Celiac Disease Triggers

Using Mendelian Randomization to infer causality

The largest conducted Mendelian Randomization screening

AUTHOR
M. Knigge

INSTITUTE
Hanze University
Life Science and Technology

ORGANIZATION
Eriba
Deparment of Genetics

DATE
Wednesday, January
10-01-2018

Celiac Disease Triggers

Using Mendelian Randomization to infer causality

umcg

Hanze University of Applied Sciences Groningen

AUTHOR
M. Knigge

SUPERVISOR
Asst. Prof. Dr. S. Sanna

LECTURER
Dr. M. A. Noback

INSTITUTE
Hanze University
Life Science and Technology

ORGANIZATION
Eriba
Deparment of Genetics

DATE
Wednesday, January 10-01-2018

PREFACE

ABSTRACT

Celiac disease is a complex autoimmune disorder of the small intestine characterized by gluten-sensitivity. When people with celiac disease digest gluten their body triggers an immune response that attacks the small intestine. Celiac disorder affects up to two percent of the European population, and it is hereditary. The inheritance model of celiac disease is oligogenic. In this model, HLA-DQ2 is a genetic variant in the major histocompatibility complex (MHC) region that accounts for 40% of the inheritable risk. But, this genetic variant alone is not sufficient enough to develop celiac disease. 30% of the Europeans are carriers of the variant, however 3% express celiac disease.[2]

The aim of this study is the identification of factors that cause or protect against celiac disease, to explain what is triggering disease manifestation in HLA-DQ2 carriers. To determine potential causal or protective relationships we investigated over 500.000 clinical parameters and molecular phenotypes. We used the concept of Mendelian Randomization[3] (MR), and applied two-samples MR methods: Inverse Variance Weighted[4] (IVW) and MR-Egger method[5] to infer causality links by combining summary statistics from genome-wide association studies (GWAS).

We identified 1133 significant (FDR<0.05) clinical parameters that cause or protect for celiac disease, and others that show feedback mechanisms . For example autoimmune diseases such as Type 1 Diabetes (T1D), have been largely described to co-occur with celiac disease.[45] We also confirmed a causative effect of Eosinophils (and protective of lymphocytes) in increasing the risk of celiac disease, which was recently described using the same approach[46]. Here we also observed that there is feedback mechanisms, with the disease also affecting, in turn, the number of Eosinophils and lymphocytes. Finally, in addition to parameters that are supported by previous epidemiological studies, we found an interesting causative role for infection diseases and allergies. At the molecular level, we observed evidence of causality for changes in expression of 25 genes and for changes in methylation patterns at other 12 (FDR<0.05). Interestingly, 21 of those 37 genes are in loci not previously associated with Celiac Disease at the genome-wide level, but the majority (13) is known to be associated with other autoimmune diseases, tonsillitis and/or asthma. Those results not only strength the links that we detected with those clinical parameters, but also suggest novel candidate genetic loci to be assessed in future genetic studies.

We have carried out a systematic application of Mendelian Randomization to hundred of thousands of phenotypes, making this study the largest of its kind. We have identified causal and protective factors for celiac disease, recapitulating knowledge from epidemiological studies and also highlighting new key players. We expect that application to other diseases can bring new insights in the understanding of pathophysiology as we well as of biological mechanism, especially as more genome-wide association studies on -omics measurements will be available in the future.

ABBREVIATIONS

GWAS	Genome-Wide Association Study
HLA	Human Leukocyte Antigen
IV	Instrumental Variable
IVW	Inverse-Variance Weighted
LD	Linkage Disequilibrium
MHC	Major Histocompatibility Complex
MR	Mendelian Randomization
OR	Odds Ratio
RCT	Randomized Controlled Trial
SNP	Single-Nucleotide Polymorphism
UMCG	University Medical Center Groningen

ORGANISATION

This internship was provided and guided by the celiac disease research group in the department of Genetics at the University Medical Center Groningen (UMCG). The hierarchy can been seen as a structure divided in multiple sections (A-F), each with its own director (figure 1). Above these directors is a main director. All sections are divided into departments, and each department has a head. For the Department of Genetics this is prof. Richard Sinke. Under his coordination, several associated and full professors lead different research groups, including prof. Cisca Wijmenga who is the leader of the Celiac Research group.[37]

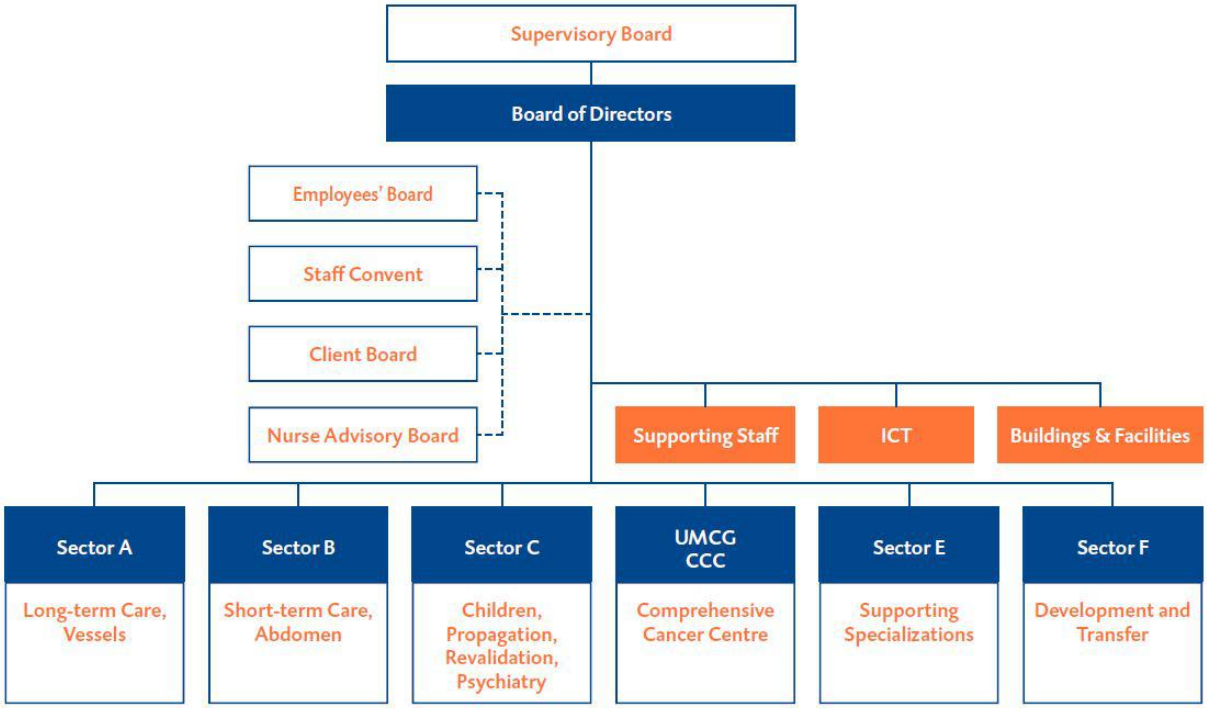

Figure 1: Organogram of the UMCG organizational structure.[37]

TABLE OF CONTENTS

INTRODUCTION

Celiac disease is a complex autoimmune disorder of the small intestine characterized by gluten-sensitivity. When people with celiac disease digest gluten their body triggers an immune response that attacks the small intestine. This causes damage to the villi that are lined up in the small intestine for nutrient uptake.[1] Celiac disorder affects up to two percent of the European population, and it is hereditary. People with a first-degree relative with celiac disease have 30 times higher risk of developing celiac disease. The inheritance model of celiac disease is oligogenic, which means that a phenotypic outcome is determined by multiple genes, with one having a much stronger impact than others. In this model, HLA-DQ2 is a genetic variant in the MHC region that accounts for 40% of the inheritable risk. But, this genetic variant alone is not sufficient enough to develop celiac disease. 30% of the Europeans are carriers of the variant, however 3% express celiac disease.[2]

The aim of this study is the identification of factors that cause or protect against celiac disease, and to quantify the impact of the causal or protective relationship. To determine potential causal or protective relationships more than 500.000 public available clinical parameters and molecular mechanisms -- like Crohn's disease, Inflammatory bowel disease, Hemoglobin concentration, and Eosinophil percentage of granulocytes -- will be screened with the MR approach. MR is an approach that uses genetic variants associated with a phenotype of interest to estimate a causal or protective relationship between this phenotype (exposure) and a relevant medical condition (outcome).[3] This study uses two-samples MR[4] methods: IVW[5] method and MR-Egger method[5]. Such methods can infer causality links by statistical methods that combine summary statistics data from GWAS, and thus have a great potential to be applied today in the era of data sharing.

We expect to find multiple causal or protective clinical parameters or molecular mechanisms that will be quantified to understand the impact of the causal or protective relationship, that can potentially explain what is triggering celiac disease in the 3% from the 30% of the Europeans that are carriers of HLA-DQ2.

THEORY

CELIAC DISEASE

Celiac disease is an autoimmune disorder that has the possibility to occur genetically in predisposed people where the ingestion of gluten cause damage. When people with celiac disease digest gluten their body triggers an immune response that attacks the small intestine.[1] This is characterized by small intestinal damage with the loss of function from the villi to take up nutrients, which leads to malabsorption.[6]

MENDELIAN RANDOMIZATION

MR refers to the process of random segregation and assortment of alleles from an ancestor to offspring, that occur during gamete formation, in meiosis.[3, 4, 5]

This process gives us the advantage of using these genetic variants in observational settings to predict causal or protective relationships between exposure and outcome.[3, 4, 5] This process is comparable to Randomized Controlled Trials (RCT). In RCTs, the process of random and evenly distributing the sample into two arms ensures that known and unknown confounders are distributed evenly across both arms, see figure 2, stage (A). In MR, the random segregation and assortment of alleles is analogous the process of randomization in RCTs. In stage (B) for RCT one arm will be the case group, where some form of experiment will be conducted, and the other is the control group. At this stage for MR, both arms will represent different genotypes which in stage (C) will lead to different products. For RTC at this stage, in the case arm there will take place on-off-target effects. At the final stage (D), differences between the two arms from both RCT, and MR can be studied and should only reflect the differences between the two arms.

The MR approach uses genetic variants as Instrumental Variables (IVs) to infer a causal or protective relationship between an exposure and outcome. IVs are variables that are associated with the exposure phenotype of interest and does not suffer from reverse causality or confounding factors.[3, 4, 5] Reverse causality is the phenomena where it is expected that for example X causes changes in Y, but it can be that Y causes a change in X.Take this example: an individual who is characterized by being a heavy alcohol consumer, and because of the over consumption is that individual depressed. But, it is possible that this individual is a heavy alcohol consumer due to being depressed. Confounding is the variable that is not accounted for and can bias the analysis by suggesting that there is correlation when that is not the case.

MR uses genetic variants for a variety of reasons. (I) genetics variants are less amenable for confounding factors. As denoted by Mendel's first law, the law of segregation; genetic variants segregate randomly and independently from environmental factors, and Mendel's second law, the law of independent assortment; genetic variants segregate independently from other traits. (II) reverse causality does not affect the genetic instruments because an individuals germline genotype precedes the medically

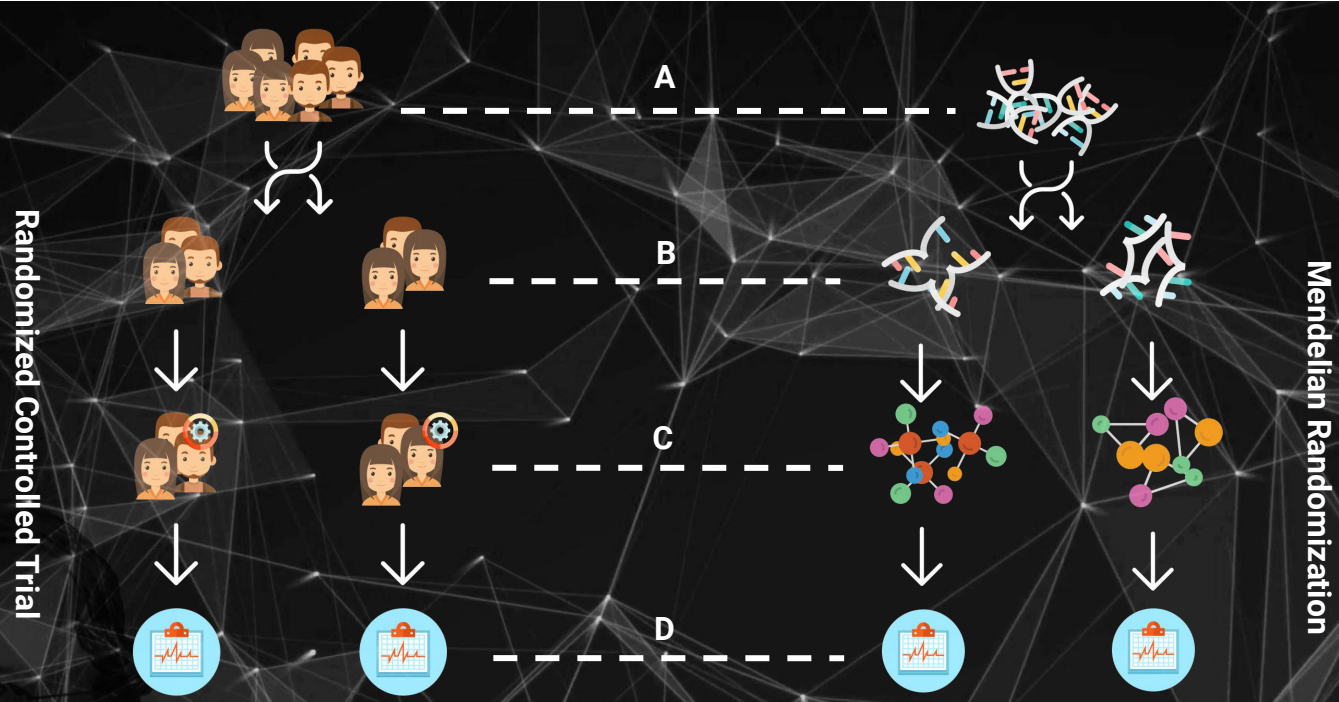

Figure 2: RCT in comparison to MR. The left diagram is an overview of a RCT, and the right diagram is an overview of a MR. (A) describes the random allocation and segregation of alleles in MR, and in RCT the random distribution of the population into two arms. At stage (B); here in RCT there will be a case group and a control group, and for MR there will be two different genotypes. After that, stage (C); for RCT here will take place different target effects for one of the arms, for MR there will be two different products from the genotype which in turn in stage (D) give rise to a different effect, as does this take place for RCT. And these differences in arms can be studied.

relevant disease.[3, 4, 5] (III) genetic variants are subjected to little measurement error or bias. (IV) MR does not force to use the actual causal variant but is satisfied with a marker that is in Linkage Disequilibrium (LD) with the causal genetic variant. (V) MR gives us the advantage to use the genetic variants from GWAS summary statistics, which are routinely available on large well-phenotyped scale.[7, 8]

Linkage Disequilibrium is the association of alleles at different loci in a population that are not random. Loci are in LD when the frequency of alleles of the associations is higher or lower than is expected to occur when the loci were independent and associated at random.[11]

For a genetic variable to be used as an IV, the variant must meet three core assumptions. (I) the genetic variant must be associated with the exposure, (II) the genetic variant must not be associated with confounders, (III) the genetic variable can only be associated with the outcome through the exposure[3, 4, 5], see figure 3.

