel genetic loci associated with celiac disease

Celiac disease (CeD) is a common immune-mediated disease of the small intestine that is triggered by exposure to dietary gluten. While the HLA locus plays a major role in disease susceptibility, 39 non-HLA loci were also identified in a study of 24,269 individuals. We now build on this earlier study by adding 4,125 additional Caucasian samples including an Argentinian cohort. In doing so, we not only confirm the previous associations, we also identify three novel independent genome-wide significant associations at loci: 3p14.1, 12p13.31 and 22q13.1. By applying a genomics approach and differential expression analysis in CeD intestinal biopsies, we prioritize potential causal genes at these novel loci, including *LTBR*, *FRMD4B*, *CYTH4* and *RAC2*. Nineteen prioritized causal genes are overlapping known drug targets. Pathway enrichment analysis and expression of these genes in CeD biopsies suggests they have roles in regulating multiple pathways such as the tumor necrosis factor (TNF) mediated signaling pathway and positive regulation of I- κ B kinase/NF- κ B signaling.

Introduction

Celiac disease (CeD) is a common immune-mediated disease (IMD) present in approximately 1% of the Western population that is characterized by inflammation of the small intestine, villous atrophy and crypt hyperplasia. CeD is caused by an interaction of environmental and genetic factors (1). The main environmental factor is exposure to dietary gluten and the only available treatment is life-long adherence to a gluten-free diet. An estimation of the genetic component of CeD in twins is 75% (2). The main genetic risk factors for the development of CeD are human leukocyte antigen (HLA) molecules, specifically the HLA-DQ2.5 and HLA-DQ2.2 haplotypes, which are responsible for 40% of disease heritability (3). However, while the presence of the HLA-DQ2.5 and HLA-DQ2.2 haplotypes are necessary to develop the disease, they are not sufficient in themselves to promote the disease; many HLA carriers do not develop CeD, indicating that additional genetic factors may play a role.

Previous genome-wide association studies (GWAS)(4)-(5) have identified 26 loci outside the HLA-region that increase the risk of developing CeD. In 2011, we fine-mapped more than 50% of these previously CeD-associated loci and identified 13 novel non-HLA loci using the ImmunoChip platform, which led to the identification of 57 independent SNPs in 39 loci. Although it remains a challenge to pinpoint the causal variants and genes at these 39 CeD-associated loci, much progress has been made using integrative functional genomic approaches that combine multiple layers of omics information. Analysis of the candidate causal genes for CeD from these loci has led to a better understanding of disease pathology and identified new causal pathways, such as interferon gamma signaling (6) and autophagy (7).

Adapted from Ricaño-Ponce et al. ImmunoChip meta-analysis in European and Argentinian populations identifies two novel genetic loci associated with celiac disease Eur. J. of Human Genetics (2020)

*: Corresponding author

Bigger sample sizes and integrative omics approaches for other diseases have not only led to a better understanding of disease biology (8–10), they have also pinpointed new treatment options (11). In the present study, we aimed to identify new loci contributing to CeD by increasing the sample size (adding > 4000 new samples) and adding ethnic diversity of our patient cohort. We then used a system genetics approach to identify new pathways playing a role in disease pathogenesis.

Material and methods

Subjects

Aside from the two cohorts from Latin America of self-reported Argentinian origin, all individuals included in our analysis are Europeans from the Netherlands, Spain, Italy, Ireland and Poland (**Supplementary Table 1**). All cases were diagnosed according to standard clinical criteria, positive tissue transglutaminase antibodies or endomysial antibodies and, in all cases, small intestinal biopsy with Marsh stages II or III. Written informed consent was obtained for all individuals and the study was approved by the ethics committee or institutional review board of all participating institutions. We used the British, Italian, Polish, Spanish and Dutch cohorts that were included and described in our previous Immunochip analyses (3), and added additional samples to the Dutch, Spanish, Italian and Polish cohorts (for a total of 3 925 cases and 4 743 controls) following the same inclusion criteria (**Supplementary Table 1**). The Indian cohort included in the Trynka *et al.* study was instead excluded being too genetically different from our cohorts. We also added an Irish cohort (393 cases and 455 controls) that was described previously by Coleman *et al* (12), and two Argentinian cohorts specifically collected for this study. The Argentinian cases were included after diagnosis by the presence of tissue transglutaminase IgA or anti-deamidated gliadin peptide IgG antibodies and positive endomysial antibodies, and intestinal biopsy class Marsh IIIA or above. Argentinian controls were unselected blood donors and population controls. Blood samples for DNA isolation were collected in the Gastroenterology hospital “Dr. C. Bonorino Udaondo” in Buenos Aires, Argentina and in the OSEP in Mendoza, Argentina after written informed consent was given.

Genotyping and quality control

DNA isolation of the Argentinian samples was carried out at the University Medical Center Groningen (UMCG) by the salting out procedure (13). The additional samples were genotyped on the Immunochip at the UMCG, following Illumina’s standard protocols. Variant calling was performed using Genome Studio with the same cluster used by Trynka *et al* (3). All quality control (QC) checks and filters were performed per cohort using PLINK versions 1.07(14) and 1.9 (15). Specifically, non-polymorphic markers and markers with duplicated rs identifiers were removed and data mapped to the human reference 19 (build 37) using the LiftOver tool from UCSC (<http://genome.ucsc.edu/cgi-bin/hgLiftOver>) (16). Samples with a call rate < 98% and single nucleotide polymorphism (SNPs) with a

chapter 3

call rate < 99% or Hardy-Weinberg equilibrium exact test < 0.001 were discarded.

Hidden relationships between samples within cohorts were identified by calculating identity by descent estimates, which were derived using 14,453 non-HLA independent variants (two rounds of linkage-disequilibrium (LD) pruning using the “indep” option (window size of 50, step $\geq$ 2 size, $r^2 > 0.2$). When we identified duplicated individuals, the sample with the best call rate was kept. Individuals with first and second degree of consanguinity (pihat > 0.2) were excluded. Population outliers within each cohort were detected by multi-dimensional scaling (MDS) plots using R Studio (<http://www.rstudio.com/>) with the previously described set of SNPs and excluded.

Due to its confirmed association to CeD, extended LD and high complexity, we excluded the HLA region (chr6:19,892,021-39,892,022); we also excluded the X chromosome. After applying sample and SNP QC filters, we obtained 12,948 cases and 14,826 controls from eight different cohorts (**Supplementary Table 1**) and 127,855 SNPs for association analysis.

Statistical analysis.

Logistic regression was implemented per cohort using PLINK 1.9 (15) including gender and 3 MDS components as covariates to correct for population stratification. We used a sample-size-weighted Z-score meta-analysis on the association results of the eight cohorts (**Supplementary table 1**) in PLINK 1.9 (15). Manhattan plots of $-\log_{10}P$ and the QQ plot were generated using RStudio (<http://www.rstudio.com/>). The inflation factor in the non-CeD associated regions was 1.67 (**Supplementary Figure 1A**) as expected by the ImmunoChip designed that included mainly immune-related genes, but similar to our previous ImmunoChip study (3) non-excess of associations was observed in three densely genotyped loci selected for bipolar disorder ($\lambda = 1.055$, **Supplementary Figure 1B**). We used the standard genome-wide significance threshold of $p < 5 \times 10^{-8}$, as PLINK reports p-values below 2.22×10^{-16} as 0 for the meta-analysis, we assigned an arbitrary value of 9.99×10^{-17} for potting and results are indicate as $p < 2.22 \times 10^{-16}$ in the manuscript . Regional association plots for genome-wide significant loci were generated using LocusZoom (<http://locuszoom.org/>).

SNP annotation

We used the SNP2GENE function in FUMA (<http://fuma.ctglab.nl/>) to perform functional mapping and annotation of results from the association analysis. We selected the European population of the 1000 Genomes Project phase 3 to calculate LD and included all SNPs with a minor allele frequency (MAF) > 0.001. For gene mapping, we selected the eQTL option, including all databases except GTEx v6 (because it already includes the latest version (v7) and the samples overlap between the two databases). We included only significant eQTLs (FDR < 0.05). We used all genes in Ensembl version 92 for gene mapping. To explore tissue-specificity and biological context of identified genes and perform pathway enrichment

analysis, we used the function GENE2FUNC.

We used Immunobase (www.immunobase.org) to explore if the associated regions were associated to other immune-mediated diseases (IMDs) or to CeD by other studies. To identify the closest gene, we annotated the strongest associated SNP at each locus (TopSNPs) using Haploreg v4 (17), and retrieved the effect on gene expression (eQTLs) from a selected list of 12 studies. We also extracted eQTL information from GTEx (<http://www.gtexportal.org/home/>) and mapped eQTLs using peripheral blood RNA-seq data from 2,116 unrelated individuals, as described by Zhernakova et al (18). Finally, we used GeneNetwork (<https://www.genenetwork.nl/>) to predict the function of the genes.

Figure 1
Regional plots of genome-wide significant loci.

SNP with the strongest association in the region is shown in purple. SNPs in LD with the strongest associated SNP are shown in red ($r^2 < 1$ and > 0.8), orange ($r^2 < 0.8$ and > 0.6), green ($r^2 < 0.6$ and > 0.4), light blue ($r^2 < 0.4$ and > 0.2), and dark blue ($r^2 < 0.2$). Lower panel shows the genes located within the region. (A) Association signals at the 1q25.3 locus. (B) Association signals at the 3p14.1 locus. (C) Association signals at the 12p13.31 locus. (D) Association signals at the 22q13.1 locus

Differential gene expression analysis in intestinal biopsies.

We assessed the expression of the genes affected by the TopSNPs as detected by HaploReg and eQTL annotation, in intestinal biopsies of 12 celiac patients and 12 controls. The biopsies were selected according to United European Gastroenterology criteria; the biopsy sampling, RNA isolation (19) and microarray hybridization have been described previously (20).

CHR	SNP Chr position	SNP	Risk allele	Other allele	N	OR	P	MAF in EUR popu- lation	Closest gene	Is the region asso- ciated to other IMDs?	All eQTLs
1	183532580	rs17849502*	T	G	7	1.43	<2.22 x 10 ⁻¹⁶	0.06	<i>NCF2</i>	SLE, CeD	none
3	69252899	rs6806528	T	C	8	1.19	9.10 x 10 ⁻⁹	0.13	<i>FRMD4B</i>		FRMD4B
12	6511996	rs2364484	C	A	6	1.13	5.31E x 10 ⁻⁹	0.28	<i>RP1-102E24.8, LTBR</i>	AS, JIA, MS, PBC	LTBR
22	37633851	rs9610686	C	T	7	1.11	3.28 x 10 ⁻⁸	0.39	<i>RAC2</i>	T1D, VIT	C1QTNF6, SSTR3, CYTH4, RAC2

The raw data has been deposited in EBI ArrayExpress with the accession ID ‘E-MTAB-4613’(20). Expression data were quantile normalized using the Illumina Beadstudio program. Quantile-normalized and log2-transformed expression values were used for differential expression analysis and differences were assessed with a T-test. Significance was defined as a $p < 0.05$. Boxplots were generated using R Studio.

Protein-QTLs in plasma

We used existing imputed genotype data and *LTBR* concentrations from 1,179 individuals from the Life lines-DEEP cohort generated in a previous study by Zhernakova *et al.* (21). We performed Spearman correlation analysis between the SNP (rs2364484) dosage and *LTBR* levels to test the association between SNP genotypes and protein levels.

Results

Validation non-HLA loci previously associated to CeD

In a meta-analysis of the eight CeD-cohorts (Supplementary Table 1), we confirmed the association of all 38 previously reported autosomal non-HLA CeD loci (3). In 35 of the associated loci, the previously reported TopSNP was also our most significantly associated SNP. In the other three loci, the previously reported TopSNP was excluded during QC in six of the cohorts, thus other SNPs in high LD with them showed the most significant associations. A Manhattan plot illustrating the results of the associations after excluding the HLA region is shown in Supplementary figure 2.

Identification of three new loci associated to CeD

Four novel loci outside of the previously reported associations (3) reached genome wide significance ($p < 5 \times 10^{-8}$, Table 1) in this analysis: 1q25.3, 3p14.1, 12p13.31 and 22q13.1.

The top SNP, rs17849502 at the 1q25.3 locus (Risk allele = T, OR = 1.43, $p < 2.22 \times 10^{-16}$, Figure 1A), is located in an exon of the Neutrophil Cytosolic Factor 2 gene (*NCF2*). The association to this missense variant

Table 1
Genome-wide significant loci associated to celiac disease
Abbreviations: CHR = chromosome; SNP = single nucleotide polymorphism; N = number of cohorts included in the analysis; OR = odds ratio; P = P value from the weighted meta-analysis; MAF = minor allele frequency; EUR = European population from 1000 genomes project.

Figure 2
***FRMD4B* (3p14.1) locus.**
 (A) The risk allele (underlined in red) of the Top-SNP rs6806528: C > T increases the expression of the *FRMD4B* gene ($p = 3.36 \times 10^{-6}$). The number of individuals analyzed is shown under each genotype. (B) *FRMD4B* gene function predictions based on GeneNetwork, a co-expression-based network

(rs17849502) was missed in our previous Immunochip study (3) because it was a single association and there are not SNPs in LD with them within the locus. However, this low frequency variant (MAF = 0.07) was later identified in a re-sequencing study of CeD patients (22), and its association was replicated in a cross-disease meta-analysis of CeD and RA (23). This region has also been associated to systemic lupus erythematosus (24). The other three loci that reached genome-wide significance are all novel, with TopSNPs located in non-coding regions of the genome.

The TopSNP rs6806528 (Risk allele = T, OR = 1.18, $p = 9.1 \times 10^{-9}$) at the 3p14.1 locus is located within an intron of the FERM domain containing the 4B gene (*FRMD4B*); we did not find any proxies of the TopSNP ($r^2 > 0.8$) to be coding. This locus has not been associated to any other IMDs, but it did show modest association to CeD (25) in a cohort of 1,550 North American CeD cases and 3,084 controls ($p = 0.0012$). The risk allele of rs6806528 (T) increases the gene expression levels of *FRMD4B* in blood based on exon level eQTL analysis ($p = 3.36 \times 10^{-9}$, **Figure 2A**). The *FRMD4B* gene functions as a scaffolding protein (26) and is predicted by GeneNetwork to be involved in riboflavin metabolism, the Fc epsilon RI signaling pathway, the T cell receptor signaling pathway and axon guidance (**Figure 2B**).

The TopSNP rs9610686 (Risk allele = C, OR = 1.107, $p = 3.28 \times 10^{-9}$) of the 22q13.1 locus (**Figure 1D**) is a common variant located in an intron of the Ras-related C3 botulinum toxin substrate 2 gene (*RAC2*). This locus has previously been associated to type 1 diabetes (27) and vitiligo (28). TopSNP rs9610686 affects gene expression of multiple nearby genes: *C1QTNF6*, *CYTH4*, *RAC2* and *SSTR3*. In artery aorta, blood, breast mammary tissue and skin, it affects the gene expression of Complement C1q and tumor necrosis factor-Related protein 6 (*C1QTNF6*). *C1QTNF6*

Figure 3
12p13.31 locus.
(A) Functional predictions based on GeneNetwork for genes affected by the most-associated SNP in the locus (rs9610686:C > T). (B) Expression of the *CYTH4* and *RAC2* genes is significantly higher in CeD cases with a Marsh III diagnosis, as compared to healthy controls

modulates inflammation and insulin sensitivity in obese and diabetic mice and humans (29), and is predicted to be mainly involved in glycan-related processes and cell adhesion (**Figure 3A**). The somatostatin receptor 3 gene (*SSTR3*), which regulates antiproliferative signaling and apoptosis (30) and is predicted to be involved in glycosphingolipid biosynthesis and diabetes, is expressed in several tissues, including brain, ovary, pituitary, uterus, blood and testis. *SSTR3* expression level is affected by the rs9610686 genotype only in testis, where its expression is higher than in the other tissues. Additionally, the risk allele of rs9610686 decreases the levels of expression in blood of two other genes, cytohesin 4 (*CYTH4*) and *RAC2*, both involved in immune-related processes. *CYTH4*'s strongest pathway predictions are for toll-like receptor signaling, leukocyte transendothelial migration, natural killer cell mediated cytotoxicity, Fc gamma R-mediated phagocytosis and chemokine signaling. Mutations in *RAC2* cause neutrophil immunodeficiency syndrome (31), which is characterized by severe bacterial infections and poor wound healing. *RAC2* is involved in actin-based cellular functions of phagocyte cells, as well as cell proliferation and cell survival (32). It is also predicted to be involved in primary immunodeficiency, hematopoietic cell lineages, the Fc epsilon RI signaling pathway, the B cell receptor signaling pathway and natural killer cell mediated cytotoxicity. B cells (33) and natural killer cells (34) are important players in CeD pathogenesis. In addition to their immunity-related functions, *CYTH4* and *RAC2* are overexpressed in intestinal biopsies of celiac patients ($p = 0.00024$ and $p = 6.77 \times 10^{-6}$, respectively, **Figure 3B**), further suggesting they have a role in the disease.

Figure 4
Functional annotation
of the 12p13.31 locus
containing *LTBR*.

(A) The risk allele (underlined in red) of Top-SNP rs2364484:C > R, the strongest association in the 12p13.31 locus, increases the expression of the *LTBR* gene ($p=1.51 \times 10^{-9}$). Number of individuals analyzed is shown under each genotype. (B) The expression of the *LTBR* gene is significantly higher in CeD cases with a Marsh III diagnosis, compared with healthy controls. (C) The risk allele of Top-SNP rs2364484:C > R (underlined in red) significantly increases the concentration of *LTBR* in plasma of healthy individuals

At the chromosome 12p13.31 locus, the TopSNP rs2364484 (rs2364484, Risk allele = C, OR = 1.13, $p = 5.31 \times 10^{-99}$) is an intergenic variant between the Lymphotoxin Beta Receptor gene (*LTBR*) and the CD27 antisense RNA 1 gene (*CD27-AS1*) (Figure 1C). This locus has previously been associated to ankylosing spondylitis (35), juvenile idiopathic arthritis (36), multiple sclerosis (MS)(10) and primary biliary cirrhosis (37). TopSNP rs2364484*C increases the expression of *LTBR* in multiple tissues including blood (Figure 4A), brain, stomach, testis, adipose, artery, breast, colon, esophagus and pancreas. TopSNP rs2364484*C has also been shown to affect *LTBR* expression in ileal biopsies of 173 individuals (38) ($p = 1.22 \times 10^{-12}$), and the expression of *LTBR* was increased in intestinal biopsies of 12 CeD patients compared to controls ($p = 0.045$, Figure 4B). Furthermore, the risk allele C of rs2364484 increased the concentration of *LTBR* in plasma of 1,179 healthy individuals from the Lifelines-DEEP cohort ($p = 4.28 \times 10^{-6}$, Figure 4C). *LTBR* is involved in cell death, chemokine release and inflammation (39), all important pathways in CeD, and the role of *LTBR* in the non-canonical NF κ B activation cascade is well established(40,41).

Functional annotation and pathway enrichment analyses on all CeD loci

To explore the functional impact of all CeD-associated loci, we performed functional annotation of significant loci and gene-mapping using FUMA. The SNP2GENE function identified 34 loci reaching genome-wide significance in this study, as some loci from our previous Immunochip study were only suggestive here (Supplementary figure 4AV), comprising 4,045 candidate SNPs including our TopSNPs and SNPs in high LD with them ($r^2 > 0.8$). Thirty-six candidate SNPs were exonic within coding genes, 45 were exonic within ncRNAs, 53 were located in 3' UTR and 19 in 5' UTR (Supplementary Figure 4B). Using multiple independent eQTLs datasets, FUMA mapped the candidate SNPs to 212 genes. The expression of these candidate genes was analyzed with MAGMA tissue expression analysis implemented in FUMA using 30 general tissue types from GTEx v7. We found significant enrichment of candidate genes to be expressed

chapter 3

in blood, spleen and small intestine (**Supplementary figure 5**). It has been established that cells present in blood are important players in CeD (42), and that the disease leads to small-intestinal mucosal injury (1). One third of CeD patients have defective spleen function and the prevalence of this dysfunction increases to 80% as the severity of the disease increases (43), indicating that the factors causing CeD also affect the spleen.

We explored if the results from the meta-analysis would lead to the discovery of new treatment options for CeD. While there are still no reported drugs for the treatment of CeD, 19 of the 212 candidate genes prioritized in the FUMA analysis are reported drug targets, including *RAC2* from the novel 22q13.1 locus (**Supplementary Table 2**). Some of these drugs reduce inflammation or are immune-suppressants, and they are indicated for use in IMDs. Reported drugs include vedolizumab and CCX282 for the treatment for IBD, natalizumab for MS, abatacept for RA and JIA, galiximab for RA and PS, 2-Methoxyestradiol for RA and INCB3284 that is being investigated for use/treatment in inflammatory disorders, which might indicate a potential effect in CeD, but this requires follow-up study.

Using the 212 candidate genes as input, we looked for gene enrichment in multiple data sets and found enrichment of 286 Gene Ontology biological terms (**Supplementary Table 3**). We were able to confirm the enrichment to many well-known CeD pathways (**Supplementary Table 3**), including regulation of alpha beta T cell activation and proliferation, regulation of cell-cell adhesion, regulation of lymphocytes and leucocytes, production of multiple cytokines including interferon gamma, regulation of inflammatory response and regulation of B cell mediated immunity. Some of the pathways that contain novel associated genes (**Supplementary Table 4**) popped up for the first time, including TNF-mediated signaling, response to TNF regulation of I- κ B kinase/NF- κ B signaling, positive regulation of I-kappaB kinase/NF- κ B signaling and apoptotic signaling.

LTBR locus links NF- κ B pathway to celiac disease

Pro-inflammatory cytokines, adhesion molecules and enzymes whose gene expression is known to be regulated by NF κ B are involved in CeD (44). There is also a deregulation of the NF κ B -pathway in the intestine of CeD patients (45). As mentioned before, LTBR is well known for its role in the NF κ B -pathway. In addition to LTBR, three other genes involved in the NF κ B pathway have also been prioritized as CeD genes: Receptor activator of nuclear factor kappa-B ligand (RANKL), TNF Alpha Induced Protein 3 (TNFAIP3) and protein kinase C gamma (PRKCG). To formally test whether the NF κ B -pathway was involved in CeD pathogenesis, we compared the expression of 95 genes involved in the NF κ B signaling pathway according to the Kyoto Encyclopedia of Genes and Genomes (http://www.genome.jp/kegg-bin/show_pathway?hsa04064) in intestinal biopsies of patients with active CeD with those of healthy controls. We observed that 37 of the 95 genes showed significant differences in their levels of expression ($p < 0.05$, **Supplementary figure 6**) as shown in **Supplementary figure 6a** and **6b**. These differentially expressed genes are involved in both the canonical and non-canonical NF κ B pathway.

Discussion

We report here the largest meta-analysis of celiac cases and controls analyzed to date ($n = 27,774$: 12,948 cases and 14,826 controls), adding 4,125 new samples to our previous Immunochip study (3). We identified four novel loci at genome-wide significance and another 18 showing evidence for suggestive association. As expected, most of the novel loci are within regions previously associated to other IMDs such as IBD, type 1 diabetes, psoriasis and MS. These results may imply a high level of genetic sharing of IMDs, but could reflect the design of the Immunochip, which was designed to densely genotype regions associated by previous GWAS for fine-mapping purposes and suggestive variants for replication. To clarify on similarity and differences of CeD and other IMDs, association studies across the whole genome using large cohorts are needed. Such whole genome association approaches will also allow the discovery of additional new loci that are not fully covered on the Immunochip. These discoveries would lead to a better understanding of the disease-specific genetic and molecular mechanisms.

Our study implicates three new CeD-associated regions. The locus on 3p14.1 containing the *FRMD4B* gene has not been associated to any other IMD. This lack of association in previous studies might be caused by the poor coverage within the locus, as in the initial Immunochip analyses the 3p14.1 variants were excluded during the QC process, similarly to the association identified to *NCF2* in the 1q25.3 locus. The potential role of *FRMD4B* in CeD needs more functional investigation, as the role of this gene is currently unclear.

The most plausible candidate genes in the 22q13.1 locus are *CYTH4* and *RAC2*. While *CYTH4* has been mainly associated with schizophrenia and bipolar disorder, an evolutionary analysis (46) looking at regulatory elements conserved across mammals within the *RAC2* gene identified three major haplogroups present in the population. One of these was associated to an increased risk for MS and inflammatory bowel disease, suggesting an important role for *RAC2* in the pathogenesis of IMDs. *RAC2* also activates T helper (Th) 1-specific signaling and IFN- γ gene expression (47). In CeD, gliadin-specific CD4 $^{+}$ T cells respond to gliadin peptides presented via HLA-DQ2 or HLA-DQ8, which represent the strongest genetic risk for the disease. Upon activation, gliadin-specific CD4 $^{+}$ T cells polarize towards the Th1-type pathway and produce IFN- γ , whose expression is also up-regulated in intestinal biopsies of untreated celiac patients (48), further implying a role for *RAC2* in CeD.

The 12p13.31 locus containing *LTBR* has also been associated to ankylosing spondylitis (49), however through an independent variant that leads to the splicing of exon 6 of *TNFRSF1A*, resulting in loss of the trans-membrane domain. Our CeD TopSNP was not in high LD with the ankylosing spondylitis variants. It is on the same haplotype as a non-synonymous coding variant in *LTBR* associated to juvenile idiopathic arthritis (50) (rs2364480, $r^2 = 0.9$, $D' = 0.96$), which suggests that *LTBR* is an important causal gene for multiple autoimmune diseases. Furthermore, the CeD TopSNP increases the expression of the *LTBR* gene in blood

chapter 3

and is differentially expressed in biopsies of celiac patients. *LTBR* is well known to be involved in multiple immune pathways, including the non-canonical NFκB pathway. Although the role of the NFκB pathway in CeD is well known and has been validated by experimental studies (44,45), it was not clear whether the deregulation of this pathway is a cause or a consequence of CeD. Our study, however, suggest a causal role for NFκB in CeD pathogenesis as we find strong association of four NFκB genes and their differential expression in CeD intestinal biopsies.

Our systematic annotation of loci from the meta-analysis lead to the identification of drug targets for 19 prioritized genes. Some of these drugs reduce inflammation or are immune-suppressants, and they are indicated for use in RA, inflammatory bowel disease, psoriasis, juvenile idiopathic arthritis and MS, while the reposition of such drugs to CeD may need further investigation, our results might help to prioritize drugs for further studies.

We acknowledge limitations of our study. Firstly, use of the ImmunoChip restricted our analysis to loci already implicated in autoimmune diseases, which could be one reason we did not discover novel non-immune pathways. Secondly, although we included a non-European population, the design of the ImmunoChip is based on the European population and does not include population-specific variants from other ethnicities, thus a more suitable platform should be used to study the Argentinian population.

In conclusion, we have shown that increasing the sample size of our previous study allowed us to not only map new regions associated to CeD, but also to identify new disease pathways. The integration of multiple layers of omics information provided more insight into the individual loci and into the pathways involved in disease pathogenesis.

Acknowledgements

We thank the Argentinian clinicians for recruiting individuals with CeD to provide blood samples; the genotyping facility of the UMCG for help in generating the Immunochip data; Jeffrey Barrett, Rinse Weersma and Ross McManus for providing genotypes from extra controls; all the participating CeD patients and controls; and Kate Mc Intyre for editing the manuscript.

Supplementary information

The Supplementary Material for this article can be found online at: <https://www.nature.com/articles/s41431-019-0520-4#Sec17>

Funding

This work was supported by an ERC Advanced grant [FP/2007-2013/ERC grant 2012-322698], an NWO Spinoza prize grant [NWO SPI 92-266], NWO-VIDI grants [864.13.013] to J.F. and [016.178.056] to A.Z., Hypatia grant to V. K. from Rodboud UMC, NWO Gravitation Netherlands Organ-on-Chip Initiative [024.003.001] to C.W., a European Union Seventh Framework Programme grant (EU FP7) TANDEM project [HEALTH-F3-2012-305279] to C.W and a ERC Starting Grant [715772] to

A.Z. GT is supported by the Wellcome Trust grant WT206194.

1. Tack GJJ, Verbeek WHMH, Schreurs MWJWJ, Mulder CJJJ. The spectrum of celiac disease: epidemiology, clinical aspects and treatment. *Nature reviews Gastroenterology & hepatology*. 2010;7(4):204–13.
2. Kuja-Halkola R, Lebowitz B, Halfvarson J, Wijmenga C, Magnusson PKE, Ludvigsson JF. Heritability of non-HLA genetics in coeliac disease: a population-based study in 107 000 twins. *Gut*. 2016 Nov 1;65(11):1793–8.
3. Trynka G, Hunt KA, Bockett NA, Romanos J, Mistry V, Szperl A, et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nature Genetics*. 2011;43(12):1193–201.
4. Dubois P, Trynka G, Franke L, Hunt K, Romanos J, Curtotti a, et al. Multiple common variants for celiac disease influencing immune gene expression. *Nat Genet*. 2010;42(4):295–302.
5. van Heel DA, Franke L, Hunt KA, Gwilliam R, Zhernakova A, Inouye M, et al. A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21. *Nature Genetics*. 2007 Jul 10;39(7):827–9.
6. Kumar V, Westra HJ, Karjalainen J, Zhernakova D V., Esko T, Hrdlickova B, et al. Human Disease-Associated Genetic Variation Impacts Large Intergenic Non-Coding RNA Expression. *PLoS Genetics*. 2013;9(1).
7. Ricaño-Ponce I, Zhernakova D V., Deelen P, Luo O, Li X, Isaacs A, et al. Refined mapping of autoimmune disease associated genetic variants with gene expression suggests an important role for non-coding RNAs. *Journal of Autoimmunity*. 2016 Apr;68:62–74.
8. Tsoi LC, Spain SL, Knight J, Ellinghaus E, Stuart PE, Capon F, et al. Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. *Nature genetics*. 2012;44(12):1341–8.
9. Jostins L, Ripke S, Weersma R, Duerr RH, Dermot P, Hui K, et al. Host-microbe interactions shape genetic risk for inflammatory bowel disease. 2012;
10. Beecham AH, Patsopoulos NA, Xifara DK, Davis MF, Kempainen A, Cotsapas C, et al. Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nature genetics*. 2013;45(11):1353–60.
11. Okada Y, Wu D, Trynka G, Raj T, Terao C, Ikari K, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*. 2014;506(7488):376–81.
12. Coleman C, Quinn EM, Ryan AW, Conroy J, Trimble V, Mahmud N, et al. Common polygenic variation in coeliac disease and confirmation of ZNF335 and NIFA as disease susceptibility loci. *European Journal of Human Genetics*. 2015;353(October 2014):1–7.
13. Miller SA, Dykes DD, Polesky HF. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic acids research*. 1988 Feb 11;16(3):1215.
14. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*. 2007;81(3):559–75.
15. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. 2015;4(7).
16. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. *Genome research*. 2002 Jun 1;12(6):996–1006.
17. Ward LD, Kellis M. HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic acids research*. 2016 Jan 4;44(D1):D877–81.

18. Zhernakova D V, Deelen P, Vermaat M, van Iterson M, van Galen M, Arindrarto W, et al. Identification of context-dependent expression quantitative trait loci in whole blood. *Nature genetics*. 2016;49(November).
19. Diosdado B, Wapenaar MC, Franke L, Duran KJ, Goerres MJ, Hadithi M, et al. A microarray screen for novel candidate genes in coeliac disease pathogenesis. *Gut*. 2004 Jul;53(7):944–51.
20. Hunt KA, Zhernakova A, Turner G, Heap GAR, Franke L, Bruinenberg M, et al. Newly identified genetic risk variants for celiac disease related to the immune response. *Nature Genetics*. 2008 Apr 2;40(4):395–402.
21. Zhernakova D V., Le TH, Kurilshikov A, Atanasovska B, Bonder MJ, Sanna S, et al. Individual variations in cardiovascular-disease-related protein levels are driven by genetics and gut microbiome. *Nature Genetics*. 2018 Nov 24;50(11):1524–32.
22. Hunt KA, Mistry V, Bockett NA, Ahmad T, Ban M, Barker JN, et al. Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. *Nature*. 2013;498(7453):232–5.
23. Gutierrez-Achury J, Zorro MM, Ricaño-Ponce I, Zhernakova DV, Diogo D, Raychaudhuri S, et al. Functional implications of disease-specific variants in loci jointly associated with coeliac disease and rheumatoid arthritis. *Human Molecular Genetics*. 2016;25(1):180–90.
24. Bentham J, Morris DL, Cunninghame Graham DS, Pinder CL, Tomblinson P, Behrens TW, et al. Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nature Genetics*. 2015 Dec 26;47(12):1457–64.
25. Garner C, Ahn R, Ding YC, Steele L, Stoven S, Green PH, et al. Genome-Wide Association Study of Celiac Disease in North America Confirms FRMD4B as New Celiac Locus. *PLoS ONE*. 2014;9(7).
26. Klarlund JK, Holik J, Chawla A, Park JG, Buxton J, Czech MP, et al. Signaling Complexes of the FERM Domain-containing Protein GRSP1 Bound to ARF Exchange Factor GRP1*. *JBC Papers in Press* Published on July. 2001;9(23).
27. Cooper JD, Smyth DJ, Smiles AM, Plagnol V, Walker NM, Allen JE, et al. Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. *Nature Genetics*. 2008;40(12):1399–401.
28. Jin Y, Birlea SA, Fain PR, Ferrara TM, Ben S, Riccardi SL, et al. Genome-wide association analyses identify 13 new susceptibility loci for generalized vitiligo. *Nature Genetics*. 2012 Jun 6;44(6):676–80.
29. Lei X, Seldin MM, Little HC, Choy N, Klonisch T, Wong GW. C1q/TNF-related protein 6 (CTRP6) links obesity to adipose tissue inflammation and insulin resistance. *Journal of Biological Chemistry*. 2017 Sep 8;292(36):14836–50.
30. Sharma K, Patel YC, Srikant CB. Subtype-selective induction of wild-type p53 and apoptosis, but not cell cycle arrest, by human somatostatin receptor 3. *Molecular Endocrinology*. 1996 Dec;10(12):1688–96.
31. Ambruso DR, Knall C, Abell AN, Panepinto J, Kurkchubasche A, Thurman G, et al. Human neutrophil immunodeficiency syndrome is associated with an inhibitory Rac2 mutation. *Proceedings of the National Academy of Sciences of the United States of America*. 2000 Apr 25;97(9):4654–9.
32. Yang FC, Kapur R, King AJ, Tao W, Kim C, Borneo J, et al. Rac2 stimulates Akt activation affecting BAD/Bcl-XL expression while mediating survival and actin function in primary mast cells. *Immunity*. 2000 May;12(5):557–68.
33. Mesin L, Sollid LM, Niro R Di. The intestinal B-cell response in celiac disease. *Frontiers in Immunology*. 2012;3:313.
34. Marafini I, Imeneo MG, Monteleone G. The Role of Natural Killer Receptors in Celiac Disease. *Immunome Research*. 2017 Jan 20;13(1):1–2.
35. Cortes A, Hadler J, Pointon JP, Robinson PC, Karaderi T, Leo P, et al. Identification of multiple risk variants for ankylosing spondylitis through high-density genotyping of immune-related loci. *Nature genetics*. 2013;45(7):730–8.

chapter 3

36. Hinks A, Cobb J, Marion MC, Prahalad S, Sudman M, Bowes J, et al. Dense genotyping of immune-related disease regions identifies 14 new susceptibility loci for juvenile idiopathic arthritis.
37. Mells GF, Floyd JAB, Morley KI, Cordell HJ, Franklin CS, Shin S-Y, et al. Genome-wide association study identifies 12 new susceptibility loci for primary biliary cirrhosis. *Nature genetics*. 2011;43(4):329–32.
38. Kabakchiev B, Silverberg MS. Expression quantitative trait loci analysis identifies associations between genotype and gene expression in human intestine. *Gastroenterology*. 2013 Jun;144(7):1488–96, 1496.e1-3.
39. Chang Y-H, Hsieh S-L, Chen M-C, Lin W-W. Lymphotoxin beta receptor induces interleukin 8 gene expression via NF-kappaB and AP-1 activation. *Experimental cell research*. 2002 Aug 15;278(2):166–74.
40. VanArsdale TL, VanArsdale SL, Force WR, Walter BN, Mosialos G, Kieff E, et al. Lymphotoxin-beta receptor signaling complex: role of tumor necrosis factor receptor-associated factor 3 recruitment in cell death and activation of nuclear factor kappaB. *Proceedings of the National Academy of Sciences of the United States of America*. 1997 Mar 18;94(6):2460–5.
41. Li C, Norris PS, Ni C-Z, Havert ML, Chiong EM, Tran BR, et al. Structurally Distinct Recognition Motifs in Lymphotoxin- β Receptor and CD40 for Tumor Necrosis Factor Receptor-associated Factor (TRAF)-mediated Signaling. *Journal of Biological Chemistry*. 2003 Dec 12;278(50):50523–9.
42. Meresse B, Malamut G, Cerf-Bensussan N. Celiac Disease: An Immunological Jigsaw. *Immunity*. 2012;36(6):907–19.
43. Di Sabatino A, Brunetti L, Carnevale Maffii G, Giuffrida P, Corazza GR. Is it worth investigating splenic function in patients with celiac disease? *World journal of gastroenterology*. 2013 Apr 21;19(15):2313–8.
44. Maiuri MC, De Stefano D, Mele G, Iovine B, Bevilacqua MA, Greco L, et al. Gliadin increases iNOS gene expression in interferon- γ -stimulated RAW 264.7 cells through a mechanism involving NF- κ B. *Naunyn-Schmiedeberg's Archives of Pharmacology*. 2003 Jul 4;368(1):63–71.
45. Fernandez-Jimenez N, Castellanos-Rubio A, Plaza-Izurieta L, Irastorza I, Elcoroaristizabal X, Jauregi-Miguel A, et al. Coregulation and modulation of NF κ B-related genes in celiac disease: uncovered aspects of gut mucosal inflammation. *Human molecular genetics*. 2014 Mar 1;23(5):1298–310.
46. Sironi M, Guerini FR, Agliardi C, Biasin M, Cagliani R, Fumagalli M, et al. An Evolutionary Analysis of RAC2 Identifies Haplotypes Associated with Human Autoimmune Diseases. *Molecular Biology and Evolution*. 2011 Dec 1;28(12):3319–29.
47. Li B, Yu H, Zheng W, Voll R, Na S, Roberts AW, et al. Role of the guanosine triphosphatase Rac2 in T helper 1 cell differentiation. *Science (New York, NY)*. 2000 Jun 23;288(5474):2219–22.
48. Nilsen EM, Jahnsen FL, Lundin KE, Johansen FE, Fausa O, Sollid LM, et al. Gluten induces an intestinal cytokine response strongly dominated by interferon gamma in patients with celiac disease. *Gastroenterology*. 1998 Sep;115(3):551–63.
49. Braun J, Sieper J, Devlam K, al. et, Calin A, Braun J, et al. Ankylosing spondylitis. *Lancet (London, England)*. 2007 Apr 21;369(9570):1379–90.
50. Hinks A, Cobb J, Marion MC, Prahalad S, Sudman M, Bowes J, et al. Dense genotyping of immune-related disease regions identifies 14 new susceptibility loci for juvenile idiopathic arthritis. *Nature Genetics*. 2013 Jun;45(6):664–9.

part two

causal inference of molecular traits

Alexander Kurilshikov†, Carolina Medina-Gomez†, Rodrigo Bacigalupé†, Djawad Radjabzadeht, Jun Wang†, Ayse Demirkan, Caroline I. Le Roy, Juan Antonio Raygoza Garay, Casey T. Finnicum, Xingrong Liu, Daria V. Zhernakova, Marc Jan Bonder, Tue H. Hansen, Fabian Frost, Malte C. Rühlemann, Williams Turpin, Jee-Young Moon, Han-Na Kim, Kreete Lüll, Elad Barkan, Shiraz A. Shah, Myriam Fornage, Joanna Szopskina-Tokov, Zachary D. Wallen, Dmitrii Borisevich, Lars Agreus, Anna Andreasson, Corinna Bang, Larbi Bedrani, Jordana T. Bell, Hans Bisgaard, Michael Boehnke, Dorret I. Boomsma, Robert D. Burk, Annique Claringbould, Kenneth Croitoru, Gareth E. Davies, Cornelia M. van Duijn, Liesbeth Duijts, Gwen Falony, Jinyuan Fu, Adriaan van der Graaf, Torben Hansen, Georg Homuth, David A. Hughes, Richard G. Ijzerman, Matthew A. Jackson, Vincent W.V. Jaddoe, Marie Joossens, Torben Jørgensen, Daniel Keszthelyi, Rob Knight, Markku Laakso, Matthias Laudes, Lenore J. Launer, Wolfgang Lieb, Aldons J. Lusis, Ad A.M. Masclee, Henriette A. Moll, Zlatan Mujagic, Qi Qibin, Daphna Rothschild, Hocheol Shin, Søren J. Sørensen, Claire J. Steves, Jonathan Thorsen, Nicholas J. Timpson, Raul Y. Tito, Sara Vieira-Silva, Uwe Völker, Henry Völzke, Urmo Vösa, Kaitlin H. Wade, Susanna Walter, Kyoko Watanabe, Stefan Weiss, Frank U. Weiss, Omer Weissbrod, Harm-Jan Westra, Gonke Willemsen, Haydesh Payami, Daisy M.A.E. Jonkers, Alejandro Arias Vasquez, Eco J.C. de Geus, Katie A. Meyer, Jakob Stokholm, Eran Segal, Elin Org, Cisca Wijmenga, Hyung-Lae Kim, Robert C. Kaplan, Tim D. Spector, Andre G. Uitterlinden, Fernando Rivadeneira, Andre Franke, Markus M. Lerch, Ulde Franke, Serena Sanna, Mauro D'Amato, Oluf Pedersen, Andrew D. Paterson, Robert Kraaij†, Jeroen Raes†, Alexandra Zhernakova†.*

chapter 4

Large-scale association analyses identify host factors influencing human gut microbiome composition

To study the effect of host genetics on gut microbiome composition, the MiBioGen consortium curated and analyzed genome-wide genotypes and 16S fecal microbiome data from 18,340 individuals (24 cohorts). Microbial composition showed high variability across cohorts: only 9 out of 410 genera were detected in more than 95% samples. A genome-wide association study (GWAS) of host genetic variation in relation to microbial taxa identified 31 loci affecting microbiome at a genome-wide significant ($P < 5 \times 10^{-8}$) threshold. One locus, the lactase (*LCT*) gene locus, reached study-wide significance (GWAS signal $P = 1.28 \times 10^{-20}$), and it showed an age-dependent association with *Bifidobacterium* abundance. Other associations were suggestive ($1.95 \times 10^{-10} < P < 5 \times 10^{-8}$) but enriched for taxa showing high heritability and for genes expressed in the intestine and brain. A phenotype-wide association study and Mendelian randomization identified enrichment of microbiome trait loci in the metabolic, nutrition and environment domains and suggested the microbiome has causal effects in ulcerative colitis and rheumatoid arthritis.

Introduction

The gut microbiome is an integral part of the human holobiont. In recent years, many studies have highlighted the link between its perturbations and immune, metabolic, neurologic and psychiatric traits, drug metabolism and cancer (1). Environmental factors, like diet and medication, play a significant role in shaping the gut microbiome composition (2–4), although twin, family and population-based studies have shown that the genetic component also plays a role in determining gut microbiota composition, and a proportion of bacterial taxa are heritable (5,6).

Several studies (7–9) have investigated the effect of genetics on microbiome composition through genome-wide association studies (GWAS) and identified dozens of associated loci. However, little cross-replication across these studies has been observed so far (10,11). This may be due to a number of factors. First, methodological differences in the collection, processing and annotation of stool microbiota are known to have significant effects on the microbiome profiles obtained (12–14) and can generate heterogeneity and a lack of reproducibility across studies. Second, most association signals are rather weak, which suggests that existing studies of 1,000–2,000 samples (7–9) are underpowered. Finally, some of the GWAS signals related to microbiome compositions may be population-specific, i.e. they may represent *bona fide* population differences in genetic structure and/or environment.

To address these challenges and obtain valuable insights into the relationship between host genetics and microbiota composition, we set up the international consortium MiBioGen (11). In this study, we have coordinated 16S rRNA gene sequencing profiles and genotyping data from 18,340 participants from 24 cohorts from the USA, Canada, Israel,

Adapted from Kurilshikov et al. Large-scale association analyses identify host factors influencing human gut microbiome composition. *Nat Genet* 53, 156–165 (2021)

†: Equal contribution first

‡: Equal contribution last

*: Corresponding author

Figure 1
Diversity of microbiome composition across the MiBioGen cohorts.

(a) Sample size, ethnicity, genotyping array and 16S rRNA gene profiling method. The SHIP/SHIP-TREND and GEM_v12/GEM_v24/GEM_1CHIP subcohorts are combined in SHIP and GEM, respectively (**Online Methods**; see **Supplementary Note** for cohort abbreviations). This merge resulted in the total of 21 cohorts depicted in the figure. (b)* Total richness (number of genera with mean abundance over 0.1%, i.e. 10 reads out of 10,000 rarefied reads) by number of cohorts investigated. (c)* Number of core genera (genera present in > 95% of samples from each cohort) by number of cohorts investigated. (d) Pearson correlation of cohort sample size with total number of genera. Confidence band represents the standard error of the regression line. (e)* Unweighted mean relative abundance of core genera across the entire MiBioGen dataset. (f)* Per-sample richness across the 21 cohorts. Asterisks indicate cohorts that differ significantly from all the others (pairwise Wilcoxon rank-sum test; FDR < 0.05). (g) Diversity (Shannon index) across the 21 cohorts, with the DanFund and PNP cohorts presenting higher and lower diversity in relation to the other cohorts (pairwise Wilcoxon rank sum test; FDR < 0.05). (*) For all boxplots, the central line, box and whiskers represent the median, IQR and 1.5 times the IQR.

South Korea, Germany, Denmark, the Netherlands, Belgium, Sweden, Finland and the UK. We performed a large-scale, multi-ethnic, genome-wide meta-analysis of the associations between autosomal human genetic variants and the gut microbiome. We explored the variation of microbiome composition across different populations and investigated the effects of differences in methodology on the microbiome data. Through the implementation of a standardized pipeline, we then performed microbiome trait loci (mbTL) mapping to identify genetic loci that affect the relative abundance (mbQTLs) or presence (microbiome Binary Trait loci, or mbBTLs) of microbial taxa. Finally, we focused on the biological interpretation of GWAS findings through Gene Set Enrichment Analysis (GSEA), Phenome-wide association studies (PheWAS) and Mendelian randomization (MR) approaches.

Results

Landscape of microbiome composition across cohorts

Our study included cohorts that were heterogeneous in terms of ethnic background, age, male/female ratio and microbiome analysis methodology. Twenty cohorts included samples of single ancestry, namely European (16 cohorts, N = 13,266), Middle-Eastern (1 cohort, N = 481), East Asian (1 cohort, N = 811), American Hispanic/Latin (1 cohort N = 1,097) and African American (1 cohort, N = 114), whereas four cohorts were multi-ancestry (N = 2,571) (see **Supplementary Note**, **Supplementary Tables 1,2**).

Twenty-two cohorts comprised adult or adolescent individuals ($N = 16,632$), and two cohorts consisted of children ($N = 1,708$). The microbial composition was profiled by targeting three distinct variable regions of the 16S rRNA gene: V4 (10,413 samples, 13 cohorts), V3-V4 (4,211 samples, 6 cohorts) and V1-V2 (3,716 samples, 5 cohorts) (**Figure 1a**). To account for differences in sequencing depth, all datasets were rarefied to 10,000 reads per sample. Next, we performed taxonomic classification using direct taxonomic binning instead of OTU clustering methods (see **Online Methods**) (11,15,16).

In general, cohorts varied in their microbiome structure at multiple taxonomic levels (**Figure 1b-g**). This variation may largely be driven by the heterogeneity between populations and differences in technical protocols (**Supplementary Tables 1-3**). Combining all samples ($N = 18,340$) resulted in a total richness of 385 genus-level taxonomic groups that had a relative abundance higher than 0.1% in at least one cohort. This observed total richness appears to be below the estimated saturation level (**Figure 1b**), suggesting that a further increase in sample size and a higher sequencing depth are needed to capture the total gut microbial diversity (**Figure 1d**). As expected, the core microbiota (the number of bacterial taxa present in over 95% of individuals) decreased with the inclusion of additional cohorts (**Figure 1c**, **Online Methods**). The core microbiota comprise nine genera, of which seven were previously identified as such (3), plus the genera *Ruminococcus* and *Lachnospirillum* (**Figure 1e**). Of these nine genera, the most abundant genus was *Bacteroides* (18.65% (SD:8.65)), followed by *Faecalibacterium* (6.19% (SD:2.35)), *Blautia* (3.36% (SD:2.84)) and *Alistipes* (3.05% (SD:1.47)). Among the European cohorts that compose the largest genetically and environmentally homogeneous cluster, the core microbiota also included *Ruminiclostridium*, *Fusicatenibacter*, *Butyricicoccus* and *Eubacterium*, genera which typically produce short-chain fatty acids (17).

The DNA extraction method was the principal contributor to heterogeneity, with a non-redundant effect size of 29% on the microbiome variation (measured as average genus abundance per cohort; stepwise distance-based redundancy analysis $R^2_{\text{adj}} = 0.27$, $P_{\text{adj}} = 7 \times 10^{-4}$) (**Supplementary Table 4**). Richness and Shannon diversity also differed significantly across cohorts. The cohorts with the lowest richness (HCHS/SOL) and highest diversity (DanFund) used specific DNA extraction kits that were not used by other studies, possibly contributing to their outlying alpha diversities (**Figure 1f,g**, **Supplementary Table 3**). Overall, the 16S rRNA domain sequenced and the DNA extraction methods used, together with cohort ethnicity, accounted for 32.74% of richness variance.

Given the high heterogeneity of microbial composition across cohorts, we applied both per-cohort and whole study-filters for taxa inclusion in GWAS (see **Online Methods**).

Heritability of microbial taxa and alpha diversity

We performed estimation of heritability (H^2) of gut microbiome composition based on the two twin cohorts included in our study

(Supplementary Table 5). The TwinsUK cohort, composed of 1,176 samples, including 169 monozygotic (MZ) and 419 dizygotic (DZ) twin pairs, was used to estimate H^2 using the ACE (additive genetic variance (A)/ shared environmental factors (C)/ non-shared factors plus error (E)) model. The Netherlands Twin Registry (NTR) cohort (only MZ twins, $N = 312$, 156 pairs) was used to replicate the MZ intraclass correlation coefficient (ICC). None of the alpha diversity metrics (Shannon, Simpson and inverse Simpson) showed evidence for heritability ($A < 0.01$, $P = 1$). Among the 159 bacterial taxa that were present in more than 10% of twin pairs, 19 taxa showed evidence for heritability ($P_{\text{nominal}} < 0.05$) (Figure 2a). The ICC shows concordance between TwinsUK and NTR for these 19 bacterial taxa ($R = 0.25$, $P = 0.0018$, Figure 2b).

The SNP-based heritability calculated from mbQTL summary statistics using linkage disequilibrium (LD) score regression showed two bacterial taxa, genus *Ruminiclostridium* and family *Peptostreptococcaeae*, passing the significance threshold given the number of 211 taxa tested ($Z < 3.68$, Supplementary Table 5). The results of the SNP-based heritability and twin-based heritability showed significant correlation across the tested taxa ($R = 0.244$, $P = 7.2 \times 10^{-4}$).

Thirty one loci associated with gut microbes through GWAS

First, we studied the genetic background of the alpha diversity (Simpson, inverse Simpson and Shannon diversity indices). We identified no significant hits in the meta-GWAS ($P > 5 \times 10^{-8}$; Supplementary Table 6, Supplementary Figure 1), which is in line with the observed lack of heritability for these indices.

Next, we used two separate GWAS meta-analysis approaches (18–20) to explore the effect of host genetics on the abundance levels (mbQTL) or presence/absence (mbBTL) of bacterial taxa in the gut microbiota (see Online Methods).

