Supplementary Note to Qin et al. (2021) “Combined effects of host genetics and diet on human gut microbiota and incident disease in a single population cohort”.

Due to space constraints, details about various parts of the results and methods were added to this associated Supplementary Note.

Supplementary Results

Replication of previous microbiome GWAS signals

We assessed whether previously reported association signals were replicated in FINRISK02. Briefly, our replication approach was almost identical to that of Rotschild et al. (2018)[1], in which we examined whether matched associations between our results and previous work surpassing a Bonferroni-corrected significance threshold also involved identical microbial taxa or equivalent GTDB-NCBI taxon synonyms. In total, we considered 547 SNPs reported associated with gut microbial features in 8 studies[1–8], regardless of their passing genome-wide significance, or their proxies in high LD ($r^2>0.8$). We observed frequent phylum-level replication of previously reported variants (**Supplementary Table 3**), indicating that statistical power of microbiome GWAS is a key issue. In particular, 150 previously reported associations and the same or proxy SNPs in FINRISK02 surpassed Bonferroni-corrected significance ($p<9.12\times10^{-5}$); these encompassed the *LCT* locus as well as 20 other loci.

Supplementary Methods

Taxonomic profiling, quality filtering and data transformation (detailed section)

Taxonomic profiling of FR02 metagenomes was performed as follows: briefly, raw shotgun metagenomic sequencing reads were mapped using the *k*-mer-based metagenomic classification tool Centrifuge[9] to an index database custom-built to encompass reference genomes that followed the taxonomic nomenclature introduced and updated in the GTDB release 89[10–12]. This implies that unless specified otherwise, all taxonomic names in our study refer to their nomenclature in GTDB, which can be related to the original NCBI nomenclature using the GTDB database server: https://gtdb.ecogenomic.org/taxon_history/. The same profiling approach has also been used and described in recent studies from our consortium[3,4,13].

Gut microbial composition was represented as the relative abundance of taxa. For each metagenome at phylum, class, order, family, genus and species levels, the relative abundance of a taxon was computed as the proportion of reads assigned to the clade rooted at this taxon among total classified reads. The relative abundance of a taxon with no reads assigned in a metagenome was considered as zero in the corresponding profile. For the purpose of this association study and because of reduced accuracy and power when considering rare taxa, we focused on common and relatively abundant microbial taxa, defined as prevalent in >25% studied individuals, and defined with at least 10 mapped reads per individual. For the purpose of association, and as previous studies have reported that only some microbial taxa are inheritable[14], we also removed taxa with zero SNP-heritability. This filtering resulted in a microbial dataset composed of a total of 2,801 taxa, including 59 phyla, 95 classes, 187 orders, 415 families, 922 genera and 1,123 species.

Taxonomic profiles derived from sequencing data are by nature compositional because of an arbitrary total imposed by the instrument[15]. The compositional data of microbial taxa is not independent and can lead to inappropriate use of linear regression. To overcome this artificial bias, all relative abundance values were transformed by centre-log-ratio (CLR)[16]. CLR transformed data can vary in real space and better fit the normality assumption of linear regression. To minimize the impact of zeros, the reads count profiles were shifted by +1 before the transformation. The choice of zero modification method depends on our understanding of what would constitute a "zero" in our data. As we stringently focus on taxa present in >25% individuals, it is reasonable to assume that a lot of the taxa below this threshold could likely be detected if sequencing depth was higher than our study. We conceptually used the +1 shift of matrix as a proxy for the effect of increasing sequencing effort, which would affect all taxa. This process was performed using the R package *compositions*.

When visually comparing relative abundances in groups of individuals throughout the manuscript, we used untransformed relative abundances, for better interpretability. Alpha (Shannon index) and beta (Bray-Curtis distance) diversity were calculated at genus level used functions in the R package *vegan*. We did not find a correlation between sequencing depth and Shannon diversity index (Spearman's ρ=-0.001598, p=0.90) in n=5,959 samples (**Supplementary Figure 9**). Additionally, to define CLR-transformed abundances of higher taxonomic levels than species, we summed the raw abundances of all taxa (e.g. species) belonging to a specific higher taxonomic taxon (e.g. genus), and then applied a CLR transformation.

Additionally, we observed that Eastern and Western Finnish populations did not have different microbiome diversity, despite having overall slightly different lifestyles, and mortality rates. To further investigate this, we visualised potential geographical effects using a PCoA plot on beta-diversity (Bray-Curtis dissimilarity) from metagenomic profiles of samples used in the GWAS from our study (n=5,959; **Supplementary Figure 10**).

Our study presents results involving the bacterium Faecalicatena lactaris, a taxonomic definition introduced in GTDB release 89, and which was used throughout this study. *F. lactaris* was reclassified in the latest (to date) release of GTDB (release 95) as *Mediterraneibacter lactaris*, a new taxonomic definition. Future taxonomic reclassifications of bacterial species according to the GTDB taxonomic system can be checked on the GTDB website at: https://gtdb.ecogenomic.org/.

Note on GTDB nomenclature of *Faecalicatena lactaris*

Our study presents results involving the bacterium *Faecalicatena lactaris*, which was used throughout this study. Reference genomes from NCBI used to define this taxa were initially belonging to a Candidate *Ruminococcus lactaris* species and as such, *F. lactaris* was initially called *Ruminococcus_B lactaris* in GTDB releases 80, 83 and 86. Release 89 (used in our study) introduced the *F. lactaris* nomenclature, which was finally reclassified in GTDB releases 95 and 202 as *Mediterraneibacter lactaris*. Future taxonomic reclassifications of bacterial species according to the GTDB taxonomic system can be checked on the GTDB website at: https://gtdb.ecogenomic.org/.

