Bacterial metabolic signatures in MASLD predicted through gene-centric studies in stool metagenomes

Juan Manuel Medina-Méndez[1,2], Paula Iruzubieta[2], Raúl Fernández-López[1], Javier Crespo[2], Fernando de la Cruz[1,*]

Author affiliations

[1] Instituto de Biomedicina y Biotecnología de Cantabria (IBBTEC), Universidad de Cantabria-CSIC, Santander, Spain

[2] Servicio de Gastroenterología y Hepatología. Grupo de Investigación Clínica y Traslacional en Enfermedades Digestivas. Instituto de Investigación Valdecilla (IDIVAL). Hospital Universitario Marqués de Valdecilla, Santander, Spain

[*] Corresponding author: delacruz@unican.es; Tel.: +34-942-201942

Abstract

Metabolic dysfunction-associated steatotic liver disease (MASLD) is a multifactorial condition influenced by the gut microbiome (GM). While previous studies have reported inconsistent associations between MASLD and key microbial clades using low-resolution 16S rRNA profiling, we employed high-resolution metagenomic sequencing and multi-marker taxonomic classification across three independent cohorts to identify robust microbial and functional signatures of MASLD. We consistently detected a depletion of *Agathobacter rectalis*, a known butyrate producer, in MASLD patients. Functionally, MASLD was characterized by a depletion of genes involved in butyrate and methane biosynthesis -particularly within the crotonyl-butyryl-CoA axis- alongside an enrichment of genes

driving the production of endogenous alcohols such as ethanol and 1-propanol. Genes encoding these fermentative pathways, often organized in operons like *pdu* and *tor*, were more abundant in MASLD samples, indicating a potential shift toward alcohol-producing metabolism. These geno-metabolic changes were accompanied by a broader displacement of beneficial taxa and an increase in accessory gene content across the GM, underscoring the limitations of taxonomy-based disease associations. Many of the differentially abundant genes were also found on plasmids, suggesting that horizontal gene transfer contributes to strain-level metabolic variability relevant to MASLD progression. Our findings support a model in which GM-driven metabolic shifts -rather than taxonomic changes alone- play a central role in MASLD pathogenesis, highlighting the importance of functional and mobile genetic element (MGEs) profiling for uncovering mechanistic links between the microbiome and liver disease.

Keywords

MASLD, gut microbiome, metagenomic sample, metabolic gene, RPKSM, differential gene abundance

1. Introduction

Metabolic dysfunction-associated steatotic liver disease (MASLD) is the most common liver disease worldwide, affecting approximately one quarter of the global population[1]. Clinically, MASLD is defined by the presence of steatosis (i.e., accumulation of intrahepatic triglycerides) in more than 5% of hepatocytes, in association with metabolic risk factors (particularly, obesity and type 2 diabetes) and in the absence of excessive alcohol consumption (≥30 g per day for men and ≥20 g per day for women) or other chronic liver diseases[2]. However, the term encompasses a spectrum of pathological situations, ranging from simple steatosis to metabolic dysfunction-associated steatohepatitis (MASH), which can further progress to fibrosis, cirrhosis and hepatocellular carcinoma[3]. MASLD also significantly increases

42 the risk of cardiovascular disease, extrahepatic malignancies (e.g., colorectal or breast cancer), liver-
43 related complications or chronic kidney disease[4]. The pathogenesis of MASLD is multifactorial, involving
44 genetic, environmental, inflammatory and metabolic factors, and it has a bidirectional association with
45 components of the metabolic syndrome, particularly insulin resistance[5].

46 The gut microbiome (GM) interacts with the liver via the gut-liver axis (GLA) through the portal
47 vein, which transports GM-derived metabolites to the liver. This axis can be disrupted by multiple
48 interconnected factors, affecting the progression of MASLD and fibrosis[6]. Alterations of the GM
49 composition and/or their associated metabolic production can lead to an increased production of
50 harmful metabolites such as pathogen-associated molecular patterns, which induce inflammatory
51 responses in the liver when they reach it through the bloodstream, leading to liver injury[7,8]. The
52 transient of these endotoxins into portal circulation is facilitated by an increased intestinal permeability
53 caused by a disruption of tight junction proteins in the intestinal membrane[9]. Additionally, alterations
54 of the bile acid composition and circulation affect liver metabolism and function. Since bile acids play a
55 role in the regulation of glucose and lipid metabolism, their dysregulation contributes to gut
56 inflammation and MASLD[10,11].

57 Several studies have attempted to establish a causal link between bacterial metabolism and
58 MASLD[12,13]. However, metabolic levels can also be influenced by host metabolism, diet or
59 environmental factors. Similarly, metataxonomic analysis have explored the relationship between GM
60 composition and MASLD[14,15]. Although 16S rRNA gene sequencing provides taxonomic resolution at the
61 genus or species level, it fails to capture strain-level metabolic variability. Bacterial genomic plasticity
62 allows key metabolic genes, such as those potentially relevant to MASLD pathogenesis, to be encoded

63 in their accessory genome, including MGEs, further complicating the association between taxonomy
64 and metabolic function[16].

65 In this study, we used the genes encoding metabolic enzymes involved in the final steps of
66 microbial pathways leading to the production of butyrate, short-chain alcohols (SCAs) ethanol and
67 propanol, methane, trimethylamine (TMA) and trimethylamine N-oxide (TMAO) as proxies to
68 determine their abundance in MASLD. To achieve this, we isolated these gene families and quantified
69 them in over 550 metagenomes obtained from fecal samples of MASLD and healthy patients, stratified
70 across three cohorts. We also explored taxonomic signatures to gain a comprehensive view of bacterial
71 diversity and abundance. Finally, we assessed the presence of these genes within the human GM
72 genomes and plasmids. This approach provided insights into the role of microbial pathways in MASLD
73 progression and highlighted key genetic signatures in both healthy and MASLD patients.

74 **2. Results**

75 ***Agathobacter rectalis* is depleted in MASLD**

76 Certain bacterial clades have been positively correlated with MASLD progression, including
77 *Enterobacteriaceae*[15,17] such as *Escherichia / Shigella spp*.[17,18], *Bacteroidaceae* like *Bacteroides spp.*[18,19],
78 *Veillonellaceae*[15] or *Streptococcus spp.*[20]. Conversely, other clades have been reported to exhibit an
79 inverse correlation with MASLD progression, including *Ruminococcaceae*[15] such as *Faecalibacterium*
80 *prausnitzii*[18,21], *Lachnospiraceae* like *Eubacterium rectale*[18,21] or *Dorea longicatena*[21], *Clostridiaceae* like
81 *Clostridium spp.*[17], and even *Pseudomonas spp.*[20]. However, these studies often yield conflicting results.
82 For example, a reported a positive correlation between *Ruminoccocus spp.* abundance and MASLD
83 progression[19] contradicts one of the main findings in two posterior studies[18,21]. Similarly, *Prevotella spp.*
84 was found to decrease with MASLD development[22] and fibrosis stage[19], yet another study reported an

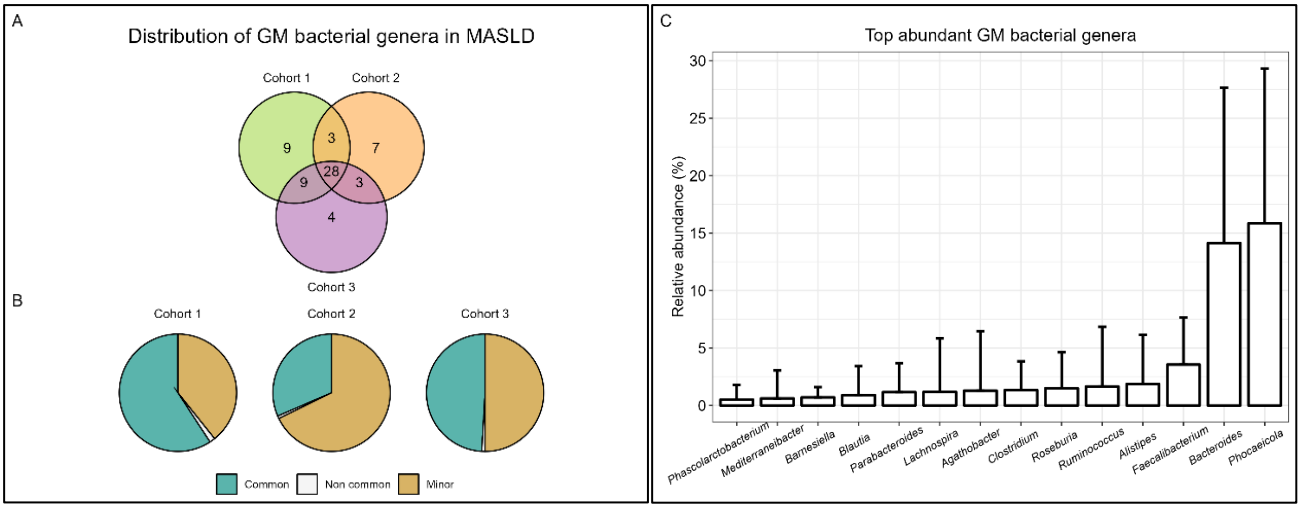
85 increase in this genus in MASLD patients[20]. There are additional examples of discordant taxonomic
86 signatures in MASLD, such as *Blautia spp.* or *Roseburia spp.*[23].

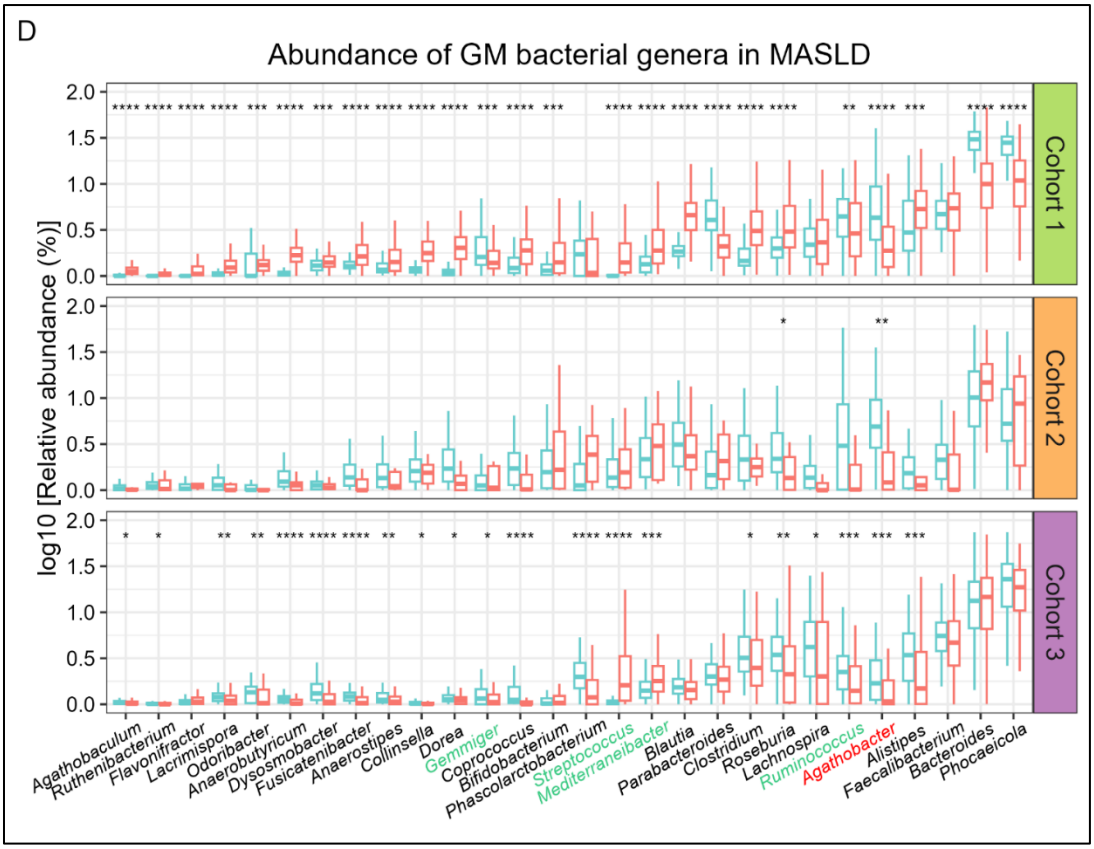
87 To investigate whether any bacterial genera is systematically altered in MASLD, we profiled the
88 taxonomic composition of the GM communities across all samples from the three cohorts at both genus
89 and species level using MetaPhlAn[24], as detailed in Materials and Methods. This analysis identified 28
90 bacterial genera shared across all three cohorts (Fig. 1A), which accounted for 59%, 31.5% and 49% of
91 the GM fraction in each cohort, respectively (Fig. 1B). Thus, nearly half of the bacterial genera identified
92 in at least one cohort were shared across all three (28 out of 63), representing between one-third and
93 two-thirds of the total GM abundance.