In total, 18,340 samples and 211 taxa were included in the mbQTL mapping analysis (Online Methods, Supplementary Table 3). We identified genetic variants that mapped to 20 distinct genetic loci associated with the abundance of 27 taxa (Figure 3; Supplementary Figure 2,3; Supplementary Table 7, 8). MbBTL mapping covered 177

Figure 2
Heritability of microbiome taxa and its concordance with mbQTL mapping.

a, Microbial taxa that showed significant heritability in the TwinsUK cohort (ACE model; nominal $P < 0.05$, no adjustment for multiple comparison). Taxa with at least one genome-wide significant (GWS) mbQTL hit are marked in red. Only taxa present in more than 10% of pairs (> 17 MZ pairs and > 41 DZ pairs) are shown. Circles and diamonds represent heritability values. Error bars represent 95% CIs. b, Correlation of MZ ICC between TwinsUK and NTR cohort. Only taxa with significant heritability (ACE model $P < 0.05$) that were present in both TwinsUK and NTR are shown. Red and blue dots indicate bacterial taxa with and without GWS mbQTLs ($P < 5 \times 10^{-8}$), respectively. Segments represent 95% CIs. c, Correlation between heritability significance ($-\log_{10} P_{H^2}$ PH2 TwinsUK) and the number of loci associated with microbial taxon at relaxed threshold ($P_{\text{mbQTL}} < 1 \times 10^{-5}$). Taxa with at least one GWS-associated locus are marked in red. Error bars represent 95% CIs.

taxa, and 10 loci were found to be associated with presence/absence of bacterial taxa (Figure 3, Supplementary Table 7, 9). For one taxon, family *Peptococcaceae*, two independent mbBTLs were detected (Figure 3, Supplementary Table 7). Two out of 30 mbTLs showed heterogeneity in mbTL effect-sizes (Supplementary Note).

In both the mbQTL and mbBTL mapping, only one out of 31 loci (*LCT* locus – *Bifidobacterium*, $P = 8.63 \times 10^{-21}$) passed the strict correction for the number of taxa tested ($P < 1.95 \times 10^{-10}$ for 257 taxa included in the analysis). However, the remaining loci include functionally relevant variants (i.e. the *FUT2* gene suggested by earlier studies (21)) and, overall, showed concordance with the heritability of microbial taxa. Seven out of the nine taxa that showed the strongest evidence for heritability in the TwinsUK cohort ($P < 0.01$) also have genome-wide significant mbTLs (Figure 2b). For the taxa with genome-wide significant mbTLs, the number of independent loci associated with a relaxed threshold of 1×10^{-5} strongly correlated with heritability significance ($R = 0.62$, $P = 1.9 \times 10^{-4}$, Figure 2c), suggesting that more mbTLs would be identified for this group of bacteria using a larger sample size.

LCT mbQTL effect shows age and ethnic heterogeneity

The strongest association signal was seen for variants located in a large block of about 1.5Mb at 2q21.3, which includes the *LCT* gene and 12 other protein-coding genes. This locus has previously been associated with the abundance of *Bifidobacterium* in Dutch (7), UK (6) and US (22) cohorts. Previous studies have also shown a positive correlation of

Figure 3
Manhattan plot of the mbTL mapping meta-analysis results.

MbQTLs are indicated by letters. MbBTLs are indicated by numbers. For mbQTLs, the Spearman correlation test (two-sided) was used to identify loci that affect the covariate-adjusted abundance of bacterial taxa, excluding samples with zero abundance. For mbQTLs, p-values (two-sided) were calculated by logistic regression. Horizontal lines define nominal genome-wide significance ($P = 5 \times 10^{-8}$, red) and suggestive genome-wide ($P = 1 \times 10^{-5}$, blue) thresholds.

Figure 4
Association of the LCT locus (rs182549) with the genus *Bifidobacterium*.

(a) Forest plot of effect sizes of rs182549 and abundance of *Bifidobacterium*. Effect sizes and 95% CI are defined as circles and error bars. Effect sizes were calculated from Spearman correlation p-values (**Online Methods**). (b) Meta-regression of the association of mean cohort age and mbQTL effect size. Confidence bands represent the standard error of the meta-regression line. (c) Meta-regression analysis of the effect of linear, squared and cubic terms of age on mbQTL effect size. Confidence bands represent the standard error of the meta-regression line. (d) Age-dependence of mbQTL effect size in the GEM cohort. Blue boxes include samples in the age range 6–16 years old. Red boxes include samples with age ≥ 17 years. The C/C (rs182549) genotype is a proxy for the NC_000002.11:g.136608646= (rs4988235) allele, which is associated to functional recessive hypolactasia. The central line, box and whiskers represent the median, IQR and 1.5 times the IQR, respectively. See Supplementary Note for cohort abbreviations.

Bifidobacterium abundance with the intake of milk products, but only in individuals homozygous for the low-function LCT haplotype, thereby indicating that gene–diet interaction regulates *Bifidobacterium* abundance (7). In our study, the strongest association was seen for rs182549 ($P = 1.28 \times 10^{-20}$), which is a perfect proxy for the functional LCT variant rs4988235 ($r^2 = 0.996$, $D' = 1$ in European populations). This association showed evidence for heterogeneity across cohorts ($I^2 = 62.73\%$, Cochran's Q $P = 1.4 \times 10^{-4}$). A leave-one-out strategy showed that the COPSAC2010 cohort, which includes children 4–6 years of age range, contributed the most to the detected heterogeneity (**Figure 4a,b**; **Supplementary Table 2**). When this study was excluded from the meta-analysis, the heterogeneity was reduced ($I^2 = 51.9\%$, Cochran's Q $P = 0.004$). A meta-regression analysis showed that linear effects of age and ethnicity accounted for 11.84% of this heterogeneity. Including quadratic and cubic terms of age in the model explained 39.22% of the heterogeneity, and the residual heterogeneity was low (Cochran's Q $P = 0.01$) (**Figure 4c**).

Following these observations, we decided to investigate the effect of age and ethnicity in the multi-ethnic GEM cohort, comprising 1,243 individuals with an age range between 6 and 35 years, of which nearly half of the participants are 16 years or younger. Our analysis showed a significant SNP–age interaction on the level of *Bifidobacterium* abundance ($P < 0.05$, see **Online Methods**). Individuals homozygous for the NC_000002.11:g.136616754CC (rs182549) genotype showed a higher abundance of the genus *Bifidobacterium* in the adult group, but not in the younger group (**Figure 4d**). The age–genotype interaction was significant

in the GEM_v12 and GEM_ICHIP subcohorts, both comprising mostly European individuals, while the GEM_v24 cohort, mainly composed of individuals of different Israeli subethnicities (see **Online Methods**) who live in Israel, showed neither an mbQTL effect (Beta = -0.002 [95%CI: -0.21, 0.21]) nor an interaction with age ($P > 0.1$). The lack of an *LCT* mbQTL effect in adults was also observed in another Israel cohort in the study (Personalized Nutrition Project (PNP), 481 adults, Beta = -0.20 [95%CI: -0.61, 0.20]). Altogether, the cohorts that reported the lowest *LCT* effect sizes were the two cohorts of Israeli ethnicity volunteered in Israel (GEM_v24, PNP) and a child cohort (Copenhagen Prospective Studies on Asthma in Childhood (COPSAC), Beta = -0.18 [95%CI: -0.36, -0.01]).

mbTLs are enriched for genes related to metabolism

Several loci detected at genome-wide significance level were enriched for genes related to metabolism. In the mbQTL analysis, the *FUT2-FUT1* locus was associated to the abundance of the *Ruminococcus torques* genus group, a genus from the *Lachnospiraceae* family. The leading SNP (rs35866622 for *R. torques* group, $P = 2.21 \times 10^{-8}$) is a proxy for the functional variant rs601338 ($r^2 = 0.8$; $D' = 0.9$ in European populations) that introduces a stop-codon in *FUT2* (23). Another proxy of the functional *FUT2* SNP, rs281377, showed association to the *Ruminococcus gnavus* genus group in the binary analysis, however this signal was just above the genome-wide significance threshold ($P = 5.79 \times 10^{-8}$) (**Supplementary Table 9**). *FUT2* encodes the enzyme alpha-1,2-fucosyltransferase, which is responsible for the secretion of fucosylated mucus glycans in the gastrointestinal mucosa (24). Individuals homozygous for the stop-codon (rs601338*A/A, non-secretors) do not express ABO antigens on the intestinal mucosa. We observed that the tagging NC_000019.9:g.49218060C > T (rs35866622 non-secretor) allele was associated with a reduced abundance of the *R. torques* group and a decreased presence of the *R. gnavus* group. *Ruminococcus sp.* are specialized in the degradation of complex carbohydrates (25), thereby supporting a link between genetic variation in the *FUT2* gene, levels of mucus glycans and the abundance of this taxa. When assessing the link between this variant and phenotypes in the LifeLines-DEEP (LLD; $N = 875$) and Flemish Gut Flora Project (FGFP, $N = 2,259$) cohorts (**Online Methods**), the strongest correlation for the *R. torques* group was seen with fruit intake (LLD: $R_{sp} = -0.19$, $P_{adj} = 3.1 \times 10^{-5}$; FGFP: $R_{sp} = -0.10$, $P_{adj} = 1.4 \times 10^{-4}$, **Supplementary Table 10, 11**), in line with the association of *FUT2* with food preferences, as discussed in the results of the PheWAS (see below).

Several other suggestive mbQTLs can be linked to genes potentially involved in host-microbiome crosstalk. One of them includes three SNPs in 9q21 (top-SNP rs602075, $P = 3.57 \times 10^{-8}$) associated with abundance of *Allisonella*. The 9q21 locus includes the genes *PCSK5*, *RFK* and *GCNT1*, of which *RFK* encodes the enzyme that catalyzes the phosphorylation of riboflavin (vitamin B₂) and *GCNT1* encodes a glycosyltransferase involved in biosynthesis of mucin. These products play major roles in the host-microbiota interactions within the intestine, where they are used by bacteria for their metabolism and involved in the regulation of the host immune defense (26). Another association signal

10p13 (rs61841503, $P = 9.8 \times 10^{-9}$), which affects the abundance of the heritable family *Peptostreptococcaceae*, is located in the *CUBN* gene, the receptor for the complexes of cobalamin (vitamin B₁₂) with gastric intrinsic factor (the complex required for absorption of cobalamin). *CUBN* is expressed in the kidneys and the intestinal epithelium and is associated with B₁₂-deficient anemia and albuminuria (27). Cobalamin is required for host–microbial interactions (28), and supplementation with cobalamin induced a substantial shift in the microbiota composition of an *in vitro* colon model (29). These associations suggest that some members of the gut microbiome community might be affected by genetic variants that regulate the absorption and metabolism of vitamins B₂ and B₁₂.

Among mbBTLs, the strongest evidence for association was seen for a block of 10 SNPs (rs7574352, $P = 1.42 \times 10^{-9}$) associated with the family *Peptococcaceae*, a taxon negatively associated with stool levels of the gut inflammation markers chromogranin A (LLD: $R_{sp} = -0.31$, $P_{adj} = 4.4 \times 10^{-18}$, **Supplementary Table 10**) and calprotectin (LLD: $R_{sp} = -0.11$, $P_{adj} = 0.058$) and with ulcerative colitis (FGFP: $R_{sp} = -0.06$, $P_{adj} = 0.09$, **Supplementary Table 11**). The association block is located in the intergenic region in the proximity (220kb apart) of *IRF1*, which is involved in insulin resistance and susceptibility to type 2 diabetes (30). Other highlights of identified mbTLs are given in the **Supplementary Note**.

GSEA, FUMA and PheWAS analysis

To explore the potential functions of the identified mbTLs, we performed FUMA (Functional Mapping and Annotation of GWAS, Methods) (31), GSEA and PheWAS, followed by Bayesian colocalization analysis and genetic correlation of *Bifidobacterium* abundance to its PheWAS-related traits. FUMA of 20 mbQTL loci returned 139 positional and eQTL genes. GSEA on these genes suggested an enrichment for genes expressed in the small intestine (terminal ileum) and brain (*substantia nigra* and putamen basal ganglia) (**Supplementary Figure 4**). The positional candidates for mbBTLs did not show any enrichment in GSEA analysis.

To systematically assess the biological outcomes of the mbTLs, we looked up the 31 mbTLs in the summary statistics for 4,155 complex traits and diseases using the GWAS ATLAS (32). Five out of 31 leading SNPs were associated with one or more phenotypes at $P < 5 \times 10^{-8}$ (**Supplementary Table 12**): rs182549 (*LCT*) and rs35866622 (*FUT1/ FUT2*), followed by rs4428215 (*FNDC3B*), rs11647069 (*PM FBP1*) and rs9474033 (*PKHD1*).

The variant showing highest pleiotropy, rs182549 (*LCT*, *Bifidobacterium*), was associated with multiple dietary and metabolic phenotypes, and the causal involvement of the SNP across pairs of traits was confirmed by colocalization test ($PP.H4.abf > 0.9$) for 49 out of 51 tested phenotypes. The NC_000002.11:g.136616754= (rs182549) allele, which predisposes individuals to lactose intolerance, was negatively associated with obesity (33) and positively associated with Type 2 diabetes mellitus (T2DM) diagnosis ($OR = 1.057$ [95%CI:1.031, 1.085], $P = 1.74 \times 10^{-5}$), family history of T2DM (paternal: $OR = 1.054$ [95%CI:1.035, 1.073], $P = 1.41 \times 10^{-8}$;

maternal: OR = 1.035 [95%CI:1.016, 1.053], P = 0.0002, siblings: OR = 1.03 [95%CI:1.009, 1.052]), and several nutritional phenotypes in the UK Biobank cohort (32). Moreover, the functional *LCT*SNP rs4988235 variant is associated with 1,5-anhydroglucitol ($P = 4.23 \times 10^{-28}$) (34), an indicator of glycemic variability (35). There was a nominally significant correlation of *Bifidobacterium* with raw vegetable intake ($r = 0.36$, $P = 0.0016$), but this correlation was not statistically significant after correction for multiple testing.

NC_000019.9:g.49218060= (rs35866622, *FUT1/FUT2* locus) was positively associated with fish intake and height. The secretor allele was negatively associated with the risks of cholelithiasis and Crohn's disease, alcohol intake frequency, high cholesterol and waist-to-hip ratio (adjusted for body mass index (BMI), with PP.H4.abf > 0.9).

Consistent with the single SNP analysis, gene-based PheWAS also showed a strong link of the *LCT* locus with metabolic traits (e.g. $P = 5.7 \times 10^{-9}$ for BMI), whereas several nutritional (e.g. $P = 1.26 \times 10^{-20}$ for oily fish

Figure 5
Phenome-wide association study (PheWAS) domain enrichment analysis.

The analysis covered top-SNPs from 30 mbTLs and 20 phenotype domains. Three thresholds for multiple testing were used: 0.05, 8.3×10^{-5} (Bonferroni adjustment for number of phenotypes and genotypes studied) and 5×10^{-8} (an arbitrary genome-wide significance threshold). Only categories with at least one significant enrichment signal are shown.

intake), immune-related (e.g. $P = 1.73 \times 10^{-12}$ for mean platelet volume), gastrointestinal (e.g. $P = 8.77 \times 10^{-14}$ for cholelithiasis) and metabolic signals (e.g. $P = 1.13 \times 10^{-13}$, high cholesterol) mapped to the *FUT1/FUT2* locus (Figure 5, Supplementary Table 13).

Finally, we performed a phenotype domain enrichment analysis (Methods). We observed that top loci were enriched with signals associated with the metabolic domain supported by 4 mbTLs, followed by nutritional, cellular, immunological, psychiatric, ophthalmological, respiratory and reproductive traits, and the activities domain (Figure 5, Supplementary Table 14).

Mendelian Randomization analysis

To identify the potential causal links between gut microbial taxa and phenotypes, we performed bi-directional two-sample MR analyses using the TwoSampleMR package (36). We focused on two groups of phenotypes: diseases (autoimmune, cardiovascular, metabolic and psychiatric) and nutritional phenotypes (37–42). The complexity of the mechanisms by which host genetics affect microbiome composition, and the limited impact of genetic variants on microbial taxa variability, require caution when performing and interpreting causality estimation using MR analysis (43). We therefore carried out several sensitivity analyses and excluded any results that showed evidence of being confounded by pleiotropy (Methods). Only pairs supported by three or more SNPs were considered. With these strict cut-offs, no evidence for causal relationships between microbiome taxa and dietary preferences was identified (Supplementary Tables 15, 16). However, our results suggest that a higher abundance of the class *Actinobacteria* and its genus *Bifidobacterium* may have a protective effect on ulcerative colitis (Actinobacteria: OR = 0.56 [95%CI: 0.44–0.71] for each SD increase in bacterial abundance, $P_{\text{BHadj}} = 8.8 \times 10^{-4}$; Bifidobacterium: OR = 0.51 [95%CI: 0.39–0.71], $P_{\text{BHadj}} = 9.8 \times 10^{-5}$) (Figure 6a,b). We also observed that higher abundance of family Oxalobacteraceae has a protective effect on rheumatoid arthritis (OR = 0.82, [95%CI: 0.74–0.91], $P_{\text{BHadj}} = 0.028$, Figure 6c).

Figure 6
Mendelian randomization (MR) analysis.

The X-axes show the SNP-exposure effect and the Y-axes show the SNP-outcome effect (SEs denoted as segments). (a) MR analysis of class *Actinobacteria* (exposure) and ulcerative colitis (outcome). (b) MR analysis of genus *Bifidobacterium* (exposure) and ulcerative colitis (outcome). (c) MR analysis of family *Oxalobacteraceae* (exposure) and rheumatoid arthritis (outcome).

Discussion

We report here on the relationship between host genetics and gut microbiome composition in 18,340 individuals from 24 population-based cohorts of European, Hispanic, Middle Eastern, Asian and African ancestries. We have estimated the heritability of the human gut

microbiome and the effect of host genetics on the presence and abundance of individual microbial taxa. We studied the heterogeneity of the mbTL signals and characterized the impact of technical and biological factors on their effect magnitude. In addition, we explored the relevance of the identified mbTLs to health-related traits using GSEA, PheWAS and MR approaches.

Our large, multi-ethnic study allowed for an informative investigation of the human gut microbiome. However, there was large heterogeneity in the data, which reflects biological differences across the cohorts and methodological differences in the processing of samples. Overall, seven different methods of fecal DNA extraction and three different 16S rRNA regions were used (12,44). In addition, differences in the ethnicities, ages and BMIs of the participants led to a remarkable variation in microbiome richness, diversity and composition across cohorts. Diet, medication and lifestyle, among other factors (2,3), are known to influence the microbiome but were not included in our analysis because this data were not available for all cohorts. Large variation in the microbiome composition may have reduced the power of our mbTL analysis (see **Supplementary Note**).

We did not detect a host genetic effect on bacterial diversity, in line with a lack of its detectable heritability. Thirty-one taxon-specific mbTLs (20 mbQTLs and 11 mbBTLs) were identified at $P < 5 \times 10^{-8}$. Even with our large sample size, the number of mbTLs identified is rather modest. Only the association of *LCT* locus with *Bifidobacterium* ($P = 1.28 \times 10^{-20}$) passed the conservative study-wide significance threshold of $P > 1.95 \times 10^{-10}$. However, we observed that heritable taxa tend to have more genome-wide significant loci and suggestively associated loci, and twin-based heritability is significantly correlated with SNP-based heritability. Our results confirm that only a subset of gut bacteria is heritable, and that the genetic architecture affecting the abundance of heritable taxa is complex and polygenic.

The association between the *LCT* locus and the *Bifidobacterium* genus was the strongest in our study. It has been shown that the functional SNP in the *LCT* locus, rs4988235, determines not only the abundance of the *Bifidobacterium* genus, but also the strength of the association between this genus and milk/dairy product intake (7). Here, we showed the ethnic heterogeneity and age-dependent nature of the *LCT*-*Bifidobacterium* association – the effect is weaker in children and adolescents – consistent with existing knowledge on lactose intolerance (45,46). The strongest mbQTL effect was observed in the Hispanic Community Health Study/Study of Latinos (HCHS/SOL) cohort that comprises individuals of Hispanic/Latin American ethnicity and shows the highest prevalence of the lactose intolerant NC_000002.11:g.136616754CC (rs182549) genotype (683 out of 1,097 individuals).

To explore the potential functional effects of mbTLs on health-related traits, we used GSEA, PheWAS and MR approaches. The GSEA indicated enrichment of mbQTLs for genes expressed in the small intestine and brain. These results support the existence of the gut–brain axis mediated by microbiome and likely influencing gastrointestinal, brain and mood disorders (47–49). In addition, the PheWAS analysis identified a

significant overlap between the genetic variants affecting gut microbes and a broad range of host characteristics, including psychiatric, metabolic and immunological traits, and nutritional preferences, amongst other phenotype groups (**Supplementary Table 14**). Moreover, genetic determinants of bacterial abundance are involved in regulating host metabolism, particularly obesity-related traits. Among the interesting bacteria, earlier studies have linked the relative abundances of *Ruminococcus* (50), *Lachnospiraceae* (51) and *Ruminococcaceae* (52) to obesity. PheWAS analysis also indicated that SNPs from the LCT and FUT2 loci that associated with bacterial taxa are also associated to dietary preference factors, including fish, cereal, bread, alcohol, vegetable and ground coffee intake, along with other dietary phenotypes. Interestingly, other genes found to be associated with mbTLs also included olfactory receptors (*OR1F1*) and genes involved in the absorption and metabolism of B₂ and B₁₂ vitamins (*RFK* and *CUBN*).

Genetic anchors to microbiome variation also allow for estimation of causal links with complex traits through MR approaches (53–55). MR results indicate that Actinobacteria and *Bifidobacterium* might have a protective effect in ulcerative colitis. Cross-sectional studies have reported an increased abundance of Actinobacteria in healthy individuals as compared to inflammatory bowel disease patients (56,57), although these results have not always been consistent (58,59). *Bifidobacterium* was also previously shown to have a beneficial effect on ulcerative colitis in a clinical trial (58,60). We also revealed that abundance of family Oxalobacteraceae in the gut microbiome might be protective for rheumatoid arthritis; the abundance of this family in lung was previously shown to be negatively associated with rheumatoid arthritis (61). Protective effects of the bacterial taxa on these diseases support the potential of microbiome-based therapy. In summary, we report the largest study to date to investigate the genetics of human microbiome across multiple ethnicities. Microbiome heterogeneity and high inter-individual variability substantially reduces the statistical power of microbiome-wide analyses: similar to earlier microbiome GWAS studies, we report a limited number of associated loci. Nevertheless, our results point to causal relationships between specific loci, bacterial taxa and health-related traits. Heritability estimates suggest that these associations are likely part of a larger spectrum that is undetectable in the current study sample size. This warrants future studies that should take advantage of larger sample sizes, harmonized protocols and more advanced microbiome analysis methods, including metagenomics sequencing instead of 16S profiling and quantification of bacterial cell counts. Given the essential role of the gut microbiome in the metabolism of food and drugs, our results contribute to the development of personalized nutrition and medication strategies based on both host genomics and microbiome data.

Methods

Data collection

A total of 25 cohorts, comprising 18,340 participants of different ethnicities and ages, participated in the microbiome GWAS analysis (**Supplementary Tables 1, 2**). The **Supplementary Note** provides detailed descriptions of data collection per cohort.

16S microbiome data processing

The rationale behind the selection of the 16S rRNA processing pipeline was described previously (11). In short, the divergence in the 16S rRNA gene domains between cohorts makes operational taxonomic unit (OTU)-level analysis impossible, while the use of a direct taxonomic classification of the reads and an up-to-date reference database allowed us to achieve good between-domain concordance of taxonomic composition and a higher mapping rate.

The participating cohorts varied in their sample collection protocol, selection of DNA purification kits used to extract DNA from fecal samples, the 16S domain selected for PCR (**Supplementary Table 1**), read length, depth, post-sequencing quality control (QC) and the software used to merge tags of paired-end sequencing. After processing the QC-filtered merged reads, all cohorts implemented the standardized 16S processing pipeline (https://github.com/alexakur/miQTL_cookbook) that uses SILVA release 128 (15) as a reference database, with truncating the taxonomic resolution of the database to genus level.

Briefly, the procedure was as follows. First, all samples were rarefied to 10,000 reads using a predefined random seed to allow for rarefaction reproducibility. Samples with fewer than 10,000 reads were discarded. Second, RDP classifier v.2.12(16) was used to bin the reads to a reference database. For each taxonomic level, the posterior probability of 0.8 was used as a cutoff to bin each read to the corresponding taxon. The posterior cutoff probability was traced for each taxonomic level separately. For example, if the posterior probability passed the cutoff on family level but not on genus level, the read was binned to taxonomy on the family level (all corresponding upper taxonomic levels) and discarded on the genus level. It was also assigned to a special “NOTAX_genus” pseudo-taxon to maintain data compositionality.

To characterize the contribution of cohort-wise metadata (16S domain, DNA extraction method, cohort ethnicity, lysis temperature and type of lysis buffer) to the microbiome composition, we used a distance-based redundancy analysis test in which each cohort represented a sample and variables represented mean abundances of genera in the corresponding cohort (taxa with prevalence below 20% discarded). The association of metadata with richness was performed by multivariate linear regression analysis.

The alpha diversity indices, including Shannon, Simpson and inverse Simpson indices, were calculated on genus level with non-adjusted, non-transformed taxa counts. For all other analyses, the taxonomic counts of non-zero samples were natural log-transformed and adjusted for potential covariate effects using linear regression. The list of covariates used in the regression models varied between cohorts, but always included sex, age, genetic principal components (PCs) calculated on non-imputed genetic data (3 PCs for monoethnic cohorts, 10 PCs for multiethnic cohorts and 5 PCs for the HCHS/SOL cohort as a multi-ethnic population of different, but closely related ethnicities; see **Supplementary Note** for Cohort descriptions) and cohort-specific potential microbiome batch effects, if applicable. Variables such as the length of time in non-frozen storage, the 16S sequencing batch, etc. were also included. The residuals of the adjustment were then scaled and centered (mean = 0 and SD = 1).

In the analysis of microbiome composition heterogeneity, the cohorts SHIP/SHIP-TREND and GEM_HCE_v12/GEM_HCE_v24/GEM_HCE_ICHIP were merged to SHIP and GEM, respectively, because they were analyzed with exactly the same protocols in the same laboratories. In the microbiome-genetics analysis, these five cohorts were included individually as they differed in the genotyping arrays and/or general populations they represented.

For each cohort, only the taxa present in more than 10% of the samples were included in the quantitative microbiome trait loci (mbQTL) mapping, whereas taxa present in more than 10% but less than 90% of the samples were included in binary trait loci (mbBTL) mapping (**Supplementary Table 3**). Study-wide cutoffs for mbQTL mapping

included an effective sample size of at least 3,000 samples and presence in at least three cohorts. For mbBTLs, a mean abundance higher than 1% in the taxon-positive samples was required. This resulted in 211 taxa (131 genera, 35 families, 20 orders, 16 classes and 9 phyla) that passed taxon inclusion cutoffs for mbQTL analysis, and 177 taxa (108 genera, 34 families, 16 orders, 12 classes and 7 phyla) for mbBTL analysis.

Genetic data processing

Despite the difference in genotyping array platforms, most cohorts used similar procedures for imputation and post-imputation filtering steps. Twenty-three out of 24 cohorts used the Michigan Imputation Server (<https://imputationserver.sph.umich.edu/index.html>) for imputation, using the HRC 1.0 or 1.1 reference panel (62). Due to restrictions in manipulating data, the PNP study employed an in-house pipeline for imputation instead, using IMPUTE2 (63,64) software (v.2.3.2) and 1000G reference panel with addition of population-matched genotypes of Jewish individuals (65). The post-imputation cutoffs were the same for PNP and the other cohorts. Post-imputation VCFs were transformed into TriTyper format and filtered using GenotypeHarmonizer v.1.4.20 software (66). The following cutoffs were applied for inclusion: minor allele frequency > 0.05, pointwise imputation QC > 0.4 and SNP-wise call rate filtering > 0.95.

Heritability analysis

Heritability was calculated using data collected on 169 MZ and 419 DZ pairs of twins from the TwinsUK cohort (total of 1,176 individuals). Twin-based heritability was calculated by fitting an ACE model using the OpenMx package (v.2.8.3), as previously described (5). Prior to heritability estimation, the taxonomic abundance was normalized using inverse rank sum transformation. Since the NTR cohort comprised only MZ twins, the between-cohort heritability concordance was calculated as the correlation of intraclass correlation coefficient (ICC) for MZ twins. Pearson's correlation between NTR's and TwinsUK's ICCs was used to estimate the concordance. For mbQTLs, SNP-based heritability was calculated by LD score regression using 'LDSC' tool (67).

Microbiome GWAS analysis

The modified version of the eQTL mapping pipeline (<https://github.com/molgenis/systemsgenetics/tree/master/eqtl-mapping-pipeline>) was used to perform mbQTL mapping (68).

The microbiome GWAS was performed in three ways. First, we performed GWAS on three microbiome alpha diversity metrics (Shannon, Simpson and Inverse Simpson), using Spearman correlation between SNP dosages and alpha diversity metrics after adjustment for age, sex, technical covariates and genetic principal components. Second, we used Spearman correlation to identify loci that affect the covariate-adjusted abundance of bacterial taxa, excluding samples with zero abundance (mbQTLs).

Third, we identified the loci associated with probability of presence vs absence of the bacterial taxon (mbBTLs). To perform mbBTL analysis, we used a two-stage approach composed of fast correlation screening followed by logistic regression analysis as a robust method for binary traits GWAS (19). First, we calculated the Pearson correlation between SNP dosage and bacterial presence encoded as 0/1, without adjusting for any covariate effect and using the previously mentioned eQTL mapping pipeline, and used weighted Z-score meta-analysis to calculate non-centrality for SNP-taxon association. Finally, all SNP-taxon pairs with a first stage meta P-value < 1×10^{-4} were recalculated using multiple logistic regression (R base package, versions from 3.2.0 to 3.5.1 depending on the group) with bacterial presence as an outcome and using SNP dosage along with the list of covariates as predictors. All the mbBTLs that reached nominal genome-wide significance threshold ($P < 5 \times 10^{-8}$) in logistic regression analysis had a Pearson correlation P-value (at first stage) more significant than $P < 10^{-6}$, presuming the completeness of two-stage procedure in revealing genome-wide significant mbBTL using $P < 10^{-4}$ cutoff at the first stage of analysis.

mbTL meta-analysis

Meta-analysis was performed using a weighted Z-score method implemented in BinaryMetaAnalyzer (v.1.0.13B available on MiBioGen Cookbook), a part of the eQTL mapping pipeline that was used in large-scale eQTL meta-analyses (20,68). Per-cohort, Z-scores were calculated from Spearman correlation p-values using inverse normal transformation, transforming two-tailed p-values to one-tailed p-values and tracing the effect directions using the following formula:

$$\text{sign}(R_{\text{sp}}) * \text{qnorm}(1-P/2)$$

Where $\text{sign}(R_{\text{sp}})$ denotes the sign of Spearman correlation, qnorm denotes the quantile function for the normal distribution and P denotes the two-tailed p-value of Spearman correlation. For quantitative mbQTLs, the cohorts were weighted by the square root of the effective sample size (the number of samples having the bacterial taxon). For binary mbQTLs, the square root of the reported cohort size was used as a weighting for each study. The summary statistics generated for mbQTLs also include meta-effect sizes and standard errors. These were generated using the inverse variance weighted meta-analysis method performed on the per-cohort effect sizes and standard errors, backtracked from association Z-scores and minor allele frequencies using the strategy proposed and implemented by Zhu et al (69), where they also give the detailed derivation of the following equations:

$$b = z S$$

$$S = 1/\sqrt{2p(1-p)(n+z^2)}$$

Where, b is the estimated effect size, S is the estimated standard error, p is the allele frequency and n is the sample size.

Heterogeneity exploration analysis

Cross-study heterogeneity of the effects of genetic variants in the relative abundance of taxonomical units was assessed using Cochran's Q-test for heterogeneity (70), as implemented in METAL v2018-08-28 (71), for all genome-wide significant variants ($P < 5 \times 10^{-8}$) found in our main analysis. To avoid reporting false-positive associations due to different study designs or data collection methods, we used a stringent threshold of $P < 0.05$ to reject the null hypothesis of no heterogeneity. This threshold is conservative considering that several variants were tested simultaneously, and no correction for multiple testing was applied. When there was evidence of heterogeneity, a random effect model was also implemented at the meta-analysis level to confirm the association results, using the metaphor R package v.2.0-0 (<https://cran.r-project.org/web/packages/metafor/metafor.pdf>).

Additionally, when there was evidence for heterogeneity of a SNP-effect across cohorts, we implemented a meta-regression approach using the same package to assess whether variables such as age, ethnicity or sequenced region could explain the observed effect-size heterogeneity.

Analysis of SNP-age interaction analysis in the LCT locus

To discover whether the association of functional SNPs in the *LCT* locus to the abundance of the *Bifidobacterium* genus varied between groups of adults and infants, we performed age-SNP interaction analysis in the GEM cohort, which comprises three sub-cohorts that each have a comparable number of individuals above and below puberty age. The age of 17 years was selected to split the cohort into the age groups: adolescents or adults. Since the GEM cohort was composed of three sub-cohorts of different ethnic composition, we evaluated the interaction in both joint analysis and in each subcohort separately, using the following formula:

$$Bac = Sex + PC[1-3] + Age_{group} + Cohort + SNP_{dos} + SNP_{HZ} + SNP_{GT}:Age_{group}$$

where *Bac* is the log-transformed count of genus *Bifidobacterium*, *PC*[1-3] are three floats with the first 3 genetic PCs, *Cohort* is a batch variable that determines the cohort the sample belongs to, *SNP_{dos}* is a float-encoded dosage of alternative allele, *SNP_{HZ}* is a Boolean variable describing heterozygosity, *SNP_{GT}* is a genotype encoded as an unordered factor and *Age_{group}* is a two-level factor (above or below split level). The inclusion of a numeric dosage variable and a Boolean *SNP_{HZ}* variable allowed us to properly adjust for the recessive effect of the SNP on *Bifidobacterium* abundance without neglecting SNP imputation uncertainty as embedded in SNP dosage.

The analysis was then repeated for each GEM subcohort separately, using the same model.

Association of mbTL-associated taxa with host phenotypes

Bacterial taxa found to be significantly associated with genetic determinants were correlated with 207 host phenotypes, including the intrinsic host properties, diet, disease and medication information, in the LLD and FGFP cohorts. We used Spearman correlation with Benjamini-Hochberg (BH)-adjustment for multiple testing to assess the correlation between phenotypes and bacteria that had mbQTLs. For the taxa with mbQTLs, samples with zero abundance were truncated. For the taxa with mbBTLs, the abundance was transformed to a binary trait encoding presence/absence.

FUMA analyses of meta-analysis results

Functional mapping and annotation of 30 meta-analysis results were performed with FUMA (v1.3.5), an integrated web-based platform (31). Summary statistics from the mbQTL analyses for each of the 20 independent association signals were used in the analysis. Genome-wide significant loci and their boundaries were defined as non-overlapping genomic regions that extend across an LD window of $r^2 \geq 0.4$ (based on the 1000G EUR reference panel) (72) from the association signals with $P < 5.0 \times 10^{-8}$. Independent ($r^2 < 0.1$) lead SNPs from each locus were defined as those most strongly associated with a microbial trait (i.e. with the lowest P value) at the specific region. Multiple risk loci were merged into a single genomic locus if the distance between their LD blocks was < 250 kb.

Functional annotation of all candidate risk SNPs was obtained from different repositories integrated in FUMA. Furthermore, these functionally annotated SNPs were mapped to protein-coding genes using the following two strategies: i) positional mapping, with the maximum distance of 10 kb to protein-coding genes, and ii) eQTL mapping, using information from data repositories such as GTEx v7 and Blood eQTL browser (<http://genenetwork.nl/bloodeqtlbrowser/>) (20).

As the mbBTL mapping procedure provides accurate statistics for only the subset of SNPs (see Microbiome GWAS analyses paragraph), and we thus lack full summary statistics, we only performed positional mapping for mbBTLs, taking in the protein-coding genes within 10 kb distance of the 10 leading SNPs per trait.

All mapped protein-coding genes were combined into one list for either mbQTL or mbBTL analysis prior to performing GSEA integrated in FUMA. In further investigations, hypergeometric tests of enrichment of all mapped genes were performed not only in tissue-specific (differentially expressed) gene sets, but also in gene sets curated from various sources, e.g. MsigDB. We reported all enriched gene sets (≥ 2) with an FDR adjusted P-value < 0.05 .

PheWAS, genetic correlation and colocalization analysis

We performed the PheWAS look-ups in the summary statistics results of 4,155 traits collected by the GWASATLAS (32) (<http://atlas.ctglab.nl/>, accessed on: 25-09-2019) database for the top SNPs per mbQTL locus that were revealed by either mbQTL or mbBTL mapping. GWASATLAS includes 600 traits from the UK Biobank and is enriched with extensive phenotypes on proteomics ($n = 1124$ proteins), hematology ($n = 36$), metabolomics ($n = 1145$ metabolic features) and immune markers ($n = 241$), studied across variable sample sizes. It also contains 1,009 GWASs performed prior to the UK Biobank effort, all categorized under 27 phenotype domains. Next, we tested if any of these 27 domains were enriched by the phenotypes associated to one of the SNPs of interest (using a liberal P-value threshold of 0.05 for the SNP-phenotype association) as compared to the expected distributions under the null hypothesis. In order to obtain the distributions under the null hypothesis, we selected matching 1000 SNPs for each top SNP using SNPSNAP (73) matched by allele frequency, gene density, number of LD pairs and distance from the closest gene.

We then extracted corresponding results from the GWASATLAS for the matched 30,000 SNPs (1000 matching SNPs per each top mbTL SNP). The enrichment of each domain was tested by comparing the proportions of observed and expected significant results for the SNPs of interest using the `prop.test` function in R. This resulted in one-sided P-values and odds ratios. Seven domains (Aging, Body structures, Connective tissue, Ear-Nose-Throat, Infection, Muscular and Social Interactions) that included fewer than 20 GWAS tables were excluded from the enrichment tests, resulting in 20 domains. We used a conservative Bonferroni-based P-value threshold of 8.06×10^{-5} for the enrichment testing, accounting for 20 domains and a total of 30 mbTL top SNPs coming from both the mbQTL and mbBTL mapping. In addition, we performed gene-based PheWAS look-ups in the GWASATLAS for candidate genes of interest within 250 kb around the association peaks, as defined by the FUMA algorithms.

The genetic correlation between *Bifidobacterium* and its PheWAS-related traits (from **Supplementary Table 12**) was estimated following a LD-score regression approach (67) using the 'ldsc' tool. For testing colocalization of the PheWAS signals, we used the approximate Bayes factor approach as implemented by the "coloc.abf" function from the "coloc" library in R (74), using genetic variants within ± 250 kb around the top signals.

Mendelian Randomization analysis

MR analyses were performed in R using TwoSampleMR package (v.0.5.5) (36). Causality direction was tested between the microbiome and two data types: i) autoimmune, cardiovascular, metabolic (including weight-related phenotypes) and psychological diseases (GWAS summary statistics from MRBase (36)) known to be associated with microbiome composition (2,3,37–42,47) and ii) 42 nutritional phenotypes and alcohol intake frequency from the UK Biobank round 2 (<http://www.nealelab.is/uk-biobank/>).

For MR analyses, the combined meta-effects and standard errors from inverse variance meta-analysis were used. To test if a complex trait affected microbiome composition, we selected independent genetic variants associated with complex traits at the genome-wide significant level ($P < 5 \times 10^{-8}$) and used these as instruments in our MR analyses. For complex diseases, we transformed Odds Ratios (ORs) and C.I. to effect sizes and standard errors using the built-in function of the TwoSampleMR package. To test if microbiome changes were causally linked to complex traits, we first confined ourselves to bacteria with genome-wide significant QTLs. For these, we selected all SNPs with a less stringent cut-off of $P < 1 \times 10^{-5}$ in our MR analyses as instruments. This strategy was used to increase the number of SNPs available in order to perform sensitivity analyses, as shown previously (53). Independent SNPs were selected as instrumental variables based on $r^2 < 0.001$ in 1000G EUR data, within the TwoSampleMR package. When no shared SNPs were available between exposure and outcome, proxies from the 1000G EUR data ($r^2 > 0.8$) were added. We kept only the results based on at least three shared SNPs. MR causality tests were performed using the Wald ratio, and Wald ratios were meta-analyzed using the inverse-variance weighted (IVW) method (75). We also estimated the causality using additional methods: the weighted mode method (76), which provides an alternative approach to IVW; MR-Egger (77), which estimates

the degree of horizontal pleiotropy in the data; and MR PRESSO (78), which estimates the pleiotropy and corrects for it by removing outliers from the IVW model. We also assessed the heterogeneity of the results using Cochran's Q statistics (75) and using leave-one-out analyses (36). We estimated instrument variable (IV) strengths using F statistics: the amount of variance explained by IVs was calculated for each exposure using the TwoSampleMR package (get_r_from_lor function) for binary traits and VPE as defined in Shi *et al.* (79). F statistics were then calculated as $(r^2 * (N - 1 - k)) / ((1 - r^2) * k)$, where r^2 is the variance explained, N is the sample size and k is the number of IVs. We kept the results for the conventional threshold of F statistics > 10 (80).

After performing the MR tests, we excluded duplicated GWAS traits, as the same phenotype is often studied in multiple GWAS. To remove the duplicates, we kept the study with the largest sample size among all the tested GWAS studies for each trait.

After excluding duplicates and tests performed with weak instruments (F statistics < 10), we applied a BH correction for multiple testing to the results obtained from the IVW MR test, and subsequently used a stringent filtering procedure on the significant results to avoid false-positives. Specifically, we removed the MR results that were based on fewer than three SNPs and thus could not be further investigated with sensitivity analyses. We also removed the MR results that were not supported by other MR tests (weighted mode method $P > 0.05$, MR PRESSO $P > 0.05$) and those that showed substantial pleiotropy or heterogeneity as estimated by MR-Egger (MR-Egger intercept $P < 0.05$) or MR PRESSO outliers-adjusted test ($P > 0.05$), as well as those where leave-one-out analysis identified one SNP driving the signal (all but one leave-one-out configurations had $P < 0.05$). Of note, MR-Egger slope, which represents the causal estimate, was not used as a filtering step given the reduced power to detect causal effects. It is also worth noting that for all but one of the reported MR results that passed all the filters above, the MR-Egger slope p-value was greater than 0.05, therefore an MR-Egger intercept $P < 0.05$ cannot be used to exclude presence of pleiotropy. Even though many of our MR-Egger intercept results provided little evidence of directional pleiotropy, it is worth noting that a $P < 0.05$ cannot exclude the presence of pleiotropy and requires further understanding of the biological mechanisms underpinning the relationship between genetic variation, the gut microbiome and health outcomes. To exclude more complex causality scenarios, we also removed those results for which the reverse MR P-value was below 0.05. Of note, the causal relationship identified for the microbiome feature class Actinobacteria (as exposure) and ulcerative colitis (outcome) showed a consistent effect direction when using only the only genome-wide significant SNP, but with wider confidence interval (OR = 0.40 [95% CI: 0.22-0.71] $P_{\text{nominal}} = 0.002$).

Supplementary information

The Supplementary Material for this article can be found online at: <https://www.nature.com/articles/s41588-020-00763-1#Sec26>

Acknowledgements

We thank Jackie Senior and Kate McIntyre for editing the manuscript. The cohorts funding and acknowledgements information is given in the **Supplementary Note**.

Author contributions

A.K., A.Z., R.Kr., C.M.G., L.F. and J.R. conceived and designed the study. A.K., C.M.G., R.B., D.R. and J.W. were responsible for coordinating and performing meta-analysis. A.D., C.L.R., J.A.R.G., C.T.F., X.L., D.Z., M.J.B. lead the specific downstream analyses, and should be considered as shared second authors. Specifically, A.D. performed the PheWAS analysis, C.L.R. and C.T.F. performed the heritability analysis in TwinsUK and NTR cohorts, respectively, and J.A.R.G. performed the age-related analysis of LCT locus. X.L. ran and interpreted the FUMA analysis, D.Z. ran and interpreted the mendelian randomization analysis. M.J.B. substantially contributed to the development of the

analysis pipeline and protocols. RK, JR and AZ jointly supervised the project. A. vd G., A.C., H.J.W., Ur.V., M.J.B., S.S. and L.F. developed the pipeline for the meta-analysis and contributed to the methodology and statistical analysis. K.W. contributed to the PheWAS enrichment analysis. A.K., C.M.G., R.B., D.R., J.W., A.D., C.L.R., J.A.R.G., C.T.F., X.L., D.Z., M.J.B., M.D.A., S.S., R.Kr., J.R. and A.Z. wrote the manuscript, with contributions from all authors.

K.A.M, L.J.L and M.F collected and managed the CARDIA cohort. A.D.P, J.A.R.G., K.C., L.B. and W.T. collected and managed the GEM cohort. H.B., J.S., J.T., S.A.S. and S.J.S collected and managed the COPSAC study. D.B., O.P., T.H., T.J., and T.H.H. collected and managed the DanFunD study. D.A.H., G.F., J.R., J.W., K.H.W., M.J., N.J.T., R.Y.T., R.B. and S.V.S. collected, genotyped and managed the FGFP study. C.M.G, F.R., H.A.M., L.D. and V.W.V.J. collected and managed the Generation R study. H.N.K., H.S. and H.L.K. collected and managed the KSCS study. C.W., J.F., A.Z., L.F., S.S. and A.K. collected and managed the LLD cohort. A.J.L., E.O., K.L., M.Lk. and M.B. collected and managed the METSIM cohort. A.A.M.M., D.M.A.E.J., D.K. and Z.M. collected and managed the MIBS-CO cohort. H.P. and ZDW collected and managed the NGRC cohort. C.T.F., D.I.B., E.J.C.G., G.E.D., G.W. and R.G.J collected and managed the NTR cohort. Da.R., E.B., E.S. and O.W. collected and managed the PNP cohort. A.A., L.A., M.D.A., Su.W. and X.L. collected and managed the PopCol cohort. A.F., C.B., M.C.R., M.Ld and W.L collected and managed the BSPSPC and FOCUS cohorts. A.G.U., C.Mv.D, Dj.R. and R.Kr. collected and managed the RS cohort. F.F., F.U.W., G.H., H.V., M.M.L, St. W. and Uw.V. collected and managed the SHIP and TREND cohorts. L.Y.M., Q.Q., R.Kn., R.C.K. and R.D.B collected and managed the SOL cohort. C.I.L.R, C.J.S., J.T.B., M.A.J. and T.D.S. collected and managed the TwinsUK cohort. A.A.V. and J.S.T contributed to the discussion. All authors approved the final manuscript.

References

1. Gilbert, J. A. et al. Current understanding of the human microbiome. *Nat. Med.* 24, 392–400 (2018).
2. Zhernakova, A. et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* 352, 565–569 (2016).
3. Falony, G. et al. Population-level analysis of gut microbiome variation. *Science* 352, 560–564 (2016).
4. Rothschild, D. et al. Environment dominates over host genetics in shaping human gut microbiota. *Nature* 555, 210–215 (2018).
5. Goodrich, J. K. et al. Human Genetics Shape the Gut Microbiome. *Cell* 159, 789–799 (2014).
6. Goodrich, J. K. et al. Genetic Determinants of the Gut Microbiome in UK Twins. *Cell Host Microbe* 19, 731–743 (2016).
7. Bonder, M. J. et al. The effect of host genetics on the gut microbiome. *Nat. Genet.* 48, 1407–1412 (2016).
8. Wang, J. et al. Genome-wide association analysis identifies variation in vitamin D receptor and other host factors influencing the gut microbiota. *Nat. Genet.* 48, 1396–1406 (2016).
9. Turpin, W. et al. Association of host genome with intestinal microbial composition in a large healthy cohort. *Nat. Genet.* 48, 1413–1417 (2016).
10. Kurilshikov, A., Wijmenga, C., Fu, J. & Zhernakova, A. Host Genetics and Gut Microbiome: Challenges and Perspectives. *Trends Immunol.* 38, (2017).
11. Wang, J. et al. Meta-analysis of human genome-microbiome association studies: the MiBioGen consortium initiative. *Microbiome* 6, 101 (2018).
12. Sinha, R. et al. Assessment of variation in microbial community amplicon sequencing by the Microbiome Quality Control (MBQC) project consortium. *Nat. Biotechnol.* 35, 1077–1086 (2017).
13. Sinha, R., Abnet, C. C., White, O., Knight, R. & Huttenhower, C. The microbiome quality control project: baseline study design and future directions. *Genome Biol.* 16, 276 (2015).
14. Vandeputte, D., Tito, R. Y., Vanleeuwen, R., Falony, G. & Raes, J. Practical considerations for large-scale gut microbiome studies. *FEMS Microbiol. Rev.* 41, S154–S167 (2017).
15. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 41, D590–D596 (2012).
16. Cole, J. R. et al. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.* 37, D141–5 (2009).
17. Louis, P., Young, P., Holtrop, G. & Flint, H. J. Diversity of human colonic butyrate-producing bacteria revealed by analysis of the butyryl-CoA:acetate CoA-transferase gene. *Environ. Microbiol.* 12, 304–314 (2010).
18. Westra, H.-J. et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* 45, 1238–1243 (2013).
19. Wason, J. M. S. & Dudbridge, F. A general framework for two-stage analysis of genome-wide association studies and its application to case-control studies. *Am. J. Hum. Genet.* 90, 760–73 (2012).
20. Vosa, U. et al. Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. *bioRxiv* 447367 (2018). doi:10.1101/447367
21. Zhernakova, D. V et al. Individual variations in cardiovascular-disease-related protein levels are driven by genetics and gut microbiome. *Nat. Genet.* 50, 1524–1532 (2018).
22. Blekhnman, R. et al. Host genetic variation impacts microbiome composition across human body sites. *Genome Biol.* 16, 191 (2015).
23. Machiela, M. J. & Chanock, S. J. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* 31, 3555–7 (2015).
24. Kashyap, P. C. et al. Genetically dictated change in host mucus carbohydrate landscape exerts a diet-dependent effect on the gut microbiota. *Proc. Natl. Acad. Sci.* 110, 17059–17064 (2013).
25. Crost, E. H. et al. Mechanistic Insights Into the Cross-Feeding of *Ruminococcus gnavus* and *Ruminococcus bromii* on Host and Dietary Carbohydrates. *Front.*

- Microbiol. 9, 2558 (2018).
26. Yoshii, K., Hosomi, K., Sawane, K. & Kunisawa, J. Metabolism of Dietary and Microbial Vitamin B Family in the Regulation of Host Immunity. *Front. Nutr.* 6, (2019).
27. Haas, M. E. et al. Genetic Association of Albuminuria with Cardiometabolic Disease and Blood Pressure. *Am. J. Hum. Genet.* 103, 461–473 (2018).
28. Rowley, C. A. & Kendall, M. M. To B12 or not to B12: Five questions on the role of cobalamin in host-microbial interactions. *PLOS Pathog.* 15, e1007479 (2019).
29. Xu, Y. et al. Cobalamin (Vitamin B12) Induced a Shift in Microbial Composition and Metabolic Activity in an in vitro Colon Simulation. *Front. Microbiol.* 9, (2018).
30. Gysemans, C. et al. Interferon regulatory factor-1 is a key transcription factor in murine beta cells under immune attack. *Diabetologia* 52, 2374–2384 (2009).
31. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* 8, 1826 (2017).
32. Watanabe, K. et al. A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.* 51, 1339–1348 (2019).
33. Nicklas, T. A. et al. Self-perceived lactose intolerance results in lower intakes of calcium and dairy foods and is associated with hypertension and diabetes in adults. *Am. J. Clin. Nutr.* 94, 191–8 (2011).
34. Shin, S.-Y. et al. An atlas of genetic influences on human blood metabolites. *Nat. Genet.* 46, 543–550 (2014).
35. Suhre, K. et al. Metabolic Footprint of Diabetes: A Multiplatform Metabolomics Study in an Epidemiological Setting. *PLoS One* 5, e13953 (2010).
36. Hemani, G. et al. The MR-Base platform supports systematic causal inference across the human phenome. *Elife* 7, (2018).
37. Qin, J. et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* 490, 55–60 (2012).
38. Koeth, R. a et al. Intestinal microbiota metabolism of L-carnitine, a nutrient in red meat, promotes atherosclerosis. *Nat. Med.* 19, 576–85 (2013).
39. Coit, P. & Sawalha, A. H. The human microbiome in rheumatic autoimmune diseases: A comprehensive review. *Clin. Immunol.* 170, 70–79 (2016).
40. Vatanen, T. et al. Variation in Microbiome LPS Immunogenicity Contributes to Autoimmunity in Humans. *Cell* 165, 842–853 (2016).
41. O'Mahony, S. M., Clarke, G., Borre, Y. E., Dinan, T. G. & Cryan, J. F. Serotonin, tryptophan metabolism and the brain-gut-microbiome axis. *Behav. Brain Res.* 277, 32–48 (2015).
42. Karlsson, F. H. et al. Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* 498, 99–103 (2013).
43. Wade, K. H. & Hall, L. J. Improving causality in microbiome research: can human genetic epidemiology help? *Wellcome Open Res.* 4, 199 (2019).
44. Brooks, J. P. et al. The truth about metagenomics: quantifying and counteracting bias in 16S rRNA studies. *BMC Microbiol.* 15, 66 (2015).
45. Coluccia, E. et al. Congruency of Genetic Predisposition to Lactase Persistence and Lactose Breath Test. *Nutrients* 11, 1383 (2019).
46. Lapidis, R. A. & Savaiano, D. A. Gender, Age, Race and Lactose Intolerance: Is There Evidence to Support a Differential Symptom Response? A Scoping Review. *Nutrients* 10, (2018).
47. Valles-Colomer, M. et al. The neuroactive potential of the human gut microbiota in quality of life and depression. *Nat. Microbiol.* 4, (2019).
48. Lloyd-Price, J. et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* 569, 655–662 (2019).
49. Vich Vila, A. et al. Gut microbiota composition and functional changes in inflammatory bowel disease and irritable bowel syndrome. *Sci. Transl. Med.* 10, eaap8914 (2018).
50. Ottosson, F. et al. Connection Between BMI-Related Plasma Metabolite Profile and Gut Microbiota. *J. Clin. Endocrinol. Metab.* 103, 1491–1501 (2018).
51. Tun, H. M. et al. Roles of Birth Mode and Infant Gut Microbiota in Intergenerational Transmission of Overweight and Obesity From Mother to Offspring. *JAMA Pediatr.* 172, 368–377 (2018).
52. Finnicum, C. T. et al. Metataxonomic Analysis of Individuals at BMI Extremes

- and Monozygotic Twins Discordant for BMI. *Twin Res. Hum. Genet.* 21, 203–213 (2018).
53. Sanna, S. et al. Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. *Nat. Genet.* 51, 600–605 (2019).
  54. Jia, J. et al. Assessment of Causal Direction Between Gut Microbiota–Dependent Metabolites and Cardiometabolic Health: A Bidirectional Mendelian Randomization Analysis. *Diabetes* 68, 1747–1755 (2019).
  55. Yang, Q., Lin, S. L., Kwok, M. K., Leung, G. M. & Schooling, C. M. The Roles of 27 Genera of Human Gut Microbiota in Ischemic Heart Disease, Type 2 Diabetes Mellitus, and Their Risk Factors: A Mendelian Randomization Study. *Am. J. Epidemiol.* 187, 1916–1922 (2018).
  56. Rinninella, E. et al. What is the Healthy Gut Microbiota Composition? A Changing Ecosystem across Age, Environment, Diet, and Diseases. *Microorganisms* 7, 14 (2019).
  57. Plichta, D. R., Graham, D. B., Subramanian, S. & Xavier, R. J. Therapeutic Opportunities in Inflammatory Bowel Disease: Mechanistic Dissection of Host–Microbiome Relationships. *Cell* 178, 1041–1056 (2019).
  58. Frank, D. N. et al. Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases. *Proc. Natl. Acad. Sci.* 104, 13780–13785 (2007).
  59. Morgan, X. C. et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol.* 13, R79 (2012).
  60. Tursi, A. et al. Treatment of Relapsing Mild-to-Moderate Ulcerative Colitis With the Probiotic VSL#3 as Adjunctive to a Standard Pharmaceutical Treatment: A Double-Blind, Randomized, Placebo-Controlled Study. *Am. J. Gastroenterol.* 105, 2218–2227 (2010).
  61. Scher, J. U. et al. The lung microbiota in early rheumatoid arthritis and autoimmunity. *Microbiome* 4, 60 (2016).
  62. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* 48, 1279–1283 (2016).
  63. Howie, B., Marchini, J. & Stephens, M. Genotype Imputation with Thousands of Genomes. *G3* 1, 457–470 (2011).
  64. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* 44, 955–959 (2012).
  65. Carmi, S. et al. Sequencing an Ashkenazi reference panel supports population-targeted personal genomics and illuminates Jewish and European origins. *Nat. Commun.* 5, 4835 (2014).
  66. Deelen, P. et al. Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration. *BMC Res. Notes* 7, 901 (2014).
  67. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* 47, 1236–41 (2015).
  68. Westra, H.-J. et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* 45, 1238–1243 (2013).
  69. Zhu, Z. et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* (2016). doi:10.1038/ng.3538
  70. Cochran, W. G. The Combination of Estimates from Different Experiments. *Biometrics* 10, 101 (1954).
  71. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 26, 2190–2191 (2010).
  72. Auton, A. et al. A global reference for human genetic variation. *Nature* 526, 68–74 (2015).
  73. Pers, T. H., Timshel, P. & Hirschhorn, J. N. SNPsnap: a Web-based tool for identification and annotation of matched SNPs. *Bioinformatics* 31, 418–420 (2015).
  74. Giambartolomei, C. et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* 10, e1004383 (2014).
  75. Bowden, J. et al. A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Stat. Med.* 36, 1783–1802 (2017).
  76. Hartwig, F. P., Davey Smith, G. & Bowden, J. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *Int. J. Epidemiol.* 46, 1985–1998 (2017).

