1. Project Titel

Celiac Disease Triggers

2. Location

UMCG ERIBA, Genetic Department, A. Deusinglaan 1

3.

X

4. Project Summary

Celiac disease is a autoimmune disorder that has the possibility to occur in genetically predisposed people where the ingestion of gluten causes damage in the small intestine. When people with celiac disease digest gluten their body triggers an immune response that attacks the small intestine. This causes damage to the vili that are lined up in the small intestine for nutrient absorption. In addition, this leads to improperly nutrient uptake. Celiac disorder affects up to 2% of the European population. Celiac disease is hereditary. People with a first-degree relative with celiac disease have a risk of developing celiac disease. The inheritance model of the disease is oligogenic, a phenotypic outcome that is determined by more than on gene. 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 is not sufficient enough to develop celiac disease. 30% of the Europeans are carriers of the variant, however 3% are affected by the celiac disease.[4]

The aim of this study is to determine what factors triggers celiac disease. Public data will be used to identify factors that can explain variability in the disease manifestation. Systematically hundreds of potential triggers will be screened using the “Mendelian Randomization” approach. Mendelian randomization (MR) is an approach that uses genetic variants associated with a modifiable exposure or biological intermediate to estimate the casual relationship between these variables and a medically relevant outcome free from the influence of confounding. It was initially developed to examine the relationship between modifiable exposures/bio-markers and disease. Its use has been expanded to encompass applications in other areas. This is important because an increasing number of studies are investigating relationships between high-throughput molecular intermediates; DNA expression, gene methylation, metabolites. These investigations suffer from all the same issues of confounding and reverse causality.[1]

Reverse causality is the association between variable X and Y, but it an other way as expected. For example, it is assumed that X causes a change in Y. But this can also be the other way around, Y causing a change in X. This applies when the exposure disease process is reversed, in a different point of view, the exposure causes the risk factor. For example: someone is a heavy alcohol consumer, and because of the overly consumption is that person depressed, or is this person a heavy consumer of alcoholic products due to their depression?[3] Confounding causality is a variable that is correlated to both the dependent variable and the independent variable. X is the independent variable, Y is the dependent variable. To estimate the effect of X on Y. the statistician must suppress the effects of extraneous variables that influence both X and Y. We say that, X and Y are confounded by some other variable Z whenever Z is a cause of both X and Y. For example: z is gender, x is choice of drug, and y is the outcome. The variable x is dependent of z because a difference in gender can cause a change in drug choice, due to this y depends on x but is independent from z.[2]

MR refers to the random segregation and assortment of genes from parents to offspring that occur during gamete formation and provides a method of using genetic variants in observational settings to make casual inferences regarding the relationship between exposure and outcomes. The basic principle utilized in the MR framework is that if genetic variants either alter the level of or mirror the biological effects of a modifiable exposure that itself alters disease risk, then these genetic variants should be related to disease risk. MR studies aim to provide evidence for or against a causal relationship between a modifiable exposure variable and a disease. Genetic variants are used because these are less susceptible for confounding because of it is subjected to Mendel’s first law, the law of of segregation. These genetic variants segregate independently and randomly from environmental factors, and it can be assumed that genetic variants segregate independently from other traits.[1]

This public data comes in the form of Genome Wide Association (GWAS) summary data, which is for example available at GWAS Central. GWAS Central provides a centralized compilation of summary genetic association studies.[6] For the first part in this study, as much of available public GWAS data will be used, the reason for this is that a genetic variable must meet three core assumptions before it can be used as a instrumental variable (IV) for the MR framework. A big part of this study will be the screening of potential IVs for determination if they meet these core assumptions which are: (a) the confounders of the exposure-outcome must not be associated with the instrument, (b) the exposure of interest must be associated with the instrument, (c) the outcome must not be affected by the instrument, except through the exposure variable.[1] The testing for these assumptions will be done with RStudio[5] The methods used for testing whether the IVs meets the three core assumptions is not known, this will be further discussed during the internship. After screening a database with GWAS summary data for potential IVs in the form of single-nucleotide polymorphisms (SNP) comes building the pipeline, the mechanism between screening and performing MR. Which will be done in Python.[7]

The Mendelian Randomization Approach will be executed in RStudio, with the package MendelianRandomization.[9] This package is developed for various Mendelian Randomization analyses on summarized genetic data. This package has a diverse choice of methods to use in determination if an exposure has a causal effect on an outcome. For this package is a special class developed called MRInput. This object has multiple slots in which values can be assigned to pass all necessary information. **Table 1** describes the slots in the MRInput object. In order ta make an MRInput object, the following can be done: assign values to each slot separately or extract the values from a PhenoScanner.csv output file. This file can be collected with the bioinformatician tool PhenoScanner.[10] This is a curated database of public available results of GWAS. Currently the PhentoScanner can not be queried directly from the MendelianRandomization package. The PhenoScanner takes the output from the web version and converts it into an MRInput object.[8]

Table 1 Description of the MRInput object slots.

