

1 Combined effects of host genetics and diet on human gut microbiota and 2 incident disease in a single population cohort

3
4 Youwen Qin^{1,2}, Aki S. Havulinna³, Yang Liu^{1,4}, Pekka Jousilahti³, Scott C. Ritchie^{1,5-7}, Alex Tokolyi⁸,
5 Jon G. Sanders^{9,10}, Liisa Valsta³, Marta Brożyna¹, Qiyun Zhu¹¹, Anupriya Tripathi^{11,12}, Yoshiki
6 Vazquez-Baeza^{13,14}, Rohit Loomba¹⁵, Susan Cheng¹⁶, Mohit Jain^{11,13}, Teemu Niiranen^{3,17}, Leo Lahti¹⁸,
7 Rob Knight^{11,13,14}, Veikko Salomaa³, Michael Inouye^{1,2,5-7,19-21*}§, Guillaume Méric^{1,22*}§

8
9 ¹Cambridge Baker Systems Genomics Initiative, Baker Heart and Diabetes Institute, Melbourne, Victoria,
10 Australia; ²School of BioSciences, The University of Melbourne, Melbourne, Victoria, Australia; ³Department of
11 Public Health Solutions, Finnish Institute for Health and Welfare, Helsinki, Finland; ⁴Department of Clinical
12 Pathology, The University of Melbourne, Melbourne, Victoria, Australia; ⁵Cambridge Baker Systems Genomics
13 Initiative, Department of Public Health and Primary Care, University of Cambridge, UK; ⁶British Heart
14 Foundation Centre of Research Excellence, University of Cambridge, UK; ⁷National Institute for Health Research
15 Cambridge Biomedical Research Centre, University of Cambridge and Cambridge University Hospitals,
16 Cambridge, UK; ⁸Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK; ⁹Department of Ecology
17 and Evolutionary Biology, Cornell University, Ithaca, NY, USA; ¹⁰Cornell Institute for Host-Microbe Interaction
18 and Disease, Cornell University, Ithaca, NY, USA; ¹¹Department of Pediatrics, School of Medicine, University of
19 California San Diego, La Jolla, CA, USA; ¹²Division of Biological Sciences, University of California San Diego,
20 La Jolla, California, USA; ¹³Center for Microbiome Innovation, University of California San Diego, La Jolla, CA,
21 USA; ¹⁴Department of Computer Science & Engineering, Jacobs School of Engineering, University of California
22 San Diego, La Jolla, CA, USA; ¹⁵NAFLD Research Center, Department of Medicine, University of California San
23 Diego, La Jolla, CA, USA; ¹⁶Smidt Heart Institute, Cedars-Sinai Medical Center, Los Angeles, CA, USA;
24 ¹⁷Department of Medicine, Turku University Hospital and University of Turku, Turku, Finland; ¹⁸Department of
25 Future Technologies, University of Turku, Turku, Finland; ¹⁹British Heart Foundation Cardiovascular
26 Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, UK; ²⁰Health Data
27 Research UK Cambridge, Wellcome Genome Campus & University of Cambridge, UK; ²¹The Alan Turing
28 Institute, London, UK; ²²Department of Infectious Diseases, Central Clinical School, Monash University,
29 Melbourne, Victoria, Australia.

30
31 § These authors contributed equally

32 *Corresponding authors: Michael Inouye: mi336@medschl.cam.ac.uk; Guillaume Méric:
33 guillaume.meric@baker.edu.au.

34

35 **Abstract**

36

37 Co-evolution between humans and the microbial communities colonizing them has resulted in
38 an intimate assembly of thousands of microbial species mutualistically living on and in their
39 body and impacting multiple aspects of host physiology and health. Several studies examining
40 whether human genetic variation can affect gut microbiota suggest a complex combination of
41 environmental and host factors. Here, we leverage a single large-scale population-based cohort
42 of 5,959 genotyped individuals with matched gut microbial shotgun metagenomes, dietary
43 information and health records up to 16 years post-sampling, to characterize human genetic
44 variations associated with microbial abundances, and predict possible causal links with various
45 diseases using Mendelian randomization (MR). Genome-wide association study (GWAS)
46 identified 583 independent SNP-taxon associations at genome-wide significance ($p < 5.0 \times 10^{-8}$),
47 which included notable strong associations with *LCT* ($p = 5.02 \times 10^{-35}$), *ABO* ($p = 1.1 \times 10^{-12}$), and
48 *MED13L* ($p = 1.84 \times 10^{-12}$). A combination of genetics and dietary habits was shown to strongly
49 shape the abundances of certain key bacterial members of the gut microbiota, and explain their
50 genetic association. Genetic effects from the *LCT* locus on *Bifidobacterium* and three other
51 associated taxa significantly differed according to dairy intake. Variation in mucin-degrading
52 *Faecalicatena lactaris* abundances were associated with *ABO*, highlighting a preferential
53 utilization of secreted A/B/AB-antigens as energy source in the gut, irrespectively of fibre
54 intake. *Enterococcus faecalis* levels showed a robust association with a variant in *MED13L*,
55 with putative links to colorectal cancer. Finally, we identified putative causal relationships
56 between gut microbes and complex diseases using MR, with a predicted effect of *Morganella*
57 on major depressive disorder that was consistent with observational incident disease analysis.
58 Overall, we present striking examples of the intricate relationship between humans and their
59 gut microbial communities, and highlight important health implications.

60

61 **Introduction**

62

63 Humans have co-evolved with the microbial communities that colonize them, resulting in a
64 complex assembly of thousands of microbial species mutualistically living in their
65 gastrointestinal tract. A fine-tuned interplay between microbial and human physiologies can
66 impact multiple aspects of development and health to the point that dysbiosis is often
67 associated with disease¹⁻³. As such, increasing evidence points to the influence of human
68 genetic variation on the composition and modulation of their gut microbiota.

69

70 Past genetic studies have collectively revealed important host-microbe interactions⁴⁻¹⁴.
71 Previous twin studies detected significant heritability signal from the presence and abundance
72 of only a few microbial taxa, such as some *Firmicutes*¹⁵, suggesting a strong transientness and
73 variability in gut microbial composition, as well as an important influence from external
74 factors^{6,15-18}. Nonetheless, a well-described association between *Bifidobacterium* levels and
75 *LCT-MCM6*, governing the phenotype of lactase persistence throughout adulthood in
76 Europeans, was uncovered in 2015⁴ and subsequently replicated by later studies^{6,7,9-12},
77 suggesting a very strong influence of the evolution of dairy diet in modern humans on their gut
78 bacteria. Additionally, genes involved in immune and metabolic processes⁹ but also disease¹⁹
79 were also associated with gut microbial variation. Despite several promising findings,
80 reproducibility across studies varying in sampling and methods is generally poor, and most
81 previously reported associations lose significance after multiple testing corrections²⁰. The
82 individual gut microbiota is largely influenced by environmental variables, mostly diet and
83 medication²¹⁻²³, which could explain a larger proportion of microbiome variance than
84 identifiable host genetic factors^{9,10}. Biological factors could also influence the cross-study
85 reproducibility of results. GWAS would typically not reproducibly identify genetic
86 associations with taxa harbouring microbial functions potentially shared by multiple unrelated
87 species^{24,25}. Indeed, a certain degree of functional redundancy has been observed in human gut
88 microbial communities²⁵, which is believed to play a role in the resistance and resilience to
89 perturbations²⁶⁻²⁸. However, both assembly and functioning in human gut microbial
90 communities seem to be driven by the presence of a few particular and identifiable keystone
91 taxa²⁹, which exert key ecological and modulatory roles on gut microbial composition
92 independently of their abundance^{30,31}. Such taxa are relatively prevalent across individuals and
93 thought to be part of the human “core” microbiota^{30,31}, which makes them potentially
94 identifiable through GWAS.

95

96 Increasing sample size in studied populations could yield novel and robustly associated results,
97 and alleviate the effect of confounding technical or biological factors. This could be achieved
98 either by performing meta-analyses of GWAS conducted in various populations¹², or by using
99 larger cohort datasets. In this study, we used a large single homogenous population cohort with
100 matching human genotypes and shotgun faecal metagenomes (N=5959; FINRISK 2002
101 (FR02)) to identify novel genome-wide associations between human genotypes and gut
102 microbial abundances (**Figure S1**). We further leveraged additional and extensive health
103 registry and dietary individual data to investigate the effects of diet and genotype on particular
104 host-microbial associations, and to predict incident disease linked to gut microbial variation.

105

106 **Results**

107

108 **Genome-wide association analysis of gut microbial taxa**

109

110 Genome-wide association tests were applied to 2,801 microbial taxa and 7,979,834 human
111 genetic variants from 5,959 individuals enrolled in the FR02 cohort, which includes all taxa

112 discovered to be prevalent in >25% of the cohort (**Methods**). Using a genome-wide
113 significance threshold ($p < 5.0 \times 10^{-8}$), a total of 478 distinct GTDB taxa, which represented 17%
114 of all tested taxa and included 11 phyla, 19 classes, 24 orders, 63 families, 148 genera and 213
115 species, were found to be associated with at least one genetic variant (**Figure 1, Table S1**).
116 Conditional analysis found 583 independent SNP-taxon associations at genome-wide
117 significance (**Table S1**). Heritability across the 2,801 taxa ranged between $h^2 = 0.001$ to 0.214,
118 with the highest values observed for taxa belonging to the *Firmicutes* and *Firmicutes_A* GTDB
119 phyla, both of which encompassed half (241/476, 50.4%) of all associated taxa with genetic
120 variation (**Figure S2**). There were no differences in SNP heritability between groups of
121 associated or non-associated taxa at genome-wide significance ($p = 0.23$).

122
123 Three loci were strongly associated with microbial variation at study-wide significance, as
124 shown on a Manhattan plot showing the lowest resulting p-value for each SNP tested against
125 each of the 2,801 taxa (**Figure 1, Table 1**). There was no evidence of excess false positive rate
126 in the GWAS (median $\lambda_{GC} = 1.0051$) (**Figure 1B**). After conditional analysis, the strongest
127 association by far ($p = 5.0 \times 10^{-35}$) involved members of class *Actinobacteria* and rs3940549, a
128 variant in the *LCT-MCM6-ZRANB3* locus region which is in high LD ($r^2 = 0.87$) with the well-
129 described *LCT* variant rs4988235 causing lactase persistence in adults of European ancestry
130 (**Figure S3**). In total, 29 taxa were associated with the *LCT-MCM6* region, including 18 below
131 study-wide significance (**Figure 1, Table S1**). These involved *Bifidobacterium*-related
132 *Actinobacteriota* and three taxa from the GTDB *Firmicutes_A* phylum which included 2
133 uncultured species defined from metagenome-assembled reference genomes (*UBA3855*
134 *sp900316885* and *CAG-81 sp000435795*) (**Table 1**). The association of these three
135 *Firmicutes_A* with *LCT* was still genome-wide significant after adjusting for *Bifidobacterium*
136 abundances (**Table S2**). A variant in *ABO* (rs545971), expressing the histo-blood
137 group ABO system transferase, was strongly associated ($p = 1.1 \times 10^{-12}$) with levels of
138 *Faecalicatena lactaris*. There was evidence for a second independent signal at *ABO* associated
139 with the *Collinsella* genus (chr9:133271182; $p = 2.5 \times 10^{-8}$) (**Table S1, Figure 1**). Rs187309577
140 and rs143507801 in *MED13L*, expressing the Mediator complex subunit 13L, were found to be
141 associated with genus *Enterococcus* ($p = 1.8 \times 10^{-12}$) and the *Enterococcus faecalis* species
142 ($p = 7.26 \times 10^{-11}$), respectively (**Table S1, Figure 1**).

143

144 **Human gut microbiome keystone taxa are associated with genetic variation**

145

146 In total, we identified 31 distinct genetic variants associated ($p < 5.0 \times 10^{-8}$) with 39 microbial
147 taxa related to identified keystone species as listed by Banerjee *et al.* (2018)^{29,32}, which
148 included the *Actinobacteria* class³⁰, *Helicobacter pylori*²⁹, *Bacteroides stercoris*³³, *Bacteroides*
149 *thetaiotaomicron*³⁴, *Ruminococcus bromii*³⁵, *Klebsiella pneumoniae*³⁶, *Proteus mirabilis*³⁶,
150 *Akkermansia muciniphila*³¹, and the archaeon *Methanobrevibacter smithii*^{37,38} (**Figure 1C**,
151 **Table S1**). Only one documented keystone species from Banerjee *et al.*²⁹, *Bacteroides*
152 *fragilis*³⁹, was not associated with genetic variation in our study. This observation suggests that
153 keystone species, although defined as exerting selective modulation and not broad effects on
154 microbiome composition variation, generally associates with human genetic variation,
155 suggesting an intimate association with the human gut niche, in line with their reported key
156 ecological roles in microbiome modulation and functioning. Our work highlights novel human
157 genotypes possibly associated with keystone taxa (**Table S1**), which could further improve our
158 understanding of their ecology.

159

160 **Combined effect of host genetics and dietary dairy intake on gut levels of *LCT*-associated** 161 **bacteria**

162

163 We compared the abundances of 4 bacterial taxa strongly associated with the *LCT* locus
164 (*Bifidobacterium* genus, *Negativibacillus* genus, *UBA3855 sp900316885* and *CAG-81*
165 *sp000435795*) in individuals with different rs4988235 genotypes and dairy diets (**Figure 2A**).
166 The abundance of *Bifidobacterium* in individuals producing lactase through adulthood
167 (rs4988235:TT) was unaffected by dairy intake. However, lactose-intolerant individuals
168 (rs4988235:CC) self-reporting a regular dairy diet had a significant increase in *Bifidobacterium*
169 abundance ($p=1.75\times 10^{-13}$; Wilcoxon-rank test). An intermediate genotype (rs4988235:CT) was
170 linked to an intermediate increase (**Figure 2A**). This trend did not seem to be affected by age⁴⁰
171 (**Figure S4**).