94 Among these shared genera, *Gemmiger*, *Streptococcus*, *Mediterraneibacter*, *Ruminococcus* and
95 *Agathobacter* exhibited a consistent pattern of differential abundance between groups across all
96 cohorts, reaching statistical significance in at least two (Fig. 1D). This exclusion criterion was applied to
97 enhance the reliability of the results, ensuring that differentially abundant genera were not solely
98 driven by heterogeneous demographic variables (e.g. geographical location, age, sex, or ethnicity) or
99 environmental factors (e.g. diet, physical activity, comorbidities, or drug exposure), rather than being
100 directly related to the MASLD phenotype[25,26]. However, this threshold was not applied to Cohort 2
101 because all patients in this group had MASLD, and applying it could have been overly restrictive,
102 masking potential taxonomic differences associated with the disease. Of these five genera, only
103 *Agathobacter spp.* showed statistically significant differences across the three cohorts (Fig. 1D). Note
104 that *Agathobacter*, along with *Ruminococcocus* and *Mediterraneibacter*, belongs to the fraction of the
105 highest abundant GM bacterial genera (Fig. 1C), highlighting their potential contribution to the
106 phenotype in specific patient cohorts.

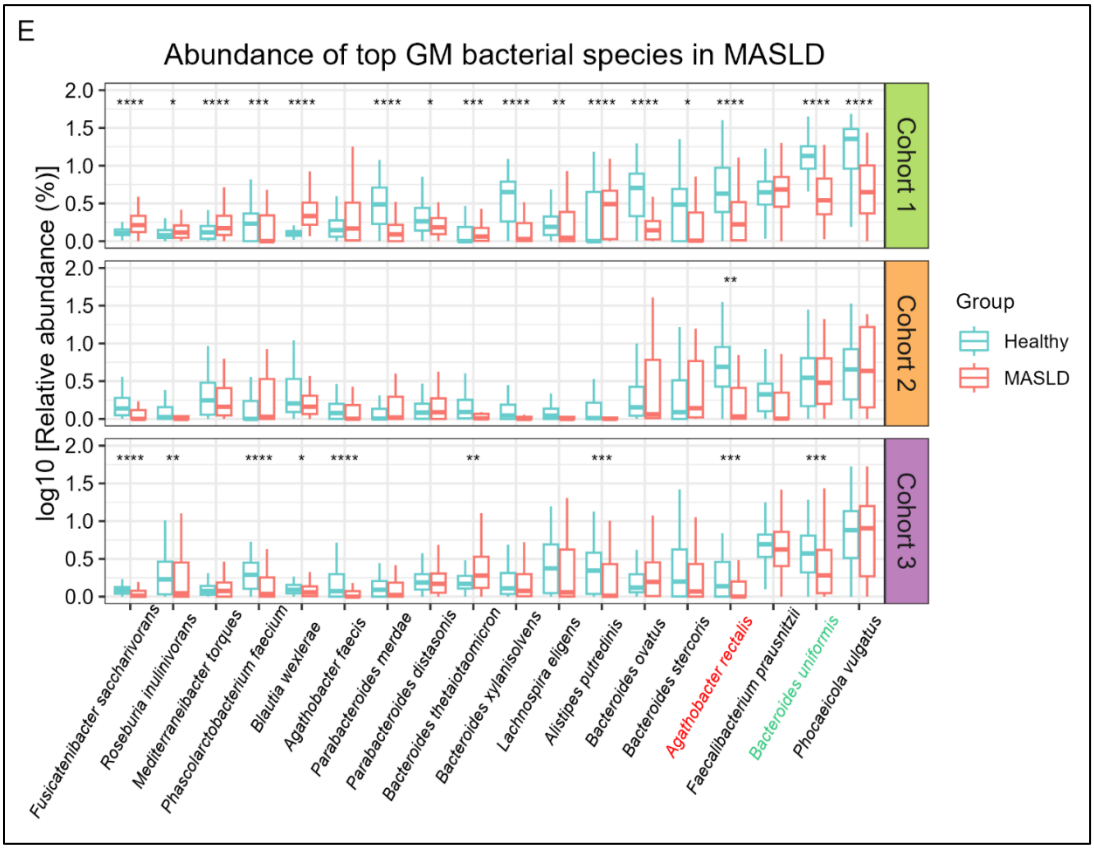
Analogously, the most abundant bacterial species within the GM were analyzed for differences in abundances between groups across the three cohorts (Fig. 1E). Only *Agathobacter rectalis* and *Bacteroides uniformis*, two of the most abundant GM bacterial species (Supplementary Fig. 1), exhibited a consistent pattern of differential abundance across all cohorts, with statistical significance in at least two. However, only *A. rectalis* showed a significant difference between groups in all cohorts, identifying it as a robust taxonomic marker depleted in MASLD.

Figure 1. Taxonomic profiling of GM in MASLD. (A) Number of bacterial genera common to the three cohorts. **(B)** Average relative abundance of bacterial genera predicted by MetaPhlAn across the three study cohorts. The blue section of the pie charts reflects the average relative abundance of bacterial genera common to the three cohorts (n=28), whereas the grey and green sections reflect, respectively, the abundances of non-common and minor genera (i.e., with an average relative abundance of 0 in at least one of the cohorts). **(C)** Top abundant GM genera (average absolute abundance >0.5%) **(D-E)** Relative abundance of bacterial genera **(D)** and species **(E)** in the metagenomes of the fecal samples, shown on a logarithmic scale. Blue and red boxplots correspond to healthy and MASLD groups, respectively. Each facet represents one cohort. Differences in abundance were evaluated using pairwise Mann-Whitney tests with Benjamini-Hochberg adjustment for multiple testing. Horizontal bars represent median values, while boxes and whiskers represent the interquartile range and the full data spread (minimum to maximum), respectively. Clades depicted in green across the x-axis have a consistent differential abundance pattern between groups over the three cohorts. Only *Agathobacter rectalis*, depicted in red, has a consistent and statistically significant decrease in the MASLD group across cohorts. (*) p<0.05, (**) p<0.01, (***) p<0.001, (****) p<0.0001. Non-starred genera present non-significant differences in abundance between groups.




D
Abundance of GM bacterial genera in MASLD
log10 [Relative abundance (%)]
Cohort 1
Cohort 2
Cohort 3
Agathobaculum
Ruthenibacterium
Flavonifractor
Lacrimispora
Odoribacter
Anaerobutyricum
Dysosmobacter
Fusicatenibacter
Anaerostipes
Collinsella
Dorea
Gemmiger
Coprococcus
Bifidobacterium
Phascolarctobacterium
Streptococcus
Mediterraneibacter
Blautia
Parabacteroides
Clostridium
Roseburia
Lachnospira
Ruminococcus
Agathobacter
Alistipes
Faecalibacterium
Bacteroides
Phocaeicola

129


E
Abundance of top GM bacterial species in MASLD
log10 [Relative abundance (%)]
Cohort 1
Cohort 2
Cohort 3
Group
Healthy
MASLD
Fusicatenibacter saccharivorans
Roseburia inulinivorans
Mediterraneibacter torques
Phascolarctobacterium faecium
Blautia wexlerae
Agathobacter faecis
Parabacteroides merdae
Parabacteroides distasonis
Bacteroides thetaiotaomicron
Bacteroides xylanisolvens
Lachnospira eligens
Alistipes putredinis
Bacteroides ovatus
Bacteroides stercoris
Agathobacter rectalis
Faecalibacterium prausnitzii
Bacteroides uniformis
Phocaeicola vulgatus

130

131 **Butyrate-producing genes are depleted in MASLD**

132 The identification of *A. rectalis* as a consistent taxonomic signature across multiple MASLD
133 cohorts underscores its potential protective role in MASLD development. However, taxonomy alone
134 may not fully capture the functional contributions of the GM. Certain microbial metabolites regulate
135 GLA integrity and have been suggested to influence MASLD progression[27–29]. These compounds derive
136 from dietary carbohydrates and proteins that are not fully digested in the upper gastrointestinal tract
137 and are instead catabolized by the GM through fermentative pathways that are predominant in the
138 anaerobic environment of the colon[30,31]. Once absorbed into the bloodstream, they are transported to
139 the liver via the portal vein, where they can exert both beneficial and detrimental effects[32,33].

140 The most well-characterized of these metabolites are short chain fatty acids, produced by the
141 colonic fermentation of complex carbohydrates present in dietary fiber that escape digestion in the
142 small intestine. These metabolic pathways have been relatively well understood for nearly thirty years[34]
143 and their impact on GM homeostasis has been reviewed in several articles[35–37]. For this reason, we
144 interrogated the metagenomic fecal samples from the patient cohorts for metabolic signatures that
145 could explain the MASLD phenotype. First, we focused on the genes encoding enzymes involved in
146 butyrate formation as a proxy for the bacteria that carry them, independently of their taxa. To validate
147 our hypothesis, we analyzed the genic abundance of the final enzymes involved in butyrate production.

148 We selected butyrate as a target because it is crucial for maintaining GLA homeostasis by
149 supporting the colonic epithelial barrier[38,39]. It also promotes lipid oxidation[40,41], induces hepatic
150 gluconeogenesis[42], and prevents insulin resistance[43], all of which are key factors in MASLD. In mice,
151 both butyrate-producing probiotics[44] and sodium butyrate administration[45–47] have been shown to

152 prevent steatosis and slow MASLD progression. Additionally, fecal microbiota transplantation (FMT)
153 from healthy to MASH mice increased cecal butyrate levels and reduced steatosis[48].

154 Bacterial butyrate can be formed through two major terminal reactions (Fig. 2A). The first
155 involves a dephosphorylation of butyryl-P that yields one ATP through an energetically favorable
156 reaction that is catalyzed by a butyrate kinase encoded by *buk.* In the precursor reaction, butyryl-P is
157 formed through a phosphorylation of butyryl-CoA catalyzed by a phosphate butyryl transferase
158 encoded by *ptb*. Butyrate also accumulates after the transfer of a CoA cofactor from butyryl-CoA to
159 acetate through butyryl-CoA:acetate CoA transferase, encoded by *but*. This CoA-transferase pathway
160 also conserves the energy of the CoA bond in the newly formed CoA-moiety of the co-substrate. Butyryl-
161 CoA is produced from crotonyl-CoA through two different enzymes: butyryl-CoA trans-2-
162 oxidorreductase, encoded by *fabV*, and butyryl-CoA dehydrogenase, encoded by *bcd.* This pathway is
163 further reviewed in the human GM elsewhere[49].

164 We quantified the abundance of these genes in the metagenomic fecal samples of both groups
165 across the three MASLD cohorts, as described in Materials and Methods. Genes involved in the
166 production of butyrate and its precursors, butyryl-P and butyryl-CoA, were significantly depleted in the
167 GM of patients with cirrhosis MASLD compared to the healthy group, and their abundance also
168 decreased significantly with MASLD progression (Fig. 2B). The precursor gene *fabV* was nearly an order
169 of magnitude less abundant than *ptb* and *bcd* in both cohorts, underscoring the greater prevalence of
170 *bcd* over *fabV* in the CoA-transferase route. Additionally, *ptb* did not exhibit significant differences in
171 Cohort 2. Gene *but* was significantly depleted in the control group in both cohorts, whereas *buk* did not
172 show differences in Cohort 2 either. Notably, the depletion of butyrate-related genes was not observed
173 in MASLD patients from Cohort 1.

