Peer-reviewed version available at *Nutrients* **2020**, *12*, 1420; doi:10.3390/nu12051420

Article

A Two-Sample Mendelian Randomization analysis investigates associations between gut microbiota and celiac disease

Iraia García-Santisteban[1], Ariadna Cilleros-Portet[1], Elisabet Moyua-Ormazabal[1], Alexander Kurilshikov[2], Alexandra Zhernakova[2], Koldo Garcia-Etxebarria[3,4], Nora Fernandez-Jimenez[1*] and Jose Ramon Bilbao[1,5*]

[1] Department of Genetics, Physical Anthropology and Animal Physiology, University of the Basque Country (UPV/EHU) and Biocruces-Bizkaia Health Research Institute, Leioa, Basque Country, Spain; iraia.garcia@ehu.eus (I.G.S.); ariadna.cilleros@uvic.cat (A.C.P.); emoyua001@ikasle.ehu.eus (E.M.); nora.fernandez@ehu.eus (N.F.J.); joseramon.bilbao@ehu.eus (J.R.B.).

[2] Departament of Genetics, University of Groningen and University Medical Center Groningen, Groningen, The Netherlands; sashazhernakova@gmail.com (A.Z.); alexa.kur@gmail.com (A.K.).

[3] Gastrointestinal genetics group, Biodonostia Health Research Institute, San Sebastian, Spain; koldo.garcia@biodonostia.org

[4] Spanish Biomedical Research Center in Liver and Digestive Diseases (CIBERehd).

[5] Spanish Biomedical Research Center in Diabetes and associated Metabolic Disorders (CIBERDEM).

* Correspondence: nora.fernandez@ehu.eus; Tel.: +34-94-601-2909; joseramon.bilbao@ehu.eus; Tel.: +34-94-601-5289

Abstract: Celiac disease (CeD) is a complex immune-mediated inflammatory condition triggered by ingestion of gluten in genetically predisposed individuals. Literature suggests that alterations in gut microbiota composition and function precede the onset of CeD. Considering that microbiota is partly determined by host genetics, we speculate that the genetic makeup of CeD patients could elicit disease development through alterations in the intestinal microbiota. To evaluate potential causal relationships between gut microbiota and CeD, we performed a Two-Sample Mendelian Randomization analysis (2SMR). Exposure data were obtained from the raw results of a previous Genome Wide Association Study (GWAS) of gut microbiota, and outcome data from summary statistics of CeD GWAS and Immunochip studies. We have identified a number of putative associations between gut microbiota SNPs associated with CeD. Regarding bacterial composition, most of the associated SNPs are related to Firmicutes phylum, whose relative abundance has been previously reported to be altered in CeD patients. In terms of functional units, we have linked a number of SNPs to several bacterial metabolic pathways that seem to be related to CeD. Overall, this study represents the first 2SMR approach to elucidate the relationship between microbiome and CeD.

Keywords: celiac disease; gut microbiota; mendelian randomization.

1. Introduction

Celiac disease (CeD), the most common food intolerance, is a chronic immune-mediated systemic disorder triggered by an aberrant immune response to dietary gluten in genetically predisposed individuals [1]. Virtually all CeD cases harbor specific genetic variants located in the Human Leucocyte Antigen (HLA) region that encode for the HLA-DQ2/DQ8 heterodimers capable of presenting gluten peptides to T-cells, thus activating an inflammatory immune response in the intestine [2]. In addition to HLA risk alleles, Genome-Wide Association Studies (GWASs) have identified more than 40 non-HLA loci with modest contributions to CeD risk [3,4]. Whereas genetic predisposition and gluten exposure are necessary, they seem to be insufficient for the development

of CeD, suggesting that other factors might serve as crucial triggers for disease onset and progression. Gut microbiota could be one of such factors.

The human gut microbiota is composed of the microbial communities, mainly bacteria, which live in the digestive tract. It is acquired at birth from the environment, and diversity builds up over the first few years of life [5,6]. Afterwards, composition of the gut microbiota is largely shaped not only by environmental factors such as diet [7], but also by host genetic components [8–10]. Studies carried out in twins support the idea that host genetics influences gut microbial diversity [11]. Specific genetic variations have been linked to a vast array of medical conditions [12] including CeD. Cross-sectional studies indicate that the composition of the gut microbiota is altered in subjects with CeD compared with controls [13–16]. Prospective studies performed so far report differences in microbiota composition and diversity between children with a genetic predisposition for CeD compared to those with a non-selected genetic background [17–20]. Considering that host genetics contributes to the composition and function of gut microbiota, it is feasible to think that risk genotypes for CeD act in part via the microbiota. However, the limited number of studies carried out in patients, together with the small sample sizes, make it difficult to mechanistically correlate host genetics, gut microbiota and CeD.

Mendelian Randomization (MR) is an increasingly used statistical tool that can help establish a causal relationship between an exposure and an outcome of interest by employing single nucleotide polymorphisms (SNPs) as instrumental variables [21]. Two-Sample Mendelian Randomization (2SMR) refers to the application of MR in non-overlapping sets of individuals. These data can be obtained from public summary statistics of GWASs, or estimated directly from own genomic datasets [22]. Landmark genomic studies have identified many SNPs associated with gut microbiota structure and function [8–10], which could serve as valid instruments for the exposure in 2SMR analyses. Regarding outcomes of interest, SNPs could be selected from relevant GWA studies on CeD [4]. Thus, it is possible to investigate how gut microbiota composition and function may affect CeD risk by using publicly available data.

This study represents the first 2SMR approach to examine the interplay between host genetics, gut microbiota and CeD. Our analysis has identified a number of genetic variants associated to bacterial composition and function that could be driving CeD pathogenesis. These findings not only improve our understanding about the disease, but also elicit new research lines towards diagnosis and clinical management.