172
173 An inverse pattern was observed for the abundance distributions of *Negativibacillus* and
174 uncultured *CAG-81 sp000435795*, for which abundances decreased in lactose intolerant
175 individuals reporting dairy intake, as compared to rs4988235:TT individuals consuming dairy
176 products ($p=0.049$ and $p=0.041$, respectively) (**Figure 2A**). Levels of *UBA3855 sp900316885*
177 were unaffected by a dairy diet in lactose-intolerant individuals but were surprisingly lower in
178 rs4988235:TT individuals who reported dairy intake ($p=8.23\times 10^{-5}$) (**Figure 2A**). These
179 opposite and contrasting effects of dairy intake on associated bacterial abundances in lactose-
180 intolerant individuals could reflect competition for lactose in the gut. Genus *CAG-81*
181 abundances were the most negatively correlated with those of the other *LCT*-associated taxa
182 (**Figure S5**), which suggests that this competition could be strong and prevalent enough to
183 drive co-association at the *LCT* locus, possibly mediated by lactose intake (**Figure 2B**).

184

185 **Functional profiling of CAZymes in 11 *Bifidobacterium* species**

186

187 Of all 11 *Bifidobacterium* species prevalent enough in our study population to be included in
188 the GWAS, only *B. dentium* was not associated with the *LCT* locus ($p=1.70\times 10^{-2}$), nor was it
189 co-abundant with any other *Bifidobacterium* species (**Figure S6A**). *B. dentium* has previously
190 been suggested to have different metabolic abilities⁴¹. A clustering of carbohydrate-active
191 enzymes (CAZyme) profiles from reference genomes of all 11 *Bifidobacterium* species
192 revealed that *B. dentium* clustered apart from the 10 other species, which grouped consistently
193 with their co-abundance patterns (**Figure S6B**). *B. dentium* harboured more genes encoding
194 CAZyme families with preferred fiber/plant-related substrates (GH94, GH26, GH53) than
195 other *Bifidobacterium* species, which seemed to harbour more milk oligosaccharide-targeting
196 CAZyme families (GH129, GH112) than *B. dentium* (**Figure S6B**), which could relate to the
197 observed association differences. This suggests that bacterial metabolic abilities can be strong
198 drivers of co-abundance, and of association with human genetic variation.

199

200 **Functionally distinct *ABO*-associated bacteria are impacted differently by genotype and** 201 **dietary fiber intake**

202

203 A variety of bacteria metabolize blood antigens, with potential applications in synthetic
204 universal donor blood production^{42,43}. Gut bacteria are particularly exposed to A- and B-
205 antigens in the gut mucosa of secretor individuals⁴⁴. Our associations of *Faecalicatena lactaris*
206 ($p=1.10\times 10^{-12}$) and *Collinsella* ($p=2.59\times 10^{-8}$) with *ABO* suggest a possible metabolic link with
207 blood antigens. A comparison of CAZyme profiles across a set of reference genomes revealed
208 3 CAZymes with blood-related activities in *F. lactaris* (GH110⁴⁵, GH136⁴⁶, CBM32⁴⁷), but
209 none in any of 9 *Collinsella* species (**Figure 3A**). More mucus-targeting and less fiber-
210 degrading enzymes were found in *F. lactaris* than *Collinsella* (**Figure 3A**), suggesting distinct
211 functions in the gut.

212

213 As previously reported⁵, neither ABO blood types, nor secretor status had an impact on alpha
214 and beta diversity (**Figure S7**). However, we observed that the effect of *ABO* genotypes on *F.*
215 *lactaris* levels, underlying the association, were largely driven by secretor status, with
216 increased abundances in secretor individuals from genotype groups rs545971:CT ($p=3.6\times 10^{-4}$)
217 and rs545971:TT ($p=9\times 10^{-4}$), A ($p=1.24\times 10^{-5}$) and AB blood type groups ($p=1.24\times 10^{-5}$), but
218 not in rs545971:CC genotype ($p=0.4339$), or B and O blood types individuals (**Figure 3B**).
219 Levels in non-secretors did not vary across *ABO* genotypes or blood types (**Figure 3B**).
220 Despite a slight increase in blood type A secretors, *Collinsella* only remained minimally
221 affected by secretor status or blood group (**Figure S8A**). Taken together, this suggests that the
222 secretion of soluble A and B-antigens strongly affects *F. lactaris* in the gut, possibly through
223 reduced opportunity to use them as substrate. Both levels of *F. lactaris* and *Collinsella* were
224 significantly higher when individuals were predicted to secrete A-, B- and AB-antigens in their
225 gut mucosa ($p<2.2\times 10^{-16}$ and $p=1.3\times 10^{-8}$, respectively) (**Figure S8B**).

226
227 A high fiber diet is thought to induce a metabolic transition from mucus-degrading to fiber-
228 degrading activities in the colon, as carbohydrates from fiber are more easily metabolized⁴⁸.
229 The increase in *F. lactaris* abundances in A/B/AB-secretors (defined as secreting A-, B- and
230 AB-antigens) compared to non- A/B/AB-secretors remained strongly significant irrespective of
231 fiber intake ($p=1.15\times 10^{-9}$ in the low-fiber diet group, and $p=4.4\times 10^{-3}$ in the high-fiber diet
232 group), suggesting that either *F. lactaris* has a strong affinity for secreted A/B/AB-antigens,
233 does not efficiently degrade dietary fiber, or will not easily switch to it as an energy source
234 (**Figure 3C**). *F. lactaris* levels were increased in non-A/B/AB-secretors with a high fiber diet,
235 implying a switch to fiber degradation or interaction with fiber-degrading bacteria (**Figure**
236 **3C**). *Collinsella* variation in both A/B/AB-secretors and non-A/B/AB-secretors with high- and
237 low-fiber diets was similar to the compounded abundances of 13 major mucin-degrading
238 species in the human gut⁴⁹, suggesting a similar ecological response in stark contrast with *F.*
239 *lactaris* (**Figure 3C, Figure 3D**).

240 241 **MED13L association with *Enterococcus faecalis* as a putative link with CRC development**

242
243 The allele frequency of the *MED13L* rs143507801 variant (A>G), associated with levels of
244 *Enterococcus faecalis* ($p=7.26\times 10^{-11}$), was low (MAF=0.0111), consistent with reported allele
245 frequencies in the gnomAD database⁵⁰. In our study population, 131 individuals carried
246 rs143507801:G allele, 130 being heterozygous (GA) and only one being homozygous (GG).
247 We observed that *E. faecalis* levels were increased in heterozygous rs143507801:GA
248 individuals (**Figure 4**). *E. faecalis* is a gut commensal, but also an opportunist pathogen
249 believed to play a role in colorectal cancer (CRC) development, possibly through direct
250 damaging of colorectal cells⁵¹⁻⁵⁶. *MED13L* and *MED13* encode for Mediator transcriptional
251 coactivator complex modules associating with RNA polymerase II⁵⁷, and as such specifically
252 interact with cyclin-dependent kinase 8 (CDK8) modules described for their oncogenic
253 activation of transcription during colon tumorigenesis⁵⁸. Consequently, we observed slightly
254 higher levels of *E. faecalis* ($p=0.014$) in 14 individuals enrolled in FR02 who had prevalent
255 CRC at the time of sampling (**Figure 4**). Groups of individuals segregated by allelic variant
256 and CRC status could not be compared robustly due to small sample size. Taken together, these
257 results suggest a possible link between *E. faecalis* and CRC through the MED13 activation of
258 CDK8 in colorectal tumours, which will need to be investigated further.

259 260 **Causal inference predictions between microbes and diseases highlight causal effect of** 261 ***Morganella* on MDD**

262

263 Interpreting results of causal inference prediction using bacterial information entails to
264 particular caution, due to the possibility of multiple and unaccounted confounding factors¹¹,
265 but can be useful to highlight potential focus for future research. Here, we predicted 96 causal
266 effects in both microbe to disease and disease to microbe directions using bidirectional
267 Mendelian Randomization (MR). Of these, 34 were from microbial levels as exposure to
268 disease as outcome, with a large proportion of causal effects in psychiatric and neurological
269 diseases (**Table S5**). For example, MR suggested an increased abundance of *Faecalicoccus*
270 may have a causal effect on anorexia nervosa (OR=1.8 per SD increase in bacterial
271 abundance; CI_{95%}=1.3-2.5; $p=2.0\times 10^{-4}$, MR method IVW)(**Methods**). Other examples included
272 increasing abundances of *Morganella* and *Raoultella* predicted to have causal effects on major
273 depressive disorder (MDD) (**Table S5**). When MR was performed in the reverse direction,
274 using disease risk as an exposure and microbial levels as an outcome, most predicted causal
275 effects involved autoimmune and inflammatory diseases but the strongest predicted causal
276 effect involved type 2 diabetes (T2D) (**Table S6**). Doubling the genetic risk of T2D (possibly
277 accompanied by external factors such as hypoglycaemic medications or metformin intake) was
278 predicted to reduce levels of the uncultured *CAG-345 sp000433315* species (*Firmicutes*
279 phylum) by 0.14 SD (SE=0.04, $p=3.0\times 10^{-4}$, MR method IVW). A few other examples included
280 some degree of literature validation, such as the higher genetic risk for primary sclerosing
281 cholangitis (PSC) causally impacting levels of the cholesterol-reducing *Eubacterium_R*
282 *coprostanoligenes*⁵⁹. Furthermore, a higher genetic risk for coeliac disease (CD) was predicted
283 to increase abundances in 4 species previously reported to be more abundant in CD patients
284 than controls⁶⁰ (**Table S6**). Finally, a higher genetic risk for multiple sclerosis (MS) was
285 predicted to cause a reduction in the abundance of *Lactobacillus_B ruminis*, consistent with the
286 report that *Lactobacillus sp.* can reduce symptom severity in an animal model of MS⁶¹.

287
288 The availability in our study dataset of up to 16 years of electronic health record follow-up
289 after the initial sampling of the microbiota allowed for observational validation of predicted
290 effects using MR. Of all causal predictions identified using MR, only the effect of *Morganella*
291 on MDD could be validated by a statistically significant association with incident MDD
292 (HR=1.11, CI_{95%}=1.01-1.22, per SD increase of bacterial abundance), after accounting for age,
293 sex and BMI (**Figure 5**). In our GWAS, *Morganella* variation in the study population
294 associated with a variant (rs192436108; $p=6.16\times 10^{-8}$) in the *PDE1A* locus, which has
295 previously been linked to depression^{62,63} and psychiatric disorders⁶⁴. Taken together, these
296 predicted links between *Morganella* and MDD suggest more efforts should be deployed into
297 exploring the possible roles of this bacterium as part of the brain-gut axis metabolic
298 modulation of health.

299

300 **Discussion**

301

302 Here, through GWAS and the subsequent investigation of functional and ecological factors
303 contributing to the most robust human-microbe associations, we present a diverse and global
304 picture of human-microbe interactions in a single cohort of ~6,000 European individuals. We
305 find 3 genetic loci to be strongly associated with gut microbial variation. Two of these loci,
306 *LCT* and *ABO*, are well-known and very segregated in human populations, possibly explaining
307 why our homogenous European cohort identified them as being associated so strongly. A third
308 more mysterious association with the *MED13L* locus highlights possible links with cancer
309 while predictive causal inference highlights several diseases as being causally linked to gut
310 microbes.

311

312 **Lactase persistence as a recently evolved strong modulator of gut bacterial abundances**

313

314 Lactase persistence, or the continued ability to digest lactose into adulthood, is the most
315 strongly selected single-gene trait over the last 10,000 years in multiple human populations⁶⁵,
316 believed to have spread amongst humans with the advent of animal domestication and the
317 culturally transmitted practice of dairying⁶⁶. In our study, as in previous work^{4,6,7,11,12}, the
318 association of *LCT* variants with *Actinobacteria*, more specifically *Bifidobacterium*, is by far
319 the most statistically significant, suggesting a profound interaction between *Actinobacteria* and
320 the human gut, in line with their reported keystone activities³⁰. We reported a strong increase
321 of *Bifidobacterium* levels in genetically lactose intolerant people reporting a regular
322 consumption of dairy products⁹. This increase was not confounded by age in adults, despite
323 *Bifidobacterium* levels generally decreasing with age in our cohort. While self-reported dietary
324 information is not entirely reliable due to various social reasons^{67,68}, our study population was
325 large, and the differences were significant enough to consider this a robust observation. These
326 observations can be explained by the evolutionary adaptation of *Bifidobacterium* species to
327 specifically use human and bovine milk oligosaccharides as an energy source⁶⁹. In adults
328 unable to produce lactase in their small intestines, consumed lactose is likely to become
329 available for colonic bacteria as an energy source to compete for (**Figure 3A**). Hints of a
330 possible competitive relationship between *Bifidobacterium* and *Negativibacillus*, another *LCT*-
331 associated taxon were revealed, which could be mediated by lactose intake and will need to be
332 investigated further in functional studies.