77. 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 (2015).
78. 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.* 50, 693–698 (2018).
79. Shim, H. et al. A Multivariate Genome-Wide Association Analysis of 10 LDL Subfractions, and Their Response to Statin Treatment, in 1868 Caucasians. *PLoS One* 10, e0120758 (2015).
80. Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian Randomization Analysis With Multiple Genetic Variants Using Summarized Data. *Genet. Epidemiol.* 37, 658–665 (2013).
81. Swertz, M. A. et al. The MOLGENIS toolkit: rapid prototyping of biosoftware at the push of a button. *BMC Bioinformatics* 11, S12 (2010)

E.A. Lopera-Mayat, A. Kurilshikov†, A. van der Graaft, S. Hut, S. Andreu-Sánchez, L. Chen, A. Vich Vila, R. Gacesa, T. Sinha, V. Collij, M.A.Y. Klaassen, L.A. Bolte, M.F. Brandao Gois, P.B.T. Neerincx, M.A. Swertz, LifeLines Cohort Study*, H.J.M. Harmsen, C. Wijmenga, J. Fu, R.K. Weersma, A. Zhernakova#, S. Sanna#

chapter 5

Effect of host genetics on the gut microbiome in 7,738 participants of the Dutch Microbiome Project

Host genetics are known to influence the gut microbiome, yet their role remains poorly understood. To robustly characterize these effects, we performed a genome-wide association study of 207 taxa and 205 pathways representing microbial composition and function within the Dutch Microbiome Project, a population cohort of 7,738 individuals from the northern Netherlands. Two robust, study-wide significant ($p < 1.89 \times 10^{-10}$) signals near the LCT and ABO genes were found to affect multiple microbial taxa and pathways, and were replicated in two independent cohorts. The LCT locus associations were modulated by lactose intake, while those at ABO reflected participant secretor status determined by FUT2 genotype. Eighteen other loci showed suggestive evidence ($p < 5 \times 10^{-8}$) of association with microbial taxa and pathways. At a more lenient threshold, the number of loci identified strongly correlated with trait heritability, suggesting that much larger sample sizes are needed to elucidate the remaining effects of host genetics on the gut microbiome.

Introduction

The human intestinal microbial community contains trillions of microorganisms that play an important role in maintaining normal gut function and immune homeostasis (1). Emerging evidence has shown that alterations in gut microbial composition are associated with the pathogenesis of many human diseases, including gastrointestinal disorders, metabolic syndrome, cardiovascular diseases and other conditions (2,3).

The gut microbiome is influenced by many factors, such as environmental factors including diet and medication usage (4,5), but also by host genetics. Heritability studies in twins and families have highlighted that human genetics contributes to gut microbial variation, showing heritability estimates ranging from 1.9% to 8.1% (6,7). This observation encouraged the first efforts to identify genomic loci that influence gut microbiota through genome-wide association studies (GWAS). These early gut microbiome GWAS identified several microbial quantitative trait loci (mbQTL) located in genes related to intestinal mucosal barrier, immune response and drug and food metabolism (8–11). However, reproducibility of these findings has been limited by differences in data processing methodologies, modest sample sizes and strong environmental effects, which, taken together, limit the detection of robust host genetic associations (12). A recent large-scale genome-wide meta-analysis of 24 cohorts replicated the association between *Bifidobacterium* abundance and the lactase (*LCT*) gene locus (13), which had previously only been reported in single-cohort studies (6,14). Other suggestive mbQTLs identified in this broad meta-analysis were proportional to heritability estimates from independent twin studies, indicating that additional loci found at lenient levels of significance are

*Adapted from: Lopera-Maya, et al.
Effect of host genetics on the gut
microbiome in 7,738 participants of the
Dutch Microbiome Project.
BioRxiv (2021)*

Accepted in Nature Genetics

*†: equal contribution first
*: a list of authors appears at the end of
this chapter
: equal contribution last

Figure 1
Genome-wide association scan results.

Manhattan plot of host genomic associations to bacterial taxa and bacterial pathway abundances with at least one genome-wide significant association ($p < 5 \times 10^{-8}$). Y-axis shows the $-\log_{10}$ transformation of the association p-value observed at each tested variant. X-axis shows the genomic position of variants. The thresholds of study-wide ($p = 1.89 \times 10^{-10}$) and genome-wide ($p = 5 \times 10^{-8}$) significance are shown as horizontal red and blue lines, respectively. Independent SNP-trait associations reaching genome-wide significance are listed in the tables and labelled on the Manhattan plot. The colors of associated hits indicate whether they represent an association with taxonomy (orange) or a bacterial pathway (blue).

likely to be real but that larger sample sizes are needed to reach sufficient statistical power (13). Nonetheless, meta-analyses of mbQTL studies still lack resolution due to the high levels of heterogeneity between cohorts, which can reduce power in meta-analysis. On top of this, many existing cohorts rely on 16S rRNA measurements, but this method does not allow for identification at bacterial species-level resolution or of bacterial pathway abundances. Indeed, measuring both species- and pathway-level abundances is essential for a further understanding of the function of an individual's microbiome because two bacterial species can have very different functions in the gut microbiome, yet pathways may be shared across distant microbial species and have roughly the same biological effect (15). mbQTL studies using shotgun metagenomic sequencing in large population cohorts are therefore needed to overcome the high variability in microbiome definition and reveal more robust and refined associations.

To enable a broader and deeper understanding of host-microbiota interactions, we used shotgun metagenomic sequencing on feces from 7,738 individuals of the Dutch Microbiome Project (DMP) (16) and matched their imputed genotypes to differences in taxa and pathway abundances. By comparing our results with summary statistics from other independent studies (13,17,18), we reveal novel host-microbiota interactions. Furthermore, we explore the impact of both intrinsic and environmental factors in confounding or modulating these genetic effects and identify diet-dependent host-microbiota interactions. We further assess the potential causal relationships between the gut microbiome and dietary habits, biomarkers and disease, using Mendelian randomization. Finally, we carry out a power analysis that shows how microbiome studies, even at the current sample size, are not well powered to reveal the complex genetic architecture by which host genetics can regulate the gut microbiome.

Results

Genome-wide association identifies known and novel associations with several bacterial taxonomies and bacterial pathways

We investigated 5.5 million common (minor allele frequency (MAF) > 0.05) genetic variants on all autosomes and the X chromosome, using linear mixed models (19) to test for additive genetic effects on 207 taxa and 205 bacterial pathways in 7,738 individuals from the DMP cohort (**Methods**) (19). Our quality control steps did not detect evidence for test statistic inflation (median genomic lambda 1.003 for taxa and 1.01 for pathways). We identified 33 independent SNP-trait associations at 20 loci at a genome-wide p -value threshold of 5×10^{-8} (**Figure 1, Supplementary Table 1**). Two loci passed the more stringent study-wide threshold of 1.89×10^{-10} that accounts for the number of independent tests performed (**Methods**). The other 18 loci were associated at p -values between 5×10^{-8} and 1.89×10^{-10} , with 4 associated to taxa and 14 to pathways. We refer to these 18 loci as “suggestive” and discuss them below.

The strongest signal was seen at rs182549 located in an intron of *MCM6*, a perfect proxy of rs4988235 ($r^2 = 1$, 1000 Genomes European populations), a variant known to regulate the *LCT* gene and responsible for lactase persistence in adults (ClinVar accession RCV000008124). The T allele of rs182549, which confers lactase persistence through a dominant model of inheritance, was found to decrease the abundance of species *Bifidobacterium adolescentis* ($p = 7.6 \times 10^{-14}$) and *Bifidobacterium longum* ($p = 3.2 \times 10^{-08}$) as well as those of higher-level taxa (phyla *Actinobacteria* $p = 2.3 \times 10^{-13}$; class *Actinobacteria*, $p = 2.3 \times 10^{-13}$; order *Bifidobacteriales*, $p = 5.9 \times 10^{-15}$; family *Bifidobacteriaceae*, $p = 5.86 \times 10^{-15}$; genus *Bifidobacterium*, $p = 8.59 \times 10^{-15}$). Associations at this locus were also seen for other taxa of the same genus, albeit at lower levels of significance (*Bifidobacterium catenulatum*, $p = 3.9 \times 10^{-5}$), and for species

Each node shows a taxonomic level (from outside to inside: phylogenetic group, phylum, class, order, family, genus and species). Note that branch lengths do not represent phylogenetic distance. Inner labels represent genetic locus. External labels represent the clade. Nodes with dotted line type indicate that the GWAS for that taxa was not performed. Node color corresponds to different levels of significance as described in the legend.

(a) depicts associations detected at the MCM6/LCT locus, using as the most significant p-value observed for rs4988235 and rs182549.

(b) depicts associations to the ABO locus, using the most significant p-value observed for rs8176645 and rs550052.

The second locus that passed study-wide significance consists of genetic variants near the *ABO* gene. *ABO* encodes the BGAT protein, a histo-

Figure 3
Association at the LCT locus and interaction with lactose intake.

Comparison of *f. Bifidobacteriaceae* relative abundance (a) between groups of Lactase Persistent (LP, rs182549 C/T or T/T) and Lactose Intolerant (LI, rs182549 C/C) participants and (b) stratified among individuals with low or high daily lactose intake levels. Lactose intake was corrected for daily calorie consumption. Relative abundance was natural log-transformed and adjusted by age and sex. Distribution is displayed with violin plots, where each inner box represents the first and third quartiles of the distribution and the middle line represents the median. Lactose intake levels were defined as low if < first quartile and high if $\geq$ first quartile. Lactose intake was only available for 5,801 of the 6,809 LP samples and for 376 of the 443 LI samples. P-values were obtained with a two-sided Wilcoxon rank test. Species in the same family shared similar distributions, although the difference in distribution within the LI group is not significant (Supplementary Figure 2).

blood group ABO system transferase. Associations found at this locus include species *B. bifidum* (rs8176645, $p = 5.5 \times 10^{-15}$) and *Collinsella aerofaciens* (rs550057, $p = 2.0 \times 10^{-8}$, $r^2 = 0.59$ with rs8176645 in 1000 Genomes Europeans) and higher order taxa (rs550057, genera *Collinsella*, $p = 9.3 \times 10^{-11}$; family Coriobacteriaceae, $p = 3.01 \times 10^{-9}$; order Coriobacteriales, $p = 3.03 \times 10^{-9}$) (Figure 2B). Interestingly, the metabolic pathway representing the bacterial degradation of lactose and galactose is also associated to the ABO locus (Metacyc ID: “LACTOSECAT-PWY: lactose and galactose degradation I”, rs507666, $p = 5.38 \times 10^{-15}$). The association at this locus with the genus *Collinsella* and with metabolic pathway LACTOSECAT-PWY have been recently described (18,20). To our knowledge, the associations with *B. bifidum* and *C. aerofaciens* are novel.

Association at *LCT* reflects the lactase persistence inheritance model and affects multiple taxonomic levels and bacterial pathways

Given that lactose tolerance is inherited in a dominant fashion, we tested the associations found in this locus using a dominant model for SNP rs182549 and thereby compared lactase persistent (LP) to lactose intolerant (LI) individuals. Indeed, all seven taxa associated to the *LCT* locus at genome-wide significance showed a stronger association signal when we used a dominant model for the alternative allele (all associations $p < 2 \times 10^{-27}$), with decreased taxa abundance in LP individuals (Supplementary Table 2). We further tested the other 200 taxa for this SNP using a dominant transmission model. Intriguingly, we also observed suggestive association ($p < 1 \times 10^{-4}$) with rs182549 for taxa that were associated to ABO locus in our GWAS analysis (genera *Collinsella* and species *B. bifidum* and *C. aerofaciens*)

and for the species *Roseburia inulinivorans* of the family Lachnospiraceae (**Supplementary Table 2**). With the exception of *Roseburia inulinivorans*, there was a consistent direction of effect across the associated taxa, with increased abundance in LI compared to LP individuals (**Figure 3**).

Furthermore, when analyzing LI and LP individuals separately in the bacterial pathways, rs182549 was also associated with an increased abundance of the “*LACTOSECAT-PWY*” bacterial pathway in LI individuals (effect = +0.300 in standard deviation (SD) units, standard error (SE) = 0.49, $p = 1.02 \times 10^{-9}$). This is not surprising given that the taxa in our dataset that were most highly correlated with this pathway are of the class Actinobacteria and species *B. adolescentis* (Spearman correlation of 0.73 and 0.69, respectively), which are both associated with SNPs near *LCT*.

Associations at *ABO* are dependent on secretor status

To further understand the mechanisms underlying the association signal at the *ABO* locus, we derived blood group types based on the genotype status of three genetic variants (see **Methods**). The majority of the individuals were either type A (40%) or type O (48%), reflecting estimates from previous studies (21). We observed that the associations at this locus were driven by individuals producing either A or B antigens compared to individuals with blood type O. The abundances of the genera *Collinsella* and the metabolic pathway “*LACTOSECAT PWY*” were higher in individuals producing A or AB antigens (Wilcoxon test for *Collinsella*, blood type O vs. A: $p = 2.8 \times 10^{-9}$, O vs. AB: $p = 0.11$; Wilcoxon test for metabolic pathway, O vs. A: $p = 5 \times 10^{-14}$, O vs. AB: $p = 0.038$). In contrast, the abundance of *B. bifidum* was lower compared to individuals with blood type O (Wilcoxon test blood type, A vs. O: $p = 2.3 \times 10^{-14}$, B vs. O: $p = 0.007$, AB vs. O: $p = 0.006$) (**Figure 4**). Secretion of the antigens produced by the A or B blood groups in body fluids, including the gut mucosa, is dependent on the individual’s secretor status. Secretor status is controlled by the fucosyltransferase 2 (*FUT2*) gene in a recessive manner and determined by nonsense mutation rs601338 (OMIM 182100)(22). Non-secretors carry the A/A genotype, which leads to a non-functional *FUT2* gene and inability to present ABO antigens on the intestinal mucosa. We observed that the associations of taxa and bacterial pathways with blood types were only present in secretors and absent in non-secretors (**Figure 4**). This indicates that the association at the *ABO* locus is driven by the exposure of blood antigens to gut bacteria. These results recapitulate recent evidence for an mbQTL effect of this locus on the *Collinsella* genus and on other bacteria reported for a Finnish population cohort (18). Associations of a functional variant in the *FUT2* gene to other bacterial taxa were observed in a recent meta-analysis, but none of these bacterial taxa or pathways showed significant association for this locus in our cohort (13).

Genetic associations are not confounded by intrinsic factors but may be modulated by diet

Gut microbiome composition and function is known to be affected by intrinsic factors such as sex or body-mass index (BMI) and by external

Figure 4. Association with blood types and interaction with FUT2.

Relative abundances of microbiome features found to be associated with SNPs at the ABO locus are compared between the inferred blood types (O, A, B and AB) with O as reference group (panels a, b and c) and between secretors and non-secretors stratified again by the blood types grouped by the presence (A/B/AB) or absence (O) of terminal sugars (panels d, e and f). The natural log-transformed relative abundance, adjusted by age and sex, and is shown using a violin plot in which the inner box represents the first and third quartiles of the distribution and the middle line represents the median. In the top panels, the distribution across all blood types is also significantly different for all features (ANOVA: $p = 6.7 \times 10^{-14}$ for species *B. Bifidum*, $p = 5.4 \times 10^{-9}$ for *LACTOSECAT-PWY* and $p = 1.4 \times 10^{-13}$ for genus *Collinsella*). P-values are obtained from a two-sided Wilcoxon ranked test. S.: secretors, N.s.: non-secretors, ns: $p > 0.05$, * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$, **** $p \leq 0.0001$.

factors such as diet and medication usage (4). We first explored the impact of BMI, medication usage and stool frequency and consistency. None of our genome-wide significant associations were attenuated when including these factors as covariates (**Methods**, all comparisons, Cochran's Q for difference in effect size $p > 0.05$), indicating that the genetic associations are independent and not confounded by these factors (**Supplementary Table 3a**). Finally, we did not find sex-specific associations for the 33 SNP-trait associations, although five associations did exhibit a smaller genetic effect in males compared to females (**Supplementary Table 3b**). We further investigated the effect of diet at the *LCT* and *ABO* loci, taking into account the dominant inheritance model at *LCT* and the dependence on secretor status at *ABO*. From 21 dietary factors derived from a food questionnaire, we considered those previously found to be associated ($FDR < 0.05$) with the microbial taxa and pathway abundances showing genome-wide association signals in these two loci (see **Methods**)(16). For the genus *Collinsella* and pathway *LACTOSECAT-PWY*, no dietary factors were found at this FDR threshold, we therefore considered the same diet factors associated to *B. bifidum* in the analyses, given that they showed similar patterns of genetic association. In an analysis that included age, gender, genetic and dietary factors, the dietary factors did not alter the effect of the genetic components, indicating that diet is not a source of bias in these genetic associations (**Supplementary Table 4**). Diet also remained an important factor after including genetics, with four taxa and one pathway associated at *LCT* and *ABO* being affected by at least one dietary factor ($p < 0.05$)(**Supplementary Table 4**), with the maximum of 16 factors found for *B. longum*. Among all the diet factors that showed

association with the microbiome features after accounting for the genetic signals (44 diet-microbiome pairs), we found evidence for a gene-diet interaction for one bacteria at the *LCT* locus: an increased abundance of Bifidobacteriaceae family in LI individuals who consumed larger amounts of lactose or dairy (interaction term $p=0.03$) (**Figure 4, Supplementary Table 5**), a finding consistent with previous reports (8,18). In contrast, there was no evidence for interaction with diet at the *ABO* locus for either the *B. bifidum* species, the *Collinsella* genus or the bacterial pathway “LACTOSECAT-PWY” (interaction term $p > 0.05$) (**Supplementary Table 5**).

Association signals reveal a polygenic architecture of microbiome taxa and pathways and replicate in other independent cohorts

Apart from the study-wide significant associations at the *LCT* and *ABO* gene regions, 18 other loci showed suggestive association at $p < 5 \times 10^{-8}$ (**Supplementary Table 1**), and none of these have been reported previously. The majority (14) were associated with bacterial pathways, which could not have been directly quantified in studies using 16S rRNA data, the technology predominantly used in microbiome genetic studies to date. These associations included several genomic loci located in genes associated with human immune or metabolic disorders and mainly affected microbial amino acids (3), vitamin (4), purine (3), TCA cycle (2) and enterobactin toxin (1) pathways. These SNP-pathway associations provide insights into the relationships between host genetics and gut microbial metabolic functions. For instance, SNP rs9927590 in 16q23.1 (in gene *WWOX*), which has been reported to be involved in obesity and stenosis in Crohn’s disease (23,24), was associated with a metabolic pathway involved in short chain fatty acid (SCFA) production (*PWY-5088*, $p = 9.2 \times 10^{-9}$). SCFAs have been shown to boost intestinal barrier functions and suppress metabolic diseases, including type 1 and type 2 diabetes (17,25). A variant (rs2714053) in 11q24.1 (in gene *GRAMD1B*) resulted in lower abundance of bacterial genes involved in the folic acid biosynthesis (*PWY-6147*, $p = 4.58 \times 10^{-8}$), and this gene was previously identified to be related to ulcerative colitis in a Polish cohort (26). Another example is the association near the type 2 diabetes-associated gene *TSC22D2*(27) (rs1584586, 3q25.1) with the pathway of purine processing (*PWY-841*, $p = 3.47 \times 10^{-8}$). Four suggestive associations for taxa were identified, including signals in 5q35.3 (in the *COL23A1* gene), 8p22 (in the *SGCZ* gene), 14q32.32 (in the *TNFAIP2* gene) and Xp21.1 (in the *DMD* gene).

We sought to replicate our genome-wide significant signals using summary statistics from other independent cohorts in which microbiome data was characterized using either 16S rRNA (MiBioGen study) (13) or metagenomic sequencing (LL-DEEP cohort) (17). In the MiBioGen study, a meta-analysis of 24 cohorts comprising up to 18,340 individuals, 16S rRNA measurements do not allow for the evaluation of abundance of bacterial species and pathways and the X chromosome was not analyzed. Consequently, only 10 of our 18 SNP-taxa pairs could be tested in MiBioGen, while none of the pathways were testable. In LL-DEEP, a

Microbiome taxa
 $r_s=0.248$, $P=3.1 \times 10^{-4}$

Microbiome pathways
 $r_s=0.263$, $P=1.3 \times 10^{-4}$

genome-wide microbiome association study on 952 individuals, we extracted information for the majority of the SNP-taxa pairs (14/18) and SNP-pathway associations (14/15). Unfortunately, the power to replicate the associations in LL-DEEP was limited due to the small sample size. In both studies, we observed clear replication of the study-wide significant loci, *LCT* and *ABO*. All 7 taxa associated with SNPs near *LCT* were replicated with consistent allelic effect directions (all $p < 3.7 \times 10^{-6}$). For the *ABO* locus, we found convincing replication significance for the *Collinsella* genus ($p < 2 \times 10^{-5}$ in MiBioGen) and replication at nominal significance for the metabolic pathway *LACTOSECAT-PWY* and *B. bifidum* species ($p < 0.05$ in LL-DEEP) (Supplementary Table 6 and Supplementary Table 7). None of the other SNP-taxa or pathway pairs were replicated in MiBioGen or LL-DEEP. Interestingly, another independent SNP in the *COL23A1* gene (rs11958296; $r^2 = 0.1$ with rs10447306 from our study), shows association in MiBioGen to the abundance of the same taxa: family Rikenellaceae ($p = 2.4 \times 10^{-5}$) and genus *Alistipes* ($p = 9.3 \times 10^{-6}$).

To explore if the association signals at lower levels of significance are enriched in heritable bacteria, which would indicate if it were possible to detect more genome-wide significant mbQTLs by further increasing the sample size, we investigated the correlation of taxa and pathway heritability estimations gained from family-based analysis with the number of suggestively associated loci for each taxon and pathway. Here we observed a positive significant correlation for both taxonomy (Spearman correlation $r_s = 0.322$, $p = 2.2 \times 10^{-6}$) and pathway ($r_s = 0.316$, $p = 3.8 \times 10^{-6}$) heritability and the number of suggestive ($p < 5 \times 10^{-4}$) loci identified in our GWAS (Figure 5). The correlation remained significant when increasing the mbQTL threshold to $p < 1 \times 10^{-4}$ (taxa: $r_s = 0.308$, $p = 6.1 \times 10^{-6}$; pathways: $r_s = 0.259$, $p = 1.7 \times 10^{-4}$) and to $p < 1 \times 10^{-5}$ (taxa: $r_s = 0.195$, $p = 4.8 \times 10^{-3}$; pathways: $r_s = 0.148$, $p = 0.034$) and when removing the *LCT* and *ABO* loci from analyses (taxa: $r_s = 0.314$, $p = 3.9 \times 10^{-6}$; pathways: $r_s = 0.311$, $p = 5.7 \times 10^{-6}$).

Figure 5
Weighted Spearman correlation between estimated heritability and number of suggestive loci.

Each point represents one taxon or pathway. Bold red line represents the locally estimated scatterplot smoothing fit, with orange shaded bands indicating the 95% regression confidence intervals. The number of independent suggestive mbQTLs was obtained using LD pruning (Methods). (a) Weighted Spearman correlation of number of suggestive mbQTLs and family-based heritability for microbiome taxa. (b) Weighted Spearman correlation of number of suggestive mbQTLs and family-based heritability for microbiome functional pathways.

We also evaluated if we could replicate any of the genome-wide significant signals outside the *LCT* and *ABO* loci reported in the very recent and similarly sized Finnish population study (18). After extracting all associations with $p < 1 \times 10^{-4}$ in our dataset, we identified 3 out of 451 genome-wide significant SNPs from the Finnish study using direct or proxy ($r^2 > 0.8$) information (279 of the SNPs reported in this study were not included in our dataset because their MAF was < 0.05). For one SNP, rs642387, we identified an association with consistent allelic effect for similar taxa: the minor allele at this variant (through proxy SNP rs632222) was associated with a decreased abundance of the *Desulfovibrio* genera ($p = 8.7 \times 10^{-5}$) in our study. This corroborates the allelic effects seen in the Finnish study for the abundances of phylum Desulfobacterota A, class Desulfovibrionia, order Desulfovibrionales and family Desulfovibrionaceae. For these taxa, we observed consistent direction in our dataset, although the signals were not significant ($p = 0.14$) (**Supplementary Table 8**).

Food preference and cardio-metabolic phenotypes link to gut microbiome genetic associations

To investigate the causal relationships between microbiome composition/function and complex traits, and *vice-versa*, we used publicly available summary statistics in conjunction with our mbQTL results to perform Mendelian randomization (MR) analyses (see **Methods**). We focused on 73 phenotypes of autoimmune disease, cardiometabolic disease traits and food preferences (**Supplementary Table 9**) and on 33 microbiome features that were associated with at least one variant at genome-wide significance ($p < 5 \times 10^{-8}$). None were significant at $\text{FDR} < 0.05$. At $\text{FDR} < 0.1$ (corresponding to a $p = 2.7 \times 10^{-5}$ obtained from the inverse-variance weighted (IVW) MR test), we observed one causal relationship in the direction from microbiome to phenotypes. A genetic predisposition to higher abundance of family Rikenellaceae was causally linked to lower consumption of salt ('Salt added to food' ordinal questionnaire phenotype, $\text{IVW } p = 2.4 \times 10^{-5}$, causal effect = -0.05 in SDs of salt intake for each SD increase in Rikenellaceae abundance) based on three genetic instruments on different chromosomes (**Supplementary Table 10**) (**Supplementary Figure 2a**). This correlation was not significant in the opposite direction (IVW $p = 0.32$), suggesting that Rikenellaceae abundance may have an effect on the increased consumption of salt without being tagged by a shared causal factor. The causal estimate was consistent when using a different MR test (weighted median $p = 0.00129$, causal effect = -0.0554 (SD Salt intake)/(Rikenellaceae abundance SD)). Furthermore, there was no evidence for the causal effect being affected by pleiotropy (Egger intercept $p = 0.75$) or dependent on a single variant (highest p -value in a leave-one-out analysis: 8.8×10^{-4}) (**Supplementary Figure 1**). At a lower threshold ($\text{FDR} < 0.2$), we observed putative causal effects of two other microbial features on phenotypes: a negative effect of genera *Collinsella* abundance on the level of triglycerides ($\text{FDR} = 0.10$) and a positive effect of the histidine degradation pathway (MetaCyc: "*HISDEG-PWY: L-histidine degradation I*") on intake of processed meat ($\text{FDR} = 0.12$) (**Supplementary Figure 1**). While all these potential causal effects are intriguing and exhibit

robustness to sensitivity analyses, we recognize that the limited availability of genetic variants strongly associated with microbiome features dictates caution in interpretation.

Power analysis indicates that larger sample sizes are necessary to identify host genetic effects for microbial features with low prevalence

One of the complicating factors in any genome-wide association study of the gut microbiome is the variable detection rate of the pathways and taxa measured, which reduces the effective sample size and consequently statistical power. In our study, the additive effect of associated variants at *LCT* explain 0.8% of the variance for species *B. longum* and *B. adolescentis*, which are present in >80% of the samples, and this allowed us to have ~70% power to detect the association at $p < 1 \times 10^{-10}$ (Figure 6). In contrast, we had no power (<5%) to detect the association for species *B. bifidum* and *B. catenulatum* at this level of significance, as they were present in only 14% and 26% of samples, respectively. For *B. bifidum*, despite being a rare species, we were able to detect a study-wide significant association at *ABO* because the effect was substantial (2.5% of the variance explained by SNP rs8176645, power = 83%), but we were underpowered to see the association at this locus with species *C. aerofaciens*, which was present in only 45% of the samples, and for which the additive effect of *ABO* variants was smaller (0.5% variance explained) (Figure 6). In a power analysis, we estimated that our sample size (7,738 individuals) is underpowered to detect genetic effects for taxa that are present in < 80% of samples when considering an effect size comparable or smaller to the effect of *LCT* for *B. adolescentis*. This is striking, considering that 93% of

Figure 6
Power analysis taking into account bacterial presence levels

(a) Statistical power (at $\alpha = 1 \times 10^{-10}$) in our dataset for SNPs at the *ABO* and *LCT* loci, considering the associated taxonomies. 80% detection power is shown by the red horizontal line. (b) Power (y-axis) to detect an effect at significance level $\alpha = 1 \times 10^{-10}$ depending on the variance explained by a genetic variant (x-axis). Colored lines distinguish power under different levels of bacterial or pathway prevalence, which reduces effective sample size. (c) Sample size needed to identify an association at 80% power level across different bacterial presence levels and $\alpha = 1 \times 10^{-10}$. Horizontal dashed line indicates the sample size of the DMP cohort.

the taxa we identified in this cohort are present in < 80% of the samples (**Methods**). We estimate that to find a similar effect size for at least 20% of gut microbiome composition (approximately 20% of the taxa are present in at least 20% of the samples in our cohort) would require a cohort of ~50,000 individuals.

Discussion

We carried out the largest genome-wide association study of the gut microbiome composition and function in a single population by analyzing metagenomic sequencing data in 7,738 volunteers from the northern Netherlands. This recapitulated genetic associations at two known loci, *LCT* and *ABO*, and highlighted novel associations with species and bacterial pathways in these loci. Furthermore, we identified associations ($p < 5 \times 10^{-8}$) at 18 other loci for 4 taxa and 14 bacterial pathways. None of these associations were affected by major confounders of the gut microbiome such as drug usage, diet and BMI. Finally, we used an MR approach to pinpoint potential causal links between gut microbiome composition, complex traits and food intake habits.

The strongest association with the gut microbiome was found at the *LCT* locus, which remains the most robust genetic association with gut microbiome identified to date. Associations at this locus with *Bifidobacteria* have been consistently reported in studies of different ethnicities, across a range of sample sizes, and in studies using different technologies and protocols for gut microbiome characterization (6,8,13,14,28). We also confirmed that the increase of *Bifidobacterium* was more evident in LI individuals who were consuming milk or milk-derived products (13,29). In addition, given that the resolution of metagenomic sequencing allows for species-level characterization of microbiome profiles, we could show that this effect was mainly attributable to species *B. longum*, *B. adolescentis*, *B. catenulatum* and *B. bifidum*, an observation that was corroborated by two very recent studies in Finnish and US Hispanic/Latino populations (18,30).

We found another convincing association in the *ABO* gene locus that was responsible for changing abundance in several taxa and pathways. Associations with microbiome composition at the *ABO* locus have been reported previously for populations of different ethnicities and in non-human species. For example, associations with *Bacteroides* and *Faecalibacterium* were reported in a study of five German cohorts that used 16S rRNA sequencing for gut microbiome characterization (28) and with the microbial pathway of lactose and galactose degradation in a cohort of 3,432 Chinese individuals (31). The importance of *ABO* in determining host-microbiome interaction has also been recently reported in pigs (32). A deletion at this locus that inactivates the ABO acetyl-galactosaminyl-transferase has been shown to change the porcine microbiome composition by altering intestinal N-acetyl-galactosamine concentrations and consequently reducing the abundance of *Erysipelotrichaceae* strains, which have the capacity to import and catabolize N-acetyl-galactosamine (33). In our analysis, the strongest association was with *B. bifidum*, which

has not been reported before, and with the *Collinsella* genus. We did not detect any evidence of interaction with diet at this locus, although this could be due to limitations in available information as the recording of dietary information was done at different times than stool collection time. We did, however, confirm that association at this locus depends on secretor status determined by the *FUT2* gene (28) and thus on the ability to incorporate antigens into bodily fluids that are released in the gut. Intriguingly, we observed that taxa associated with the *ABO* locus also showed evidence of association at the *LCT* locus, indicating a common action of these two loci in contributing to the growth of these bacteria. The most compelling hypothesis is that the availability of sugars in the gut, via undigested lactose in LI individuals or secretion of glycans in non-O blood type secretors, provides direct energy sources for these bacteria. This is further supported by the observation that LI individuals and non-O blood type secretors were both associated with increased abundance of a bacterial pathway for lactose and galactose degradation. This mechanism would not explain however the opposite pattern of association at ABO for *B. bifidum*.

The strongest mbQTLs we identified reside in genes under selective pressure. The *LCT* gene is a highly differentiated gene among human populations due to positive selection conferred by the lactase persistence phenotype. It has been estimated that strong selection occurred within the past 5,000–10,000 years, consistent with there being an advantage to lactase persistence and the ability to digest milk in the setting of dairy farming (34). Variants at this locus have been linked through GWAS to not only food habits and metabolic phenotypes, but also to immune cell populations (35). The *ABO* locus is evolutionarily highly differentiated; it has been shown to have experienced balancing selection for the last 3 million years in many primate species (36). Several evolutionary sources of selective pressure have been proposed, including via infections by pathogens such as malaria (37) and cholera (38). *ABO* variants have also been linked to cardiometabolic traits, white and red blood cells and cytokine levels (39,40). Therefore, host-microbiome interactions are likely to be shaped by human-microbe co-evolution and survival, probably through a balance between food availability for gut bacteria and enhanced immune response of the host.

From this perspective, it will be crucial to compare genetic studies of the gut microbiome in diverse populations with different genetic backgrounds in order to understand the complexity of host-microbe interactions. This requires community efforts to standardize definition of taxonomies and measurement methodologies in order to facilitate comparison between cohorts. For example, in our attempt to replicate the findings from the Finnish population cohort, only a limited number of taxa could be directly matched or connected through the Genome Taxonomy Database (41).

To explore the causality of the relations of the microbiome with complex traits and food preferences, we performed bi-directional MR analysis using 56 dietary traits, 16 diseases and 5 biomarkers. At $FDR < 0.2$, we observed three causal relationships. Although the limited impact of genetic variants on both microbiome composition and dietary preferences requires caution when interpreting causality estimation by MR (42,43), it is intriguing that

two observations indicate the causal role of microbial composition on food preferences. We observed that an increase in Rikenellaceae abundance led to decreased consumption of salt and an increase in the bacterial pathway of histidine degradation led to increased intake of processed meat. Although multiple studies have shown that dietary changes, such as variations in salt and meat intake, have a strong effect on microbiome composition (44), it is intriguing to suggest that genetically determined variations in the microbiome might affect individual food preferences. This would be supported by bacterial genetic variations in salt tolerance and carnitine metabolism genes (45) and by the established knowledge that the composition of gut microbiome can accurately predict the effect of food items on host metabolism (46). There was additional evidence from our results to support a role for the microbiome in influencing food preferences. A perfect proxy of rs642387 (rs503397, $r^2 = 0.99$ in 1000 Genomes Europeans) that was recently reported to be associated with the family Desulfovibrionaceae and related taxa in a Finnish population, and replicated in our study, had been previously associated to bitter alcoholic beverage consumption in an independent cohort (47). Another study found an increase in family Desulfovibrionaceae in individuals with high alcohol consumption (48), and family Desulfovibrionaceae, genus *Desulfovibrio* and other related taxa were also associated with increased consumption of alcohol in our DMP cohort (16), supporting a pleiotropic effect of this locus on both microbiome and alcohol intake. These studies, together with our findings, suggest that gut microbiota could indirectly influence an individual's food preferences by mediating the downstream effect of consumption of different products.

Understanding the differences in microbial pathway abundance is crucial for deeply understanding the underlying function of the gut microbiome. Microbiome measurements that target single organisms or taxonomic groups may be inefficient because two strains of one species may differ in their functional capacities. It is therefore essential to evaluate not only the taxonomic but also the functional composition of the microbiome. To our knowledge, ours is the largest single study investigating host genetics effects on the human gut microbiome function using metagenomics sequencing, and thus perform genome-wide association analysis of microbial pathways. Our analyses revealed several loci associated with microbial pathways, although only one passed stringent study-wide significance.

We also observed 18 loci at genome-wide significance and many more at a more lenient threshold. The correlation we observe between the heritability of microbial taxa and pathways and the number of suggestively associated loci indicates that mbQTLs with smaller effects are likely to exist. Those loci remained under the detection limit in the current study, and the sample size would need to be increased by orders of magnitude to discover their role. In fact, gut microbiome composition is characterized by high inter-individual variation. Analysis of core microbiota in different populations indicates that only a few bacteria are present in >95% of studied individuals (5,13,16), drastically reducing the effective sample sizes for analyses. We performed power calculations taking into account the inter-individual variations in microbiome composition and concluded

that a sample size comparable to our study (~8,000 participants) is only sufficient to identify associations to taxa present in >80% individuals, which comprise only 7% of all identified taxa in the cohort. For bacteria present in >20% of the samples, >50,000 participants would be necessary to identify an effect size similar to that of *LCT* and *ABO*. Much larger cohorts are required to identify genetic contribution to rarer bacteria. We conclude that joint efforts that combine tens of thousands of individuals, combined with harmonized methodology to reduce the technical bias, will be needed to characterize more than a few major loci, as has also been the case for genetic studies of much more heritable quantitative traits such as BMI (heritability ~40%), height (heritability ~80%) and others (49–51).

Methods

Cohort description

Lifelines is a multi-disciplinary prospective population-based cohort study with a unique three-generation design that is examining the health and health-related behaviors of 167,729 people living in the north of the Netherlands. Lifelines employs a broad range of investigative procedures to assess the biomedical, socio-demographic, behavioral, physical and psychological factors that contribute to the health and disease of the general population, with a special focus on multi-morbidity and complex genetics (52,53). During the first follow-up visit, a subset of 8,208 Lifelines participants were enrolled in a parallel project, The Dutch Microbiome Project (DMP). The goal of this cohort is to evaluate the impact of different exposures and life-styles on gut microbiota composition (16).

The Lifelines study was approved by the medical ethical committee from the University Medical Center Groningen (METc number: 2017/152). Additional written consents were signed by the DMP participants or legal representatives for children aged under 18.

Genome characterization

Genotyping of 38,030 Lifelines participants was carried out using the Infinium Global Screening array® (GSA) MultiEthnic Diseases version, following the manufacturer's protocols, at the Rotterdam Genotyping Center and the department of Genetics of the University Medical Center Groningen. Here, we used available quality controlled genotyping data imputed with Haplotype Reference Consortium (HRC) panel v1.1(54), as described elsewhere (55). For this study, we analyzed European samples, identified through a Principal Component Analysis (PCA) that projected 43,587 genetic markers previously pruned from our genotyping data (sliding window of 1Mb, linkage disequilibrium $r^2 < 0.2$, step = 5) onto the 1000 Genomes (all populations) and GoNL cohort as a population reference (56,57). Europeans were identified as samples clustering together with European populations according to the first two principal components (<3 standard deviations from any of the 1000G European samples or GoNL samples), and 35 non-European participants were

removed. Quality-controlled genotype information was obtained for 7,738 of the DMP participants for whom quality-controlled microbiome data and BMI were also available. Of these, 58.1% were females and had ages ranging from 8 to 84 (mean of 48.5 years). The mean BMI value was 25.58 (range from 13.10 to 63.70).

Microbiome characterization

The gut microbiome was characterized from stool samples as described in Gacesa et al.(16): In brief, stool samples were collected by participants and frozen within 15 minutes after production and then transported on dry ice to the Lifelines facility to be stored at -80°C. Microbial DNA was extracted using the QIAamp Fast DNA Stool Mini Kit (Qiagen, Germany), following the manufacturer's instructions. Shotgun metagenomic sequencing was carried out using the Illumina HiSeq 2000 platform at Novogene, China. Metagenomic sequencing data was profiled following methods previously used in other cohorts, as described in (16,58). Low quality reads (PHRED quality ≤ 30), adapters and host sequences were removed using KneadData tools v0.5.1. Taxonomic composition was determined with MetaPhlAn2 v2.7.2 (59). Characterization of biochemical pathways was performed with HUMAnN2 pipeline v0.11.1(60), integrated with the UniRef90 v0.1.1 protein database (61), the ChocoPhlAn pangenome database and the DIAMOND alignment tool v0.8.22 (62). After quality control (samples with eukaryotic or viral abundance $\leq 25\%$ and total read depth ≥ 10 million were retained), we had information on 950 microbial taxa and 559 functional pathways. For this study, we focused only on bacterial taxa and functional pathways with mean relative abundance $> 0.001\%$ that are present in at least 1000 of the 7,739 participants, which resulted in a list of 207 taxonomies (5 phyla, 10 classes, 13 orders, 26 families, 48 genera and 105 species) and 328 bacterial pathways. Furthermore, we removed redundant pathways by discarding one pathway among pairs that were highly correlated (Spearman $r_2 > 0.95$), as well as pathways not previously described in bacteria that could thus be coming from sources other than bacteria, resulting in 205 pathways for genetic analyses.

Diet phenotypes definition

Dietary habits were assessed using a semi-quantitative Food Frequency Questionnaire (FFQ) designed and validated by the division of Human Nutrition of Wageningen University as described before in Siebelink et al and Gacesa et al.(16,63). The FFQ data was collected 4 years prior to fecal sampling, and supplementary questionnaires were collected concurrent with fecal sampling, with the stability of long-term dietary habits was assessed as described in Gacesa et al.(16). We analyzed the dietary factors that were previously found to be associated (False discovery rate (FDR) < 0.05) to the microbiome features in our study that had a genome-wide significant signal in the *ABO* and *LCT* loci. We also analyzed lactose intake for the species *Bifidobacterium longum* and *Bifidobacterium adolescentis*, given their association with the *LCT* region in our study. Participants with an implausible caloric intake (< 800 or > 3934 kcal/day for males, < 500 or > 2906 kcal/day for females)(64) were not included in these analyses.

GWAS analysis method

Genome-wide association analysis was performed in 7,738 European samples for 412 features (205 functional pathways and 207 microbial taxa), investigating genetic additive effects using allele dosages for 5,584,686 genetic variants with minor allele frequency (MAF) > 0.05 and information score > 0.4 on the autosomes (chromosomes 1-22) and the X chromosome. We focused on the quantitative dimensions of relative bacterial abundance, treating all zero values as missing data. We used log-transformed taxonomies and regressed these in a linear mixed model using SAIGE v.0.38(19), with age, sex and the genetic relationship matrix (GRM) among participants as covariates. We used the standard settings of SAIGE, which applies inverse-rank normalization to the traits prior to the association analyses. The genetic relationship matrix was built using a set of 54,565 SNPs selected from the total set of quality-controlled SNPs directly genotyped and filtered for allele frequency and redundancy (MAF $\geq$ 0.05, $r^2 < 0.2$, sliding window = 500Kb). For pathways, we first screened associations with SAIGEgds, an implementation of SAIGE in a computationally efficient framework that allows faster analyses by using best-guess genotypes instead of dosages (65). We then recalculated genome-wide associations statistics using SAIGE v.0.38 for the 60 pathways that showed at least one genetic variant at $p < 1 \times 10^{-7}$.

Definition of the study-wide significant threshold

To estimate the number of independent phenotypes assessed, we used principal component analysis on the matrix of 412 microbiome features (207 taxonomies and 205 pathways) available for GWAS analysis to decompose variability in independent components (axes). We estimated that 264 components are needed to explain 90% of the microbiome variance. We then defined our study-wide significant p -value threshold by correcting the genome-wide significance threshold for this factor ($5 \times 10^{-8} / 264 = 1.89 \times 10^{-10}$).

Association using dominant model

To evaluate association using a dominant model on SNP rs182549 at the *LCT* locus, we used best-guess genotypes and converted T/C to T/T. Association analysis was then run for all taxa as well as the “*LACTOSECAT-PWY*” pathway using SAIGEgds, and the same covariates and transformation used for the GWAS analysis.

Inference of blood groups

We estimated blood groups from genotyped and imputed data following the scheme of Franke et al, 2020(66). Specifically, we used the absence of the rs8176719 insertion to define blood-type allele O1, the T allele of rs41302905 to define blood-type allele O2 and the T allele of rs176746 to define the blood-type allele B (instead of rs8176747). Diploid individuals O1O1, O2O2 and O1O2 were considered blood type O. Diploid individuals O1B, O2B and BB were considered blood type B. Absence of the alleles mentioned above was used to define blood-type

allele A. To evaluate differences across blood types, we compared the mean relative abundance of microbiome features in individuals with A, B and AB blood type to that in individuals with the O blood type using a two-sided Wilcoxon test. To evaluate the interaction with rs601338 *FUT2* (secretor/non-secretor) locus, we grouped individuals in two groups (non-O blood type and blood type O) to distinguish production or non-production of antigens, and compared pairs using a two-sided Wilcoxon test. All analyses were done using base R v3.6.1 (<https://www.R-project.org/>).

Effects of intrinsic and external factors on significantly associated loci

We evaluated the robustness of genome-wide associated signals by incorporating the following potential confounders into our statistical model: medication usage, anthropometric data and stool frequency and consistency data (collection and processing was described in Gacesa et al.(16)). We analyzed the effects of the following drug groups: proton pump inhibitors (ATC A02BC, N = 130), laxatives (osmotic ATC A06AD, N = 44; volume increasing ATC A06AC, N = 77), one group of anti-bacterials (ATC J01, N = 24) and other anti-infectives (ATC J, N = 39). The other group of drugs considered was antibiotic use in the 3 months prior to stool collection (N = 450). For each of these medications, we created dichotomous variables for all participants coded as 0 (non-user) or 1 (user). The list of intrinsic factors included BMI, stool frequency and stool consistency (mean Bristol stool scale). All models also incorporated age and sex as covariates and were run only for the genome-wide-significant SNP-trait pairs (**Supplementary Table 1**) using the same software used for GWAS (SAIGE, Zhou et al. (19)). To evaluate the impact of these covariates on the genetic signals, we used Cochran's Q heterogeneity test to compare the effect size obtained by the covariate-inclusive model and the basic model (that only includes age, sex and the genetic variant). To evaluate the impact of sex, we ran the SNP-association analysis in SAIGE separately for males and females using only age as a covariate. For each genetic variant, differences in effect size in males and females were tested using Cochran's Q heterogeneity test.

Interaction analyses

We used a three-step procedure to evaluate gene-diet interactions for all the taxa and pathways associated with SNPs at the *LCT* and *ABO* loci. First, we extracted the variables representing dietary habits that had previously shown significant association with these microbial traits (16). Next, we added these variables to the basic genetic model (feature ~ age + sex + genetic variant) to confirm their association at (at least) nominal significance level ($p < 0.05$) while accounting for the associated genetic variant (s). Finally, for the dietary variables showing nominal significance, we evaluated the interaction with the genetic variant (s) by including an interaction term into the association model. For the *LCT* locus, we used SNP rs182549 and considered two groups of genotypes (C/C vs. C/T and T/T) according to the dominant inheritance model. For the *ABO* locus, we used a binary definition of blood type (blood type O and A/B/AB), and

also considered the effect of the rs601338 genotype in the *FUT2* gene (defining secretor/ non-secretor individuals). All microbiome features were inverse-rank normalized before analyses, and age and sex were added as covariates, as in the main GWAS analysis. All models were fit using base R v3.6.1, function *lm()*, statistical tests were as implemented in packages *rstatix* v0.5.0 and *ggpubr* v0.3.0, and package *RNOmni* v0.7.1 was used for the inverse-rank normalization.

Replication in other cohorts and data sets

We looked for replication of our results using summary statistics from two independent studies: a genome-wide meta-analysis of 16S rRNA data from 24 cohorts and a genome-wide study on metagenomics data in the LL-DEEP cohort, another subset of the Lifelines cohort with data generated 4 years before the DMP and in which 255 participants were also later enrolled in DAG3(13,17). In the first study, we could not look at SNPs associated with species or pathways, as these microbiome features cannot be defined using 16S data, or SNPs on the X chromosome, as they were not analyzed in this study. In the second study, we searched for the exact same taxonomy or pathway, but similarly to MiBioGen, X chromosomal variants were not tested and some taxonomies were not defined due to the differences in metagenomics data processing pipelines.

Next to the replication of our findings, we also evaluated whether the genome-wide signals reported in a recent genome-wide study of microbiome taxa from a Finnish cohort were replicable in our data(18). We searched for all SNPs outside the *LCT* and *ABO* loci or any proxy ($r^2 > 0.8$) in our dataset and selected all SNP-taxonomy pairs that showed a $p < 1 \times 10^{-4}$ with at least one taxonomy in our cohort (**Supplementary Table 8**). We then looked at these associated taxa in the respective cohorts and compared them visually and with the aid of the Genome Taxonomy Database (<https://gtdb.ecogenomic.org/>) to determine if they were the same bacterial taxa or taxa from the same taxonomic branch.