2. Materials and Methods

Two-sample MR analysis was conducted using the "TwoSampleMR" R package [23], following the guidelines provided by the developers (https://mrcieu.github.io/TwoSampleMR), and in-house developed R scripts (available upon request). Figure 1a displays a flowchart describing the whole procedure.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2020 doi:10.20944/preprints202003.0388.v1

Peer-reviewed version available at Nutrients 2020, 12, 1420; doi:10.3390/nu12051420

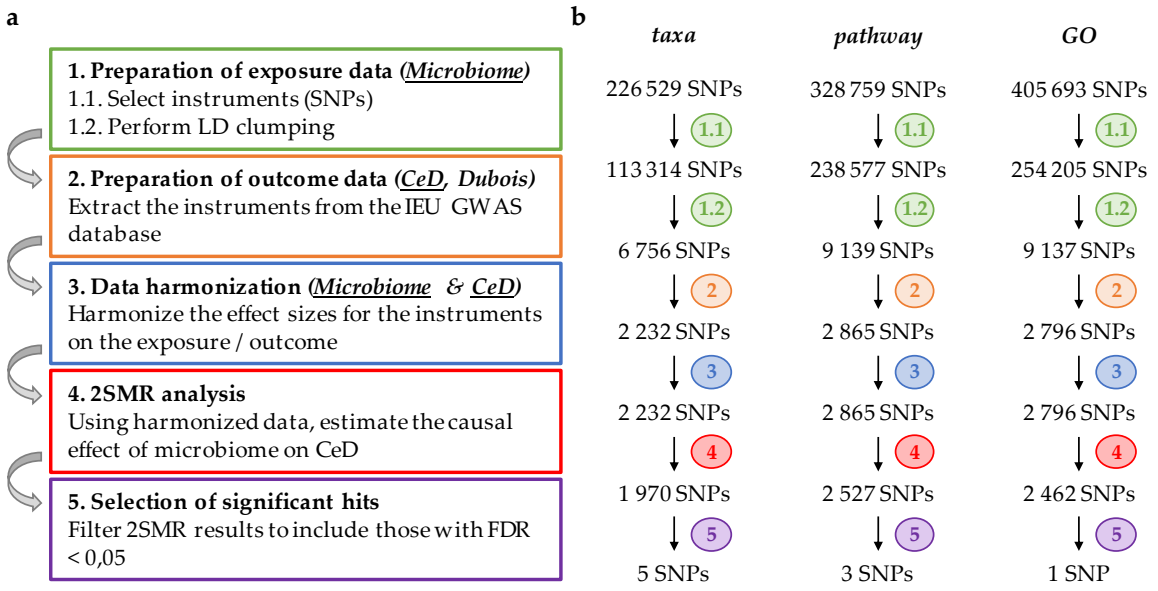

Figure 1. Schematic representation of 2SMR analysis using Bonder microbiome and Dubois CeD GWAS as exposure and outcome datasets, respectively. (**a**) Flowchart of the step-by-step analysis: after preparing exposure data (Step 1), outcome data is extracted (Step 2), both datasets are harmonized (Step 3), and 2SMR analysis is performed (Step 4); finally, significant hits are selected based on their False Discovery Rate (FDR) (Step 5); (**b**) Diagram representing the number of SNPs selected in each category (taxa, pathway, GO) after performing each step of the analysis.

2.1. Preparation of exposure data

Genetic instruments from the exposure data were obtained from raw results of the microbiome GWAS carried out by Bonder and colleagues in 1514 subjects [10]. Raw data comprised three independent datasets associating genetic variants (SNPs) with different bacterial traits, namely taxonomies (taxa), MetaCyc pathways (pathway) and Gene Ontology (GO) terms. Bacterial taxonomies included information about phyla, classes, orders, families, genera and species. MetaCyc pathways refer to experimentally elucidated metabolic pathways from bacteria retrieved from the MetaCyc metabolic pathway database (https://metacyc.org/) while GO-terms comprised a subset of 'informative' GO terms, associated with more than 2000 proteins for which no descendant term was associated. A more detailed description is available in the original article from Bonder and collaborators [10]. We filtered the three SNP sets by p-value and further investigated only those showing suggestive associations at a genome-wide significance level ($p<10^{-5}$).

Next, datasets were converted to "exposure data" format using the *format_data* function of the TwoSampleMR package. In this process, SNPs that did not have the standard "rs" nomenclature (less than 0.02% of all the variants), were dropped. In the case of SNPs that appeared more than once (due to association to more than one bacterial trait), only the one with the lowest p-value was retained.

Once exposure data had been formatted, independent SNPs were selected with the *clump_data* function in the same package, which uses the 1000 Genomes Project as a reference panel. Stringent clumping criteria were set to retain only those SNPs with the lowest p-value above a linkage disequilibrium (LD) threshold of $r^2 = 0.01$ in 10000 Kb windows.

2.2. Preparation of outcome data

Summary-level data of two independent association studies on CeD were retrieved from the IEU GWAS database. As discovery dataset, we selected the largest-to date GWAS on CeD, published by Dubois and colleagues in 2010 (ebi-a-GCST-000612). In this work, more than 500000 SNPs were genotyped in 4533 CeD individuals and 10750 controls [4]. As replication dataset, we used the Celiac Immunochip study, performed by Trynka and collaborators (ebi-a-GCST005523) in 2011. The investigators carried out a dense-genotyping of immune-mediated disease loci in 12041 CeD subjects

and 12228 controls, mapping in depth previously identified GWAS peaks and discovering several new CeD-associated regions [3]. Instrument SNPs from the GWAS and Immunochip results were obtained and formatted with the *extract_outcome_data* function, using previously prepared exposure data as input.

2.3. Harmonization of exposure and outcome data

To ensure that the effect of a SNP on the exposure and on the outcome correspond to the same allele, both exposure and outcome datasets were harmonized using the *harmonise_data* function. In this process, ambiguous and/or palindromic SNPs were removed, creating a new data frame that has the exposure and outcome data combined.