333

334 Two interesting questions stem from our findings. First, the genetic determinants of lactose
335 intolerance are known to vary across ethnicity⁷⁰ and cross-population heterogeneity in the
336 *LCT-Bifidobacterium* association was recently reported¹². As more non-European-centric
337 genetic studies are conducted worldwide^{12,71,72}, examining this combined interaction between
338 dairy diet and *Bifidobacterium* in different genetic backgrounds could bring new insights.
339 Secondly, despite recent progresses, lactose intolerance is still largely underdiagnosed, and
340 genetic prediction rates from large population studies exceed lactose intolerance prevalence
341 rates obtained using physical tests⁷⁰. In our work, we lacked information on lactose
342 malabsorption symptoms in lactose intolerant individuals reporting a regular dairy diet. These
343 people could experience discomfort symptoms without knowingly implicating their own
344 lactose intake, but another possibility could be that the ability of *Bifidobacterium* to degrade
345 lactose may alleviate the perceived symptoms of discomfort associated with lactose
346 intolerance, therefore encouraging individuals to unknowingly continue consuming lactose that
347 they would otherwise not be able to digest⁷³. This possible probiotic effect would be interesting
348 to investigate in controlled studies.

349

350 **Blood antigen secretion can influence levels of specific gut microbial commensals**

351

352 The *ABO* gene expresses a glycosyltransferase in many cell types, which determines the ABO
353 blood group of an individual by modifying the oligosaccharides on cell surface glycoproteins.
354 A comparison of humans and non-human primates has identified *ABO* (along with the MHC)
355 as harbouring ancient multiallelic polymorphisms that are maintained across species^{74,75}.
356 Evolutionary selective pressures at this locus have been proposed to be linked to pathogen
357 infection. Indeed, many infectious diseases such as norovirus infection, bacterial meningitis,
358 malaria, cholera⁷⁶, or even more recently SARS-CoV-2^{77,78} are associated with host blood type
359 and secretor status⁷⁶, suggesting that infection could be a driver of a strong balancing selection
360 that has maintained *ABO* polymorphisms. Furthermore, blood type variation has been
361 intriguingly linked to various chronic diseases⁷⁶, such as heart and vascular diseases, gastric
362 cancers, diabetes, asthma or even dementia⁷⁶. Many of these chronic diseases are also
363 associated with dysbiosis of the gut microbiota, which prompts interesting but largely
364 unexplored parallel between gut commensals, blood types and disease⁴⁴. Our study confirms

365 previous findings⁵ that secretor status or blood types do not seem to globally affect gut
366 microbial alpha- or beta-diversity. It also confirms reports from two very recent studies: the
367 first of these studies, a meta-analysis across five German cohorts, using 16S rRNA sequencing
368 to characterize the gut microbiota, linked *Bacteroides* and *Faecalibacterium* to *ABO* and
369 *FUT2*⁷⁹. The second study, taking a functional approach, intriguingly associated bacterial
370 lactose and galactose degradation genes to *ABO* variation in a cohort of 3,432 Chinese
371 individuals⁸⁰. Taken together, these findings suggest a broad association of *ABO*
372 polymorphisms with microbial variation in various human populations.

373
374 An important research effort aiming to enzymatically produce synthetic universal donor blood
375 has driven a push for screening a large diversity of CAZymes, including bacteria, revealing
376 substrate affinities for blood antigens across various microbes^{42,43}. Here we highlight *F.*
377 *lactaris* (formerly *Ruminococcus lactaris*), as a mucin-degrading commensal likely able to
378 digest blood antigens through its predicted harbouring of GH110, GH136 and CBM32
379 CAZyme family genes⁴⁵⁻⁴⁷. *F. lactaris* is strongly associated with *ABO* genetic variation in our
380 European cohort, and is differentially abundant in people according to their predicted gut
381 mucosal secretion of A/B/AB-antigens. Interestingly, our findings are not consistent with *F.*
382 *lactaris* switching to a fiber-degrading activity in individuals reporting a high fiber diet, unlike
383 other mucin-degrading bacteria in our study and in the literature⁴⁸ and *Collinsella*, another
384 *ABO*-associated taxon (**Figure 3B**). Our work suggests that some gut commensals such as *F.*
385 *lactaris* appear to be very efficient and adapted metaboliser of A/B/AB-antigens in the gut,
386 despite their predicted ability to degrade simpler carbohydrates in fiber. This could be an
387 example of ecological niche differentiation in the gut, with impacts on associated *F. lactaris*
388 microbial communities, of which *Collinsella*, also associated with *ABO*, may belong.

389 390 **Unexplored links with disease and the nervous system**

391
392 Although validation of the association is inconclusive because of the low prevalence of CRC
393 cases and genetic variation in our study population, the association of *MEDI3L* rs143507801
394 variant with *Enterococcus faecalis* suggested a putative link with CRC. It has been shown that
395 *MEDI3* could directly link a cyclin-dependent kinase 8 (CDK8) module to Mediator^{81,82},
396 which is a colorectal cancer oncogene, amplified in colorectal tumours and activating
397 transcription driving colon tumorigenesis leading to CRC⁵⁸. This could explain a long
398 suspected link between *Enterococcus faecalis* and development of CRC after having been
399 found in higher concentrations in CRC patients than healthy individuals⁵¹⁻⁵⁵. The suspected
400 mode of action of *E. faecalis* on CRC development is currently unclear, but could be linked to
401 extracellular free radical production directly leading to DNA break, point mutation and
402 chromosomal instability in colorectal cells⁵⁶. Although we saw a trend of *E. faecalis* being
403 increased in abundance in prevalent CRC patients, and in *MEDI3L* variation, more focused
404 work and a larger sample size will be required to precisely pinpoint a link between this
405 bacterium and CRC through the Mediator complex, if any.

406
407 Causal inference analysis highlighted a very promising example of interplay between a gut
408 microbe and a complex disease. Among other suggested links with psychiatric diseases, we
409 predicted that increasing abundances of *Morganella* and *Raoultella* could have causal effects
410 on MDD. Members of the *Enterobacteriaceae* family, such as these two genera, have
411 previously been found in higher levels in MDD patients⁸³. Although caution is required when
412 interpreting predictions of causality⁸⁴, several studies elaborated the gut-brain axis hypothesis,
413 and increasing evidence suggests that gut microbes are likely to influence host behavior via a
414 systemic modulation of hormones and metabolites⁸⁵⁻⁸⁷. Most importantly, our MR-based
415 observation was consistent with observed hazards using follow-up observational data up to 16

416 years after initial sampling. This observation supports previous experimental results showing
417 an increase of IgM and IgA-related immune response against *Morganella* secreted
418 lipopolysaccharide in major depression⁸⁸. This finding potentially highlights the intimate
419 influence of the gut-brain axis on humans.

420

421 Our MR analysis suggested that known genetic risks of autoimmune and inflammatory
422 diseases could also influence gut microbes. One explanation could be that disease susceptibility
423 would affect host immunity and gut barrier integrity, which may favor an increase in some key
424 microbes. However, several studies have shown that manipulating gut microbial composition
425 could be a potential therapy for autoimmune and inflammatory diseases⁸⁹, which would
426 suggest that composition variation in specific gut microbe maybe a requirement for the
427 penetration of a disease phenotype⁹⁰. Further mechanistic studies are needed to untangle host-
428 microbe interactions in disease, and further interpret these predictions.

429

430 **The case for larger datasets and including uncultured novel species in metagenomic** 431 **studies**

432

433 Our study highlights the benefits of increasing sample size to increase the statistical power for
434 discovery. Although the *LCT* locus has been reported multiple times to be associated with
435 bacterial taxa, our work is the first to report study-wide significant associations in a single
436 cohort, at the strongest significance ever reported. The association with *Bifidobacterium* in our
437 study was even stronger than the recent findings that used integrative data from 18,473
438 individuals in 28 different cohorts¹², emphasizing the importance of standardized methodology
439 and homogeneity in participant ethnicity (especially when studying highly geographically
440 distributed traits such as lactose intolerance traits⁹¹). *ABO* allelic variation is also notoriously
441 affected by geography⁹², which could explain why some meta-analyses in non-homogenous
442 populations could miss it or not. Importantly, metagenomic sequencing with standardized,
443 robust taxonomic definitions^{93,94} can provide species-level characterization of microbial
444 profiles in the gut of individuals, which is challenging when using 16S rRNA-based studies.
445 An example from our work is the observation that *Bifidobacterium dentium* was prevalent but
446 not associated with the *LCT* locus like all other *Bifidobacterium* species in the population.
447 Observed difference in carbohydrate-active enzymes that are commonly found in other
448 *Bifidobacterium* species may explain this difference⁴¹. Furthermore, GTDB taxonomic
449 standardization results in greater taxon granularity, i.e. smaller, more discrete clades of similar
450 phylogenetic depth than commonly known lineages or species^{93,94}. In theory, this would
451 increase overall accuracy⁹⁵, as a weak association with a poorly-defined lineage may be caused
452 by a strong association with a well-defined subset of that lineage, defined as a coherent group
453 using GTDB⁹⁴. Finally, a myriad of microbial taxa that are to date solely defined and
454 represented by uncultured metagenome-assembled genomes (MAGs) in the GTDB database
455 were found to be independently associated with various loci. Along with recent reports that the
456 more gut microbiome diversity is explored, the more novel, unknown species are
457 discovered^{96,97}, this suggests that many discoveries are yet to be made in the field of human
458 microbiome studies.

459

460 **Material and methods**

461

462 **Study population**

463

464 The FINRISK study population has been extensively described elsewhere⁹⁸. FINRISK
465 population surveys have been performed every 5 years since 1972 to monitor trends in
466 cardiovascular disease risk factors in the Finnish population^{98,99}. The FINRISK 2002 (FR02)
467 study population has been extensively described elsewhere^{98,100}. Briefly, it was based on a
468 stratified random sample of the Finnish population aged between 25 and 74 years from six
469 geographical areas of Finland¹⁰¹. The sampling was stratified by sex, region and 10-year age
470 group so that each stratum had 250 participants. The overall participation rate was 65.5% (n =
471 8,798). Selected participants filled out a questionnaire, then participated in a clinical
472 examination carried out by specifically trained nurses and gave a blood sample from which
473 various laboratory measurements were performed. They also received a sampling kit and
474 instructions to donate a stool sample at home and mailed it to the Finnish Institute for Health
475 and Welfare in an overnight mail. The follow-up of the cohort took place by record linkage of
476 the study data with the Finnish national electronic health registers (Hospital Discharge Register
477 and Causes of Death Register), which provide in practice 100% coverage of relevant health
478 events in Finnish residents. For present analyses involving follow-up data, we used a follow-up
479 which extended until 31/12/2018.

480

481 The study protocol of FR02 was approved by the Coordinating Ethical Committee of the
482 Helsinki and Uusimaa Hospital District (Ref. 558/E3/2001). All participants signed an
483 informed consent. The study was conducted according to the World Medical Association's
484 Declaration of Helsinki on ethical principles.

485

486 **Cohort phenotype metadata and specific dietary information**

487

488 The phenotype data in this study comprised of demographic characteristics, life habits, disease
489 history, laboratory test results and follow-up electronic health records (EHRs). More
490 specifically, baseline dietary factors were collected. Participants were asked to provide answers
491 to exhaustive diet questionnaires when they were enrolled in the study. Details of the method
492 have been described previously⁹⁹. To broadly assess diet information within the cohort
493 participants, a binary variable was used to indicate whether individuals were self-reporting to
494 follow various possible dietary restrictions. Dietary consumption of specific food product
495 categories was also reported.

496

497 **Self-reporting of lactose-free diet and dietary fibre consumption**

498

499 Allelic distribution at the *LCT-MCM6*:rs4988235 variant responsible for lactase persistence in
500 Europeans was as following in our study population: 1,936 (35%) individuals had the T/T
501 allele conferring a lactase persistence phenotype through adulthood, allowing them to digest
502 lactose, while 981 (18%) individuals had the C/C allele conferring lactose intolerance. Most
503 individuals (n=2,611, 47%) had the intermediate allele C/T making them likely to be able to
504 digest lactose. Most individuals reported a regular dairy intake in their diet (n=5,002, 89%),
505 while 706 (12.5%) individuals reported a regular lactose-free diet.

506

507 A total fiber consumption score was calculated from the questionnaires, reflecting the overall
508 consumption of a combination of various fiber-rich foods such as high-fiber bread, vegetables
509 (vegetable foods, fresh and boiled) and berries (fruits, berries and natural juices). The resulting
510 total fiber index values ranged from 9 (low dietary fiber intake) to 48 (high dietary fiber

511 intake), with a median of 33. Comparisons of the effects of low- vs. high-fiber diets were made
512 between the 1st (n=1,213) and 4th (n=1,132) quartiles of the total fiber index.