Correlation between heritability estimates and the number of associated loci

To analyze the correlation between family-based heritability and the number of suggestive mbQTLs, we used heritability estimates previously derived from this cohort(16). We then calculated the number of independent mbQTLs per microbial trait by performing linkage disequilibrium pruning ($r^2 < 0.1$ in our data set, window size 1Mb using Plink v1.9(67)) for all SNPs at the three different thresholds: $p < 5 \times 10^{-4}$, $< 1 \times 10^{-4}$ and $< 5 \times 10^{-5}$. The association of heritability and the number of mbQTLs was calculated in R v.4.0.3 using a weighted Spearman correlation from `wCorr`` v1.9.1 package, with each taxon or pathway treated as a data point. The weights used in calculating the correlation were inversely proportional to the standard errors of heritability estimates. The regression lines in **Figure 5** were fit using the *loess* function (locally estimated scatterplot smoothing, base R package v4.0.3) with span and degree parameters set to 1.

Mendelian randomization analysis

To evaluate causal relationships between the gut microbiome and other common traits, we performed Mendelian randomization (MR) analyses that combined the summary statistics of the microbiome with publicly available summary statistics on food preferences, autoimmune and cardiovascular diseases and other cardio-metabolic traits. We analyzed the 33 microbiome features (pathways and taxa) with at least one variant passing the $p < 5 \times 10^{-8}$ threshold (**Supplementary Table 1**) and combined these with 77 publicly available summary statistic datasets retrieved using the IEU GWAS DATABASE(68).

We performed a bi-directional MR analysis, first testing if microbiome traits causally affect a phenotype and then testing if phenotypes can causally affect the microbiome traits. For each comparison we intersected the microbiome variants (MAF > 0.05) by rsID, position and alleles with the publicly available summary statistic variants. We then selected instruments using the 'clump_data()' function of the TwoSampleMR package (v0.5.5)(69). The publicly available summary statistics were clumped using a p -value threshold of $< 5 \times 10^{-8}$ and otherwise standard settings ($r^2 < 0.001$, 10mb window size). Due to the limited statistical significance of the microbiome traits, we performed p -value clumping at a less stringent $p < 5 \times 10^{-6}$ threshold. If fewer than three variants were clumped, we removed the trait combination from analysis.

The MR analysis was done using the TwoSampleMR v0.5.5 package. We first selected trait combinations that passed the Benjamini-Hochberg FDR threshold of 0.1 ($p = 2.805 \times 10^{-5}$) in the inverse-variance weighting test, resulting in one suggestively causal trait combination. We further checked that the significant trait combinations were unlikely to be driven by pleiotropy based on two criteria: i) the Egger regression intercept is nominally significant (indicating the presence of horizontal pleiotropy) and ii) the weighted median results are not nominally significant (indicating that no single variant influences the result)(70,71).

Power analysis

We calculated the variance explained at our loci using the formula described in Teslovich et al. (72), which takes into account MAF, effect size, standard error and sample size. We then performed a power analysis based on a linear model of association considering different genetic effect sizes (variance explained) and sample sizes (N) <https://genome.sph.umich.edu/wiki/Power_Calculations:_Quantitative_Traits>. We performed a sample size calculation by doing a grid search in the sample size sequence: [1000, 1050, ... , 50000] and keeping the lowest sample size that had power > 80%.

Data availability

Raw sequencing microbiome data, genotyping data and participant metadata are not publicly available to protect participants' privacy and research agreements in the informed consent. The data can be accessed

by all bona-fide researchers with a scientific proposal by contacting the Lifelines Biobank (instructions at: <https://www.lifelines.nl/researcher>). The authors declare that all other data supporting the findings of this study are available within the paper and its supplementary information files.

LifeLines Cohort Study - group authors genetics

Raul Aguirre-Gamboa (1), Patrick Deelen (1), Lude Franke (1), Jan A Kuivenhoven (2), Esteban A Lopera Maya (1), Ilja M Nolte (3), Serena Sanna (1), Harold Snieder (3), Morris A Swertz (1), Judith M Vonk (3), Cisca Wijmenga (1)

(1) Department of Genetics, University of Groningen, University Medical Center Groningen, the Netherlands

(2) Department of Pediatrics, University of Groningen, University Medical Center Groningen, the Netherlands

(3) Department of Epidemiology, University of Groningen, University Medical Center Groningen, the Netherlands

Acknowledgements

The authors wish to acknowledge the services of the Lifelines Cohort Study, the contributing research centers delivering data to Lifelines and all the study participants. The Lifelines initiative has been made possible by subsidy from the Dutch Ministry of Health, Welfare and Sport; the Dutch Ministry of Economic Affairs; the University Medical Center Groningen (UMCG); the University of Groningen (UG) and the Provinces of the North of the Netherlands (Drenthe, Friesland and Groningen). This project was carried out under Lifelines Project number OV18_0464. We thank Mathieu Plateel and Jody Geelderloos-Arends for their contribution in genotyping the Lifelines samples, Kate McIntyre for editorial support and Patrick Deelen for discussion of results. We also thank the UMCG Genomics Coordination Center, the UG Center for Information Technology and their sponsors BBMRI-NL & TarGet for storage and computational infrastructure.

The generation and management of genotype data for the Lifelines Cohort Study was supported by the UMCG Genetics Lifelines Initiative (UGLI). Metagenomics sequencing of the cohort was funded by the CardioVascular Onderzoek Nederland (CVON) grant (CVON 2012-03) to MH, JF and AZ. RG and RKW are supported by the collaborative TIMID project (LSHM18057-SGF) financed by the PPP allowance made available by Top Sector Life Sciences & Health to Samenwerkende Gezondheidsfondsen (SGF) to stimulate public-private partnerships and co-financing by health foundations that are part of the SGF. RKW is supported by the Seerave Foundation. AZ is supported by a European Research Council (ERC) Starting Grant 715772, a Netherlands Organization for Scientific Research (NWO) VIDI grant 016.178.056, a CVON grant 2018-27 and the NWO Gravitation grant ExposomeNL 024.004.017. JF is supported by the NWO Gravitation grant The Netherlands Organ-on-Chip Initiative (024.003.001) and CVON grant 2018-27. CW is further supported by an

ERC advanced grant (ERC-671274) and an NWO Spinoza award (NWO SPI 92-266). HJMHd is supported by the INTERREG V A (202085) funded project EurHealth-1Health. LC is supported by a joint fellowship from the UMCG and China Scholarship Council (CSC201708320268) and a Foundation De Cock-Hadders grant (20:20-13). EA is supported by Colciencias fellowship ed.783.

Author Contributions

ELM, AK and SH performed genetic analyses.
AvdG performed Mendelian randomization and power analyses.
AK, SH, LC, AVV and RG processed microbiome data.
SAS, TS, VC, MAYK, LAB and MBG processed samples meta-data.
PBTN and MAS provided resources on the HP computing cluster and assisted with data management.
HJMH, CW, JF, RKW and AZ provided funding resources and designed the DMP.
SZ and SS supervised the study.
ELM, AK, AvdG, SH, SAS, SZ and SS drafted the manuscript.
All authors were involved in data interpretation and provided critical input to the manuscript draft.

Competing Interests statement

The authors declare no conflicts of interest.

References

1. Valdes AM, Walter J, Segal E, Spector TD. Role of the gut microbiota in nutrition and health. *BMJ*. 2018 Jun 13;361:k2179.
2. Hall AB, Tolonen AC, Xavier RJ. Human genetic variation and the gut microbiome in disease. *Nat Rev Genet*. 2017;18(11):690–9.
3. Fan Y, Pedersen O. Gut microbiota in human metabolic health and disease. *Nat Rev Microbiol*. 2020 Sep 4;
4. Zhernakova A, Kurilshikov A, Bonder MJ, Tigchelaar EF, Schirmer M, Vatanen T, et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science*. 2016 Apr 29;352(6285):565–9.
5. Falony G, Joossens M, Vieira-Silva S, Wang J, Darzi Y, Faust K, et al. Population-level analysis of gut microbiome variation. *Science*. 2016 Apr 29;352(6285):560–4.
6. Goodrich JK, Davenport ER, Beaumont M, Jackson MA, Knight R, Ober C, et al. Genetic Determinants of the Gut Microbiome in UK Twins. *Cell Host Microbe*. 2016 May 11;19(5):731–43.
7. Rothschild D, Weissbrod O, Barkan E, Kurilshikov A, Korem T, Zeevi D, et al. Environment dominates over host genetics in shaping human gut microbiota. *Nature*. 2018 Mar;555(7695):210–5.
8. Bonder MJ, Kurilshikov A, Tigchelaar EF, Mujagic Z, Imhann F, Vila AV, et al. The effect of host genetics on the gut microbiome. *Nat Genet*. 2016;48(11):1407–12.
9. Wang J, Thingholm LB, Skiecevičienė J, Rausch P, Kummen M, Hov JR, et al. Genome-wide association analysis identifies variation in vitamin D receptor and other host factors influencing the gut microbiota. *Nat Genet*. 2016;48(11):1396–406.
10. Turpin W, Espin-Garcia O, Xu W, Silverberg MS, Kevans D, Smith MI, et al. Association of host genome with intestinal microbial composition in a large healthy cohort. *Nat Genet*. 2016;48(11):1413–7.
11. Hughes DA, Bacigalupe R, Wang J, Röhlemann MC, Tito RY, Falony G, et al. Genome-wide associations of human gut microbiome variation and implications for causal inference analyses. *Nat Microbiol*. 2020;5(9):1079–87.
12. Kurilshikov A, Wijmenga C, Fu J, Zhernakova A. Host Genetics and Gut Microbiome: Challenges and Perspectives. *Trends Immunol*. 2017;38(9):633–47.
13. Kurilshikov A, Medina-Gomez C, Bacigalupe R, Radjabzadeh D, Wang J, Demirkan A, et al. Genetics of human gut microbiome composition. *bioRxiv*. 2020 Jun 28;2020.06.26.173724.
14. Blekhnman R, Goodrich JK, Huang K, Sun Q, Bukowski R, Bell JT, et al. Host genetic variation impacts microbiome composition across human body sites. *Genome Biol*. 2015 Sep 15;16:191.
15. Vieira-Silva S, Falony G, Darzi Y, Lima-Mendez G, Garcia Yunta R, Okuda S, et al. Species–function relationships shape ecological properties of the human gut microbiome. *Nat Microbiol*. 2016 Jun 13;1(8):1–8.
16. Gacesa R, Kurilshikov A, Vila AV, Sinha T, Klaassen M a. Y, Bolte LA, et al. The Dutch Microbiome Project defines factors that shape the healthy gut microbiome. *bioRxiv*. 2020 Nov 27;2020.11.27.401125.
17. Sanna S, van Zuydam NR, Mahajan A, Kurilshikov A, Vich Vila A, Vösa U, et al. Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. *Nat Genet*. 2019 Apr;51(4):600–5.
18. Qin Y, Havulinna AS, Liu Y, Jousilahti P, Ritchie SC, Tokolyi A, et al. Combined effects of host genetics and diet on human gut microbiota and incident disease in a single population cohort. *medRxiv*. 2020 Sep 13;2020.09.12.20193045.
19. Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME, Wolford BN, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet*. 2018;50(9):1335–41.
20. M-GWAS for the gut microbiome in Chinese adults illuminates on complex diseases | *bioRxiv* [Internet]. [cited 2020 Nov 27]. Available from: <https://www.biorxiv.org/content/10.1101/736413v1>
21. Van Der Heide HM, Magnee W, Van Loghem JJ. Blood group frequencies in the

- Netherlands. *Am J Hum Genet.* 1951 Dec;3(4):344–7.
22. Hazra A, Kraft P, Selhub J, Giovannucci EL, Thomas G, Hoover RN, et al. Common variants of FUT2 are associated with plasma vitamin B12 levels. *Nat Genet.* 2008 Oct;40(10):1160–2.
  23. P719. Identification of a genetic risk variant in the WWOX gene associated with recurrent fibrostenotic Crohn's disease. *J Crohns Colitis.* 2016 Mar 1;10(suppl_1):S472–S472.
  24. Wang K, Li W-D, Zhang CK, Wang Z, Glessner JT, Grant SFA, et al. A genome-wide association study on obesity and obesity-related traits. *PLoS One.* 2011 Apr 28;6(4):e18939.
  25. Kim CH. Microbiota or short-chain fatty acids: which regulates diabetes? *Cell Mol Immunol.* 2018 Feb;15(2):88–91.
  26. Ostrowski J, Paziewska A, Lazowska I, Ambrozkiwicz F, Goryca K, Kulecka M, et al. Genetic architecture differences between pediatric and adult-onset inflammatory bowel diseases in the Polish population. *Sci Rep.* 2016 23;6:39831.
  27. Xue A, Wu Y, Zhu Z, Zhang F, Kemper KE, Zheng Z, et al. Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nat Commun.* 2018 27;9(1):2941.
  28. Röhlemann MC, Hermes BM, Bang C, Doms S, Moitinho-Silva L, Thingholm LB, et al. Genome-wide association study in 8,956 German individuals identifies influence of ABO histo-blood groups on gut microbiome. *medRxiv.* 2020 Sep 18;2020.07.09.20148627.
  29. Bonder MJ, Kasela S, Kals M, Tamm R, Løkk K, Barragan I, et al. Genetic and epigenetic regulation of gene expression in fetal and adult human livers. *BMC Genomics.* 2014;15(1):1–13.
  30. Moon Jee-Young, Wang Zheng, Usyk Mykhaylo, Vazquez-baeza Yoshiki, Isasi Carmen R, Mossavar-Rahmani Yasmin, et al. Abstract P459: Milk Intake, Host LCT Genotype and Gut Bifidobacteria in Relation to Obesity: Results From the Hispanic Community Health Study/Study of Latinos (HCHS/SOL). *Circulation.* 2020 Mar 3;141(Suppl_1):AP459–AP459.
  31. Liu X, Tong X, Zou Y, Lin X, Zhao H, Tian L, et al. Inter-determination of blood metabolite levels and gut microbiome supported by Mendelian randomization. *bioRxiv.* 2020 Jul 1;2020.06.30.181438.
  32. Motta V, Luise D, Bosi P, Trevisi P. Faecal microbiota shift during weaning transition in piglets and evaluation of AO blood types as shaping factor for the bacterial community profile. *PLoS One.* 2019;14(5):e0217001.
  33. An ancient deletion in the ABO gene affects the composition of the porcine microbiome by altering intestinal N-acetyl-galactosamine concentrations | *bioRxiv [Internet].* [cited 2020 Nov 27]. Available from: <https://www.biorxiv.org/content/10.1101/2020.07.16.206219v1>
  34. Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, et al. Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene. *Am J Hum Genet.* 2004 Jun 1;74(6):1111–20.
  35. Auer PL, Johnsen JM, Johnson AD, Logsdon BA, Lange LA, Nalls MA, et al. Imputation of exome sequence variants into population- based samples and blood-cell-trait-associated loci in African Americans: NHLBI GO Exome Sequencing Project. *Am J Hum Genet.* 2012 Nov 2;91(5):794–808.
  36. Séguirel L, Thompson EE, Flutre T, Lovstad J, Venkat A, Margulis SW, et al. The ABO blood group is a trans-species polymorphism in primates. *Proc Natl Acad Sci.* 2012 Nov 6;109(45):18493–8.
  37. Band G, Le QS, Clarke GM, Kivinen K, Hubbart C, Jeffreys AE, et al. Insights into malaria susceptibility using genome-wide data on 17,000 individuals from Africa, Asia and Oceania. *Nat Commun.* 2019 Dec 16;10(1):5732.
  38. Barua D, Paguio AS. ABO blood groups and cholera. *Ann Hum Biol.* 1977 Jan 1;4(5):489–92.
  39. Watanabe K, Stringer S, Frei O, Umičević Mirkov M, de Leeuw C, Polderman TJC, et al. A global overview of pleiotropy and genetic architecture in complex traits. *Nat Genet.* 2019 Sep;51(9):1339–48.
  40. Naitza S, Porcu E, Steri M, Taub DD, Mulas A, Xiao X, et al. A Genome-Wide Association Scan on the Levels of Markers of Inflammation in Sardinians Reveals Associations That Underpin Its Complex Regulation. *PLOS Genet.* 2012 Jan

- 26;8(1):e1002480.
41. Parks DH, Chuvochina M, Chaumeil P-A, Rinke C, Mussig AJ, Hugenholtz P. A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat Biotechnol*. 2020 Sep;38(9):1079–86.
42. Wade KH, Hall LJ. Improving causality in microbiome research: can human genetic epidemiology help? *Wellcome Open Res*. 2020 Apr 24;4:199.
43. Pirastu N, McDonnell C, Grzeszkowiak EJ, Mounier N, Imamura F, Day FR, et al. Using genetics to disentangle the complex relationship between food choices and health status. *bioRxiv*. 2019 Nov 8;829952.
44. Wang C, Huang Z, Yu K, Ding R, Ye K, Dai C, et al. High-Salt Diet Has a Certain Impact on Protein Digestion and Gut Microbiota: A Sequencing and Proteome Combined Study. *Front Microbiol* [Internet]. 2017 [cited 2020 Nov 27];8. Available from: <https://www.frontiersin.org/articles/10.3389/fmicb.2017.01838/full#B5>
45. Culligan EP, Marchesi JR, Hill C, Sleator RD. Combined metagenomic and phenomic approaches identify a novel salt tolerance gene from the human gut microbiome. *Front Microbiol* [Internet]. 2014 [cited 2020 Nov 27];5. Available from: <https://www.frontiersin.org/articles/10.3389/fmicb.2014.00189/full>
46. Zeevi D, Korem T, Zmora N, Israeli D, Rothschild D, Weinberger A, et al. Personalized Nutrition by Prediction of Glycemic Responses. *Cell*. 2015 Nov 19;163(5):1079–94.
47. Zhong VW, Kuang A, Danning RD, Kraft P, van Dam RM, Chasman DI, et al. A genome-wide association study of bitter and sweet beverage consumption. *Hum Mol Genet*. 2019 15;28(14):2449–57.
48. Björkhaug ST, Aanes H, Neupane SP, Bramness JG, Malvik S, Henriksen C, et al. Characterization of gut microbiota composition and functions in patients with chronic alcohol overconsumption. *Gut Microbes*. 2019;10(6):663–75.
49. Sanna S, Jackson AU, Nagaraja R, Willer CJ, Chen W-M, Bonnycastle LL, et al. Common variants in the GDF5-UQC region are associated with variation in human height. *Nat Genet*. 2008 Feb;40(2):198–203.
50. Weedon MN, Lettre G, Freathy RM, Lindgren CM, Voight BF, Perry JRB, et al. A common variant of HMGA2 is associated with adult and childhood height in the general population. *Nat Genet*. 2007 Oct;39(10):1245–50.
51. Loos RJF, Lindgren CM, Li S, Wheeler E, Zhao JH, Prokopenko I, et al. Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nat Genet*. 2008 Jun;40(6):768–75.
52. Tigchelaar EF, Zernakova A, Dekens JAM, Hermes G, Baranska A, Mujagic Z, et al. Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. *BMJ Open*. 2015 Aug 1;5(8):e006772.
53. Scholtens S, Smidt N, Swertz MA, Bakker SJ, Dotinga A, Vonk JM, et al. Cohort Profile: LifeLines, a three-generation cohort study and biobank. *Int J Epidemiol*. 2015 Aug 1;44(4):1172–80.
54. the Haplotype Reference Consortium, McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet*. 2016 Oct;48(10):1279–83.
55. Lopera Maya EA, van der Graaf A, Lanting P, van der Geest M, Fu J, Swertz M, et al. Lack of Association Between Genetic Variants at ACE2 and TMPRSS2 Genes Involved in SARS-CoV-2 Infection and Human Quantitative Phenotypes. *Front Genet* [Internet]. 2020 [cited 2020 Nov 27];11. Available from: <https://www.frontiersin.org/articles/10.3389/fgene.2020.00613/full>
56. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*. 2015 Oct 1;526(7571):68–74.
57. Francioli LC, Menelaou A, Pulit SL, van Dijk F, Palamara PF, Elbers CC, et al. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet*. 2014 Aug;46(8):818–25.
58. Tigchelaar EF, Zernakova A, Dekens JAM, Hermes G, Baranska A, Mujagic Z, et al. An introduction to LifeLines DEEP: study design and baseline characteristics. *bioRxiv*. 2014 Sep 16;009217.
59. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat Methods*. 2015 Oct;12(10):902–3.

60. Franzosa EA, McIver LJ, Rahnavard G, Thompson LR, Schirmer M, Weingart G, et al. Species-level functional profiling of metagenomes and metatranscriptomes. *Nat Methods*. 2018;15(11):962–8.
61. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, UniProt Consortium. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinforma Oxf Engl*. 2015 Mar 15;31(6):926–32.
62. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods*. 2015 Jan;12(1):59–60.
63. Siebelink E, Geelen A, de Vries JHM. Self-reported energy intake by FFQ compared with actual energy intake to maintain body weight in 516 adults. *Br J Nutr*. 2011 Jul;106(2):274–81.
64. Willett WC. Nutritional Epidemiology [Internet]. Nutritional Epidemiology. Oxford University Press; [cited 2020 Nov 30]. Available from: <https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780195122978.001.0001/acprof-9780195122978>
65. Zheng X, Davis JW. SAIGEgds—an efficient statistical tool for large-scale PheWAS with mixed models. *Bioinformatics* [Internet]. [cited 2020 Nov 29]; Available from: <https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btaa731/5902828>
66. Genomewide Association Study of Severe Covid-19 with Respiratory Failure I NEJM [Internet]. [cited 2020 Nov 29]. Available from: <https://www.nejm.org/doi/full/10.1056/NEJMoa2020283>
67. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*. 2015;4:7.
68. Elsworth B, Lyon M, Alexander T, Liu Y, Matthews P, Hallett J, et al. The MRC IEU OpenGWAS data infrastructure. *bioRxiv*. 2020 Aug 10;2020.08.10.244293.
69. 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. Loos R, editor. *eLife*. 2018 May 30;7:e34408.
70. Burgess S, Thompson SG. Mendelian Randomization Methods for Using Genetic Variants in Causal Estimation [Internet]. CRC Press; 2015 [cited 2018 Dec 3]. 210 p. Available from: <https://www.crcpress.com/Mendelian-Randomization-Methods-for-Using-Genetic-Variants-in-Causal-Estimation/Burgess-Thompson/p/book/9781466573178>
71. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol*. 2015 Apr 1;44(2):512–25.
72. Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, Koseki M, et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*. 2010 Aug;466(7307):707–13.

Adriaan van der Graaft, Maria Zorrot, Annique Claringbould, Urmo Vösa, Raul Aguirre-Gamboa, Chan Li, Joram Mooiweer, Isis Ricano-Ponce, Zuzanna Borek, Frits Koning, Yvonne Kooy-Winkelaar, Ludvig Sollid, Shuo-Wang Qiao, BIOS consortium, Vinod Kumar, Yang Li, Lude Franke, Sebo Withoff, Cisca Wijmenga, Serena Sanna*, Iris Jonkers*.

chapter 6

Systematic prioritization of candidate genes in disease loci identifies TRAFD1 as a master regulator of IFN γ signalling in celiac disease

Celiac disease (CeD) is a complex T cell-mediated enteropathy induced by gluten. Although genome-wide association studies have identified numerous genomic regions associated with CeD, it is difficult to accurately pinpoint which genes in these loci are most likely to cause CeD. We used four different *in silico* approaches – Mendelian Randomization inverse variance weighting, COLOC, LD overlap and DEPICT – to integrate information gathered from a large transcriptomics dataset. This identified 118 prioritized genes across 50 CeD-associated regions. Co-expression and pathway analysis of these genes indicated an association with adaptive and innate cytokine signalling and T cell activation pathways. 51 of these genes are targets of known drug compounds and/or likely druggable genes, suggesting that our methods can be used to pinpoint potential therapeutic targets. In addition, we detected 172 gene-combinations that were affected by our CeD-prioritized genes *in trans*. Notably, 41 of these *trans*-mediated genes appear to be under control of one master regulator, TRAFD1, and were found to be involved in IFN γ signalling and MHC I antigen processing/presentation. Finally, we performed *in vitro* experiments in a human monocytic cell line that validated the role of TRAFD1 as an immune regulator acting *in trans*. Our strategy confirmed the role of adaptive immunity in CeD and revealed a genetic link between CeD and IFN γ signalling as well as with MHC I antigen processing, both major players of immune activation and CeD pathogenesis.

Introduction

Celiac disease (CeD) is an auto-immune disease in which patients experience severe intestinal inflammation upon ingestion of gluten peptides. CeD has a large genetic component, with heritability estimated to be approximately 75% (1). The largest CeD-impacting locus is the HLA region, which contributes approximately 40% of CeD heritability (2). While the individual impacts of CeD-associated genes outside the HLA region are smaller, they jointly account for an additional 20% of heritability. Previous genome-wide association studies (GWAS) have identified 42 non-HLA genomic loci associated with CeD. Yet, identification of these non-HLA genetic components and an understanding of the molecular perturbations associated with them are necessary to understand CeD pathophysiology.

Understanding the biological mechanisms of non-HLA CeD loci is difficult: only three of these loci point to single nucleotide polymorphisms (SNPs) located in protein-coding regions (3). The other CeD-risk loci cannot be explained by missense mutations, making it necessary to look at other biological mechanisms such as gene expression to explain their role in CeD pathogenicity. Several studies have been performed to integrate expression quantitative trait loci (eQTLs) with CeD GWAS associations

*Adapted from van der Graaf et al.
Systematic prioritization of candidate
genes in disease loci identifies TRAFD1
as a master regulator of IFN γ signalling
in celiac disease, Front. Genet. 11,
(2021).*

†: Equal contribution
*: Corresponding author

(4–6), and several candidate genes, including *UBASH3A*, *CD274*, *SH2B3* and *STAT4* (7), have been identified, implicating T cell receptor, NFκB and interferon signalling pathways as biological pathways associated with CeD pathology. Unfortunately, these eQTL studies had limited sample sizes, which reduced their power to identify *cis*- and (especially) *trans*-eQTLs. Furthermore, previous attempts to integrate eQTLs have mostly annotated genomic loci based on catalogued eQTLs without testing the causality of the genes in the onset or exacerbation of CeD (5,8,9).

Gene expression and GWAS data can also be integrated using methodologies that identify shared mechanisms between diseases. These methods can be roughly divided into three classes: variant colocalization methods, causal inference methods and co-expression methods. Colocalization methods consider the GWAS and eQTL summary statistics at a locus jointly and probabilistically test if the two signals are likely to be generated by the same causal variant (10). Causal inference methods test if there is a causal relationship between expression changes and the disease, using genetic associations to remove any confounders (11,12). Finally, co-expression methods do not use eQTL information, but rather test if there is significant co-expression between the genes that surround the GWAS locus (13). Unfortunately, there is no current “gold standard” method for finding the causal gene behind a GWAS hit, as all the methods discussed here are subject to their respective assumptions, drawbacks and caveats. However, it is worthwhile to use all these methods in parallel to find the most likely causal genes for CeD.

Here, we systematically applied four prioritization methods to the latest meta-analysis for CeD that we have performed previously (14) and coupled them with eQTL results from the Biobank Integrative Omics Study (BIOS) cohort (15), one of the largest cohorts for which there is genotype and RNA-seq expression data of peripheral blood mononuclear cells. We focused on 58 CeD associated loci ($p < 5 \times 10^{-6}$) outside the HLA region. Our approach prioritized 118 genes in 50 loci and identified one gene, *TRAFD1*, as a master regulator of *trans*-effects. We then experimentally validated the role of *TRAFD1*-mediated genes using a *TRAFD1* knockdown through RNA-seq in a disease-relevant cell type. Our study yields novel insights into the genetics of CeD and is proof-of-concept for a systematic approach that can be applied to other complex diseases. A schematic overview of our study is shown in **Supplementary Figure 1A**.

Materials and Methods

Genotypes for eQTL analysis

We used the BIOS cohort (15) to map eQTLs in 3,746 individuals of European ancestry. The BIOS cohort is a collection of six cohorts: the Cohort on Diabetes and Atherosclerosis Maastricht (16), the Leiden Longevity Study (17), Lifelines DEEP (18), the Netherlands Twin Registry (19), the Prospective ALS Study Netherlands (20) and the Rotterdam Study (21). As described in Vosa et al. (22), each cohort was genotyped separately using different arrays and genotypes were subsequently

chapter 6

imputed to the Haplotype Reference Consortium panel (HRC v1.0) on the Michigan imputation server (23).

While genotyping data was generated and imputed independently of this study (22), here we applied additional filters at the marker and sample level. Specifically, we considered only biallelic SNPs with a minor allele frequency (MAF) > 0.01 , a Hardy-Weinberg test p value $>$ and an imputation quality RSQR > 0.8 . We also removed related individuals using a genetic relationship matrix (GRM) created with Plink 1.9 (24). Pairs of individuals with a GRM value > 0.1 were considered related, and one individual was removed from each of these pairs. Furthermore, population outliers were identified using a principal component analysis on the GRM, and individuals who were more than 3 standard deviations from the means of principal component 1 or 2 were removed.

Expression quantification and eQTL analysis

We used pre-processed RNA-seq data from a previous study (15,22). After matching the RNA-seq data with the genotyping data filtered described above, 3,503 individuals, 19,960 transcripts and 7,838,327 autosomal SNPs remained for analyses. We performed genome-wide eQTL mapping for the transcripts using plink 1.9 (24) with the `--assoc` command. We defined *cis*-eQTL variants as those located within ± 1.5 Mb of the transcript and *trans*-eQTLs as variants located outside these boundaries. Of note, we opted for this larger window for defining *cis*-eQTLs (usually is ± 500 Kb) so that eQTL associations would fully overlap the associated CeD GWAS peak even when a gene is on the edge of the CeD-associated region (chosen to be 1Mb). To select variants that could explain the *cis*-eQTL signal of a gene, we used GCTA-COJO (25) v1.26. For this analysis, we required selected variants to reach a p -value threshold of 5×10^{-6} and included the BIOS cohort genotypes as LD reference. This identified 707 genes with at least one eQTL reaching this threshold, 357 of which had more than one conditionally independent eQTL variant.

CeD summary statistics associated regions and candidate genes

We used summary statistics from a CeD GWAS meta-analysis of 12,948 cases and 14,826 controls that analyzed 127,855 variants identified using the ImmunoChip array (14). SNP positions were lifted over to human genome build 37 using the UCSC *liftover* tool. We first identified lead associated variants in the CeD meta-analysis by performing p -value clumping; we used plink 1.9 (24) to select variants at a p -value threshold of 5×10^{-6} and pruned variants in LD with these selected variants using standard plink settings ($r^2 > 0.5$, utilizing 1000 Genomes European sample LD patterns and ± 250 Kb window) (24,26). We removed variants in an extended HLA region (chromosome 6, 25Mb to 37Mb) due to the complex long-range LD structure in this region and because we aim to understand the function of the non-HLA genetic component of CeD. We looked for candidate genes around the clumped variants as follows. First, we considered all genomic regions around (± 1 Mb) every clumped variant, to also capture variants in long range LD that are outside the standard ± 500 Kb window (27). We then joined all overlapping CeD-associated regions together and looked for gene transcripts that partly or

fully overlapped with the associated regions. This approach identified 58 CeD-associated regions and 1,235 candidate genes that are potentially causal for CeD.

Gene prioritization using Mendelian Randomization–Inverse Variance Weighting (MR-IVW), COLOC, LD overlap and DEPICT

We prioritized CeD-associated genes using three eQTL-based methods – MR-IVW (28), COLOC (10) and LD overlap – and one co-regulation-based method, DEPICT (13). For the MR-IVW method, we used the independent variants identified by GCTA-COJO as instrumental variables (11,27) to test causal relationships between changes in gene expression and CeD, as we have demonstrated that this procedure is preferred over p -value clumping (12). MR-IVW was only performed when there were three or more independent eQTLs available (164 genes). A gene was significant for the MR-IVW test if the causal estimates passed a Bonferroni threshold p -value of 3.1×10^{-4} . Heterogeneity of causal estimates was evaluated using Weighted Median MR analysis and Cochran's Q test (29). For the COLOC method, we used the 'coloc' v3 R package and considered a gene significant for the COLOC analysis if the posterior probability of shared variants (H4) was larger than 0.9. For the LD overlap method, a gene was considered significant if there was high LD ($r^2 > 0.8$) between the top independent eQTL and the top CeD variant in the region. Finally, we applied DEPICT (13) to the clumped CeD GWAS variants described in 'CeD summary statistics associated regions and candidate genes. Genes identified by the DEPICT analysis were considered significant if a False Discovery Rate (FDR) < 0.05 was found with DEPICT's own permutation measure.

We scored each gene in the CeD-associated loci by considering each of the four prioritization methods. A gene was prioritized as 'potentially causal' in CeD pathology when one of the four methods was significant (one line of evidence). If multiple lines of evidence were significant, the gene was prioritized more highly than when only a single line of evidence was available.

To explore how the prioritized genes affect CeD risk, we gave each gene an effect direction based on the effect direction of the top variants in the eQTL and the CeD GWAS. We assigned positive ('+') when increased expression increases CeD risk, negative ('-') when increased expression decreases CeD risk and '?' when it was not clear from our results how the expression affects CeD risk. To assign these directions, we used the following procedure:

1. If there was a concordant effect that was significant in the top variants of both the eQTLs and the GWAS, the direction of the concordant effect was chosen.
2. If there was a concordant effect, but no significance of the SNP in one of the datasets, we could not be sure of an effect direction, and a question mark was chosen. The only exception to this was if the MR-IVW was significant, in which instance we chose the direction of the MR-IVW effect as effect direction.

chapter 6

3. If there was a discordant effect between the top SNPs, and both were significant in both datasets, a question mark was chosen. The only exception to this was when the MR-IVW was significant, when the MR-IVW effect was chosen.
4. If the effect between the eQTL top SNP and CeD top SNP was not in the same direction and only the eQTL top SNPs were genome-wide significant, the eQTL direction was chosen.
5. If the effect between the eQTL top SNP and CeD top SNP was not in the same direction and only the GWAS top SNPs were genome-wide significant, the GWAS direction was chosen.
6. Otherwise, a question mark was chosen.

Co-regulation clustering

The genes that have been prioritized may have some shared function in CeD pathology. To identify possible shared pathways, we performed co-regulation clustering analysis based on 1,588 normalized expression co-regulation principal components identified from RNA-seq information across multiple human tissues by Deelen et al (30). We performed pairwise Pearson correlation of our prioritized genes with these 1,588 principal components and derived a correlation Z score for each prioritized gene pair. We then performed hierarchical clustering of this Z score matrix using Ward distances and identified 4 clusters from the resulting dendrogram.

Trans-eQTL and mediation analysis

238 autosomal genes that were not located in, but were associated with, a significant *trans*-eQTL variant ($p < 5 \times 10^{-8}$) in the CeD-associated regions were identified and used as potential targets for mediation by our associated genes in the CeD-associated loci (86 potential *cis* mediating genes). We first selected *trans*-eQTL genes that were co-expressed (Pearson $r > 0.1$, 197 gene combinations) with prioritized genes, then performed mediation analysis by running the *trans*-eQTL association again using the expression of the *cis*-eQTL gene as a covariate. We defined a *trans*-mediated gene if, after mediation analysis, the change (increase or decrease) in the effect size of the top *trans*-eQTL variant was significant according to the statistical test described in Freedman and Schatzkin (31). For this analysis, we used a Bonferroni-adjusted p -value of 3.0×10^{-4} .

Cell type proportion and SH2B3 expression mediation analysis

To assess if the *cis*-eQTL effect of *TRAFD1* was not a proxy for cell-type composition, we performed mediation analyses in a fashion similar to the *trans*-mediation analysis above using cell proportions measured in a subset of individuals in the BIOS cohort. To ensure that there was no residual effect of *SH2B3*-expression on the mediating effect of *TRAFD1*, we corrected the original *TRAFD1* expression levels for the expression levels of *SH2B3*, leaving *TRAFD1* expression independent of *SH2B3*, and reran the mediation analysis.

Literature review

We performed a REACTOME pathway (32) analysis to determine the potential function of the prioritized genes. This was complemented with a literature search (research and review papers) in Pubmed. For the coding and non-coding genes for which no studies were found, Genecards (www.genecards.org) and Gene Network v2.0 datasets (30) were used, respectively. Information regarding the potential druggability of the prioritized genes was obtained from DrugBank (33), the pharmacogenetics database (34) and a previous study that catalogued the druggability of genes (35).

Culturing of the THP-1 cell line

The cell line THP-1 (Sigma Aldrich, ECACC 88081201) was cultured in RPMI 1640 with L-glutamine and 25mM HEPES (Gibco, catalogue 52400-025), and supplemented with 10% fetal bovine serum (Gibco, catalogue 10270) and 1% penicillin/ streptomycin (Lonza, catalog DE17602E). The cells were passed twice per week at a density lower than 0.5×10^6 cells/ml in a humidified incubator at 5% CO₂, 37°C.

siRNA treatment of THP-1 culture

THP-1 cells were plated at 0.6×10^6 cells/ml and transfected with 25 nM siRNA using Lipofectamine RNAimax transfection reagent (Invitrogen, catalogue 13788), according to the manufacturer's protocol. Cells were treated with an siRNA to target *TRAFD1* (Qiagen catalogue 1027416, sequence CCCAGCCGACCCATTAACAAT) (Knockdown (KD)), and cells treated with transfection mix without siRNA (Wild type (WT)) or non-targeting control siRNA (scrambled (SCR)) (Qiagen catalog SI03650318, sequence undisclosed by company) were included as controls. All the treatments were performed in triplicate. 72 hours after transfection, a small aliquot of cells was stained for Trypan Blue exclusion to determine cell viability and proliferation. The cells were stimulated with either LPS (10 ng/ml) from *E. coli* (Sigma catalogue 026:B6) or media alone (unstimulated) for 4h. Subsequently, the cells were centrifuged, and the cell pellets suspended in lysis buffer and stored at -80°C until used for RNA and protein isolation.

qPCR of THP-1 culture

The total RNA from THP-1 cells was extracted with the mirVana™ miRNA isolation kit (AMBION, catalogue AM1561) and subsequently converted to cDNA using the RevertAid H Minus First Strand cDNA Synthesis Kit (Thermo scientific, catalogue K1631). qPCR was done using the Syber green mix (Bio-Rad, catalogue 172-5124) and run in a QuantStudio 7 Flex Real-Time system (Applied Biosystems, catalogue 448598). Primer sequences to determine KD levels of *TRAFD1* were 5' GCTGTAAAGAAGCATGAGGAGAC and 3' TTGCCACATAGTCCGTCCG. *GAPDH* was used as endogenous qPCR control with primers 5' ATGGGGAAGGTGAAGGTCG and 3' GGGGTCATTGATGGCAACAATA. Relative expression values of *TRAFD1* were normalized to the endogenous control *GAPDH* and calculated using the $\Delta\Delta CT$ method, then given as a percentage relative to SCR expression levels.

Western blot (WB) THP-1 culture

Cell pellets from THP-1 cells were suspended on ice-cold lysis buffer (PBS containing 2% SDS and complete protease inhibitor cocktail (Roche, catalog 11697498001)). Protein concentration of cell extracts was determined using the BCA protein kit (Pierce, catalog 23225). Proteins were separated on 10% SDS-polyacrylamide electrophoresis gel and transferred to a nitrocellulose membrane. After 1 hour of blocking with 5% fat-free milk in Tris-Tween-Buffer-Saline, the membranes were probed for 1 hour at room temperature with mouse mono-clonal TRAFD1 antibody 1:1000 (Invitrogen, catalog 8E6E7) or mouse monoclonal anti-actin antibody 1:5000 (MP Biomedicals, catalog 08691001), followed by incubation with goat anti-mouse horseradish peroxidase-conjugated secondary antibodies 1:10000 (Jackson Immuno Research, catalog 115-035-003). After three 10-minute washes, the bands were detected by Lumi light WB substrate (Roche, catalogue 12015200001) in a Chemidoc MP imaging system (Bio-Rad) and quantified using Image Lab™ software (Bio-Rad). The band intensity of TRAFD1 was normalized to actin, and the TRAFD1 SCR control level was set as 100%.

Statistical analysis for in vitro experiments in THP-1

The statistical analyses of proliferation, qPCR and WB were performed using Prism 5 software (GraphPad Software, Inc.). Results are presented as mean $\pm$ SEM from a representative experiment. Statistical differences were evaluated using a one-tailed *t*-test.

RNA sequencing (RNA-seq) in THP-1 cells

RNA from THP-1 cells was extracted with the mirVana™ miRNA isolation kit (AMBION, catalog AM1561). Prior to library preparation, extracted RNA was analysed on the Experion Stdsend RNA analysis kit (Bio-Rad, catalog 7007105). 1 μ g of total RNA was used as input for library preparation using the quant seq 3' kit (Lexogen, catalog 015.96), according to the manufacturer's protocol. Each RNA library was sequenced on the Nextseq500 (Illumina). Low quality reads, adaptors and poly-A tail reads were removed from FASTQ files. The QC-ed FASTQ files were then aligned to the human_g1k_v37 Ensembl Release 75 reference genome using HISAT default settings (36), and sorted using SAMtools (37). Gene-level quantification was performed by the *featurecounts* function of the RSubread R package v1.6.2 (38). A modified Ensembl version 75 gtf file mapping only to the last 5' 500 bps per gene was used as gene-annotation database to prevent counting of reads mapping to intra-genic A-repeats. Gene-level differential expression analysis between conditions was performed using the DESeq2 R package (39) after removing genes with zero counts. Differentially expressed genes (DEGs) were defined as genes presenting an absolute log2 fold change (llog2 FCI) >1 and an FDR ≤ 0.01 across treatment (WT vs. SCR or KD unstimulated cells). To identify the genes responding to LPS stimulation, the DEGs between unstimulated samples and their respective stimulated sample were determined. Venn diagrams were used to depict the relationship between these genes. REACTOME pathway analyses were performed to identify biological

processes and pathways enriched in different sets of DEGs using the enrichr API. Enrichments were considered significant if they were below a 0.05 FDR-threshold defined by the enrichr API (32). Raw data and count matrix are available under GEO accession number GSE146284.

Gene set permutation analysis

It can be difficult to determine if a set of genes is 'on average' more or less differentially expressed due to co-expression between the genes within the set. To mitigate this, we performed a permutation test that considers the median absolute T statistic calculated by DESeq2 (39). This allowed us to compare the expected differential expression of an arbitrary set of genes when exposed to a non-targeting siRNA with that of the observed differential expression of a siRNA targeting *TRAFD1*. We compare the median differential expression in the WT-SCR comparison, with the observed differential expression of the same set of genes in the SCR-KD comparison. This will still incorporate the co-expression structure of the data. To do this, we randomly selected a same-sized set of genes 1,000,000 times in each relevant experiment (WT-SCR or SCR-KD), and determined the observed median absolute T statistic. We calculated a ratio of how often the permuted value is higher than the observed value. For example, the observations can be that 1% of permuted gene sets are more differentially expressed in the WT-SCR experiment, while only 0.01% of permuted genes sets are more differentially expressed in the SCR-KD experiment. Finally, we divide these values by one another, (percentage SCR-KD)/(percentage WT-SCR), to calculate a fold change in differential expression. In the example given above, this indicates that the KD is 100 times ($0.01/1 = 100$) more differentially expressed than expected.

Available RNA-seq datasets

Four RNA-seq datasets available to us were included to study the pattern of expression of prioritized genes. A brief description of each dataset is provided below.

1. *Whole biopsy samples:* Duodenal biopsies were obtained from 11 individuals (n = 6 CeD patients and n = 5 controls) who underwent upper gastrointestinal endoscopy (previously described) (40). Data can be found under accession number GSE146190.
2. *Intra epithelial cytotoxic lymphocytes (IE-CTLs)* CD8⁺ TCRαβ IE-CTLs cell lines were isolated from intestinal biopsies and expanded, as described previously (41). Stimulations (IL-15, IFNβ or IL21) and DEGs analysis were described by Zorro et al 2020 (40). Data can be found under accession number GSE126409.
3. *Gluten specific T cells (GS T cells)* Stimulations with anti-CD3 and anti-CD28 (n = 22 samples per conditions were done in 6-well plates coated overnight with anti-CD3 (2.5 µg/ml; Biolegend, San Diego, CA, USA) and anti-CD28 (2.5 µg/ml; Biolegend, San Diego, CA, USA) or PBS (negative control) for 0, 10, 30 and 180 minutes. At each time point cells were harvested for RNA isolation. RNA was extracted with the mirVana RNA isolation kit (Ambion, Life Technologies, Carlsbad, CA, USA) according to the manufacturer's instructions. The

quantity and quality of RNA was determined by Bioanalyzer (Agilent technologies, Santa Clara, CA, USA). The sequencing libraries were prepared from 1 µg of total RNA using the TruSeq Stranded Total RNA with Ribo-Zero Globin kit (Illumina, San Diego, CA, USA) according to the manufacturer's instructions. Sequencing was done with the Illumina HiSeq2500 (Illumina, San Diego, CA, USA). Before alignment the reverse complement of the fastQ sequences were taken using the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit). The alignment was done using Hisat2 version 2.0.4 against the forward strand with default alignment parameters (42). The reference genome index was made using hisat2-build indexer and the 1000 genomes reference genome version GRCh37 v75, with default parameters. For the samples that had paired end data only the first mate file was used for alignment. Reads mapping to multiple positions were removed. The genes were quantified using HTSeq (43) version 0.6.1.p1, with options -m union, -t exon, --stranded yes, and other options on default. DE effects were quantified using the R package DEseq2 (version 1.26.0) (43) and filtered on having an absolute log2 fold change (log2FC) of at least 1 and an FDR < 0.05. Data is available under accession number GSE146441 (Bakker, Ramirez-Sanchez and Borek et al. manuscript in preparation).

4. iv) *Caco-2 cells* Stimulation of the Caco-2 cell line with 60 ng/ml IFN γ (PeproTech) and DEG analysis were previously described (44). Data can be found under accession number GSE146893.

Results

Gene prioritization identifies 118 likely causal CeD genes.

To identify genes that most likely play a role in CeD (prioritized genes), we combined our recently published genome-wide association meta-analysis (14) with (1) eQTLs derived from whole-blood transcriptomes of 3,503 Dutch individuals (15) and (2) a co-regulation matrix derived from expression data in multiple different tissues and 77,000 gene expression samples (13). We selected 1,258 genes that were within 1Mb of the 58 CeD-associated non-HLA variant regions ($p < 5 \times 10^{-6}$) (see **Methods**), and prioritized the genes that are the most likely related to CeD using four different gene prioritization methods: MR-IVW (11), COLOC (10), LD overlap and DEPICT (13) (**Supplementary Figure 1 B-C**, **Supplementary Table 1**).

The first three methods use summary statistics of eQTLs and the CeD GWAS. MR-IVW is a test that assesses the causal relationship between a gene expression and the disease using the gene eQTLs as instruments (11,12) (see **Methods**). MR-IVW was applied to 162 genes, for which three of more independent *cis*-eQTL variants (at $p < 5 \times 10^{-6}$) were identified (see **Methods**) (27), and shown to be consistent after heterogeneity correction (**Supplementary Table 2**) (**Methods**). COLOC is a variant colocalization test that assesses if there is a shared causal variant in the locus, using all the variants in a locus as instruments (10) (**Methods**). LD overlap is an annotation approach that tests if the most associated variants in a GWAS

A Gene prioritization of CeD loci prioritizes 118 genes potentially causal to CeD.

B Biological interpretation of CeD associated genes.

Figure 1
Celiac disease prioritized genes and their proposed function and cell type.

(A) A chromosome ideogram depicting the location of each prioritized gene identified in a CeD-associated GWAS locus. Loci are marked with red bars. Genes depicted by a square are the target of an approved drug or a drug in development. All other genes are depicted by a circle. Each circle or square is coloured according to the lines of evidence (Methods) supporting its causal role. (B) Functions and cell types highlighted by the prioritized genes, according to our literature review (see Methods) (n = 118 genes, for 37 genes neither a function nor a specific cell type on which the gene may operate could be specified). All genes contributing to a specific function are listed under the sub-heading and coloured according to the change that leads to increased CeD risk: increased expression (red), decreased expression (blue), or undefined (black). The symbols + or - denote if a biological process is thought to be induced or repressed by the gene, respectively, according to literature.

locus are in strong LD with an eQTL (Methods). DEPICT was the fourth method we used to prioritize genes. DEPICT is a gene-prioritization method based on co-regulation in expression datasets across multiple different tissues. DEPICT identifies enrichment for co-regulated genes from genes in a GWAS locus. In contrast to the other methods, DEPICT assessed the potential role of all 1,258 genes independently of the presence of an eQTL.

In total, 118 out of the 1,258 assessed genes were prioritized by at least one of the four methods. Of these 118 genes, 28 had two lines of evidence, 7 genes (*CD226*, *NCF2*, *TRAFD1*, *HM13*, *COLCA1*, *CTSH*, *UBASH3A*) had three lines of evidence, and one gene (*CSK*) was supported by all four methods (Supplementary Table 1) (Figure 1A). Overall, we identified potentially causal genes in 50 out of 58 CeD-associated regions.

The four different gene prioritization methods complement each other in different ways. DEPICT prioritized the most genes: 66 in total, but also had the highest proportion of 38 of them uniquely prioritized (38/66, 58% unique). One reason for this is that DEPICT is based on co-expression, not on eQTLs and thus more genes could be tested. Indeed, 16 genes prioritized by DEPICT do not have a significant eQTL. Overall, the highest concordance between prioritization methods was found between COLOC and LD overlap (30% and 26% unique genes, respectively), while MR-IVW uniquely prioritized a relatively large proportion of genes (9/20, 45% unique). Thus, each method prioritizes genes somewhat similarly, but adds a unique set of genes based on the peculiarities and assumptions of the specific method.

Interestingly, 26 of the 118 prioritized genes are targeted by an approved drug or a drug in development according to DrugBank and Finan et al. (35) (Figure 1A) (Supplementary Table 3) and therefore could lead to therapeutic intervention in CeD. For example, drugs such as Natalizumab

and Basiliximab that target the proteins encoded by *ITGA4* and *IL21R*, respectively, are currently approved or under study for the treatment of immune-mediated diseases including rheumatoid arthritis (45), Crohn's disease (46) and multiple sclerosis (47) or as an immune-suppressor to avoid kidney transplant rejection. An additional 25 genes encode proteins that are similar to proteins targeted by already approved drugs following Finan et al. (35) (**Supplementary Table 3**).

Through a systematic literature we found that these 118 genes may participate in general biological processes such as cell adhesion and proliferation (*C1orf106* and *FASLG*, respectively) (**Figure 1B** and **Supplementary Table 4**) as well as immune associated processes. For instance genes such as *THEMIS*, *IL2*, *CD28*, *CTLA4* and *UBASH3A* are involved in T cell activation and co-stimulation. Other genes including *CCR1*, *CC2* and *IL21* participate in inflammation by activating and recruiting monocytes, dendritic cells, neutrophils and B cells (48). Interestingly, we also find genes that encode for transcription factors (e.g. *IRF4* and *ETS1*) that are essential for the differentiation of T helper 1 (Th1) cells (49,50), the key players in the pathogenesis of CeD.

Co-regulation patterns of cis-eQTL-prioritized loci reveals four functional clusters

We sought to further understand the biological role of these 118 genes using a guilt-by-association co-regulation approach to identify clusters of shared molecular function (see **Methods**). We identified 4 different clusters

Figure 2
Co-expression pattern of cis-eQTL prioritized genes.

(A) Heatmap showing the Spearman correlations between gene expression patterns of each prioritized gene. Blue squares indicate negative correlation. Red squares indicate positive correlation. Both are shaded on a gradient scale according to the Z score of the correlation. A dendrogram computed with Ward distances between the correlations is shown on top of the heatmap. Branches of the dendrogram are coloured differently to mark separate clusters. (B) Results of the REACTOME gene set enrichment analysis of the genes belonging to each of the clusters identified in (A). Colour key denotes the significance (-log 10 multiple testing adjusted *p* value) of each biological pathway.

based on their expression co-regulation with 1,588 principal components that were identified from the co-expression of 31,499 RNA-seq samples across multiple tissues (30) (**Figure 2A, Supplementary Table 5**).

In three out of four co-regulation clusters, we find shared immune function. For instance, *STAT1*, *CD274* and *IL12A* are in the same cluster, and are all involved in interferon gamma (IFN γ) and interleukin (IL-6) signalling, later referred to as the cluster enriched for IFN γ signalling (**Figure 2B, Supplementary Table 6**). In another cluster, we find genes involved in T cell receptor mediated activation and CD28 co-stimulation (e.g. *CD28*, *CTLA4* and *ICOS*), we refer to this cluster as the cluster enriched for TCR activation (**Figure 2B**). In the third cluster we find genes involved in cytokine and chemokine signalling genes, namely *CCR1*, *CCR2*, *CCR3*, *IL2EA*, *IL21* and *IL18R1*: the cluster enriched for chemokine receptor binding (**Figure 2B, Supplementary Table 6**). The biological processes found in these clusters are essential for the activation and function of both the innate and adaptive immune system.