2.4. Two-sample Mendelian Randomization (2SMR) and statistical analysis

2SMR analysis was performed on the harmonized data frame using the *mr()* function, which returns a data frame of estimates of the causal effect of the exposure on the outcome for a range of different MR methods (Wald ratio, MR Egger, weighted median, inverse variance weighted and weighted mode). The multiple-testing-adjusted p-value, (False Discovery Rate, FDR) was calculated with the Benjamini–Hochberg procedure with the function *p.adjust()* of the "stats" package in R.

3. Results

Using a 2SMR approach, we set out to identify associations between gut microbiota (exposure) and CeD (outcome) using genetic variants. Exposure data were obtained from raw results of one of the most complete GWAS on gut microbiome, where Bonder and colleagues examine the influence of host genetics on gut microbiota composition (by interrogating microbial taxonomies) and function (by assessing bacterial MetaCyc pathways and GO-terms) [10]. These data were used to prepare three exposure datasets: taxa, pathway and GO, for our 2SMR analysis. After a stringent selection procedure of genetic variants (see Materials and Methods section for details), we retained 6756, 9179 and 9137 SNPs in taxa, pathway and GO categories, respectively (Figure 1b). Outcome data was prepared using summary-level data from the to-date largest CeD GWAS, which includes genetic variants encompassing the whole genome [4]. This led to the selection of 2232, 2865 and 2796 SNPs in the taxa, pathway and GO categories, respectively, which were subsequently harmonized and finally analyzed with 2SMR. We selected SNPs with a FDR<0.05, identifying 5, 6 and 1 hits associated with the taxa, pathway and GO categories (Table 1).

Table 1. Two-Sample Mendelian Randomization estimates between gut microbiome and CeD. Chr.: Chromosome; SE: Standard Error; FDR: False Discovery Rate.

SNP	Effect/Other allele	Chr.	Position	p-value	Effect size ± SE	FDR
taxa						
rs4396302	A/G	11	128420926	7.88×10^{-7}	0.80 ± 0.16	0.001
rs7594065	T/C	2	204814676	1.29×10^{-6}	-0.61 ± 0.13	0.001
rs10093096	C/T	8	64907701	3.56×10^{-5}	-0.84 ± 0.20	0.027
rs11545016	T/C	8	22438313	6.58×10^{-5}	-0.96 ± 0.24	0.037
rs12913063	T/C	15	75424593	9.97×10^{-5}	1.09 ± 0.28	0.044
pathway						
rs7585642	A/C	2	61217542	2.56×10^{-7}	-0.92 ± 0.18	0.001
rs131659	G/A	22	21964761	4.07×10^{-5}	1.04 ± 0.25	0.046
rs11867190	A/G	17	5261220	4.82×10^{-5}	0.59 ± 0.14	0.046
GO						
rs6848139	C/A	4	123395041	7.42×10^{-6}	-1.95 ± 0.43	0.021

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 March 2020 doi:10.20944/preprints202003.0388.v1

Peer-reviewed version available at Nutrients 2020, 12, 1420; doi:10.3390/nu12051420

To validate our results, we replicated the 2SMR analysis following the same criteria with the Celiac Immunochip results [3] as outcome data (Figure S1a,b). Since this study performed a dense genotyping of specific regions of the genome (including the HLA, the main CeD-predisposing locus), the number of SNPs finally analyzed in this case was around 10 times smaller, and the overlap with the SNPs included in Dubois analysis was low (Supplementary figure 1c). Even with this small overlap, 8 out of the 9 significant SNPs from the discovery study were replicated in the analysis with the Immunochip data, and many of them showed suggestive associations at a genome-wide significance level ($p<10^{-5}$). More importantly, forest plots comparing the beta values from both 2SMR analyses showed similar effect sizes and directions (Figure 2). Consistency in the results across the two independent 2SMR analyses builds confidence in the obtained estimates.