The first assumption can be assessed by investigation of the strength of the association between the genetic instrument and the exposure. The second assumption can be assessed by examining the relationship between the genetic variant and potential confounders. Final, the third assumption can be addressed by examining the biological pathways of the genetic instruments.[3]

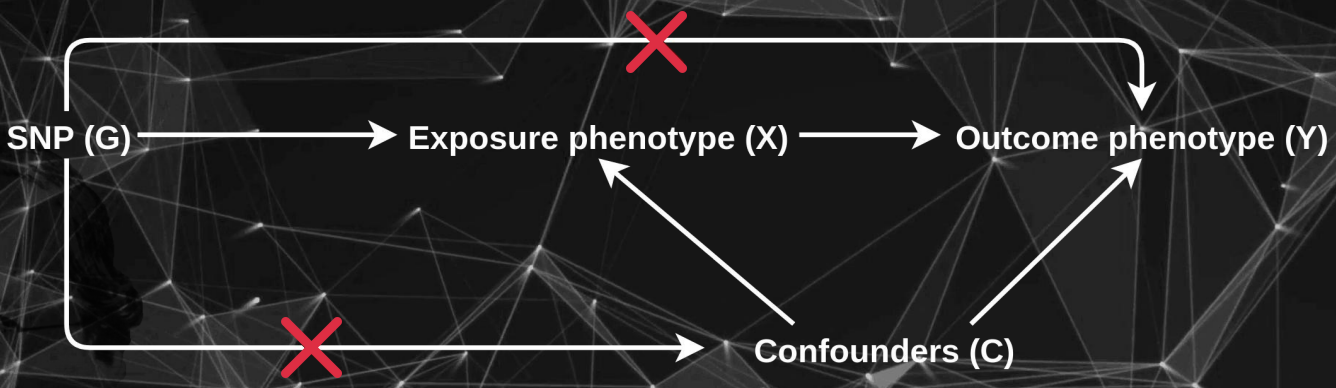

Figure 3: graph showing the assumptions made by MR. The nodes in this graph represent the genetic instrument: Single-nucleotide polymorphism (SNP), the exposure, the outcome, and confounders. This graphs shows the relationships between variables that meet the three core assumptions. (I) Genetic variant must be associated with exposure, (II) no association between genetic variant and confounders, (III) only association with outcome through exposure. The red crosses show the relationships that are not allowed in the MR framework.

ONE-SAMPLE MENDELIAN RANDOMIZATION

One-sample MR is the basic implementation of MR on SNPs, exposure, and outcome from a single data set, with the addition of genetic data from participants. The causal effect of the exposure on the outcome can be estimated by a two-step regression analysis. In the first step , a linear regression is fitted on the genetic variants from the exposure. In the second step,the genetic variants from the outcome are regressed over the predicted values from step one by linear or logistic regression. The coefficient from the second stage can be interpreted as the change in the outcome per unit increase in the exposure. If this coefficient is significant, than the exposure is causal (or protective, depending on the sign of the coefficient) for the disease.[9]

TWO-SAMPLES MENDELIAN RANDOMIZATION

Two-samples MR takes advantage of the fact that it is not needed to obtain both, the exposure and outcome from the same population sample, nor the addition of genetic data. The two-sample approach method makes it possible to use publicly available GWAS. And needs only the genetic variant identifier, effect allele, effect size, and standard error.

To calculate the causal estimate between the outcome (Y) and exposure (X) one needs to take the ratio of the effect from the genetic variant (G) on outcome divided by the effect from the genetic variant on the exposure.[4] This is referred to as the Wald method ($\hat{\beta}_{IV}$), see equation 1. Where $\hat{\beta}_{Y|G}$ is the effect from the genetic variant on outcome, and $\hat{\beta}_{X|G}$ is the effect from the genetic variant on the exposure.[3, 4]

$$\hat{\beta}_{IV} = \frac{\hat{\beta}_{Y|G}}{\hat{\beta}_{X|G}} \qquad (1)$$

In addition, the standard error of $\hat{\beta}_{IV}$ ($\sigma_{\hat{\beta}_{IV}}$) can be calculated as follow (equation 2):

$$\sigma_{\hat{\beta}_{IV}} = \sqrt{\hat{\beta}_{IV}^2 / \frac{Z^2_{\hat{\beta}_{Y|G}} Z^2_{\hat{\beta}_{X|G}}}{Z^2_{\hat{\beta}_{Y|G}} + Z^2_{\hat{\beta}_{X|G}}}} \qquad (2)$$

where $Z_{\hat{\beta}_{Y|G}}$, and $Z_{\hat{\beta}_{X|G}}$ are the number of standard deviations from the mean of the effect from the genetic variant on outcome or exposure.[3, 5]

The use of one single variant in two-samples MR has however several limitations. The first is lack of power: one single variant in fact is often not sufficient to represent the entire variation in exposure, thus limiting the ability to assess causality: the link between exposure variation and outcome. One single variant is also highly prone to biased estimates. In fact, one variant may be associated with (or maybe in LD with another variant associated with) another exposure (this phenomena is known as pleiotropy), and thus the estimated causal effect can be due to both or only one of the two exposure phenotypes. Distinguish which exposure is causal is not possible with only one single variant, this will be significant because of the causal link of the other exposure. To overcome those limitations, other methods have been introduced that can combine multiple genetic variants to increase power and disentangle the issue of pleiotropy.[3]

INVERSE-VARIANCE WEIGHTED MENDELIAN RANDOMIZATION

With multiple genetic variants, the causal estimate ($\hat{\beta}_{IV}$) for each variant can be averaged by using an inverse-variance weighted approach to compute an overall causal estimate, which is denoted as the inverse-variance weighted (IVW) estimate ($\hat{\beta}_{IVW}$). This method assumes that all genetic variants bring independent evidence, that there is no correlation between ratio estimates. The pooled estimate can be calculated as follow (equation 3)[3, 4, 5]:

$$\hat{\beta}_{IVW} = \frac{\sum_G \hat{\beta}_{Y|G}\,\hat{\beta}_{X|G}\,\sigma^{-2}_{\hat{\beta}_{Y|G}}}{\sum_G \hat{\beta}^2_{X|G}\,\sigma^{-2}_{\hat{\beta}_{Y|G}}} \qquad (3)$$

This pooled estimate can also be calculated by deploying a weighted linear regression of the genetic variants from the outcome on the genetic variants from the exposure with using inverse-variance weights ($\sigma^{-2}_{\hat{\beta}_{Y|G}}$), this method does not contain an intercept term (equation 4)[3, 4, 5]:

$$\hat{\beta}_{Y|G} = \hat{\beta}_{IVW}\hat{\beta}_{X|G} + \epsilon_G; \quad \epsilon_G \sim \mathcal{N}(0, \sigma^2_{\epsilon_G}\sigma^2_{\hat{\beta}_{Y|X}}) \qquad (4)$$

Here $\hat{\beta}_{IVW}$ is the parameter, ϵ_G is the residual term, and σ_{ϵ_G} is the residual standard error.[3, 4, 5]

In addition, to compute the standard error of the pooled estimate ($\sigma_{\hat{\beta}_{IVW}}$) see equation 5:

$$\sigma_{\hat{\beta}_{IVW}} = \frac{1}{\sqrt{\sum_G \hat{\beta}^2_{X|G}\,\sigma^{-2}_{\hat{\beta}_{Y|G}}}} \qquad (5)$$

MENDELIAN RANDOMIZATION-EGGER METHOD

MR-Egger is a method which can be deployed when the core assumptions do not hold, because it uses a weighted regression with an unconstrained intercept to regress the outcome on the exposure. The unconstrained intercept from this does not take into account the assumption that all genetic variants meet the assumptions, due to this, this method is

less biased by potential confounding. MR-egger contains three parts: (I) a test for violations of the core assumptions, (II) a test for testing the causal effect, and (III) an estimate of the causal effect.[5]

MR-Egger is similar to IVW with respect to a simple modification of the weighted linear regression. Instead of setting the intercept term zero, it is estimated as part of the analysis (equation 6)[3, 4, 5]:

$$\hat{\beta}_{Y|G} = \hat{\beta}_{0E} + \hat{\beta}_{1E}\ \hat{\beta}_{X|G} + \epsilon_G; \quad \epsilon_G \sim \mathcal{N}(0, \sigma^2_{\epsilon_G}\sigma^2_{\hat{\beta}_{Y|X}}) \qquad (6)$$

Here $\hat{\beta}_{0E}$ is the intercept, and $\hat{\beta}_{1E}$ is the slope, that is the causal estimate of the method. The standard deviation from the MR-Egger estimate can be calculated by dividing the slope ($\hat{\beta}_{1E}$) by the residuals standard deviation (σ_{ϵ_G}) from the weighted linear regression.

The MR-Egger method contains a test for the estimate, which is a test to determine if the estimate differs from zero (MR-Egger causal test). The intercept is interpreted as the overall pleiotropic effect. If the computed intercept is zero, than the estimate will be the same as the estimate from IVW. But with a non-zero intercept it is indicated that the IVW estimate can not be trusted, this is known as the MR-Egger intercept test.[5]

COCHRAN'S Q TEST

To test if the single variant estimates from MR are sufficiently similar to warrant their combined result, the homogeneity (measure of the similarity between estimates) of those estimates is assessed by the Cochran's Q statistic (Q). This Q statistic is applied to determine if there is a subset of the data that differs from the rest of the pool, which has the potential to mask the true identity of the direction of the causality or reflect causality when it is not present.[12, 15, 16]

To apply Q statistic it needs estimates of the causal effect from the single variants ($\hat{\beta}_{IV}$) and the estimated standard errors of those effects ($\sigma_{\hat{\beta}_{IV}}$). Which are the the causal estimates calculated by the Wald method. Cochran's Q measures heterogeneity of $\hat{\beta}_{IV}$ by inspecting the variation around the estimate from the IVW method ($\hat{\beta}_{IVW}$). where the weight for $\hat{\beta}_{IV}$ is the estimated variance ($1/\sigma_{\hat{\beta}_{IV}}$), see equation 7.[12]

$$Q = \sum_G \frac{1}{\sigma_{\hat{\beta}_{IV|G}}} (\hat{\beta}_{IV|G} - \hat{\beta}_{IVW})^2 \qquad (7)$$

Here, G is the amount of causal estimates between exposure and outcome. Q is a test for homogeneity, which under the null hypothesis follows a chi-squared distribution, with G-1 degrees of freedom where G is the number of variants. The null hypothesis is rejected when the value of Q reaches the critical value of chi-square statistics corresponding to a probability of results to occur by chance of .05 (this threshold is different for for each distribution with different degrees of freedom), then heterogeneity is present.[12, 13, 14, 15, 16]

PATHWAY SCORING ALGORITHM

The Cochran's Q method, when significant will identify a subset of variants that deviates from the average causal estimate, and