Slots	Description	Datatype
betaX	The associations of the genetic variants with the exposure. Is the beta-coefficient from regression analysis of the exposure on each genetic variant	Numeric vector
betaXse	The associations of the genetic variants with the exposure. Is the standard error.	Numeric vector
betaY	The associations of the genetic variants with the outcome. Beta-coefficients from regression analyses of the outcome on each genetic variant	Numeric vector
betaYse	The associations of the genetic variants with the outcome. Is the standard errors	Numeric vector
correlation	Correlations between variants. If not provided, it is assumed that variants are not correlated	Matrix
exposure	Name of the risk factor, e.g. LDL-cholesterol	Character string
outcome	Name of the outcome, e.g. coronary heart disease.	Character string
SNPs	the names of the various genetic variants (SNPs) in the dataset, e.g. rs12785878.	Character vector

For a basic introduction in performing MR analysis the next part describes some functionality of the MR package, which is of course oversimplified, or will probably not reach the degree of statistical analysis that is expected of the real project during the internship. It is more an expectation of what steps take place in an analysis with the MR package. So, the first part of this study will be screening genetic variables or instrumental variables, with the PhenoScanner[10], or collecting summarized data from the GWAS central database.[6] Which is now unknown, for example it is unclear which genetic variants take part in Celiac Disease. The data acquisition part is skipped, and pre-compiled data is used for this example. The second step is creating an MRInput object, slot by slot, this can be done with the `mr_input()` function:

Table 2 shows the slots contained by the MRInput object, recall that betaX and betaXse are the associations between genetic variant and exposure and that betaY and betaYse are the associations between genetic variant and outcome. This data that is used, are the associations of 28 genetic variants with LDL-cholesterol, HDL-cholesterol, triglycerides, and coronary heart disease (CHD) risk.[11] For analyzing the MRInput object the MendelianRandomization package has four methods for causal estimation; inverse-variance weighted method, median-based method, maximum likelihood method, and MR-Egger method. These functions can be used with:

Table 2 The slots of the MRInput object. BetaX en BetaXse are the associations of genetic variant and exposure. BetaY and BetaYse are the associations of genetic variant with outcome.

SNP	betaX	betaXse	betaY	betaYse
snp_1	0.0260	0.004	0.0677	0.0286
snp_2	-0.0440	0.004	-0.1625	0.0300
snp_3	-0.0380	0.004	-0.1054	0.0310
snp_4	-0.0230	0.003	-0.0619	0.0243
snp_6	-0.0310	0.006	-0.1278	0.0667
snp_7	-0.0180	0.004	-0.0408	0.0373
snp_8	0.0460	0.007	0.0770	0.0543
snp_9	0.0590	0.004	0.1570	0.0306
snp_10	0.0040	0.003	-0.0305	0.0236
snp_12	-0.0050	0.005	0.1823	0.0403
snp_13	0.0040	0.005	-0.0408	0.0344
snp_14	0.0220	0.005	0.1989	0.0335
snp_15	-0.0050	0.004	0.0100	0.0378
snp_16	-0.0020	0.004	0.0488	0.0292
snp_18	0.0040	0.004	-0.0408	0.0319
snp_19	0.0110	0.004	-0.0305	0.0316
snp_20	0.0090	0.003	-0.0408	0.0241
snp_21	-0.0110	0.004	-0.0202	0.0285

snp_22	-0.0030	0.003	-0.0619	0.0217
snp_23	-0.0120	0.004	0.0296	0.0298
snp_24	0.0003	0.003	0.0677	0.0239
np_25	-0.0150	0.003	-0.0726	0.0220
snp_26	-0.0080	0.004	-0.0726	0.0246
snp_27	0.0090	0.003	0.0000	0.0255
snp_28	-0.0360	0.007	0.0198	0.0647

These methods can be used with: *mr_ivw(MRInputObject)*, *mr_median(MRInputObject)*, *mr_egger(MRInputObject)*, *mr_maxlik(MRInputObject)*. Or one could use the *mr_allmethods(MRInputObject, method = "all")* function to easily compare results from multiple methods. After deploying one of the methods or all, one could plot the results to explore the associations of the different genetic variants, figure 1. Second, one could plot all methods estimates to compare them graphically, figure 2. For plotting the MendelianRandomization package has a plot function: *mr_plot()* which can be used for plotting the MRInput object, or the plot function can be applied to the output of of the testing methods which generates graphically estimates of the methods.