513

514 **Genotyping, imputation and quality control**

515

516 The genotyping was performed on Illumina genome-wide SNP arrays (the HumanCoreExome
517 BeadChip, the Human610-Quad BeadChip and the HumanOmniExpress) and has been
518 described previously¹⁰². Stringent criteria were applied to remove samples and variants of low
519 quality. Samples with call rate <95%, sex discrepancies, excess heterozygosity and non-
520 European ancestry were excluded. Variants with call rate <98%, deviation from Hardy-
521 Weinberg Equilibrium ($p < 1 \times 10^{-6}$), and minor allele count < 3 were filtered. Data was pre-
522 phased by using Eagle2 v2.3¹⁰³. Imputation was performed using IMPUTE2 v2.3.0¹⁰⁴ with two
523 Finnish-population-specific reference panels: 2,690 high-coverage whole-genome sequencing
524 and 5,092 whole-exome sequencing samples. To evaluate the imputation quality, we compared
525 the sample allele frequencies with reference populations and examined imputation quality
526 (INFO scores) distributions. Imputed SNPs with INFO >0.7 were kept for analysis. Post
527 imputation quality control was carried out by using plink v2.0¹⁰⁵. Samples with >10% missing
528 rate were removed. Individuals with extreme height or BMI values were further excluded (31
529 individuals with height < 1.47m; 5 with BMI > 50 were removed). Both genotyped and imputed
530 SNPs were kept for analysis if they met the following criteria: call rate >90%, no significant
531 deviation from Hardy-Weinberg Equilibrium ($p > 1.0 \times 10^{-6}$), and minor allele frequency >1%.
532 The post-QC dataset comprised 7,980,477 SNPs.

533

534 **Metagenomic sequencing from stool samples**

535

536 Stool samples were collected by participants and mailed overnight to Finnish Institute for
537 Health and Welfare for storing at -20°C; the samples were sequenced at the University of
538 California San Diego in 2017. The gut microbiome was characterized by shallow shotgun
539 metagenomics sequencing with Illumina HiSeq 4000 Systems. We successfully performed
540 stool shotgun sequencing in n=7,231 individuals. The detailed procedures for DNA extraction,
541 library preparation and sequence processing have been previously described¹⁰¹. Adapter and
542 host sequences were removed. To preserve the quality of data while retaining most of the
543 disease cases, samples with a total number of sequenced reads lower than 400,000 were
544 removed.

545

546 **Taxonomic profiling, quality filtering and data transformation**

547

548 Taxonomic profiling of FR02 metagenomes has been described elsewhere^{100,106}. Briefly, raw
549 shotgun metagenomic sequencing reads were mapped using the *k*-mer-based metagenomic
550 classification tool Centrifuge¹⁰⁷ to an index database custom-built to encompass reference
551 genomes that followed the taxonomic nomenclature introduced and updated in the GTDB
552 release 89⁹³⁻⁹⁵. This implies that unless specified otherwise, all taxonomic names in our study
553 refer to their nomenclature in GTDB, which can be related to the original NCBI nomenclature
554 using the GTDB database server: https://gtdb.ecogenomic.org/taxon_history/.

555

556 Gut microbial composition was represented as the relative abundance of taxa. For each
557 metagenome at phylum, class, order, family, genus and species levels, the relative abundance
558 of a taxon was computed as the proportion of reads assigned to the clade rooted at this taxon
559 among total classified reads. The relative abundance of a taxon with no reads assigned in a
560 metagenome was considered as zero in the corresponding profile. For the purpose of this
561 association study and because of reduced accuracy and power when considering rare taxa, we

562 focused on common and relatively abundant microbial taxa, defined as prevalent in >25%
563 studied individuals, and defined with at least 10 mapped reads per individual. For the purpose
564 of association, and as previous studies have reported that only some microbial taxa are
565 inheritable¹⁰⁸, we also removed taxa with zero SNP-heritability. This filtering resulted in a
566 microbial dataset composed of a total of 2,801 taxa, including 59 phyla, 95 classes, 187 orders,
567 415 families, 922 genera and 1,123 species.

568

569 Taxonomic profiles derived from sequencing data are by nature compositional because of an
570 arbitrary total imposed by the instrument¹⁰⁹. The compositional data of microbial taxa is not
571 independent and can lead to inappropriate use of linear regression. To overcome this artificial
572 bias, all relative abundance values were transformed by centre-log-ratio (CLR)¹¹⁰. CLR
573 transformed data can vary in real space and better fit the normality assumption of linear
574 regression. To minimize the impact of zeros, the reads count profiles were shifted by +1 before
575 the transformation. This process was performed using the R package *compositions*. When
576 visually comparing relative abundances in groups of individuals throughout the manuscript, we
577 used untransformed relative abundances, for better interpretability. Alpha (Shannon index) and
578 beta (Bray-Curtis distance) diversity were calculated at genus level used functions in the R
579 package *vegan*.

580

581 **Genome-wide association analysis**

582

583 The protocol followed in this study was described elsewhere¹¹¹. Briefly, linear mixed model
584 (LMM) implemented in BOLT-LMM¹¹² was used to search for genome-wide associations
585 accounting for the individual similarity. Since BOLT-LMM only accepts <1 million SNPs in
586 modelling the genetic relationship matrix, SNPs were pruned at the threshold of $r^2 < 0.1$
587 (plink2¹⁰⁵, command *--indep-pairwise 1000 80 0.1*), resulting in 106,201 independent SNPs.
588 BOLT-LMM automatically performs leave-one-chromosome-out (LOCO) analysis to avoid
589 proximal contamination. Although LMM accounts for the cryptic relatedness in individuals,
590 there are still large population structure cannot be addressed. Thus, the top 10 genetic principal
591 components (calculated by FlashPCA2¹¹³ based on the pruned SNPs mentioned above) were
592 included as covariates. Age, gender, and genotyping batch were adjusted. As no genetic variant
593 was reported to have large effect size on gut microbiota, statistic estimates were based on
594 infinitesimal model which assumes small non-zero effect for large number of genetic variants.
595 To identify independent associations, GCTA-COJO¹¹⁴ was used to conduct approximate
596 conditional and joint analysis using individual genetic data. Window size was set to 10 Mb,
597 assuming SNPs on different chromosomes or more than 10 Mb distance are uncorrelated. The
598 resulting effect size (beta coefficient) indicated the number of standard deviation changes of a
599 taxon's CLR transformed abundance corresponding to one effective allele increase of SNP.

600

601 As microbes interact non-independently with each other in the gut, as part of larger ecological
602 and functional communities, matSpDlite^{115,116} was used to estimate the number of independent
603 tests based on eigenvalue variance, the larger the eigenvalue variance the smaller the number
604 of effective tests. The number of independent tests was 1,328 for 2,801 tested taxa. We used
605 this information to calculate a Bonferroni-adjusted study-wide significant level for significant
606 associations, which was set to $5 \times 10^{-8} / 1328 = 3.8 \times 10^{-11}$. A genome-wide significant threshold
607 was set as 5×10^{-8} .

608

609 **Prediction of ABO blood groups and secretor status**

610

611 SNP-based typing of ABO histo-blood group was performed. A combination of four SNPs¹¹⁷
612 was used for the prediction, and a 98% concordance with phenotypically typed ABO histo-

613 blood group has been reported for this method⁵. For blood group allele A, the two different
614 types A1 and A2 were predicted by rs507666 and rs8176704 respectively. Blood group allele B
615 was inferred from rs8176746 and blood group allele O was predicted by rs687289. As the
616 combination of these SNPs are exclusive, no haplotype information was needed. To validate
617 the accuracy of prediction, we compared it with the prediction using a different combination of
618 SNPs⁷⁷. The two predictions were highly consistent, with over 99.9% concordance. In addition,
619 the distribution of ABO groups was consistent with the population distribution found in public
620 database. Secretor status was predicted by the genotype of *FUT2* variant rs601338, where AA
621 or AG genotypes are secretors and GG genotypes are non-secretors. An 100% concordance
622 between the variation in rs601338 and secretor status was reported in a study on Finnish
623 individuals¹¹⁸.

624

625 **Bidirectional two-sample Mendelian randomization (MR) analysis**

626

627 Causal relationships between diseases and gut microbiota were investigated at genus and
628 species levels only to maximise interpretability. In total, 213 species and 148 genera associated
629 with at least one variant at genome-wide significant level ($p < 1 \times 10^{-8}$) were included. GWAS
630 summary results were collected for 46 diseases from MR-Base¹¹⁹ (**Table S4**). These included
631 12 autoimmune or inflammatory diseases, 9 cardiometabolic diseases, 13 psychiatric or
632 neurological diseases, cardiovascular diseases, 4 bone diseases and 8 cancers. For disease with
633 more than one GWAS records, the record with the largest sample size was kept.

634

635 Bi-directional causal inference was performed as follows to infer causal effects of microbial
636 abundance variation (exposure) on disease risk (outcome), and of disease (exposure) on
637 microbial abundance levels (outcome). To select the SNP instruments for microbial exposures
638 in our study, we followed recommendations from a previous study showing that associated
639 SNPs below a significance threshold of $p < 1 \times 10^{-5}$ had the largest explained variance on
640 microbial features¹²⁰. For each taxon, GCTA-COJO was used to perform a conditional analysis
641 to select independently associated SNPs at $p < 1 \times 10^{-5}$. SNP instruments for disease exposures
642 were selected at genome-wide significant threshold ($p < 5 \times 10^{-8}$). Subsequently LD-clumping
643 with a strict threshold ($r^2 < 0.001$ in 1000G EUR within 10 Mb windows) was conducted to
644 select independent instruments with the lowest p values for taxa and diseases, respectively.

645

646 Effective alleles of all genetic variants were oriented to the risk-increasing alleles of exposures.
647 For each inference, five different MR methods were used to estimate the causal effect: (1)
648 inverse variance weighted (IVW)¹²¹, (2) weighted median¹²², (3) simple mode¹²³, (4) weighted
649 mode¹²³ and (5) MR-Egger¹²⁴. IVW is the most sensitive method which requires all
650 instruments are valid. But in reality, it is hard to verify that no any genetic instrument violates
651 any instrumental assumptions. Weighted median only requires at least half of the instruments
652 are valid, making its inference robust to the cases where some instruments violating the
653 assumptions. Simple mode and weighted mode rely on the largest group of similar instruments,
654 reducing the effects of other instruments especially outliers. MR-Egger allows instruments
655 having non-zero pleiotropy and provides way to test and estimate the pleiotropy effect in
656 addition to causal estimate. As these methods are based on different assumptions, the
657 consistency among them indicates a credible estimate¹²⁵, even if discrepancy in these methods
658 does not necessarily suggest the absence of causality. A predicted causal estimate was deemed
659 interesting in our study if: (1) it reached a nominal $p < 0.05$ for at least three of the five tested
660 methods (**Table S7**), (2) directionality testing supported the causal direction, and (3) no
661 significant casual effect in the reverse direction. In addition, MR-PRESSO¹²⁶ was used to

662 formally detect and correct for the pleiotropic outliers. Analyses were conducted using the R
663 package *TwoSampleMR*¹¹⁹.

664

665 **Cox proportional hazards regression**

666

667 Cox proportional hazards regression was conducted to test the association between baseline
668 abundance of gut microbe and incident major depression (16 years follow-up, n=181 incident
669 events). Microbial abundances were CLR-transformed and standardized to zero-mean and unit-
670 variance. The Cox models were stratified by sex and adjusted for age and log-transformed
671 BMI, with time-on-study as the time scale. Participants with prevalent major depression at
672 baseline were excluded. R function *coxph()* in the R package *survival* was used for this
673 analysis.

674

675 **Profiling of carbohydrate-active enzymes (CAZymes) in bacterial genomes**

676

677 The standalone run_dbCAN2 v2.0.11 tool¹²⁷ (https://github.com/linnabrown/run_dbcan) was
678 used to scan for the presence of CAZyme genes from public assembled bacterial genomes
679 taken from the GTDB release 89 reference. We used a CAZyme reference database taken from
680 the CAZy database¹²⁸ (31st July 2019 update). In total, we scanned 327 *Bifidobacterium sp.*, 2
681 *Faecalicatena lactaris* and 15 *Collinsella sp.* reference genomes included in GTDB release 89.
682 Three methods were compared as part of the run_dbCAN2 procedure (HMMER, DIAMOND,
683 and Hotpep). We considered a positive detection result when all three methods agreed on a
684 CAZyme family identification. Identification of preferred reported substrates for the various
685 CAZyme families was done manually from key publications^{48,129}, from literature searches and
686 from the CAZypedia website¹³⁰. Certain CAZyme families have a broad range of substrates,
687 many of which are still unknown, which results in our reported preferred substrates to be as
688 accurate as possible, but non-exhaustive.

689

690 **Carbon impact and offsetting**

691

692 We used GreenAlgorithms v1.0¹³¹ to estimate that the main computational work in this study
693 had a carbon impact of at least 531.94 kg CO₂e, corresponding to 560 tree-months. As a
694 commitment to the reduction of carbon emissions associated with computation in research, we
695 consequently funded planting of 30 trees through a local Australian charity, which across their
696 lifetime will sequester a combined estimated 8,040 kg CO₂e, or 15 times the amount of CO₂e
697 generated by this study.

698

699

700 **Acknowledgements**

701

702 We thank all participants of the FINRISK 2002 survey for their contributions to this work. The
703 FINRISK surveys are mainly funded by budgetary funds from the Finnish Institute for Health
704 and Welfare with additional funding from several domestic foundations. MI was supported by
705 the Munz Chair of Cardiovascular Prediction and Prevention. VS was supported by the Finnish
706 Foundation for Cardiovascular Research. LL was supported by Academy of Finland (decision
707 295741). ASH was supported by the Academy of Finland, grant no. 321356. RL receives
708 funding support from NIEHS (5P42ES010337), NCATS (5UL1TR001442), NIDDK
709 (U01DK061734, R01DK106419, P30DK120515, R01DK121378, R01DK124318), and DOD
710 PRCRP (W81XWH-18-2-0026). This study was supported by the Victorian Government's
711 Operational Infrastructure Support (OIS) program, and by core funding from: the UK Medical
712 Research Council (MR/L003120/1), the British Heart Foundation (RG/13/13/30194;
713 RG/18/13/33946) and the National Institute for Health Research [Cambridge Biomedical
714 Research Centre at the Cambridge University Hospitals NHS Foundation Trust] [*]. This work
715 was supported by Health Data Research UK, which is funded by the UK Medical Research
716 Council, Engineering and Physical Sciences Research Council, Economic and Social Research
717 Council, Department of Health and Social Care (England), Chief Scientist Office of the
718 Scottish Government Health and Social Care Directorates, Health and Social Care Research
719 and Development Division (Welsh Government), Public Health Agency (Northern Ireland),
720 British Heart Foundation and Wellcome. *The views expressed are those of the authors and not
721 necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

722

723 **Author declaration**

724

725 The study protocol of FINRISK 2002 was approved by the Coordinating Ethical Committee of
726 the Helsinki and Uusimaa Hospital District (Ref. 558/E3/2001). All participants signed an
727 informed consent. The study was conducted according to the World Medical Association
728 Declaration of Helsinki on ethical principles. All necessary patient/participant consent has been
729 obtained and the appropriate institutional forms have been archived.