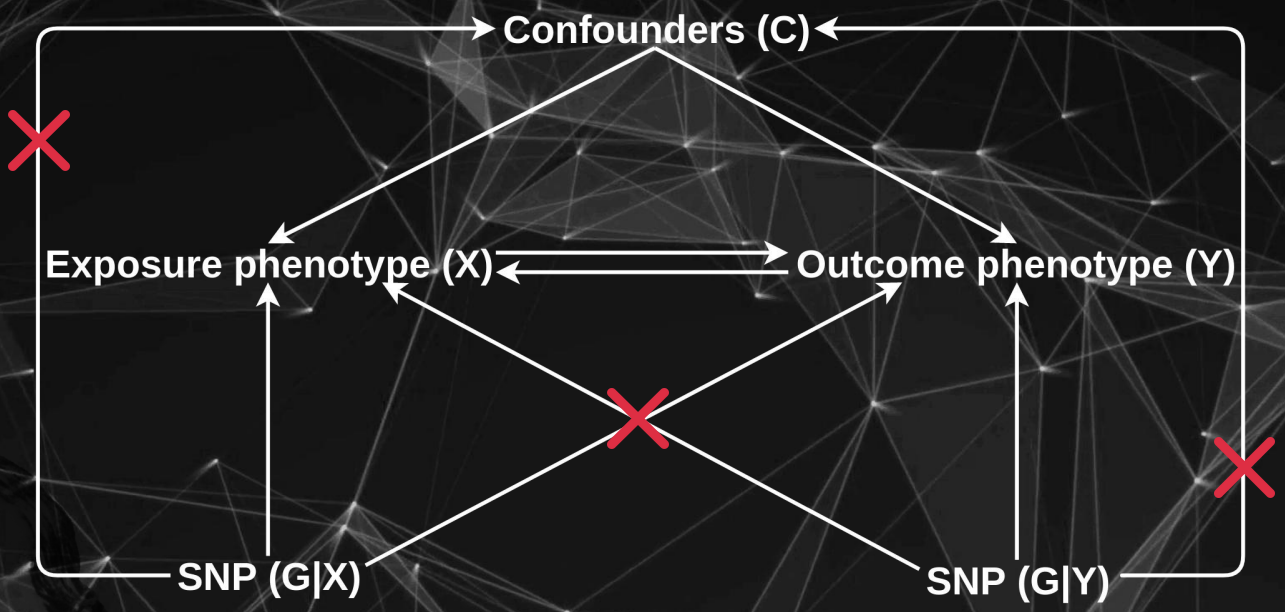

Figure 4: this graph shows the framework of a two-sample bidirectional MR. This framework is comparable to that of two-samples MR, but in addition the relationship between an exposure phenotype and outcome phenotype is exploited in both directions to gain insight in the link between two phenotypes. As in the MR framework there are assumptions that must be met, which are also applied in this framework. (I) genetic variants must be associated with the exposure, (II) no association between genetic variants and confounders, and (III) only association with outcome through exposure. This is here exactly the same, but only if the direction changes, the phenotypes for exposure and outcome will be flipped. The red crosses represent the relationships that violate the core assumptions.

another subset of variants that lead to homogeneous estimates. Both subsets are variants associated with the same exposure, but they may act through different biological pathways. Understanding the biological difference identified by the two subsets is key to better understand the link between the exposure and outcome

To gain biological insight a tool for computing gene and pathway scores from SNP-phenotype summary statistics can be used: Pathway scoring algorithm (Pascal). This tool allows gene and pathway-level analysis of GWAS without the restriction to access original genotypic data.[17]

TWO-SAMPLES BIDIRECTIONAL MENDELIAN RANDOMIZATION

The factor hypothesized to be an exposure may instead be a consequence of the disease, or the same factor may be both causal and a consequence of the disease through circular feedback mechanisms. Therefore, two-samples bi-direction MR, or in other words, MR test where each tested, exposure is in turn considered an outcome and the outcome the exposure, can be used to clarify the complete relation between a factor and a disease.[18]

MATERIAL & METHODS

The aim of this study is to find factors that protect or cause celiac disease, and to quantify the protective or causal relationship. More than 500.000 publicly available clinical and molecular phenotypes have been analyzed with the two-samples bi-directional approach.

In order to perform a MR analysis at this scale several steps have been made. The first step is data acquisition (public available GWAS summary statistics), (II) pre-processing of the data, (III) clumping of SNPs that are in LD, (IV) harmonization of the outcome on exposure or vice versa if direction is to be changed, (V) formal testing of two different two-samples MR methods, (VI) test for heterogeneity in the data set, (VII) investigation of biological differences between heterogeneous sets identified (in point VI), (VIII) visualization results by plotting the outcome against the exposure, (IX) interpretation of the causal or protective impact from exposure on outcome, or vice versa.

COMPUTATIONAL INFRASTRUCTURE

To ensure reproducibility and efficiency of the analyses, all steps are implemented in a software package for the R open-source environment. The package, named mendelian-Randomization, performs MR analyses using summarized data including data pre-processing and post-results sensitivity analyses as described in the following paragraphs.

The mendelianRandomization package can be downloaded from the web-based hosting service Bitbucket:

https://bitbucket.org/MatthijsKnigge/mendelianRandomization.

Or can be directly installed into R by using libraries that are described in the tutorial on the source page.

This package provides the user with several options to perform a MR analysis. The core functionality that is implemented are the two-samples MR methods: Wald ratio, IVW, and MR-Egger. However, there is also the functionality to pre-process the data, clump the data, harmonization of the data, a test for assessing the homogeneity, and a function for predicting biological pathways, and other functionalities which are discussed in detail on the source page of the package.

DATA

To perform a MR analysis, the genetic variants for the outcome and exposure parameters must be defined. For the outcome, two GWAS summary statistics were used: a genome-wide association study of 4533 individuals with celiac disease consisting of 505.721 genetic variants, and a GWAS study of 12.041 cases and 12.228 controls characterized with a densely genotyped array, having 97.434 genetic variants.

The exposures were acquired from publicly available databases (GWAS catalog[24], GTExPortal[25], UKBiobank[22, 23], and mQTL[26] database) and additional GWAS summary statistics from single research group websites cited in their scientific publications. The exposures consist of a variety of clinical parameters and molecular mechanisms such as

cholesterol, body size, psoriasis, hormones, and gene expression and methylation levels profiles in several tissues.

The GWAS catalog is a database that contains GWAS summary statistics of published SNP-trait associations. In total 4.063 GWAS summary statistics were acquired from this database, like hormone levels, arthritis rheum, diabetes, body fat percentage, and hematocrit.

GTExPortal database contains correlations between genotype and tissue-specific gene expression. GTExPortal contains gene expression levels for 46 different tissues evaluated in up to 400 individuals. These tissue-specific profiles have been divided in tissue-gene traits, which gave 280.226 different traits in total. Also included is a genotype to gene-expression association summary statistics from a large expression study of 5.311 individuals, which assessed expression variation in whole blood.[47]

The UK Biobank is a database of GWAS associations for hundreds of traits carried out in the UKBiobank, a population cohort of 500,000 samples from the UK, containing the results of 3202 GWAS summary statistics and heritability analyses, like bread intake, nucleated red blood cell percentage, and monocyte count.

mQTL database contains a large-scale genome-wide DNA methylation analysis of 1.000 mother-child pairs at serial time points across different stages in life: birth, childhood (seven years), adolescence (15-17 years), pregnancy, and middle age. Every DNA methylation analysis has been divided in unique stage-gene-SNPs traits, which in total gave 222.542 different genes for all methylation profiles.

The additional downloaded exposures from publicly available publications are in total 1.310, this are traits like asthma, coronary artery disease, High density lipoprotein (HDL) cholesterol, platelet count, and mean cell volume.

HUMAN LEUKOCYTE ANTIGEN REGION

The MHC is within the human genome the most dense region in terms of polymorphism, and exhibits the most complex LD patterns. LD may vary drastically from one population to the other, and may also extend for several hundreds of Kb or even Mb. Therefore variants in this region (chromosome 6, from 20Mb to 40Mb) have been discarded to avoid an over-estimation of causality for factors that are influenced by the same HLA haplotypes of Celiac disease, or by haplotypes in LD with it.[35]

The variants in this region for the two outcome GWAS summary statistics have been discarded, the Dubois GWAS[20] went from 505.721 genetic variants to 502.783 genetic variants, and the Trynka GWAS[21] went from 97.434 to 89.620genetic variants.

PRE-PROCESSING

In addition to defining genetic variants to use for a MR analysis, these variants should be chosen appropriately, meaning that the study should not violate the core assumptions. In this step the first core assumption was

addressed (the genetic variant must be associated with the exposure[3, 4, 5]) by selecting for genome-wide significance: only SNPs with a p-value of 5e-8 (the conventional threshold for GWAS statistical significance[27]) or smaller where used.

Furthermore, genetic variants where removed if it was not possible to derive their effect size or Odds Ratio (OR), when there was allelic information missing and it was not possible to query them by the 1000 Genomes phase 3 reference panel[28], when variants were listed twice in the same file but with different effect sizes, when it was not possible to infer the direction of the allele, specifically in the case of palindromic SNPs (A/T and C/G).

LINKAGE DISEQUILIBRIUM CLUMPING

LD is defined as the difference between the observed frequency of an allele at two loci and the frequency expected from random segregation.[11] Gene loci are presumed to be independent of one another, however LD suggests otherwise, it describes the situation in which they are not, and genetic variants could be coinherited. This creates a situation where LD is introducing pleiotropy. This leads to the associations of SNPs related to more than one post-transcriptional process.[29] And

Clearly this violates the second core assumption: the genetic variant must not be associated with confounders.[3, 4, 5] To counter attack SNPs that are in LD which can introduce pleiotropy, these SNPs have been clumped with PLINK 1.9[30].

Clumping was performed on the exposure data in order to find LD structures and to disentangle pleiotropy. With PLINK LD is estimated between SNP-SNP pairs based on a reference genotype data set, where a LD score is calculated (the squared correlation based on genotypic allele counts: r^2).[31] And only one SNP per region of LD structure is used, and this SNP must also be present in the outcome data set. The clumping analysis was performed using the following PLINK command:

```
./PLINK
    --bfile EUR_1000G_phase3_vcf
    --clump exposure.data
    --clump-p1 5*10^-8
    --clump-p2 5*10^-8
    --clump-r2 .1
    --clump-kb 1000
    --clump-verbose
    --out exposure.clump.data
```

Clumps are formed around index variants which by default have a p-value not greater than 1e-4, this threshold can be adjusted with --clump-p1. Sites that are n kilobase (kb) from an index variant, and have r^2 larger than m, and have association p-value lesser than given are assigned to that index variant, if they have not been previously assigned to another LD structure. These parameters can be adjusted with --clump-p2, --clump-r2, and --clump-kb. The --bfile parameter expects the reference genotype data. The --clump-verbose parameter is an option for returning an extended report. Last, --out is used for defining the output file.[31]

As stated earlier, if MR wants to access the causal or protective relationship between an exposure and outcome, the genetic variants

from the exposure must be present in the outcome, however, this is not always true. This is where LD has been used in advantage of the study. Namely, MR does not force to use the actual causal variant but is satisfied with a marker that is in LD. When a certain SNP in the exposure is not present in the outcome, a proxy SNP has been used that is in LD if its r^2 is equal or greater than .8 , with respect to the given parameters.[32]

HARMONIZATION

When using two or more independently generated data sets data harmonization is an essential step for two-samples MR, because GWAS summary statistics from different studies do not have harmonized alleles.[33]

In the harmonization from the outcome on the exposure, several steps have been made. (I) all the signs from genetic variants from the exposure where flipped if they where negative, and if the effect was negative, the major allele and minor allele also have been flipped if true. This has been done because MR-Egger is very sensitive to changes in direction of the exposure effects (which are the independent variable in the model)[5], (II) all variants from the outcome must be associated with the exposure in the same direction. When a variant is not coded in the same direction, then the variant has been flipped, which means that in the outcome the alleles, and genetic effect are flipped :

outcome

effect allele	other allele	effect size
A	**C**	**0.5**

exposure

effect allele	other allele	effect size
C	**A**	**0.5**

↓

outcome

effect allele	other allele	effect size
C	**A**	**- 0.5**

exposure

effect allele	other allele	effect size
C	**A**	**0.5**

(III) When the variant in exposure and outcome do not share the same allelic information, they where discarded from the data set:

outcome

effect allele	other allele	effect size
A	**G**	**0.5**

exposure

effect allele	other allele	effect size
C	**A**	**0.5**

TWO-SAMPLES MENDELIAN RANDOMIZATION

The basic principle that is deployed in MR is that genetic variants either amend the level of, or mirror the biological effects of, an exposure that in turn alters the disease risk, should be associated with the disease risk.[34]

Two-samples MR is applied to identify the factors that cause or protect for celiac disease, to determine potential causal or protective relationships more than 500.000 clinical parameters and molecular phenotypes have

been screened with the MR approach.

The two-samples MR Wald methods has been applied to calculate causal estimate between the variants in exposure and outcome (equation 1). Multiple genetic variants have then been combined using IVW (equation 3) and MR-Egger (equation 6) to increase power and estimate the overall effect.

COCHRAN'S Q TEST

To warrant the combined causal estimate from the IVW method the Cochran's Q test was used. Cochran's Q assesses the homogeneity by testing the chi-square distribution from the association between exposure and outcome by computing a Q. term for every SNP-SNP association between exposure and outcome.

The sum of the Q terms is taken and tested by the chi-square (χ^2) test, the null hypothesis of variants being homogeneous is rejected when the corresponding p-value of the test is smaller than .05. If the null hypothesis is rejected, then the max Q term will be removed, and the IVW causal estimate will be re-calculated with the remaining genetic variants. This is recursively executed until the test is not significant anymore (indicating lack of heterogeneity).[36]

When Cochran's Q method was significant a subset of variants that deviates from the average causal estimate, and another subset of variants that identifies homogeneous variants, were identified. On the subset that deviates from the average causal estimate, the IVW, and MR-Egger methods have been executed again. This test is used only for IVW because MR-Egger has an incorporated test for heterogeneity (the intercept).[5]

PATHWAY SCORING ALGORITHM

When between the exposure ~ outcome association Cochran's Q was significant two distinct subsets of data sets where identified: one that deviates from the average causal estimate, and another that leads to homogeneous estimates.

Both subsets are associated with the same exposure, however they may act through different biological pathways. To better understand the link between the exposure and the outcome, the biological difference between these genetic variant subsets is accessed by Pascal.

Pascal is a powerful tool that computes gene and pathway scores from SNP-phenotype summary statistics. The goal of pathway analysis is to gain insight into the biological processes by the aggregation of a collection of observed SNP association signals into a pathway signal.[17] The pathways analysis was performed using the following Pascal command:

```
./Pascal
    --pval SNP.file
    --genescoring=sum
    --runpathway=on
```

Here --pval expects a file containing SNPs with the associated GWAS p-value, --genescoring=sum defines the genescoring method, the default is sum, and --runpathway should be turn on if Pascal needs to calculate

pathway scores. Pascal gives pathways associated with the given genetic variants, with a p-value describing the significance of the association as output.