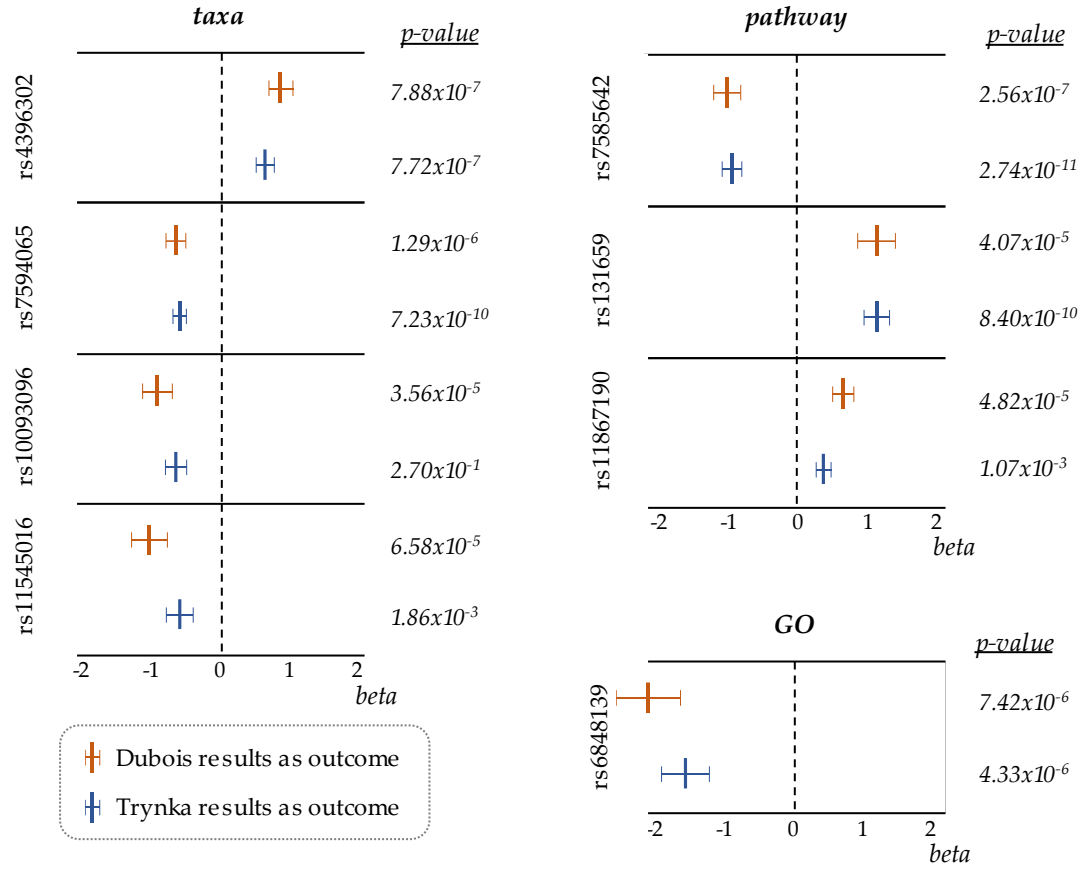

Figure 2. Forest plots comparing effect-sizes and p-values of SNPs identified in taxa, pathway and GO categories, in two independent 2SMR analyses where either Dubois or Trynka CeD datasets were used as outcome data.

The majority of the SNPs identified were located in introns and intergenic regions of CeD-related genes, and each of them was associated with one microbiome feature, either structural (taxa) or functional (pathway or GO) (Table 2). Traits related with bacterial composition were mainly linked to the Clostridiales order, a group of bacteria that has been observed to be altered in CeD individuals [24–26]. These data support the link between the genetic variants and CeD through the modulation of gut microbiota diversity. Regarding functional units, associated bacterial pathways and GO-terms are related to the metabolism of certain amino acids or their derivatives, supporting the implication of microbial metabolism on CeD pathogenesis.

Table 2. Associated microbiota traits of 2SMR hits. p: phylum; c: class; o: order; f: family; g: genus; s: species.

SNP	Associated microbiota trait
taxa	
rs4396302	Firmicutes (p), Clostridia (c), Clostridiales (o), *Peptostreptococcaceae* (f), *Peptostreptococcaceae* (g), *Peptostreptococcaceae* unclassified (s)
rs7594065	Firmicutes (p), Closridia (c), Clostridiales (o), *Clostridiales* noname (f), *Pseudoflavonifractor* (g)
rs10093096	Proteobacteria (p)
rs11545016	Firmicutes (p), Clostridia (c), Clostridiales (o), *Lachnospiraceae* (f), *Lachnospiraceae* noname (g)
pathway	
rs7585642	PWY-6060 (malonate degradation II, biotin-dependent)
rs131659	ARG+POLYAMINE-SYN (superpathway of arginine and polyamine biosynthesis)
rs11867190	PWY-3081 (L-lysine biosynthesis V)
GO	
rs6848139	GO:0016831 (MF, carboxy-lyase activity)

4. Discussion

To our knowledge, this work represents the first attempt to explore the interplay between host genetics, gut microbiota and CeD using a 2SMR approach. One of the key points for performing a successful 2SMR analysis is the appropriate selection of datasets to be used as exposure and outcome. In our study, instead of limiting the exposure data to selected instruments, we included the complete summary statistics of the microbiome GWAS by Bonder and collaborators in our analysis, thus increasing the coverage of our study. We then applied a stringent clumping protocol that selected genome-wide significant and independent variants in 10000 Kb windows, to end up with less than 3% of the genomic variants from the original datasets. In addition, compared to other genome-wide studies on gut microbiome [8,9], the GWAS from Bonder and colleagues evaluates the contribution of host genetics not only to bacterial composition, but also in terms of functional units [10]. In fact, recent whole-metagenome shotgun sequencing has revealed that fecal metabolic profiles are associated with only a few key species but with many common microbial functions, stressing the importance of the functional role of microbial communities on top of the microbial species present in the flora [27].

Regarding outcome data, we selected two breakthrough studies on CeD: the GWAS performed by Dubois and collaborators was selected as a discovery dataset, since it covered genetic variants encompassing the whole genome [4]. As a "replication" dataset we selected the celiac Immunochip study, that contains a smaller number of SNPs from regions known to be associated with immune disorders, but performs a much more dense genotyping of those particular loci [3]. Despite the limited overlap between the SNPs included in each of the two 2SMR analyses, 8 out of the 9 significant associations from the discovery study were replicated, with similar effect sizes, underscoring the validity of our results. On the other hand, the SNPs identified are located mainly in immune-related loci, and this reinforces the idea that host immune system plays a relevant role shaping microbial communities [28] also in the context of CeD.

In our analysis, we have pinpointed a number of significant associations between host genetics, gut microbiome and CeD. One of the most interesting findings is that all the identified taxa-related SNPs were linked to the Firmicutes and Proteobacteria phyla, two groups of bacteria whose abundance has been shown to be altered in CeD patients that harbor HLA risk alleles [17,24,29–31]. In this sense, it has been shown that Type I Diabetes risk variants in the IL2 pathway genes are

associated with microbial shifts in mice and humans, including the decrease of the *Lachnospiraceae* and *Clostridiales* families of the Firmicutes phylum [32]. Additionally, the oral administration of these particular strains is able to reduce disease severity in mouse models of colitis and allergic diarrhea [33], possibly through a bacteria-induced upregulation of *ICOS* and a consequent increased production of regulatory T-cells that reduce exacerbated immune-responses and inflammation. In the present work we have identified rs7594065 and rs6848139 (located close to the well-known IL2 pathway genes, *ICOS*/*CTLA4* and *IL2*, respectively) to be associated with both microbial features and CeD. In particular, the CeD-risk allele of rs7594065 appears to be negatively correlated with *Clostridiales* abundance. We propose that these SNPs somehow reduce the presence of *ICOS* and regulatory T cell-inducing strains in the microbial flora, contributing to the characteristic inflammation of the celiac intestine.