730

731 **Data Availability**

732

733 The data for the present study are available with a written application to the THL Biobank as
734 instructed in the website of the Biobank: <https://thl.fi/en/web/thl-biobank/for-researchers>.

735

736 **Conflicts of interest**

737

738 VS has consulted for Novo Nordisk and Sanofi and received honoraria from these companies.
739 He also has ongoing research collaboration with Bayer AG, all unrelated to this study. RL
740 serves as a consultant or advisory board member for Anylam/Regeneron, Arrowhead
741 Pharmaceuticals, AstraZeneca, Bird Rock Bio, Boehringer Ingelheim, Bristol-Myer Squibb,
742 Celgene, Cirus, CohBar, Conatus, Eli Lilly, Galmed, Gemphire, Gilead, Glympse bio, GNI,
743 GRI Bio, Inpharm, Intercept, Ionis, Janssen Inc., Merck, Metacrine, Inc., NGM
744 Biopharmaceuticals, Novartis, Novo Nordisk, Pfizer, Prometheus, Promethera, Sanofi,
745 Siemens and Viking Therapeutics. In addition, his institution has received grant support from
746 Allergan, Boehringer-Ingelheim, Bristol-Myers Squibb, Cirus, Eli Lilly and Company,
747 Galectin Therapeutics, Galmed Pharmaceuticals, GE, Genfit, Gilead, Intercept, Grail, Janssen,
748 Madrigal Pharmaceuticals, Merck, NGM Biopharmaceuticals, NuSirt, Pfizer, pH Pharma,
749 Prometheus, and Siemens. He is also co-founder of Liponex, Inc.

750

751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811

References

1. Khosravi, A. & Mazmanian, S. K. Disruption of the gut microbiome as a risk factor for microbial infections. *Current Opinion in Microbiology* **16**, 221–227 (2013).
2. Belizário, J. E. & Napolitano, M. Human microbiomes and their roles in dysbiosis, common diseases, and novel therapeutic approaches. *Front. Microbiol.* **6**, (2015).
3. Levy, M., Kolodziejczyk, A. A., Thaiss, C. A. & Elinav, E. Dysbiosis and the immune system. *Nat Rev Immunol* **17**, 219–232 (2017).
4. Blekhman, R. *et al.* Host genetic variation impacts microbiome composition across human body sites. *Genome Biol* **16**, 191 (2015).
5. Davenport, E. R. *et al.* ABO antigen and secretor statuses are not associated with gut microbiota composition in 1,500 twins. *BMC Genomics* **17**, 941 (2016).
6. Goodrich, J. K. *et al.* Genetic Determinants of the Gut Microbiome in UK Twins. *Cell Host & Microbe* **19**, 731–743 (2016).
7. Bonder, M. J. *et al.* The effect of host genetics on the gut microbiome. *Nat Genet* **48**, 1407–1412 (2016).
8. Turpin, W. *et al.* Association of host genome with intestinal microbial composition in a large healthy cohort. *Nat Genet* **48**, 1413–1417 (2016).
9. 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).
10. Rothschild, D. *et al.* Environment dominates over host genetics in shaping human gut microbiota. *Nature* **555**, 210–215 (2018).
11. 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).
12. Kurilshikov, A. *et al.* Genetics of human gut microbiome composition. <http://biorxiv.org/lookup/doi/10.1101/2020.06.26.173724> (2020) doi:10.1101/2020.06.26.173724.
13. Kolde, R. *et al.* Host genetic variation and its microbiome interactions within the Human Microbiome Project. *Genome Med* **10**, 6 (2018).
14. Rühlemann, M. C. *et al.* Application of the distance-based F test in an mGWAS investigating β diversity of intestinal microbiota identifies variants in *SLC9A8* (NHE8) and 3 other loci. *Gut Microbes* **9**, 68–75 (2018).
15. Goodrich, J. K. *et al.* Human Genetics Shape the Gut Microbiome. *Cell* **159**, 789–799 (2014).
16. Xie, H. *et al.* Shotgun Metagenomics of 250 Adult Twins Reveals Genetic and Environmental Impacts on the Gut Microbiome. *Cell Systems* **3**, 572–584.e3 (2016).
17. Lim, M. Y. *et al.* The effect of heritability and host genetics on the gut microbiota and metabolic syndrome. *Gut* **66**, 1031–1038 (2017).
18. Le Roy, C. I. *et al.* Heritable components of the human fecal microbiome are associated with visceral fat. *Gut Microbes* **9**, 61–67 (2018).
19. Goodrich, J. K., Davenport, E. R., Clark, A. G. & Ley, R. E. The Relationship Between the Human Genome and Microbiome Comes into View. *Annu. Rev. Genet.* **51**, 413–433 (2017).
20. Kurilshikov, A., Wijmenga, C., Fu, J. & Zhernakova, A. Host Genetics and Gut Microbiome: Challenges and Perspectives. *Trends in Immunology* **38**, 633–647 (2017).
21. David, L. A. *et al.* Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505**, 559–563 (2014).
22. Falony, G. *et al.* Population-level analysis of gut microbiome variation. *Science* **352**, 560–564 (2016).
23. Zhernakova, A. *et al.* Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* **352**, 565–569 (2016).
24. Eng, A. & Borenstein, E. Taxa-function robustness in microbial communities. *Microbiome* **6**, 45 (2018).
25. Ferrer, M. *et al.* Microbiota from the distal guts of lean and obese adolescents exhibit partial functional redundancy besides clear differences in community structure: Metaproteomic insights associated to human obesity. *Environ Microbiol* **15**, 211–226 (2013).
26. Moya, A. & Ferrer, M. Functional Redundancy-Induced Stability of Gut Microbiota Subjected to Disturbance. *Trends in Microbiology* **24**, 402–413 (2016).
27. Louca, S. *et al.* Function and functional redundancy in microbial systems. *Nat Ecol Evol* **2**, 936–943 (2018).
28. Louca, S. *et al.* High taxonomic variability despite stable functional structure across microbial communities. *Nat Ecol Evol* **1**, 0015 (2017).
29. Banerjee, S., Schlaeppli, K. & van der Heijden, M. G. A. Keystone taxa as drivers of microbiome structure and functioning. *Nat Rev Microbiol* **16**, 567–576 (2018).
30. Trosvik, P. & de Muinck, E. J. Ecology of bacteria in the human gastrointestinal tract—identification of keystone and foundation taxa. *Microbiome* **3**, 44 (2015).
31. Shetty, S. A., Hugenholtz, F., Lahti, L., Smidt, H. & de Vos, W. M. Intestinal microbiome landscaping: insight in community assemblage and implications for microbial modulation strategies. *FEMS Microbiol. Rev.* **41**, 182–199 (2017).

- 812 32. Chia, L. W. *et al.* Deciphering the trophic interaction between Akkermansia muciniphila and the
813 butyrogenic gut commensal Anaerostipes caccae using a metatranscriptomic approach. *Antonie van*
814 *Leeuwenhoek* **111**, 859–873 (2018).
- 815 33. Fisher, C. K. & Mehta, P. Identifying Keystone Species in the Human Gut Microbiome from Metagenomic
816 Timeseries Using Sparse Linear Regression. *PLoS ONE* **9**, e102451 (2014).
- 817 34. Curtis, M. M. *et al.* The Gut Commensal Bacteroides thetaiotaomicron Exacerbates Enteric Infection
818 through Modification of the Metabolic Landscape. *Cell Host & Microbe* **16**, 759–769 (2014).
- 819 35. Ze, X., Duncan, S. H., Louis, P. & Flint, H. J. Ruminococcus bromii is a keystone species for the
820 degradation of resistant starch in the human colon. *ISME J* **6**, 1535–1543 (2012).
- 821 36. Garrett, W. S. *et al.* Enterobacteriaceae Act in Concert with the Gut Microbiota to Induce Spontaneous and
822 Maternally Transmitted Colitis. *Cell Host & Microbe* **8**, 292–300 (2010).
- 823 37. Hajishengallis, G., Darveau, R. P. & Curtis, M. A. The keystone-pathogen hypothesis. *Nat Rev Microbiol*
824 **10**, 717–725 (2012).
- 825 38. Ley, R. E., Peterson, D. A. & Gordon, J. I. Ecological and Evolutionary Forces Shaping Microbial Diversity
826 in the Human Intestine. *Cell* **124**, 837–848 (2006).
- 827 39. Wu, S. *et al.* A human colonic commensal promotes colon tumorigenesis via activation of T helper type 17
828 T cell responses. *Nat Med* **15**, 1016–1022 (2009).
- 829 40. Kato, K. *et al.* Age-Related Changes in the Composition of Gut Bifidobacterium Species. *Curr Microbiol*
830 **74**, 987–995 (2017).
- 831 41. Engevik, M. A. *et al.* Bifidobacterium dentium Fortifies the Intestinal Mucus Layer via Autophagy and
832 Calcium Signaling Pathways. *mBio* **10**, e01087-19, /mbio/10/3/mBio.01087-19.atom (2019).
- 833 42. Rahfeld, P. & Withers, S. G. Toward universal donor blood: Enzymatic conversion of A and B to O type. *J.*
834 *Biol. Chem.* **295**, 325–334 (2020).
- 835 43. Liu, Q. P. *et al.* Bacterial glycosidases for the production of universal red blood cells. *Nat Biotechnol* **25**,
836 454–464 (2007).
- 837 44. Arnolds, K. L., Martin, C. G. & Lozupone, C. A. Blood type and the microbiome- untangling a complex
838 relationship with lessons from pathogens. *Current Opinion in Microbiology* **56**, 59–66 (2020).
- 839 45. Liu, Q. P. *et al.* Identification of a GH110 Subfamily of α 1,3-Galactosidases: NOVEL ENZYMES FOR
840 REMOVAL OF THE α 3GAL XENOTRANSPLANTATION ANTIGEN. *J. Biol. Chem.* **283**, 8545–8554
841 (2008).
- 842 46. Pichler, M. J. *et al.* Butyrate producing colonic Clostridiales metabolise human milk oligosaccharides and
843 cross feed on mucin via conserved pathways. *Nat Commun* **11**, 3285 (2020).
- 844 47. Ficko-Blean, E. & Boraston, A. B. The Interaction of a Carbohydrate-binding Module from a *Clostridium*
845 *perfringens* N -Acetyl- β -hexosaminidase with Its Carbohydrate Receptor. *J. Biol. Chem.* **281**, 37748–37757
846 (2006).
- 847 48. Desai, M. S. *et al.* A Dietary Fiber-Deprived Gut Microbiota Degrades the Colonic Mucus Barrier and
848 Enhances Pathogen Susceptibility. *Cell* **167**, 1339-1353.e21 (2016).
- 849 49. Tailford, L. E., Crost, E. H., Kavanaugh, D. & Juge, N. Mucin glycan foraging in the human gut
850 microbiome. *Front. Genet.* **6**, (2015).
- 851 50. Genome Aggregation Database Consortium *et al.* The mutational constraint spectrum quantified from
852 variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
- 853 51. Jahani-Sherafat, S., Alebouyeh, M., Moghim, S., Ahmadi Amoli, H. & Ghasemian-Safaei, H. Role of gut
854 microbiota in the pathogenesis of colorectal cancer; a review article. *Gastroenterol Hepatol Bed Bench* **11**,
855 101–109 (2018).
- 856 52. Amamani, R. & Rapose, A. Colon cancer and enterococcus bacteremia co-affection: A dangerous alliance.
857 *Journal of Infection and Public Health* **10**, 681–684 (2017).
- 858 53. Pillar, C., M. Enterococcal virulence - pathogenicity island of E. Faecalis. *Front Biosci* **9**, 2335 (2004).
- 859 54. Khan, Z., Siddiqui, N. & Saif, M. W. *Enterococcus Faecalis* Infective Endocarditis and Colorectal
860 Carcinoma: Case of New Association Gaining Ground. *Gastroenterol Res* **11**, 238–240 (2018).
- 861 55. De Almeida, C. *et al.* Differential Responses of Colorectal Cancer Cell Lines to Enterococcus faecalis'
862 Strains Isolated from Healthy Donors and Colorectal Cancer Patients. *JCM* **8**, 388 (2019).
- 863 56. Huycke, M. M., Abrams, V. & Moore, D. R. Enterococcus faecalis produces extracellular superoxide and
864 hydrogen peroxide that damages colonic epithelial cell DNA. *Carcinogenesis* **23**, 529–536 (2002).
- 865 57. Allen, B. L. & Taatjes, D. J. The Mediator complex: a central integrator of transcription. *Nat Rev Mol Cell*
866 *Biol* **16**, 155–166 (2015).
- 867 58. Firestein, R. *et al.* CDK8 is a colorectal cancer oncogene that regulates β -catenin activity. *Nature* **455**, 547–
868 551 (2008).
- 869 59. Li, L., Batt, S. M., Wannemuehler, M., Dispirito, A. & Beitz, D. C. Effect of feeding of a cholesterol-
870 reducing bacterium, Eubacterium coprostanoligenes, to germ-free mice. *Lab. Anim. Sci.* **48**, 253–255
871 (1998).
- 872 60. Marasco, G. *et al.* Gut Microbiota and Celiac Disease. *Dig Dis Sci* **61**, 1461–1472 (2016).