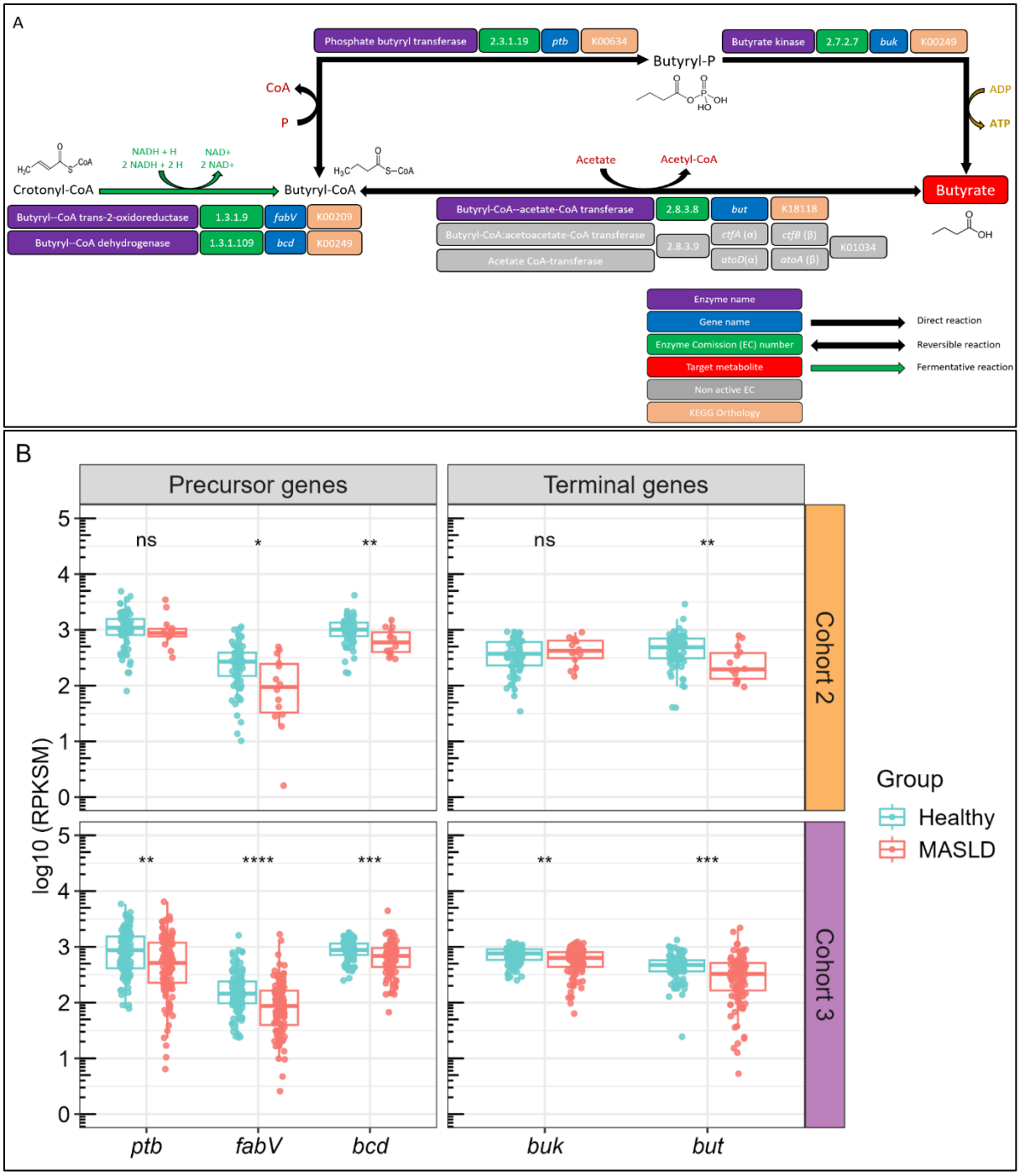


Figure 2. Abundance of butyrate-producing genes in patient cohorts. (A) Characterization of the terminal reactions involved in the formation of bacterial butyrate from crotonyl-CoA. Enzymes, EC numbers, coding genes and KOs, as well as reaction types, are indicated for each metabolic step according to the legend (lower part). **(B)** Abundance of butyrate-producing genes in Cohorts 2 and 3, shown on a logarithmic scale. Boxplots represent the gene abundance from the metagenomic measures of the fecal samples, and differences in abundances were evaluated using pairwise Mann-Whitney tests with Benjamini-Hochberg adjustment for multiple testing. Upper and lower facets represent abundances in Cohorts 2 and 3, respectively. Left and right facets represent abundances of precursor and terminal coding genes, respectively, as described in (A). RPKSM: reads per kilobase per genic size per million reads. (*) $p<0.05$, (**) $p<0.01$, (***) $p<0.001$, (****) $p<0.0001$, ns: non-significant.

185 **SCA-producing genes are increased in MASLD**

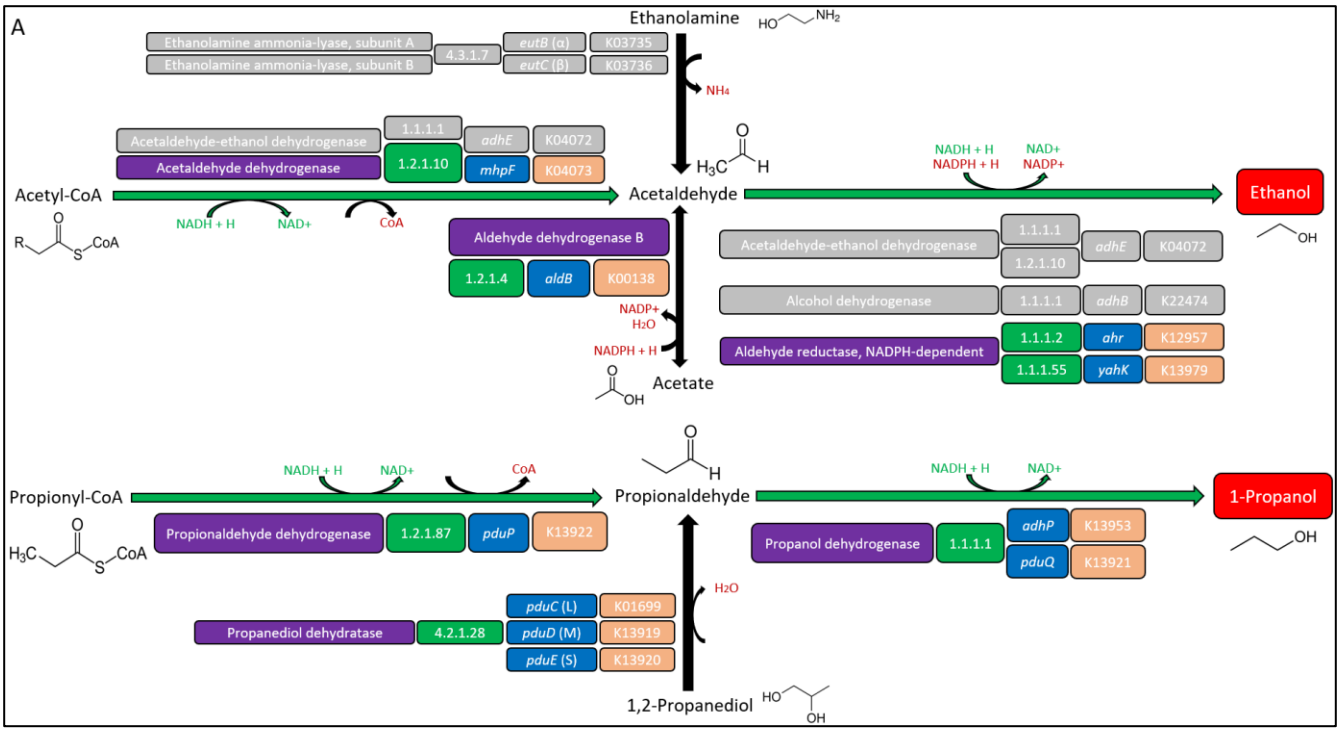
186 Short-chain alcohols (SCAs) are a metabolic family also suggested to regulate GLA integrity, with
187 several studies linking GM endogenous alcohol production to MASLD. One of the first reports, published
188 over a decade ago, described elevated peripheral blood ethanol and increased fecal *Escherichia spp.*
189 16S rRNA levels in patients with MASH[50]. Similar findings were reported in a pediatric cohort, where
190 children with MASLD exhibited higher levels of fecal ethanol, *Prevotella spp.*, and
191 Gammaproteobacteria 16S rRNA[51]. A more recent study showed that postprandial peripheral blood
192 ethanol and fecal *Lactobacillaceae* 16S rRNA levels increased in MASH patients after selective inhibition
193 of host ADH. This effect was reversed with broad-spectrum antibiotic treatment[52].

194 In 2019, Yuan and colleagues identified a high-alcohol-producing strain of *Klebsiella pneumoniae*
195 capable of causing MASLD through ethanol production[53]. Specifically, they found that FMT from a MASH
196 patient containing this strain induced MASLD in healthy mice, whereas removing the strain before FMT
197 prevented disease development. Furthermore, ethanol produced by this *K. pneumoniae* strain was
198 shown to disrupt lipid homeostasis in hepatic cells and to cause mitochondrial dysfunction, oxidative
199 stress, and the accumulation of reactive oxygen species[54]. These factor are critical to MASH
200 development through their role in lipid peroxidation and inflammation[55]. Additionally, a recent
201 computational study identified a positive correlation between the abundance of two fermentative
202 pathways in certain families from class Clostridia and fatty-liver disease in patients. These pathways
203 include the production of ethanol from pyruvate and the heterolactic fermentative pathway[56].

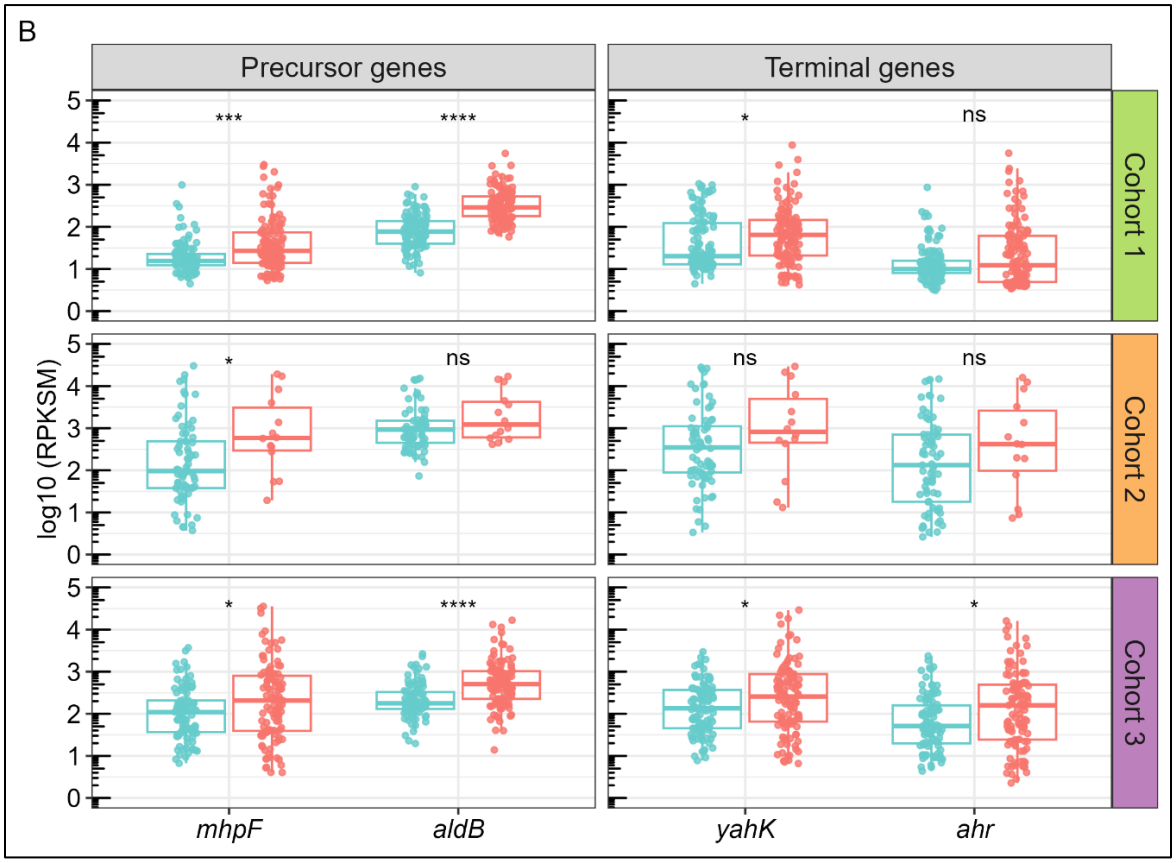
204 With this ample evidence, we hypothesized that genes encoding enzymes involved in the
205 formation of endogenous SCAs may represent another metabolic mechanism contributing to MASLD
206 pathology. We focused on ethanol and propanol, which are produced by the reduction of their

207 respective aldehydes through ethanol and propanol dehydrogenases. These enzymes help to reduce
208 the levels of toxic aldehydes in the GM[57,58] and replenish NAD^+ levels to maintain fermentative growth.
209 Bacterial ethanol is formed from acetaldehyde through a fermentative reaction that can be
210 catalyzed by 1) an NADH-dependent, bifunctional acetaldehyde-ethanol dehydrogenase encoded by
211 *adhE*, 2) an NADPH-dependent alcohol dehydrogenase encoded by *adhB*[59,60] and 3) a NADPH-
212 dependent aldehyde reductase encoded by two different genes: *ahr* and *yahK*, with the corresponding
213 liberation of phosphorylated reducing power in form of $NADP^+$ [61,62]. The precursor acetaldehyde can be
214 formed from 1) ethanolamine degradation, through a reaction catalyzed by ethanolamine ammonia-
215 lyase which is encoded by *eutB* and *eutC* genes and liberates ammonia as a by-product[63], 2) acetyl-CoA,
216 through a fermentative reaction catalyzed by bifunctional AdhE and by a CoA-acetylating acetaldehyde
217 dehydrogenase encoded by *mhpF*[64,65] and 3) acetate, through a dehydration catalyzed by aldehyde
218 dehydrogenase B, encoded by *aldB*[66] (Fig. 3A, top). Bacterial 1-propanol is formed from
219 propionaldehyde through a fermentative reaction catalyzed by a propanol dehydrogenase encoded by
220 two different genes: *adhP* (a propanol-preferring alcohol dehydrogenase) and *pduQ.* Propionaldehyde
221 can be generated through two distinct reactions: 1) a CoA transference from propionyl-CoA catalyzed
222 by a propionaldehyde dehydrogenase encoded by *pduP*; and 2) a dehydration of 1,2 propanediol
223 catalyzed by a propanediol dehydratase composed by three subunits encoded by *pduC*, *pduD* and *pduE.*
224 These genes are co-located within the same operon and encode the corresponding enzymes
225 responsible for the metabolism of 1,2-propanediol within the propanediol-utilizing (Pdu)
226 microcompartment, a specialized bacterial structure found in certain enteric species[67,68] (Fig. 3A,
227 bottom).