TWO-SAMPLES BIDIRECTIONAL MENDELIAN RANDOMIZATION

To further investigate the direction of the relationship between exposure and outcome, two-samples bidirectional MR was performed to evaluate all possible relationships between the phenotypes.

This was done by swapping the exposure and outcome, thus outcome becomes the exposure, and the exposure becomes the outcome and all previously steps described in performing an MR analysis where conducted again.

RESULTS

There have been analyzed 511.344 exposures against celiac diseases using a bi-directional MR approach. All summary statistics file underwent several data pre-processing and harmonization steps, as described in the methods section. After harmonization, it was possible to analyze 149.758 exposure as potential risk factor for Celiac disease, and 9.154 exposure as potential consequences of the disease (reverse MR), for the full tables see: https://bioinf.nl/~mknigge/index.html

TWO-SAMPLES MENDELIAN RANDOMIZATION

Two-samples MR is applied to all remaining traits to identify factors that cause or protect for celiac disease. Two-samples MR Wald methods has been applied to calculate causal estimates between genetic variants in exposure and outcome. Then, multiple genetic variants have been combined with IVW and MR-egger method to increase power and estimate the overall effect.

For example, the MR analysis celiac disease Trynka (outcome)~ thyroid gland disorders (exposure). Both exposure and outcome contain 27 genetic variants, both with different effect sizes, and between these is the causal effect estimated by the Wald method, and the overall effect is obtained by IVW, and MR-egger, see figure 5. This figure shows the genetic variants from exposure and outcome plotted against each other. This MR analysis indicates that when there is an increased risk for thyroid glands disorders that the risk for celiac disease increases. And the MR-Egger intercept is not significant different than zero, meaning that the IVW estimate can be trusted.

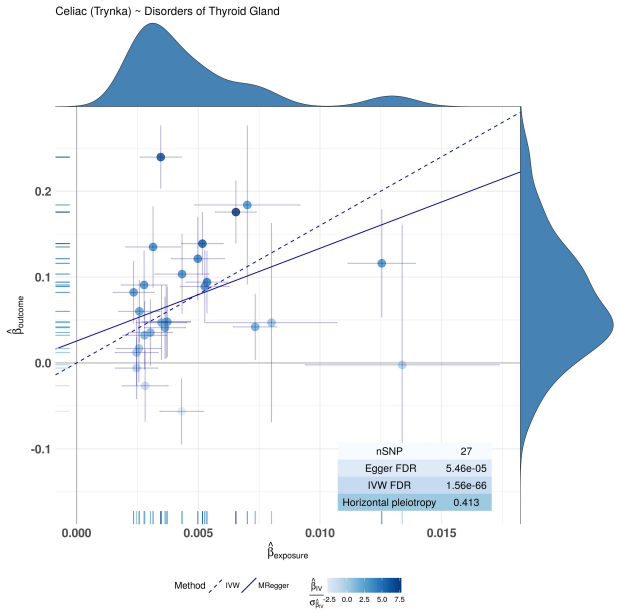

Figure 5: MR analysis celiac ~ disorders of thyroid glands. The dots are the genetic variants. The color gradients shows the Z score, the confidence interval around the genetic variants is denoted, the dotted line is the IVW estimate, the solid line is the MR-Egger estimate. Around the plot is the density drawn. In the table is denoted: amount of SNPs, significance of the IVW and Egger estimate, and the significance of the MR-Egger intercept. The x-axis shows the exposure, and the y-axis the outcome. This MR analysis describes that a higher risk for disorders of thyroid glands causes a higher risk for celiac.

If the causal relationship between celiac ~ hypertensive disease is investigated, the same process is undertaken as described before, see figure 6. In total between outcome and exposure are 21 genetic variants present. Here both, the causal estimate from MR-egger and IVW are significant, and the direction indicates that a higher risk for hypersensitive disease increases the risk for celiac disease. And that the MR-Egger intercept is not significant, meaning that the analysis does not

suffer from pleiotropic effects.

In evaluating celiac disease ~ melanoma (figure 7), the IVW method implies that there is a protective relationship between celiac and Melanoma, that a lower risk for Melanoma protects for celiac disease. Of note, the MR-Egger estimate (the slope of the MR-egger) is not significant, there is not evidence for pleiotropy (intercept of MR-Egger). This discrepancy of results can be attributable to the small number of variants and the different power and sensitivity of the two statistical tests.

This approach has been applied to all traits between celiac (Trynka, and Dubois) and exposures.

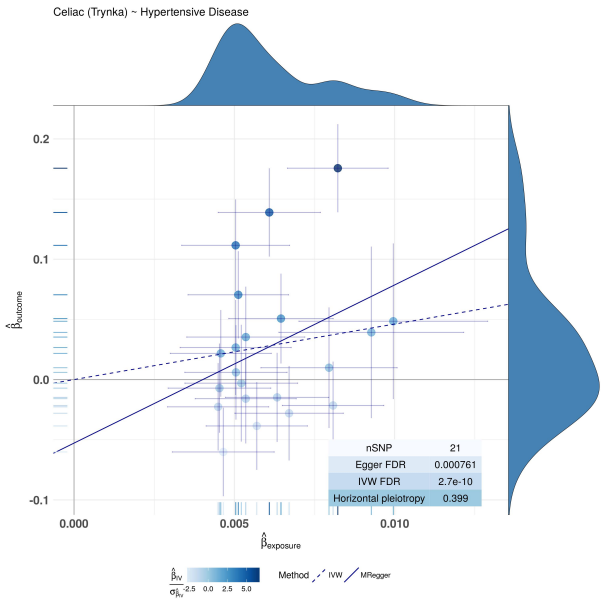

Figure 6: MR analysis celiac disease ~ hypertensive disorders. Showing that an increased risk for Hypersensitive disease increases the risk for celiac disease. Both, MR-Egger and IVW are significant, and the MR-Egger intercept is not significant meaning that the IVW estimate can be trusted.

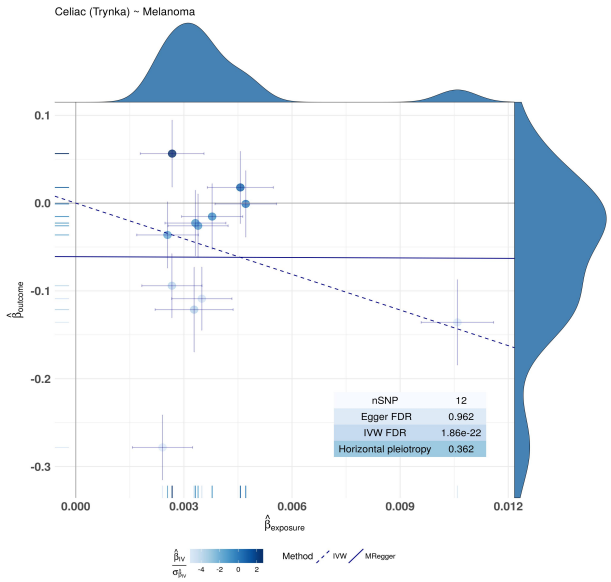

Figure 7: MR analysis celiac disease ~ melanoma. Showing a protective relationship. Here the IVW estimate is significant but the MR-Egger intercept not. But the IVW estimate can be trusted because the intercept from the MR-Egger is not significant non-zero.

COCHRAN'S Q TEST

The IVW method does not include a test for pleiotropy, but sensitivity analyses can be performed to assess robustness of the combined result. Here we used the Cochran's Q test to assesses the homogeneity of causal estimates derived from each single variant with the Wald test (which are then combined with IVW method into a overall estimate). The Cochran's Q test, when significant, indicates presence of homogeneity. In this case, we identify the variant corresponding to the most extreme causal estimate compared to the combined IVW estimated. We then remove this variant from the set of variants associated with the exposure, and we recalculate again the IVW and Cochran's Q test. We repeat this recursively until the Cochran's Q test is not significant anymore, thus suggesting the absence of heterogeneity. With this procedure,

a subset of genetic variants is identified that deviates from average causal estimate, and another subset of genetic variants that identifies homogeneous variants.

When applying Cochran's Q test on celiac ~ thyroid gland disorders it identifies 22 genetic variants as diverging from the average causal estimate and five genetic variants as homogeneous, see figure 8.

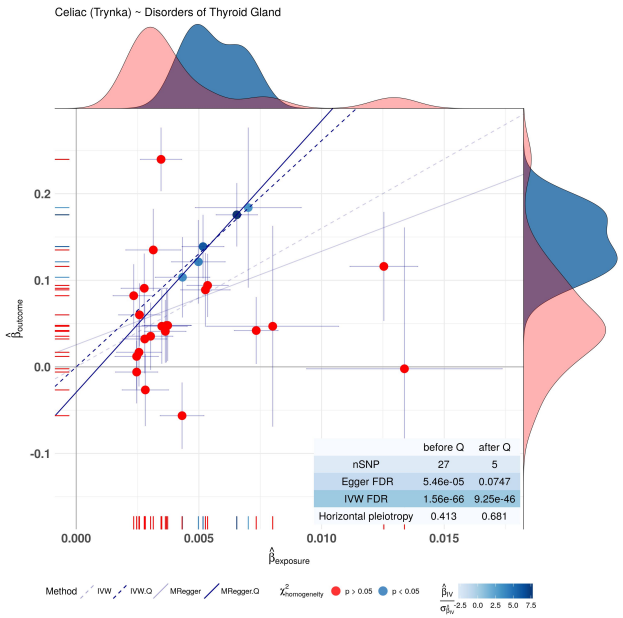

Figure 8: MR analysis celiac disease ~ thyroid gland disorders. In blue are the genetic variants that are defined as homogeneous by Cochran's Q, and in red are the genetic variants that deviate from the average causal estimate. In transparent blue are the causal estimates from both MR-Egger and IVW before Cochran's Q test, and in non-transparent the causal estimates after Cochran's Q test. The MR-Egger becomes non-significant, and the IVW estimate stays significant, and no form of pleiotropy is introduced by inspecting the intercept from MR-Egger. And, the direction of the relationship stays the same.

In blue are the genetic variants that are considered homogeneous by Cochran's Q test, and in red are the ones that are considered deviating from the causal estimate. The transparent blue lines shows both the MR-Egger and IVW method before Cochran's Q test, and the non-transparent lines the causal estimates after applying Cochran's Q test.

The true direction of the relationship was not masked by genetic variants that deviate from the average causal estimate, that may act through an other biological pathway. The significance is lowered for the IVW method but still significant, and the MR-Egger is changed to non-significant, and the intercept test from the MR-Egger method has changed to being closer to zero, meaning that the IVW estimate can be trusted.

In the MR analysis celiac disease ~ hypersensitive disease (figure 9) Cochran's Q test identified eight genetic variants as homogeneous and 13 genetic variants as deviating from the average causal estimate. The causal estimate from IVW is still significant, but from MR-Egger not, and by visual inspection it shown that the true identify of the direction was masked by genetic outliers. The MR analysis after Cochran's Q test shows a protective relationship instead of causal. This change in direction is suspicious and further analyses are therefore needed to clarify the role of hypertension in Celiac Disease.

Cochran's Q test in celiac disease ~ melanoma (figure 10) identifies three genetic variants as homogeneous and nine as deviating from the average causal estimate.The direction of the causality is not changed, and the estimate from IVW is yet significant, the estimate from

MR-Egger is not changed, still non-significant but the intercept from MR-Egger indicates that the IVW causal estimate can be trusted.

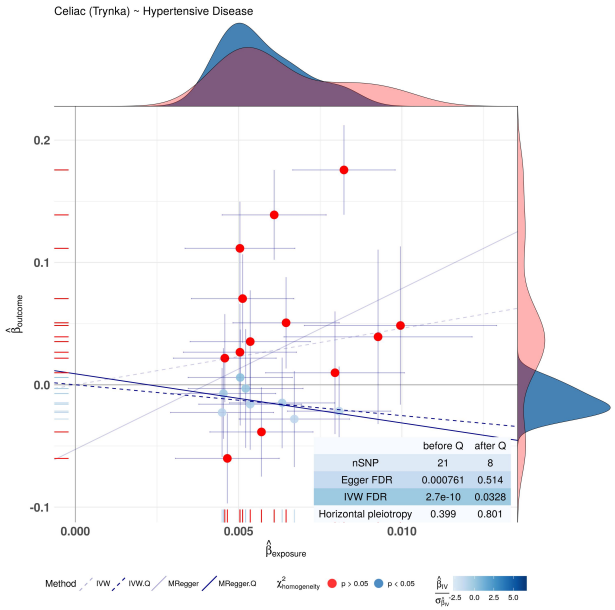

Figure 9: MR analysis celiac disease ~ hypertension disease after applying Cochran's Q test. Eight genetic variants are defined as homogeneous and 13 as deviating from the average causal estimate. By pruning genetic outliers it is shown that the causality is driven by genetic outliers. Instead of having a causal relationship, it is changed to a protective relationship.