Another SNP from the pathway category, namely rs131659, is associated with the bacterial arginine/polyamine biosynthesis pathway. As illustrated in Figure 3a, the first step of this pathway is the conversion of arginine to ornithine by arginase, an enzyme whose activity has been shown to be induced by gluten peptides in human monocytes. In the second step, L-ornithine carboxy-lyase (a GO-term that has also been identified in our analysis) transforms ornithine into polyamines, which have been reported to increase permeability and inflammation *in vitro* [34]. Gluten peptides exert the same level of activation of the arginine metabolism in CeD and healthy individuals [35], suggesting that the increased activity of this pathway corresponds to the celiac microbiota, and in turn causes increased permeability and inflammation of the intestine.

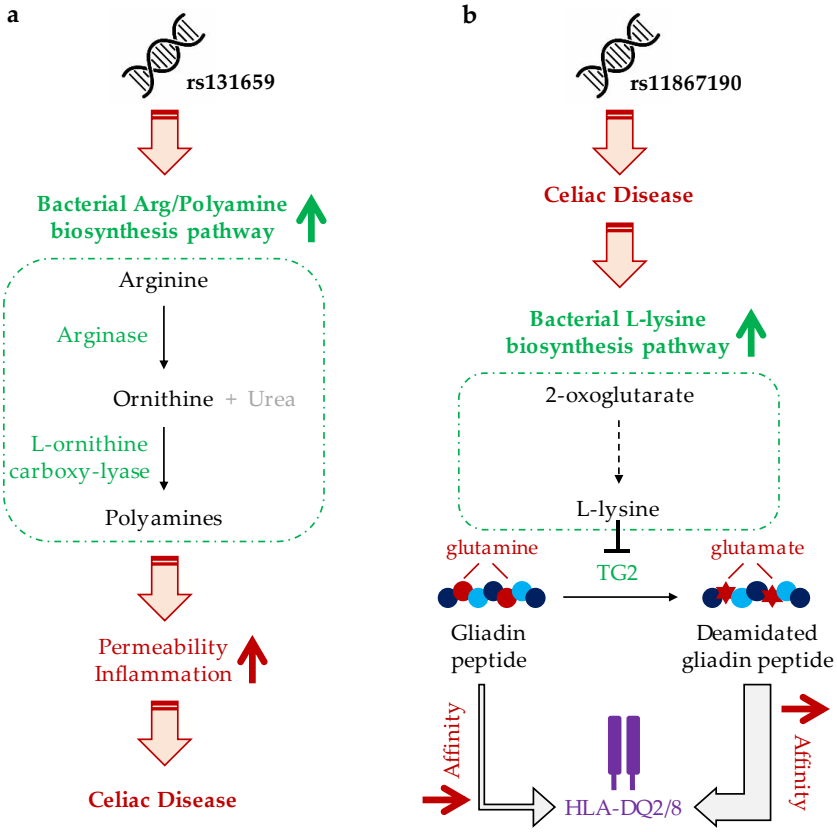

Figure 3. Schematic representation of hypothetical mechanism of action of the identified SNP-microbiota associations. (a) rs131659 is associated with increased bacterial arginine/polyamine biosynthesis pathway, where arginine is converted to polyamines; by increasing permeability and inflammation, polyamines could play a role in CeD manifestation. (b) rs11867190 SNP is associated with CeD; an increased production of L-lysine amino acid in the celiac intestine could be an adaptation of gut bacteria to counteract the activity of type 2 transglutaminase (TG2), which converts glutamines into glutamates; impaired TG2 activity would prevent the increase in the affinity of gliadin peptides for HLA-DQ2/8 receptors.

The last example among our functional candidates is rs11867190, which is associated with increased L-lysine biosynthesis (Figure 3b). A study carried out with gluten treated with transglutaminase type 2 (TG2) in the presence of a saturating amount of L-lysine showed an interruption of the deamidation reaction catalyzed by TG2, and consequently a lower affinity of these gluten-derived peptides for HLA-DQ2 molecules. In fact, enzymatic modification of gluten with TG2 plus lysine was able to suppress its immunologic effects on the duodenal mucosa of CeD patients [36]. These data suggest that a lysine-producing flora in CeD patients could represent a mechanism by which TG2 activity is weakened as a response to the disease.

Nevertheless, some limitations of this study should also be considered. Our study uses microbiome data from a cohort of adult individuals, whereas several microbiome studies on CeD have been carried out in children. It would be interesting to perform a similar analysis in an infant population when genome-wide studies on childhood microbiota become available. Also, we cannot exclude the possibility that non-significant associations might arise from a lack of power due to the limited sample size (1514 individuals) of the microbiome GWAS. In this sense, the recently established MiBioGen consortium aims to meta-analyze large-scale data from 18 independent cohorts in order to investigate host genetics and microbiota associations in almost 20000 subjects. This large-scale resource will also be very useful to assess the biological impact of gene-microbiome interactions in different diseases, including CeD [37]. Finally, the 2SMR approach cannot rule out the possibility that the associations discovered are pleiotropic rather than causal, and even then, one cannot not fully discard the possibility of reverse causation. Actually, we propose two opposite models: one in which variants associate to CeD through the modification of bacterial arginine and polyamine synthesis and/or through the reduction of regulatory T cell-inducing bacterial strains (direct causation), as well as another in which SNPs are associated with CeD and the bacterial flora adapts to the pathological condition of the host by producing lysine, in an attempt to impair gliadin presentation by HLA molecules (reverse causation). Further studies will certainly be needed in order to clarify this complex scenario.

In summary, this is the first work to identify genetic variants that could mediate CeD pathogenesis through gut microbiota. Our results should be interpreted with caution, pending epidemiological and experimental confirmation of the mechanisms proposed, but put forward interesting and plausible hypotheses that can explain the complex interactions between the host and microbial communities in the celiac gut, that might be potentially useful for the prediction, prevention and treatment of the disease.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, Figure S1: Schematic representation of 2SMR analysis using Bonder microbiome and Trynka CeD GWASs as exposure and outcome datasets, respectively.