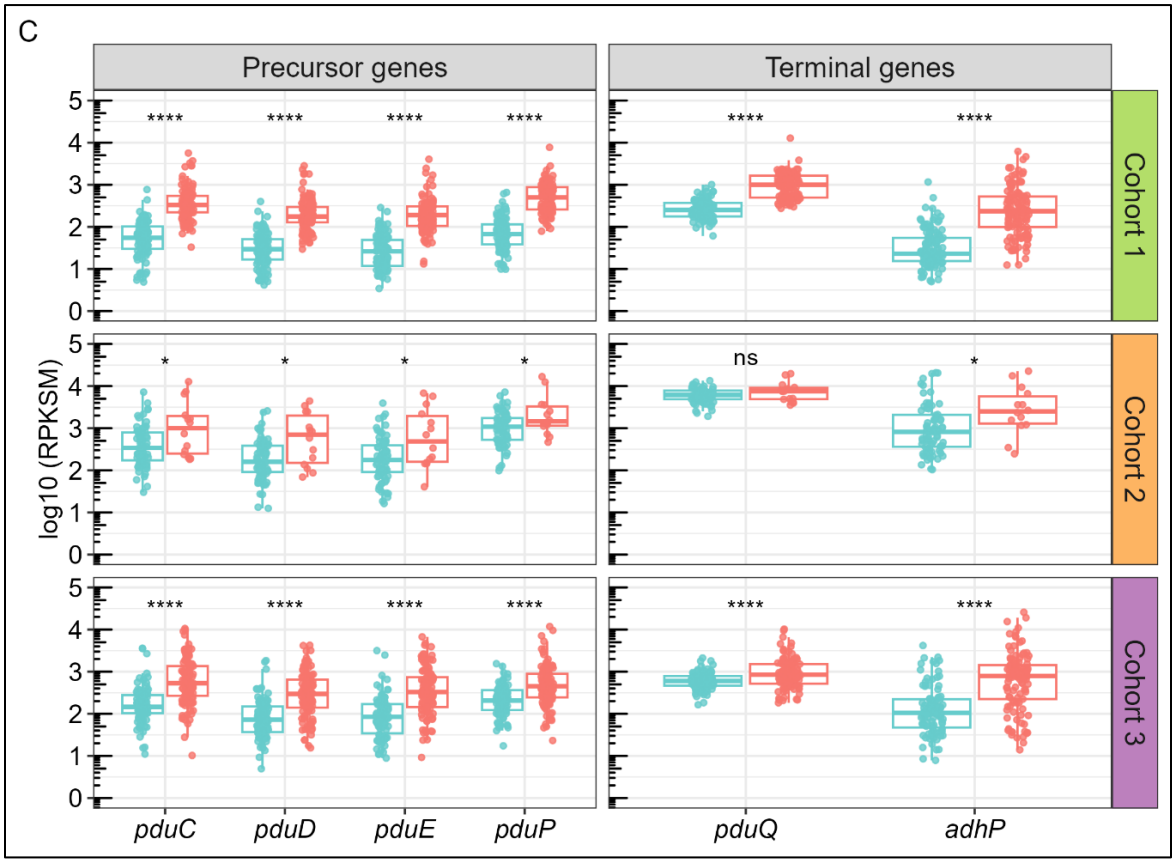
228 We quantified the abundance of genes involved in bacterial SCA formation in the metagenomic
229 samples of the three cohorts as described above. The genes *mhpF* and *aldB*, involved in ethanol
230 production, and *yahK* and *ahr*, involved in acetaldehyde production, were significantly enriched in the
231 GM of MASLD patients from Cohort 1 and Cohort 3 compared to healthy groups. Their abundance also
232 increased with MASLD progression (Fig. 3B). Notably, *mhpF* and *ahr* were depleted compared to *aldB*
233 and *yahK* respectively in all three cohorts. All genes involved in the formation of 1-propanol and
234 propionaldehyde were also significantly enriched in MASLD patients by more than half order of
235 magnitude, increasing with MASLD progression (Fig. 3C, Supplementary Fig. 4A). These findings reveal
236 a systematic increase of SCA-producing genes in MASLD.



237


B
Precursor genes
Terminal genes
Cohort 1
Cohort 2
Cohort 3
log10 (RPKSM)
***
****
*
ns
mhpF
aldB
yahK
ahr

238


C
Precursor genes
Terminal genes
Cohort 1
Cohort 2
Cohort 3
log10 (RPKSM)
****
*
ns
pduC
pduD
pduE
pduP
pduQ
adhP

239

240 **Figure 3. Abundance of SCA-producing genes in patient cohorts. (A)** Characterization of the terminal reactions involved in
241 the formation of bacterial ethanol (top) and propanol (bottom). Enzymes, EC numbers, coding genes and KOs, as well as
242 reaction types, are indicated for each metabolic step according to the legend in Fig. 2A. Non-candidate genes are colored in
243 grey. **(B)** Abundance of *mhpF, aldB, yahK* and *ahr* genes involved in the last two reactions from pyruvate to ethanol in the
244 three cohorts. **(C)** Abundance of *pduC-E, pduP, pduQ* and *adhP* genes, involved in the last two reactions from 1,2-propanediol
245 and propionyl-CoA to 1-propanol in the three cohorts. Boxplot layout, distribution in facets and statistical tests were
246 performed as in Fig. 2C.

247 **Genes involved in methane production are decreased in MASLD, whereas *tor* operons driving TMA**
248 **accumulation are elevated**

249 Trimethylamine N-oxide (TMAO) is a metabolite derived from the oxidation of trimethylamine
250 (TMA), which is produced by the microbial degradation of dietary compounds such as choline and L-
251 carnitine. Circulating TMAO levels have also been linked to MASLD severity[69] and all-cause mortality in
252 MASLD patients[70]. Higher plasmatic TMAO levels also increase cardiovascular disease risk -a condition
253 associated with MASLD- in both mice[71] and clinical patients[72]. Therefore, we investigated whether
254 bacterial TMAO could represent another metabolic mechanism contributing to MASLD.

255 Methane production in the GM occurs through a two-step process: a simultaneous TMA
256 demethylation and CoM methylation, followed by the reduction of methyl-CoM (Fig. 4A). TMA
257 originates from choline degradation via the *cut* operon[73] and is oxidized in the liver to TMAO[74] or
258 reduced back to TMA under anaerobic conditions via the *torCAD* and *torZY* operons. Specifically, *torA*
259 and *torZ* encode two TMAO reductases, whose molecular mechanisms are reviewed elsewhere[75].

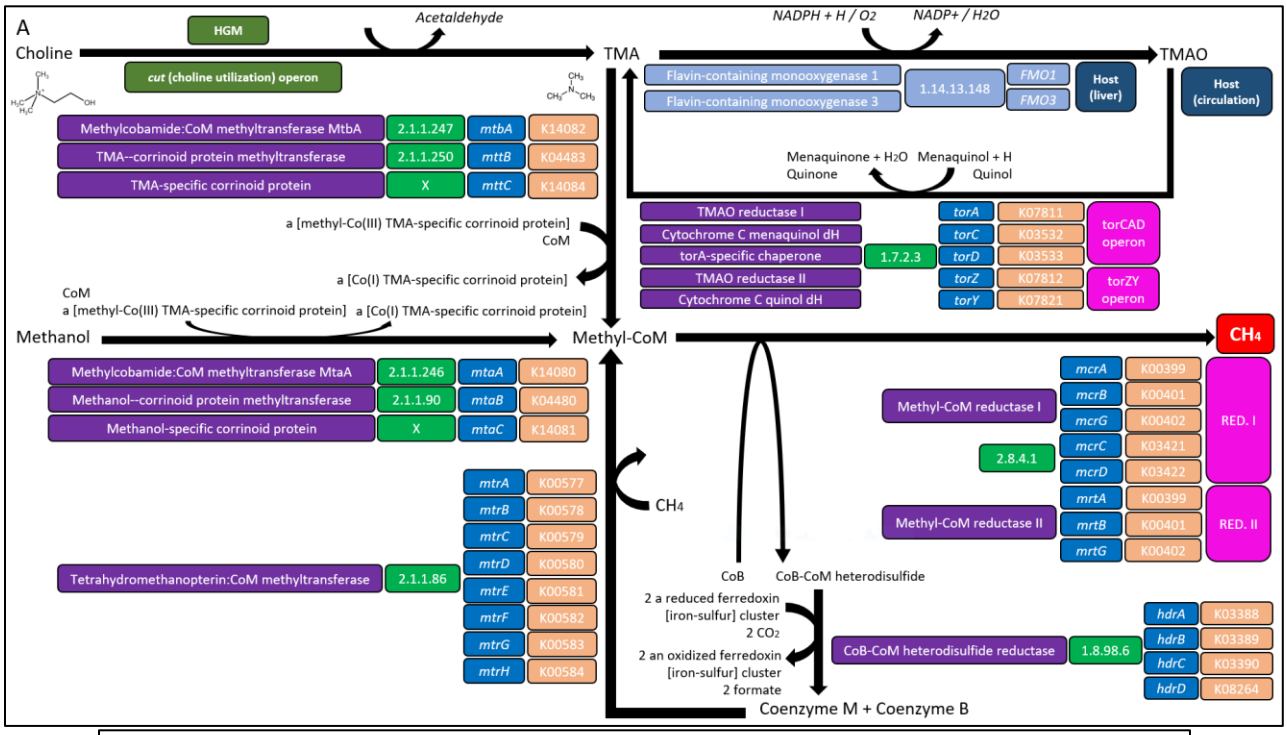
260 CoM methylation occurs through two well-characterized enzymatic systems: a TMA-specific
261 (composed of *mttC*, *mtbA* and *mttB*) and a methanol-specific pathway (composed of *mtaC, mtaA* and
262 *mtaB*)[76]. The final reduction of methyl-CoM to methane is catalyzed by methyl-CoM reductase (MCR) I,
263 which is encoded by the *mcrBDCGA* operon[77]. MCR I has an alternative isozyme termed MCR II, encoded
264 by *mrtBGA* operon[78]. The resulting heterodisulfide (CoB-S-S-CoM) is then reduced to CoM and CoB by
265 a heterodisulfide reductase encoded by operon *hdrABC*, thereby regenerating both cofactors[79].