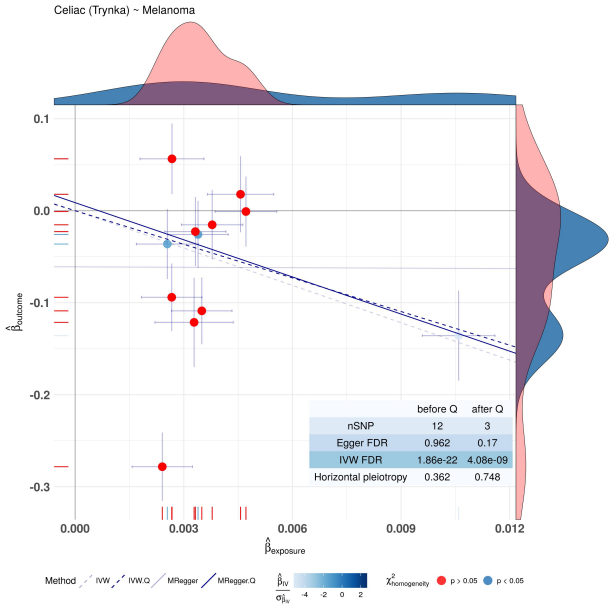

Figure 10: Cochran's Q test prunes nine SNPs as deviating from the causal estimate, and three as being homogeneous. This causes no change in direction of the causal relationship.

TWO-SAMPLES BIDIRECTIONAL MENDELIAN RANDOMIZATION

The phenotype used in the MR analysis celiac disease ~ exposure can be a consequence of the disease or can be both causal and a consequence of the disease through a feedback mechanism. In order to research the direction of the relationship between outcome and exposure two-samples bidirectional MR was performed to investigate all possible relationships between phenotypes. For the MR analysis this means that the following was performed: celiac disease ~ exposure, and exposure ~ celiac disease. This section presents the results of exposure ~ celiac disease (reverse MR).

In the MR analysis celiac disease ~ thyroid gland disorders, this has been performed in the other direction: thyroid gland disorders ~ celiac disease, see figure 11. The forward direction (celiac disease ~ thyroid gland disorders) shows that a higher risk for thyroid gland disorders increases the risk for celiac disease, and in the reverse direction (thyroid gland disorders ~ celiac disease) it shows the same relationship meaning that there is a circular feedback mechanism, it may be the cause and a consequence of the disease.

The setting celiac disease ~ melanoma and melanoma ~ celiac disease (figure 12) shows in the forward direction a relationship where a higher risk for melanoma decreases the risk for celiac disease, and in the reverse direction

that a higher risk for celiac disease increases the risk for melanoma. So here the exposure is a consequence of the outcome.

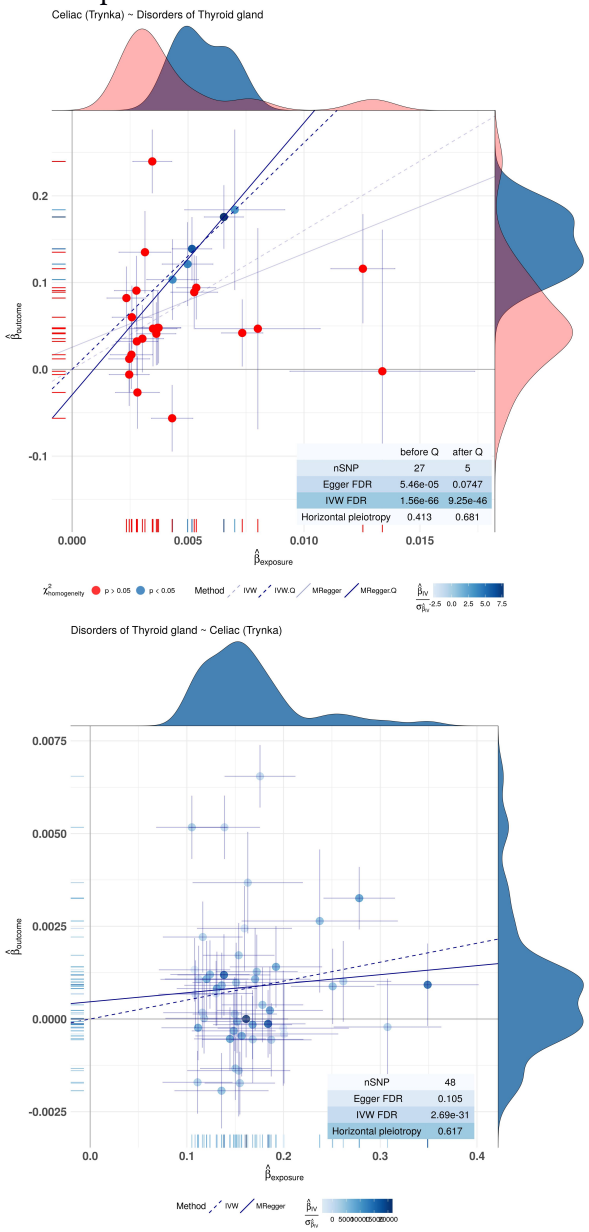

Figure 11:MR analysis in the forward direction: celiac disease ~ thyroid gland disorders, and in the reverse thyroid gland disorders ~ celiac disease. This MR analysis shows that there is circular feedback mechanisms at play, the exposure can be both, the cause and the consequence of the disease, but the reverse is non-significant..

In celiac disease ~ hypertensive diseases, and hypertensive diseases ~celiac disease (figure 13) there is a protective circular feedback mechanism. In the forward direction: a higher for hypertensive diseases lowers the risk for celiac disease. And the reverse direction shows that a higher risk for celiac disease lowers the risk for hypertensive diseases, but the reverse is non-significant..

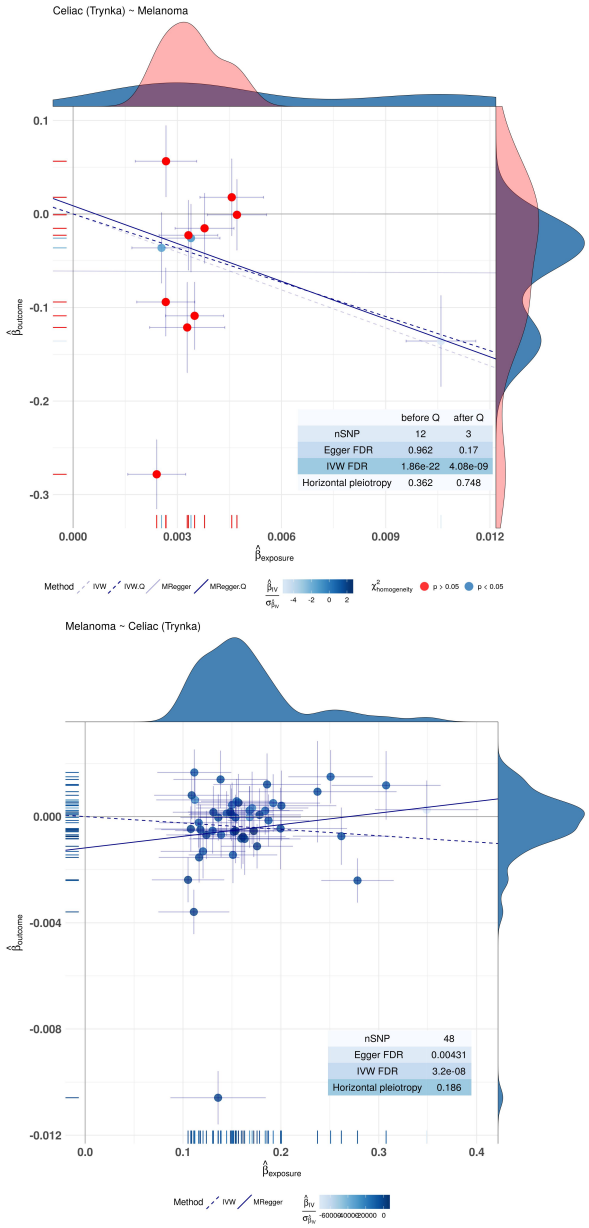

Figure 12: celiac disease ~ melanoma shows that an increased risk for melanoma lowers the risk for celiac disease, and melanoma ~ celiac disease shows that an increased risk for celiac disease increases the risk for

melanoma, making it a consequence of the outcome.

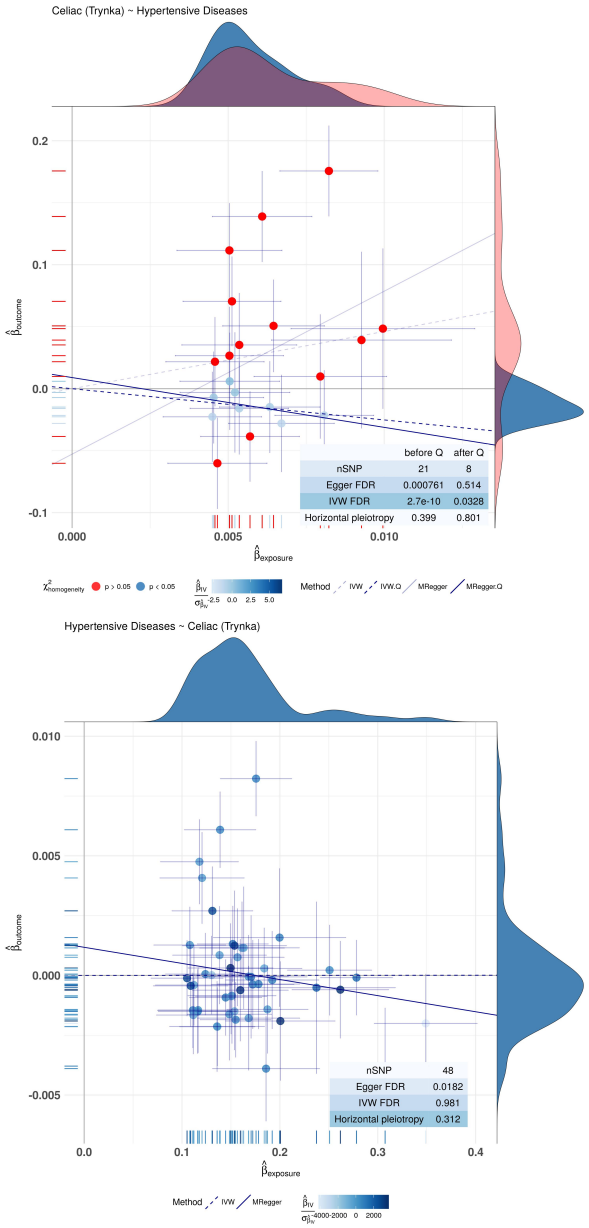

Figure 13: celiac disease ~ hypertensive disease shows that a higher risk for hypertensive disease is protective for celiac disease, and the reverse direction: hypertensive disease ~ celiac disease shows that a higher risk for celiac disease lowers the risk for hypertensive disease. Here a protective circular feedback mechanism is at play.

In total for both directions, and both outcome GWAS summary statistics, over two million times a MR analysis was conducted. To correct for multiple testing, we adjusted the p-values with FDR Benjamini–Hochberg procedure (using the function p.adjust() in R), and considered as significant tests showing a corrected p-value < 0.05. This results in 1133 significant clinical parameters that cause or protect for celiac disease, and others that show feedback mechanisms.

PATHWAY SCORING ALGORITHM

In the case when Cochran's Q test was significant, the test identified two subsets of genetic variants: one that deviates from the average causal estimate and one that leads to homogeneous estimates. Both these subsets are associated with the exposure but can act through a different biological pathway. Pascal is used to do pathways analyses to infer the role of such variants in biological mechanisms. The aim is to better understand the link between exposure and outcome by looking at the biological differences between the two sets of identified variants.

When considering the two subsets identified by Cochran's Q in celiac disease ~ thyroid gland disorder, one set of five genetic variants are identified as homogeneous and 22 as deviating from the average causal estimate. Pascal identifies multiple pathways for both sets. For the homogeneous variants the following paths were identified:

- REACTOME SIGNALING BY SCF KIT
- REACTOME REGULATION OF KIT SIGNALING
- REACTOME FACTORS INVOLVED IN MEGA KARYOCYTE DEVELOPMENT AND PLATELE T PRODUCTION

- REACTOME HEMOSTASIS

And for the genetic variants that deviate from the causal estimate:

- KEGG CELL ADHESION MOLECULES CAMS
- BIOCARTA CTLA4 PATHWAY
- REACTOME CTLA4 INHIBITORY SIGNALING
- REACTOME BETA DEFENSINS
- KEGG CYTOKINE CYTOKINE RECEPTOR INTERACTION
- BIOCARTA TOB1 PATHWAY
- KEGG SPLICEOSOME
- REACTOME COSTIMULATION BY THE CD28 FAMILY
- REACTOME DEFENSINS
- REACTOME CHEMOKINE RECEPTORS BIND CHEMOKINES
- KEGG HEMATOPOIETIC CELL LINEAGE
- BIOCARTA TH1TH2 PATHWAY

The homogeneous genetic variants are involved in pathways that signal Stem Cell Factors (SCF), which is a growth factor for promoting proliferation, migration, survival and differentiation of hematopoetic progenitors[38], pathways that signal Kit receptors which is a cytokine receptor, that plays an important role in erythropoiesis, lymphopoiesis, mast cell development, and function[39], and the development of megakaryocytes,and platelets, and hemostasis.

The genetic variants that deviate from the average causal estimate are involved in the production of cell adhesion molecules, the co-stimulatory signaling during T-cell activation[40], the inhibition of T-cell activation[41], cytokine-cytokine receptor interactions, regulation of T-cell activation[42], involvement of optimal activation of T-cells[43], blood-cell development from a hematopoetic stem cell[44], differentiation from helper T-cells in Th1 and Th2[45].

Pascal is applied to all MR results that were considered significant FDR < 0.05.