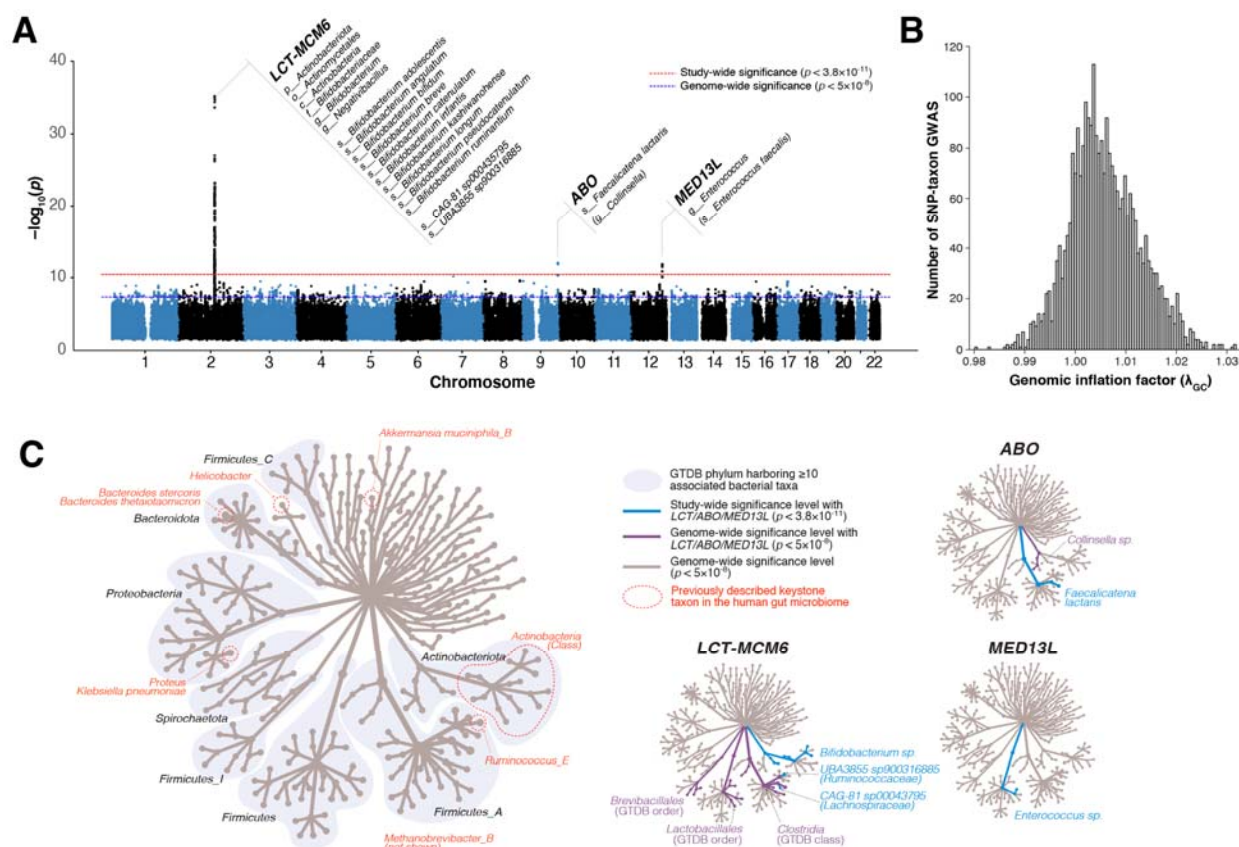
- 873 61. Lavasani, S. *et al.* A Novel Probiotic Mixture Exerts a Therapeutic Effect on Experimental Autoimmune
874 Encephalomyelitis Mediated by IL-10 Producing Regulatory T Cells. *PLoS ONE* **5**, e9009 (2010).
- 875 62. Tomita, H. *et al.* G protein-linked signaling pathways in bipolar and major depressive disorders. *Front.*
876 *Genet.* **4**, (2013).
- 877 63. Wong, M.-L. *et al.* Phosphodiesterase genes are associated with susceptibility to major depression and
878 antidepressant treatment response. *Proceedings of the National Academy of Sciences* **103**, 15124–15129
879 (2006).
- 880 64. Schork, A. J. *et al.* A genome-wide association study of shared risk across psychiatric disorders implicates
881 gene regulation during fetal neurodevelopment. *Nat Neurosci* **22**, 353–361 (2019).
- 882 65. Burger, J. *et al.* Low Prevalence of Lactase Persistence in Bronze Age Europe Indicates Ongoing Strong
883 Selection over the Last 3,000 Years. *Current Biology* S0960982220311878 (2020)
884 doi:10.1016/j.cub.2020.08.033.
- 885 66. Gerbault, P. *et al.* Evolution of lactase persistence: an example of human niche construction. *Phil. Trans. R.*
886 *Soc. B* **366**, 863–877 (2011).
- 887 67. Hebert, J. R. *et al.* Social Desirability Trait Influences on Self-Reported Dietary Measures among Diverse
888 Participants in a Multicenter Multiple Risk Factor Trial. *The Journal of Nutrition* **138**, 226S–234S (2008).
- 889 68. Schoeller, D. A. How Accurate Is Self-Reported Dietary Energy Intake? *Nutrition Reviews* **48**, 373–379
890 (2009).
- 891 69. Sakanaka, M. *et al.* Evolutionary adaptation in fucosyllactose uptake systems supports bifidobacteria-infant
892 symbiosis. *Sci. Adv.* **5**, eaaw7696 (2019).
- 893 70. Storhaug, C. L., Fosse, S. K. & Fadnes, L. T. Country, regional, and global estimates for lactose
894 malabsorption in adults: a systematic review and meta-analysis. *The Lancet Gastroenterology &*
895 *Hepatology* **2**, 738–746 (2017).
- 896 71. Liu, X. *et al.* M-GWAS for the gut microbiome in Chinese adults illuminates on complex diseases.
897 <http://biorxiv.org/lookup/doi/10.1101/736413> (2019) doi:10.1101/736413.
- 898 72. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The Missing Diversity in Human Genetic Studies. *Cell* **177**,
899 26–31 (2019).
- 900 73. Szilagy, A. Adaptation to Lactose in Lactase Non Persistent People: Effects on Intolerance and the
901 Relationship between Dairy Food Consumption and Evaluation of Diseases. *Nutrients* **7**, 6751–6779 (2015).
- 902 74. Ségurel, L., Gao, Z. & Przeworski, M. Ancestry runs deeper than blood: The evolutionary history of ABO
903 points to cryptic variation of functional importance: Insights & Perspective. *BioEssays* n/a-n/a (2013)
904 doi:10.1002/bies.201300030.
- 905 75. Segurel, L. *et al.* The ABO blood group is a trans-species polymorphism in primates. *Proceedings of the*
906 *National Academy of Sciences* **109**, 18493–18498 (2012).
- 907 76. Ewald, D. R. & Sumner, S. C. J. Blood type biochemistry and human disease: Blood type biochemistry and
908 human disease. *WIREs Syst Biol Med* **8**, 517–535 (2016).
- 909 77. Ellinghaus, D. *et al.* Genomewide Association Study of Severe Covid-19 with Respiratory Failure. *N Engl J*
910 *Med* NEJMoa2020283 (2020) doi:10.1056/NEJMoa2020283.
- 911 78. Shelton, J. F. *et al.* Trans-ethnic analysis reveals genetic and non-genetic associations with COVID-19
912 susceptibility and severity. <http://medrxiv.org/lookup/doi/10.1101/2020.09.04.20188318> (2020)
913 doi:10.1101/2020.09.04.20188318.
- 914 79. Rühlemann, M. C. *et al.* ABO histo-blood groups influence gut microbiome, with causal relationship
915 between *Bacteroides* and inflammatory bowel disease.
916 <http://medrxiv.org/lookup/doi/10.1101/2020.07.09.20148627> (2020) doi:10.1101/2020.07.09.20148627.
- 917 80. Liu, X. *et al.* Inter-determination of blood metabolite levels and gut microbiome supported by Mendelian
918 randomization. <http://biorxiv.org/lookup/doi/10.1101/2020.06.30.181438> (2020)
919 doi:10.1101/2020.06.30.181438.
- 920 81. Knuesel, M. T., Meyer, K. D., Bernecky, C. & Taatjes, D. J. The human CDK8 subcomplex is a molecular
921 switch that controls Mediator coactivator function. *Genes & Development* **23**, 439–451 (2009).
- 922 82. Tsai, K.-L. *et al.* A conserved Mediator–CDK8 kinase module association regulates Mediator–RNA
923 polymerase II interaction. *Nat Struct Mol Biol* **20**, 611–619 (2013).
- 924 83. Jiang, H. *et al.* Altered fecal microbiota composition in patients with major depressive disorder. *Brain,*
925 *Behavior, and Immunity* **48**, 186–194 (2015).
- 926 84. Wade, K. H. & Hall, L. J. Improving causality in microbiome research: can human genetic epidemiology
927 help? *Wellcome Open Res* **4**, 199 (2020).
- 928 85. Foster, J. A. & McVey Neufeld, K.-A. Gut–brain axis: how the microbiome influences anxiety and
929 depression. *Trends in Neurosciences* **36**, 305–312 (2013).
- 930 86. Fung, T. C., Olson, C. A. & Hsiao, E. Y. Interactions between the microbiota, immune and nervous systems
931 in health and disease. *Nat Neurosci* **20**, 145–155 (2017).
- 932 87. Valles-Colomer, M. *et al.* The neuroactive potential of the human gut microbiota in quality of life and
933 depression. *Nat Microbiol* **4**, 623–632 (2019).

- 934 88. Maes, M., Kubera, M. & Leunis, J.-C. The gut-brain barrier in major depression: intestinal mucosal
935 dysfunction with an increased translocation of LPS from gram negative enterobacteria (leaky gut) plays a
936 role in the inflammatory pathophysiology of depression. *Neuro Endocrinol. Lett.* **29**, 117–124 (2008).
937 89. Marchesi, J. R. *et al.* The gut microbiota and host health: a new clinical frontier. *Gut* **65**, 330–339 (2016).
938 90. Ruff, W. E., Greiling, T. M. & Kriegel, M. A. Host–microbiota interactions in immune-mediated diseases.
939 *Nat Rev Microbiol* **18**, 521–538 (2020).
940 91. Mattar, R., Mazo & Carrilho. Lactose intolerance: diagnosis, genetic, and clinical factors. *CEG* 113 (2012)
941 doi:10.2147/CEG.S32368.
942 92. Bodmer, W. Genetic Characterization of Human Populations: From ABO to a Genetic Map of the British
943 People. *Genetics* **199**, 267–279 (2015).
944 93. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the
945 tree of life. *Nat Biotechnol* **36**, 996–1004 (2018).
946 94. Parks, D. H. *et al.* A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat Biotechnol*
947 (2020) doi:10.1038/s41587-020-0501-8.
948 95. Méric, G., Wick, R. R., Watts, S. C., Holt, K. E. & Inouye, M. *Correcting index databases improves*
949 *metagenomic studies*. <http://biorxiv.org/lookup/doi/10.1101/712166> (2019) doi:10.1101/712166.
950 96. Pasolli, E. *et al.* Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes
951 from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176**, 649–662.e20 (2019).
952 97. Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat*
953 *Biotechnol* (2020) doi:10.1038/s41587-020-0603-3.
954 98. Borodulin, K. *et al.* Cohort Profile: The National FINRISK Study. *International Journal of Epidemiology*
955 **47**, 696–696i (2018).
956 99. Borodulin, K. *et al.* Forty-year trends in cardiovascular risk factors in Finland. *The European Journal of*
957 *Public Health* **25**, 539–546 (2015).
958 100. Liu, Y. *et al.* *Early prediction of liver disease using conventional risk factors and gut microbiome-*
959 *augmented gradient boosting*. <http://medrxiv.org/lookup/doi/10.1101/2020.06.24.20138933> (2020)
960 doi:10.1101/2020.06.24.20138933.
961 101. Salosensaari, A. *et al.* *Taxonomic Signatures of Long-Term Mortality Risk in Human Gut Microbiota*.
962 <http://medrxiv.org/lookup/doi/10.1101/2019.12.30.19015842> (2020) doi:10.1101/2019.12.30.19015842.
963 102. FinnGen *et al.* Polygenic and clinical risk scores and their impact on age at onset and prediction of
964 cardiometabolic diseases and common cancers. *Nat Med* **26**, 549–557 (2020).
965 103. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet* **48**,
966 1443–1448 (2016).
967 104. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype
968 imputation in genome-wide association studies through pre-phasing. *Nat Genet* **44**, 955–959 (2012).
969 105. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaSci*
970 **4**, 7 (2015).
971 106. Ruuskanen, M. O. *et al.* *Links between gut microbiome composition and fatty liver disease in a large*
972 *population sample*. <http://medrxiv.org/lookup/doi/10.1101/2020.07.30.20164962> (2020)
973 doi:10.1101/2020.07.30.20164962.
974 107. Kim, D., Song, L., Breitwieser, F. P. & Salzberg, S. L. Centrifuge: rapid and sensitive classification of
975 metagenomic sequences. *Genome Res.* **26**, 1721–1729 (2016).
976 108. Goodrich, J. K., Davenport, E. R., Waters, J. L., Clark, A. G. & Ley, R. E. Cross-species comparisons of
977 host genetic associations with the microbiome. *Science* **352**, 532–535 (2016).
978 109. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome Datasets Are
979 Compositional: And This Is Not Optional. *Front. Microbiol.* **8**, 2224 (2017).
980 110. Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J. A. & Pawlowsky-Glahn, V. [No title found].
981 *Mathematical Geology* **32**, 271–275 (2000).
982 111. Qin, Y. *et al.* *Genome-wide association and Mendelian randomization analysis prioritizes bioactive*
983 *metabolites with putative causal effects on common diseases*.
984 <http://medrxiv.org/lookup/doi/10.1101/2020.08.01.20166413> (2020) doi:10.1101/2020.08.01.20166413.
985 112. Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat*
986 *Genet* **47**, 284–290 (2015).
987 113. Abraham, G., Qiu, Y. & Inouye, M. FlashPCA2: principal component analysis of Biobank-scale genotype
988 datasets. *Bioinformatics* **33**, 2776–2778 (2017).
989 114. Genetic Investigation of ANthropometric Traits (GIANT) Consortium *et al.* Conditional and joint multiple-
990 SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat*
991 *Genet* **44**, 369–375 (2012).
992 115. Li, J. & Ji, L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation
993 matrix. *Heredity* **95**, 221–227 (2005).