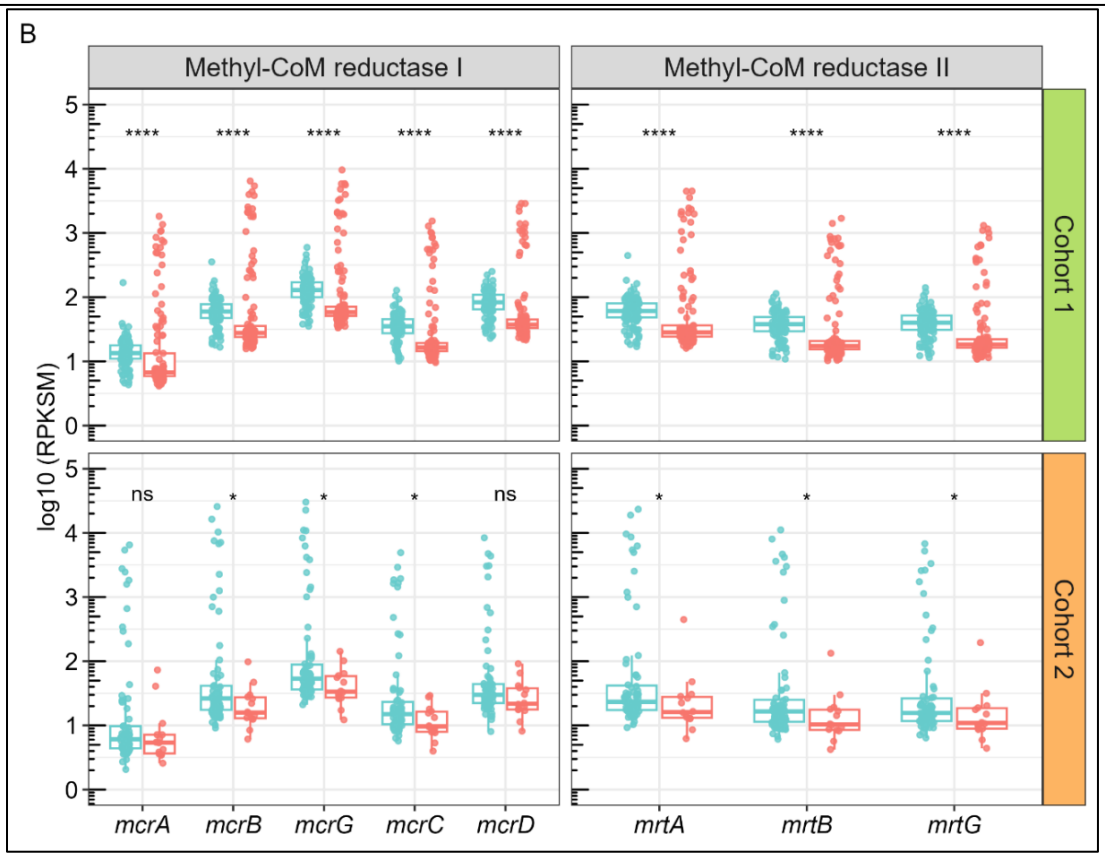
266 Additionally, methyl-CoM is replenished via tetrahydromethanopterin CoM-methyltransferase
267 (encoded by operon *mtrABCDEFGH*), which methylates CoM[80].

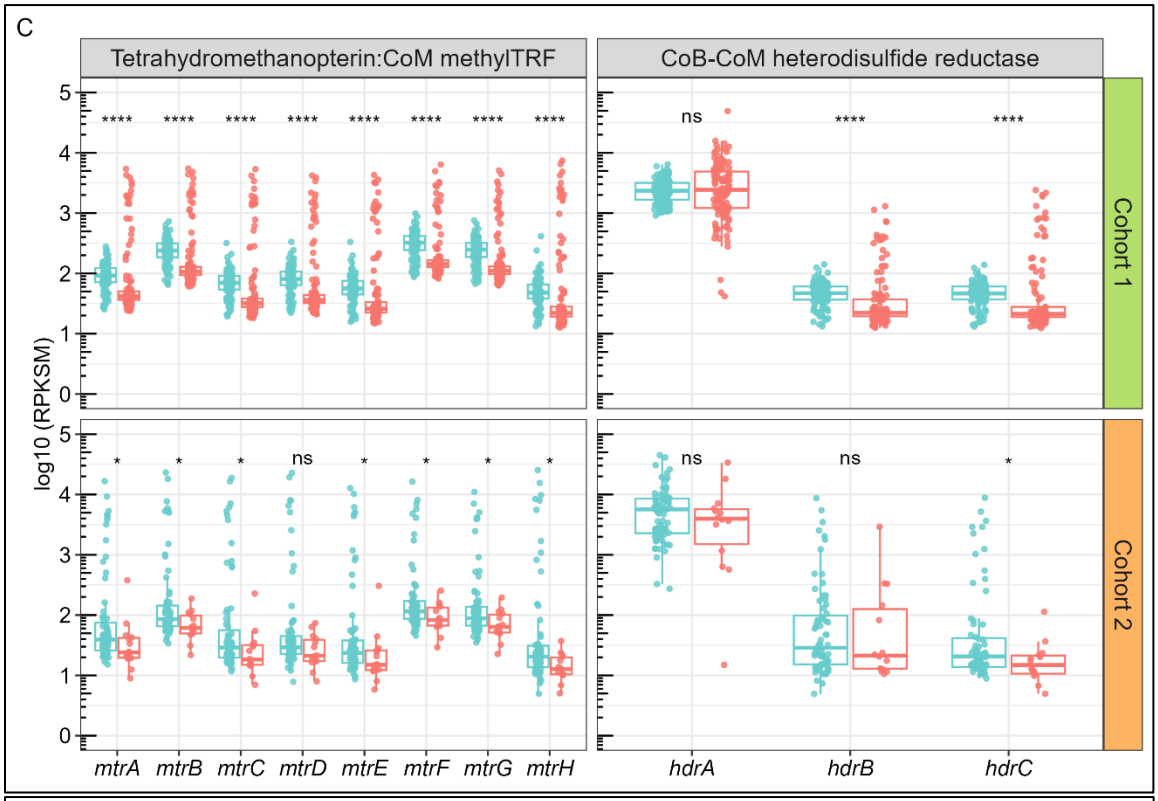
268 We measured the abundance of these genes in the samples of the three MASLD cohorts as
269 described above. Genes encoding MCR I and II, which catalyze the reduction of methyl-CoM to
270 methane, were significantly decreased in the GM of MASLD patients. Their abundance also declined
271 with MASLD progression (Fig. 4B). Concordantly, *hdr* and *mtr* genes, responsible for replenishing CoM
272 and transferring it to a methyl group to form the methyl-CoM precursor, were significantly decreased
273 in MASLD and showed a progressive depletion with MASLD advancement (Fig. 4C). Interestingly, *hdrA*
274 was two orders of magnitude more abundant than the rest of the genes from the operon. Genes *mtbA,*
275 *mttB, mtaA* and *mtaB* were significantly decreased in MASLD. However, they did not exhibit differences
276 between the initial and advanced stages of MASLD (Supplementary Fig. 2). Notably, the higher
277 abundance of methane-producing genes observed in the healthy group was not replicated in Cohort 3.
278 In contrast, bacterial genes involved in TMA regeneration showed the opposite trend. The *torCAD* and
279 *torZY* operons were significantly increased in the GM of MASLD patients, and their abundance increased
280 with MASLD progression (Fig. 4D, Supplementary Fig. 4B). These findings suggest that methane
281 production from TMA and methanol, which typically occurs in a healthy state, is disrupted in MASLD,
282 revealing another potential microbial pathway involved in MASLD pathogenesis.

283 Finally, we quantified the abundance of five universal single-copy genes (USCGs), as described
284 in Material and Methods, to confirm that observed differences in genic abundance between sample
285 groups were truly associated with the MASLD phenotype and not confounded by biological variability
286 or spurious technical artifacts related to the quantification method. No significant differences were

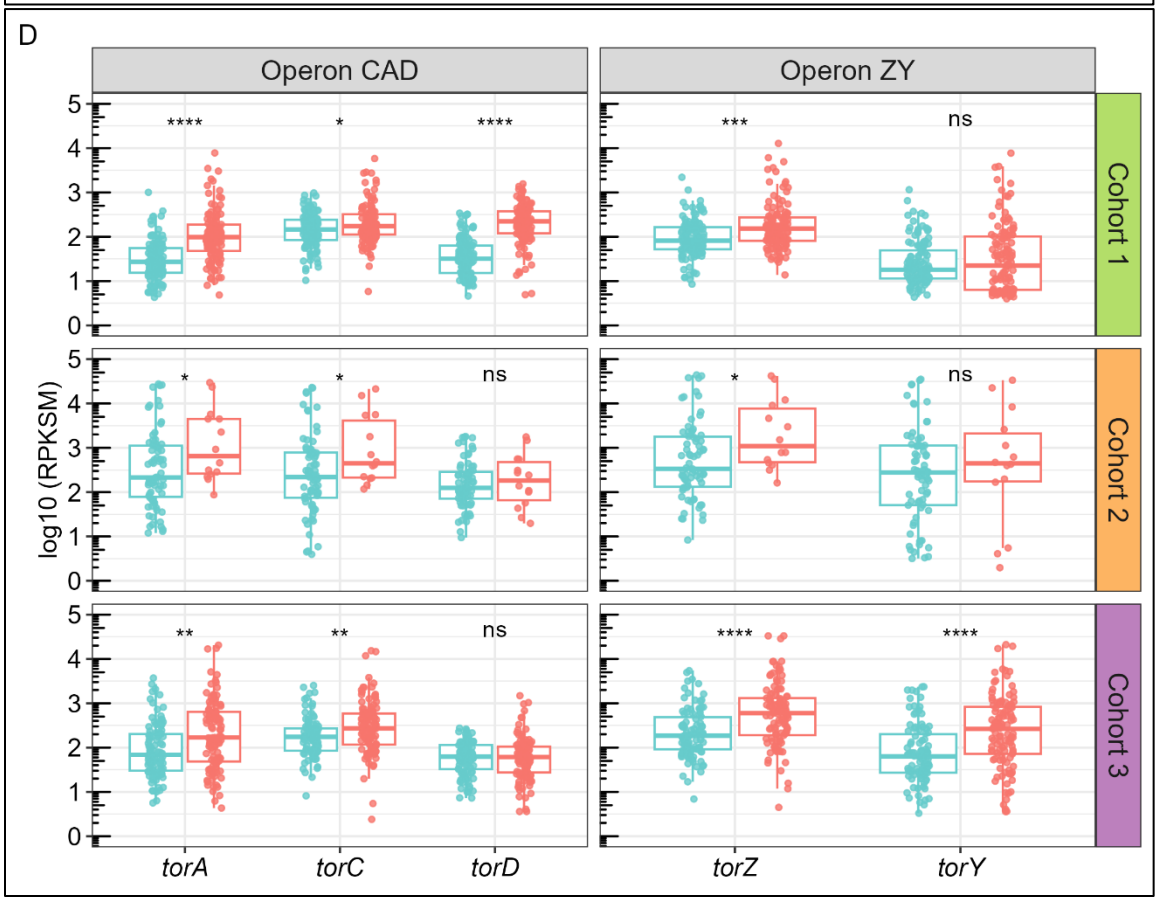
found in gene abundance between groups for any USCG in any of the three cohorts (Supplementary Fig. 1).






C
Tetrahydromethanopterin:CoM methylTRF
CoB-CoM heterodisulfide reductase
Cohort 1
Cohort 2
log10 (RPKSM)
mtrA
mtrB
mtrC
mtrD
mtrE
mtrF
mtrG
mtrH
hdrA
hdrB
hdrC

291


D
Operon CAD
Operon ZY
Cohort 1
Cohort 2
Cohort 3
log10 (RPKSM)
torA
torC
torD
torZ
torY

292
293

294 **Figure 4. Abundance of methane and TMA-producing genes in patient cohorts. (A)** Characterization of the terminal
295 reactions involved in the formation of microbial methane. Enzymes, EC numbers, coding genes and KOs, as well as reaction
296 types, are indicated for each metabolic step according to the legend annexed to Fig. 2A. **(B)** Abundance of genes encoding
297 methyl-coM reductase I (*mcr*) and methyl-CoM reductase II (*mrt*). **(C)** Abundance of genes encoding CoB-CoM
298 heterodisulfide reductase (*hdr*) and tetrahydromethanopterin S-methyltransferase (*mtr*). **(D)** Abundance of *tor* operons.
299 Facets represent genic operons. Boxplot organization and statistical tests were conducted as in Fig. 2C.

300 **Candidate metabolic genes are accessory in the GM**

301 Although this analysis identifies *A. rectalis* and multiple metabolic genes involved in the
302 formation of butyrate, SCAs, TMAO and methane as consistent taxonomic and functional signatures
303 associated with MASLD, these markers alone are not sufficiently robust to establish a causal link
304 between GM composition and MASLD phenotype. GM bacterial species have pleomorphic genomes,
305 consisting of up to thousands of genomically distinct strains that share a conserved genetic core but
306 possess an accessory genome that is highly variable among them[81,82]. This accessory genome is often
307 encoded in plasmids and other MGEs which are subject to frequent changes and can readily transfer
308 between different bacterial strains[16,83].

309 To test this hypothesis, we assessed whether candidate metabolic genes involved in MASLD are
310 core or accessory across some of the most abundant human GM genomes. Genome selection and genic
311 presence were determined as described in Materials and Methods. Briefly, the presence of a target
312 gene in a clade was defined as the percentage of strains encoding it, classifying it as core if present in
313 >80% of congeneric or conspecific genomes, as accessory if found in 20-80%, and as highly accessory if
314 present in <20% genomes.