Author Contributions: Conceptualization, I.G.-S., N.F.-J., J.R.-B.; resources and data curation, A.K., A.Z., K.G.-E.; methodology, formal analysis and validation, I.G.-S., A.C.-P., E.M.-O., K.G.-E., N.F.-J., J.R.-B.; writing—original draft preparation, I.G.-S., N.F.-J., J.R.-B; writing—review and editing, I.G.-S., A.C.-P., E.M.-O., A.K., A.Z, K.G.-E., N.F.-J., J.R.-B.; supervision, project management and funding acquisition, N.F.-J., J.R.-B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Basque Department of Health, grant numbers GVSAN2018/111086 and GVSAN2019/111085 to J.R.-B. and N.F.-J., respectively.

Acknowledgments: The authors thank for technical and human support provided by SGIker (UPV/EHU/ ERDF, EU), for the allocation of computational resources provided by the Scientific Computing Service.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results

References

1. Lindfors, K.; Ciacci, C.; Kurppa, K.; Lundin, K.E.A.; Makharia, G.K.; Mearin, M.L.; Murray, J.A.; Verdu, E.F.; Kaukinen, K. Coeliac disease. *Nat. Rev. Dis. Prim.* 2019, *5*, 1–18.
2. Bevan, S.; Popat, S.; Braegger, C.P.; Busch, A.; O'Donoghue, D.; Falth-Magnusson, K.; Ferguson, A.; Godkin, A.; Hogberg, L.; Holmes, G.; et al. Contribution of the MHC region to the familial risk of coeliac disease. *J. Med. Genet.* **1999**, *36*, 687–690.
3. Trynka, G.; Hunt, K.A.; Bockett, N.A.; Romanos, J.; Mistry, V.; Szperl, A.; Bakker, S.F.; Bardella, M.T.; Bhaw-Rosun, L.; Castillejo, G.; et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat. Genet.* **2011**, *43*, 1193–1201.
4. Dubois, P.C.A.; Trynka, G.; Franke, L.; Hunt, K.A.; Romanos, J.; Curtotti, A.; Zhernakova, A.; Heap, G.A.R.; Ádány, R.; Aromaa, A.; et al. Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet.* **2010**, *42*, 295–302.
5. Tamburini, S.; Shen, N.; Wu, H.C.; Clemente, J.C. The microbiome in early life: Implications for health outcomes. *Nat. Med.* 2016, *22*, 713–722.
6. Bäckhed, F.; Roswall, J.; Peng, Y.; Feng, Q.; Jia, H.; Kovatcheva-Datchary, P.; Li, Y.; Xia, Y.; Xie, H.; Zhong, H.; et al. Dynamics and stabilization of the human gut microbiome during the first year of life. *Cell Host Microbe* **2015**, *17*, 690–703.
7. David, L.A.; Maurice, C.F.; Carmody, R.N.; Gootenberg, D.B.; Button, J.E.; Wolfe, B.E.; Ling, A. V.; Devlin, A.S.; Varma, Y.; Fischbach, M.A.; et al. Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **2014**, *505*, 559–563.
8. Wang, J.; Thingholm, L.B.; Skiecevičie, J.; Rausch, P.; Kummen, M.; Hov, J.R.; Degenhardt, F.; Heinsen, F.A.; Rühlemann, M.C.; Szymczak, S.; et al. Genome-wide association analysis identifies variation in Vitamin D receptor and other host factors influencing the gut microbiota. *Nat. Genet.* **2016**, *48*, 1396–1406.
9. Turpin, W.; Espin-Garcia, O.; Xu, W.; Silverberg, M.S.; Kevans, D.; Smith, M.I.; Guttman, D.S.; Griffiths, A.; Panaccione, R.; Otley, A.; et al. Association of host genome with intestinal microbial composition in a large healthy cohort. *Nat. Genet.* **2016**, *48*, 1413–1417.
10. Bonder, M.J.; Kurilshikov, A.; Tigchelaar, E.F.; Mujagic, Z.; Imhann, F.; Vila, A.V.; Deelen, P.; Vatanen, T.; Schirmer, M.; Smeekens, S.P.; et al. The effect of host genetics on the gut microbiome. *Nat. Genet.* **2016**, *48*, 1407–1412.
11. Goodrich, J.K.; Davenport, E.R.; Beaumont, M.; Jackson, M.A.; Knight, R.; Ober, C.; Spector, T.D.; Bell, J.T.; Clark, A.G.; Ley, R.E. Genetic Determinants of the Gut Microbiome in UK Twins. *Cell Host Microbe* **2016**, *19*, 731–743.
12. Jackson, M.A.; Verdi, S.; Maxan, M.E.; Shin, C.M.; Zierer, J.; Bowyer, R.C.E.; Martin, T.; Williams, F.M.K.; Menni, C.; Bell, J.T.; et al. Gut microbiota associations with common diseases and prescription medications in a population-based cohort. *Nat. Commun.* **2018**, *9*.
13. D'Argenio, V.; Casaburi, G.; Precone, V.; Pagliuca, C.; Colicchio, R.; Sarnataro, D.; Discepolo, V.; Kim, S.M.; Russo, I.; Del Vecchio Blanco, G.; et al. Metagenomics reveals dysbiosis and a potentially pathogenic N. flavescens strain in duodenum of adult celiac patients. *Am. J. Gastroenterol.* **2016**, *111*, 879–890.
14. Bodkhe, R.; Shetty, S.A.; Dhotre, D.P.; Verma, A.K.; Bhatia, K.; Mishra, A.; Kaur, G.; Pande, P.; Bangarusamy, D.K.; Santosh, B.P.; et al. Comparison of small gut and whole gut microbiota of first-degree relatives with adult celiac disease patients and controls. *Front. Microbiol.* **2019**, *10*, 137–140.
15. De Palma, G.; Nadal, I.; Medina, M.; Donat, E.; Ribes-Koninckx, C.; Calabuig, M.; Sanz, Y. Intestinal

dysbiosis and reduced immunoglobulin-coated bacteria associated with coeliac disease in children. *BMC Microbiol.* **2010**, *10*.