- 994 116. Nyholt, D. R. A Simple Correction for Multiple Testing for Single-Nucleotide Polymorphisms in Linkage
995 Disequilibrium with Each Other. *The American Journal of Human Genetics* **74**, 765–769 (2004).
996 117. Paré, G. *et al.* Novel Association of ABO Histo-Blood Group Antigen with Soluble ICAM-1: Results of a
997 Genome-Wide Association Study of 6,578 Women. *PLoS Genet* **4**, e1000118 (2008).
998 118. Wacklin, P. *et al.* Secretor Genotype (FUT2 gene) Is Strongly Associated with the Composition of
999 Bifidobacteria in the Human Intestine. *PLoS ONE* **6**, e20113 (2011).
1000 119. Hemani, G. *et al.* The MR-Base platform supports systematic causal inference across the human phenome.
1001 *eLife* **7**, e34408 (2018).
1002 120. Sanna, S. *et al.* Causal relationships among the gut microbiome, short-chain fatty acids and metabolic
1003 diseases. *Nat Genet* **51**, 600–605 (2019).
1004 121. Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian Randomization Analysis With Multiple Genetic
1005 Variants Using Summarized Data: Mendelian Randomization Using Summarized Data. *Genet. Epidemiol.*
1006 **37**, 658–665 (2013).
1007 122. Bowden, J., Davey Smith, G., Haycock, P. C. & Burgess, S. Consistent Estimation in Mendelian
1008 Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet. Epidemiol.* **40**,
1009 304–314 (2016).
1010 123. Hartwig, F. P., Davey Smith, G. & Bowden, J. Robust inference in summary data Mendelian randomization
1011 via the zero modal pleiotropy assumption. *International Journal of Epidemiology* **46**, 1985–1998 (2017).
1012 124. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect
1013 estimation and bias detection through Egger regression. *International Journal of Epidemiology* **44**, 512–525
1014 (2015).
1015 125. Burgess, S. *et al.* Guidelines for performing Mendelian randomization investigations. *Wellcome Open Res*
1016 **4**, 186 (2020).
1017 126. Verbanck, M., Chen, C.-Y., Neale, B. & Do, R. Detection of widespread horizontal pleiotropy in causal
1018 relationships inferred from Mendelian randomization between complex traits and diseases. *Nat Genet* **50**,
1019 693–698 (2018).
1020 127. Zhang, H. *et al.* dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic*
1021 *Acids Research* **46**, W95–W101 (2018).
1022 128. Cantarel, B. L. *et al.* The Carbohydrate-Active EnZymes database (CAZy): an expert resource for
1023 Glycogenomics. *Nucleic Acids Research* **37**, D233–D238 (2009).
1024 129. Cantarel, B. L., Lombard, V. & Henrissat, B. Complex Carbohydrate Utilization by the Healthy Human
1025 Microbiome. *PLoS ONE* **7**, e28742 (2012).
1026 130. The CAZypedia Consortium. Ten years of CAZypedia: a living encyclopedia of carbohydrate-active
1027 enzymes. *Glycobiology* **28**, 3–8 (2018).
1028 131. Lannelongue, L., Grealey, J. & Inouye, M. Green Algorithms: Quantifying the carbon emissions of
1029 computation. *arXiv:2007.07610 [cs]* (2020).
1030

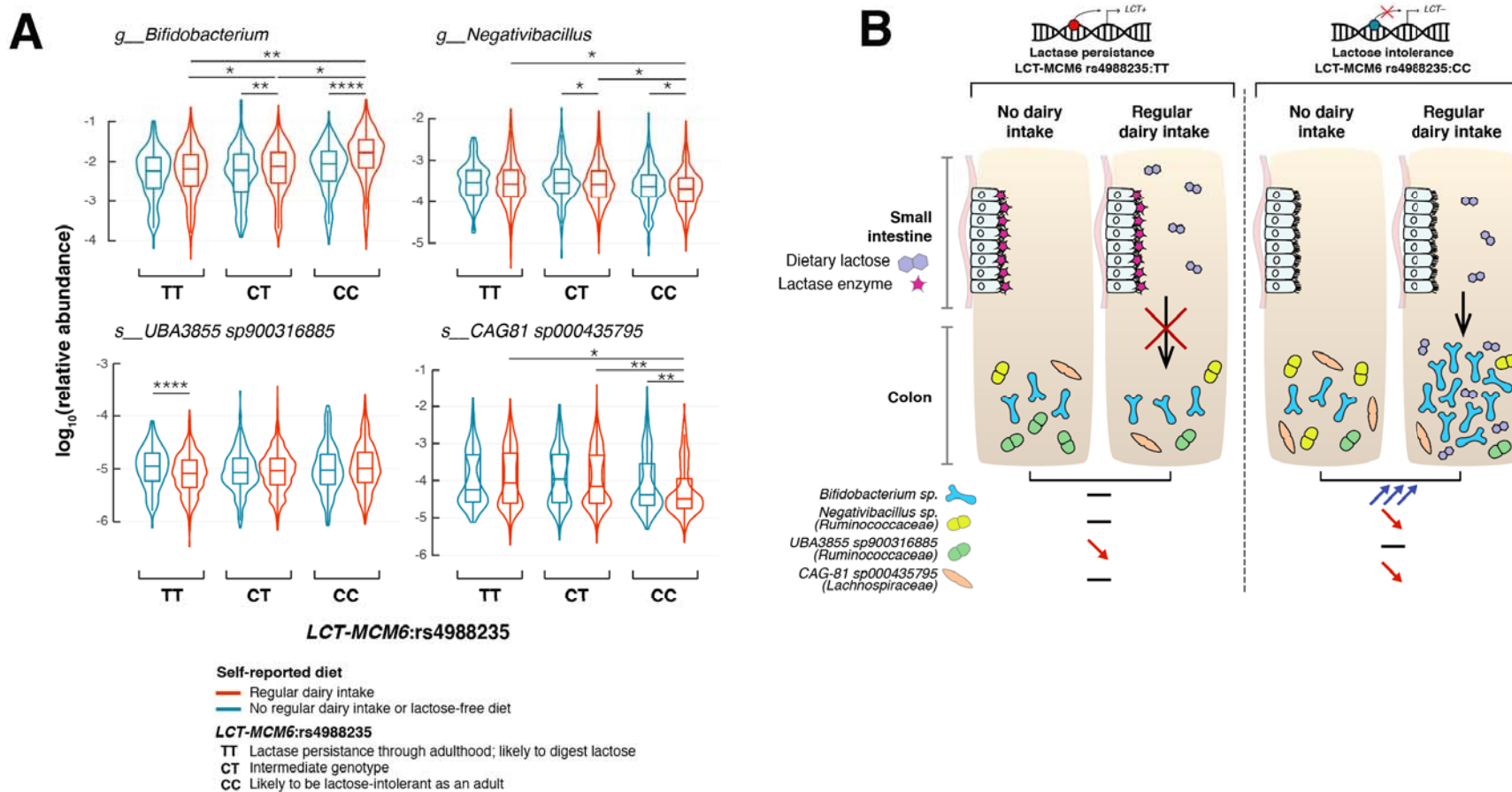
1031 **Main figure and tables**

1032
 1033 **Figure 1. Genome-wide association of human genetic and gut microbial variations.** (A) Manhattan plot aggregating the top associations with microbial variation. Each
 1034 SNP was tested against each of the 2,801 taxa and the Manhattan plot shows the lowest resulting p-value for each SNP. Loci with associations above study-wide significance
 1035 level ($p < 3.8 \times 10^{-11}$; red dashed line) are annotated with the human locus name and the corresponding associated microbial taxa. The blue dashed line denotes genome-wide
 1036 significance level ($p < 5 \times 10^{-8}$). (B) The distribution of genomic inflation factor (λ_{GC}) in 2,801 tested taxa [median(λ_{GC})=1.0051; mean(λ_{GC})=1.0059]. (C) Tree-based
 1037 visualization of the taxonomic diversity of genome-wide associated microbial taxa. The central root of the tree represents the Bacteria domain, the first connected node
 1038 represents phylum, the second connected node class, the third order and the fourth family. Every node represents at least one associated taxa in the GWAS at genome-wide
 1039 significance level. The three smaller trees on the right highlight all taxonomic groups containing at least one taxon identified as associated with the *LCT-MCM6*, *ABO*, and
 1040 *MED13L* loci (blue edges and nodes denote taxa associated at study-wide significance level and purple edges and nodes denote taxa associated at genome-wide significance
 1041 level). The main tree is annotated to indicate phyla harbouring >10 distinct genome-wide associated taxa, as well as previously described keystone taxa.



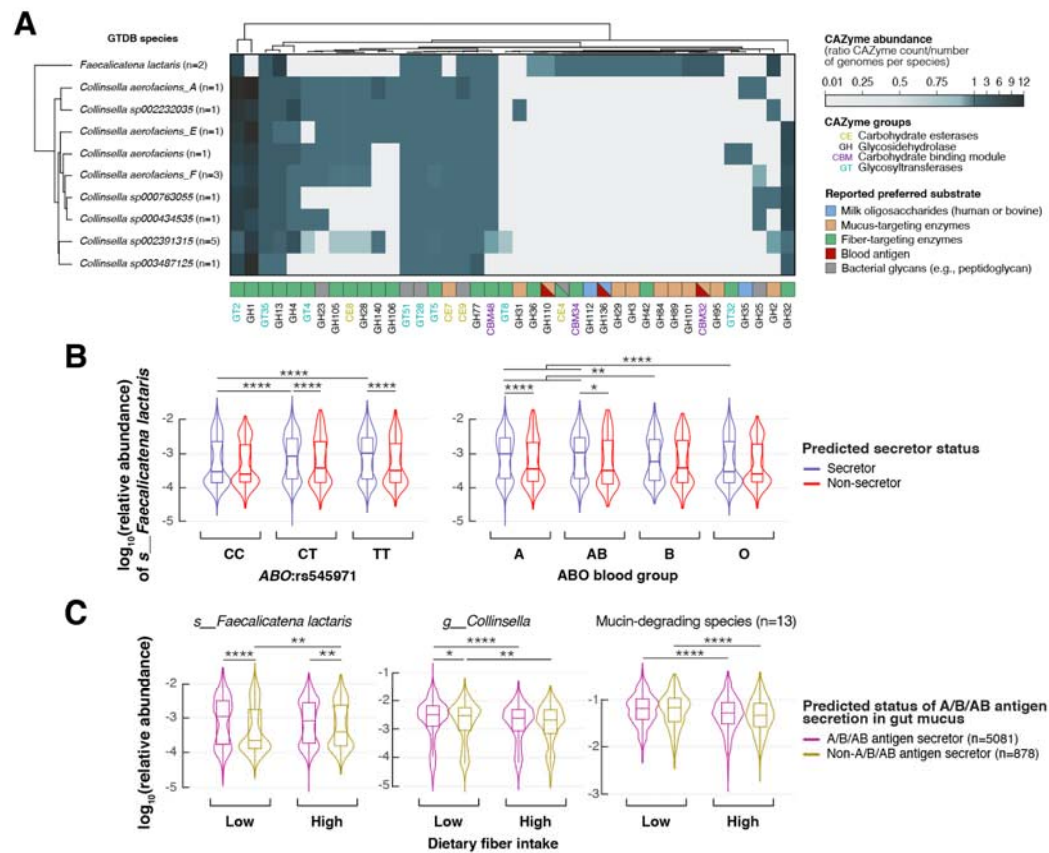
1043
1044
1045
1046
1047
1048
1049
1050

Figure 2. Interaction of human genotype, dairy diet and gut bacterial variation with the *LCT* locus. (A) The 4 panels present variation in microbial abundances for the 4 most significantly associated taxa with the *LCT* locus: *Bifidobacterium*, *Negativibacillus*, *UBA3855 sp900316885* and *CAG-81 sp000435795*. Abundances are compared across stratified groups of individuals from the FR02 cohort according to *LCT-MCM6:rs4988235* genotype and self-reported dietary lactose intake (red: regular dairy diet, blue: lactose-free diet). Sample sizes for groups of individuals self-reporting a regular dairy diet: rs4988235:TT (n=1,786), CT (n=2,413), CC (n=736); self-reporting a non-regular dairy diet or lactose-free diet: TT (n=150), CT (n=198), CC (n=245). All statistical comparisons denote the p-values of Wilcoxon rank test on the distributions of untransformed relative abundances. P-values thresholds are abbreviated as follow: *:p≤0.05; **:p≤0.01; ***:p≤0.001; ****:p≤0.0001. Only significantly different comparisons are indicated. (B) Host genetics and gut microbes interact in the context of dairy intake and lactose intolerance.