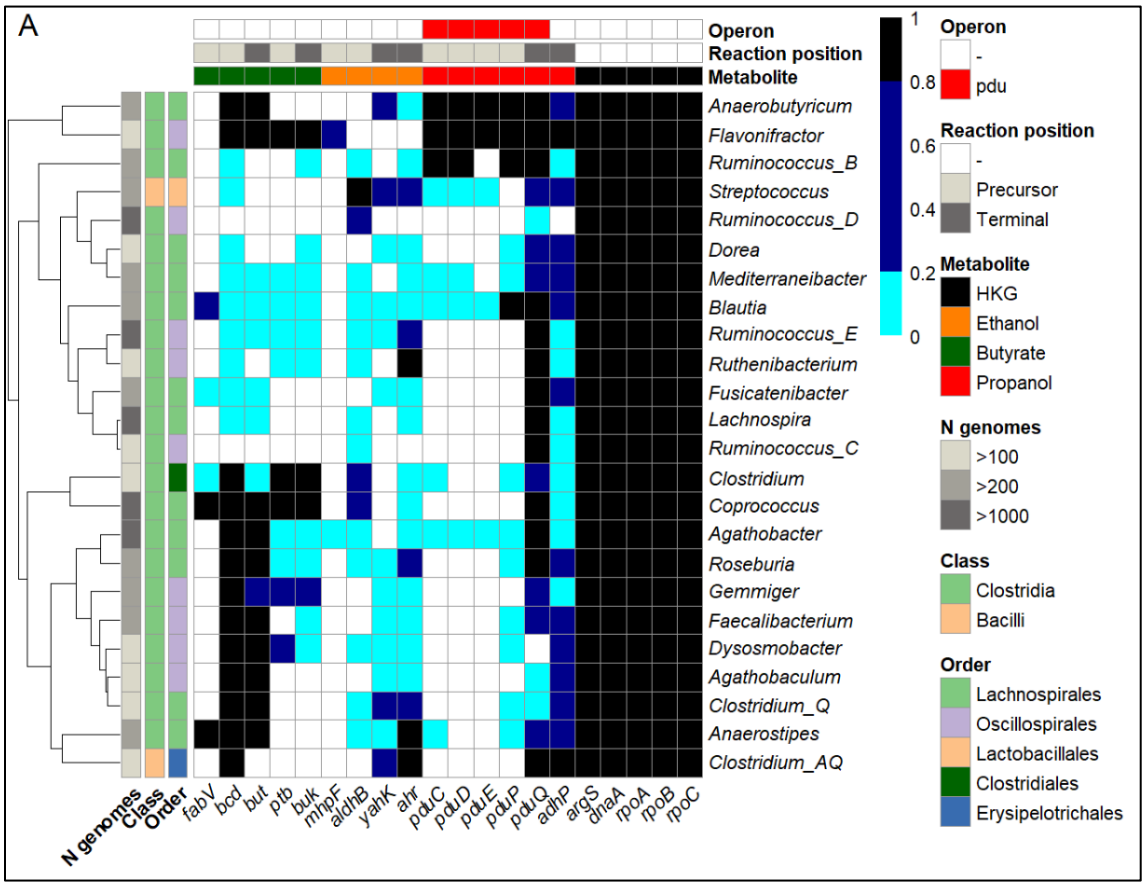
315 Most of the selected GM genera (n=24) belong to class Clostridia, specifically to orders
316 Lachnospirales and Oscillospirales (Fig. 5A). Butyrate-producing genes from the butyryl-CoA pathway
317 (*bcd* and *but*) were core in nearly half of these genera, while those from the butyryl-P pathway (*ptb* and
318 *buk*) were core only in *Flavonifractor, Clostridium* and *Coprococcus.* Notably, butyrate-producing genes

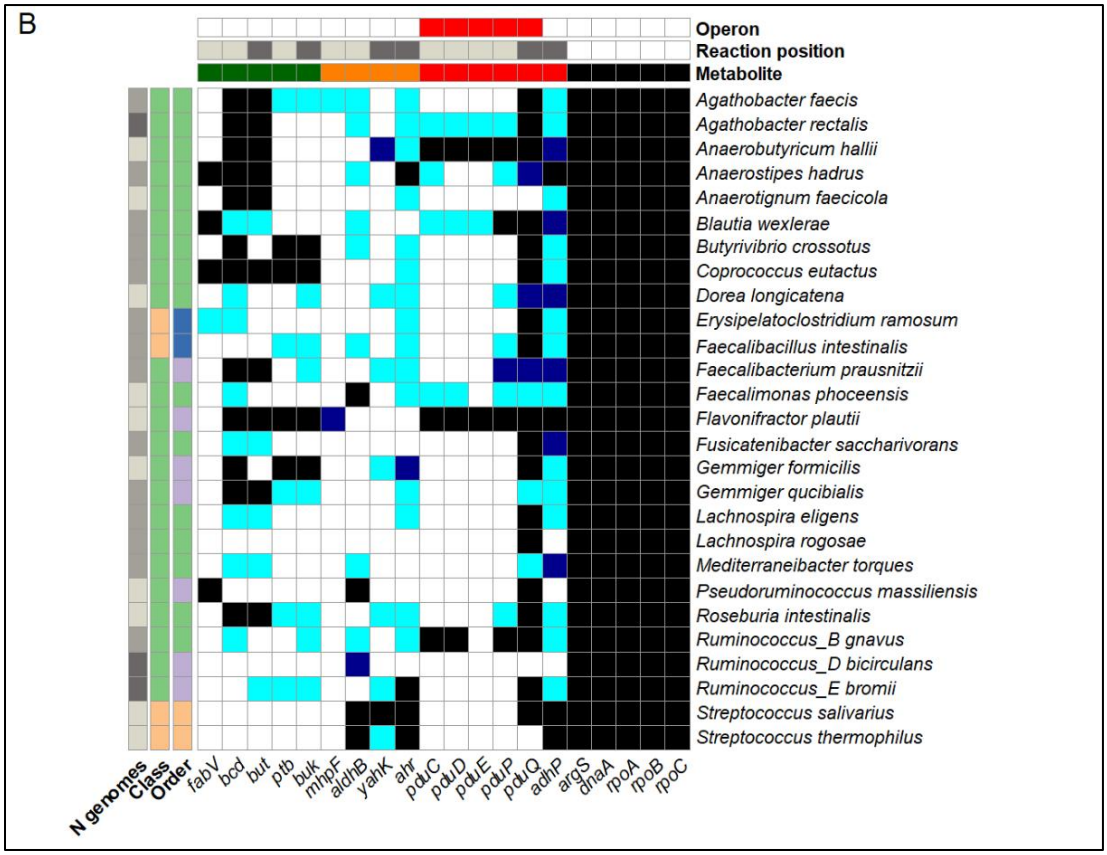
319 were highly accessory across several genera, including *Lachnospira, Ruminococcus_B* and *E,*
320 *Ruthenibacterium, Blautia*, *Fusicatenibacter*, *Dorea,* and *Mediterraneibacter*, and were completely
321 absent in *Ruminococcus_C* and *D*. In contrast, acetaldehyde and ethanol-producing genes were highly
322 accessory across all genera except *Streptococcus*, an heterolactic genus that constitutively encoded
323 *aldhB* in all its genomes (specifically in *S. salivarius* and *S. thermophilus,* Fig. 5B) and exhibited
324 moderated presence (20-80%) of *yahK*/*ahr*. The acetaldehyde-producing gene *mhpF* was nearly absent
325 in the GM, except for a few genomes of *Flavonifractor* and *Agathobacter*. Propionaldehyde and
326 propanol-producing genes from the *pdu* operon were core in three genera: *Ruminococcus_B* (*R_B.*
327 *gnavus*), *Anaerobutyricum* (*A. hallii*) and *Flavonifractor* (*F. plautii*), though *pduE* was unexpectedly
328 absent in *Ruminococcus_B*. These genes were highly accessory in *Blautia* (*B. wexlerae*), *Streptococcus*,
329 *Medierraneibacter* or *Agathobacter* (*A. rectalis*). Notably, *A. hallii* and *F. plautii* were the only species
330 encoding both core butyrate and propanol-producing genes.

331 Note that the nomenclature used in Fig. 5 follows the Genome Taxonomy Database standards,
332 which redefine taxonomic ranks based on genome-wide phylogenetic analyses[84]. As a result, traditional
333 genera such as *Ruminococcus* and *Clostridium*, which were found to be polyphyletic (i.e., composed of
334 species that do not share a single common ancestor) are split into multiple monophyletic clades (e.g.,
335 *Ruminococcus_A, Ruminococcus_B* or *ClostridiumQ*).

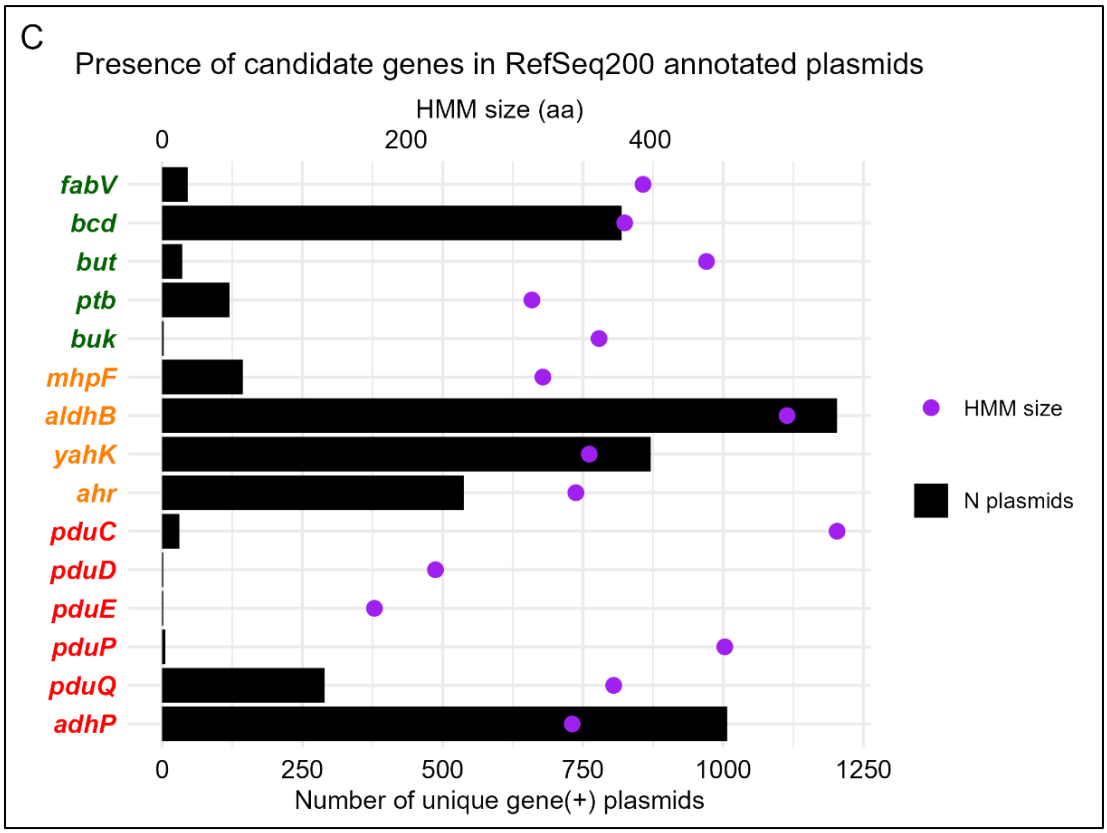
336 Beyond their accessory nature within GM pangenomes, candidate metabolic genes were also
337 encoded on plasmids (Fig. 5C). Genes involved in ethanol formation such as *aldhB, yahK* and *ahr* were
338 among the most widely distributed, being present in hundreds of different plasmids. In contrast,
339 butyrate-producing genes (*fabV, but, ptb,* and *buk*) and propanol-related genes (*pduC, pduD, pduE,*
340 *pduP,* and *pduQ*) were detected in fewer plasmids, further supporting their variable presence across

GM genomes. A notable exception was *bcd*, which exhibited a broader presence across both chromosomal and plasmid-encoded sequences. These findings reinforce the idea that non-essential metabolic functions, especially ethanol production, are encoded in MGEs, contributing to their accessory nature within the GM.