16. Nadal, I.; Donant, E.; Ribes-Koninckx, C.; Calabuig, M.; Sanz, Y. Imbalance in the composition of the duodenal microbiota of children with coeliac disease. *J. Med. Microbiol.* **2007**, *56*, 1669–1674.
17. Sellitto, M.; Bai, G.; Serena, G.; Fricke, W.F.; Sturgeon, C.; Gajer, P.; White, J.R.; Koenig, S.S.K.; Sakamoto, J.; Boothe, D.; et al. Proof of concept of microbiome-metabolome analysis and delayed gluten exposure on celiac disease autoimmunity in genetically at-risk infants. *PLoS One* **2012**, *7*, 33387.
18. Sánchez, E.; De Palma, G.; Capilla, A.; Nova, E.; Pozo, T.; Castillejo, G.; Varea, V.; Marcos, A.; Garrote, J.A.; Polanco, I.; et al. Influence of environmental and genetic factors linked to celiac disease risk on infant gut colonization by Bacteroides species. *Appl. Environ. Microbiol.* **2011**, *77*, 5316–5323.
19. Olivares, M.; Benítez-Páez, A.; de Palma, G.; Capilla, A.; Nova, E.; Castillejo, G.; Varea, V.; Marcos, A.; Garrote, J.A.; Polanco, I.; et al. Increased prevalence of pathogenic bacteria in the gut microbiota of infants at risk of developing celiac disease: The PROFICEL study. *Gut Microbes* **2018**, *9*, 551–558.
20. Olivares, M.; Walker, A.W.; Capilla, A.; Benítez-Páez, A.; Palau, F.; Parkhill, J.; Castillejo, G.; Sanz, Y. Gut microbiota trajectory in early life may predict development of celiac disease. *Microbiome* **2018**, *6*.
21. Lawlor, D.A.; Harbord, R.M.; Sterne, J.A.C.; Timpson, N.; Smith, G.D. Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Stat. Med.* **2008**, *27*, 1133–1163.
22. Hartwig, F.P.; Davies, N.M.; Hemani, G.; Smith, G.D. Counterfactual causation: Avoiding the downsides of a powerful, widely applicable but potentially fallible technique. *Int. J. Epidemiol.* 2016, *45*, 1717–1726.
23. Hemani, G.; Zheng, J.; Elsworth, B.; Wade, K.H.; Haberland, V.; Baird, D.; Laurin, C.; Burgess, S.; Bowden, J.; Langdon, R.; et al. The MR-base platform supports systematic causal inference across the human phenome. *Elife* **2018**, *7*.
24. Olivares, M.; Neef, A.; Castillejo, G.; De Palma, G.; Varea, V.; Capilla, A.; Palau, F.; Nova, E.; Marcos, A.; Polanco, I.; et al. The HLA-DQ2 genotype selects for early intestinal microbiota composition in infants at high risk of developing coeliac disease. *Gut* **2015**, *64*, 406–417.
25. Collado, M.C.; Calabuig, M.; Sanz, Y. Differences between the fecal microbiota of coeliac infants and healthy controls. *Curr. Issues Intest. Microbiol.* 2007, *8*, 9–14.
26. Ercolini, D.; Francavilla, R.; Vannini, L.; De Filippis, F.; Capriati, T.; Di Cagno, R.; Iacono, G.; De Angelis, M.; Gobbetti, M. From an imbalance to a new imbalance: Italian-style gluten-free diet alters the salivary microbiota and metabolome of African celiac children. *Sci. Rep.* **2015**, *5*, 18571.
27. Visconti, A.; Le Roy, C.I.; Rosa, F.; Rossi, N.; Martin, T.C.; Mohney, R.P.; Li, W.; de Rinaldis, E.; Bell, J.T.; Venter, J.C.; et al. Interplay between the human gut microbiome and host metabolism. *Nat. Commun.* **2019**, *10*, 1–10.
28. Hooper, L. V.; Littman, D.R.; Macpherson, A.J. Interactions between the microbiota and the immune system. *Science (80-.).* 2012, *336*, 1268–1273.
29. Wacklin, P.; Kaukinen, K.; Tuovinen, E.; Collin, P.; Lindfors, K.; Partanen, J.; Mäki, M.; Mättuö, J. The duodenal microbiota composition of adult celiac disease patients is associated with the clinical manifestation of the disease. *Inflamm. Bowel Dis.* **2013**, *19*, 934–941.
30. Sánchez, E.; Donat, E.; Ribes-Koninckx, C.; Fernández-Murga, M.L.; Sanz, Y. Duodenal-mucosal bacteria associated with celiac disease in children. *Appl. Environ. Microbiol.* **2013**, *79*, 5472–5479.
31. Verdu, E.F.; Galipeau, H.J.; Jabri, B. Novel players in coeliac disease pathogenesis: Role of the gut microbiota. *Nat. Rev. Gastroenterol. Hepatol.* 2015, *12*, 497–506.
32. Mullaney, J.A.; Stephens, J.E.; Costello, M.E.; Fong, C.; Geeling, B.E.; Gavin, P.G.; Wright, C.M.; Spector, T.D.; Brown, M.A.; Hamilton-Williams, E.E. Type 1 diabetes susceptibility alleles are associated with

distinct alterations in the gut microbiota. *Microbiome* **2018**, *6*.