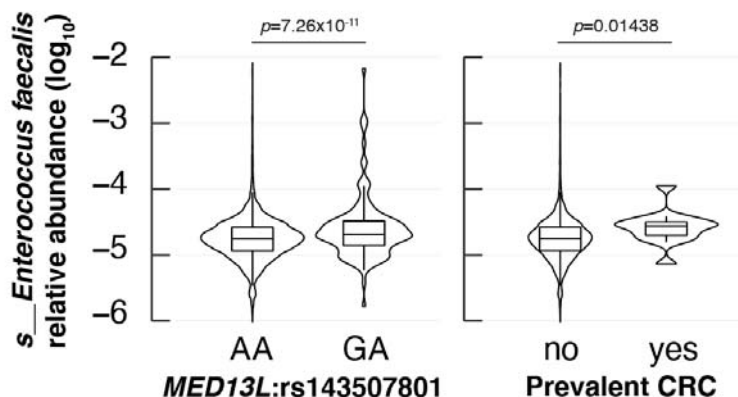


1051

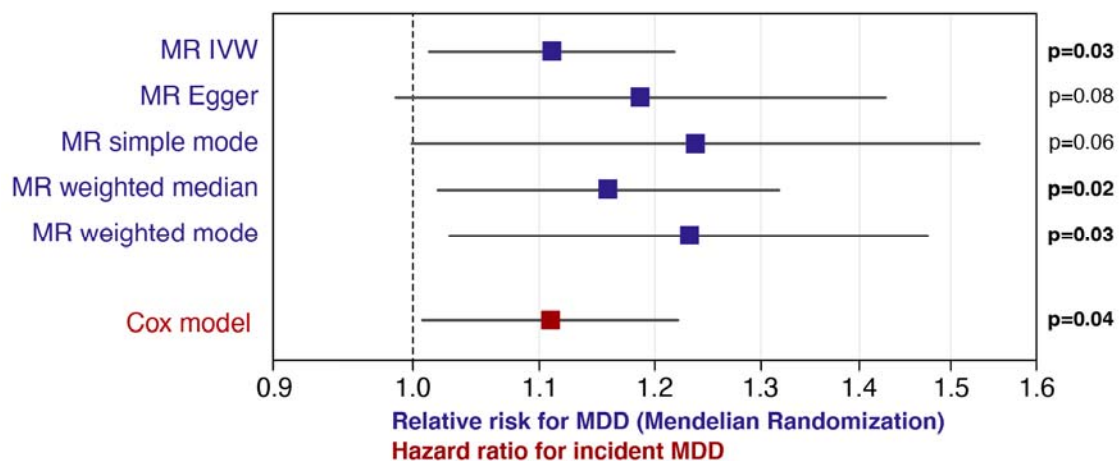
1052 **Figure 3. Functional profiling and effect of host genetics and dietary fiber intake on gut abundance variation of two bacterial taxa associated with the ABO locus**
1053 (A) Carbohydrate-active enzymes (CAZyme) distribution patterns in previously published *F. lactaris* and *Collinsella* reference genomes which were included in the GTDB
1054 release 89 index used to classify metagenomes in this study. The heatmap indicates species abundance in corresponding CAZyme families, corresponding to the total count
1055 of detected families for each species divided by the number of reference genomes examined for the same species. Values <1 indicate that less than one copy per genome of
1056 the corresponding CAZyme family was detected for each, values >1 indicate that more than one copy per genome was detected. Preferred substrate groups are based on
1057 literature search and descriptions on CAZypedia.org. (B) ABO-associated *F. lactaris* abundances are compared across stratified groups of individuals from the FR02 cohort
1058 according to (left panel): ABO:rs4988235 genotype and predicted secretor status (blue: secretor status conferred by FUT2 rs601338:AG/AA genotype; red: non-secretor
1059 status conferred by FUT2 rs601338:GG genotype) and (right panel) according to predicted A, AB, B and O blood types, and predicted secretor status. Sample sizes for
1060 compared groups of individuals: secretor status with rs545971:C/C (n=1,538), C/T (n=2,493), T/T (n=1,050) and blood group A (n=2,178), AB (n=460), B (n=900), O
1061 (n=1,543); non-secretor status with rs545971:C/C (n=266), C/T (n=437), T/T (n=175) and blood group A (n=383), AB (n=80), B (n=148), O (n=267). (C) ABO-associated
1062 *F. lactaris* and *Collinsella sp.* abundances, as well as compounded abundances from 13 mucin-degrading species from Tailford *et al.* (2015), are compared across stratified
1063 groups of individuals from the FR02 cohort according to the predicted A/B/AB-antigen secretion status and dietary fiber intake. The A/B/AB-antigen secretion status was
1064 defined to segregate individuals according to the predicted phenotype of releasing soluble A/B/AB oligosaccharides branched onto a H-antigen into the gut mucosa.
1065 A/B/AB-antigen secretors were defined as secretor individuals from blood types A, AB and B. Non- A/B/AB-antigen secretors were defined as non-secretor individuals and
1066 O-antigen secretors. Fiber intake was compared in individual groups from the top and bottom quartiles of total fiber score based on dietary questionnaires and approximating
1067 the amount of fiber in an individual's diet. Sample sizes for compared groups of individuals: A/B/AB-antigen secretors (n=1393) following a low-fiber diet (n=723) or a
1068 fiber-rich diet (n=670), or non- A/B/AB-antigen secretors (n=952) following a low-fiber diet (n=490) or a fiber-rich diet (n=462). All statistical comparisons denote the p-
1069 values of Wilcoxon rank test on the distributions of untransformed relative abundances. P-values thresholds are abbreviated as follow: *, $p \leq 0.05$; **, $p \leq 0.01$; ***, $p \leq 0.001$;
1070 ****, $p \leq 0.0001$. Only significantly different comparisons are indicated. (D) Host genetics and gut microbes interact in the context of fiber intake, secretor status and blood
1071 types.
1072



1074 **Figure 4. Effect of host genetics and prevalent colorectal cancer on gut levels of *Enterococcus faecalis* associated with *MED13L* variation across participants of the**
 1075 **FR02 cohort.** Abundances are compared across individuals grouped according to (left panel): *MED13L*:rs143507801 genotype, (right panel): colorectal cancer (CRC,
 1076 prevalence according to the Finnish Cancer Registry). The comparison between *E. faecalis* variation and *MED13L*:rs143507801 reflects the GWAS results (Table S1). The
 1077 comparison of *E. faecalis* abundances in individuals with or without CRC at baseline was performed using a Wilcoxon rank test. Sample sizes for compared groups of
 1078 individuals: rs143507801:A/A (n=5,825), G/A (n=130) (Note: only 1/5959 individual in our cohort was G/G); with CRC (n=14), without CRC at baseline (n=5,941).



1079 **Figure 5. MR-based causal effects and incident depression analysis link *Morganella* with major depressive disorder.** The plot shows results for 5 concurring MR
 1080 methods and hazard ratio for incident MDD in the FR02 cohort up to 16 years after baseline sampling using Cox model.
 1081
 1082



1083

1084
1085
1086

Table 1. Study-wide significant SNP-taxon associations after GWAS. A full table including the associated genotypes and bacterial taxa at genome-wide significance level as well as the full GTDB taxonomic path of all taxa are included in Table S1.

Locus	Chromosome	Top associated variant (rsID)	Effect allele	Non-effect allele	Associated bacterial taxon (GTDB release 89 nomenclature)	GTDB phylum	If different, best matching taxon synonym(s) from NCBI nomenclature*	beta	SE	P
<i>LCT-MCM6</i>	2	rs3940549	G	A	<i>c_Actinobacteria</i>	<i>p_Actinobacteriota</i>	-	0.086	0.007	5.02×10 ⁻³⁵
<i>LCT-MCM6</i>	2	rs3940549	G	A	<i>o_Actinomycetales</i>	<i>p_Actinobacteriota</i>	<i>o_Bifidobacteriales</i> ; <i>o_Micrococcales</i>	0.123	0.010	3.74×10 ⁻³²
<i>LCT-MCM6</i>	2	rs3940549	G	A	<i>g_Bifidobacterium</i>	<i>p_Actinobacteriota</i>	-	0.180	0.015	1.69×10 ⁻³¹
<i>LCT-MCM6</i>	2	rs3940549	G	A	<i>s_Bifidobacterium_ruminantium</i>	<i>p_Actinobacteriota</i>	-	0.150	0.014	2.00×10 ⁻²⁸
<i>LCT-MCM6</i>	2	rs3940549	G	A	<i>s_Bifidobacterium_breve</i>	<i>p_Actinobacteriota</i>	-	0.118	0.011	8.01×10 ⁻²⁸
<i>LCT-MCM6</i>	2	rs62168795	C	T	<i>f_Bifidobacteriaceae</i>	<i>p_Actinobacteriota</i>	-	0.160	0.015	8.96×10 ⁻²⁸
<i>LCT-MCM6</i>	2	rs3940549	G	A	<i>p_Actinobacteriota</i>	<i>p_Actinobacteriota</i>	<i>p_Actinobacteria</i>	0.051	0.005	1.20×10 ⁻²²
<i>LCT-MCM6</i>	2	rs3940549	G	A	<i>s_Bifidobacterium_infantis</i>	<i>p_Actinobacteriota</i>	<i>s_Bifidobacterium_longum</i>	0.112	0.011	2.37×10 ⁻²²
<i>LCT-MCM6</i>	2	rs3940549	G	A	<i>s_Bifidobacterium_angulatum</i>	<i>p_Actinobacteriota</i>	-	0.137	0.014	4.40×10 ⁻²²
<i>LCT-MCM6</i>	2	rs3940549	G	A	<i>s_Bifidobacterium_kashiwanohense</i>	<i>p_Actinobacteriota</i>	-	0.134	0.015	5.92×10 ⁻²⁰
<i>LCT-MCM6</i>	2	rs182549	C	T	<i>s_Bifidobacterium_adolescentis</i>	<i>p_Actinobacteriota</i>	-	0.207	0.023	6.11×10 ⁻²⁰
<i>LCT-MCM6</i>	2	rs62168795	C	T	<i>s_UBA3855_sp900316885</i>	<i>p_Firmicutes_A</i>	Uncultured <i>Clostridiales</i> bacterium	0.055	0.006	7.17×10 ⁻¹⁹
<i>LCT-MCM6</i>	2	rs3940549	G	A	<i>s_Bifidobacterium_pseudocatenulatum</i>	<i>p_Actinobacteriota</i>	-	0.154	0.017	8.65×10 ⁻¹⁹
<i>LCT-MCM6</i>	2	rs3940549	G	A	<i>s_Bifidobacterium_catenulatum</i>	<i>p_Actinobacteriota</i>	-	0.169	0.019	1.20×10 ⁻¹⁸
<i>LCT-MCM6</i>	2	rs3940549	G	A	<i>s_CAG-81_sp000435795</i>	<i>p_Firmicutes_A</i>	Uncultured <i>g_Clostridium</i> sp. CAG:58	-0.151	0.017	3.72×10 ⁻¹⁸
<i>LCT-MCM6</i>	2	rs3940549	G	A	<i>s_Bifidobacterium_longum</i>	<i>p_Actinobacteriota</i>	-	0.131	0.015	1.41×10 ⁻¹⁷
<i>LCT-MCM6</i>	2	rs4988235	G	A	<i>s_Bifidobacterium_bifidum</i>	<i>p_Actinobacteriota</i>	-	0.166	0.020	1.31×10 ⁻¹⁶
<i>ABO</i>	9	rs545971	T	C	<i>s_Faecalicatena_lactaris</i>	<i>p_Firmicutes_A</i>	<i>s_Ruminococcus lactaris</i> (NCBI) <i>s_Ruminococcus_B lactaris</i> (GTDB release 86) <i>s_Mediterraneibacter lactaris</i> (GTDB release 95)	0.105	0.015	1.10×10 ⁻¹²
<i>MED13L</i>	12	rs187309577	T	C	<i>g_Enterococcus</i>	<i>p_Firmicutes</i>	-	0.177	0.025	1.84×10 ⁻¹²
<i>LCT-MCM6</i>	2	rs182549	C	T	<i>g_Negativibacillus</i>	<i>p_Firmicutes_A</i>	<i>g_Negativibacillus</i> ; <i>g_Clostridium</i>	-0.081	0.012	5.22×10 ⁻¹²

* NCBI-GTDB taxonomic synonyms were retrieved using the GTDB Taxon History tool: https://gtdb.ecogenomic.org/taxon_history/

1087