The enrichments found for these clusters do not fully define the biological function of these clusters. However, this co-regulation clustering approach has the benefit that it can assign putative function to otherwise unknown genes. For example, long non-coding RNAs (lncRNAs) genes are usually hard to assign function to, even though they represent ~10% of our prioritized genes. Based on their cluster membership and the principle of guilt-by-association to assign function to unknown genes, most of the prioritized lncRNAs are likely to be involved in cytokine/chemokine signalling (**Figure 2A, B**) (**Supplementary Table 6**).

Mediation analysis uncovers TRAFD1 as a major trans-eQTL regulator

To further understand the potential regulatory function of the 118 prioritized genes, we identified downstream regulatory effects by performing a *trans*-mediation analysis using a two-step approach (**Methods**) (**Supplementary Figure 2A**). We first considered all genes with a *trans*-eQTL ($p < 5 \times 10^{-8}$) located in any of the 58 CeD-associated regions, then performed a mediation analysis by re-assessing the *trans*-eQTL effect after adjusting the expression levels for the expression of the prioritized gene(s) in the same locus.

Of the 497 possible prioritized gene–*trans*-eQTL gene combinations, we found 172 that exhibited significant mediation effects. These combinations map to 13 associated regions and represent 21 unique mediating *cis*-eQTL genes and 79 unique mediated *trans*-eQTL genes (**Supplementary Table 7**). Among all the associated regions, the CeD-associated region on chromosome 12 contained the largest number of both *cis*-mediating genes ($N = 5$) and *trans*-mediated genes ($N = 60$). In this region, *TRAFD1* (which had 3 lines of evidence in our prioritization analysis described in 3.1) mediated more *trans* genes ($N = 41$) than all of the other regional *cis*-regulators and also had the highest mediation impact (average Z-score difference in effect size between mediated and unmediated analysis = 2.79) (**Methods**) (**Supplementary Table 7, Supplementary Figure 2B-C**). Of

chapter 6

A *TRAFD1* mediates 41 genes in *trans*

B Reactome 2016 enrichments of 41 *trans* genes

Figure 3
***trans* genes action and function.**

(A) Circle genomic plot depicting the location of the 41 genes trans-mediated by *TRAFD1*. The three genes also prioritized by our *cis*-eQTL analysis are named (red). (B) Results of the REACTOME gene set enrichment analysis of *TRAFD1*-mediated genes. Colour code denotes the significance ($-\log_{10}$ adjusted p value) of each biological pathway.

note, the top eQTL variant of *TRAFD1* is a missense variant in the nearby gene *SH2B3*. This SNP has previously been described to be a *trans*-eQTL in blood by Westra et al. (51). In a larger *trans*-eQTL analysis, Vosa et al. found 502 significant *trans* genes (22). Furthermore, this missense variant has been also associated to a number of complex traits, including blood cell types and platelets, and autoimmune diseases (52,53). However, our *TRAFD1* results are not confounded by cell type, as cell type composition did not affect the eQTL association of *TRAFD1* in our cohort ($p > 0.044$ for 24 different cell-type traits) (Methods) (Supplementary Table 8). To ensure that the mediated *trans* genes of *TRAFD1* were not also mediated by *SH2B3*, we corrected *TRAFD1* expression levels for *SH2B3* expression levels and re-ran the mediation analysis. Here we found that the mediating effect of *TRAFD1* was still significant for all 41 *trans*-mediated genes and that the median Z-score difference between mediated and unmediated was higher than that of *SH2B3*, although it was slightly attenuated compared to the original *TRAFD1* signal (Supplementary Table 9, Supplementary Figure 2C). Based on these results, we conclude that *TRAFD1* mediates 41 other genes in *trans* independently of *SH2B3* expression (Figure 3A).

Strikingly, three of the *TRAFD1* *trans*-mediated genes – *STAT1*, *CD274* and *PDCD1LG2* – are also prioritized *cis* genes in their respective loci (Figure 3A). These results suggest that the *trans*-mediated *TRAFD1* effects may have an additional additive effect in these CeD-associated loci.

TRAFD1 is a poorly characterized gene that has been suggested to act as a negative regulator of the NF κ B pathway (54). To further elucidate the biological processes in which the 41 *TRAFD1* *trans*-mediated genes could be involved, we performed a REACTOME 2016 gene set enrichment analysis (Supplementary Table 10). Here we found that IFN γ signalling, cytokine signalling and major histocompatibility complex class I (MHC I) antigen processing / presentation are enriched pathways, which points to a role for *TRAFD1* and *TRAFD1* *trans*-mediated genes in antigen presentation and immune response (Figure 3B). Many of the *TRAFD1* *trans*-mediated genes, including *GBP1/2/4/5/6*, *TAP1*, *PSME2*, *PSMB10*, *UBE2L6* and *FBXO6*, seem involved in the processing of antigens via phagocytosis,

A TRAFD1 protein levels upon KD

B qPCR *TRAFD1* RNA levels

C RNA-seq *TRAFD1* RNA levels

D

Figure 4, *in vitro* knockdown of TRAFD1 validates trans-mediation network

THP-1 cell line based knockdown of *TRAFD1* (KD) compared to non-specific siRNA (SCR) and untransfected cells (WT) (A) Western blot based protein levels of *TRAFD1* compared to the B actin control. p -value ≤ 0.0001 (****). (B) qPCR RNA levels and (C) RNA-seq levels of *TRAFD1* expression. (D) Comparative differential expression experiment of *TRAFD1*, comparing the 41 *trans*-mediated genes and *TRAFD1* differential expression in the WT vs. SCR to the SCR vs. KD in the LPS stimulated condition. The WT vs. SCR is 52.8 times less differentially expressed as the SCR vs. KD, indicating that *TRAFD1* KD affects these 41 genes in *trans*.

ubiquitination and proteasomal processing (55,56). However, some of the *trans*-mediated genes may also be involved in IgG-dependent class II MHC antigen presentation, including *FCGR1A*, *e* and *TRIM21* (57). This confirms that IFN signalling, cytokine signalling and antigen processing and presentation are associated to CeD through genetics, independently of the HLA locus (4).

siRNA knockdown of TRAFD1 confirms trans-mediated genes are differentially expressed

We performed a siRNA KD experiment on *TRAFD1* to gain more insights into the biological function of this gene and to validate the *TRAFD1 trans*-mediated genes in a functional assay. We evaluated the transcriptional changes of knocking down *TRAFD1* in the monocyte-like cell line THP-1 under resting conditions (unstimulated) or in the presence of LPS, a known inducer of the NFκB pathway (58). We selected this cell line considering the role of monocytes in modulating of the epithelial barrier function and cytokine secretion in a CeD-specific context (59), as well as for the technical ability to alter the expression of *TRAFD1* using siRNA and the quick response of monocytes to LPS (60).

After siRNA treatment, we observed no significant differences in cell viability or proliferation among the controls (WT and SCR) and the KD treatment (**Supplementary Figure 3A, B**). However, as expected for the KD cell line, we noted a significant reduction in the expression of *TRAFD1* compared to the controls in both WB and qPCR analyses (**Figure 4A-C**). KD of *TRAFD1* was also confirmed in the RNAseq data, with *TRAFD1* expression levels reduced by 41% in unstimulated KD cells compared to unstimulated SCR cells (adjusted $p = 0.004$) and by 34% in LPS-stimulated KD cells compared to LPS-stimulated SCR cells (not significant) (**Supplementary Table 11**). The reduced KD effect upon LPS stimulation is consistent with our expectation that *TRAFD1* acts as negative regulator of the NFκB pathway, which is activated by several stimuli, including LPS (58). Therefore, we considered the KD as successful and neither the transfection method nor a reduced expression of *TRAFD1* had a toxic effect (**Figure 4A-C, Supplementary Figure 3**).

Next, we tested if the 41 *TRAFD1 trans*-mediated genes were more differentially expressed than expected after LPS stimulation (**Figure 4D**). To disentangle differential expression from the co-expression inherently present in a gene expression dataset, we devised a permutation scheme that compared the control (WT vs. SCR) observations with the KD (SCR vs. KD) observations (see **Methods**). This scheme takes into account the co-expression of a gene set, as this co-expression is present in both the control and the experimental observation. After performing 1,000,000 permutations of 42 genes (41 *trans* mediated genes and *TRAFD1*) in the LPS-stimulated comparison, the median test statistic in the control observations was observed 52.8 times more often than in the KD observations (0.264% for WT-SCR vs. 0.005% for SCR-KD, **Supplementary Figure 4**). This indicates that the 41 *trans*-mediated genes and *TRAFD1* as a set are 54 times more differentially expressed than expected. We did not find increased differential expression of the same gene set in the

unstimulated condition (1.120% for WT-SCR vs. 0.307% for SCR-KD, **Figure 4D**, **Supplementary Figure 4**), indicating that *TRAFD1* mainly regulates genes in an LPS-stimulated state.

RNA-seq in CeD relevant cell types identifies tissue specific biological function

To complement our REACTOME gene set enrichment analysis and dig deeper into the biological processes and cell types in which the 118 prioritized genes may act, we analyzed their expression profiles in available RNA-seq datasets from disease-relevant cell types including 1) small intestinal biopsies of active CeD patients and healthy controls, 2) intra-epithelial cytotoxic lymphocytes (IE-CTLs) stimulated with disease-relevant cytokines IL-21, IL-15 and IFN β and 3) gluten specific CD4⁺ T cells (gsCD4⁺ T cells) stimulated with antiCD3-antiCD28, which mimics the disease-specific response to gluten peptides (**Supplementary Figure 5**) (differentially expressed genes for each dataset are available in **Supplementary Table 12**).

We observed that the genes contained in the cluster not showing enrichment for particular biological functions, and those contained in the cluster enriched for IFN γ signalling (green and red, **Figure 2**) identified in the analysis described in 3.2, are most expressed in CeD patients derived small intestinal biopsies (**Supplementary Figure 5A**). Similarly, some *TRAFD1 trans*-mediated genes such as *STAT1*, *CXCL10* and *TAP1*, which are essential for IFN response (61), chemotaxis (62) and antigen processing (63), respectively (**Supplementary Figure 5B**), were found to be upregulated mainly in the biopsy samples derived from CeD patients.

Genes contained in these two clusters were also highly expressed in stimulated IE-CTLs, which is in line with the IFN γ pathway enrichment observed (**Figure 2**). IFN γ is mainly produced by gsCD4⁺ T cells and IE-CTLs and is known to disrupt the integrity of the intestinal epithelial cells in CeD-associated villous atrophy (64–66). Moreover, most *TRAFD1 trans*-mediated genes exhibit an increase in expression in response to IFN γ in intestinal epithelial cells (Caco-2) or IFN β in IE-CTLs (**Supplementary Figure 5B**).

The genes in the co-regulation clusters enriched for TCR signalling and chemokine and cytokine signalling (cyan and pink clusters, **Figure 2**), were instead highly expressed in gsCD4⁺T cells, especially after stimulation with anti-CD3/anti-CD28, indicating that these prioritized genes may be biologically relevant in the immediate T cell receptor response to gluten ingestion (**Supplementary Figure 5A**). In contrast, antiCD3-antiCD28 stimulation in gsCD4⁺T cells resulted in both up- and downregulation of the *TRAFD1 trans*-mediated genes, implying that *TRAFD1 trans*-mediated genes respond more strongly to IFN signalling (IFN γ or IFN β) than to TCR activation by anti-CD3/anti-CD28 (**Supplementary Figure 5B**).

Taken together the gene expression patterns of the 118 prioritized genes, when combined with information from our literature search (**Figure 1B**) (**Supplementary Tables 4** and **12**), suggests that these genes may

control general biological processes (e.g. apoptosis, gene regulation and cytoskeleton remodeling) as well as specific immune functions (e.g. cell adhesion, cell differentiation and TCR signalling) in diverse cell types (e.g. T cells, neutrophils, B cells, monocytes, epithelial cells) (**Supplementary Figure 5** and **Supplementary Table 12**). The non-HLA genetic loci associated to CeD thus seem to affect a complex network of cells and biological processes. Of note, when we analyzed the enrichment of the 41 *TRAFD1* *trans*-mediated genes in significantly differentially expressed genes in CeD relevant cell types, we found that the enrichment was strongest in IE-CTLs and epithelial cells upon IFN signalling (**Supplementary Table 12**), suggesting that *TRAFD1* and *TRAFD1* mediated genes modulate IFN signalling possibly via regulation of NFκB in the context of the CeD inflammatory environment.

Discussion

In the present study we aimed to identify CeD candidate genes using four *in silico* methods (MR-IVW, COLOC, LD overlap and DEPICT) and whole blood transcriptomics data from a population-based cohort. While previous studies have used at least one of these methods (3,9,14,67), to our knowledge this is the first effort that integrates these four different statistical approaches. This systematic prioritization approach resulted in 118 prioritized likely causal genes, including 26 that are direct targets of an approved drug or of drugs under development for other complex diseases, including autoimmune diseases. The co-expression pattern within a large RNA-seq dataset from blood (30) suggests these genes are involved in cytokine signalling in innate and adaptive cells as well as in T-cell receptor activation pathways.

We also ran a *trans*-mediation analysis at CeD loci by incorporating trans-eQTLs; these have mostly been ignored by other prioritization studies which have thus missed long-distance co-regulation interactions (68). With this analysis, we identified one of the 118 prioritized genes, *TRAFD1*, to be *trans*-regulator of 41 genes, a set of genes showing a strong enrichment in IFNγ signalling and MHC I antigen processing/presentation pathways, which are pivotal for the disease pathogenesis. Using si-RNA experiments we confirmed a significant perturbation of these 41 genes after blocking *TRAFD1* expression, a signal that was not seen instead for the 502 genes identified by Vosa et al. (22). *trans*-regulated by one of the *TRAFD1* eQTL variants (data not shown). This may indicate that our mediation analysis restricted the gene set to reflect a more defined role of *TRAFD1* in CeD.

TRAFD1 has not been related to CeD previously (14,67). This gene is thought to be a regulator of the NFκB signalling pathway (54), a pathway that is abnormally activated in the intestinal mucosa of CeD patients (69). Our results suggest for the first time a role of *TRAFD1* in signalling response to potentially both type I and type II interferons. IFNγ-mediated type II interferon and IFNβ-mediated type I interferon activation both involve activation of the NFκB pathway (70–74), and *TRAFD1* and *TRAFD1*-mediated genes respond strongly to both interferons in IE-CTLs

and intestinal epithelial cells. Interestingly, the *IFNG* or *IFNB* locus are not associated with CeD directly. We therefore hypothesize that *TRAFD1* provides a genetic link to the IFN response in CeD patients.

IE-CTLs, which are the effector cells in CeD, have not thus far been directly genetically associated with the disease. However, given the activation of at least half of the 41 *trans*-mediated genes in IE-CTLs upon IFN stimulation, we propose that the IE-CTLs may also be genetically linked to the disease through the action of *TRAFD1*.

We acknowledge the limitations of our study that could have restricted our findings. For example, a drawback is the limited genome coverage of the CeD summary statistics used in this study. Immunochip only measures genotypes in regions known to be associated with immune function. Therefore, our current interpretation of CeD loci is biased toward immune-related mechanisms. Only when comprehensive whole-genome CeD association analyses will become available, can this bias be removed. Another constraint is the lack of a golden standard method to select relevant genes at an associated locus. In our combined gene prioritization approach we observed that the four different statistical methods applied to our data identified unique and jointly prioritized genes. While we believe that the genes prioritized in this study represent robustly prioritized genes for CeD, it is difficult to prove that all the prioritized genes are truly causal based on statistical methodology alone. For instance, in the MR-IVW method for causality, it is impossible to rule out pleiotropic effects, even though we do find that our estimates are not overly heterogeneous (75). Moreover, for many genes there were not enough eQTLs to apply this causality test. Functional validation of these genes in disease context is needed to rule out false positives. In this study we used eQTLs derived from whole blood, which is a relevant tissue in CeD, but it will not capture all causal genes across different tissues (76). We have attempted to mitigate this bias by including DEPICT in our analyses, which uses co-expression data from many tissues regardless of eQTLs. Tissue limitations also apply to the translation of the prioritized genes to the differential expression experiments performed in the disease relevant cell types. Genes found in blood do not directly translate to other cell types.

In conclusion, this study provides a framework for predicting candidate genes and their function using a systematic *in silico* approach that could be extended to other complex diseases. Using this approach, we not only confirmed previous association between adaptive cells (gsCD4⁺ T cells and B cells) and CeD, but we also unveil a link between specific genes that may contribute to the disease via innate immune cells, epithelial cells and IE-CTLs. Finally, we identified a novel master regulator, *TRAFD1*, influencing a set of genes enriched for two major pathways of immune activation IFN γ signalling and antigen processing, which could thus be a potential target for therapeutic interventions in CeD.

Supplementary information

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2020.562434/full#supplementary-material>

Nomenclature

Underlined words are definitions that have been explained in the preceding lines.

eQTL – expression quantitative trait locus, a location on the genome that is statistically associated to changes in gene expression.

***cis*-eQTL** – an eQTL located in the same locus of the gene that is being interrogated (within 1.5Mb from gene transcript start or end).

***trans*-eQTL** – an eQTL that is not physically close to the gene that is being interrogated (>1.5Mb from transcript start/end or on a different chromosome).

***cis*-eQTL gene** – a gene that is associated with a change in expression as a consequence of a *cis*-eQTL.

***trans*-eQTL gene** – a gene that is associated with a change in expression as a consequence of a *trans*-eQTL.

CeD – celiac disease

CeD-associated region – a genomic region that is associated to CeD based on results from genome-wide association studies on CeD.

Prioritized gene – a gene prioritized as being potentially causal for CeD according to the four statistical methods depicted in **Supplementary Figure 1B-C**. In this study, prioritized genes are always within the CeD-associated regions.

Mediating *cis* gene – a prioritized gene that is statistically responsible for the change in expression of a *trans*-eQTL gene. Of note, while the *trans*-eQTL is located in the same CeD-associated region of the mediating *cis*-gene, the mediated *trans*-gene is not.

Mediated *trans* gene – a gene located outside CeD-associated regions that is statistically mediated by a mediating *cis* gene located in the same region of the corresponding *trans*-eQTL.

Author contributions

A.v.d.G., M.Z., S.S. and I.J. designed the study, analyzed and interpreted results. A.v.d.G., M.Z., U.V., R.A.G., I.R.P., S.S. and I.J. performed data quality control and statistical analysis. M.Z., C.L., J.M., Z.B. performed *in vitro* experiments. A.v.d.G. F.K., Y.K.W., L.S., S.W.Q., B.C., V.K., Y.L., L.F., S.W., C.W. S.S. and I.J. contributed data to and performed *in silico* experiments for this study. A.v.d.G., M.Z. S.S. and I.J. wrote the manuscript with critical input from V.K., Y.L., L.F., S.W. and C.W. All authors read and approved the manuscript.

Funding

This work was supported by an ERC Advanced grant [FP/2007-2013/ERC grant 2012-322698] and an NWO Spinoza prize grant [NWO SPI 92-266] to C.W. As well as a grant from Stiftelsen K.G. Jebsen to C.W., S.W.Q. and L.M.S. [project SKGJ-MED-017]. I.J. is supported by a Rosalind Franklin Fellowship from the University of Groningen and an NWO VIDI grant [016.171.047].

Acknowledgements

We are very grateful for the altruistic donation of biological materials by the study participants, without them this study would not be possible. In addition, we thank the UMCG Genomics Coordination center, and the UG Center for Information Technology, and their sponsors BBMRI-NL & TarGet, for storage and computing infrastructure. We thank BBMRI-NL for providing the transcriptome and genotyped data for the BIOS cohort. We would like to thank Prof. Bana Jabri for providing IE-CTL cell lines, Prof. Morris Swertz for data storage and cluster facilities and Kate McIntyre for editing the manuscript. A preliminary version of this manuscript has been released as a pre-print at BioRxiv (77)

References

1. Kuja-Halkola R, Lebowitz B, Halfvarson J, Wijmenga C, Magnusson PKE, Ludvigsson JF. Heritability of non-HLA genetics in coeliac disease: A population-based study in 107 000 twins. *Gut*. 2016;65(11):1793–8.
2. Bevan S, Popat S, Braegger CP, Busch A, O'Donoghue D, Falth-Magnusson K, et al. Contribution of the MHC region to the familial risk of coeliac disease. *J Med Genet*. 1999;36(9):687–90.
3. Trynka G, Hunt KA, Bockett NA, Romanos J, Mistry V, Szperl A, et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat Genet*. 2011;43(12):1193–201.
4. Kumar V, Gutierrez-Achury J, Kanduri K, Almeida R, Hrdlickova B, Zhernakova D V., et al. Systematic annotation of celiac disease loci refines pathological pathways and suggests a genetic explanation for increased interferon-gamma levels. *Hum Mol Genet*. 2015;24(2):397–409.
5. Ricaño-Ponce I, Zhernakova D V., Deelen P, Luo O, Li X, Isaacs A, et al. Refined mapping of autoimmune disease associated genetic variants with gene expression suggests an important role for non-coding RNAs. *J Autoimmun*. 2016;68:62–74.
6. Dubois PCA, Trynka G, Franke L, Hunt KA, Romanos J, Curtotti A, et al. Multiple common variants for celiac disease influencing immune gene expression. *Nat Genet* [Internet]. 2010 Apr [cited 2019 Oct 19];42(4):295–302. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/20190752>
7. Zhernakova A, Stahl EA, Trynka G, Raychaudhuri S, Festen EA, Franke L, et al. Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci. *PLoS Genet*. 2011;7(2).
8. Jonkers IH, Wijmenga C. Context-specific effects of genetic variants associated with autoimmune disease. Vol. 26, *Human Molecular Genetics*. 2017. p. R185–92.
9. Fernandez-Jimenez N, Bilbao JR. Mendelian randomization analysis of celiac GWAS reveals a blood expression signature with diagnostic potential in absence of gluten consumption. *Hum Mol Genet*. 2019;28(18):3037–42.
10. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. Williams SM, editor. *PLoS Genet* [Internet]. 2014 May 15 [cited 2019 Oct 19];10(5):e1004383. Available from: <https://dx.plos.org/10.1371/journal.pgen.1004383>
11. Burgess S, Butterworth A, Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet Epidemiol* [Internet]. 2013 Nov [cited 2018 Jul 19];37(7):658–65. Available from: <https://onlinelibrary.wiley.com/doi/abs/10.1002/gepi.21758>

12. van der Graaf A, Claringbould A, Rimbert A, Heijmans BT, Hoen PAC', van Meurs BJ, et al. Mendelian randomization while jointly modeling cis genetics identifies causal relationships between gene expression and lipids. *Nat Commun* [Internet]. 2020 Dec 1 [cited 2020 Oct 13];11(1):1–12. Available from: <https://doi.org/10.1038/s41467-020-18716-x>
13. Pers TH, Karjalainen JM, Chan Y, Westra HJ, Wood AR, Yang J, et al. Biological interpretation of genome-wide association studies using predicted gene functions. *Nat Commun* [Internet]. 2015 May 19 [cited 2019 Oct 19];6(1):5890. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/25597830>
14. Ricaño-Ponce I, Gutierrez-Achury J, Costa AF, Deelen P, Kurilshikov A, Zorro MM, et al. Immunochip meta-analysis in European and Argentinian populations identifies two novel genetic loci associated with celiac disease. *Eur J Hum Genet* [Internet]. 2020 Oct 7 [cited 2019 Oct 22];28(3):313–23. Available from: <http://www.nature.com/articles/s41431-019-0520-4>
15. Zhernakova D V., Deelen P, Vermaat M, Van Iterson M, Van Galen M, Arindarto W, et al. Identification of context-dependent expression quantitative trait loci in whole blood. *Nat Genet*. 2017;49(1):139–45.
16. van Greevenbroek MMJ, Jacobs M, van der Kallen CJH, Vermeulen VMMJ, Jansen EHJM, Schalkwijk CG, et al. The cross-sectional association between insulin resistance and circulating complement C3 is partly explained by plasma alanine aminotransferase, independent of central obesity and general inflammation (the CODAM study). *Eur J Clin Invest* [Internet]. 2011 [cited 2019 Feb 11];41(4):372–9. Available from: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2362.2010.02418.x>
17. Deelen J, van den Akker EB, Trompet S, van Heemst D, Mooijaart SP, Slagboom PE, et al. Employing biomarkers of healthy ageing for leveraging genetic studies into human longevity. *Exp Gerontol*. 2016;82:166–74.
18. Tigchelaar EF, Zhernakova A, Dekens JAM, Hermes G, Baranska A, Mujagic Z, et al. Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: Study design and baseline characteristics. *BMJ Open* [Internet]. 2015 Aug [cited 2018 Nov 7];5(8):e006772. Available from: <https://bmjopen.bmj.com/content/5/8/e006772>
19. Lin BD, Willemsen G, Abdellaoui A, Bartels M, Ehli EA, Davies GE, et al. The Genetic Overlap between Hair and Eye Color. *Twin Res Hum Genet*. 2016;19(6):595–9.
20. Huisman MHB, De Jong SW, Van Doormaal PTC, Weinreich SS, Schelhaas HJ, Van Der Kooi AJ, et al. Population based epidemiology of amyotrophic lateral sclerosis using capture-recapture methodology. *J Neurol Neurosurg Psychiatry*. 2011 Oct;82(10):1165–70.
21. Hofman A, Brusselle GGO, Murad SD, van Duijn CM, Franco OH, Goedegebure A, et al. The Rotterdam Study: 2016 objectives and design update. *Eur J Epidemiol* [Internet]. 2015 [cited 2018 Nov 7];30(8):661–708. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4579264/>
22. Vösa U, Claringbould A, Westra H-J, Bonder MJ, Deelen P, Zeng B, et al. Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. *bioRxiv* [Internet]. 2018 Oct [cited 2018 Nov 7];18:447367. Available from: <https://www.biorxiv.org/content/10.1101/447367v1>
23. Das S, Forer L, Schunherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. *Nat Genet* [Internet]. 2016 Oct [cited 2019 Feb 11];48(10):1284–7. Available from: <https://www.nature.com/articles/ng.3656>
24. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience*. 2015;4(1):7.
25. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: A tool for genome-wide complex trait analysis. *Am J Hum Genet*. 2011;88(1):76–82.
26. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, et al. A global reference for human genetic variation. *Nature* [Internet]. 2015 Oct [cited 2016 Feb 17];526(7571):68–74. Available from: <http://www.nature.com/nature/journal/v526/n7571/full/nature15393.html>
27. Yang J, Ferreira T, Morris AP, Medland SE, Madden PAF, Heath AC, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics

- identifies additional variants influencing complex traits. *Nat Genet* [Internet]. 2012 Apr [cited 2018 Feb 6];44(4):369–75. Available from: <https://www.nature.com/articles/ng.2213>
28. Burgess S, Thompson SG. Interpreting findings from Mendelian randomization using the MR-Egger method. *Eur J Epidemiol*. 2017;32(5):377–89.
29. Bowden J, Hemani G, Davey Smith G. Invited Commentary: Detecting Individual and Global Horizontal Pleiotropy in Mendelian Randomization-A Job for the Humble Heterogeneity Statistic? *Am J Epidemiol*. 2018 Sep;187(12):2681–5.
30. Deelen P, van Dam S, Herkert JC, Karjalainen JM, Brugge H, Abbott KM, et al. Improving the diagnostic yield of exome-sequencing by predicting gene-phenotype associations using large-scale gene expression analysis. *Nat Commun*. 2019;10(1).
31. Freedman LS, Schatzkin A. Sample size for studying intermediate endpoints within intervention trials or observational studies. *Am J Epidemiol* [Internet]. 1992 Nov 1 [cited 2019 Oct 19];136(9):1148–59. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/1462974>
32. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles G V., et al. Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* [Internet]. 2013;14(1):128. Available from: <https://doi.org/10.1186/1471-2105-14-128>
33. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Res*. 2018;46(D1):D1074–82.
34. Whirl-Carrillo M, McDonagh EM, Hebert JM, Gong L, Sangkuhl K, Thorn CF, et al. Pharmacogenomics knowledge for personalized medicine. *Clin Pharmacol Ther* [Internet]. 2012 Oct [cited 2019 Oct 13];92(4):414–7. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/22992668>
35. Finan C, Gaulton A, Kruger FA, Lumbers RT, Shah T, Engmann J, et al. The druggable genome and support for target identification and validation in drug development. *Sci Transl Med* [Internet]. 2017 Mar 29;9(383):eaag1166. Available from: <http://stm.sciencemag.org/content/9/383/eaag1166.abstract>
36. Kim D, Langmead B, Salzberg SL. HISAT: A fast spliced aligner with low memory requirements. *Nat Methods* [Internet]. 2015 Mar 9;12(4):357–60. Available from: <https://doi.org/10.1038/nmeth.3317>
37. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* [Internet]. 2009 Aug 15 [cited 2019 Sep 21];25(16):2078–9. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/19505943>
38. Liao Y, Smyth GK, Shi W. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res*. 2019;47(8).
39. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol* [Internet]. 2010 [cited 2019 Sep 21];11(10):R106. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/20979621>
40. Zorro MM, Aguirre-Gamboa R, Mayassi T, Ciszewski C, Barisani D, Hu S, et al. Tissue alarmins and adaptive cytokine induce dynamic and distinct transcriptional responses in tissue-resident intraepithelial cytotoxic T lymphocytes. *J Autoimmun* [Internet]. 2020 Feb 4 [cited 2020 Feb 24];108:102422. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0896841120300330>
41. Jabri B, De Serre NPM, Cellier C, Evans K, Gache C, Carvalho C, et al. Selective expansion of intraepithelial lymphocytes expressing the HLA-E- specific natural killer receptor CD94 in celiac disease. *Gastroenterology*. 2000;118(5):867–79.
42. Kim D, Langmead B, Salzberg SL. HISAT: A fast spliced aligner with low memory requirements. *Nat Methods*. 2015 Mar;12(4):357–60.
43. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* [Internet]. 2014 Dec [cited 2015 Jun 12];15(12):550. Available from: <http://genomebiology.com/2014/15/12/550/abstract>
44. Zorro Manrique MM. Celiac disease: From genetic variation to molecular culprits. University of Groningen; 2020.
45. Chiu YG, Ritchlin CT. Denosumab: targeting the RANKL pathway to treat

- rheumatoid arthritis. *Expert Opin Biol Ther* [Internet]. 2017 Jan 2 [cited 2019 Oct 7];17(1):119–28. Available from: <https://www.tandfonline.com/doi/full/10.1080/14712598.2017.1263614>
46. Rutgeerts P, Vermeire S, Van Assche G. Biological Therapies for Inflammatory Bowel Diseases. *Gastroenterology* [Internet]. 2009 Apr [cited 2019 Oct 7];136(4):1182–97. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/19249397>
47. Baldassari LE, Rose JW. Daclizumab: Development, Clinical Trials, and Practical Aspects of Use in Multiple Sclerosis. *Neurotherapeutics* [Internet]. 2017 Oct 13 [cited 2019 Oct 7];14(4):842–58. Available from: <http://link.springer.com/10.1007/s13311-017-0553-8>
48. Esche C, Stellato C, Beck LA. Chemokines: Key players in innate and adaptive immunity. *J Invest Dermatol*. 2005;125(4):615–28.
49. Mahnke J, Schumacher V, Ahrens S, Kading N, Feldhoff LM, Huber M, et al. Interferon Regulatory Factor 4 controls TH1 cell effector function and metabolism. *Sci Rep* [Internet]. 2016 Dec 20 [cited 2019 Oct 19];6(1):35521. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/27762344>
50. Grenningloh R, Bok YK, Ho IC. Ets-1, a functional cofactor of T-bet, is essential for Th1 inflammatory responses. *J Exp Med* [Internet]. 2005 Feb 21 [cited 2019 Oct 19];201(4):615–26. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/15728239>
51. Westra HJ, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet* [Internet]. 2013 Oct [cited 2015 Oct 21];45(10):1238–43. Available from: <http://www.nature.com.proxy-ub.rug.nl/ng/journal/v45/n10/full/ng.2756.html>
52. Westra HJ, Martínez-Bonet M, Onengut-Gumuscu S, Lee A, Luo Y, Teslovich N, et al. Fine-mapping and functional studies highlight potential causal variants for rheumatoid arthritis and type 1 diabetes. *Nat Genet* [Internet]. 2018 Oct 17 [cited 2020 Feb 25];50(10):1366–74. Available from: <http://www.nature.com/articles/s41588-018-0216-7>
53. Astle WJ, Elding H, Jiang T, Allen D, Ruklisa D, Mann AL, et al. The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell*. 2016 Nov;167(5):1415–1429.e19.
54. Sanada T, Takaesu G, Mashima R, Yoshida R, Kobayashi T, Yoshimura A. FLN29 deficiency reveals its negative regulatory role in the toll-like receptor (TLR) and retinoic acid-inducible gene I (RIG-I)-like helicase signaling pathway. *J Biol Chem* [Internet]. 2008 Dec 5 [cited 2019 May 17];283(49):33858–64. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/18849341>
55. Wlynebreten I. Involvement of autophagy in MHC class I antigen presentation. *Scand J Immunol* [Internet]. 2020 Nov 19 [cited 2020 Nov 1];92(5). Available from: <https://onlinelibrary.wiley.com/doi/10.1111/sji.12978>
56. Meunier E, Broz P. Interferon-inducible GTPases in cell autonomous and innate immunity. *Cell Microbiol* [Internet]. 2016 Feb 1 [cited 2020 Nov 1];18(2):168–80. Available from: <http://doi.wiley.com/10.1111/cmi.12546>
57. Lu LL, Suscovich TJ, Fortune SM, Alter G. Beyond binding: Antibody effector functions in infectious diseases [Internet]. Vol. 18, *Nature Reviews Immunology*. Nature Publishing Group; 2018 [cited 2020 Nov 1]. p. 46–61. Available from: <https://www.nature.com/articles/nri.2017.106>
58. Dorrington MG, Fraser IDC. NF-κB signaling in macrophages: Dynamics, crosstalk, and signal integration. *Front Immunol* [Internet]. 2019 [cited 2019 Sep 21];10(APR):705. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/31024544>
59. Delbue D, Cardoso-Silva D, Branchi F, Itzlinger A, Letizia M, Siegmund B, et al. Celiac Disease Monocytes Induce a Barrier Defect in Intestinal Epithelial Cells. *Int J Mol Sci* [Internet]. 2019 Nov 9 [cited 2020 Oct 13];20(22):5597. Available from: <https://www.mdpi.com/1422-0067/20/22/5597>
60. Sharif O, Bolshakov VN, Raines S, Newham P, Perkins ND. Transcriptional profiling of the LPS induced NF-κB response in macrophages. *BMC Immunol* [Internet]. 2007 Jan 12 [cited 2020 Oct 14];8(1):1. Available from: <http://bmccimmunol.biomedcentral.com/articles/10.1186/1471-2172-8-1>
61. Kim HS, Kim DC, Kim HM, Kwon HJ, Kwon SJ, Kang SJ, et al. STAT1 deficiency

redirects IFN signalling toward suppression of TLR response through a feedback activation of STAT3. *Sci Rep.* 2015;5.

62. Majumder S, Bhattacharjee S, Paul Chowdhury B, Majumdar S. CXCL10 Is Critical for the Generation of Protective CD8 T Cell Response Induced by Antigen Pulsed CpG-ODN Activated Dendritic Cells. *PLoS One.* 2012;7(11).
63. Seyffer F, Tampai R. ABC transporters in adaptive immunity. Vol. 1850, *Biochimica et Biophysica Acta - General Subjects.* 2015. p. 449–60.
64. Abadie V, Discepolo V, Jabri B. Intraepithelial lymphocytes in celiac disease immunopathology. *Semin Immunopathol.* 2012;34(4):551–6.
65. Nilsen EM, Lundin KEA, Krajci P, Scott H, Sollid LM, Brandtzaeg P. Gluten specific, HLA-DQ restricted T cells from coeliac mucosa produce cytokines with Th1 or Th0 profile dominated by interferon γ . *Gut.* 1995;37(6):766–76.
66. Wapenaar MC, Van Belzen MJ, Fransen JH, Fariña Sarasqueta A, Houwen RHJ, Meijer JWR, et al. The interferon gamma gene in celiac disease: Augmented expression correlates with tissue damage but no evidence for genetic susceptibility. *J Autoimmun.* 2004;23(2):183–90.
67. Withoff S, Li Y, Jonkers I, Wijmenga C. Understanding Celiac Disease by Genomics [Internet]. *Trends in Genetics* May, 2016 p. 295–308. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/26972670>
68. Brynedal B, Choi JM, Raj T, Bjornson R, Stranger BE, Neale BM, et al. Large-Scale trans-eQTLs Affect Hundreds of Transcripts and Mediate Patterns of Transcriptional Co-regulation. *Am J Hum Genet* [Internet]. 2017 Apr [cited 2019 Oct 19];100(4):581–91. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S000292971730071X>
69. Fernandez-jimenez N, Castellanos-rubio A, Plaza-izurieta L, Irastorza I, Elcoroaristizabal X, Jauregi-miguel A, et al. Coregulation and modulation of NF κ B-related genes in celiac disease: Uncovered aspects of gut mucosal inflammation. *Hum Mol Genet* [Internet]. 2014 Mar 1 [cited 2019 Oct 19];23(5):1298–310. Available from: <https://academic.oup.com/hmg/article-lookup/doi/10.1093/hmg/ddt520>
70. Deb A, Haque SJ, Mogensen T, Silverman RH, Williams BRG. RNA-Dependent Protein Kinase PKR Is Required for Activation of NF- κ B by IFN- γ in a STAT1-Independent Pathway. *J Immunol.* 2001;166(10):6170–80.
71. Mitchell S, Mercado EL, Adelaja A, Ho JQ, Cheng QJ, Ghosh G, et al. An NF κ B Activity Calculator to Delineate Signaling Crosstalk: Type I and II Interferons Enhance NF κ B via Distinct Mechanisms. *Front Immunol* [Internet]. 2019 Jun 25 [cited 2020 Oct 23];10(JUN):1425. Available from: <https://www.frontiersin.org/article/10.3389/fimmu.2019.01425/full>
72. Pfeffer LM. The role of nuclear factor κ b in the interferon response [Internet]. Vol. 31, *Journal of Interferon and Cytokine Research. J Interferon Cytokine Res;* 2011 [cited 2020 Oct 23]. p. 553–9. Available from: <https://pubmed.ncbi.nlm.nih.gov/21631354/>
73. Thapa RJ, Basagoudanavar SH, Nogusa S, Irrinki K, Mallilankaraman K, Sliker MJ, et al. NF- B Protects Cells from Gamma Interferon-Induced RIP1-Dependent Necroptosis. *Mol Cell Biol* [Internet]. 2011 Jul 15 [cited 2020 Oct 23];31(14):2934–46. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC21576359/?tool=EBI>
74. Meyerovich K, Fukaya M, Terra LF, Ortis F, Eizirik DL, Cardozo AK. The non-canonical NF- κ B pathway is induced by cytokines in pancreatic beta cells and contributes to cell death and proinflammatory responses in vitro. *Diabetologia* [Internet]. 2016 Mar 1 [cited 2020 Oct 23];59(3):512–21. Available from: <http://imagej.nih.gov/>
75. Burgess S, Thompson SG. Mendelian randomization: Methods for using genetic variants in causal estimation [Internet]. *Mendelian Randomization: Methods for Using Genetic Variants in Causal Estimation.* CRC Press; 2015 [cited 2018 Dec 3]. 1–207 p. Available from: <https://www.crcpress.com/Mendelian-Randomization-Methods-for-Using-Genetic-Variants-in-Causal-Estimation/Burgess-Thompson/p/book/9781466573178>
76. Ongen H, Brown AA, Delaneau O, Panousis NI, Nica AC, Dermitzakis ET. Estimating the causal tissues for complex traits and diseases. *Nat Genet* [Internet]. 2017 Dec [cited 2018 Oct 16];49(12):1676–83. Available from: <https://www.nature.com/articles/ng.3981>

chapter 6

77. Graaf A van der, Zorro M, Claringbould A, Vosa U, Aguirre-Gamboa R, Li C, et al. Systematic prioritization of candidate genes in disease loci identifies TRAFD1 as a master regulator of IFN γ signalling in celiac disease. *bioRxiv*. 2020 Mar 5;2020.03.04.973487.

Adriaan van der Graaf, Anniq Claringbould, Antoine Rimbert, BIOS consortium*, Harm-Jan Westra, Yang Lit, Cisca Wijmenga†, and Serena Sanna†

chapter 7

Mendelian randomization while jointly modeling *cis* genetics identifies causal relationships between gene expression and lipids

Inference of causality between gene expression and complex traits using Mendelian Randomization (MR) is confounded by pleiotropy and linkage disequilibrium (LD) of gene expression quantitative trait loci (eQTL). Here we propose an MR method, MR-link, that accounts for unobserved pleiotropy and LD by leveraging information from individual-level data, even when only one eQTL variant is present. In simulations, MR-link shows false positive rates close to expectation (median 0.05) and high power (up to 0.89), outperforming all other tested MR methods and coloc. Application of MR-link to low-density lipoprotein cholesterol (LDL-C) measurements in 12,449 individuals with expression and protein QTL summary statistics from blood and liver identifies 25 genes causally linked to LDL-C. This includes the known *SORT1* and *ApoE* genes as well as *PVRL2*, located in the *APOE* locus, for which a causal role in liver was not known. Our results showcase the strength of MR-link for transcriptome-wide causal inferences.

Introduction

Mendelian randomization (MR) is a method that can infer causal relationships between two heritable complex traits from observational studies (1,2). In recent years, MR has gained popularity in the epidemiological field and has provided valuable insights into the risk factors that cause diseases and complex traits (1–3). MR has, for example, successfully identified causal relationships between low-density lipoprotein cholesterol (LDL-C) and coronary artery disease, in turn informing therapeutic strategies (4,5). MR analysis has also shown that a causal relationship between high-density lipoprotein cholesterol (HDL-C) and coronary artery disease is unlikely, which is in contrast to previous epidemiological associations (6). The same approach has been applied to identify molecular marks that are causal to disease (7–10). Since gene expression is one of these marks, investigating its causal role in complex traits is of particular interest given that complex trait loci are enriched for expression quantitative trait loci (eQTLs) (11).

MR infers a causal relationship between an exposure (e.g. a risk factor) and an outcome (e.g. a complex trait) by leveraging QTL variants of the exposure as instrumental variables (IVs). The mathematical model behind MR relies on three main assumptions to correctly infer causality: the IVs have to be i) associated with the exposure, ii) independent of any confounder of the exposure–outcome association and iii) conditionally independent of the outcome given the exposure and confounders. One major challenge of applying MR to gene expression is correcting for deviations from the third assumption, which can occur in the presence of linkage disequilibrium (LD) between the eQTL variants used as IVs, or

Adapted from van der Graaf et al. Mendelian randomization while jointly modeling cis genetics identifies causal relationships between gene expression and lipids, Nat. Comms. 11, (2020).

†: Equal contribution last
*: *A list of authors and their affiliations appears at the end of this paper

Figure 1
Graphical representation of the study

The Biobank Integrative Omics Study (BIOS) cohort was used to identify expression quantitative trait loci (eQTLs) and characterize the genetic architecture of gene expression. Dashed outbox: Knowledge used in a simulation scheme that mimicked gene expression traits, including linkage disequilibrium (LD) between eQTL single nucleotide polymorphism (SNPs). We used this simulation to assess the false positive rates and power for widely used Mendelian randomization (MR) methods. We applied our MR method, MR-link, to both the simulations and to individual-level data of low density lipoprotein cholesterol (LDL-C) in 12,449 individuals (Lifelines) combined with BIOS and GTEx eQTL as well as protein quantitative trait loci (pQTL) summary statistics to identify gene-expression changes and protein level changes that are causally linked to LDL-C within or outside a genome-wide association study (GWAS) locus.

in the presence of pleiotropy, i.e. when IVs affect the outcome through pathways other than the exposure of interest. Accounting for LD is necessary when gene expression is the exposure trait in MR because, in contrast to the majority of complex traits, the genetic architecture of gene expression is characterized by the presence of strong-acting eQTLs located proximal to their transcript (in *cis*), which are often correlated through LD (12,13). On top of this, the presence of pleiotropy cannot be excluded *a priori* given that the majority of variants in our genome are likely to affect one or multiple phenotypes (14–16). There are MR methods (7,17–21) that extend standard MR analysis to correct for LD and pleiotropy, however the application of these methods is not optimal because they require either the removal of pleiotropic IVs from the statistical model (7,19,20), that all sources of pleiotropy are measured and incorporated into the model (22,23), or that both the exposure and the outcome are measured in the same cohort (21). These constraints limit robust inference of gene expression traits as there are often only a limited number of IVs (i.e. eQTL variants) available, and subsequent removal of outliers will substantially reduce power. Likewise, it is not always possible to measure all sources of pleiotropy because it could come from expression of a gene in a different tissue or even from other unobserved molecular marks or phenotypes.

Here we introduce MR-link, an MR method that allows for causal inference in the presence of LD and an unobserved pleiotropic effect, without requiring the removal of pleiotropic IVs or measuring all sources of pleiotropy. MR-link uses summary statistics of an exposure combined with individual-level data on the outcome to estimate the causal effect of an exposure from IVs (i.e. eQTLs if the exposure is gene expression), while at the same time correcting for pleiotropic effects using genetic variants

that are in LD with these IVs (*cis*-genetics) (**Figure 1**).

We assess the performance of MR-link using simulated data in 100 different scenarios that mimic the genetic architecture of gene expression. This information is derived from eQTL association patterns in a large cohort of samples with genetic and transcriptomics data (13). Subsequently we apply MR-link to individual-level data for LDL-C measurements in 12,449 individuals with four different eQTL summary statistic datasets: blood eQTLs identified in the BIOS cohort (**Figure 1**) and eQTLs from blood, liver and cerebellum from the GTEx Consortium (24) (**Figure 1**). We further explore the performance of MR-link on another molecular layer, protein levels, through the application of MR-link on protein quantitative trait loci (pQTL) summary statistics from Sun et al. combined with our LDL-C measurements (25) Our results in simulated and real data show that MR-link can robustly identify causal relationships between molecular traits - such as gene expression and protein levels - and an outcome (e.g. a complex trait), even when the information for causal inference is very limited.

Figure 2
Typical scenarios of pleiotropy in causal inference of gene expression changes as an exposure

Typical scenarios to consider when performing causal inference in gene expression: (A) expression quantitative trait locus (eQTL) single nucleotide polymorphisms (SNPs) used as instrumental variables (IVs) for the same gene (exposure) are in linkage disequilibrium (LD) and pleiotropic effects are absent, (B) pleiotropy is present through LD between IVs for different exposures (pleiotropy through LD) and (C) pleiotropy is present through overlap of the IVs (pleiotropy through overlap). In each panel, the left image shows the genomic context while the right image is a schematic diagram of the corresponding causal effects. Please note that the unobserved exposure trait does not necessarily need to be a protein product: it could be any measured or unmeasured phenotype that is regulated by the genetic locus.

Results

eQTL variants between different genes are often in LD

In a standard MR analysis, IVs need to be independent (not in LD) and have to affect the outcome only through the exposure (absence of pleiotropy). Even in absence of pleiotropy, correlated IVs in the *cis* locus may negatively influence an MR analysis (**Figure 2A**). We distinguish two types of pleiotropy: (i) pleiotropic variants that are in LD with an IV

(pleiotropy through LD, **Figure 2B**) and (ii) when the IV and the pleiotropic variant are the same and affect the outcome through two distinct mechanisms (pleiotropy through overlap **Figure 2C**). If pleiotropy through LD is prevalent, genetic variants in the *cis* region other than those selected as IVs can be used to explain the pleiotropic effects. Incorporating these variants in an MR model can then account for this pleiotropy through LD (**Figure 2B**).

We investigated how often pleiotropy through LD occurs in gene expression by looking at how frequently eQTL variants are shared between genes in *cis*. Using data from the BIOS cohort, a cohort of 3,503 Dutch individuals whose genome and whole blood transcriptome has been characterized (**Figure 1**), we searched for eQTLs located 1.5 megabases (Mb) on both sides of the translated region of 19,960 genes (Methods) (26). We then applied a summary statistics–based stepwise linear regression approach (GCTA-COJO) to identify jointly significant variants, e.g. one or more variants that jointly associate significantly with expression changes of a gene (27) (Methods). We observed that 54% of the genes with an eQTL at $p < 5 \times 10^{-8}$ (13,778 genes) had two or more jointly significant eQTL variants at $p < 5 \times 10^{-8}$ (Methods) (**Figure 1** and **Figure 2A**). These genetic effects were mostly non-overlapping: only 13.4% of the genes have overlapping top eQTL variants ($r^2 > 0.99$). In contrast, genetic variants regulating gene expression of a gene were very often in LD with other eQTLs: 40.6% of top variants are in LD ($r^2 > 0.5$) between genes, and this percentage increased to 60.3% if all jointly significant eQTL variants were considered (Methods).

To strengthen our inferences on the genetic regulation of gene expression in *cis*, we performed statistical fine-mapping using FINEMAP v1.3.1 (28) on 13,276 genes (Methods). Only 373 (2.8%) genes have full eQTL overlap (all variants in the top configuration of a gene are identical or in high LD ($r^2 > 0.99$)), while 33.2% of the genes have at least one variant in $r^2 > 0.5$ LD with a variant in the top configuration of another gene. These percentages are higher for configurations with larger posterior inclusion probabilities (Methods) (**Supplementary Data 1**), but overall the results are similar to our observations from the GCTA-COJO analysis, i.e. the genetics of gene expression in whole blood is mostly regulated by variants that do not overlap but are in moderate LD with variants associated with gene expression changes of another gene. Based on these results, it seems likely that pleiotropy through LD is more common than pleiotropy through overlap in gene expression traits.

MR-link outperforms other methods in discriminative ability

We have developed an MR-method, MR-link, that uses the genetic region surrounding IVs as a covariate to correct for pleiotropic effects (Methods, **Figure 2** and **Supplementary Note 1**). The model underlying MR-link is informed by the observation that the genetic regulation of gene expression is characterized mostly by eQTLs that are in LD, but not overlapping, between genes. This suggests that the variants in the genetic vicinity of the

chapter 7

IVs can be used to correct for pleiotropic effects. MR-link gathers information from genetic variants in LD with an IV to jointly model the outcome through the IVs and their genetic vicinity (Methods). Compared to other MR methods that require summary statistics of both the exposure and the outcome (two-sample MR), our approach adds a requirement of individual-level data for the outcome, but has the advantage that it can perform causal inference even when only a single IV is available. Strictly speaking, MR-link corrects for pleiotropy under the assumption that pleiotropy can be better explained by variants in LD with the IV (pleiotropy through LD) (**Figure 2B**) and that pleiotropy through overlap is absent (**Figure 2C**). In case of a single IV, this assumption needs to be fully accounted for, but when multiple IVs are available, this assumption can be relaxed somewhat. Differences in effect-sizes between IVs can be used to distinguish the causal effect of interest from a pleiotropic effect in the same way that multivariable MR corrects for pleiotropy (22). Of note, MR-link does not require the source of pleiotropy to be specified in the model; MR-link can account for pleiotropic effects arising from, for instance, gene expression in other tissues or from other molecular layers or phenotypes.

We assessed the performance of MR-link under different scenarios and compared it to four other MR methods: Inverse variance weighting (IVW), which assumes absence of LD and pleiotropy, and the pleiotropy-robust methods MR-Egger, LDA-MR-Egger and MR-PRESSO (**Table 1**) (17–19,29). In addition, we compared MR-link to the widely used Bayesian colocalization method *coloc* (30), although this is not a formal test for assessing causal relationships, but rather a way to evaluate if two traits share the same causal variant(s) in a locus (30).

We simulated causal relationships between an exposure and an outcome in a 5Mb region, based on LD structure estimated for 403 European samples from the 1000 Genomes project (31) (Methods). All tested MR methods were assessed in 1,500 simulated data sets for 100 different scenarios that varied with respect to the absence or presence of causality, the absence or presence of pleiotropy and the number of causal eQTL variants. We initially evaluated two approaches to select QTL variants as IVs: GCTA-COJO (v1.26.0) and *p* value clumping (Methods) (27,32). We observed that GCTA-COJO was best suited for IV selection because: (i) the median number of IVs identified by GCTA-COJO better represented the number of simulated causal variants (**Supplementary Data 2**) and (ii) the false positive rates (FPRs) in the MR analysis using the IVW method were lower (median FPR was 0.057 using GCTA-COJO versus 0.115 using

Table 1
MR Methods assessed in this study

MR methods assessed in our simulation with information about the type of data needed to make a causal estimate, ability to correct for pleiotropy or linkage disequilibrium and the minimum number of instrumental variables required. Abbreviations: IVW: inverse variance weighting, Sumstats: summary statistics, ILD: individual-level data, O: outcome, E: Exposure, LDR: linkage disequilibrium reference (LDR) panel. For more details, see Methods. ^aWe refer here to the MR-PRESSO test that reports estimates after identifying and removing outliers, as the test without outliers generalizes to IVW estimates.