DISCUSSION

Celiac disorder affects up to two percent of the European population, and it is hereditary. The inheritance model of celiac disease is oligogenic.[1] In this model, a genetic variant in the Major Histocompatibility Complex (MHC) region, the HLA-DQ2 haplotype, accounts for 40% of the inheritable risk, but it is not sufficient enough to develop the disease. 30% of the Europeans are in fact carriers of the variant, but only 3% express celiac disease. We aimed to systematically assess multiple clinical parameters and molecular mechanisms to identify causal and protective factors, that can explain what is triggering disease manifestation in HLA-DQ2 carriers.[2]

We identified 1133 significant (figure 14) (FDR < 0.05) clinical parameters that cause or protect for celiac disease, and others that show feedback mechanisms . We found an interesting causative role for infection diseases and allergies. At the molecular level, we observed evidence of causality for changes in expression of 25 genes and for changes in methylation patterns at other 12 (FDR<0.05). Interestingly, 12 of those 37 genes are in loci not previously associated with Celiac Disease at the genome-wide level, but the majority (13) is known to be associated with other autoimmune diseases, tonsillitis and/or asthma. Those results not only strength the links that we detected with those clinical parameters, but also suggest novel candidate genetic loci to be assessed in future genetic studies.

We investigated over 500.000 clinical parameters and molecular phenotypes, for which summary statistics from genome-wide association studies (GWAS) were available. We used the concept of Mendelian Randomization[3] (MR) approach, and applied two-samples MR methods: Inverse Variance Weighted[4] (IVW) and MR-Egger method[5] to infer causality links by combining summary statistics from GWAS. For our analyses, we discarded variants located in the MHC region to avoid over-estimation of causality for factors that also influenced by the same or other HLA haplotypes.

ABNORMAL ERYTHROPOIESIS

The significant results in abnormal erythropoiesis like haemoglobin concentration, shows that it a is a consequence of the disease, see figure 15. Which is described by Brusco et al[48], they found that the increase of haemoglobin is an indicator of celiac disease, and that it can be considered as a new predictor of celiac disease.

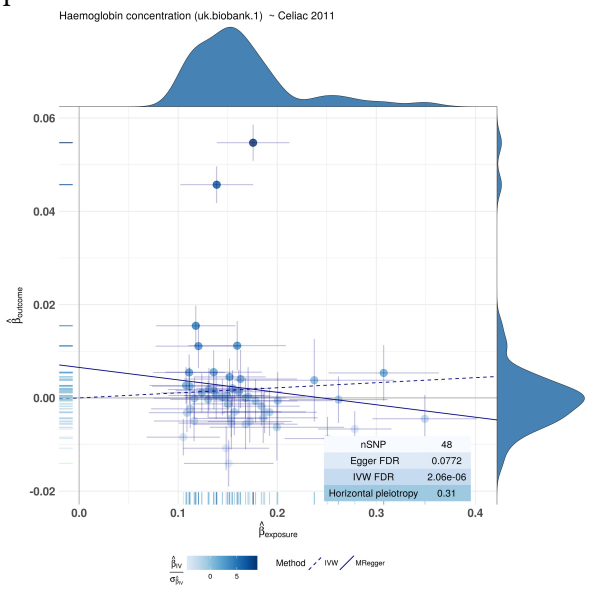

Figure 15: haemoglobin~celiac disease shows that a higher risk for haemoglobin increases the risk for celiac .

ANEMIA

Anemia is a condition were there are not enough healthy red blood cells to transport oxygen to tissues, and it is a secondary for malabsorption of iron, folic acid, and vitamin B12. Halfdanarson et al[49] describes that celiac disease may be associated with anemia as a consequence. These findings are consistent with multiple MR analysis on independent GWAS summary statistics where it is shown that anemia is a consequence of the disease, see appendix C. The traits used are: treatment medication: folic acid product, other anaemias, treatment medication: vitamin B12, anemia, diagnose: iron deficiency, diagnose: other anaemias, and vitamin B12 deficiency anemia.

THYROID GLAND DISORDERS

Thyroid Gland disorders are one of the most common autoimmune disorders were an autoimmune-mediated destruction of the thyroid glands takes place. This disorder has an increase presence in patients diagnosed with celiac disease, and vice versa.[50]

By Lerner et al[50], the relationship between the autoimmune disorder Thyroiditis and celiac disease is described as a co-occurring phenomena, were circular feedback is present, that in both directions increases the risk for having the phenotype as a disease[50], which is consistent with results from MR. In the MR analysis celiac disease ~ thyroid disorders it is shown that there is a feedback mechanisms that increases the risk for having the disease. This pattern is confirmed in multiple independent GWAS studies, in both directions, from the UKBiobank there were extracted three GWAS summary statistics that regressed the genetic data on the phenotype thyroid disorders. For accessing both directions for the three GWAS summary statistics six times a MR analysis has been performed, see appendix B. The results describe for all summary statistics that having celiac disease increases the risk for having thyroid disorders, and that having thyroid disorders gives a higher risk for having celiac disease.

HYPERTENSION DISEASE

Hypertension disease refers to the increase in blood pressure. Lim et al[51] describes that observations suggested that celiac disease might cause endothelial dysfunction giving rise to a form of hypertension disease, and it is proven in an observational study that treatment of celiac disease is associated with reduction of endothelial function, which "cures" hypertension. Here is expected to find a causal link between hypertensive ~ celiac where a higher risk for celiac disease causes hypertensive, or that a lower risk for celiac disease protects for hypertension, which is confirmed here, by multiple traits related to hypertension: blood pressure medicatior, treatment medication: ozaar 25mg tablet, diastolic blood pressure, and medication treatment: doxazosin, see appendix D.

THROMBOCYTOSIS

Thrombocytosis (high platelets counts) is a component of blood with the function to stop bleeding by clumping and clotting blood vessels. The effect from thrombocytopenia (low platelets counts) upon celiac disease is rarely been reported. However treating celiac disease can restore platelet count in some cases.

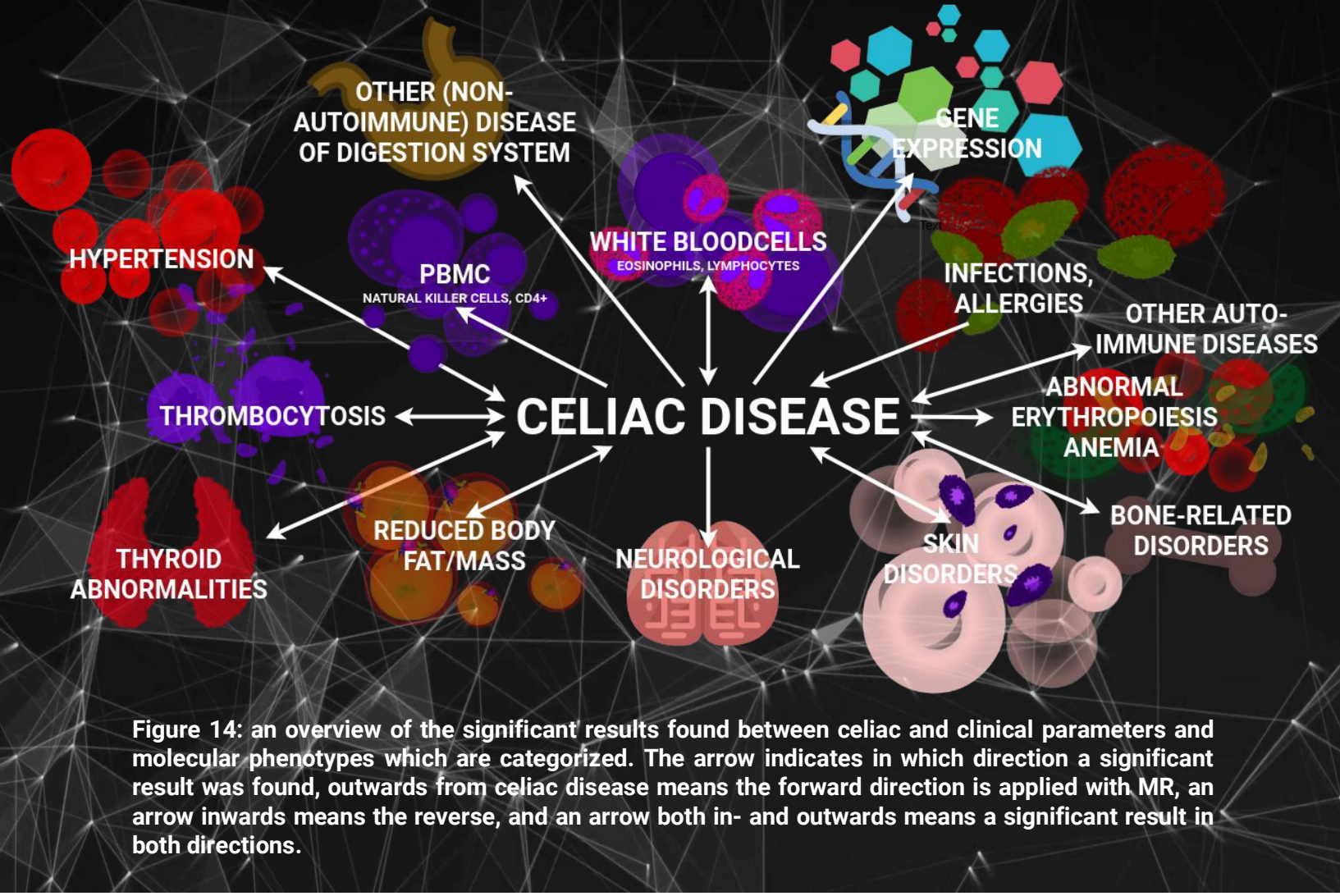

Figure 14: an overview of the significant results found between celiac and clinical parameters and molecular phenotypes which are categorized. The arrow indicates in which direction a significant result was found, outwards from celiac disease means the forward direction is applied with MR, an arrow inwards means the reverse, and an arrow both in- and outwards means a significant result in both directions.

Thrombocytosis in association with celiac disease is more common and reported but the origin of causation is unknown. But the thrombocytosis can disappear after treating celiac disease.[52] Which suggests that a lower risk for celiac disease lowers the risk for thrombocytosis, see figure 18.

CD4+ T-CELLS

OTHER AUTOIMMUNE DISEASES

EOSINOPHILS AND LYMPHOCYTES

INFECTION DISEASES

GENE EXPRESISON AND METHYLATION PATTERNS

CONCLUSION

REFERENCES

1. Robert Di Niro, et al. High abundance of plasma cells secreting transglutaminase 2-specific IgA autoantibodies with limited somatic hypermutation in celiac disease intestinal lesions. Nature Medicine 18, 441-445 (2012). https://doi.org/10.1016/S0016-5085(98)70008-3

2. Gujral, N., Freeman, H. J., & Thomson, A. B. (2012). Celiac disease: Prevalence, diagnosis, pathogenesis and treatment. World Journal of Gastroenterology : WJG, 18(42), 6036–6059. http://doi.org/10.3748/wjg.v18.i42.6036

3. David M. Evans and George Davey Smith. Mendelian Randomization: New Applications in the Coming Age of Hypothesis-Free Causality. Annual Review of Genomics and Human Genetics 16, 327-350. http://www.annualreviews.org/doi/10.1146/annurev-genom-090314-050016

4. Debbie A. Lawlor. Commentary: Two-sample Mendelian randomization: opportunities and challenges. International Journal of Epidemiology, Volume 45, Issue 3, 1 June 2016, Pages 908–915, https://doi.org/10.1093/ije/dyw127

5. Stephan Burgess, Simon G. Thompson. Interpreting findings from Mendelian randomization using the MR-Egger method. S.G. Eur J Epidemiol (2017) 32: 377. https://doi.org/10.1007/s10654-017-0255-x

6. Jerry S, Trier, M.D. Celiac Sprue. N Engl J Med 1991; 325:1709-1719December 12, 1991 http://www.nejm.org/doi/full/10.1056/NEJM199112123252406

7. George Davey Smith, Gibran Hemani. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Human Molecular Genetics, Volume 23, Issue R1, 15 September 2014, Pages R89–R98, https://doi.org/10.1093/hmg/ddu328

8. Caroline L Relton, George Davey Smith. Two-step epigenetic Mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease. International Journal of Epidemiology, Volume 41, Issue 1, 1 February 2012, Pages 161–176, https://doi.org/10.1093/ije/dyr233

9. University of BRISTOL. (2017). MR Methods: Single-sample MR. http://www.bristol.ac.uk/integrative-epidemiology/mr-methods/introduction-to-mr/single-sample-mr/ [accessed: December 15, 2017]

10. Scitable by Nature Education. (2017). Pleiotropy: One Gene Can Affect Multiple Traits. https://www.nature.com/scitable/topicpage/pleiotropy-one-gene-can-affect-multiple-traits-569 [accessed: December 15, 2017]

11. Slatkin M. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. Nature reviews Genetics. 2008;9(6):477-485. http://doi.org/10.1038/nrg2361

12. David C. Hoaglin. Misunderstandings about Q and 'Cochran's Q test' in meta-analysis. Statistics in Medecine, Volume 35, Issue 4 20 February 2016 Pages 485–495. https://doi.org/10.1002/sim.6632

13. Stat Trek. (2017). Chi-Square Distribution. http://stattrek.com/probability-distributions/chi-square.aspx [accessed December 17, 2017]

14. NCBI. (2017) Chi-Square Distribution. https://www.ncbi.nlm.nih.gov/mesh?Db=mesh&Cmd=DetailsSearch&Term=%22Chi-Squar

e+Distribution%22%5BMeSH+Terms%5D [accessed December 17, 2017]
15. Philiph Sedwick. Meta-analyses: what is heterogeneity? BMJ 2015; 350 doi: https://doi.org/10.1136/bmj.h1435
16. Elena Kulinskaya, Michael B. Dollinger, Kirsten Bjørkestøl. Testing for Homogeneity in Meta-Analysis I. Biometrics, journal of the international biometric society, The One-Parameter Case: Standardized Mean Difference. Testing for Homogeneity in Meta-Analysis I. The One-Parameter Case: Standardized Mean Difference. http://dx.doi.org/10.1111/j.1541-0420.2010.01442.x
17. Lamparter D, Marbach D, Rueedi R, Kutalik Z, Bergmann S (2016) Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. PLoS Comput Biol 12(1): e1004714. doi:10.1371/journal.pcbi.1004714
18. Amy E. Taylor, Stephan Burgess, Jennifer J. Ware, et al. Investigating causality in the association between 25(OH)D and schizophrenia. Scientific Reports **6**, Article number: 26496 (2016). https://www.nature.com/articles/srep26496
19. R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.. [online] available from: https://www.r-project.org/about.html [accessed: December 25, 2017]
20.Patrick C A Dubois, Gosia Trynka, Lude Franke, et al. Multiple common variants for celiac disease influencing immune gene expression. Nature Genetics 42, 295–302 (2010). https://www.nature.com/articles/ng.543
21.Gosia Trynka, Hunt Karen A, Bockett, Nicholas A, et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nature Genetics 43, 1193–1201 (2011). https://www.nature.com/articles/ng.998
22. Gene ATLAS (2017) The Rosalin Institute, University of Edinburgh. [online] available from: http://geneatlas.roslin.ed.ac.uk/downloads/ [accessed: November 14, 2017]
23. UK Biobank GWAS Results (2017) Neale Lab. [online] available from: https://sites.google.com/broadinstitute.org/ukbbgwasresults/home?authuser=0 [accessed: November 17, 2017]
24. GWAS catalog (2017). The NHGRI-EBI Catalog of published genome-wide association studies. [online] available from: https://www.ebi.ac.uk/gwas/home [accessed: January 2, 2018]
25. GTExPortal (2017) Dataset Summary of Analysis Samples of the V7 Release. [online] available from: https://www.gtexportal.org/home/ [accessed: November 20, 2018]
26. mQTL Database (2017) mQTLdb Large-scale genome-wide DNA methylation analysis of 1,000 mother-child pairs at serial time points across the life-course (ARIES).[online] available from: http://www.mqtldb.org/ [accessed: November 25, 2018]
27. Duggal P, Gillanders EM, Holmes TN, Bailey-Wilson JE. Establishing an adjusted p-value threshold to control the family-wide type 1 error in genome wide association studies. BMC Genomics. 2008;9:516. doi:10.1186/1471-2164-9-516.