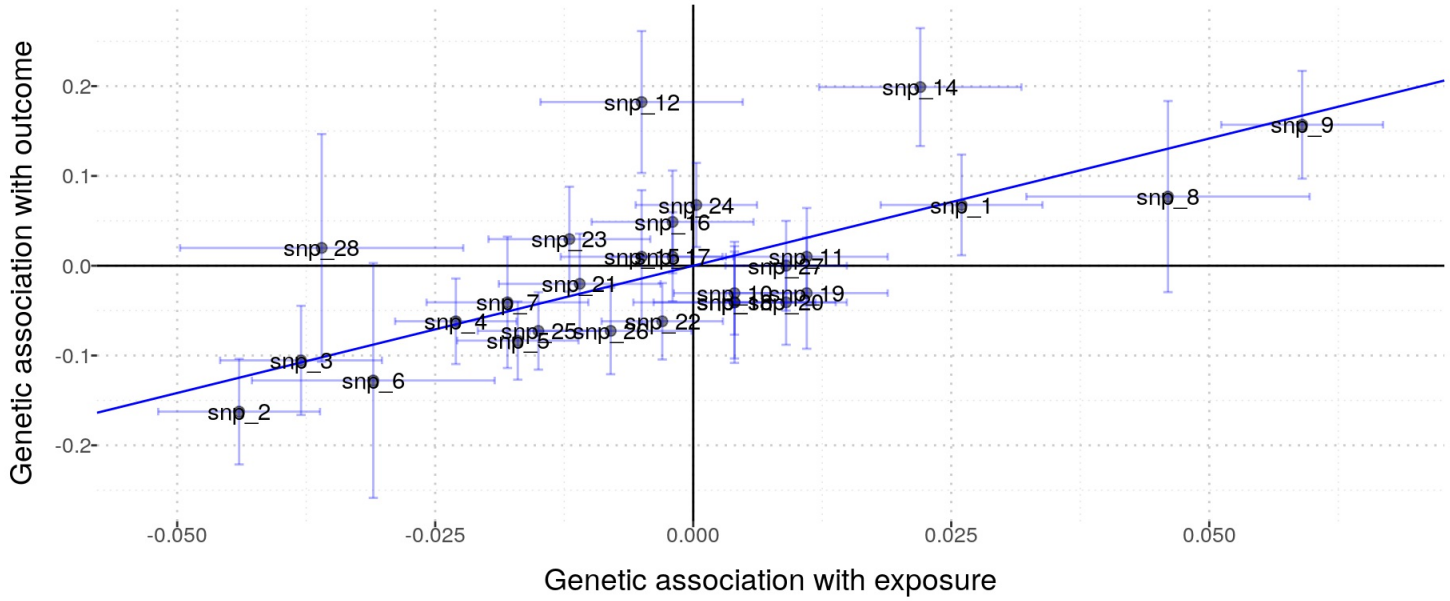

Figure 1 The mr_plot function applied to the MRInput object for exploration of the association between genetic variants.