MR method	Data required	Pleiotropy correction	LD correction	Minimum no. of IVs required
MR-link	Sumstats: E; ILD: O	Yes	Yes	1
Inverse variance weighted (IVW) (29)	Sumstats: O, E	No	No	1
LDA-MR-Egger (17)	Sumstats: O, E, LDR	Yes	Yes	3
MR-Egger (18)	Sumstats: O, E	Yes	No	3
MR-PRESSO ^a , (19)	Sumstats: O, E	Yes	No	3 ^a

Figure 3
Relative performance of different MR methods

The figure shows performance of MR methods on simulations representing the pleiotropy through linkage disequilibrium (LD) scenario (depicted in **Figure 2B**) when 1, 3, 5 or 10 causal expression quantitative trait locus (eQTL) single nucleotide polymorphisms (SNPs) were simulated (Methods). (A) False positive rates (at $\alpha = 0.05$) in scenarios where no causal relationship is simulated. (B) Power to detect a small causal effect (at $\alpha = 0.05$). (C) Power to detect a large causal effect (at $\alpha = 0.05$). Note that MR-link is the only MR method that can adjust for pleiotropy when only one or two instrumental variables are available. MR methods that had fewer than 100 out of 1,500 estimates in a scenario are not shown (Methods). Extended results depicted here can be found in **Supplementary Data 3**.

clumping) (**Supplementary Figure 1** and **Supplementary Data 2**). We therefore selected IVs for the exposure using the GTCA-COJO approach in subsequent analyses.

When we simulated pleiotropy through LD with no causal effect of the known exposure on the outcome (**Figure 2B, 3A, Supplementary Data 3** and Methods), all existing MR-methods showed inflated FPRs (up to 0.71, 0.15, 0.13 and 0.27 for IVW, MR-Egger, LDA-MR-Egger and MR-PRESSO, respectively), whereas MR-link presented an FPR close to expectation (median: 0.05, maximum: 0.058). In addition, for LDA-MR-Egger, MR-Egger and MR-PRESSO, the FPR was undesirably dependent on the number of causal SNPs simulated (**Figure 3A**).

In the scenarios of pleiotropy through LD and non-null causal effects ($=0.05$, $=0.1$, $=0.2$ and $=0.4$), MR-link has high detection power (up to 0.89) and strongly outperforms all other pleiotropy-robust methods (maximum detected power was 0.28 for MR-Egger, 0.26 for LDA-MR-Egger and 0.65 for MR-PRESSO) (**Figure 3B-C, Supplementary Data 3** and Methods). Among all the methods tested, including MR-link, and for all scenarios, IVW had the greatest detection power but also an inflated FPR (minimum FPR: 0.63), making this MR method unreliable in such pleiotropic scenarios (Methods).

When we simulated increasing levels of pleiotropy through overlap (**Figure 2C** and Methods), a situation we expect to be rare in real-world scenarios based on our observation in the BIOS cohort, we observed that all methods including MR-link have increased FPRs (up to 0.22 for MR-link, 0.77 for IVW, 0.10 for LDA-MR-Egger, 0.13 for MR-Egger and 0.30 for MR-PRESSO) (**Supplementary Data 4**). Nonetheless, MR-link remains a powerful method when a causal effect is simulated: maximum power was 0.79 for MR-link, 0.98 for IVW, 0.29 for MR-Egger, 0.28 for LDA-MR-Egger and 0.65 for MR-PRESSO (**Supplementary Data 4**). Although IVW again had the highest power (0.98) here, the FPR was likewise highly inflated (0.77).

Finally, we compared MR-link to the *coloc* package using the area under the receiver operator characteristic curve (AUC) metric as well as FPRs and power (calculated using *coloc* PP4 > 0.9 as a threshold) (Methods). We used the AUC metric because *coloc* provides posterior probabilities of causal variant sharing and not p values (Methods). As *coloc* assumes

chapter 7

that the exposure and the outcome share only one causal variant, we also included the recently implemented *coloc* variations (*coloc-cond* and *coloc-masked*) in our comparison. These variations are expected to perform better in scenarios with multiple causal variants (33). When comparing MR-link to the *coloc* variations through the AUC metric, we find that MR-link consistently outperforms *coloc* and *coloc-masked* in all scenarios, and *coloc-cond* in pleiotropic scenarios. In non-pleiotropic scenarios, MR-link and *coloc-cond* have approximately the same performance (**Supplementary Figure 2**) (**Supplementary Data 5**). As expected, *coloc-cond* has better discriminative performance compared to the original *coloc* when multiple causal variants are simulated (**Supplementary Figure 2**) (**Supplementary Data 5**).

To illustrate detection rates in standard coloc settings as they may be used in a real-world analysis, we determined power and FPR for all coloc variations at a PP4 threshold of > 0.9 (**Supplementary Figure 3**) (**Supplementary Data 6**). In the non-pleiotropic case, *coloc* and *coloc-cond* have the best detection power (up to 0.79 for *coloc* and 0.76 for *coloc-cond*), combined with near zero FPRs (max: 0 for *coloc* and 0.0006 for *coloc-cond*) while *coloc-masked* has lower power (up to 0.40) with a zero FPR (**Supplementary Figure 3A-C**) (**Supplementary Data 6**). In simulations of pleiotropy through LD, all *coloc* methods have increased FPRs (medians: 0.026 for *coloc*, 0.142 for *coloc-cond* and 0.0037 for *coloc-masked*) with a decrease in power relative to the non-pleiotropic simulations (max: 0.37 for *coloc*, 0.43 for *coloc-cond* and 0.14 for *coloc-masked*) (**Supplementary Figure 3D-F**) (**Supplementary Data 6**). These patterns were even more apparent in cases of pleiotropy through overlap (**Supplementary Figure 3G-I**) (**Supplementary Data 6**). This comparison through FPRs and power indicates again that MR-link has superior discriminative ability over *coloc* variations, especially in the presence of pleiotropy.

MR-link identifies gene expression causal to LDL-C levels

We applied MR-link to four separate summary statistics-based eQTL datasets combined with individual-level genotype data and LDL-C measurements in 12,449 individuals from the Lifelines cohort (34) (**Figure 1**). We assessed the causal effect of gene expression changes in i) whole blood using eQTLs from BIOS ($n = 3,503$) and GTEx ($n = 369$), ii) liver as the main tissue important for cholesterol metabolism (using eQTLs from GTEx, $n = 153$) and iii) cerebellum tissue (using eQTLs from GTEx, $n = 154$) as a tissue not involved in cholesterol metabolism but with similar sample size (and thus power) to liver tissue (24,35).

Transcriptome-wide application of MR-link to these eQTL datasets identified 24 significant genes whose variation in blood (18 using BIOS eQTLs, 2 using GTEx eQTLs) or liver (4 genes) was causally related to LDL-C (**Table 2**, **Table 3**, **Supplementary Table 1** and **Supplementary Table 2**). No significant genes were found in cerebellum (**Supplementary Table 2**).

Gene Name	causal effect	p value	#IVs	Biological function and link to LDL-C
<i>IGLC5</i>	-0.053	2.08E-9	1	Immunoglobulin lambda constant 5 (<i>IGLC5</i>) is a pseudogene; three other genes in the same locus and belonging to the same family appear in this table (<i>IGLV-70</i> , <i>IGLV4-69</i> and <i>IGLC6</i>). <i>IGLC5</i> does not have a known function in LDL-C metabolism
<i>KB-1460A1.5</i>	0.154	1.35E-8	1	<i>KB-1460A1.5</i> is an RNA gene with unknown function in LDL-C metabolism
<i>ABO</i>	-0.082	4.84E-8	4	Alpha 1-3-N-Acetylgalactosaminyltransferase And Alpha 1-3-Galactosyltransferase (<i>ABO</i>) is a protein coding blood group gene. <i>ABO</i> is located in LDL-C GWAS locus (36,37,75). <i>ABO</i> is part of the KEGG pathway Glycosphingolipid biosynthesis although the direct link between <i>ABO</i> and LDL-C remains unclear.
<i>UNC5B</i>	-0.012	4.87E-8	1	Unc-5 Netrin Receptor B (<i>UNC5B</i>) is a protein-coding gene and a receptor for the NETRIN1 protein. It is associated to the disease Hyperinsulinemic Hypoglycemia familial 3. <i>UNC5B</i> is localized in lipid rafts, membrane compartments that contain high levels of cholesterol and lipids (76). The direct link of this gene to LDL-C remains unclear.
<i>TMEM176B</i>	-0.029	1.11E-7	4	Transmembrane protein 176B (<i>TMEM176B</i>) is a protein coding gene located in the <i>TMEM176B-TMEM176A-AOC1</i> GWAS locus for HDL-C (36). Two IVs for <i>TMEM176B</i> are overlapping with IVs for <i>TMEM176A</i> .
<i>REEP1</i>	-0.022	1.25E-7	1	Receptor Accessory Protein 1 (<i>REEP1</i>) is a protein-coding gene. Mutations of the N-terminal of <i>REEP1</i> lead to accumulation in lipid droplets in the endoplasmic reticulum (77). <i>REEP1</i> does not have a known function in LDL-C metabolism.
<i>KRT79</i>	-0.059	1.54E-7	2	Keratin 79 (<i>KRT79</i>) is a protein coding gene, that promotes sebaceous gland maintenance in mice hair follicles (78). The sebaceous gland produces up to 90% of the lipids present in the epidermis. Although a direct link to LDL-C levels remains unclear.
<i>IGLC6</i>	-0.077	2.21E-7	3	Immunoglobulin lambda constant 6 (<i>IGLC6</i>) is a pseudogene; three other genes in the same locus belonging to the same family appear in this table (<i>IGLV-70</i> , <i>IGLV4-69</i> and <i>IGLC5</i>). <i>IGLC6</i> does not have a known function in LDL-C metabolism.
<i>MAP1LC3A</i>	0.037	2.32E-7	1	Microtubule Associated Protein 1 Light Chain 3 Alpha (<i>MAP1LC3A</i>) is a protein coding gene. <i>MAP1LC3A</i> has no known function in cholesterol metabolism.
<i>AOC1</i>	-0.009	2.48E-7	1	Amine Oxidase, Copper Containing 1 (<i>AOC1</i>) is a protein-coding gene located in the <i>TMEM176A-TMEM176B-AOC1</i> GWAS locus for HDL-C (36).
<i>IGLV4-69</i>	0.088	3.10E-7	1	Immunoglobulin Lambda Variable 4-69 (<i>IGLV4-69</i>) is a protein-coding gene; three other genes in the same locus and belonging to the same family appear in this table (<i>IGLV-70</i> , <i>IGLC5</i> and <i>IGLC6</i>). <i>IGLV4-69</i> does not have a known function in cholesterol metabolism
<i>SYCP2L</i>	-0.019	4.09E-7	2	Synaptonemal Complex Protein 2 Like (<i>SYCP2L</i>) is a protein-coding gene, located in a GWAS locus for antihypertensive syndrome and fatty acid measurements. <i>SYCP2L</i> does not have a known function in LDL-C metabolism (39,79).
<i>C10orf10/DEPP1</i>	-0.069	4.74E-7	1	<i>DEPP1</i> (also known as <i>C10orf10</i>) is an autophagy regulator highly expressed in adipose tissue. <i>DEPP1</i> overexpression in mice reduces glucose and triglyceride levels (80), although a direct link of this gene to LDL-C metabolism remains unclear.
<i>TMEM176A</i>	-0.022	4.77E-7	3	Transmembrane protein 176A (<i>TMEM176A</i>) is a protein-coding gene located in the <i>TMEM176B-TMEM176A-AOC1</i> GWAS locus for HDL-C (36). Two IVs for <i>TMEM176A</i> are overlapping with <i>TMEM176B</i> .
<i>RP11-18H21.1</i>	0.030	5.02E-7	2	<i>RP11-18H21.1</i> is a non-coding RNA gene without a known function in LDL-C metabolism.
<i>TACSTD2</i>	-0.019	8.65E-7	2	Tumor Associated Calcium Signal Transducer 2 (<i>TACSTD2</i>) is a protein-coding gene located on the cell membrane involved in the superpathway Ca, cAMP and Lipid Signaling. The function of <i>TACSTD2</i> in LDL-C metabolism is unclear.
<i>MSLN</i>	-0.026	1.39E-6	4	Mesothelin (<i>MSLN</i>) is a protein coding gene, its link to LDL-C is unclear.
<i>IGLVI-70</i>	0.112	3.49E-6	2	Immunoglobulin Lambda Variable (I)-70 (<i>IGLVI-70</i>) is a pseudo gene; three other genes in the same locus and belonging to the same family appear in this table (<i>IGLV4-69</i> , <i>IGLC5</i> and <i>IGLC6</i>). <i>IGLVI-70</i> does not have a known function in cholesterol metabolism

Table 2
MR-link results using BIOS blood eQTLs

This table shows 18 Bonferroni significant genes identified by MR-link as causal for LDL-C levels in the analysis that included eQTLs from the BIOS cohort. Gene names are according to ENSEMBL GENES 96 database (human Genome build 37). The causal effect estimate represents the changes in LDL-C (mg per dL) per standard deviation increase in gene expression. *p* values listed in this table are adjusted and are based on a calibrated 2 sided T test (Supplementary Note 1). Full summary statistics of the genes are shown in Supplementary table 1. Abbreviations: LDL-C: low-density lipoprotein cholesterol, IV: instrumental variable, HDL-C: high-density lipoprotein cholesterol, GWAS: Genome wide association study, siRNA: small interfering RNA.

Gene Name	causal effect	p value	#IVs	Biological function and link to LDL-C
<i>PVRL2</i>	0.318	3.00E-14	1	Poliovirus Receptor-Related 2 (<i>PVRL2</i>) is a protein-coding gene also known as <i>NECTIN2</i> . It is a cell-membrane protein located in the LDL-C GWAS locus <i>APOE</i> . siRNA experiments show that LDL-C uptake is increased in cells upon its downregulation (49). <i>PVRL2</i> knockout mice also had less atherosclerosis (47). Both studies indicate a reduction in LDL-C upon downregulation of <i>PVRL2</i> .
<i>PSRC1</i>	-0.085	3.99E-9	1	Proline And Serine Rich Coiled-Coil 1 (<i>PSRC1</i>) is a protein-coding gene located in an LDL-C GWAS locus (36). <i>PSRC1</i> has not been found to have an effect on cholesterol despite being targeted in a specific functional study (44). The IV for <i>PSRC1</i> is overlapping with the IV for <i>CELSR2</i> and <i>SORT1</i> .
<i>SORT1</i>	-0.087	5.89E-9	v	Sortilin (<i>SORT1</i>) is a protein-coding gene located in an LDL-C GWAS locus(36). siRNA and knockdown experiments have functionally validated that <i>SORT1</i> has a negative effect on LDL-C levels(42,43). The IV for <i>SORT1</i> is overlapping with the IV for <i>PSRC1</i> and <i>CELSR2</i> .
<i>CELSR2</i>	-0.099	6.80E-8	1	Cadherin EGF LAG Seven-Pass G-Type Receptor 2 (<i>CELSR2</i>) is a protein-coding gene located in an LDL-C GWAS locus(36). <i>CELSR2</i> has not been found to have an effect on cholesterol despite being targeted by a specific functional study(44). The IV for <i>CELSR2</i> is overlapping with the IV for <i>PSRC1</i> and <i>SORT1</i> .

MR analysis that used whole-blood eQTLs from GTEx was, as expected, underpowered compared to the analysis using BIOS eQTLs. Only two genes were found to be significant here, but they were not-significant in the BIOS cohort, where a more robust estimate could be made thanks to higher number of IVs identified (**Supplementary Figure 4A**). Despite the limited power, we observed high concordance between effect sizes from the two analyses (using blood eQTLs from BIOS and GTEx) for all genes that showed nominal significance ($p < 0.05$) in the BIOS analysis, with 94.8% of genes showing the same effect direction (**Supplementary Figure 4B**).

Several genes located in genome-wide association study (GWAS) loci for cholesterol metabolism were found in the MR analysis that used blood eQTLs from BIOS. These include *ABO*, located in a LDL-C locus, *AOC1*, *TMEM176A* and *TMEM176B*, which are all located in the same HDL-C-associated locus (36,37), and *SYCP2L*, which is located in a GWAS locus for polyunsaturated fatty acids and related to LDL-C levels (38,39). For the other genes identified, there was no evidence in the literature for a direct role in cholesterol metabolism, although some interesting patterns were evident. For example, we observed multiple genes involved in immunoglobulin production (*IGLC5*, *IGLC6*, *IGLV4-69* and *IGLV6-70*) and insulin metabolism (*UNC5B*, *DEPP1*), mechanisms that are consistent with the role of cholesterol in inflammation and insulin resistance (40,41). For all 18 genes, the effect direction estimated by MR-link was concordant with the direction estimated by other MR-methods when they were available, except in the case of *MSLN*, where only LDA-MR-Egger gave discordant results compared to all other methods (**Table 1**, **Supplementary Figure 5** and **Supplementary Table 3**). Interestingly, 17 of the 18 genes did not pass significance after multiple testing correction using the other tested methods: only *ABO* passed Bonferroni significance and only when using the IVW method (**Table 1**, **Supplementary Figure 5** and **Supplementary Table 3**). In 13 genes, a causal effect could not be estimated by: MR-Egger, LDA-MR-Egger and MR-PRESSO because there were too few IVs. Furthermore, MR-PRESSO did not make a causal estimate in the remaining

Table 3.
MR-link results using GTEx liver eQTLs

This table lists four Bonferroni-significant genes that were identified using GTEx liver eQTLs. Gene names are according to ENSEMBL GENES 96 database (human Genome build 37). The causal effect estimate represents changes in LDL-C (mg per dL) per standard deviation increase in gene expression. p values listed in this table are unadjusted and are based on a calibrated T statistic (Supplementary Note 1). Full summary statistics of these genes are shown in **Supplementary Table 2**. Abbreviations: LDL-C: low-density lipoprotein cholesterol, GWAS: Genome wide association study, siRNA: small interfering RNA, IV: instrumental variable.

5 genes as it identified too many outliers (**Table 1**, **Supplementary Figure 5** and **Supplementary Table 3**).

In the MR analysis using eQTLs from liver, all the genes identified fall within LDL-C GWAS loci. Among these, we found a negative causal effect for the well-known *SORT1* gene ($p = 5.9 \times 10^{-9}$). Multiple functional studies have shown that this gene encodes the protein Sortilin (encoded by *SORT1*) and that it affects plasma LDL-C levels by acting on clearance of LDL-C and on secretion of very-LDL (VLDL) by the liver (42–44) (**Table 3** and **Supplementary Table 2**). We also found two other genes in the same GWAS locus, *PSRC1* and *CELSR2*, but the IV (only one was found) for these genes was identical to that of *SORT1* due to the high correlation between expression levels of these genes. Full overlap of a single IV in this locus makes it impossible to discern causal from pleiotropic genes using MR-methods, including MR-link. The fourth gene found to be significant using liver eQTLs is *PVRL2* ($p = 3 \times 10^{-14}$), which is located in the *APOE* locus associated to LDL-C (**Table 3**) (36,37). For *PVRL2*, we estimated a positive causal effect; higher expression of *PVRL2* is causally related to higher LDL-C (**Table 3**). *PVRL2* is 17.5kb downstream of the *APOE* gene, and two common missense polymorphisms in *APOE* account for a large fraction of the association signal (37,45). Interestingly, in the most recent GWAS meta-analysis for lipids, 19 jointly significant LDL-C variants were found spanning a 162kb region that encompasses *PVRL2* (37). This indicates that, while missense mutations in *APOE* play a major role, other genes in this locus are also likely involved in LDL-C regulation and that pleiotropic effects are to be expected. Our analyses indicate that *PVRL2* is one of the causal genes at this locus. The positive effect of *PVRL2* on LDL-C was also seen in the analysis that used blood eQTLs from BIOS ($p = 4.3 \times 10^{-5}$), although it did not pass our significance threshold ($p = 4.6 \times 10^{-6}$). Likewise, variation in gene expression of *PVRL2* in blood has been found to be associated with LDL-C in a transcriptome-wide association analysis carried out in a very large genetic association study (37). Of note, since the LD between IVs used in the analysis of blood and liver eQTLs was low ($r^2 < 0.2$), the results potentially indicate a dual causal role for *PVRL2* across these two tissues.

PVRL2 has mostly been studied in the context of atherosclerosis, where it has been shown to act as cholesterol-responsive gene involved in trans-endothelial migration of leukocytes in vascular endothelial cells, a key feature in atherosclerosis development (46–48). Our results indicate a role for *PVRL2* in modulating plasma levels of LDL-C via its expression variation in the liver. Biologically the role in liver could be explained by increased production of very-LDL or decreased LDL-C uptake (**Figure 4**). In line with this hypothesis, a siRNA screen in hepatic cell lines of genes in the *APOE* locus showed that down-regulation of *PVRL2* gene expression promotes LDL-C uptake (49) (**Figure 4**). Overall, our results and existing functional evidence support that *PVRL2* expression is correlated with LDL-C levels and show a causal effect in liver (**Figure 4**).

MR-link confirms ApoE changes affect LDL-C levels

Figure 4
Biological interpretation of PVRL2

Functional and statistical evidence for the causal effect of PVRL2 on low density lipoprotein cholesterol (LDL-C) levels. The teal arrow indicates a positive causal relationship between PVRL2 expression in liver and LDL-C levels in plasma – this relationship was detected in our MR analysis. The red arrow indicates a negative causal relationship between PVRL2 expression and LDL-C uptake in hepatic cells – this relationship was detected in small interfering RNA (si) experiments described in Blattman et al (49).

To assess the effectiveness of MR-link in proteomics measurements, we combined the aforementioned LDL-C measurements in the Lifelines cohort with *cis*-pQTL summary statistics of 471 plasma protein measurements (measured using the SOMAscan platform in a cohort of 3301 individuals) (Methods) (25,50). One protein passes the Bonferroni multiple testing threshold ($p < 1.05 \times 10^{-4}$): ApoE3, an isoform of ApoE (causal effect: 0.40 (+/- 0.13), $p = 4.65 \times 10^{-5}$ (calibrated T test), SOMAmer ID: APOE.2937.10.2). pQTLs were also available for ApoE2 (SOMAmer ID: APOE.5312.49.3), another isoform of ApoE but the causal effect was weaker and did not pass the Bonferroni threshold (causal effect= 0.56 (+/- 0.24), $p = 0.002$ (calibrated T test)) (45). These results are in line with the well-known causal relationship between increased ApoE plasma levels and LDL-C, and the widely described stronger impact of the E3 isoform compared to the E2 isoform (45). Interestingly, MR-link did not estimate BGAT, the protein product of *ABO*, to be significant in this dataset (SOMAmer ID: ABO.9253.52.3, $p = 0.18$ (calibrated T test)) We

compared the IVs identified for BGAT (rs9411463 and rs72775494) with those used in the *ABO* blood eQTL analysis and found that only one IV for the BGAT protein was in LD (rs9411463) with any of the four IVs for *ABO* expression in BIOS. This scenario is in line with the overall patterns observed in the original study - only a small fraction of eQTLs in blood also affect protein levels, but our results could also reflect targeting of the SOMAmer to a specific ABO protein isoform (25). Unfortunately, further isoform information for BGAT was not available in the original study.

Discussion

Identification of genes whose changes in expression are causally linked to a phenotype is crucial for understanding the mechanisms behind complex traits. While several methods exist that infer causal relationships between two phenotypes, these rely on a set of assumptions that are often violated when gene expression is the exposure. Specifically, the presence of LD and pleiotropy between the genetic variants chosen as IVs are the main cause of violations of such assumptions (17–19,29). Here we interrogated a large gene-expression dataset and showed that the eQTLs of a gene, which can be used as IVs, are very likely to be in LD, but not overlapping, with eQTLs of other genes, indicating that potential sources of pleiotropy in transcriptome-wide MR analyses are likely to come from variants in LD with the IVs.

We therefore developed MR-link, a causal inference method that is robust to unobserved pleiotropy. Our *in-silico* results show that MR-link has the best discriminative ability compared to all the other MR methods we tested, as well as to the Bayesian colocalization method coloc. MR-link jointly models the outcome using jointly significant eQTLs as IVs, combined with variants in LD, to correct for all potential sources of pleiotropy. To our knowledge, this is the first time that this approach has been used in a causal inference method.

We applied MR-link to real data by applying it to LDL-C cholesterol measurements and eQTLs derived from blood, cerebellum and liver. This identified known and previously unknown genes within and outside GWAS loci. For example, in liver we identified the well-known negative causal relationship between expression of *SORT1* in liver and LDL-C (42–44). In liver, and suggestively in blood, we detected a causal effect for *PVRL2*, a gene located in the *APOE* locus. While a role for this gene is mostly known for immune and endothelial cells and in the context of atherosclerosis (46,48), our results indicate that regulation of expression of this gene in both blood and liver causally affects LDL-C levels. Given its established role in atherogenesis, *PVRL2* has been proposed as a potential therapeutic target for atherosclerosis. Our study indicates that such strategies should not only take into account the effect on atherosclerotic plaques, but also consider the hepatic function of *PVRL2* in regulating plasma LDL-C levels in humans.

All the genes identified in the analyses that used eQTLs from blood were different from those identified using eQTLs from liver. While this is partly

chapter 7

due to statistical power, as the BIOS cohort is more than 20 times larger than the GTEx cohort used to derive eQTLs in liver, this may also be related to tissue-specific mechanisms. We expect that causal genes found in whole blood will affect LDL-C through pathways that signal for lipid changes or regulate lipid binding to erythrocytes, as hypothesized for the *ABO* gene, whereas genes found in liver are more likely to be involved in lipid metabolism (51,52).

MR-link has several advantages over other recent MR methods developed to overcome bias from LD and pleiotropy (17,23). First, MR-link can model unobserved pleiotropy, whereas sources of pleiotropy need to be specified in multivariate MR methods. This is particularly important because sources of pleiotropy may be context-dependent and may arise from a phenotype other than those being measured in a cohort (14,35). Second, MR-link can derive robust causal estimates even when only one or two IVs are available. The majority of genes tested in our large eQTL dataset have fewer than three IVs (68%), which makes it impossible for MR-PRESSO, MR-Egger and LDA-MR-Egger to make causal estimates (17–19).

One of the MR-link assumptions is that the IVs affect the outcome only through the exposure, conditional on the unmeasured pleiotropic effect. This assumption is violated when the IVs of the exposure and of the pleiotropic effect are fully overlapping. This assumption must not be violated when a single IV is available, but can be relaxed when multiple IVs are used in the model, as the relative effects of the IVs help to discriminate between a true causal effect and a pleiotropic effect, similar to multivariable Mendelian randomization methods (22). In the case of multiple IVs that are fully overlapping, we have shown that MR-link has an increased FPR, yet still maintains higher power compared to other MR-methods and superior discriminative ability compared to coloc.

The application of MR-link is not restricted to gene expression or proteomics datasets; it can also be applied to other molecular layers that are known to have a similar genetic architecture to gene expression, such as metabolites. Given the increases in sharing of summary statistics from functional genomics QTL studies, coupled with the development of very large biobanks such as the UK biobank, the Estonian Biobank, the Lifelines cohort study and the Million Veteran Program cohort (34,53–55), we foresee many opportunities for applications of MR-link to individual-level data for the identification of the molecular mechanisms underlying complex traits. Of note, while we have limited our simulations to quantitative traits as an outcome in this paper, MR-link could be applied to binary traits such as human diseases. However, we have not investigated its performance in detail for binary outcome phenotypes. Furthermore, as for all MR studies, our method can be applied to populations of any ethnicity, provided that the summary statistics of the exposure are derived from a population that is ethnically-matched with the outcome cohort.

We foresee that many causal relationships will be discovered if highly powered causal inference methods such as MR-link are applied to many human traits. This could make it possible to build extensive causal networks similar in size and complexity to metabolic networks of small molecules, which would provide valuable insights into the mechanisms

behind human traits and diseases.

Methods

BIOS consortium cohort genotype and expression analysis

We used genotype and expression measurements on 3,746 Dutch individuals from the Biobank-based Integrative Omics Study (BIOS; <http://www.bbmri.nl/acquisition-use-analyze/bios/>), a collection of six different data cohorts: Lifelines DEEP (56), Prospective ALS Study Netherlands (57), Leiden Longevity Study (58), Netherlands Twin Registry (59), The Cohort on Diabetes and Atherosclerosis Maastricht (60) and the Rotterdam Study (61). Genotyping was performed separately per cohort (see references). All combined genotypes were imputed to the Haplotype reference consortium dataset (62) using the Michigan imputation server (63). We retained only biallelic SNPs and confined our analyses to variants with minor allele frequency (MAF) > 0.01, Hardy-Weinberg equilibrium (HWE) p value > and an imputation quality RSQR > 0.8. A genetic relationship matrix (GRM) was derived based on LD-pruned genotypes using the Plink 1.9 command `--indep 50 5 2`, and one individual was kept from all pairs of individuals that had a GRM value > 0.1 using the `--rel-cutoff` Plink 1.9 command (32). Population outliers were identified using a principal component analysis of the GRM, and individuals more distant than three standard deviations from the mean of principal component 1 and principal component 2 were removed.

RNA-seq gene expression quality control and processing is the same as those of Zhernakova et al. (13). RNA extracted from whole blood was paired end sequenced using the Illumina HiSeq 2000 instrument. RNA-seq read alignment was performed using STAR (version 2.3.0e) (64). During alignment, variants with MAF < 0.01 from the Genome of the Netherlands were masked (65). Gene expression was quantified using HTSeq (v.0.6.1p1) (66). Samples with less than 80% of reads mapping to exons were considered of low quality and removed. Samples were also removed if they had less than 85% of mapped reads, or if they had a median 3' bias larger than 70% or smaller than 45%. To further account for unobserved confounders, the expression matrix was corrected for the first 25 principal components as well as 5' bias, 3' bias, GC content, intron base-pair percentage and sex following the procedure of Zhernakova et al. (13). After genotype and expression quality control filters, 3,503 individuals with expression data of 19,960 transcripts and genotype information of 7,838,327 SNPs were available for analyses. In this set, 57% were female and the average age was 52.8 years (± 16.0 Stand. Dev.). eQTL association analysis was performed for SNPs located ± 1.5 Mb of the transcript using plink 1.9 and the `--assoc` command (32). For 13,778 genes, at least one eQTL at $p < 5 \times 10^{-8}$ was identified, and those genes were used for all the analyses described in this manuscript.

We quantified how many genetic variants are necessary to explain gene expression using a conditional joint analysis approach. We identified jointly significant eQTLs by applying GCTA-COJO (v1.26.0) (27) to eQTL summary statistics, using the BIOS cohort as LD reference panel,

chapter 7

and selecting jointly significant variants that showed a $p < 5 \times 10^{-8}$ in this analysis step. To infer how often eQTLs are shared between genes, we assessed the percentage of genes with top eQTLs (or jointly significant variants) that have LD $r^2 > 0.99$. We used the $r^2 > 0.5$ threshold to see how often eQTL variants were in LD with each other.

We performed statistical finemapping of all genes using the FINEMAP v1.3.1 program (28). First, we searched for associated eQTL variants ($p < 5 \times 10^{-8}$) in the *cis* associated region. We then padded the associated regions with 100kb and only looked for variants in this extended region. FINEMAP requires the same number of individuals across all variants, therefore we analyzed only the genes with the associated variants available in all subcohorts. We ran FINEMAP on these genes with the `--sss` option, using LD computed with PLINK v1.9, with the `--r` command. Furthermore, genes were not run if they had less than 25 variants available in the region, or if a combination of variants led to an invalid posterior probability, leaving 13,276 genes which were successfully finemapped. FINEMAP provides several configurations of statistically fine-mapped variants, along with their posterior probability of being causal. Studies that identify causal variants usually use a high posterior inclusion probability of multiple causal variant configurations to make sure the causal variant is captured in analysis. In MR studies it is not necessary to identify true causal variants, as the IV only needs to explain the exposure signal the best. In our analysis of LD between FINEMAP variants we have therefore only considered the most likely configuration identified by FINEMAP, as these variants better explain the exposure variation.

Lifelines cohort genotype data and LDL-C levels

Lifelines is a multi-generational cohort study of 167,000 individuals from the north of the Netherlands. It was approved by the medical ethics committee of the University Medical Center Groningen and conducted in accordance with Helsinki Declaration Guidelines. All participants signed an informed consent form prior to enrollment. A subset of 13,436 Lifelines samples were genotyped with the cytoSNP array and underwent the quality control steps described in Scholtens et al. (34): Genotyped variants were retained based on three criteria: Minor allele frequency > 0.001 , HWE $p > 10^{-4}$ and a missingness > 0.95 . After genotype quality control, samples were imputed using the Genome of the Netherlands reference panel (65) and Minimac version 2012.10.3 (67). Variants were further excluded if they were of bad imputation quality (RSQR < 0.3), showed deviation from HWE ($p < 10^{-6}$), or if they were absent in the set of quality controlled genotyped and imputed variants of the BIOS cohort.

Low-density lipoprotein cholesterol (LDL-C) was estimated using the Friedewald equation (68), based on triglycerides, high density lipoprotein and total cholesterol levels (34). Total cholesterol levels of individuals who were prescribed cholesterol-lowering medication were divided by 0.8 prior to calculating LDL-C. Individuals with > 4.52 mmol per Liter total triglycerides were removed (68). Additionally, LDL-C levels were corrected for age, age squared and sex. After genotype and LDL-C quality control, 12,449 individuals (of which 58.8% were female and the average age was 48.7 years (± 11.5 Stand. Dev.)) and 7,336,374 variants

remained for analyses. Association analysis for additive effects on LDL-C was performed using linear regression on standardized genotypes, e.g. transforming genotypes into a distribution with mean 0 and variance 1. Summary statistics of this analysis were used to perform MR analyses using the existing MR methods listed in **Table 1**.

GTEEx download and analysis

We downloaded GTEEx version 7 eQTL summary statistics, including non-significant results, from the GTEEx website (<https://gtexportal.org/home/datasets/>) (24). For every gene with at least one eQTL at $p < 5 \times 10^{-8}$, conditional analysis using GCTA-COJO was performed to select secondary variants at the same threshold, using the BIOS cohort as an LD reference. This resulted in 4,028, 1,557 and 1,726 genes with at least one jointly significant eQTL for whole blood, liver and brain (cerebellum) tissues, respectively.

pQTL summary statistics download and analysis

We downloaded the proteomics summary statistics of Sun et al (25) from the GWAS catalog (ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/SunBB_29875488_GCST005806). We isolated *cis* regions by selecting variants +/- 1.5Mb away from each transcript. These variants already passed the quality control steps of Sun et al.: i) INFO score ≥ 0.7 ; ii) minor allele count ≥ 8 ; iii) Hardy Weinberg equilibrium $p \geq 5 \times 10^{-6}$ (25). For all these variants we extracted UK10K minor allele frequencies as this is required for GCTA-COJO IV selection. We selected IVs using Lifelines genotypes as an LD reference (37). To run MR-link, we first selected proteins with significant ($p < 5 \times 10^{-8}$) variants that were shared between the *cis* summary statistics and the Lifelines cohort. This resulted in 471 proteins with significantly associated variants ($p < 5 \times 10^{-8}$) that are overlapping with the variants in the Lifelines cohort and for which GCTA-COJO was able to identify IVs.

Simulation of genotypes

403 non-Finnish European individuals were isolated from the 1000 Genomes phase 3 release and used as a starting point for genotype simulation (31). We simulated genotype data for 25,000 individuals in a chromosomal region (Chromosome 2, 100Mb to 105Mb, human genome build 37) using the HAPGEN2 program (v.2.2.0), combined with interpolated HAPMAP3 recombination rates (69). The region was then reduced to 1Mb in length: between 102Mbp and 103Mb. Only biallelic SNPs with MAF < 0.01 were retained from simulated genotypes, leaving 3,101 variants in this region. Simulated individuals were separated into an outcome cohort of 15,000 individuals, and into an exposure cohort and an LD reference cohort of 5,000 individuals each. These cohort sizes were chosen to roughly represent the sizes of BIOS and Lifelines cohorts. Simulation of phenotypes

Simulation of phenotypes

We simulated quantitative phenotypes representing the exposures by randomly selecting SNPs from the simulated genetic region, and subsequently assigning these an effect. Causal SNPs were selected to represent both pleiotropy through LD (**Figure 2B**) and pleiotropy through overlap (**Figure 2C**). For the scenario of pleiotropy through LD (**Figure 2B**), one to ten causal SNPs (subset $\mathbf{s}_E$) for the exposure were randomly selected from the entire simulated genetic region, and the same number of causal SNPs (subset $\mathbf{s}_U$) for the unobserved (pleiotropic) exposure was randomly selected from all SNPs in moderate LD ($0.25 < r^2 < 0.95$) with SNPs in $\mathbf{s}_E$. When pleiotropy through overlap was simulated (**Figure 2C**), the causal SNPs for the observed and unobserved exposure were selected to be identical: $\mathbf{s}_E = \mathbf{s}_U$. A combination of pleiotropy through overlap and pleiotropy through linkage was simulated by choosing some or all of the SNPs of the unobserved exposure (subset $\mathbf{s}_U$) to be overlapping and some being in LD ($0.25 < r^2 < 0.95$) with SNPs in $\mathbf{s}_E$.

The mathematical framework for the simulation of phenotypes is as follows. For each selected causal SNP of the exposure (subset $\mathbf{s}_E$), we simulated an effect-size from the uniform distribution $U(-0.5, 0.5)$, and then simulated the observed exposure $\mathbf{y}_E$ as:

$$\mathbf{y}_E = \mathbf{X}\beta_E + \mathbf{C} + \epsilon_E, (1)$$

where $\mathbf{X}$ is a genotype matrix of size $n \times m$, with n being the number of individuals (5,000) and the number of variants in the region (3,101 in the simulated data), β_E is the vector of effects

$$\beta_{E,j} = \begin{cases} \sim U(-0.5, 0.5) & \text{if } j \in \mathbf{s}_E \\ 0 & \text{otherwise} \end{cases}, \forall j \in \{1, \dots, m\},$$

and $\mathbf{C} \sim N(0, 0.5)^n$ is an n -vector of independent scalar draws from $N(0, 0.5)$, representing a cohort specific confounder value per individual. Finally, ϵ_E is an n -vector of the measurement error of the exposure. Similarly, the unobserved exposure $\mathbf{y}_U$ was simulated as:

$$\mathbf{y}_U = \mathbf{X}\beta_U + \mathbf{C} + \epsilon_U, (2)$$

where β_U is the vector of effects defined as:

$$\beta_{U,j} = \begin{cases} \sim U(-0.5, 0.5) & \text{if } j \in \mathbf{s}_U \\ 0 & \text{otherwise} \end{cases}, \forall j \in \{1, \dots, m\},$$

$\mathbf{s}_U$ is the selection of SNPs for the unobserved exposure and ϵ_U are measurement errors distributed as ϵ_E . The outcome phenotype $\mathbf{y}_O$ was then simulated as a linear combination of the observed and unobserved exposures:

$$\mathbf{y}_O = \mathbf{y}_E b_E + \mathbf{y}_U b_U + \mathbf{C} + \epsilon_O, (3)$$

where the causal effect of interest is parameterized per simulation run as $b_E = \{0, 0.05, 0.1, 0.2, 0.4\}$ and the (unknown) pleiotropic effect is the parameter $b_U = \{0, 0.4\}$ reflecting absence and presence of a pleiotropic

effect in a locus. Again, the measurement error ϵ_o is drawn from $N(0, 0.5)$.

The genetic variants of the exposures ($\mathbf{s}_E, \mathbf{s}_U$) and their effect sizes β_E, β_U were drawn and used in both cohorts (exposure and outcome), while the other random variables $\mathbf{C}, \epsilon_E, \epsilon_U, \epsilon_o$ were randomly drawn in a cohort-specific manner. Since our model was built to account for unobserved pleiotropy, the observed and unobserved exposure were used to generate the outcome phenotype as in equation (3), but only the outcome phenotypes and the summary statistics of the (observed) exposure phenotype were used in the causal inference analysis.

Simulation parameters and scenarios

We simulated 1,500 runs per scenario, each with a unique outcome (O) and two exposures (E and U). The scenarios differed in the number of causal SNPs (which varied from one to ten for both the observed and unobserved exposure), the strength of the causal relationship of interest (varied from no causal effect up to a large effect ($b_E = \{0, 0.05, 0.1, 0.2, 0.4\}$) and the presence ($b_U = 0.4$) or absence ($b_U = 0.0$) of the pleiotropic effect. This resulted in $10 \times 5 \times 2 = 100$ different scenarios.

In certain cases, an estimate cannot be made by an MR method, for instance when insufficient IVs are identified or a solution is not found in the estimation method. As a result, there are sometimes fewer estimates than expected in the final results. To ensure the stability of our FPR and power estimates, we have only reported results for a MR method in a specific scenario if we had more than 100 estimates out of the 1,500 simulated runs.

Instrumental variable selection

IV selection can be difficult when there is LD between association signals. In simulations, we used two IV selection techniques: GCTA COJO (27) and p value clumping, using standard settings of plink 1.9 except for the r^2 threshold, which was set to 0.1 (32). Both selection methods used a p value threshold of $p < 5 \times 10^{-8}$. When selecting IVs for BIOS and GTEx, we only used the GCTA-COJO technique.

MR-link

MR-link is a method for causal inference that is robust to the presence of LD and unobserved pleiotropy. It is an MR approach that requires individual-level data from the outcome cohort and summary statistics (effect sizes, standard errors and MAFs) from an exposure. Conceptually, MR-link jointly models a known exposure with SNPs that are in LD with the exposure IVs (tag-SNPs). Tag-SNPs are used to account for the unobserved pleiotropic effect present in a locus.

We defined our model in the following manner. Let $\mathbf{X}$ be a genotype matrix of $n \times m$ where n is the number of individuals in the outcome study and m are all the SNPs in a *cis*-region around the transcript (± 1.5 Mb of the transcript), in which SNPs at indices $\mathbf{s}_E$ are the causal genetic

chapter 7

variants (IVs) for the exposure E . If we define the exposure E and the unobserved (pleiotropic) exposure U as in equation (1) and (2), then the outcome phenotype from equation (3) can be represented as a function of E and U with the following equation:

$$\mathbf{y}_O = \mathbf{X}\beta_E b_E + \mathbf{X}\beta_U b_U + \mathbf{C}_O + \epsilon_O, (4)$$

where b_E is the causal effect of interest of the exposure on the outcome, b_U is the causal effect of the unobserved exposure, $\mathbf{C}_O$ is a -vector of independent scalars representing specific confounder per individual and ϵ_O is the measurement error of the outcome. In the hypothetical case that the genetic effects for both the exposure E and the pleiotropic exposure U are known, we can estimate b_E by solving equation (4) in an analysis that is similar to multivariate MR (22). In a real-world scenario, only the IV(s) for the exposure are known, while the variants that contribute to the unobserved (pleiotropic) exposure and their effect on the outcome are unknown.

Under equation (4), MR-link relies on the assumption that SNPs on $\mathbf{s}_E$ influence the outcome $\mathbf{y}_O$ only through their effect on $\mathbf{y}_{E'}$ when conditioning on $\mathbf{s}_U$.

MR-link uses the following procedure to estimate causal effects:

1. A selection $\mathbf{s}_E$ of IVs for the exposure and conditional effect sizes β_E for these IVs are determined using the GCTA-COJO method (27). A vector of effect sizes for all SNPs in the region is thus defined as:

$$\underline{\beta}_E = \begin{cases} \neq 0 & \text{if } j \in \mathbf{s}_E \\ 0 & \text{otherwise} \end{cases}, \forall j \in \{1, \dots, m\}$$

2. All SNPs in LD $0.1 < r^2 < 0.99$ with the exposure IVs are potential tag-SNPs. These variants are iteratively pruned for high LD so that tag-SNPs, $\mathbf{s}_T$ are always $r^2 < 0.95$ with each other in order to reduce collinearity and computation time.
3. The following equation is solved for using ridge regression:

$$\mathbf{y}_O = \begin{pmatrix} \vdots \\ \frac{\mathbf{x}\beta_E}{m_E} \\ \vdots \end{pmatrix} \begin{pmatrix} \vdots \\ \frac{\mathbf{x}_T}{\sqrt{m_T}} \\ \vdots \end{pmatrix} \begin{pmatrix} b_E \\ \vdots \\ \beta_U b_U \\ \vdots \end{pmatrix} + \epsilon \quad (5)$$

4. where $\mathbf{X}_T$ is the genotype matrix of the outcome containing only tagging variants as defined in step (2), m_T is the number of tagging variants and is used to normalize for the number of tags in the region, and m_E represents the number of IVs selected by the selection method and is a parameter used to remove the dependency of the model on the number of IVs. The resulting coefficient vector contains the causal effect of interest $b_{E'}$ and the vector $\beta_U b_U$ of length m_T is a nuisance parameter that captures pleiotropic effects. Because individual-level data of the outcome is modeled by MR-link, MR-link does not use any summary statistics of the outcome.

We also considered solving the equation (5) using ordinary

least squares. However, due to the multicollinear nature of

$$\begin{pmatrix} \vdots & \vdots \\ \frac{\mathbf{x}\beta_E}{m_E} & \frac{\mathbf{x}_T}{\sqrt{m_T}} \\ \vdots & \vdots \end{pmatrix},$$

this approach leads to very low detection power (**Supplementary Figures 6, 7, 8, 9; Supplementary Data 2, 3, 4, 7 and Supplementary Note 1**). We therefore applied ridge regression to solve the equation and determined a T statistic and subsequent Wald test p value for ridge regression (70). Due to the over-conservative nature of the resulting p value in simulations and real data (**Supplementary Figures 6, 7, 8, 10; Supplementary Tables 2, 3, 4, 7 and Supplementary Note 1**), we adjusted the p value distribution of each different scenario by fitting a beta distribution to null estimates to calibrate the final p values (**Supplementary Note 1**). When we report results for MR-link, it is these adjusted p values that we are referring to.

Mendelian Randomization analyses

Causal relationships were estimated with MR-link and four other existing methods: Inverse variance weighting (IVW) (29), LDA-MR-Egger regression (17), MR-Egger regression (18) and MR-PRESSO (71). All methods were (re-)implemented in Python and compared to present equal results when compared with their original implementation. The corresponding code is available at https://github.com/adriaan-vd-graaf/genome_integration.

The IVW method is a weighted meta-analysis of causal estimates from single IVs using weights proportional to the inverse variance of such estimates using the two-terms definition of standard error following Burgess and Thompson (72).

MR-Egger regression adjusts for average pleiotropy by fitting a weighted linear regression between the exposure SNP-effects and the outcome SNP-effects (18). It assumes that < 50% of the variants have a pleiotropic effect. MR-Egger can be applied when three or more instruments are available. LDA-MR-Egger is similar to MR-Egger but also recognizes LD. LDA-MR-Egger can only be used when LD information between the IVs is available (17,73).

MR-PRESSO is a method of causal inference that implements an approach to identify and remove outliers from the IVW framework (19). It assumes that < 50% of the variants have a pleiotropic effect. MR-PRESSO is unable to adjust for the presence of pleiotropy if fewer than three IVs are available, or if fewer than two IVs are left after outlier correction.

We applied these four methods to both simulated and real data. For real data, we used the LDL-C full GWAS summary statistics derived from the association carried out in the Lifelines study, as described above.

Prior to MR analyses, for each IV, we select the allele with positive effect on the exposure.

Colocalization analyses

We have run colocalization analyses on the simulated data using the R package *coloc* v4, git commit [6f3cbb1e5e90f07de772339d6e4af362140af3c3](https://github.com/coloc/coloc/commit/6f3cbb1e5e90f07de772339d6e4af362140af3c3), specifically its *coloc.abf()* function for the original *coloc* functionality and the *coloc.signals()* function for the masked (*coloc-masked*) and conditional (*coloc-cond*) estimates (30,33). We used marginal effect-sizes, standard errors and the MAFs as input that were calculated separately for the exposure and outcome. The LD for the conditional and masked *coloc* analysis was derived from the simulated reference cohort. For original *coloc*, we used the H4 test statistic of the *coloc.abf()* function as our result metric, which provides the posterior probability of sharing of the causal variants between the two traits being tested. For the *coloc-cond* and *coloc-masked* results, we have used the maximum PP4 reported by the *coloc.signals()* function, as this represents the largest posterior probability that a causal variant is shared between traits. We compared the discriminative ability of all *coloc* variations with that of MR-link using i) false positive rate and power when using a PP4 > 0.9 to declare colocalization and ii) an area under the curve (AUC) statistic of the receiver operator curve, where scenarios with (null causal effect of the exposure) were considered true negative observations and were considered the true positive observations. We determined the AUC using the sklearn library (74).

Supplementary information

The Supplementary Material for this article can be found online at: <https://www.nature.com/articles/s41467-020-18716-x#Sec21>

Data availability:

Individual level data of Lifelines cohorts are available to all bona-fide researchers upon request to the Lifelines biobank (<https://www.lifelines.nl/researcher>). Individual level data (genotypes and RNA-seq data) of the BIOS Consortium cohorts can be downloaded only by researchers of Dutch Institutes only, or analyzed (but not downloaded) by any non-Dutch researcher in a Cloud environment (<https://www.bbmri.nl/acquisition-use-analyze/bios>). GTEx summary statistics can be downloaded from the GTEx website (<https://gtexportal.org/home/datasets>). pQTLs summary statistics can be downloaded from GWAS Catalog ([ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/SunBB_29875488](ftp://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/SunBB_29875488)). Simulated data can be recreated using the code at the link provided in Code availability statement. Imputation with the Haplotype reference consortium dataset can be done at the following link: (<https://imputationserver.sph.umich.edu/index.html#!>)

Raw data used to draw **Figure 3** can be found in **Supplementary Data 3**
Code availability: An implementation of MR-link, the methods to recreate the simulated data and instructions on usage can be found at https://github.com/adriaan-vd-graaf/genome_integration.

Acknowledgments

We are very grateful for the altruistic donation of biological materials and questionnaire data by our generous study participants, without them this study would not be possible. In addition, we thank the UMCG Genomics Coordination center and the UG Center for Information Technology, and their sponsors BBMRI-NL & TarGet, for storage and computing infrastructure. We thank BBMRI-NL for providing the transcriptome and genotyped data for the BIOS cohort. We thank P. Visscher, N. Wray, J. Yang, E. Lopera-Maya, O. Bakker and N. de Klein for valuable advice during the development and writing of this work. We also thank K. McIntyre for editorial assistance and Christian Benner for support on running FINEMAP. This work is financed by the Netherlands Organization for Scientific Research (NWO): NWO Spinoza Prize SPI 92-266 (to C.W.) and by Fondation Lefoulon-Delalande (to A.R.).

Author contributions

A.v.d.G. conceived and designed MR-link with critical input from S.S.; A.v.d.G. performed all simulations and data analyses on the datasets used in this study; A.v.d.G. and A.C. performed quality control analyses on the BIOS and Lifelines cohorts; B.C. and C.W. provided access to the datasets used in this study; A.v.d.G. and S.S. wrote the manuscript with critical inputs from Y.L., H.J.W., A.R. and A.C.; Y.L., S.S. and C.W. supervised the study; C.W. provided funding for this study. All authors read and approved the manuscript.

Conflict of interest

The authors declare that there is no conflict of interest regarding the publication of this article.