28. IGSR: The International Genome Sample Resource (2014). Providing ongoing support for the 1000 Genomes Project data. [online] available from: http://www.internationalgenome.org/category/phase-3/ [accessed: September 2, 2017]
29.VanderWeele, T. J., Tchetgen Tchetgen, E. J., Cornelis, M., & Kraft, P. (2014). Methodological challenges in Mendelian randomization. Epidemiology (Cambridge, Mass.), 25(3), 427–435. http://doi.org/10.1097/EDE.0000000000000081
30. Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, et al. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. AJHG Volume 81, Issue 3, September 2007, Pages 559-575. https://doi.org/10.1086/519795
31. Shi H, Medway C, Brown K, Kalsheker N, Morgan K. Using Fisher's method with PLINK "LD clumped" output to compare SNP effects across Genome-wide Association Study (GWAS) datasets. International Journal of Molecular Epidemiology and Genetics. 2011;2(1):30-35.
31. PLINK (2017). Report postprocessing. LD-based result clumping. [online] available from https://www.cog-genomics.org/plink/1.9/postproc#clump [accessed: November 12, 2017]
32. Alastair J. Noyce MRCP, Mike A. Nalls PhD. (2015). Mendelian Randomization — the Key to Understanding Aspects of Parkinson's Disease Causation? Movement Disorders, Volume 31, Issue 4 April 2016 Pages 478–483. http://dx.doi.org/10.1002/mds.26492
34. Smith GD, Ebrahim S. Mendelian Randomization: Genetic Variants as Instruments for Strengthening Causal Inference in Observational Studies. In: National Research Council (US) Committee on Advances in Collecting and Utilizing Biological Indicators and Genetic Information in Social Science Surveys; Weinstein M, Vaupel JW, Wachter KW, editors. Biosocial Surveys. Washington (DC): National Academies Press (US); 2008. 16. Available from: https://www.ncbi.nlm.nih.gov/books/NBK62433/
35. Gough SC., Simmonds M. The HLA Region and Autoimmune Disease: Associations and Mechanisms of Action. Current Genomics. 2007;8(7):453-465. doi:10.2174/138920207783591690.
36. Tallarida R.J., Murray R.B. (1987) Chi-Square Test. In: Manual of Pharmacologic Calculations. Springer, New York, NY. ISBN: 978-1-4612-4974-0. https://doi.org/10.1007/978-1-4612-4974-0_43
37. UMCG (2017). Organisatie. [online] available from: https://www.umcg.nl/NL/UMCG/overhetumcg/organisatie/Paginas/default.aspx [accessed: November 12, 2017]
38. Gene Set Enrichment Analysis (2017). Gene Set: REACTOME_SIGNALING_BY_SCF_KIT. [online] available from: http://software.broadinstitute.org/gsea/msigdb/cards/REACTOME_SIGNALING_BY_SCF_KIT [accessed: 11 January, 2018]
39. Gene Set Enrichment Analysis (2017). Gene Set: Gene Set: REACTOME_REGULATION_OF_KIT_SIGNALING. [online] available from: http://software.broadinstitute.org/gsea/msigd

b/cards/REACTOME_REGULATION_OF_KIT_SIGNALING [accessed: 11 January, 2018]
40.Gene Set Enrichment Analysis (2017). Gene Set: BIOCARTA_CTLA4_PATHWAY. [online] available from: http://software.broadinstitute.org/gsea/msigdb/geneset_page.jsp?geneSetName=BIOCARTA_CTLA4_PATHWAY [accessed: 11 January, 2018]
41.Reactome (2018). CTLA4 inhibitory signaling. [online] available from: http://reactomerelease.oicr.on.ca/content/detail/R-CFA-389513 [accessed: 11 January, 2018]
42. Gene Set Enrichment Analysis (2017). Gene Set: Gene Set: BIOCARTA_TOB1_PATHWAY. [online] available from: http://software.broadinstitute.org/gsea/msigdb/cards/BIOCARTA_TOB1_PATHWAY [accessed: 11 January, 2018]
43. Wikipathways (2018) Costimulation by the CD28 family (Homo sapiens) [online] available from: https://www.wikipathways.org/index.php/Pathway:WP1799 [accessed: 11 January, 2018]
44. KEGG (2017) KEGG pathway: Hematopoietic cell lineage [online] available from:http://csbi.ltdk.helsinki.fi/anduril/tcga-gbm/table4_rec/hsa04640.html [accessed: 11 January, 2018]
45.William Hagopian, Hye-Seung Lee, Edwin Liu, et al. Co-occurrence of Type 1 Diabetes and Celiac Disease Autoimmunity the TEDDY Study Group Pediatrics Nov 2017, 140 (5) e20171305; DOI: 10.1542/peds.2017-1305
46. William J. Astle, Heather Elding, Tao Jiang, etal. Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease, Cell, Volume 167, Issue 5, 2016, Pages 1415-1429.e19, ISSN 0092-8674, https://doi.org/10.1016/j.cell.2016.10.042
47. Westra, Harm-Jan, Peters, Marjolein J, Esko, Tõnu et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nature Genetics 45, 1238–1243 (2013). doi:10.1038/ng.2756
48.G. Brusco, M. Di Stefano, G.R. Corazza, Increased red cell distribution width and coeliac disease,Digestive and Liver Disease, Volume 32, Issue 2, 2000, Pages 128-130, ISSN 1590-8658, https://doi.org/10.1016/S1590-8658(00)80399-0.
49.
50.

APPENDIX

A. ABSTRACT(DUTCH)

B. THYROID DISORDERS

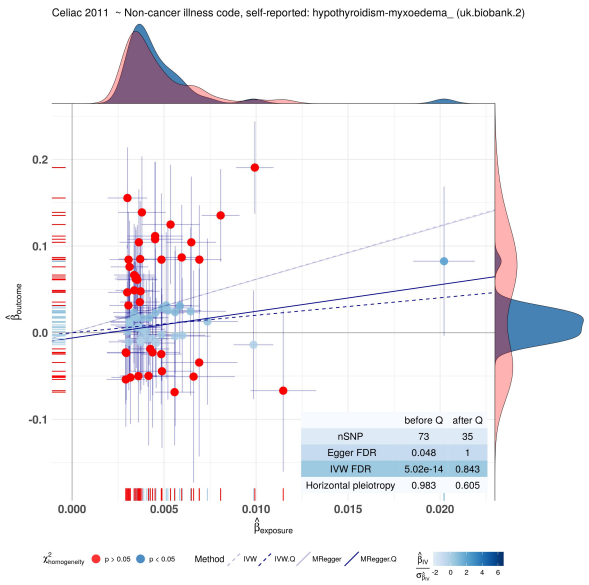

Figure B1: celiac disease ~ thyroid disorders explains that having celiac disease increases the risk for having thyroid disorders.

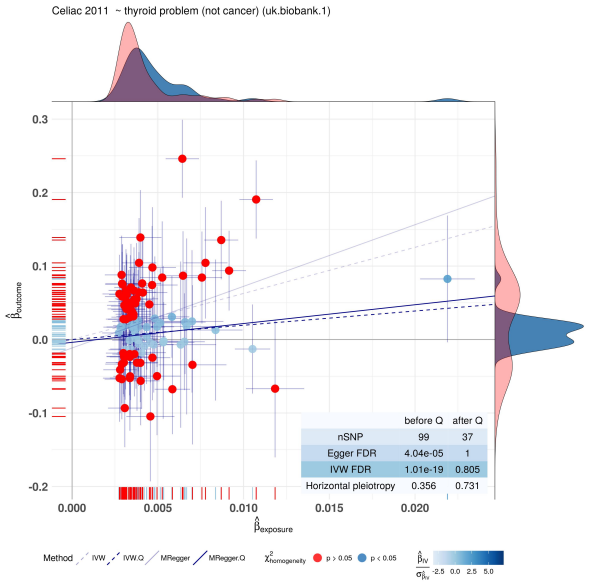

Figure B2: celiac disease ~ thyroid disorders explains that having celiac disease increases the risk for having thyroid disorders.

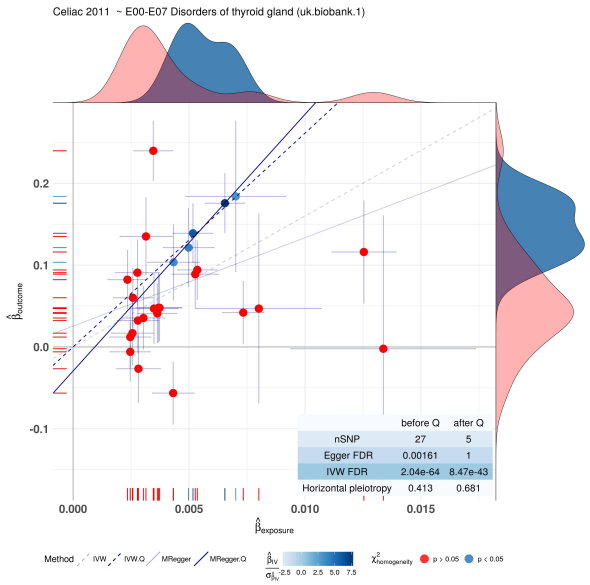

Figure B3: celiac disease ~ thyroid disorders explains that having celiac disease increases the risk for having thyroid disorders.

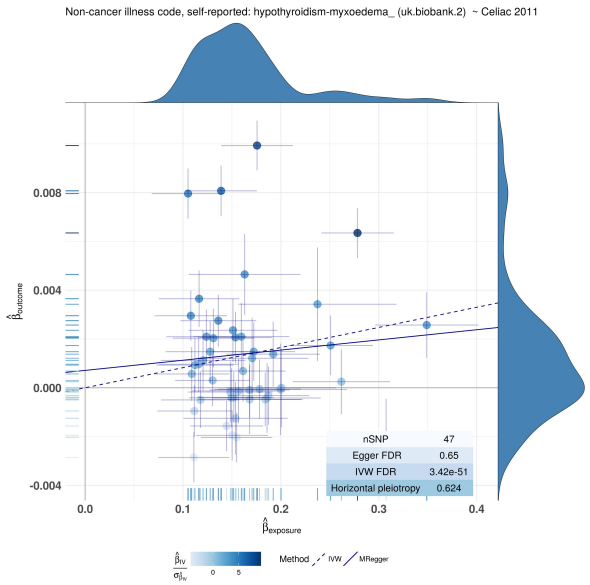

Figure B4: thyroid disorders ~ celiac disease describes that having thyroid disorders increases the risk for having thyroid disorders.

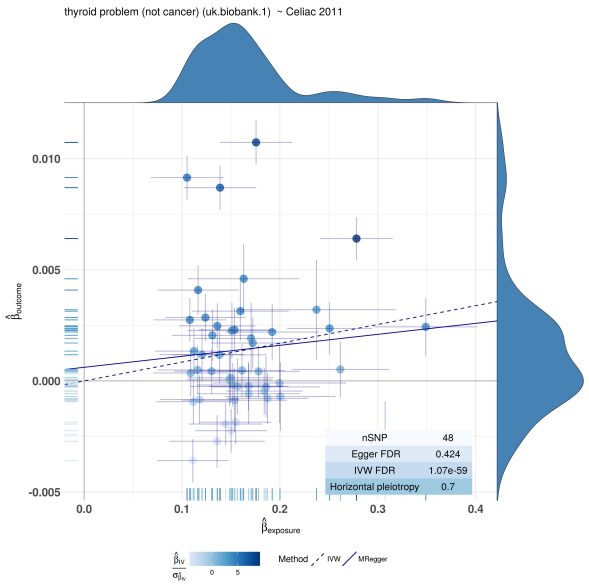

Figure B5: thyroid disorders ~ celiac disease describes that having thyroid disorders increases the risk for having thyroid disorders.