33. Atarashi, K.; Tanoue, T.; Oshima, K.; Suda, W.; Nagano, Y.; Nishikawa, H.; Fukuda, S.; Saito, T.; Narushima, S.; Hase, K.; et al. Treg induction by a rationally selected mixture of Clostridia strains from the human microbiota. *Nature* **2013**, *500*, 232–236.
34. Barilli, A.; Rotoli, B.M.; Visigalli, R.; Ingoglia, F.; Cirlini, M.; Prandi, B.; Dall'Asta, V. Gliadin-mediated production of polyamines by RAW264.7 macrophages modulates intestinal epithelial permeability in vitro. *Biochim. Biophys. Acta - Mol. Basis Dis.* **2015**, *1852*, 1779–1786.
35. Barilli, A.; Gaiani, F.; Prandi, B.; Cirlini, M.; Ingoglia, F.; Visigalli, R.; Rotoli, B.M.; De'Angelis, N.; Sforza, S.; De'Angelis, G.L.; et al. Gluten peptides drive healthy and celiac monocytes toward an M2-like polarization. *J. Nutr. Biochem.* **2018**, *54*, 11–17.
36. Elli, L.; Roncoroni, L.; Hils, M.; Pasternack, R.; Barisani, D.; Terrani, C.; Vaira, V.; Ferrero, S.; Bardella, M.T. Immunological effects of transglutaminase-treated gluten in coeliac disease. *Hum. Immunol.* **2012**, *73*, 992–997.
37. Wang, J.; Kurilshikov, A.; Radjabzadeh, D.; Turpin, W.; Croitoru, K.; Bonder, M.J.; Jackson, M.A.; Medina-Gomez, C.; Frost, F.; Homuth, G.; et al. Meta-analysis of human genome-microbiome association studies: The MiBioGen consortium initiative. *Microbiome* **2018**, *6*.
38. Rothschild, D.; Weissbrod, O.; Barkan, E.; Kurilshikov, A.; Korem, T.; Zeevi, D.; Costea, P.I.; Godneva, A.; Kalka, I.N.; Bar, N.; et al. Environment dominates over host genetics in shaping human gut microbiota. *Nature* **2018**, *555*, 210–215.
39. Benson, A.K.; Kelly, S.A.; Legge, R.; Ma, F.; Low, S.J.; Kim, J.; Zhang, M.; Oh, P.L.; Nehrenberg, D.; Hua, K.; et al. Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 18933–18938.

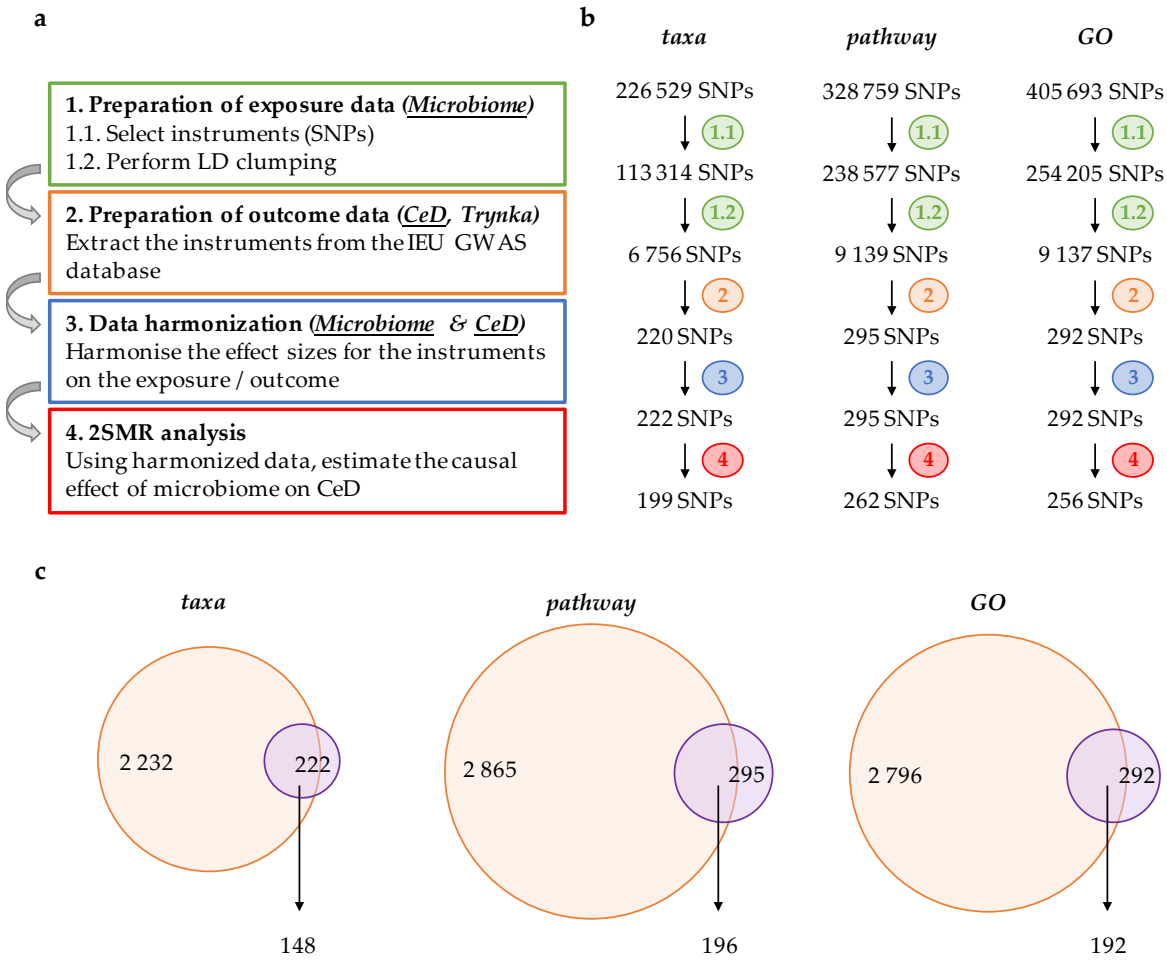

Figure S1. Schematic representation of 2SMR analysis using Bonder microbiome and Trynka CeD GWASs as exposure and outcome datasets, respectively. (**a**) Flowchart of step-by-step analysis: after preparing exposure data (Step 1), outcome data is extracted (Step 2), both datasets are harmonized (Step 3), and 2SMR analysis is performed (Step 4) (**b**) Diagram representing the number of SNPs selected in each category (taxa, pathway, GO) after performing each step of the analysis. (**c**) Venn-diagrams illustrating the number of SNPs in harmonized datasets derived form Bonder vs Dubois (orange) or Bonder vs Trynka (purple) analyses.