BIOS Consortium

Bastiaan T. Heijmans, Peter A.C. 't Hoen, Joyce van Meurs, Rick Jansen, Lude Franke

1. Burgess S, Foley CN, Zuber V. Inferring Causal Relationships Between Risk Factors and Outcomes from Genome-Wide Association Study Data. *Annu Rev Genomics Hum Genet.* 2018;19(1):303–27.
2. Pingault J-B, O'Reilly PF, Schoeler T, Ploubidis GB, Rijdsdijk F, Dudbridge F. Using genetic data to strengthen causal inference in observational research. *Nat Rev Genet.* 2018 Sep;19(9):566.
3. Evans DM, Davey Smith G. Mendelian Randomization: New Applications in the Coming Age of Hypothesis-Free Causality. *Annu Rev Genomics Hum Genet.* 2015;16(1):327–50.
4. Ference BA, Yoo W, Alesh I, Mahajan N, Mirowska KK, Mewada A, et al. Effect of long-term exposure to lower low-density lipoprotein cholesterol beginning early in life on the risk of coronary heart disease: a Mendelian randomization analysis. *J Am Coll Cardiol.* 2012 Dec 25;60(25):2631–9.
5. Ference BA, Kastelein JJP, Ginsberg HN, Chapman MJ, Nicholls SJ, Ray KK, et al. Association of Genetic Variants Related to CETP Inhibitors and Statins With Lipoprotein Levels and Cardiovascular Risk. *JAMA.* 2017 12;318(10):947–56.
6. Voight BF, Peloso GM, Orho-Melander M, Frikke-Schmidt R, Barbalic M, Jensen MK, et al. Plasma HDL cholesterol and risk of myocardial infarction: A mendelian randomisation study. *The Lancet.* 2012 Aug;380(9841):572–80.
7. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet.* 2016 May;48(5):481–7.
8. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BWJH, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet.* 2016 Mar;48(3):245–52.
9. Luijk R, Dekkers KF, Iterson M van, Arindrarto W, Claringbould A, Hop P, et al. Genome-wide identification of directed gene networks using large-scale population genomics data. *Nat Commun.* 2018 Aug 6;9(1):3097.
10. Gamazon ER, Wheeler HE, Shah KP, Mozaffari S V., Aquino-Michaels K, Carroll RJ, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet.* 2015 Sep;47(9):1091–8.
11. Li YI, Geijn B van de, Raj A, Knowles DA, Petti AA, Golan D, et al. RNA splicing is a primary link between genetic variation and disease. *Science.* 2016 Apr 29;352(6285):600–4.
12. Dobbyn A, Huckins LM, Boocock J, Sloofman LG, Glicksberg BS, Giambartolomei C, et al. Landscape of Conditional eQTL in Dorsolateral Prefrontal Cortex and Co-localization with Schizophrenia GWAS. *Am J Hum Genet.* 2018 Jun;102(6):1169–84.
13. Zhernakova DV, Deelen P, Vermaat M, Iterson M van, Galen M van, Arindrarto W, et al. Identification of context-dependent expression quantitative trait loci in whole blood. *Nat Genet.* 2017 Jan;49(1):139–45.
14. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell.* 2017 Jun;169(7):1177–86.
15. Liu B, Gloudemans MJ, Rao AS, Ingelsson E, Montgomery SB. Abundant associations with gene expression complicate GWAS follow-up. *Nat Genet.* 2019 May;51(5):768.
16. Liu X, Li YI, Pritchard JK. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. *Cell.* 2019 May;177(4):1022–1034.e6.
17. Barfield R, Feng H, Gusev A, Wu L, Zheng W, Pasaniuc B, et al. Transcriptome-wide association studies accounting for colocalization using Egger regression. *Genet Epidemiol.* 2018;42(5):418–33.
18. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol.* 2015 Apr 1;44(2):512–25.
19. 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 May;50(5):693–8.
20. Zhu Z, Zheng Z, Zhang F, Wu Y, Trzaskowski M, Maier R, et al. Causal associations between risk factors and common diseases inferred from GWAS summary data. *Nat Commun.* 2018 Jan 15;9(1):224.

21. Berzuini C, Guo H, Burgess S, Bernardinelli L. A Bayesian approach to Mendelian randomization with multiple pleiotropic variants. *Biostatistics*. 2020 Aug 1;21(1):86–101.
22. Burgess S, Thompson SG. Multivariable Mendelian randomization: The use of pleiotropic genetic variants to estimate causal effects. *Am J Epidemiol*. 2015 Feb;181(4):251–60.
23. Porcu E, Rieger S, Lepik K, Agbessi M, Ahsan H, Alves I, et al. Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. *Nat Commun*. 2019 Jul;10(1):377267.
24. Consortium Gte. Genetic effects on gene expression across human tissues. *Nature*. 2017 Oct;550(7675):204–13.
25. Sun BB, Maranville JC, Peters JE, Stacey D, Staley JR, Blackshaw J, et al. Genomic atlas of the human plasma proteome. *Nature*. 2018 Jun 7;558(7708):73–9.
26. Zernakova D V., Deelen P, Vermaat M, Van Iterson M, Van Galen M, Arindrarto W, et al. Identification of context-dependent expression quantitative trait loci in whole blood. *Nat Genet*. 2017 Jan;49(1):139–45.
27. Yang J, Ferreira T, Morris AP, Medland SE, Madden PAF, Heath AC, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet*. 2012 Apr;44(4):369–75.
28. Benner C, Spencer CCA, Havulinna AS, Salomaa V, Ripatti S, Pirinen M. FINEMAP: Efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*. 2016 Jan 14;32(10):1493–501.
29. Burgess S, Butterworth A, Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet Epidemiol*. 2013 Nov;37(7):658–65.
30. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. Williams SM, editor. *PLoS Genet*. 2014 May 15;10(5):e1004383.
31. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, et al. A global reference for human genetic variation. *Nature*. 2015 Oct;526(7571):68–74.
32. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*. 2015;4(1):7.
33. Wallace C. Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. Epstein MP, editor. *PLoS Genet*. 2020 Apr 20;16(4):e1008720.
34. Scholtens S, Smidt N, Swertz MA, Bakker SJL, Dotinga A, Vonk JM, et al. Cohort Profile: LifeLines, a three-generation cohort study and biobank. *Int J Epidemiol*. 2015 Aug;44(4):1172–80.
35. Ongen H, Brown AA, Delaneau O, Panousis NI, Nica AC, Dermitzakis ET. Estimating the causal tissues for complex traits and diseases. *Nat Genet*. 2017 Dec;49(12):1676–83.
36. Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, et al. Discovery and refinement of loci associated with lipid levels. *Nat Genet*. 2013 Nov;45(11):1274–85.
37. Klarin D, Damrauer SM, Cho K, Sun Y V., Teslovich TM, Honerlaw J, et al. Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. *Nat Genet*. 2018 Nov;50(11):1514–23.
38. Ander BP, Dupasquier CMC, Prociuk MA, Pierce GN. Polyunsaturated fatty acids and their effects on cardiovascular disease. *Exp Clin Cardiol*. 2003;8(4):164–72.
39. Lemaitre RN, Tanaka T, Tang W, Manichaikul A, Foy M, Kabagambe EK, et al. Genetic Loci Associated with Plasma Phospholipid n-3 Fatty Acids: A Meta-Analysis of Genome-Wide Association Studies from the CHARGE Consortium. *PLOS Genet*. 2011 Jul 28;7(7):e1002193.
40. Barchetta I, Cimini FA, Capoccia D, Bertocchini L, Ceccarelli V, Chiappetta C, et al. Neurotensin is a lipid-induced gastrointestinal peptide associated with visceral adipose tissue inflammation in obesity. *Nutrients*. 2018 Apr;10(4).
41. Earnest CP, Jordan AN, Safir M, Weaver E, Church TS. Cholesterol-lowering effects of bovine serum immunoglobulin in participants with mild hypercholesterolemia. *Am J Clin Nutr*. 2005 Apr;81(4):792–8.

chapter 7

42. Kjolby M, Andersen OM, Breiderhoff T, Fjorback AW, Pedersen KM, Madsen P, et al. Sort1, encoded by the cardiovascular risk locus 1p13.3, is a regulator of hepatic lipoprotein export. *Cell Metab.* 2010 Sep;12(3):213–23.
43. Musunuru K, Strong A, Frank-Kamenetsky M, Lee NE, Ahfeldt T, Sachs K V., et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature.* 2010 Aug;466(7307):714–9.
44. Wang X, Raghavan A, Peters DT, Pashos EE, Rader DJ, Musunuru K. Interrogation of the Atherosclerosis-Associated SORT1 (Sortilin 1) Locus with Primary Human Hepatocytes, Induced Pluripotent Stem Cell-Hepatocytes, and Locus-Humanized Mice. *Arterioscler Thromb Vasc Biol.* 2018 Jan;38(1):76–82.
45. Phillips MC. Apolipoprotein e isoforms and lipoprotein metabolism. *IUBMB Life.* 2014 Sep 1;66(9):616–23.
46. Erbilgin A, Siemers N, Kayne P, Yang WP, Berliner J, Lusis AJ. Gene expression analyses of mouse aortic endothelium in response to atherogenic stimuli. *Arterioscler Thromb Vasc Biol.* 2013 Nov;33(11):2509–17.
47. Rossignoli A, Shang MM, Gladh H, Moessinger C, Foroughi Asl H, Talukdar HA, et al. Poliovirus Receptor-Related 2: A Cholesterol-Responsive Gene Affecting Atherosclerosis Development by Modulating Leukocyte Migration. *Arterioscler Thromb Vasc Biol.* 2017 Mar;37(3):534–42.
48. Skogsberg J, Lundstrum J, Kovacs A, Nilsson R, Noori P, Maleki S, et al. Transcriptional profiling uncovers a network of cholesterol-responsive atherosclerosis target genes. *PLoS Genet.* 2008 Mar;4(3):e1000036.
49. Blattmann P, Schubert C, Pepperkok R, Runz H. RNAi-Based Functional Profiling of Loci from Blood Lipid Genome-Wide Association Studies Identifies Genes with Cholesterol-Regulatory Function. *PLoS Genet.* 2013 Feb;9(2):e1003338.
50. Candia J, Cheung F, Kotliarov Y, Fantoni G, Sellers B, Griesman T, et al. Assessment of Variability in the SOMAscan Assay. *Sci Rep.* 2017 Dec 1;7(1):1–13.
51. Klop B, van de Geijn GJM, Bovenberg SA, van der Meulen N, Elte JWF, Birnie E, et al. Erythrocyte-Bound Apolipoprotein B in Relation to Atherosclerosis, Serum Lipids and ABO Blood Group. Eller P, editor. *PLoS ONE.* 2013 Sep 19;8(9):e75573.
52. McLachlan S, Giambartolomei C, White J, Charoen P, Wong A, Finan C, et al. Replication and characterization of association between ABO SNPs and red blood cell traits by meta-analysis in Europeans. *PLoS ONE.* 2016 Jun 1;11(6).
53. Gaziano JM, Concato J, Brophy M, Fiore L, Pyarajan S, Breeling J, et al. Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J Clin Epidemiol.* 2016 Feb 1;70:214–23.
54. Leitsalu L, Haller T, Esko T, Tammesoo M-L, Alavere H, Snieder H, et al. Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *Int J Epidemiol.* 2015 Aug 1;44(4):1137–47.
55. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Med.* 2015 Mar;12(3):e1001779.
56. Tigchelaar EF, Zernakova A, Dekens JAM, Hermes G, Baranska A, Mujagic Z, et al. Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: Study design and baseline characteristics. *BMJ Open.* 2015 Aug;5(8):e006772.
57. Huisman MHB, de Jong SW, van Doormaal PTC, Weinreich SS, Schelhaas HJ, van der Kooi AJ, et al. Population based epidemiology of amyotrophic lateral sclerosis using capture-recapture methodology. *J Neurol Neurosurg Psychiatry.* 2011 Oct;82(10):1165–70.
58. Deelen J, van den Akker EB, Trompet S, van Heemst D, Mooijaart SP, Slagboom PE, et al. Employing biomarkers of healthy ageing for leveraging genetic studies into human longevity. *Exp Gerontol.* 2016;82:166–74.
59. Lin BD, Willemsen G, Abdellaoui A, Bartels M, Ehli EA, Davies GE, et al. The Genetic Overlap between Hair and Eye Color. *Twin Res Hum Genet.* 2016;19(6):595–9.
60. van Greevenbroek MMJ, Jacobs M, van der Kallen CJH, Vermeulen VMMJ, Jansen EHJM, Schalkwijk CG, et al. The cross-sectional association between insulin resistance and circulating complement C3 is partly explained by

plasma alanine aminotransferase, independent of central obesity and general inflammation (the CODAM study). *Eur J Clin Invest*. 2011;41(4):372–9.

61. Hofman A, Brusselle GGO, Murad SD, van Duijn CM, Franco OH, Goedegebuure A, et al. The Rotterdam Study: 2016 objectives and design update. *Eur J Epidemiol*. 2015;30(8):661–708.
62. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet*. 2016 Oct;48(10):1279–83.
63. Das S, Forer L, Schunherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. *Nat Genet*. 2016 Oct;48(10):1284–7.
64. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013 Jan 1;29(1):15–21.
65. Boomsma DI, Wijmenga C, Slagboom EP, Swertz MA, Karssen LC, Abdellaoui A, et al. The Genome of the Netherlands: Design, and project goals. *Eur J Hum Genet*. 2014 Feb;22(2):221–7.
66. Anders S, Pyl PT, Huber W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*. 2015 Jan 15;31(2):166–9.
67. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet*. 2012 Aug;44(8):955–9.
68. Friedewald WT, Levy RI, Fredrickson DS. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clin Chem*. 1972 Jun;18(6):499–502.
69. Su Z, Marchini J, Donnelly P. HAPGEN2: simulation of multiple disease SNPs. *Bioinforma Oxf Engl*. 2011 Aug 15;27(16):2304–5.
70. Cule E, Vineis P, De Iorio M. Significance testing in ridge regression for genetic data. *BMC Bioinformatics*. 2011 Sep 19;12(1):372.
71. Verbanck M, Chen CY, Neale B, Do R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat Genet*. 2018 May;50(5):693–8.
72. Burgess S, Thompson SG. Mendelian randomization: Methods for using genetic variants in causal estimation. *Mendelian Randomization: Methods for Using Genetic Variants in Causal Estimation*. CRC Press; 2015. 1–207 p.
73. Bowden J, Smith GD, Burgess S. Mendelian randomization with invalid instruments: Effect estimation and bias detection through Egger regression. *Int J Epidemiol*. 2015 Apr;44(2):512–25.
74. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. *J Mach Learn Res*. 2011;12:2825–30.
75. Hoffmann TJ, Theusch E, Haldar T, Ranatunga DK, Jorgenson E, Medina MW, et al. A large electronic-health-record-based genome-wide study of serum lipids. *Nat Genet*. 2018 Mar 1;50(3):401–13.
76. Maisse C, Rossin A, Cahuzac N, Paradisi A, Klein C, Haillot ML, et al. Lipid raft localization and palmitoylation: Identification of two requirements for cell death induction by the tumor suppressors UNC5H. *Exp Cell Res*. 2008 Aug;314(14):2544–52.
77. Falk J, Rohde M, Bekhite MM, Neugebauer S, Hemmerich P, Kiehntopf M, et al. Functional Mutation Analysis Provides Evidence for a Role of REEP1 in Lipid Droplet Biology. *Hum Mutat*. 2014 Apr;35(4):497–504.
78. Veniaminova NA, Grachtchouk M, Doane OJ, Peterson JK, Quigley DA, Lull M V., et al. Niche-Specific Factors Dynamically Regulate Sebaceous Gland Stem Cells in the Skin. *Dev Cell*. 2019 Nov 4;51(3):326–340.e4.
79. Sugiura-Ogasawara M, Omae Y, Kawashima M, Toyo-Oka L, Khor SS, Sawai H, et al. The first genome-wide association study identifying new susceptibility loci for obstetric antiphospholipid syndrome. *J Hum Genet*. 2017 Sep;62(9):831–8.
80. Li W, Ji M, Lin Y, Miao Y, Chen S, Li H. DEPP/DEPP1/C10ORF10 regulates hepatic glucose and fat metabolism partly via ROS-induced FGF21. *FASEB J*. 2018 Apr 27;32(10):5459–69.

Adriaan van der Graaf

chapter 8

Discussion

In the introduction of this thesis, I described my main aim:

To identify causal relationships between human -omics and complex traits

I have tried to achieve this aim by performing observational causal inference from genetics data using techniques like Mendelian randomization (MR). These techniques require robustly associated genetic loci before they can be applied. Thus, my activities can be separated into two sections. In the first section, we identified genetic locations that alter the risk for complex traits like diseases in the hope that we can perform observational causal inference on this data at a later stage. In the second section, we applied causal inference techniques to identify causal relationships in -omics phenotypes. We partly achieve the aim of this thesis as we describe how we have identified causal relationships between human complex traits and -omics data, including measurements of gut microbiome composition and function, mRNA expression and protein levels. In the next sections, I will discuss my work, its findings and its drawbacks. I then describe the directions in which I think the field of -omics causal inference is going. Finally, looking toward the future, I describe how newly found causal relationships might be used, and how their estimation can be improved.

Section 1: Finding robustly associated genetic locations

The first section of this thesis starts with [*chapter 2*](#), where we evaluated whether genetic variants in or nearby genes important for SARS-COV-2 infection were associated with human quantitative phenotypes or drug usage. We wrote this chapter in the early months of the COVID-19 pandemic. At the time, COVID-19 infection severity was observed to be different across age groups and the sexes. Furthermore, a spectrum of different symptoms was observed but it was poorly understood why this happened. We therefore wondered if we could explain some of this variability with genetics. Two genes are necessary for SARS-COV-2 infection, *TMPRSS2* and *ACE2* (1). Using 36,339 genotyped individuals of the Lifelines population cohort, we associated 178 quantitative phenotypes and the usage of 58 commonly used drugs to the genetics in and nearby these genes. We specifically investigated any potential sex-specific associations to identify factors that could explain any sex-dependent COVID-19 severity. Our results indicate that common genetic variation in and around these genes does not alter human phenotypes and are thus unlikely to explain differences in COVID-19 symptoms and severity. These results were later confirmed by a large COVID-19 case-control genome wide association study (GWAS) conducted by the Covid 19 host genome initiative (2). We did detect a suggestive association for the usage of drugs at common variants in *ACE2*, including angiotensin inhibitors, but with only mild effects (3). Given that none of these variants are associated with COVID19 symptoms and severity, the association we found likely represents a signal of genetic association with drug efficacy

and response, rather than being related to the SARS-COV-2 infection.

Chapter 3 describes the steps taken to identify genetic locations that are important for celiac disease susceptibility. We examined an existing cohort of 24,269 individuals of European ancestry (4,5) and combined this cohort with 4,125 new individuals from European and Argentinian descent. This increase in sample size identified two additional celiac disease associated loci. Unfortunately, we did not identify any Argentinian ancestry-specific associations, which could be attributed to the low coverage genotyping from the Immunochip measurements (less than 200,000 variants analyzed) but may also be due to the limited sample size of the Argentinian subpopulation. Nevertheless, inclusion of these two new loci suggested new pathways involved in celiac disease: the tumor necrosis factor (TNF) mediated signaling pathway and positive regulation of I- κ B kinase/NF- κ B signaling. To better understand celiac disease genetics, a non-targeted genotyping array or low-pass sequencing efforts should be undertaken to agnostically analyze the entire genome.

Section 2: Causality using -omics data

In the second section of this thesis, we use genetic locations - including those identified in the first section - to identify causal relationships between human molecular measurements and other human traits. In *chapter 4*, we mapped the genetic locations that influence 16S-based microbiome abundance in 18,340 individuals distributed across 24 cohorts. We identified a robust study-wide, age-dependent association of the *LCT* gene with *Bifidobacterium* abundance, confirming associations detected in smaller cohorts (6–8). Variations in 16S methodology, DNA isolation and cohort ethnicities likely influenced our chances of detecting loci at our stringent study-wide significant threshold. Using bacteria that showed at least genome-wide significant associations and applying MR, we tried to identify downstream effects of these associations and found that there could be causal effects of microbiome abundances of the class Actinobacteria and the genus *Bifidobacterium* on ulcerative colitis as well as a putative causal relationship of the *Oxalobacteraceae* family on rheumatoid arthritis. Interestingly, the gut microbiome has been associated to both autoimmune and gastrointestinal diseases but, given that genetic associations with microbiome were not study-wide significant and replication was not assessed, these putative causal relationships need to be interpreted with caution.

In *chapter 5*, we performed another GWAS on microbiome traits, but now using the more sensitive metagenomic sequencing (MGS) method. MGS allows for more fine-grained microbial measurement compared to 16S sequencing, as well as allowing for the mapping of functional elements of the microbiome. Performing an MGS GWAS in 7,738 Dutch individuals confirmed the association between *Bifidobacteria* and variants at the *LCT* locus that was seen in *chapter 4*, but we also found strong evidence for association with variants in the *ABO* locus, which seems to have an interaction with secretor status determined by a missense variant in the *FUT2* gene. While we identified one additional locus, our sample

chapter 8

size was limited for detecting new loci, based on a power estimation. Considering the sparsity of microbiome data and the low heritability of microbiome taxa and pathways, cohorts of at least 50,000 individuals would be needed to identify variants with small to moderate effect sizes. As in chapter 4, we applied MR and identified putative causal relationships between the microbiome and human phenotypes, in this case an effect of the relative abundance of family Rikenellaceae and genus *Alistipes* on salt intake and of the relative abundance of the genus *Collinsella* on triglyceride levels. As with our findings in [chapter 5](#), these associations are putative due to the relatively lenient p value thresholds used, and should thus be interpreted with caution.

[Chapter 6](#) uses the genetic loci we found to be associated with celiac disease in [chapter 3](#) to systematically identify the most likely causal genes in these loci for celiac disease. We mapped expression quantitative trait loci (eQTL) in blood for 3,502 individuals from the BIOS cohort (9) and combined these using multiple prioritization methods: MR, colocalization, linkage disequilibrium (LD) overlap between eQTL and GWAS variants and finally using a co-regulation enrichment approach. We identified 118 genes that are potentially causal to celiac disease, some of which were supported by only one of these methods, while one gene was supported by all prioritization approaches. Another gene from this prioritized list, *TRAFD1*, supported by three methods, was particularly interesting as it statistically mediated 41 other genes in *trans*. Upon knockdown, we could show that *TRAFD1* influences these 41 other genes using a novel permutation scheme. Interestingly, these 41 genes are enriched for celiac disease-specific biology, namely INF γ signaling and major histocompatibility complex antigen processing/presentation. These findings provide insights into the biology of celiac disease and could lead to new celiac disease treatment opportunities.

In [chapter 7](#), we tried to identify the best way to do MR on gene expression traits. We were partly motivated by the variability of the prioritization results of [chapter 6](#), where it seemed that prioritization methods were very variable in their output. In this chapter, we developed a new MR method called MR-link that is pleiotropy-robust even when only a single SNP as instrumental variable (IV) is available and can be easily applied to gene-expression traits. We benchmarked MR-link in simulations, where it had the highest discriminative ability compared to other MR methods and coloc (10). Furthermore, we applied MR-link to eQTL from different tissues and protein QTL studies where we identified gene expression and protein levels that were causal for LDL-cholesterol. We identified a causal relationship between *PVRL2* and LDL-cholesterol levels, a relationship that could not be identified by other MR methods given the high LD complexity and pleiotropy at the *PVRL2* genomic region.

In summary, we have combined i) GWAS, ii) -omics QTL studies and iii) MR approaches to identify causal relationships between complex traits and molecular phenotypes (-omics). For each of these three fields, I now discuss the state of the art, describe my vision of current limitations and explain what I think is still needed to better understand causal relationships in the future.

State of the art for GWAS

Since their inception, GWAS have steadily increased in sample size and the number of variants being assessed. Initially, GWAS were performed in hundreds to thousands of individuals and assessed tens of thousands of genetic variants (11,12). These days, GWAS are carried out in large cohort studies, sometimes comprising more than a million individuals analyzing millions of DNA markers (13–15). Most common genetic variation has now been associated with at least one complex trait or disease. The availability of genotyped samples in well-established biobanks and the ability to measure new phenotypes in the same individuals allows us to apply GWAS to any newly measured trait or to use these data to identify the genetic associations in genes that are relevant for SARS-COV-2 infection in [chapter 2](#). The same cohort, the Lifelines Cohort Study and Biobank, is also being used to identify the psychological and societal impact of the COVID-19 pandemic by re-contacting volunteers (16).

In addition to expanding sample sizes and extending the number of genetic variants tested, GWAS are evolving into two directions: i) inclusion of populations underrepresented in existing studies and ii) the investigation of rare and structural variation.

Most GWAS studies have been performed on European and East Asian populations. In consequence, the genetic variation that is most prevalent in these populations has been analyzed thoroughly and associated with human phenotypes, while variants that are prevalent in underrepresented populations have yet to be characterized, so that their function can be understood (17,18). Fortunately, underrepresented populations can still benefit from GWAS in other large populations, since the biological mechanisms underlying a specific GWAS locus are likely to be shared between populations. Furthermore, polygenic risk scores derived from European populations still have some predictive power in non-European ancestries, albeit with lower accuracy (19–21). There are many advantages in performing GWAS on underrepresented populations. For example, underrepresented populations have different LD structures and allele frequencies compared to Europeans (22,23). These characteristics allow for a better understanding of the causal variants in a GWAS locus if multiple populations are analyzed together. Inclusion of ethnically diverse individuals will also help to identify causes of disease that are specific to ethnicity, which is a necessary condition for the realization of effective personalized medicine that can be applied to individuals of all ancestries. In [chapter 3](#), we sought to enlarge our understanding of celiac disease in non-European populations by also including 4,125 subjects from Argentina in a large meta-analysis of European samples. With the additional samples, we were able to find two additional genetic loci. Unfortunately, the number of Argentinian samples was too small to detect Argentinian-specific genes (4). Perhaps this is not surprising, considering that the Argentinian population is partly of European descent. However, there is potential for finding new loci from admixed and fully Native American individuals. It stands to reason that very large sample sizes are needed to serve underrepresented populations, which will

chapter 8

require substantial investment. Fortunately, large population biobanking efforts are currently ongoing in Africa (24) and Mexico (<http://www.mxbiobankproject.org/>) and many more are expected to start.

Next to building cohorts from ethnically diverse populations, a major recent push in GWAS analysis is to move from genotyping arrays and imputation type genotyping to sequencing approaches for genotyping (25–28). In contrast to array-based genotyping, where only a limited number of genetic variants are measured and the remainder are imputed (interpolated if you will), sequencing techniques directly measure the DNA of an individual. As one would expect, array genotyping is less expensive than sequencing techniques. However, array genotypes are only suitable to identify common and known pathogenic DNA variations, and thus will not be able to paint a good picture of rare and structural variants in an individual. Nevertheless, rare variants can be very interesting if they are associated with a disease. Rare variants usually come from relatively recent mutations and have not been under selective pressure for as long as more common variation, thus their effect sizes can be much larger than those seen for common variation (29,30). With their relatively large effect sizes, rare variants can be used to understand major components of a disease. This is especially true when these variants are protein-coding changes. For example, they can aid to validate genes that would be potential therapeutic targets in both rare and common diseases (31)

For the identification of rare variants, however, very large sample sizes are needed because otherwise a rare variant will only be observed in a few individuals, leading to insufficient statistical power to identify a high confidence association signal. The fact that sequencing techniques are so expensive, combined with the need for large sample sizes and the large potential for drug target discovery when performing rare variant analyses, has prompted pharmaceutical companies to invest heavily into sequencing techniques in large biobanks such as the UK biobank and FinnGen (32–35) which are large enough, and deeply phenotyped, allowing for the identification of robust rare variant associations.

There is still substantial work to be done to map the genome for rare and structural variants and to understand how genetics shapes human traits in underrepresented populations. Fortunately, one can inform the other. A rare variant may not be rare in all populations. It could be more prevalent in other populations, thereby allowing us to assess the effect of this rare variant more easily. Interestingly, recent work has shown that low-depth sequencing of genotypes of underrepresented populations in large sample sizes may be more informative than genotyping by array followed by imputation at a comparable cost (36), potentially indicating that genotyping arrays are becoming obsolete for GWAS analysis.

The developments in the field of GWAS allow for improved causal inference. As we discussed, rare variants are more likely to be causal, making a strong argument for a causal relationship between the variant and the gene that it may be located in. Furthermore, techniques like MR and colocalization are dependent on robustly associated genetic variants. Cross-ancestry analysis is likely to strengthen inferences of causality.

However, there are some drawbacks to these new analyses. Cross-ethnic studies are likely to have inconsistent LD patterns, making it difficult to compare results across different studies. Moreover, if we use an MR method that is sensitive to this, such as MR-link, it could make incorrect causal inferences. On top of that, rare variants are unlikely to be in strong LD because they are more likely to be recent de novo mutations, which may also hamper analysis through MR-link that relies on LD patterns around the variant as IV. This issue could be solved by performing population-matched -omics QTL mapping such that both cohorts have the same underlying LD structure, which would make these issues less problematic.

State of the art for genetic association mapping in -omics analyses

Over the years, GWAS has become a powerful and versatile approach that can be applied to any quantitative or binary trait. GWAS is also used to map genetic variations that are associated with molecular phenotypes, such as variations in molecule abundance, transcription and translation. Today, this can be done with massive systematic screenings. High-throughput sequencing techniques allow us to measure not only an individual's genotype, but also their epigenetics, transcriptome, proteome and microbiome (37–41). These techniques were initially used to describe each layer of information and to perform treatment-control experiments. Over the course of time, as sufficiently large sample size cohorts were created, large-scale QTL studies have been set up to identify the genetics underlying these molecular traits derived through sequencing (42–45). These studies have been instrumental to help understand how genetics shapes human molecular phenotypes. Furthermore, there is a great push to understand all the microorganisms in and around the human body with which we have an important symbiotic relationship. Using sequencing techniques, we are now able to measure the microbiomes of the skin, oral cavity and the gut. As we described in [chapter 4](#) and [chapter 5](#), we measured the abundances of certain bacteria in the human gut and were able to identify QTL that influence the abundance of these bacteria, as well as their function (46,47).

As I see it, molecular QTL studies are expanding in four directions: i) increasing sample sizes through the careful meta-analysis of multiple single QTL studies, ii) identifying molecular QTL in difficult-to-access tissues and contexts, iii) performing molecular QTL studies in single cells and iv) performing QTL studies in the same individuals across different layers of measurements. A single -omics QTL study usually doesn't just follow a single direction, but often combines two or more of them. For ease of argument, I mainly discuss gene expression measurements because gene expression is a highly heritable trait and it has been shown that gene expression QTLs often overlap with disease loci (48). However, the developments listed here are also valid for most other -omics analysis techniques.

Current -omics QTL studies are still often underpowered. Fortunately, it's

chapter 8

possible to increase their power by performing meta-analysis of multiple cohorts. For instance, GWAS has benefited immensely from meta-analysis techniques that allow for single association results across different cohorts (49). This increases the power to detect genetic locations on the genome. In the same way, expression quantitative trait locus (eQTL) studies have been meta-analyzed by the eQTLGen consortium, finding that most of the genes expressed in blood are regulated by genetics (50).

Next to an increase in power through meta-analysis, it is important to measure transcription in the right place and time. eQTLs found in blood can be very different from eQTLs in the brain (51,52). This needs to be taken into account when trying to overlap QTL studies with disease GWAS (53). Furthermore, the right context also needs to be known. Immune function usually occurs upon a stimulus, and potential eQTL may be different for the same tissue in a stimulated versus unstimulated state (45). This is also the case for gut microbiome studies, where bacteria are often measured in fecal material. The microbiome is not stable across the gastrointestinal tract, so measuring samples from different locations inside the human gastrointestinal tract will be essential to find out how the microbiome works (54,55).

As a corollary to using the correct tissue, QTLs also need to be identified in the correct cells. Most bulk transcriptome studies use tissues that are a mixture of different cell types, and thus single cell techniques can be used to identify more ‘pure’ QTLs (56–58). One example of this was shown in Bonder et al., where induced-pluripotent stem cells had many more individual QTLs than expected based on the sample size alone (56). Interestingly, there are consortia being set up that are developing ways to meta-analyze QTL from single cells, such as the single cell eQTLGen consortium (59).

Finally, to understand how the -omics measurements are distributed across different tissues, it is prudent to identify how these associations interact with one-another. The GTEx consortium has built a cohort of individuals and measured expression in most of their tissues (52,53). This allows investigators to identify interactions between tissues. These capabilities are also present in single cell studies. For example, as identified by the context specific eQTLs found by van der Wijst et al. (58). Interestingly, this is also possible, across different -omics measurements. For instance, Eales et al. have measured DNA methylation, gene expression and alternative splicing in bulk kidney samples (60).

Research carried out in my thesis aligns with these new directions. We have used -omics QTLs in [chapter 4](#), [chapter 5](#), [chapter 6](#) and in [chapter 7](#). In [chapter 4](#), we used 16S sequencing and, in [chapter 5](#), metagenomic sequencing to identify genetic factors that determine the microbiome composition (4,46). In [chapter 6](#) and [chapter 7](#), we used expression measurements in 3,502 individuals to identify eQTL (61,62), information that we later used to search for causal gene expression for two complex traits: celiac disease and LDL cholesterol measurements.

State of the art for MR

MR is somewhat defined by its poster child: The causal relationship of LDL cholesterol and heart disease and the fact that MR has found that the causal relationship of HDL cholesterol and heart disease is unlikely (63–66). Both these results were subsequently validated in large and expensive clinical trials. These results are not the only validated results for MR. There are many MR studies that have identified novel and known causal relationships, including a causal relationship between body mass index (BMI) and type 2 diabetes. This was already known: weight loss is a common treatment for type 2 diabetes. However, MR was also able to show the novel causal relationship that reducing BMI reduces the risk for type 2 diabetes even when an individual has a very low BMI, and would otherwise not be considered at risk (67,68).

This success is likely to have influenced recent GWAS, as MR analyses are now included in standard follow-up of a GWAS analysis, with several causal relationships identified up to date (14,44,47). These results are encouraging, showing that observational studies are trying to transition to robust ways to perform causal inference in their analysis (69). It is also encouraging to see that the MR literature is converging onto best practice analysis schemes for identifying robust causal relationships through MR, so that the exposure to violations of the assumptions underlying MR is limited (70,71).

Even with the popularization and standardization of MR analyses there are still some pitfalls when performing an MR analysis. These pitfalls can be explained by comparing them to the IV assumptions of MR:

- IV-I:** The instrumental variable is robustly associated to the exposure variable (*Strong instrument*)
- IV-II:** The instrumental variable is not associated with any confounder (*Independence*)
- IV-III:** The instrumental variable is only associated to the outcome through the exposure (*Exclusion*)

Violations of IV-I means that the genetic variant that is used as an IV will not have a sufficient effect on the exposure or risk factor of interest. For instance, when the GWAS did not find robust associations. Violations of IV-II indicate that the instrument is associated with a confounder, for instance when there is a large effect of selection or population stratification or other forms of selection that affect the IV. Lastly, IV-III is violated when the genetic variant selected as an IV has effects on the outcome through pathways other than just the exposure. This can happen when the genetic variant is associated with many different traits. For a more extensive discussion of IVs, please see the introduction of this thesis, where I compared the IV assumptions to a treatment control trial.

Violations of IV-I, also known as weak instrument bias, are not so concerning, as these can mostly be mitigated by making the p value

chapter 8

selection threshold more stringent for the selection of an IV or by estimating the effect of an IV in a third dataset (removing Winner's Curse) (72–74).

IV-II is the assumption that the IV is independent from any confounders in the analysis. This is a difficult-to-test assumption, and, unfortunately, it could be present in many MR analyses. It has been shown that population cohorts are strongly stratified which can potentially violate IV-II (75,76). For instance, we know that the UK biobank contains population stratification in educational attainment measurements (75). If we determine the causal effect of human standing height on educational attainment, the genetic estimates for educational attainment will be biased by this population stratification, and the subsequent causal estimate will also be biased (77–79). Luckily there are ways to mitigate population stratification, but these require the use of family-based studies, and, indeed, the effect of height on educational attainment is attenuated when using study designs like these (77). Unfortunately, cohorts that are of a family design are not often available, although family-based cohorts exist, and are being further developed allowing for family-based MR (80).

Another way that IV-II can be violated is through collider bias (81). This can happen when large amounts of preprocessing have occurred on the exposure or outcome phenotype, making them inadvertently associated to a confounder (81,82). Fortunately, it is possible to remove this bias by simply not correcting a phenotype for confounders in the initial GWAS analysis. On the other hand, this often requires access to individual level data, which is not always available in a study. Collider bias can be especially concerning in -omics type analyses, because the omics trait is usually corrected for many different factors. In all the chapters where I have applied MR, the phenotypes were corrected for covariates and principal components, making collider bias a potential issue in these MR analyses.

Considering all the violations of the assumptions behind MR, to me, the largest issue for MR analysis is violations of the IV-III assumption, which is also known as horizontal pleiotropy (83). There is no way of knowing if a variant is associated with another factor. For example, if we consider an eQTL from blood, even though we could make sure the eQTL is not influencing other genes in the tissue, the same variant could have a different effect in another tissue, or even other cell types inside blood (45,58). There are roughly three ways to reduce the amount of violations of IV-III: i) by jointly estimating an average pleiotropic effect with the causal effect across multiple IVs (61,67,84,85), ii) by removing IVs that are likely pleiotropic and removing from analysis if there is one (83,86,87) and iii) by measuring all known sources of pleiotropy (88,89). We have introduced a new method called *MR-link* in chapter 7 that can correct for pleiotropy without applying these approaches, positioning MR-link as a new class of methods for performing MR in the presence of known and unknown pleiotropy. In this chapter we show that by using the genetic region surrounding the IV as a covariate can, in certain (but often encountered) circumstances, correct for a pleiotropic effect even when there is only a single SNP as IV available, making MR-link exceptionally suitable for applications in -omics datasets. We have applied MR-link

to identify causal relationships between gene expression and LDL-cholesterol, and between protein levels and LDL-cholesterol.

The circumstances in which MR-link functions best are situations where there is the so-called *pleiotropy through LD*: the genetic variant selected as an IV is in LD with genetic variants that have pleiotropic effects. In simulated circumstances of pleiotropy through LD, MR-link has the best discriminative ability compared to all the tested MR methods and coloc. When instead there is the so-called *pleiotropy through overlap*, i.e. the IV has both causal and pleiotropic effects, MR-link is unlikely to function correctly, and this leads to an excess of false positives in simulations. We found however that pleiotropy through overlap does not occur often in the whole blood gene expression dataset that we analyzed: only about 13% of tested genes had partial overlap of IVs, which would partially violate the pleiotropy through LD assumption, whereas 2.8% of genes are fully overlapping and could be violating the pleiotropy assumptions of MR-link fully.

Only considering whole blood gene expression is, of course, not fully sufficient to estimate how much pleiotropy through overlap exists in the human organism. Pleiotropy can come from any phenotype: expression in other tissues and other (molecular) measurements.

I believe however that we can extrapolate from our whole blood estimates. I would suggest that it is reasonable to assume that the amount of genetic variant overlap measured in whole blood gene expression should be close to the amount found in any other arbitrary tissue measurement or measurement condition. As you increase the number of other phenotypic measurements, the probability of encountering a phenotype with overlapping genetic variants would also increase, which is not beneficial for the assumptions of MR-link. However, not all phenotypes are pleiotropic, there are only a limited number of them. Their chance of having overlapping variants would therefore decrease proportionally with the increase of phenotypic measurements. In my view, this makes it reasonable to make inferences of pleiotropy through overlap based on whole blood expression.

Of course, our estimates remain imperfect, for instance, there could be pleiotropy through overlap that is unmeasured due to power issues. Fortunately, these would likely have small effects, and would have less serious consequences in violations of inferences of MR-link. These limitations provide directions on the needs to be satisfied for future research on causal inference methods.

Finding causal mechanisms in GWAS loci

Although interpreting GWAS loci remains difficult, GWAS studies have a great potential for drug target discovery. A gene underlying an associated locus could be a potential drug target, provided that its link to a phenotype or molecular mechanism that leads to the disease is validated. While genes at some of the GWAS loci have been functionally validated (90), for most

chapter 8

of the GWAS loci, we do not have functional evidence of their function (91,92). It remains difficult to validate a potential disease gene due to a lack of experimental techniques that are relevant for a disease: How does one simulate all the intricacies of a disease in a model organism? On top of that, it remains difficult to identify the causal gene that is affected by a given associated genetic variant, or to pinpoint specific causal variants that underlie a GWAS locus.

Combining GWAS and -omics QTL summary statistics allows us to identify potential links between molecular mechanisms and human phenotypes using statistical methods such as causal variant colocalization and MR. These methods have been applied not only in our studies but also widely by the genetic community (93). For instance, the Open Targets consortium publishes a curated list of genes that are potentially causal using colocalization methods and many other layers of information (94). New methods to identify causal genes in a GWAS locus are regularly being proposed (10,61,86,89,95,96). These methods usually validate their models on simulations. Unfortunately, simulations do not fully represent the entire space of biological mechanisms that underlie human traits, so good discriminative performance in simulations does not always assure correct results when applied in real data. Next to simulations, some studies use a few known causal genes to strengthen their inference. Application of the new models to known positives would be a real test, but no genetic locus is the same, and very often it is unknown how the tested models work on true negative results. On top of that, each method uses their own way of simulation, or their own set of true positive causal gene-trait combinations, making it difficult to understand how performance compares between methods.

To better understand the performance of causal gene finding methods, I believe the community should work to create a good and widely accepted gold standard for benchmarking these methods on. These gold standards should be, first and foremost, based on real data, and should test causal inference on known causal relationships, and on known true negatives. Secondly, we should also develop gold standard simulation datasets that are able to systematically test assumptions of the data. These gold-standard datasets already exist to compare deep learning models on a wide variety of data (97). To move the analysis of causal relationships in humans forward, a widely recognized gold standard dataset based on both simulations and known causal relationships should be created.

Using triangulation to confirm causal relationships

So how do we build a robust list of gold standard relationships? First off, causal conclusions should ideally be based on multiple lines of evidence that have biases that counteract each other, so that we do not rely only on a single statistical technique, or way of measuring. Such triangulation of evidence ensures that the causal relationship is robust, and not a false positive due to a violation of an assumption (98,99). By extension, this means that techniques such as MR should not be considered as a single point of truth, but rather as lines of evidence towards the goal of finding a true causal relationship. Ideally, other forms of causal inference should be

included into a causal inference framework.

Interestingly, there are ways to identify causal relationships from observational data without the use of IVs. So-called negative control experiments are able to identify causal relationships by estimating the effects of confounding on matched exposures or outcomes that are known to have no effect (100). Negative control methods can be particularly helpful when dealing with time series data (101). Interestingly, time series experiments are very often used when measuring -omics layers upon stimulation, indicating that these methods can be combined with each other nicely to identify another line of evidence for causality.

It is unlikely that triangulation of causation is possible only based on observational studies, simply due to the fact that individual cohorts are often biased in their measurements and sample ascertainment. This makes it almost a certainty that experimental results remain essential to confirm causal relationships.

Interestingly, new experimental techniques that could complement the biases found in MR and other forms of observational causality are being developed. CRISPR techniques that are used for genome editing can also be used to modulate mRNA expression by targeting transcripts' regulatory regions. These techniques are called CRISPR interference (CRISPRi) and activation (CRISPRa) (102–105) when coupled with high throughput read outs of the perturbed cells or model systems allow for the identification of the causal (downstream) consequences of gene expression differences. Interestingly, these techniques counteract the biases of MR if there is a controlled randomization experiment. This includes the presence of pleiotropy (IV-III), as the transcript itself is perturbed in CRISPRi and CRISPRa, in contrast to a SNP in CRISPR.

These experiments have mostly been performed in cell lines or single cell types, which usually map causal relationships within a cell or a pure tissue (106,107). Interestingly, these CRISPR techniques can be used with model organisms or combined with induced pluripotent stem cell (iPSC) technology and organ-on-a-chip technology to model more complex organ-like tissues or whole model organisms (108). In this way, it is possible to identify and validate causal relationships in a model that better represents the complex mechanisms of interactions and signaling of the human system (109,110). Furthermore, when combined with single cell measurements it allows for the mapping of causal networks in a tissue and cell specific manner (107,111). The major limiting factor with these perturbations and measurements is that they occur in model systems which may be biased by the experimental technique. Fortunately, combined with evidence observed from MR experiments, it does become possible to perform triangulation of evidence, as it allows for the identification of robust causal relationships.

In summary, there is no golden standard protocol for being able to claim a causal relationship from observational or experimental data. I foresee that in the future experimental and statistical methods would have to be used jointly to robustly claim a causal relationship in a scientific discourse. For instance, a new putative causal relationship can be found in an organ on

a chip system, and validated using MR or vice versa.

The difference between causal variants and causal relationships between traits

Even though MR has been used to identify causal relationships with -omics traits (86), its application has been overshadowed by investigators using colocalization methods (10,112). This may perhaps be due to the fact that there are not many methods that can perform MR on -omics traits, but is also due to methodological difficulties. One main issue explaining the relative under-utilization of MR is that specific MR methods have mostly been designed for complex traits with multiple loci. In contrast to complex traits, -omics traits like methylation, expression and protein levels are often only regulated by genetic variants *in cis*, i.e., very close to the DNA feature itself, with variants located in other loci having none or very minor effects.

When there is only a single locus, the IV-III assumption becomes difficult to account for. MR methods usually use multiple variants at different loci to test for the presence of pleiotropy and correct for it. Therefore, application of MR methods to -omics traits with one locus or only a handful of loci cannot lead to robust findings and thus MR is often discarded in favour of colocalization. While colocalization requires only a single locus, it is often ignored that colocalization does not identify a causal effect - only that these two traits share the same causal variants - and the results of colocalization cannot distinguish between causality and pleiotropy.

In [chapter 7](#), we found that we can increase the number of IVs available in a single locus by performing conditional stepwise IV selection, making it possible to use pleiotropy-robust methods. Furthermore, when individual level data is available, our newly proposed MR-link method is able to correct for most kinds of pleiotropy, even with only a single causal variant (61).

The MR-link method in [chapter 7](#) was motivated by the analysis in [chapter 6](#), where we were unable to perform robust causal inference on many genes in celiac disease loci (62) due to a lack of genetic variants used as IVs. We were further motivated to try and improve causal inference methods by the fact that colocalization, LD overlap and DEPICT provided discordant results when determining which genes were causal, making it difficult to identify truly causal genes.

Comparing colocalization to MR, I find it appealing that MR is able to identify a causal relationship between two traits, whereas colocalization tests the hypothesis that the causal variants are shared (10,112). Looking at it from a functional perspective, on the one hand, causal variants can be perturbed by CRISPR-type experiments (113). But causal variant perturbation can have many off-target effects: CRISPR perturbation of one variant close to *GFI1B* has been shown to affect up to 569 (!) other expression phenotypes (114). Compared to a causal relationship that is found in MR, the downstream consequences are more limited. A DNA variant will have the same or more edges in a network as a gene (**Figure**

Figure 1

Potential off-target effects are larger when perturbing a genetic variant compared to a gene expression trait

Considering all the potential causal links an exposure has, the causal network of a SNP (large box) is larger, compared to the network of an expression trait and a disease (with the same causal SNP) (small rectangle). This is important if we were to perturb the causal variant, this would have more off-target effects, compared to perturbing the gene expression of interest as the network is smaller.

1), making it easier to understand the effects of the -omics phenotype, and reducing potential off-target effects because the network the causal gene exists in is a subset of the network of that causal SNP.

Shifting from single causal relationships towards causal networks

Identifying single causal relationships between two traits is an essential and important task for medicine because perturbation of the exposure through a drug or some other treatment will also subsequently affect the outcome of interest. But, in the same way that we looked at the difference between causal genetic variants and causal relationships between phenotypes, it's important to keep in mind the complex network of signaling and feedback cascades in which a single causal relationship is located.

Let's imagine that the whole human organism is a black box of physiological mechanisms (**Figure 2A**). Finding a causal relationship between two factors, e.g. through a treatment-control experiment or through other methods like MR, will shed some light into the black box, namely that changing one factor influences another factor (**Figure 2B**). But how this works and what kind of other processes are involved is not elucidated by the causal relationship identified (**Figure 2B**). The arrow of the causal relationship will, in itself, be a black box with an unknown function (**Figure 2C**). Taking the role of a pharmaceutical company developing a new drug, we would of course be interested in the effect of a new drug, identified by the causal effect of interest. At the same time, we would also be interested in potential side effects of the drug that is under investigation. Understanding the full network of causal effects could potentially identify side effects of drugs.

Studies are already trying to integrate multiple causal effects into large databases of networks (115). These networks will initially be used to further our understanding of human biology, but they can also be used to identify

new perturbation targets in human disease and their side effects. These networks already exist for the metabolic pathways of small molecules in humans like the Krebs cycle (116,117), and have been instrumental in developing statin treatment, as statins target the rate limiting step of cholesterol synthesis (118,119). If we can extend these pathways of small molecules to integration of proteins, we may be in a better position to identify how human biology works and why certain side effects occur after treatments.

Interestingly, a network view of causal relationships can also be used to validate new causal relationships in the form of a gold standard. If we are to identify a causal relationship between a drug target and a disease, having knowledge of the network in which the drug acts could help identify potential side effects. Furthermore, it is likely that comorbidities of the disease would also be present in the causal network that leads to disease risk.

A holistic view of human biological knowledge

Reaching a full mechanistic network view of human causal relationships is not trivial and would require an integrated understanding of how all molecules in human cells work together across all different tissues. As we move towards this level of understanding in human biology, it is important to have a holistic view of our knowledge now. Let's consider the things we already know about human biology, some unknown factors that we know are unknown and then speculate on how we should deal with unknowns that we have no idea about (110). Using this knowledge, I propose some strategies in which we can build networks while still being mindful of the known and unknown complexity of human biology that we will not be able to identify.

To start with what we know, there is a lot we can already measure about

Figure 2

The causal networks in human biology are largely a black box. Most causal relationships that we find are black boxes in themselves

(A) If we consider all the factors measured in human biology (colored circles) and the causal relationships between them, this is largely a black box. (B) If a causal relationship between factor A and factor B is found, this means that factor A influences factor B. However, how this happens, i.e. what happens inside the causal relationships is often still a black box in itself. (C) There could be another network of interactions inside the black box that we don't know about.

human biology. We can measure the full DNA sequence of an individual (120). We are able to measure the full epigenome and transcriptome with high fidelity in bulk tissue and single cells. Furthermore, we have an almost full understanding of all the metabolic pathways of small molecules. Using questionnaire data and health records from cross-sectional and prospective cohorts, we have a fairly good understanding of human health and how it develops over time.

This is not to say that we know everything. We are still unable to measure many things with sufficient accuracy. For instance, it remains difficult to measure all the proteins and lipids in a cell or in bulk tissue, although some proteins and lipids are measurable in a bulk and single-cell context. Zooming out a little, we could argue that we need to know the interactions between cells and between tissues. But measuring these may be very difficult and has not been done in a transcriptome- or proteome-wide manner. Furthermore, even the questionnaire data and disease diagnoses used in GWAS can still be inaccurate. For instance, some individuals would show more severe symptoms of a disease than others, but in a case-control GWAS analysis, they would often be coded and analyzed as the same because severity classification remains difficult.

Finally, it's important to realize that a lot is still unknown about human biology. There are likely a lot of molecules and interactions that we have no idea about, and having this knowledge may shift the paradigm of human biology. Speculation about the future will almost certainly be wrong. However, I believe that knowing more about post-translational modifications of proteins and the electrochemistry inside a cell may well become very important in the years to come.

Improving our knowledge of human biology with what we have now

To improve our understanding of human biology, it remains vital that large observational studies continue to be generated and extended using the latest measurement technologies. I strongly recommend that these studies be coupled with genetic data, which makes MR possible for causal inference. Secondly, the main outcome is, of course, human disease. If human disease is measured as a phenotype, I believe this should be accompanied by as many disease-relevant covariates or endophenotypes as possible. We know that most human diseases are complex in nature, so being able to sub-group diseases remains important.

So how will we be able to identify mechanistic networks, when it is very likely that we are able to measure only a handful of components of said network? In short, we will not be able to solve the full extent of human biology with currently available measurement techniques, but we can construct our networks in such a way that we compartmentalize what we already know. I suggest an index phenotype-first approach. The endpoint in this analysis should be unambiguous and relevant, for instance, clinical phenotypes like disease definitions combined with from blood or urine laboratory measurements. These phenotypes are the most important points of our networks. First off, these phenotypes are often

chapter 8

measurable with high accuracy in a wide variety of labs around the globe, making their measurement unambiguous. Secondly, these phenotypes are relevant for human health, which ensures that they will remain of interest in the future.

As enough -omics data becomes available to perform causal inference analyses such as MR, my suggestion is to first identify a causal relationship between the -omics exposure and a well-defined index trait such as disease status or clinical phenotype. Upon finding a causal relationship (**Figure 3A**), we should slowly build a network around it with other known causal relationships with the index trait (**Figure 3B**). As enough of these causal relationships become known, we can then try and join these causal relationships (**Figure 3**). This approach has a few advantages. Firstly, the multiple testing burden is limited, while still constructing causal networks that are relevant to human biology. Secondly, it limits the amount of spurious causal inference due to potential sources of bias in the molecular measurements. By incorporating new measurements in this way, we don't make assumptions about any networks that are in human biology.

Next to a (clinical) index phenotype networking approach, it remains interesting to consider the causal network within a molecular layer and model it with the factors causal to our index phenotype. For instance through correlation networks and Bayesian networking (121–123). Or using specific clustering algorithms to identify specific modules in a gene regulatory network. In this way we could further identify the underlying network that resides inside a layer of molecular information (124).

The future of observational causal inference

It is very difficult to predict what will happen in the future, but as far as I can tell, the identification of causal relationships will remain an important task for human biological research in the years to come. Using techniques such as MR will help in the identification of causal links between two human traits from observational data, helping us on the way to building a better mechanistic understanding of human causal networks. I further believe that with the advent and wide application of observational causal inference we will be able to identify, for the first time, how complex human biology really is. Discovering new human biology has always been fun, but the promise of finding new insights from the integration of many different layers of data is something to look forward to.

