**S1 text. Supplementary Methods**

Table of Contents

[Real-time quantitative PCR 2](#_Toc1470921)

[Genotyping and quality control 2](#_Toc1470922)

[Covariate selection by CMS 2](#_Toc1470923)

[Simulation study 2](#_Toc1470924)

[Potential limitation 3](#_Toc1470925)

[Re-parametrization of the logistic model with a linear probability function 4](#_Toc1470926)

[Without the linear approximation of the Logistic model 8](#_Toc1470927)

[Alternative study design 9](#_Toc1470928)

# Real-time quantitative PCR

As 16S sequencing approach is known to have some limitation regarding the accuracy of the quantification [23], we performed a quantitative analysis of *F. prausnitzii* on fecal DNA extracted from weighted human stool samples via real-time quantitative PCR using an ABI 7000 Sequence Detection System apparatus with 7000 system software v. 1.2.3 (Applied Biosystems, Foster City, CA, USA). Amplification and detection were carried out in 96-well plates and with TakyonTM SYBR Green PCR kit (Eurogentec, Liege, Belgium). The probes and primers for the bacterial 16S rRNA genes described previously were used [22]. The threshold cycle for each sample was determined for each gene normalized to the CT value of the all-bacteria 16S ribosomal RNA gene. Data were calculated using the 2-ΔΔCt method [45].

# Genotyping and quality control

Patients with IBD were genotyped using Fluidigm technology (UMR CNRS 8199, Lille, France) for the following SNPs from five genes: rs12994997 (*ATG16L1*, risk allele A), rs10781499 (*CARD9*, risk allele A), rs11564258 (*LRRK2*, risk allele A), rs11209026 (*IL23R*), and three *NOD2* variants, rs2066847 (insertion), rs2066845 (risk allele C), and rs2066844 (risk allele T). The characteristics of the variants are presented in **S1 Table.** However, the very low frequency of the *IL23R* variant (minor allele frequency of 0.028) prevented from performing a valid association test, and was therefore removed from the analysis. The variants from *ATG16L1*, *CARD9*, and *LRRK2* were coded assuming additive effects of the risk allele. The *NOD2* variants had low to very low frequencies (**S1 Table**) and were therefore aggregated into a binary variable defined as the presence/absence of at least one allele associated with increased risk of IBD: SNPs rs2066844 (C > T), rs2066845 (G > C) and rs2066847 (insertion).

# Covariate selection by CMS

*CMS*(Covariates for Multiphenotype Studies) is an approach recently developed by our group for the analysis of multivariate dataset [19]. It aims that identify for each specific univariate test between a predictor and a primary outcome, a subset of variables from the data that can be used as proxy for explanatory variable of the primary outcome. Once selected, these variables can simply be included in the model as covariates. In brief, for each univariate test between a genetic variant and a given bacterium (the primary outcome), the *CMS* approach parses other bacterial levels correlated to the primary outcome and identifies those that can be added as covariate in the regression model without introducing bias [58]. More precisely, the *CMS* targets bacteria that likely vary with the primary outcome because of shared explanatory variables, and can therefore be used as proxy for unmeasured causal exposure(s) of the primary outcome. Including these bacteria as covariate in the regression model can reduce the primary outcome residual variance and increase power.

# Simulation study

We performed simulation studies for the four causal models across a wide range of scenarios to confirm the validity of our inference on effect estimate bias, and explore its robustness. In brief, we used a general framework where data were drawn based on the two following generating functions:

where parameters () were specifically set to 0 to match the desired model, i.e. in model a); in model b); and in model c) and d), while , the baseline risk was set to match the desire disease prevalence. Except when stated otherwise, the genotype was drawn from a binomial distribution with minor allele frequency of 0.1. When influenced by the genotype (model a) or the disease (model b), the bacterium was simulated using the above standard linear model while considering various distribution for the residual . However, we also considered simulating using a negative binomial distribution using the following model

where is the dispersion parameter, and , the expected value of is defined as , and and are set to zero depending on the scenario considered. Additional details on the simulation study are provided in the **SI Material and Methods**.