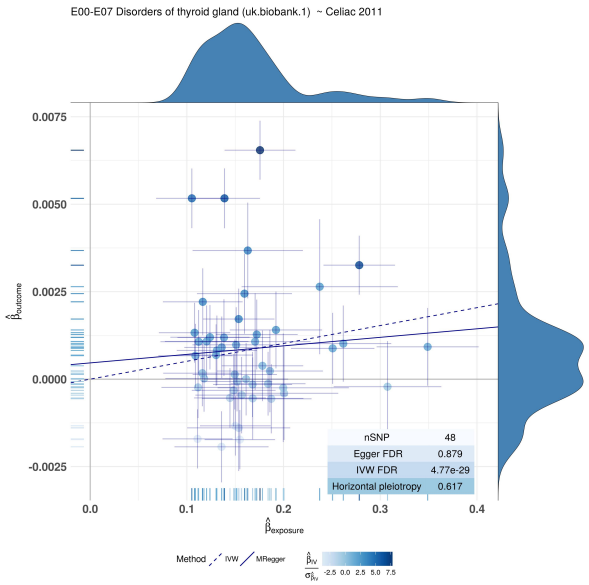

Figure B6: thyroid disorders ~ celiac disease describes that having thyroid disorders increases the risk for having thyroid disorders.

C. ANEMIA

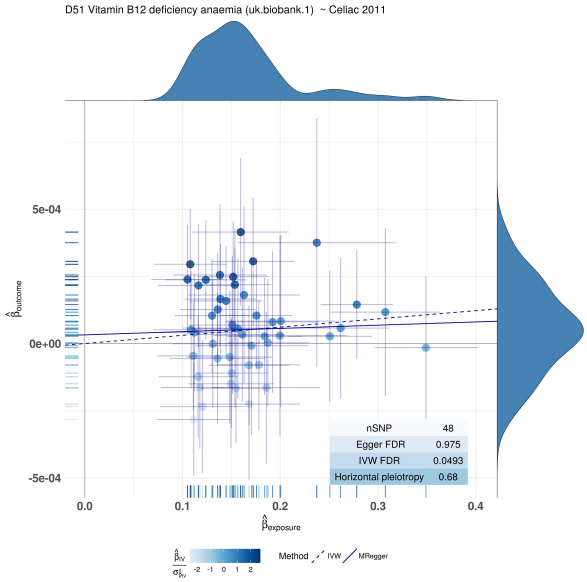

Figure C1: vitamin B12 deficiency ~ celiac disorder shows that an increased risk for celiac increases the risk for anemia.

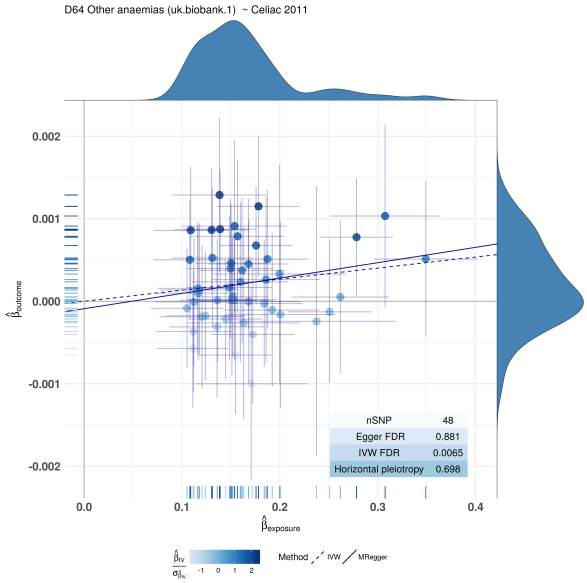

Figure C2: anemia ~ celiac disease shows that an increased risk for celiac increases the risk for anemia.

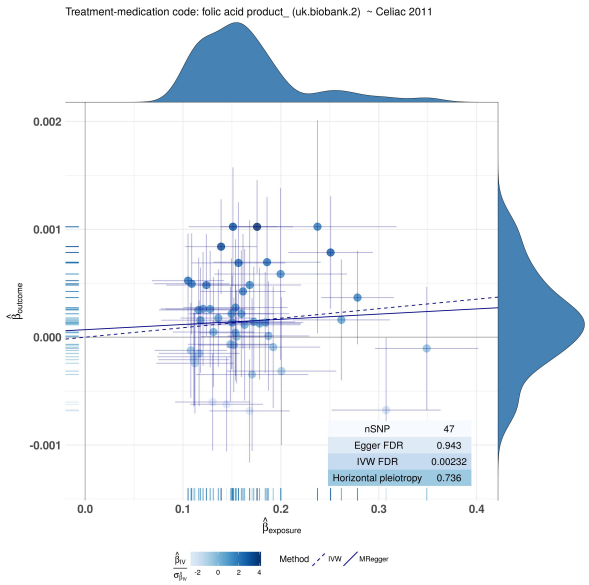

Figure C3: folic acid ~ celiac disease explains that a higher risks for celiac disease increases the level of folic acid that affects anemia.

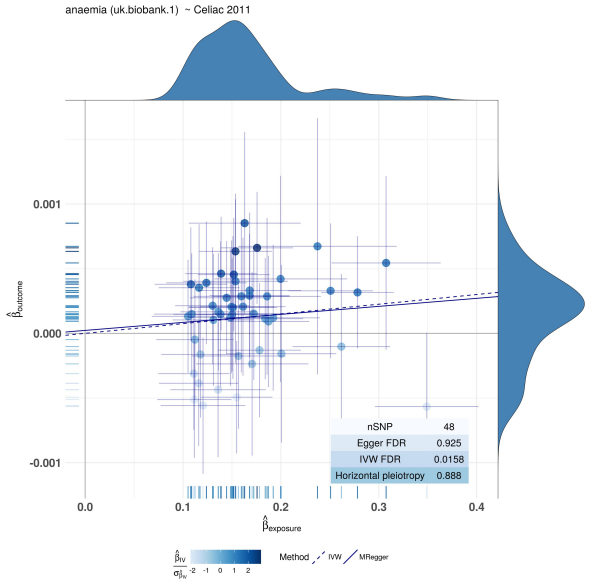

Figure C4: anemia ~ celiac disease shows that an increased risk for celiac increases the risk for anemia.

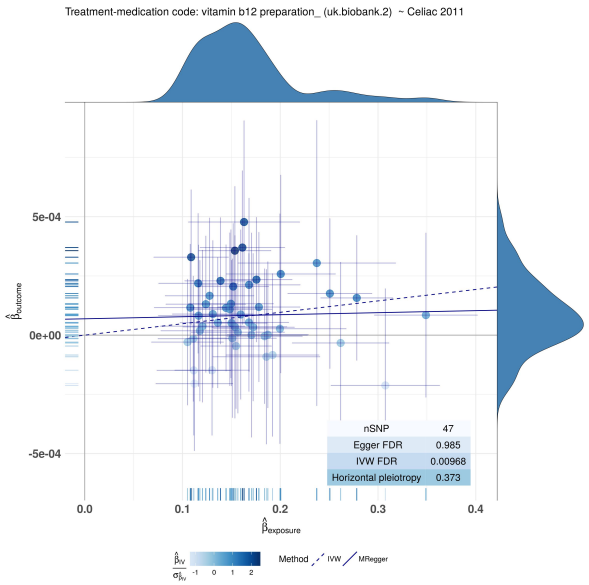

Figure C5: vitamin B12 ~ celiac disease explains that a higher risks for celiac disease increases the level of folic acid that affects anemia.

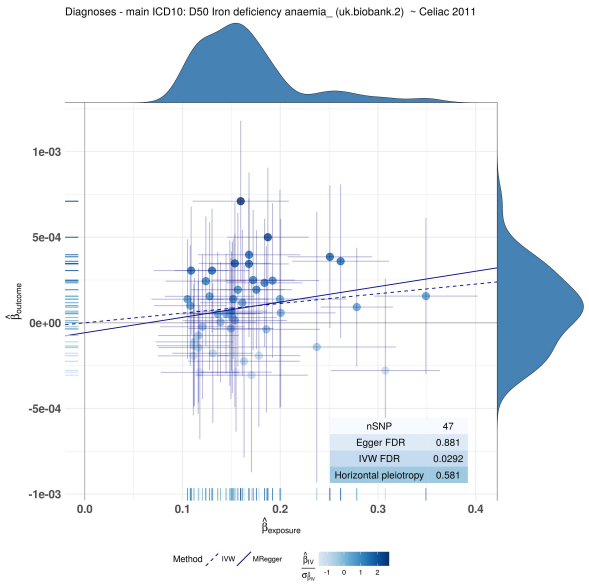

Figure C6: Iron deficiency ~ celiac disease explains that a higher risks for celiac disease increases the level of folic acid that affects anemia.

D. HYPERTENSION DISEASE

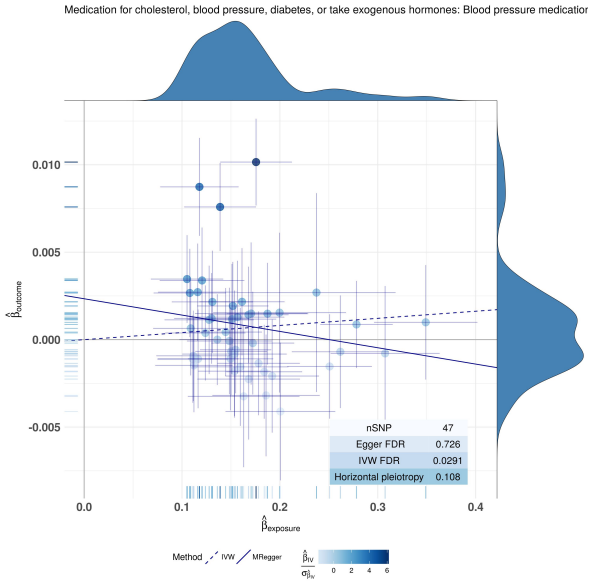

Figure D1: medication treatment ~ celiac disease reflects that a higher risk for celiac disease gives rise to a higher level in blood pressure.

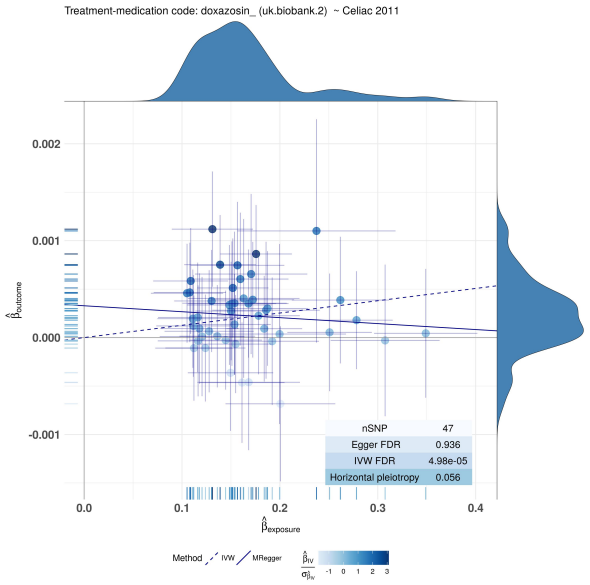

Figure D2: medication treatment ~ celiac disease shows that a higher risk for celiac disease increases blood pressure by medication.

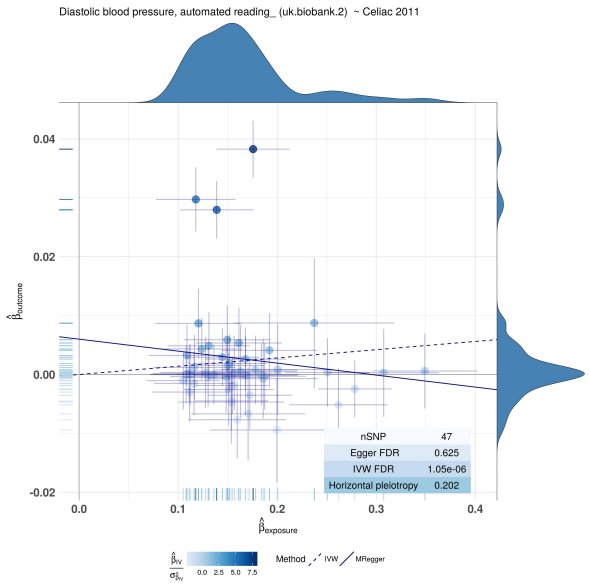

Figure D3: diastolic blood pressure ~ celiac disease depicts that a higher risk for celiac disease inflates blood pressure.

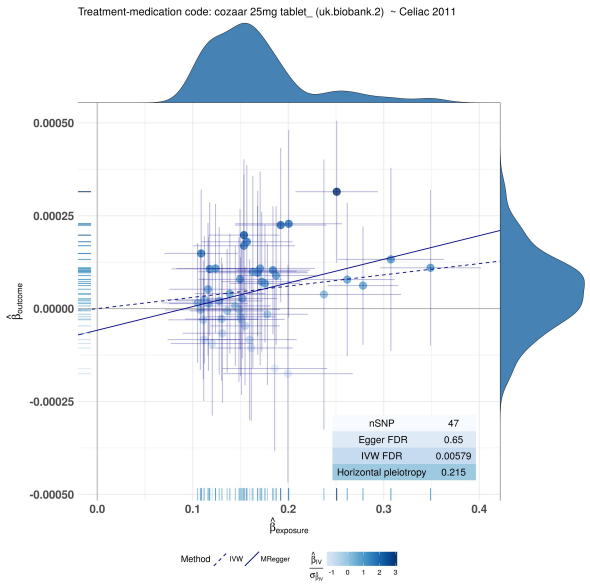

Figure D4: medication treatment ~ celiac disease shows that a higher risk for celiac disease increases blood pressure by medication.