Figure 3

A two-step approach to building causal networks

(A) Here, three layers of molecular phenotypes are presented, protein levels, gene expression, and DNA methylation. We compare each individual molecular phenotype with a well defined index, and identify three causal relationships (arrows) across the three layers. (B) Network building step. The three causal relationships that were identified in the first step are then carefully combined to identify putative causal networks between each molecular mechanism and the index trait

References

- Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S, et al. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* [Internet]. 2020 Mar 5 [cited 2020 Apr 6]; Available from: <http://www.sciencedirect.com/science/article/pii/S0092867420302294>
- The COVID-19 Host Genetics Initiative, Ganna A. Mapping the human genetic architecture of COVID-19 by worldwide meta-analysis. *medRxiv*. 2021 May 8;2021.03.10.21252820.
- Lopera Maya EA, van der Graaf A, Lanting P, van der Geest M, Fu J, Swertz M, et al. Lack of Association Between Genetic Variants at ACE2 and TMPRSS2 Genes Involved in SARS-CoV-2 Infection and Human Quantitative Phenotypes. *Front Genet* [Internet]. 2020 [cited 2020 Nov 27];11. Available from: <https://www.frontiersin.org/articles/10.3389/fgene.2020.00613/full>
- Ricaño-Ponce I, Gutierrez-Achury J, Costa AF, Deelen P, Kurilshikov A, Zorro MM, et al. Immunochip meta-analysis in European and Argentinian populations identifies two novel genetic loci associated with celiac disease. *European Journal of Human Genetics*. 2020 Mar;28(3):313–23.
- Trynka G, Hunt KA, Bockett NA, Romanos J, Mistry V, Szperl A, et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat Genet*. 2011 Dec;43(12):1193–201.
- Bonder MJ, Kurilshikov A, Tigchelaar EF, Mujagic Z, Imhann F, Vila AV, et al. The effect of host genetics on the gut microbiome. *Nat Genet*. 2016;48(11):1407–12.
- Goodrich JK, Davenport ER, Beaumont M, Jackson MA, Knight R, Ober C, et al. Genetic Determinants of the Gut Microbiome in UK Twins. *Cell Host Microbe*. 2016 May 11;19(5):731–43.
- Blekhman R, Goodrich JK, Huang K, Sun Q, Bukowski R, Bell JT, et al. Host genetic variation impacts microbiome composition across human body sites. *Genome Biol*. 2015 Sep 15;16:191.
- Zhernakova DV, Deelen P, Vermaat M, Ilterson M van, Galen M van, Arindartoro W, et al. Identification of context-dependent expression quantitative trait loci in whole blood. *Nature Genetics*. 2017 Jan;49(1):139–45.
- Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLOS Genetics*. 2014 May 15;10(5):e1004383.
- Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, Tsunoda T, et al. Functional SNPs in the lymphotoxin- α gene that are associated with susceptibility to myocardial infarction. *Nature Genetics*. 2002 Dec;32(4):650–4.
- Klein RJ, Zeiss C, Chew EY, Tsai J-Y, Sackler RS, Haynes C, et al. Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science*. 2005 Apr 15;308(5720):385–9.
- Klarin D, Damrauer SM, Cho K, Sun YV, Teslovich TM, Honerlaw J, et al. Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. *Nature Genetics*. 2018 Nov 1;50(11):1514–23.
- Sinnott-Armstrong N, Tanigawa Y, Amar D, Mars N, Benner C, Aguirre M, et al. Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nature Genetics*. 2021 Feb;53(2):185–94.
- Gaziano JM, Concato J, Brophy M, Fiore L, Pyarajan S, Breeling J, et al. Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *Journal of Clinical Epidemiology*. 2016 Feb 1;70:214–23.
- Intyre KM, Lanting P, Deelen P, Wiersma HH, Vonk JM, Ori APS, et al. Lifelines COVID-19 cohort: investigating COVID-19 infection and its health and societal impacts in a Dutch population-based cohort. *BMJ Open*. 2021 Mar 1;11(3):e044474.
- Bentley AR, Callier S, Rotimi CN. Diversity and inclusion in genomic research: why the uneven progress? *J Community Genet*. 2017 Oct 1;8(4):255–66.
- Sirugo G, Williams SM, Tishkoff SA. The Missing Diversity in Human Genetic Studies. *Cell*. 2019 Mar 21;177(1):26–31.

19. Duncan L, Shen H, Gelaye B, Meijsen J, Ressler K, Feldman M, et al. Analysis of polygenic risk score usage and performance in diverse human populations. *Nature Communications*. 2019 Jul 25;10(1):3328.
20. Kichaev G, Pasaniuc B. Leveraging Functional-Annotation Data in Trans-ethnic Fine-Mapping Studies. *Am J Hum Genet*. 2015 Aug 6;97(2):260–71.
21. Weissbrod O, Kanai M, Shi H, Gazal S, Peyrot W, Khera A, et al. Leveraging fine-mapping and non-European training data to improve trans-ethnic polygenic risk scores. *medRxiv*. 2021 Jan 20;2021.01.19.21249483.
22. Asimit JL, Hatzikotoulas K, McCarthy M, Morris AP, Zeggini E. Trans-ethnic study design approaches for fine-mapping. *European Journal of Human Genetics*. 2016 Sep;24(9):1330–6.
23. Li YR, Keating BJ. Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations. *Genome Medicine*. 2014 Oct 31;6(10):91.
24. Wonkam A. Sequence three million genomes across Africa. *Nature*. 2021 Feb;590(7845):209–11.
25. Flannick J, Mercader JM, Fuchsberger C, Udler MS, Mahajan A, Wessel J, et al. Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. *Nature*. 2019 Jun;570(7759):71–6.
26. Huglund J, Rafati N, Rask-Andersen M, Enroth S, Karlsson T, Ek WE, et al. Improved power and precision with whole genome sequencing data in genome-wide association studies of inflammatory biomarkers. *Scientific Reports*. 2019 Nov 14;9(1):16844.
27. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfuldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020 May;581(7809):434–43.
28. Liu DJ, Peloso GM, Yu H, Butterworth AS, Wang X, Mahajan A, et al. Exome-wide association study of plasma lipids in >300,000 individuals. *Nature Genetics*. 2017 Dec;49(12):1758–66.
29. Bomba L, Walter K, Soranzo N. The impact of rare and low-frequency genetic variants in common disease. *Genome Biology*. 2017 Apr 27;18(1):77.
30. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016 Aug;536(7616):285–91.
31. Saleheen D, Natarajan P, Armean IM, Zhao W, Rasheed A, Khetarpal SA, et al. Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. *Nature*. 2017 Apr;544(7649):235–9.
32. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*. 2015 Feb 12;518(7538):197–206.
33. Locke AE, Steinberg KM, Chiang CWK, Service SK, Havulinna AS, Stell L, et al. Exome sequencing of Finnish isolates enhances rare-variant association power. *Nature*. 2019 Aug;572(7769):323–8.
34. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine*. 2015 Mar 31;12(3):e1001779.
35. The road ahead in genetics and genomics | *Nature Reviews Genetics* [Internet]. [cited 2021 May 25]. Available from: <https://www.nature.com/articles/s41576-020-0272-6>
36. Martin AR, Atkinson EG, Chapman SB, Stevenson A, Stroud RE, Abebe T, et al. Low-coverage sequencing cost-effectively detects known and novel variation in underrepresented populations. *The American Journal of Human Genetics*. 2021 Apr 1;108(4):656–68.
37. Booth MJ, Branco MR, Ficiz G, Oxley D, Krueger F, Reik W, et al. Quantitative Sequencing of 5-Methylcytosine and 5-Hydroxymethylcytosine at Single-Base Resolution. *Science*. 2012 May 18;336(6083):934–7.
38. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Meth*. 2013 Dec;10(12):1213–8.
39. The 1000 Genomes Consortium. A global reference for human genetic variation.

- Nature. 2015 Oct;526(7571):68–74.
40. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*. 2008 Jul;5(7):621–8.
  41. Tessler M, Neumann JS, Afshinnkoo E, Pineda M, Hersch R, Velho LFM, et al. Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing. *Scientific Reports*. 2017 Jul 31;7(1):6589.
  42. Bonder MJ, Luijk R, Zhernakova DV, Moed M, Deelen P, Vermaat M, et al. Disease variants alter transcription factor levels and methylation of their binding sites. *Nature Genetics*. 2017 Jan;49(1):131–8.
  43. Sun BB, Maranville JC, Peters JE, Stacey D, Staley JR, Blackshaw J, et al. Genomic atlas of the human plasma proteome. *Nature*. 2018 Jun;558(7708):73–9.
  44. Zheng J, Haberland V, Baird D, Walker V, Haycock PC, Hurler MR, et al. Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. *Nature Genetics*. 2020 Oct;52(10):1122–31.
  45. Zhernakova DV, Deelen P, Vermaat M, Ijzerman M van, Galen M van, Arindrarto W, et al. Identification of context-dependent expression quantitative trait loci in whole blood. *Nature Genetics*. 2017 Jan;49(1):139–45.
  46. Kurilshikov A, Medina-Gomez C, Bacigalupe R, Radjabzadeh D, Wang J, Demirkan A, et al. Large-scale association analyses identify host factors influencing human gut microbiome composition. *Nature Genetics*. 2021 Feb;53(2):156–65.
  47. Lopera-Maya EA, Kurilshikov A, Graaf A van der, Hu S, Andreu-Sanchez S, Chen L, et al. Effect of host genetics on the gut microbiome in 7,738 participants of the Dutch Microbiome Project. *bioRxiv*. 2020 Dec 9;2020.12.09.417642.
  48. Li YI, Geijn B van de, Raj A, Knowles DA, Petti AA, Golan D, et al. RNA splicing is a primary link between genetic variation and disease. *Science*. 2016 Apr 29;352(6285):600–4.
  49. Evangelou E, Ioannidis JPA. Meta-analysis methods for genome-wide association studies and beyond. *Nature Reviews Genetics*. 2013 Jun;14(6):379–89.
  50. Võsa U, Claringbould A, Westra H-J, Bonder MJ, Deelen P, Zeng B, et al. Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis. *bioRxiv*. 2018 Oct 19;447367.
  51. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature*. 2017 Oct;550(7675):204–13.
  52. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*. 2020 Sep 11;369(6509):1318–30.
  53. Ongen H, Brown AA, Delaneau O, Panousis NI, Nica AC, GTEx Consortium, et al. Estimating the causal tissues for complex traits and diseases. *Nature Genetics*. 2017 Dec;49(12):1676–83.
  54. Ruigrok RAAA, Collij V, Sureda P, Klaassen MAY, Bolte LA, Jansen BH, et al. The Composition and Metabolic Potential of the Human Small Intestinal Microbiota Within the Context of Inflammatory Bowel Disease. *Journal of Crohn's and Colitis [Internet]*. 2021 Jan 30 [cited 2021 Jun 2];(jjab020). Available from: <https://doi.org/10.1093/ecco-jcc/jjab020>
  55. Zmora N, Zilberman-Schapira G, Suez J, Mor U, Dori-Bachash M, Bashirdes S, et al. Personalized Gut Mucosal Colonization Resistance to Empiric Probiotics Is Associated with Unique Host and Microbiome Features. *Cell*. 2018 Sep 6;174(6):1388–1405.e21.
  56. Bonder MJ, Smail C, Gloudemans MJ, Frisard L, Jakubosky D, D'Antonio M, et al. Identification of rare and common regulatory variants in pluripotent cells using population-scale transcriptomics. *Nature Genetics*. 2021 Mar;53(3):313–21.
  57. Cuomo ASE, Seaton DD, McCarthy DJ, Martinez I, Bonder MJ, Garcia-Bernardo J, et al. Single-cell RNA-sequencing of differentiating iPSCs reveals dynamic genetic effects on gene expression. *Nature Communications*. 2020 Feb 10;11(1):810.
  58. van der Wijst MGP, Brugge H, de Vries DH, Deelen P, Swertz MA, Franke L. Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nature Genetics*. 2018 Apr;50(4):493–7.
  59. van der Wijst M, de Vries D, Groot H, Trynka G, Hon C, Bonder M, et al. The single-cell eQTLGen consortium. *eLife*. 2020 Mar 9;9:e52155.

60. Eales JM, Jiang X, Xu X, Saluja S, Akbarov A, Cano-Gamez E, et al. Uncovering genetic mechanisms of hypertension through multi-omic analysis of the kidney. *Nature Genetics*. 2021 May;53(5):630–7.
61. van der Graaf A, Claringbould A, Rimbart A, Westra H-J, Li Y, Wijmenga C, et al. Mendelian randomization while jointly modeling cis genetics identifies causal relationships between gene expression and lipids. *Nature Communications*. 2020 Oct 1;11(1):4930.
62. van der Graaf, A., Zorro, M.M., Claringbould, A., Vösa, U., Aguirre-Gamboa, R., Li, C., Mooiweer, J., Ricaño-Ponce, I., Borek, Z., Koning, F., et al. (2021). Systematic Prioritization of Candidate Genes in Disease Loci Identifies TRAFD1 as a Master Regulator of IFN γ Signaling in Celiac Disease. *Frontiers in Genetics* 11, 1780.
63. Rader DJ, Hovingh GK. HDL and cardiovascular disease. *The Lancet*. 2014 Aug 16;384(9943):618–25.
64. Voight BF, Peloso GM, Orho-Melander M, Frikke-Schmidt R, Barbalic M, Jensen MK, et al. Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. *Lancet*. 2012 Aug 11;380(9841):572–80.
65. Kawashiri M, Tada H, Nomura A, Yamagishi M. Mendelian randomization: Its impact on cardiovascular disease. *Journal of Cardiology*. 2018 Oct 1;72(4):307–13.
66. Valdes-Marquez E, Parish S, Clarke R, Stari T, Worrall BB, METASTROKE Consortium of the ISGC, et al. Relative effects of LDL-C on ischemic stroke and coronary disease: A Mendelian randomization study. *Neurology*. 2019 Mar 12;92(11):e1176–87.
67. Morrison J, Knoblauch N, Marcus JH, Stephens M, He X. Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. *Nature Genetics*. 2020 Jul;52(7):740–7.
68. Wainberg M, Mahajan A, Kundaje A, McCarthy MI, Ingelsson E, Sinnott-Armstrong N, et al. Homogeneity in the association of body mass index with type 2 diabetes across the UK Biobank: A Mendelian randomization study. *PLOS Medicine*. 2019 Dec 10;16(12):e1002982.
69. Porcu E, Sjaarda J, Lepik K, Carmeli C, Darrous L, Sulc J, et al. Causal Inference Methods to Integrate Omics and Complex Traits. *Cold Spring Harb Perspect Med*. 2020 Aug 17;a040493.
70. Bowden J, Del Greco M F, Minelli C, Smith GD, Sheehan N, Thompson J. A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Statistics in Medicine*. 2017;36(11):1783–802.
71. Burgess S, Davey Smith G, Davies NM, Dudbridge F, Gill D, Glymour MM, et al. Guidelines for performing Mendelian randomization investigations. *Wellcome Open Res*. 2020 Apr 28;4:186.
72. Haycock PC, Burgess S, Wade KH, Bowden J, Relton C, Smith GD. Best (but oft-forgotten) practices: the design, analysis, and interpretation of Mendelian randomization studies. *The American Journal of Clinical Nutrition*. 2016 Apr;103(4):965.
73. Sanderson E, Windmeijer F. A weak instrument F-test in linear IV models with multiple endogenous variables. *Journal of Econometrics*. 2016 Feb 1;190(2):212–21.
74. Burgess S, Thompson SG, CRP CHD Genetics Collaboration. Avoiding bias from weak instruments in Mendelian randomization studies. *International Journal of Epidemiology*. 2011 Jun 1;40(3):755–64.
75. Haworth S, Mitchell R, Corbin L, Wade KH, Dudding T, Budu-Aggrey A, et al. Apparent latent structure within the UK Biobank sample has implications for epidemiological analysis. *Nature Communications*. 2019 Jan 18;10(1):333.
76. Hellwege J, Keaton J, Giri A, Gao X, Velez Edwards DR, Edwards TL. Population Stratification in Genetic Association Studies. *Curr Protoc Hum Genet*. 2017 Oct 18;95:1.22.1-1.22.23.
77. Brumpton B, Sanderson E, Heilbron K, Hartwig FP, Harrison S, Vie GE, et al. Avoiding dynastic, assortative mating, and population stratification biases in Mendelian randomization through within-family analyses. *Nature Communications*. 2020 Jul 14;11(1):3519.
78. Davies NM, Howe LJ, Brumpton B, Havdahl A, Evans DM, Davey Smith G. Within family Mendelian randomization studies. *Human Molecular Genetics*.

2019 Nov 21;28(R2):R170–9.

79. Belsky DW, Domingue BW, Wedow R, Arseneault L, Boardman JD, Caspi A, et al. Genetic analysis of social-class mobility in five longitudinal studies. *PNAS*. 2018 Jul 31;115(31):E7275–84.
80. Scholtens S, Smidt N, Swertz MA, Bakker SJ, Dotinga A, Vonk JM, et al. Cohort Profile: LifeLines, a three-generation cohort study and biobank. *Int J Epidemiol*. 2015 Aug 1;44(4):1172–80.
81. Paternoster L, Tilling K, Smith GD. Genetic epidemiology and Mendelian randomization for informing disease therapeutics: Conceptual and methodological challenges. *PLOS Genetics*. 2017 Oct 5;13(10):e1006944.
82. Barry C, Liu J, Richmond R, Rutter MK, Lawlor DA, Dudbridge F, et al. Exploiting collider bias to apply two-sample summary data Mendelian randomization methods to one-sample individual level data. *medRxiv*. 2020 Oct 23;2020.10.20.20216358.
83. Verbanck M, Chen C-Y, Neale B, Do R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nature Genetics*. 2018 May;50(5):693–8.
84. Berzuini C, Guo H, Burgess S, Bernardinelli L. A Bayesian approach to Mendelian randomization with multiple pleiotropic variants. *Biostatistics* [Internet]. [cited 2019 Oct 29]; Available from: <https://academic.oup.com/biostatistics/advance-article/doi/10.1093/biostatistics/kxy027/5063519>
85. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol*. 2015 Apr 1;44(2):512–25.
86. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature Genetics*. 2016 May;48(5):481–7.
87. Bowden J, Hemani G, Davey Smith G. Invited Commentary: Detecting Individual and Global Horizontal Pleiotropy in Mendelian Randomization—A Job for the Humble Heterogeneity Statistic? *American Journal of Epidemiology*. 2018 Dec 1;187(12):2681–5.
88. Burgess S, Thompson SG. Multivariable Mendelian Randomization: The Use of Pleiotropic Genetic Variants to Estimate Causal Effects. *Am J Epidemiol*. 2015 Feb 15;181(4):251–60.
89. Porcu E, Rieger S, Lepik K, Santoni FA, Reymond A, Kutalik Z. Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. *Nature Communications*. 2019 Jul 24;10(1):3300.
90. Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, Koseki M, et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*. 2010 Aug;466(7307):707–13.
91. Gallagher MD, Chen-Plotkin AS. The Post-GWAS Era: From Association to Function. *The American Journal of Human Genetics*. 2018 May 3;102(5):717–30.
92. Westra H-J, Martínez-Bonet M, Onengut-Gumuscu S, Lee A, Luo Y, Teslovich N, et al. Fine-mapping and functional studies highlight potential causal variants for rheumatoid arthritis and type 1 diabetes. *Nat Genet*. 2018 Oct;50(10):1366–74.
93. Fang H, Wolf HD, Knezevic B, Burnham KL, Osgood J, Sanniti A, et al. A genetics-led approach defines the drug target landscape of 30 immune-related traits. *Nature Genetics*. 2019 Jul;51(7):1082–91.
94. Ghousaini M, Mountjoy E, Carmona M, Peat G, Schmidt EM, Hercules A, et al. Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Research*. 2021 Jan 8;49(D1):D1311–20.
95. Wang G, Sarkar A, Carbonetto P, Stephens M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2020;82(5):1273–300.
96. Wallace C. Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. *PLOS Genetics*. 2020 Apr 20;16(4):e1008720.
97. Wiebe JM, Bruce RF, O'Hara TP. Development and use of a gold-standard data set for subjectivity classifications. In: *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* [Internet]. USA: Association for Computational Linguistics;

- 1999 [cited 2021 Jun 7]. p. 246–53. (ACL '99). Available from: <https://doi.org/10.3115/1034678.1034721>
98. Lawlor DA, Tilling K, Davey Smith G. Triangulation in aetiological epidemiology. *International Journal of Epidemiology*. 2016 Dec 1;45(6):1866–86.
99. Munafr MR, Smith GD. Robust research needs many lines of evidence. *Nature*. 2018 Jan;553(7689):399–401.
100. Lipsitch M, Tchetgen ET, Cohen T. Negative Controls: A Tool for Detecting Confounding and Bias in Observational Studies. *Epidemiology*. 2010 May;21(3):383–8.
101. Flanders WD, Strickland MJ, Klein M. A New Method for Partial Correction of Residual Confounding in Time-Series and Other Observational Studies. *American Journal of Epidemiology*. 2017 May 15;185(10):941–9.
102. Breinig M, Schweitzer AY, Herianto AM, Revia S, Schaefer L, Wendler L, et al. Multiplexed orthogonal genome editing and transcriptional activation by Cas12a. *Nat Methods*. 2019 Jan;16(1):51–4.
103. Larson MH, Gilbert LA, Wang X, Lim WA, Weissman JS, Qi LS. CRISPR interference (CRISPRi) for sequence-specific control of gene expression. *Nat Protoc*. 2013 Nov;8(11):2180–96.
104. Freimer JW, Shaked O, Naqvi S, Sinnott-Armstrong N, Kathiria A, Chen AF, et al. Systematic discovery and perturbation of regulatory genes in human T cells reveals the architecture of immune networks. *bioRxiv*. 2021 Apr 19;2021.04.18.440363.
105. Gilbert LA, Larson MH, Morsut L, Liu Z, Brar GA, Torres SE, et al. CRISPR-Mediated Modular RNA-Guided Regulation of Transcription in Eukaryotes. *Cell*. 2013 Jul 18;154(2):442–51.
106. Broekema RV, Bakker OB, Jonkers IH. A practical view of fine-mapping and gene prioritization in the post-genome-wide association era. *Open Biology*. 10(1):190221.
107. Rubin AJ, Parker KR, Satpathy AT, Qi Y, Wu B, Ong AJ, et al. Coupled Single-Cell CRISPR Screening and Epigenomic Profiling Reveals Causal Gene Regulatory Networks. *Cell*. 2019 Jan 10;176(1):361–376.e17.
108. Mandegar MA, Huebsch N, Frolov EB, Shin E, Truong A, Olvera MP, et al. CRISPR Interference Efficiently Induces Specific and Reversible Gene Silencing in Human iPSCs. *Cell Stem Cell*. 2016 Apr 7;18(4):541–53.
109. Moerkens R, Mooiweer J, Withoff S, Wijmenga C. Celiac disease-on-chip: Modeling a multifactorial disease in vitro. *United European Gastroenterology Journal*. 2019 May 1;7(4):467–76.
110. Park SE, Georgescu A, Huh D. Organoids-on-a-chip. *Science*. 2019 Jun 7;364(6444):960–5.
111. Gasperini M, Hill AJ, McFaline-Figueroa JL, Martin B, Kim S, Zhang MD, et al. A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell*. 2019 Jan 10;176(1):377–390.e19.
112. Benner C, Spencer CCA, Havulinna AS, Salomaa V, Ripatti S, Pirinen M. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*. 2016 May 15;32(10):1493–501.
113. Pickar-Oliver A, Gersbach CA. The next generation of CRISPR–Cas technologies and applications. *Nature Reviews Molecular Cell Biology*. 2019 Aug;20(8):490–507.
114. Morris JA, Daniloski Z, Domingo J, Barry T, Ziosi M, Glinos DA, et al. Discovery of target genes and pathways of blood trait loci using pooled CRISPR screens and single cell RNA sequencing. *bioRxiv*. 2021 Apr 8;2021.04.07.438882.
115. Liu Y, Elsworth B, Erola P, Haberland V, Hemani G, Lyon M, et al. EpiGraphDB: a database and data mining platform for health data science. *Bioinformatics* [Internet]. 2020 Nov 9 [cited 2021 May 29];(btaa961). Available from: <https://doi.org/10.1093/bioinformatics/btaa961>
116. Krebs HA, Johnson WA. Metabolism of ketonic acids in animal tissues. *Biochemical Journal*. 1937 Apr 1;31(4):645–60.
117. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res*. 2012 Jan;40(Database issue):D1100–7.
118. Adhyaru BB, Jacobson TA. Safety and efficacy of statin therapy. *Nat Rev Cardiol*. 2018 Dec;15(12):757–69.

119. Tobert JA. Lovastatin and beyond: the history of the HMG-CoA reductase inhibitors. *Nature Reviews Drug Discovery*. 2003 Jul;2(7):517–26.
120. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. *bioRxiv*. 2021 May 27;2021.05.26.445798.
121. Pers TH, Karjalainen JM, Chan Y, Westra H-J, Wood AR, Yang J, et al. Biological interpretation of genome-wide association studies using predicted gene functions. *Nature Communications*. 2015 Jan 19;6(1):5890.
122. Deelen P, van Dam S, Herkert JC, Karjalainen JM, Brugge H, Abbott KM, et al. Improving the diagnostic yield of exome-sequencing by predicting gene–phenotype associations using large-scale gene expression analysis. *Nat Commun*. 2019 Jun 28;10(1):2837.
123. Werhli AV, Grzegorzczak M, Husmeier D. Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. *Bioinformatics*. 2006 Oct 15;22(20):2523–31.
124. van Dam S, Vosa U, van der Graaf A, Franke L, de Magalhães JP. Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in Bioinformatics*. 2018 Jul 20;19(4):575–92.

chapter 9

English summary

Some human traits like disease are heritable, which means that they run in families. This indicates that there must be something on the DNA that affects an individual's susceptibility to developing a trait. In the last 15 years, scientists from around the world have been very successful in mapping the locations on the DNA that are associated to traits like disease, finding thousands of loci, and hundreds of DNA locations per trait, making them truly complex traits. So, we have a very good understanding about which locations on the DNA are important for developing complex traits like disease. Unfortunately, it's still unclear how these locations on the DNA affect an individual's trait. Indeed, even though we know which parts of the DNA are important for susceptibility, we have hardly any idea about how these parts of the DNA cause a disease or affect other traits.

In this thesis my colleagues and I investigate how we can best understand the DNA locations that affect trait susceptibility and in doing so, identify the causes for human traits like disease. Our main approach relies on the notion that DNA locations should be shared between the trait and its cause(s). One important technique that we have used to test for finding these causal relationships is called Mendelian randomization. Mendelian randomization can find causes to disease without having to rely on experimental data. It does so by identifying naturally occurring experiments that have happened in observational data. In principle, Mendelian randomization can conclude the same things from observational data as from an experimental study. So called 'observational causality' has many benefits as it's cheaper than an experiment, and is less burdensome on the subjects because they are not subjected to any intervention. Interestingly, it also allows more hypotheses to be tested: You don't have to develop a drug or intervention for each causal relationship that you want to find.

The causes that I'm interested in are so called '-omics' traits. -omics traits are molecular measurements that are usually strongly regulated by the DNA. This strong DNA regulation makes -omics traits interesting candidates to understand the mechanism behind the genetic loci of other traits.

In this thesis we have investigated gene expression, protein levels and microbiome measurements as our -omics traits of interest, so as to try and achieve the following main aim:

To identify causal relationships between human -omics and complex traits

The thesis is separated into 2 parts. In part one my colleagues and I try to identify the genetic locations that are associated with disease. So that we can identify causes of these and other DNA locations in part two.

Part 1, 'genetic loci for causal inference', starts with chapter 1. Here, the state of the art of current biology and -omics measurements are

introduced, describing how we identify the locations on the DNA that are associated to human traits. Then, the introduction provides a definition on causal relationships in human biology, explaining why we cannot find causal relationships without accounting for confounders, and how these are accounted for in experimental designs like a randomized control trial. Finally, a definition of Mendelian randomization is provided and how it also accounts for all confounders, so it can identify causal relationships. Chapter 2 investigates if there are genetic loci that can be associated to COVID-19 susceptibility. This chapter was written in the early months of the COVID-19 pandemic when not a lot was known about the disease or which individuals were susceptible. We do not find any locations on the DNA that pass our thresholds for significance; however, we do find some interesting results we believe warrant further follow up. In chapter 3 we identify the locations on the DNA that were associated to celiac disease. Celiac disease is an autoimmune disease of the intestine, with a large genetic component. By including an additional 4,125 samples to an already published study of 24,629 individuals, we find an additional 3 novel loci associated to celiac disease.

In the second part, *`causal inference of molecular traits`*, my colleagues and I identify which -omics traits are causal to diseases and other complex traits. Chapter 4 and chapter 5 identify the genetic loci responsible for changes in microbiome composition, and we attempt to identify potential causes for changes in microbiome composition. In chapter 4, we identify the genetic locations that affect microbiome composition in a large meta-analysis of 18,340 individuals, measured across 24 cohorts. Mendelian randomization analyses suggests that changes in microbiome composition may have a causal effect on diseases of the intestine. In chapter 5 we also identify genetic locations that affect the microbiome composition in 7,738 individuals that are measured with the more sensitive metagenomic sequencing method. We reidentify the *ABO* and *LCT* loci as well as 22 other loci with suggestive significance, and perform Mendelian randomization. Based on a power analysis, we estimate that currently available sample sizes are not sufficient to identify the remaining effects for microbiome loci.

In chapter 6 we combined the DNA locations that were found in chapter 3 with gene expression measurements in 3,502 individuals. Using a systematic gene finding approach to identify the most likely causal genes for celiac disease, we identified 118 genes with one or more lines of evidence for being causal to celiac disease. Using a mediation approach to identify the downstream effects of these 118 genes, we find that one gene, *TRAFD1*, mediates 41 other genes across the genome. These genes are then experimentally validated. Providing, for the first time, genetic evidence that IFN γ signaling and MHC I antigen processing/presentation are involved in celiac disease pathophysiology.

We were quite surprised by the variability of the initial gene prioritization findings of chapter 6, so concurrently, we tried to perform causal gene prioritization in a better way in chapter 7. Here we introduce a novel Mendelian randomization method called MR-link that identifies causal relationships between gene expression and complex traits with more

lenient statistical assumptions than previously possible. We applied MR-link to LDL-cholesterol levels and gene expression in whole blood, liver, cerebellum gene expression measurements, as well as in blood protein measurements. Combined we find 25 genes that are causally related to LDL-cholesterol, including the well-known *SORT1*, ApoE, and *ABO*, genes as well as the more obscure *PVRL2* that has experimental evidence related to cellular lipid consumption.

Finally, the findings of this thesis are discussed in [chapter 8](#). The potential issues of the findings are commented on. This chapter also looks into the future of DNA location finding, -omics studies, and causal inference. Finally we discuss how these causal findings can be integrated into human biology using a network approach.

Nederlandse samenvatting

Sommige menselijke eigenschappen, zoals ziekten, zijn erfelijk, wat betekent dat ze in families voorkomen. Dit wijst erop dat er iets in het DNA moet zijn dat van invloed is op de gevoeligheid van een individu voor het ontwikkelen van die eigenschap. De afgelopen 15 jaar hebben wetenschappers uit de hele wereld met veel succes de plaatsen in het DNA in kaart gebracht die in verband worden gebracht met eigenschappen zoals ziekten. Zij hebben duizenden locaties gevonden, met soms honderden DNA-locaties per eigenschap, waardoor het echt complexe eigenschappen zijn. We hebben op dit moment dus een zeer goed inzicht in welke locaties op het DNA belangrijk zijn voor de ontwikkeling van complexe eigenschappen zoals ziekten. Helaas is het nog steeds onduidelijk hoe deze plaatsen op het DNA van invloed zijn op de eigenschap van een individu. Sterker nog, ook al weten we welke delen van het DNA belangrijk zijn voor vatbaarheid, we hebben nauwelijks enig idee over hoe deze delen van het DNA een ziekte veroorzaken of andere eigenschappen beïnvloeden.

In dit proefschrift onderzoeken mijn collega's en ik hoe we het best de DNA-locaties kunnen begrijpen die de gevoeligheid voor eigenschappen beïnvloeden en zo de oorzaken van menselijke eigenschappen zoals ziekten kunnen identificeren. Onze belangrijkste benadering is gebaseerd op het idee dat DNA-locaties gedeeld moeten worden tussen de eigenschap en haar oorzaak (oorzaken). Een belangrijke techniek die wij hebben gebruikt om te testen op het vinden van deze causale verbanden heet Mendeliaanse randomisatie. Mendeliaanse randomisatie kan oorzaken voor ziekten vinden zonder te moeten vertrouwen op experimentele gegevens. Het doet dit door het identificeren van natuurlijk voorkomende experimenten die hebben plaatsgevonden in observationele gegevens. In principe kan Mendeliaanse randomisatie dezelfde conclusies trekken uit observationele gegevens als uit een experimentele studie. Zogenaamde 'observationele causaliteit' heeft vele voordelen: het is goedkoper dan een experiment, en het is minder belastend voor de proefpersonen, aangezien zij niet aan een interventie worden onderworpen. Interessant is dat hiermee ook meer hypothesen kunnen worden getest: je hoeft niet voor elk causaal verband een geneesmiddel of interventie te ontwikkelen.

De oorzaken waarin wij geïnteresseerd zijn, zijn zogenaamde '-omics' kenmerken. -omics kenmerken zijn moleculaire metingen die meestal sterk gereguleerd worden door het DNA. Deze sterke DNA-regulatie maakt -omics kenmerken interessante kandidaten om het mechanisme achter de genetische loci van andere kenmerken te begrijpen.

In dit proefschrift hebben we genexpressie, eiwitniveaus en microbioommetingen onderzocht als onze -omics kenmerken, om te trachten het volgende hoofddoel te bereiken:

Oorzakelijke verbanden tussen menselijke -omics en complexe kenmerken vast te stellen

Het proefschrift is opgesplitst in 2 delen. In deel 1 proberen mijn collega's en ik de genetische locaties te identificeren die geassocieerd zijn met ziekte. Zodat we in deel 2 de oorzaken van deze en andere DNA-locaties kunnen identificeren.

Deel 1, 'genetic loci for causal inference', begint met hoofdstuk 1. Hier wordt de stand van zaken van de huidige biologie en -omics metingen geïntroduceerd, en wordt beschreven hoe we de locaties op het DNA identificeren die geassocieerd zijn met menselijke eigenschappen. Vervolgens geeft de inleiding een definitie van causale verbanden in de menselijke biologie, en legt het uit waarom we geen causale verbanden kunnen vinden zonder rekening te houden met confounders en hoe daarmee rekening wordt gehouden in experimentele ontwerpen zoals een placebo-controle studie. Tenslotte geven we een definitie van Mendeliaanse randomisatie en hoe deze ook rekening houdt met alle confounders, zodat ze causale verbanden kan identificeren.

Hoofdstuk 2 onderzoekt of er genetische locaties zijn die in verband kunnen worden gebracht met de vatbaarheid voor COVID-19. Dit hoofdstuk werd geschreven in de eerste maanden van de COVID-19 pandemie toen er nog niet veel bekend was over de ziekte of welke individuen vatbaar waren. We vinden geen locaties op het DNA die onze drempels voor significantie overschrijden; maar we vinden wel een aantal interessante resultaten waarvan we denken dat ze verdere opvolging rechtvaardigen. In hoofdstuk 3 identificeren we de locaties op het DNA die in verband worden gebracht met coeliakie. Coeliakie is een auto-immuunziekte van de dunne darm, met een grote genetische component. Door 4.125 extra individuen toe te voegen aan een reeds gepubliceerde studie van 24.629 individuen, vinden we nog eens 3 nieuwe loci die geassocieerd zijn met coeliakie.

In het tweede deel, 'causal inference of molecular traits', identificeren mijn collega's en ik welke -omics traits causaal zijn voor ziekten en andere complexe eigenschappen. In hoofdstuk 4 en hoofdstuk 5 identificeren we de genetische loci die verantwoordelijk zijn voor veranderingen in microbioomsamenstelling, en proberen we mogelijke oorzaken voor veranderingen in microbioomsamenstelling te identificeren. In hoofdstuk 4 identificeren we de genetische locaties die de microbioomsamenstelling beïnvloeden in een grote meta-analyse van 18.340 individuen, gemeten over 24 verschillende cohorten. Mendeliaanse randomisatie analyses suggereren dat veranderingen in microbioomsamenstelling een causaal effect kunnen hebben op darmziekten. In hoofdstuk 5 identificeren we ook genetische locaties die van invloed zijn op de microbioomsamenstelling bij 7.738 individuen die gemeten zijn met de gevoeliger metagenomische sequencing methode. We heridentificeren de *ABO* en *LCT* loci en 22 andere loci met suggestieve significantie, en voeren Mendeliaanse randomisatie uit. Op basis van een power-analyse schatten we dat de momenteel beschikbare populatie groottes niet voldoende zijn om de resterende DNA effecten voor het microbioom te identificeren.

In hoofdstuk 6 hebben we de DNA-locaties die gevonden zijn in hoofdstuk 3 gecombineerd met genexpressie metingen bij 3.502 individuen. Met behulp van een systematische gen-vind-aanpak hebben we de meest waarschijnlijke causale genen voor coeliakie geïdentificeerd. We identificeerden 118 genen met 1 of meer lijnen van bewijs voor een oorzakelijk verband met coeliakie. Met behulp van een bemiddelingsaanpak om de downstream effecten van deze 118 genen te identificeren, vonden we dat 1 gen, *TRAFD1*, 41 andere genen over het hele genoom beïnvloed. Deze genen zijn vervolgens experimenteel gevalideerd. Dit levert, voor de eerste keer, genetisch bewijs dat IFN γ signalering en MHC I antigen verwerking en presentatie betrokken zijn bij de pathofysiologie van coeliakie.

We waren nogal verrast door de variabiliteit van de initiële gen-prioritering bevindingen van hoofdstuk 6, dus hebben we tegelijkertijd geprobeerd om causale gen-prioritering op een betere manier uit te voeren in hoofdstuk 7. Hier introduceren we een nieuwe Mendeliaanse randomisatie methode, MR-link genaamd, die causale relaties tussen genexpressie en complexe eigenschappen identificeert met mildere statistische aannames dan voorheen mogelijk was. We hebben MR-link toegepast op LDL-cholesterol niveaus en genexpressie in bloed, lever, cerebellum genexpressie metingen, evenals in bloed eiwit metingen. Gecombineerd vinden we 25 genen die causaal gerelateerd zijn aan LDL-cholesterol, waaronder de bekende *SORT1*, ApoE, en *ABO*, genen alsook het meer obscure *PVRL2* dat experimenteel bewijs heeft in verband met cellulaire lipidenconsumptie.

Tenslotte worden de bevindingen van dit proefschrift besproken in hoofdstuk 8. De mogelijke problemen van de bevindingen worden besproken, en er wordt gekeken naar de toekomst van DNA-locatiebepaling, -omics studies, en het vinden van causaliteit. Tenslotte wordt besproken hoe deze causale bevindingen kunnen worden geïntegreerd in de menselijke biologie met behulp van een netwerkbenadering.

List of Publications

Adriaan van der Graaf, Rene Wardenaar, Drexel A. Neumann, Aaron Taudt, Ruth G. Shaw, Ritsert C. Jansen, Robert J. Schmitz, Maria Colomijn-Tatche, en Frank Johannes. "Rate, spectrum, and evolutionary dynamics of spontaneous epimutations". *Proceedings of the National Academy of Sciences* 112, no 21 (2015): 6676–81.

Sipko van Dam, Urmo Vosa, Adriaan van der Graaf, Lude Franke, en Joao Pedro de Magalhaes. "Gene co-expression analysis for functional classification and gene–disease predictions". *Briefings in bioinformatics* 19, no 4 (2018): 575–92.

Isis Ricaño-Ponce, Javier Gutierrez-Achury, Ana Florencia Costa, Patrick Deelen, Alexander Kurilshikov, Maria Magdalena Zorro, Mathieu Platteel, et al. "Immunochip meta-analysis in European and Argentinian populations identifies two novel genetic loci associated with celiac disease". *European Journal of Human Genetics* 28, no 3 (2020): 313–23.

Adriaan van Der Graaf, Maria M. Zorro, Annique Claringbould, Urmo Vosa, Raul Aguirre-Gamboa, Chan Li, Joram Mooiweer, et al. "Systematic prioritization of candidate genes in disease loci identifies TRAFD1 as a master regulator of IFN γ signaling in celiac disease". *Frontiers in Genetics* 11 (2021): 1780.

Esteban A. Lopera-Maya, Adriaan van der Graaf, Pauline Lanting, Marije van der Geest, Jingyuan Fu, Morris Swertz, et al. "Lack of association between genetic variants at ACE2 and TMPRSS2 genes involved in SARS-CoV-2 infection and human quantitative phenotypes". *Frontiers in genetics* 11 (2020): 613.

Kurilshikov, Alexander, Carolina Medina-Gomez, Rodrigo Bacigalupe, Djawad Radjabzadeh, Jun Wang, Ayse Demirkan, Caroline I. Le Roy, et al. "Large-scale association analyses identify host factors influencing human gut microbiome composition". *Nature Genetics* 53, no 2 (2021): 156–65.

Adriaan van der Graaf, Annique Claringbould, Antoine Rimbart, Harm-Jan Westra, Yang Li, Cisca Wijmenga, en Serena Sanna. "Mendelian randomization while jointly modeling cis genetics identifies causal relationships between gene expression and lipids". *Nature communications* 11, no 1 (2020): 1–12.

Esteban A. Lopera-Maya, Alexander Kurilshikov, Adriaan van der Graaf, Shixian Hu, Sergio Andreu-Sánchez, Lianmin Chen, A. Vich Vila, et al. "Effect of host genetics on the gut microbiome in 7,738 participants of the Dutch Microbiome Project". Accepted in *Nature Genetics*

Serena Sanna, Alexander Kurilshikov, Adriaan van der Graaf, Jingyuan Fu, Alexandra Zhernakova "Challenges and future directions for studying effects of host genetics on the gut microbiome" Accepted in *Nature Genetics*

list of publications

Paolo Zaroni, G. Panteloglou, A. Othman, J. T. Haas, R. Meier, A. Rimbert, M. Futema, et al. "The U2-spliceosome and its interactors regulate the levels and activity of the LDL receptor in humans". *Atherosclerosis* 315 (2020): e15.

Paolo Zaroni, G. Panteloglou, A. Othman, J.T. Haas, R. Meier, A. Rimbert, M.Futema, et al. "Posttranscriptional Regulation of the Human LDL Receptor by the U2-Spliceosome". *Circulation research*, 2021.

Roeland Broekema, O.B. Bakker, A. van der Graaf, J. Gelderloos-Arends, R. Modderman, L. Paige, B. van Steensel, C. Wijmenga, S. Sanna, S. Withoff, J. van Arensbergen, I.H. Jonkers Fine-mapping of Coeliac disease associated genetic variants using SuRE-SNP implicates role of genetics in epithelial cells. *In preparation*

Adriaan van der Graaf, M. Swertz C. Wijmenga and S. Sanna. Multi-locus MR-link identifies pleiotropy robust causal relationships with high power. *In preparation*

Acknowledgements

First, I would like to thank **all the study participants** that have participated in the studies that are included in this book. You are the unsung heroes of our research. Without your altruistic donations and diligent questionnaire responses, we would have nowhere near the knowledge that we have today and this book could never have been written.

Dear **Lude**, thank you again for being my promotor and introducing me to the Genetics department in Groningen. Without you, I probably wouldn't even have known what kind of great science is possible in human genetics.

Cisca, it is with you that I started my PhD. I'm still grateful for your generous offer, even after we found out that it was likely that there would be some problems ahead. I feel privileged to have had you as my promotor, and that you have entrusted me with four wonderful years of research where I was always able to do whatever interested me. You have taught me so much about how research is done and positioned. Your drive is amazing and how you have been able to manage everything in these last few years, even in the face of so much resistance, is truly something to look up to.

I think it is fair to say that my dissertation would be nothing without your help, **Serena**. Your rigor, tact, ethics and understanding are what made the whole PhD worthwhile. Even if it wasn't clear to me you were always able to identify the best route to take. When I had a suggestion or an idea you would always take the time to understand it. Even though at times I didn't always make too much sense. But perhaps more importantly is that you were able to judge any idea by their merit and if it had any, you could be convinced, even if you initially disagreed. In life but also in science, this characteristic is rare to find. It made working together with you immensely satisfying. Thank you again for supervising me on this journey. Thanks to you, it's been the best time of my life.

Dear **Iris** and **Maria** thank you for all your patience in trying to bridge the gap between the celiac disease wet lab and the in silico work. I really liked working on the project together. I couldn't have done it without you! **Sasha** and **Jing**, thank you for taking me under your wing, and letting me work on some of your projects! I really admire how amazingly efficient you are able to do science in a collaborative way.

Mijn ouders, **Betty** en **Kees**, ik moet jullie bedanken voor alle mogelijkheden die jullie me hebben geboden. Ik heb altijd graag gebruik gemaakt van wat extra tijd, en het was super dat jullie mij die tijd gunden. De steun die ik van jullie heb gekregen nadat ik terugkwam uit Australië zal me altijd bijblijven. Het was bijzonder fijn om apart van elkaar te leven, maar toch veel contact te hebben, dat is een privilege dat helaas niet veel anderen hebben tijdens de PhD en daar ben ik jullie ontzettend dankbaar voor.

Carlos and **Esme**! Thanks for all the support you have given me finalizing the PhD! I always felt completely at ease while I was with you guys.

acknowledgements

On top of that, Carlos, thank you for being my paranymp! I'm looking forward to the ceremony!

Roeland, bedankt dat je mijn paranymp bent! Ik voel me altijd bij jou op mijn gemak. Ik kijk graag terug op het coachen samen en op die keren dat we de tweehead op de Amstel voeren. Het was volgens mij leuker voor mij dan voor jou, maar het was een gecombineerde prestatie waar we beide trots op kunnen zijn.

Lieve **Anne** and dear **Benjamin**, thank you for all your support in the last years. It was wonderful to spend time together. I still fondly remember our skiing vacation together that we had just before COVID became a pandemic. Hopefully, we will be able to do a lot more of those shared weeks of vacation in the future.

Lieve familie, **Bob, Aart, Netty, Lon, Marjo, Els, Veronique, Adriaan** en mijn **neven** en **nichten**. Het was bijzonder fijn om een fundament te hebben waar ik altijd naar terug kon keren. Elke keer als ik jullie opzocht wilde ik jullie weer vaker zien, maar helaas is dat toch makkelijker gezegd dan gedaan.

Beste **Miekel** en **Sandor**, bedankt voor jullie onvoorwaardelijke steun tijdens mijn lange tijd aan de universiteit. Het is heerlijk om een plek te hebben bij jullie waar ik altijd welkom ben. Dit kwam altijd goed uit als ik soms even klaar was met het leven in Groningen. Ik ben dankbaar dat jullie me meenamen naar concerten, opera's en dat ik altijd uitgenodigd werd voor jullie pianosalon. Onder anderen samen met **Peter** waren de concerten en de avonden echt gezellig! Ik heb door jullie klassieke muziek echt leren waarderen, maar misschien belangrijker is dat jullie me de steun gaven om te doen wat ik echt leuk vond.

Trijnie en **Evelien**! Jullie zijn mijn rode draad door het leven! Ik vind het jammer dat ik jullie niet meer zo vaak zie, maar het was heel fijn om jullie altijd ergens op de achtergrond te hebben waar ik naartoe kon trekken als het wat moeilijker ging.

Nanno en **Gree**, bedankt voor al jullie gastvrijheid in al die jaren! Ik ben zeer onder de indruk van hoe jullie voor elkaar zorgen en hoop dat ik jullie nog veel kan zien.

Julian, we zien elkaar natuurlijk te weinig, maar ik kijk met heel veel plezier terug naar al die tijd die we samen hebben verdaan. Ik moet nog heel vaak terugdenken aan onze vakantie in Iran en die taxichauffeur die ons nergens naartoe bracht in Kashan. Daarnaast moet ik natuurlijk ook aan de rest van de club denken: **Bastiaan, Casper, David, Justin, Lennart** en **Thomas**. Bedankt voor onze tijd samen, en hopelijk zien we elkaar snel weer!

Beste **Bram, Gelein, Sietce** en **Ivo**, het was ontzettend leuk om met jullie de weekenden af te spreken. We wonen niet allemaal meer in Groningen, maar ik kijk met veel plezier terug op de dagen en weekenden die we

samen hebben besteed. Op naar de volgende!

Lieve **Roos**, bedankt voor alle tijd die we samen hebben doorgebracht tijdens de PhD. Ik vond het heel fijn om samen op vakantie te gaan, wandelingen te maken en te genieten van onze avonden samen. Hoewel onze wegen zich hebben gescheiden kijk ik met veel liefde terug op de tijd die we samen hebben doorgebracht.

Edgar! Ik vond de tijd van onze PhD samen echt ongelooflijk leuk! Met jou een beetje ouwehoeren was ongekend! Veel succes als huisarts, en hopelijk zien we elkaar snel weer!

Lieve **Laurence**, bedankt dat je me onder je vleugels hebt genomen om de fijne kneepjes van het academische leven te leren kennen. Het was altijd leuk om af te spreken en een maaltijd te delen onder het genot van een wijntje.

Dan, thank you for being a wonderful friend. It's great to have someone who can make friends everywhere. **Ronald, Pierrick** and **Robert**, it was great to be your friends in Australia, I still look back at that time fondly.

De **MGLs**, **MGLD** en **alle roeiers op Gyas** het coachen, roeien en de biertjes waren super.

Jeroen en **Harold**, de potjes waren echt super! Die hebben me door deze vreemde COVID tijden heen geloodst.

Jochen and **Sven**, thanks for being my bouldering buddies! It's been wonderful to get to know you and all the best in the future! Let's meet up soon!

Hey **Aniela, Christel, Ciska, Valentijn, Iris, Jolijn, Jorrit B, Jorrit P, Kim, Loes, Mark, Hanneke, Stephanie, Niek, Hester, Paul, Hugo, Tymke, Ruby, Eline, Vera, Jan, Yvonne, Lisetten, Laura, Christian, Lotte, Geert, Johan, Alicia, Thomas, Matthijs, Chris** en iedereen van **Vera** die ik vergeten ben! Het was super leuk om met jullie alles mee te maken! Ik mis de concerten en tijden voor en achter de bar!

All my colleagues: thanks for making the work so much fun! It was great to have such a wonderful and diverse department!

George, thank you for your honest feedback on my MR work. I would like to thank you for inviting me to Bristol, and allowing me to present MR-link. It was very generous and wonderful to be there. Unfortunately, the internship didn't go through due to COVID, but it was nice to have tried it!

Esteban, thank you for being my colleague through thick and thin. I really admire your rigor, you always try to understand things to the finest detail which is a great characteristic for a scientist. I will never forget our time traveling together in Gothenburg and Cagliari. I wish you all the best in finishing your PhD. I'm very certain you will be very successful in everything you do!

acknowledgements

Arnau and **Werna**, and **Ineke** and **Nick**, thank you for all of the hospitality and time together, it was really amazing to have our dinners! I admire how you are able to balance everything that happens in life, while still remaining friendly and down to earth.

Juha, thanks for being a good friend, and see you soon in Finland. **Irene**, thanks for all the walks, the runs and the bouldering. **Kieu**, thanks for the bouldering and the barbeques. **Olivier**, thanks for all the discussion and walks, and working together, it was great to have that time together. **Marc-Jan**, thank you for hosting me near Cambridge, and always being able to have a good conversation in science, but perhaps more importantly, in how to keep being yourself in a scientific career. And **Kate**, thanks for your editing, but more importantly for your kind advice. **Raul**, thank you again for showing me how to perform a successful PhD in the department! All the best to you! **Alex**, it's been great to have you in the department, where we were able to take the time to discuss everything and nothing.

Sebo, **Eva**, **Aaron**, **Roeland**, **Iris**, **Joram**, **Renee**, **Jelle** en **Janneke** thanks for all the time together! It was really wonderful to join you all in the work, and the social gatherings, especially the barbeques at Sebo's place!

Niek, **Annique**, **Dasha**, **Patrick**, **Monique**, **Roy**, **Martijn**, **Marjolein**, **Shuang**, **Urmo**, **Tyler**, **Harm**, **Harm-Jan** en **Paulien**, bedankt dat jullie als een klankbord hebben gediend in het doen van mijn computationele en statistische werk. Het was fijn om jullie meetings bij te wonen, en jullie inzichten te horen. Daarnaast waren de sociale activiteiten altijd heel leuk!

Valerie, **Alex**, **Ranko**, **Michiel**, **Johannes**, **Marjolein**, **Laura**, **Dianne**, **Michiel**, **Marloes**, **Sergio**, **Marwah**, **Johanna**, **Siobhan**, **Renata**, thanks for all the time together, at work, dinner or next to a few cold ones. It was amazing!

Lieve **Céline**, bedankt voor alle support tijdens het schrijven en het doen van de opmaak van dit boekje! Ik vind het echt heel leuk om je te leren kennen, en hopelijk kunnen we dat nog heel lang samen doen!

And **everyone else** from which I received support, a heartfelt thank you!