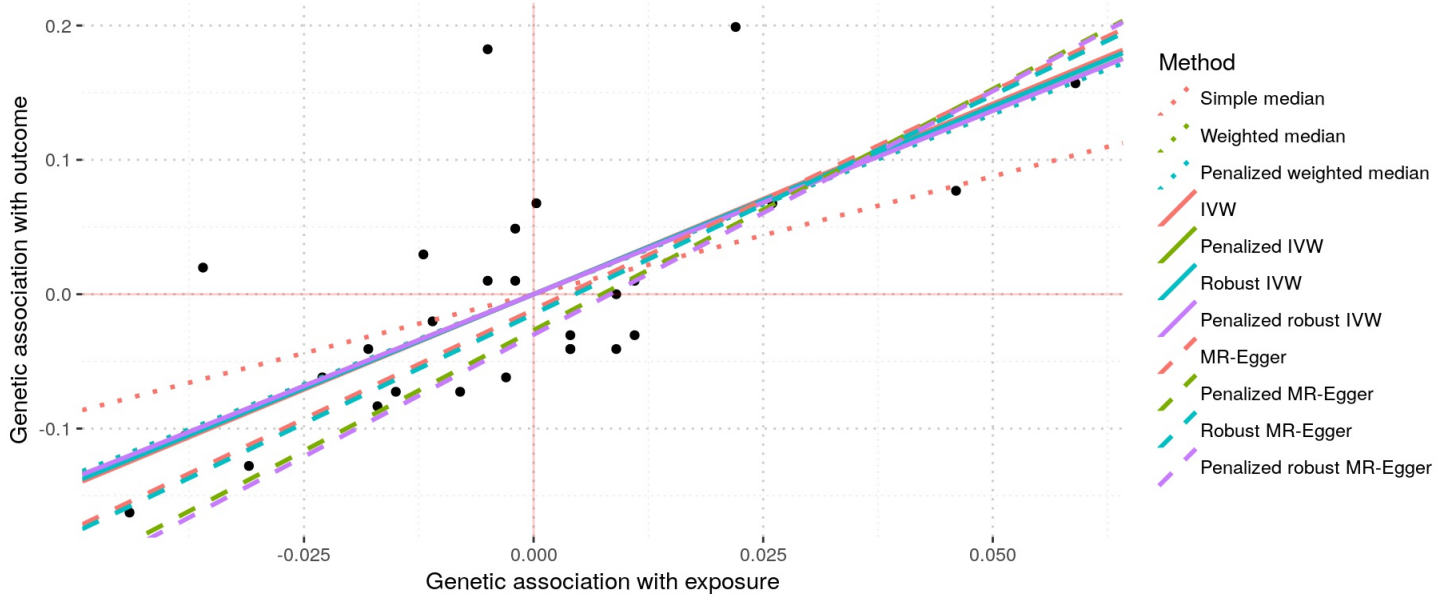

Figure 2 estimates of all the different methods applied to the MRInput object. This shows that that estimates from all methods are similar when LDL-cholesterol is the risk factor, but the MR-Egger estimates differ substantially when HDL-cholesterol is the risk factor.

It is important to note that this a simple overview of the steps that may or may not take place in this study. For example, it is at the moment unclear what tests to use for estimating association between outcome and genetic variants, or association between genetic variants and exposure. It is possible for example that the methods available in the MendelianRandomization package or not enough for this study, or that the Mendelian Randomization Approach does not fit the situation, situations where one could not simply deploy the Mendelian Randomization approach but needs an other model, like the two-sample and sub-sample Mendelian Randomization model, two-step Randomization, bidirectional Mendelian Randomization, Network Mendelian Randomization.[1] If there is a change in model, can the same methods for testing be applied?

5. Supervisor

dr. S. (Serena) Sanna, PhD

Faculty of medical Sciences/UMCG

s.sanna@umcg.nl (mailto:s.sanna@umcg.nl)

6. Daily Supervisor

Adriaan van der Graaf

7. Starting Day Internship

4/9/17

References

1. David M. Evans, George Davey Smith. 2015. Mendelian Randomization: New Applications in the Coming Age of Hypothesis-Free Causality. Annu. Rev. Genom. Hum. Genet. 2015.16:327-350.
2. Samuel Shapiro. 2008. Causation, bias and confounding: a hitchhiker's guide to the epidemiological galaxy. 2a: confounding. Available at: http://jfprhc.bmj.com/content/familyplanning/34/3/185.full.pdf (http://jfprhc.bmj.com/content/familyplanning/34/3/185.full.pdf)
3. statisticshowto. 2017. What is Reverse Causality? Available at: http://www.statisticshowto.com/reverse-causality/ (http://www.statisticshowto.com/reverse-causality/)
4. Walburg Dieterich, Tobias Ehnis, Michael Bauer. et all. 1997. Identification of tissue transglutaminsa as the autoantigen of celiac disease. Available at Nature Publishing Group: https://www.nature.com/naturemedicine (https://www.nature.com/naturemedicine)
5. RStudio 2017. Available at: https://www.rstudio.com/products/rstudio/ (https://www.rstudio.com/products/rstudio/)
6. GWAS Central. 2017. Available at gwascentral: http://www.gwascentral.org/ (http://www.gwascentral.org/)
7. Python. 2017. Available at: https://www.python.org/ (https://www.python.org/)
8. MendelianRandomization v0.2.0. 2017. an R package for performing Mendelian randomization analyses using summarized data. 1-18. Available at: https://cran.r-project.org/web/packages/MendelianRandomization/vignettes/Vignette_MR.pdf (https://cran.r-project.org/web/packages/MendelianRandomization/vignettes/Vignette_MR.pdf)
9. Olena O. Yavorska, Stephen Burgess. 2017. MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. International Journal of Epidemiology. 2017. 1–6. Available at: https://www.ncbi.nlm.nih.gov/pubmed/28398548 (https://www.ncbi.nlm.nih.gov/pubmed/28398548)
10. PhenoScanner. 2017. A database of human genotype-phenotype associations. Available at: http://www.phenoscanner.medschl.cam.ac.uk/phenoscanner (http://www.phenoscanner.medschl.cam.ac.uk/phenoscanner)
11. Waterworth et al (2011) Genetic variants influencing circulating lipid levels and risk of coronary artery disease. doi: 10.1161/atvbaha.109.201020.