Replication of previously reported associations (detailed section)

To evaluate the reproducibility of our results with previously reported associations, we collected GWAS summary results from 8 studies published in peer-reviewed journals at the time of this work[14,24–30]. These studies reported associations between 548 SNPs and microbial features. ANNOVAR was used to annotate the reported SNPs to the hg38 human reference genome[23] and we used GCTA[31] to identify a further 15,427 SNPs in high LD (r^2>0.8, within 5 Mbp) with any of these 548 SNPs. To assess replication, we first examined whether previously reported associations could be matched in our results to identical or linked SNPs, with an association below the Bonferroni-corrected suggestive significance threshold, which was set to $0.05/548 = 9.124\times10^{-5}$. Our study follows the GTDB taxonomic system[10,11], implying inherent taxonomic inconsistencies with microbial taxa named according to their NCBI taxonomic nomenclature. Similarly to the approach undertaken in previous microbiome GWAS studies[28], we then compared whether matched associations between previous studies and this work also involved microbial taxa belonging to phylogenetically related taxa, i.e. the same GTDB phyla, which we then considered as replicated below the suggestive significance threshold. As all previous studies followed the NCBI taxonomic nomenclature, we identified the most probable corresponding GTDB phylum using the Taxon History tool from the GTDB website (https://gtdb.ecogenomic.org/taxon_history/).

Bidirectional two-sample Mendelian randomization (MR) analysis (detailed section)

Causal relationships between diseases and gut microbiota were investigated at genus and species levels only to maximise interpretability. In total, 213 species and 148 genera associated with at least one variant at genome-wide significant level ($p<1\times10^{-8}$) were included. GWAS summary results were collected for 46 diseases from MR-Base[36] (**Supplementary Table 4**). These included 12 autoimmune or inflammatory diseases, 9 cardiometabolic diseases, 13 psychiatric or neurological diseases, cardiovascular diseases, 4 bone diseases and 8 cancers. For disease with more than one GWAS records, the record with the largest sample size was kept.

Bi-directional causal inference was performed as follows to infer causal effects of microbial abundance variation (exposure) on disease risk (outcome), and of disease (exposure) on microbial abundance levels (outcome). To select the SNP instruments for microbial exposures in our study (**Supplementary Table 7**), we followed recommendations from a previous study showing that associated SNPs below a significance threshold of $p<1\times10^{-5}$ had the largest explained variance on microbial features[37]. For each taxon, GCTA-COJO was used to perform a conditional analysis to select independently associated SNPs at $p<1\times10^{-5}$. F-statistics were calculated to estimate the strength of instruments for each variable, and were found to be all >10 (**Supplementary Table 5**). SNP instruments for disease exposures were selected at genome-wide significant threshold ($p<5\times10^{-8}$). Subsequently LD-clumping with a strict threshold (r^2<0.001 in 1000G EUR within 10 Mb windows) was conducted to select independent instruments with the lowest *p* values for taxa and diseases, respectively.

Effective alleles of all genetic variants were oriented to the risk-increasing alleles of exposures. For each inference, five different MR methods were used to estimate the causal effect: (1) inverse variance weighted (IVW)[38], (2) weighted median[39], (3) simple mode[40], (4) weighted mode[40] and (5) MR-Egger[41]. IVW is the most sensitive method which requires all instruments are valid. But in reality, it is hard to verify that no any genetic instrument violates any instrumental assumptions. Weighted median only requires at least half of the instruments are valid, making its inference robust to the cases where some instruments violating the assumptions. Simple mode and weighted mode rely on the largest group of similar instruments, reducing the effects of other instruments especially outliers. MR-Egger allows instruments having non-zero pleiotropy and provides way to test and estimate the pleiotropy effect in addition to causal estimate. As these methods are based on different assumptions, the consistency among them indicates a credible estimate[42], even if discrepancy in these methods does not necessarily suggest the absence of causality. A predicted causal estimate was deemed interesting in our study if: (1) it reached a nominal $p<0.05$ for at least three of the five tested methods, (2) directionality testing supported the causal direction, and (3) no significant casual effect in the reverse direction. In addition, MR-PRESSO[43] was used to formally detect and correct for the pleiotropic outliers. Analyses were conducted using the R package *TwoSampleMR*[36].

References

1. Rothschild, D. *et al.* Environment dominates over host genetics in shaping human gut microbiota. *Nature* **555**, 210–215 (2018).
2. Blekhman, R. *et al.* Host genetic variation impacts microbiome composition across human body sites. *Genome Biol* **16**, 191 (2015).
3. Bonder, M. J. *et al.* The effect of host genetics on the gut microbiome. *Nat Genet* **48**, 1407–1412 (2016).
4. Goodrich, J. K., Davenport, E. R., Waters, J. L., Clark, A. G. & Ley, R. E. Cross-species comparisons of host genetic associations with the microbiome. *Science* **352**, 532–535 (2016).
5. Hughes, D. A. *et al.* Genome-wide associations of human gut microbiome variation and implications for causal inference analyses. *Nat Microbiol* **5**, 1079–1087 (2020).
6. Turpin, W. *et al.* Association of host genome with intestinal microbial composition in a large healthy cohort. *Nat Genet* **48**, 1413–1417 (2016).
7. Wang, J. *et al.* Genome-wide association analysis identifies variation in vitamin D receptor and other host factors influencing the gut microbiota. *Nat Genet* **48**, 1396–1406 (2016).
8. Davenport, E. R. *et al.* Seasonal variation in human gut microbiome composition. *PLoS One* **9**, e90731 (2014).