# Potential limitation

How to optimize microbiome data pre-processing is challenging. Empirical bacterial levels follow non-negative, over-dispersed distribution, and often harbor a large number of zeros values. It has been suggested to use Poisson distribution, negative binomial distributions, or hurdle models to analyze microbiome data [31]. Moreover, we applied a rank-based inverse-normal transformation in our real data analysis, while there remains sparsity (after QC, bacteria had a minimum of 20% of non-zero values). To circumvent the impact of data sparsity and data transformation, we performed a stringent filtering, removing all taxa quantified in less than 20% of the individuals. We did not to use a more stringent threshold to allow for a screening as broad as possible, while we validated the most significant results through replication analysis in independent datasets. We also applied rank-based inverse normal transformation to address both non-normal distribution of the bacteria levels and avoid false signal due to outliers. We acknowledge that such non-linear transformation can affect the inference [54]. However, rank-based inverse normal transformation is very commonly used in marginal genetic effect analyses of ‘omics data (e.g. microarray, RNA-seq, metabolites, which have properties similar to our microbiome data) and the current consensus is that, despite drawbacks, it remains a simple and efficient solution in many settings as compared to more complex approaches [55, 56]. Moreover, we also noted that rank-based inverse normal transformation is also now used in some microbiome data analysis [11, 57].

To assess the impact of the aforementioned pre-processing, we performed a simulation study where bacterial levels were drawn from a negative binomial distribution keeping all parameters fixed except the dispersion parameter in order to achieve between 0% and 95% of zero-value. For each replicate, SNP-bacteria effect was estimated using standard linear regression before and after applying the rank-based inverse-normal transformation as in the real data analysis. As showed in **S14 Fig**, which presents the results from this simulation for the four scenarios from **Fig 3a-d**, the only expected impact is a very low power to identify association with taxa quantified in less than 70% of the individuals in our dataset. However, all taxa from our primary results presented in **Table 2** were present in more than 85% of the individuals (e.g. *Fprau* was quantified in more than 95% of the individuals).

Another limitation is that we assumed all effects are homogeneous in our causal inference framework, i.e. Crohn’s disease and ulcerative colitis cases were treated as a single outcome variable (IBD). However, it is well appreciated that *NOD2* variants and *F. prausnitzii* are mostly associated with CD, and ileal CD in particular, but show little evidence for association with UC. To assess the impact of genetic heterogeneity, we performed a second simulation mimicking as closely as possible the real data. Here, SNP-disease and bacteria-disease associations were present only for a subset of the cases. As showed in **S15 Fig**, in this scenario, some of the bias from model (c) and (d) illustrated in **Fig 3g-h** decreased toward the null. Our simulations also show that under model (c), and (d) we expect no association signal in neither disease subtype strata nor in the whole IBD-cases sample.

Finally, future studies with improved study design might help deciphering further the observed signal. The heterogeneity of treatment across individuals and the availabilty of covariates informations across studies can bias the results of both association screening and causal inference. Previous works already highlighted the strong effect of many drug treatment and environmental exposures on the microbiota composition. To limit the impact of these factors, we focused our discovery and replication analyses on individuals which had data on major confounding factors. Future work might instead explore imputation methods to infer missing covariates, or performed stratified analysis in more homogenous sample. Also, we merged all three *NOD2* variants into a binary variable defined as the presence/absence of at least one variant. While the effect of each of these variants is established, their combined effect is not well defined [41-43]. Understanding the relative contribution of each *NOD2* variant to the overall signal we observed is of primary interest but would require larger sample size.

# Re-parametrization of the logistic model with a linear probability function

Consider a biallelic genetic variant, generated from a binomial distribution with minor allele frequency , a bacterial level following a distribution and a disease status with probability . For mathematical convenience we further consider , and for standardized , and , respectively. In order to match standard modelization, we consider further the case where the effects of predictors on the disease status are drawn from a logistic function (but other link function, such as the liability threshold model might be explored), and the effects on are drawn from a linear model. Our four scenarios of interest can therefore be expressed with the two generating functions:

where parameters () can specifically be set to 0 to match the desired model, i.e. in model a); in model b); and in model c) and d), while , the baseline risk is set to match the desire disease prevalence.