B
Operon
Reaction position
Metabolite
Agathobacter faecis
Agathobacter rectalis
Anaerobutyricum hallii
Anaerostipes hadrus
Anaerotignum faecicola
Blautia wexlerae
Butyrivibrio crossotus
Coprococcus eutactus
Dorea longicatena
Erysipelatoclostridium ramosum
Faecalibacillus intestinalis
Faecalibacterium prausnitzii
Faecalimonas phoceensis
Flavonifractor plautii
Fusicatenibacter saccharivorans
Gemmiger formicilis
Gemmiger qucibialis
Lachnospira eligens
Lachnospira rogosae
Mediterraneibacter torques
Pseudoruminococcus massiliensis
Roseburia intestinalis
Ruminococcus_B gnavus
Ruminococcus_D bicirculans
Ruminococcus_E bromii
Streptococcus salivarius
Streptococcus thermophilus
N genomes
Class
Order
fabV
bcd
but
ptb
buk
mhpF
aldhB
yahK
ahr
pduC
pduD
pduE
pduP
pduQ
adhP
argS
dnaA
rpoA
rpoB
rpoC

346


C
Presence of candidate genes in RefSeq200 annotated plasmids
HMM size (aa)
0
200
400
fabV
bcd
but
ptb
buk
mhpF
aldhB
yahK
ahr
pduC
pduD
pduE
pduP
pduQ
adhP
HMM size
N plasmids
0
250
500
750
1000
1250
Number of unique gene(+) plasmids

347

348 **Figure 5. Presence profile of candidate metabolic genes in the human GM. (A-B)** Heatmap showing the presence of
349 candidate genes involved in MASLD across some of the most abundant clades in the human GM. The x-axis represents genes
350 associated with butyrate and SCAs production, as well as USCGs. The y-axis lists the bacterial clades inspected. Gene
351 presence is expressed as the percentage of **(A)** congeneric and **(B)** conspecific genomes encoding each target gene, with
352 colors indicating classification: black for core genes (>80% of genomes), dark blue for accessory genes (20-80% genomes),
353 light blue for highly accessory genes (<20% genomes) and white for genes completely absent in the clade. Hierarchical
354 clustering was applied to the y-axis to group genera (top) based on gene-presence similarity. **(C)** Presence of candidate genes
355 in RefSeq200 annotated plasmids. Black bars represent the number of different plasmids that encode at least one copy of
356 each gene.

357 **3. Discussion**

358 MASLD is a liver pathology influenced by multiple contributing factors, and the GM has been
359 long suspected to play a role in its progression. Numerous studies have tried to ascribe a causality
360 between GM composition and MASLD, but taxonomic signatures in key GM clades, such as *Blautia* and
361 *Roseburia spp.*, have been inconsistently associated with MASLD progression[23]. These issues may be
362 linked to the use of 16S rRNA-based profiling in metataxonomic studies, which frequently lacks
363 resolution at species and strain level[85].

364 However, the heterogeneity concomitant to bacterial genomes poses another possible cause
365 for the lack of association between disease progression and the taxonomic distribution. Bacterial
366 genomes are highly plastic, and strains from the same species often harbor different genes, many of
367 which are involved in different metabolic pathways[81]. Fermentative pathways in lactobacilli, for
368 example, are often strain-specific[86], and the products of the enterobacterial mixed acid fermentation
369 also exhibit wide variability between strains[87]. If MASLD progression is linked to the metabolic functions
370 of the GM, then the association between the disease and the taxonomic profile may be obscured by
371 the inherent metabolic diversity of bacterial species. To study these potential alternatives, we followed
372 two distinct strategies.

373 To identify species associated with MASLD and overcome the limitations of 16S rRNA analysis,
374 we utilized multiple marker genes for taxonomic classification, in an attempt to provide a more

375 comprehensive view of microbial abundances in health and disease states. Using MetaPhlan, we
376 profiled the taxonomic composition of the GM in metagenomic fecal samples of healthy and MASLD
377 donors across three independent cohorts. We identified 28 bacterial genera shared across all cohorts
378 (Fig. 1A), collectively accounting for 30-60% of the total GM abundance (Fig. 1B). Among these, five
379 candidate genera -*Gemmiger*, *Streptococcus*, *Mediterraneibacter*, *Ruminococcus* and *Agathobacter*
380 (Fig. 1C)- and two species, *Agathobacter rectalis* and *Bacteroides uniformis* (Fig. 1E), exhibited
381 consistent differences in their relative abundance between groups across all the cohorts. These taxa
382 rank among the most abundant clades in the GM (Fig. 1C, Supplementary Fig. 1). However, only
383 *Agathobacter rectalis* showed statistically significant differences. Given its high abundance in the GM,
384 its consistent abundance trend at both genus and species level, and its significant depletion in MASLD,
385 *A. rectalis* emerges as a robust taxonomic marker for the disease.

386 Although this observation has been previously reported[18,21], this result should be interpreted
387 with caution due to ongoing debates in bacterial nomenclature, as *A. rectalis* is a newer designation
388 that has replaced *Eubacterium rectale*[88]. Updates in taxonomic assignments pose challenges, as newer
389 species or genera may not be consistently reflected across different databases or studies[84,89].
390 Moreover, taxonomic classifications themselves, whether based solely on 16S rRNA or on multiple
391 markers, limit our ability to reliably infer GM changes. Bacterial genome plasticity, driven by MGEs, can
392 blur the lines of taxonomy-phenotype associations, as they contribute to a vast strain-level diversity[81]
393 that remains undetected by taxonomic profiling methods. In fact, we have previously reported that
394 geno-metabolic shifts in GM propionate production associated with inflammatory bowel disease do not
395 align with species-level variations[90].

396 GM associated metabolism contributes to MASLD by disrupting the integrity of the GLA. For this
397 reason, we focused on identifying abundance differences in the genes responsible for producing
398 metabolites potentially implicated in MASLD across several patient cohorts, i.e. butyrate, SCAs, TMA
399 and methane. We isolated the GM gene families encoding the enzymes involved in the biosynthesis of
400 these metabolites and aligned them against the metagenomic samples to quantify their abundance.
401 We found that butyrate-producing genes are depleted in the GM of MASLD patients (Fig. 2B),
402 particularly those involved in the crotonyl-butyryl CoA axis: *fabV* (butyryl-CoA oxidoreductase), *bcd*
403 (butyryl-CoA dehydrogenase) and *but* (butyryl-CoA:acetate-CoA transferase) (Fig. 2A). While *ptb* and
404 *buk* were also higher in both healthy subcohorts, their levels did not differ significantly between initial
405 and advanced MASLD groups within Cohort 2. This suggests that CoA transfer may be the predominant
406 pathway for butyrate production in the GM, as previously reported[91]. The subtler differences in *ptb* and
407 *buk* in Cohort 2 may reflect a common MASLD phenotype across different fibrosis stages, potentially
408 influenced by the smaller sample size relative to Cohort 3. Additionally, two other enzymes -structured
409 as two-subunit systems-, can transfer CoA from different cofactors to butyrate to produce butyryl-CoA:
410 butyryl-CoA:acetoacetate-CoA transferase (encoded by *atoD* and *atoA,* EC 2.8.3.9) and acetate CoA-
411 transferase (encoded by *ctfA* and *ctfB,* EC 2.8.3.9) (Fig. 2A). These were not included in this study since
412 they catalyze the reverse reaction, consuming butyrate to form butyryl-CoA[92]. Notably, the depletion
413 of butyrate genes was not observed in MASLD patients from Cohort 1.

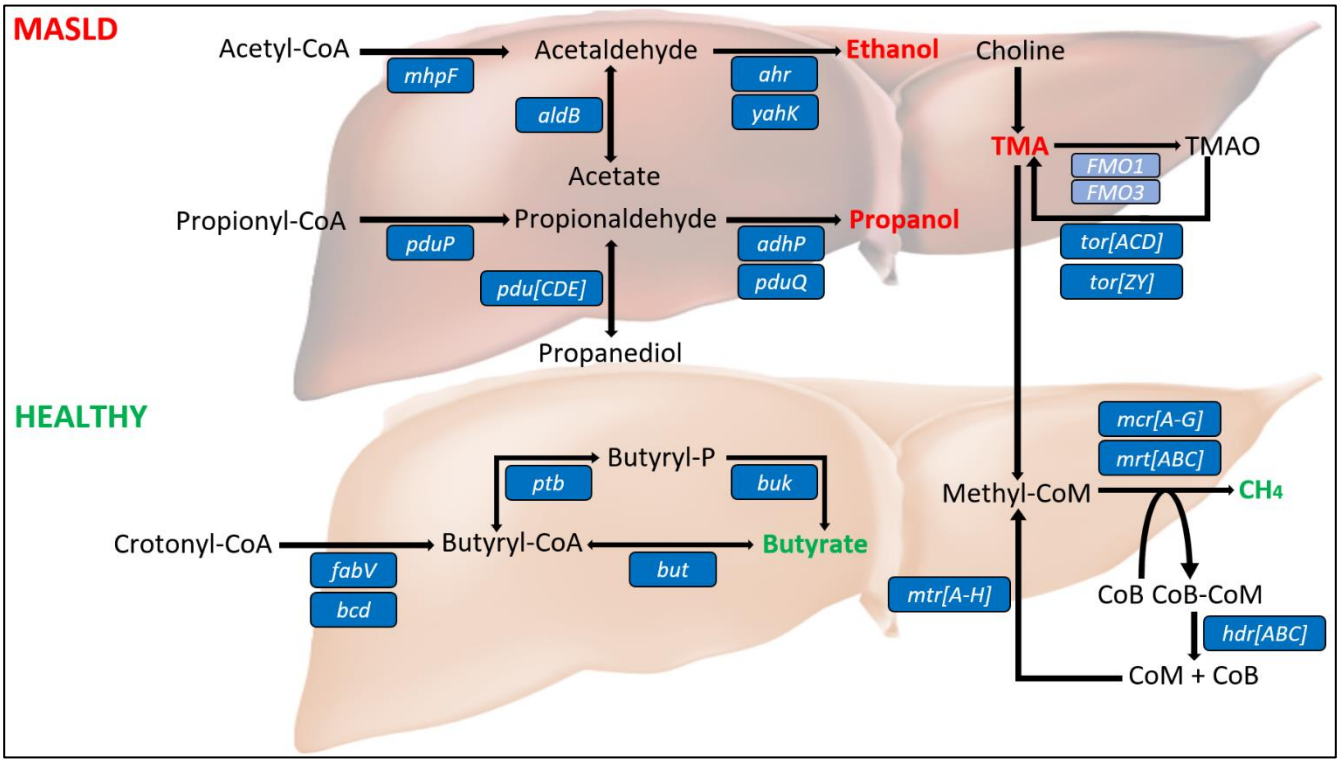
414 Contrarily, genes involved in the formation of ethanol and propanol through connected
415 fermentative reactions are generally increased in MASLD (Fig. 3). Genes *mhpF* and *aldB* -encoding two
416 aldehyde dehydrogenases that produce ethanol-, and *yahK* and *ahr* -encoding two NADPH-dependent
417 aldehyde reductases that form acetaldehyde- are significantly increased in the GM of MASLD patients

418 compared to the healthy groups (Fig. 3B). Genes involved in the production of 1-propanol structured in
419 operon *pdu* are also increased in MASLD (Fig. 3C, Supplementary Fig. 4A). Specifically, *pduCDE* -
420 encoding a propanediol dehydratase-, *pduP* -encoding a propionaldehyde dehydrogenase- and *pduQ*
421 and *adhP* -encoding two distinct propanol dehydrogenases-. The pronounced differential abundance of
422 these propanol-producing genes, organized in the *pdu* operon and *adhP*, which have been largely
423 unexplored in the context of liver disease, suggests that 1-propanol produced in bacterial
424 microcompartments may play a more significant role than ethanol in MASLD onset and progression.

425 The abundance of genes in the TMA-methane metabolic axis is also disrupted in MASLD. Genes
426 involved in methane formation (Fig. 4A) are depleted in the GM of individuals with the pathology. This
427 reduction includes both MCR I and II coding-genes (Fig. 4B), which reduce methyl-CoM to methane and
428 *mtr*, *hdr* (Fig. 4C), and Mtb / Mta-coding genes (Supplementary Fig. 2), dedicated to regenerating
429 methyl-CoM. In contrast, TMAO-reductive genes, organized in *torCAD* and *torZY* operons and
430 responsible for regenerating TMA, are increased in MASLD (Fig. 3D, Supplementary Fig. 4B). These
431 results suggest that methane production is impaired in MASLD. The opposing trend in the abundance
432 of *tor* operons further indicates that this impairment stems from a metabolic bottleneck that leads to
433 TMA accumulation instead of its conversion to methyl-CoM, effect that is driven by TMAO reductases I
434 and II.

435 Taken together, these results reveal a geno-metabolic alteration associated with MASLD,
436 characterized by an increase in genes involved in the production of SCAs and TMA, and a concurrent
437 depletion of genes responsible for butyrate and methane production (Fig. 6). We confirmed that
438 differences in gene abundance were driven by the MASLD phenotype by showing no differences in
439 USCG abundance between groups (Supplementary Fig. 3). These findings support the hypothesis that

440 the accumulation of endogenous alcohols may represent a key metabolic feature of MASLD, occurring
441 alongside the displacement of butyrate-producing taxa, as previously proposed[33]. Moreover, the
442 observed decrease in *A. rectalis* abundance partially explains the reduction in butyrate levels, as this
443 organism is a known GM butyrate producer[88].



444

445 **Figure 6. Geno-metabolic scheme involved in MASLD.** Proposed scheme showing the GM genic alterations associated with
446 butyrate, SCAs, TMA and methane production in MASLD. Genes involved in the production of ethanol, propanol and TMA
447 are increased in the pathology (top), whereas those associated with butyrate and methane are decreased (bottom). Blue
448 boxes indicate the catalyzing genes whose abundance is altered in each metabolic step.