Solving for the logistic function faces numerous challenges (see next section). Instead, we suggest the direction of the bias (i.e. the sign of ) can easily be inferred by approximating the logistic function by a linear models. When the disease is neither rare nor too common (e.g. prevalence in 0.3-0.7), and main and effects are small, the log odds are almost a linear function of the probability, so that a simple marginal model of the predictors would approximate the logistic model very well (**S8a Fig**). However, for more general cases (i.e. low or large prevalence, or large effects), fitting the probability of the disease requires adding polynomial terms of the predictors and their interactions. For illustration purposes we plotted in **S8 Fig** simulated disease probability (from a logistic model) against the predicted values from linear probability models while increasing the model complexity, i.e. increasing in the following . Importantly, the additional parameters capturing the non-linear effects of and are only of interest for prediction purposes of limited interest if the goal is causal inference.

Moreover, previous research that aimed at estimating log odds from linear models showed that for a predictor , under the assumption of normality of error of , the parameter from the linear model can be approximated as , where and are the variance of the disease and the log odds, respectively [63, 65]. It follows that under all the above assumptions (disease prevalence around 0.5, small effect, and normality of errors), the following expected values can be approximated as:

When applied to our causal models it results in the following approximations:

For model *a)*, and :

It follows that the bias can be expressed as:

For model *b)*, and :

For model *c)* :

The independence between and implies . We have therefore the following approximation:

For model *d)*  .

This is the simplest scenario and a case well established in the literature. is a direct risk factor of and its effect is invariant with . This relationship can be expressed through the linear model:

It follows that:

indicating that is invariant with in case only sample.

In our case, almost all assumptions are violated, so that an accurate estimation of the bias is not possible with the above approximation as the magnitude of will in general be misspecified. Nevertheless, as aforementioned, the signs of the estimates in the linear approximation of the conditional mean of remain valid. It follows that the direction of the bias can be inferred from the above estimators independently of the distribution of and .To confirm the validity of this approximation, we performed series of simulations across a broad range of scenarios, assuming rare or common disease, moderate or large effects, and drawing the residual of from either a normal distribution, an exponential distribution, or a uniform distribution. Our inference on the direction of the bias was validated in all scenarios considered (**S9-12 Figs**).

# Without the linear approximation of the Logistic model

The formal derivation of the bias using the logistic formulation that we explored led to intractable solutions. For example, re-writing the exponential function in the logistic model as power series we have:

The analytical form of , , and based on the above conditional expected value of , implies solving the expectation of a ratio of functions including moments of and from 0 to infinity.

Instead we explored solutions leveraging the rare disease assumption to approximate the logit function with an exponential, while using the same power decomposition. However, this solution was unsuccessful, leading in particular to an expected null bias for model c) and d), in disagreement with all simulations we conducted (see e.g. **S9-12 Figs**), and was therefore not pursued further. Indeed, following this path for model c) and d) we have:

so that,

# Alternative study design

Using these IBD-cases samples to explore causal relationship has some advantages over other study designs. Study on healthy subjects can only be used to test for association between genetic variants and bacteria but cannot be used to decipher potential mediation effect. On the other hand, while prospective cohorts in the general population have the potential to solve a number of statistical artifacts faced in our study, they require unrealistic sample size when studying relatively rare disease like IBD. A more reasonable alternative to IBD-cases only is case-control data. Case-control data provide additional information on SNP-bacteria association while only requiring a reasonable increase the sample size (e.g. matched healthy subjects). Nevertheless, case-control data does not solve the selection bias issue faced in the present study. By definition there is no simple and general solution to address selection bias. Instead the evaluation of its impact has to be done on a case by case basis after accounting for the correlation between variables in the general population, as done in our study. Our extensive simulations and sensitivity analyses confirm the robustness of our inference for the range of effect size considered.