449 Notably, results obtained at the single-gene level using curated gene families were partially
450 reproduced with the whole-genome approach applied to random sub-cohorts, particularly for genes
451 involved in the TMA-methane metabolic axis and propanol production (Supplementary Fig. 4). This
452 partial overlap supports the accuracy of the single-gene analysis. However, the lack of complete
453 concordance may be attributed either to the reduced sample size in the co-assemblies or to mismatches

454 between our custom gene families and the broader Kegg Orthology (KO) annotations required to detect
455 the same functional signals.

456 Importantly, we found that metabolic genes associated with MASLD exhibit a highly accessory
457 profile within the GM (Fig. 5). Butyrate-producing genes are highly accessory in nearly half of the most
458 abundant GM genera, while SCA-producing genes display an even more pronounced accessory pattern
459 (Fig. 5A). Among the two main butyrate biosynthesis routes, genes from the butyryl-CoA pathway are
460 more pervasive in the GM than those from the butyryl-P pathway. Notably, *Phocaeicola* and *Bacteroides*
461 -the two most abundant GM genera- were excluded from this analysis due to a limited number of high-
462 quality genomes passing our completeness filters, despite their extensive genomic representation in
463 the UHGG. Interestingly, the two key genes in the butyryl-CoA pathway (*bcd* and *but*) are core in *A.*
464 *rectalis*. Nonetheless, some strains of this known butyrate-producing species also harbor genes for
465 ethanol and propanol production, which our findings suggest may contribute to MASLD onset. This
466 highlights a critical point: taxonomic signatures alone are insufficient for associating bacterial clades
467 with disease. Genome plasticity introduces substantial strain-level metabolic variability in the GM,
468 leading to distinct phenotypic traits even within the same species: an effect well-documented in both
469 pathogenic and commensal *E. coli* strains[93,94].

470 Further supporting their mobility, candidate genes are frequently encoded in plasmids (Fig. 5C).
471 Among these, ethanol formation genes (*aldhB*, *yahK*, and *ahr*) are the most widely distributed, while
472 butyrate- and propanol-related genes are found in plasmids less frequently. Since non-essential,
473 adaptive metabolic pathways are often encoded in plasmids across many bacterial species[94–96], their
474 presence likely drives phenotypic differences in MASLD without altering the overall species composition
475 detected by conventional metataxonomic marker genes. If MASLD is actually driven by bacterial

476 metabolism, strains from different species, but harboring similar metabolic pathways, may produce
477 similar liver-impacting metabolites, potentially contributing to MASLD pathogenesis regardless of their
478 specific taxonomic position.

479 **4. Conclusions**

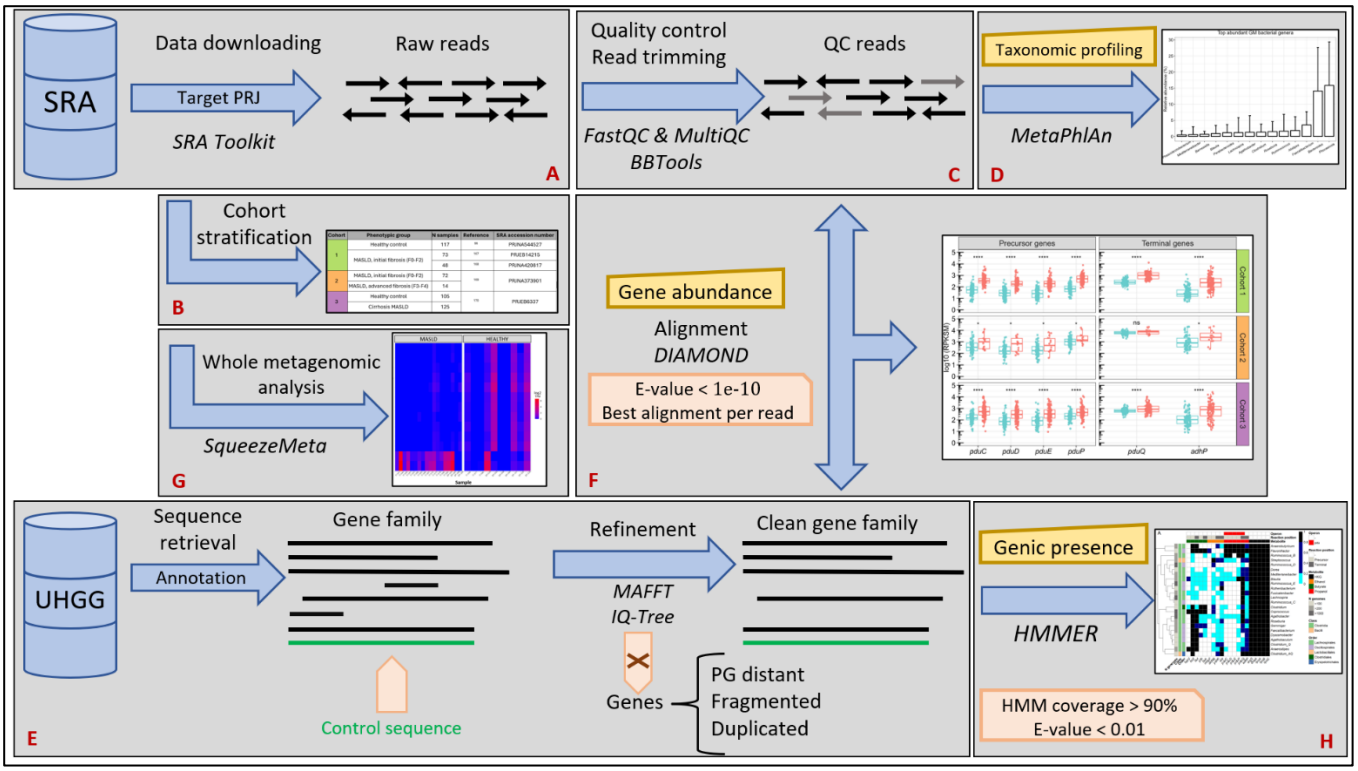
480 Our study reveals that MASLD is marked by distinct geno-metabolic alterations in the GM that
481 are not captured by conventional taxonomic profiling. Specifically, we observed an increase in genes
482 involved in the production of SCAs and TMA, coupled with a depletion of genes responsible for butyrate
483 and methane production. These results confirm multiple previous findings and suggest a broader
484 metabolic shift that may contribute to the pathogenesis of MASLD, with the displacement of beneficial
485 butyrate-producing taxa such as *Agathobacter rectalis*. Importantly, our work underscores the value of
486 analyzing metagenomes at gene level to extract additional phenotypic signatures beyond what 16S
487 rRNA-based or marker gene-based approaches can offer. By integrating metabolic pathway analysis
488 with the isolation of specific gene families, and by examining gene abundance, accessory gene profiles
489 and plasmid-mediated gene mobilization, we have significantly advanced our understanding of
490 bacterial metabolism in MASLD. This comprehensive approach not only enhances our understanding of
491 the microbial etiology of MASLD, but also paves the way for improved bacterial analysis in clinical
492 settings, offering new avenues for diagnosis and therapeutic intervention in liver pathologies.

493

494

495

496 **Materials and Methods**



497

498 **Figure 7. Overview of the methodology used in this study.** This diagram summarizes the complete computational workflow
499 for processing whole paired-end metagenomic samples, including data retrieval from the Short Read Archive (SRA), read
500 quality control and taxonomic profiling. Target gene families were isolated and their abundance quantified by alignment
501 against the sequencing libraries. Additionally, the presence of target genes was examined across GM genomes and
502 annotated plasmids was inspected. Blue arrows indicate the progression between workflow steps, with the corresponding
503 software shown in black beneath each arrow. Filtering steps are highlighted in pink boxes. Red-labeled references indicate
504 the corresponding subsection in Materials and Methods (A-H), where each step is explained in detail.

505

506 ***A. Downloading of sequencing metagenomic data***

507 A total of 554 metagenomic sequencing libraries generated from the fecal samples of MASLD
508 and healthy subjects were downloaded from the SRA with SRA Toolkit´s *fastq-dump*[97]. These samples
509 were retrieved using the accession numbers corresponding to five different studies, and they were
510 stratified in three MASLD cohorts, as detailed in Table 1. Information on the collection and processing
511 of fecal samples, bacterial DNA extraction and sequencing protocols can be found in the original studies.

512 Additionally, we retrieved the medical metadata associated with each sample donor to confirm their
513 MASLD status and classify samples into two phenotypic groups of comparison within each cohort.
514 Samples lacking a clearly defined MASLD status were excluded from this study. All data was downloaded
515 between March 2022 and March 2024

Cohort	Phenotypic group	N samples	Reference	SRA accession number
1	Healthy control	117	98	PRJNA544527
	MASLD, initial fibrosis (F0-F2)	73	99	PRJEB14215
		48	100	PRJNA420817
2	MASLD, initial fibrosis (F0-F2)	72	18	PRJNA373901
	MASLD, advanced fibrosis (F3-F4)	14		
3	Healthy control	105	101	PRJEB6337
	Cirrhosis MASLD	125		

516 **Table 1: MASLD patient cohorts analyzed in this study.**

517 Each cohort (column 1) was divided into two phenotypic comparison groups: control vs. disease (column 2). The cohorts
518 include *N* MASLD patients and corresponding whole-metagenomic stool samples derived from them (column 3). The disease
519 group in Cohort 1 consists of 73[99] and 48[100] MASLD patients with initial fibrosis. Cohort 2 was divided into initial fibrosis
520 (control) and advanced fibrosis groups. Patients were originally recruited in the listed studies (column 4), with metagenomic
521 data and medical metadata downloaded from the indicated SRA accession numbers (column 5).

522

523 ***B. Description of MASLD patient cohorts***

524 Cohort 1 consists of two sets of European, obese MASLD patients aged 20-64 years with initial
525 fibrosis (F0-F2) and varying degrees of steatosis: 73 patients from Hoyles *et al.*[99] and 10 patients from
526 Mardinoglu *et al.*[100]. In the latter group, 48 fecal samples were collected at five time points during the
527 study (two samples from day 1 were not obtained). Diagnoses were confirmed through hepatic biopsy
528 performed during obesity surgery, and none of the subjects had alcoholic liver disease, hepatitis or
529 diabetes. The control group for comparison includes 117 healthy patients from Poyet *et al.*[98]. Cohort 2

530 comprises 86 American MASLD patients from Loomba *et al.*[18], including 72 with initial (F0-F2) and 14
531 with advanced fibrosis (F3-F4). Diagnoses were confirmed through hepatic biopsy and magnetic
532 resonance imaging, and none of the patients had alcoholic liver disease, hepatitis or diabetes. Note that
533 in this study, the initial fibrosis group is referred to as "healthy" for group comparison purposes, despite
534 including patients with early-stage MASLD. Cohort 3 includes 230 Han Chinese patients from Qin *et*
535 *al.*[101]: 125 diagnosed with cirrhotic MASLD and 105 with no hepatic pathology. None of these subjects
536 had hepatitis. All cohorts consist of one metagenomic stool sample per patient, except for the 48
537 samples collected at multiple time points from 10 patients in the study from Mardinoglu *et al.*[100].

538 ***C. Quality control of metagenomic reads***

539 Sequencing reads were trimmed to remove Illumina adapter remnants and low-quality regions
540 (<Q25) using BBDuk from the BBTools suite (version 37.62)[102]. Read quality was assessed with FastQC
541 (version 0.12.1)[103] and summarized using MultiQC (version 1.22.3)[104].

542 ***D. Taxonomic signatures associated with MASLD***

543 To determine whether any bacterial clade is systematically increased or decreased in MASLD
544 across the three cohorts, we profiled the taxonomic composition of GM communities in fecal samples
545 using MetaPhlAn (version 4.0.6), which uses a wide-range of clade-specific maker genes to estimate
546 taxonomic abundances[24]. The diversity of these markers allows MetaPhlAn to provide highly accurate
547 taxonomic classification across a wide spectrum of bacterial clades, from the species level to broader
548 taxonomic ranks, without being limited to 16S rRNA alone.

549 For this study, we used the database version *mpa_vOct22_CHOCOPhlAnSGB_202212*,
550 downloaded on August 22, 2023, and added the *--ignore_eukaryotes* flag to skip profiling eukaryotic
551 organisms. Post-processing of the MetaPhlAn output involved a) removing low-abundant clades (i.e.,

552 those with an average relative abundance of 0 in at least one cohort), b) excluding taxa predicted as
553 "unclassified", and c) verifying that excluded clades were not solely present in only one of the
554 comparison groups, in order to avoid removing what could be very clear taxonomic markers. To identify
555 the most abundant clades present in the GM, we selected genera with an average relative abundance
556 above 0.5% (Fig. 1C) and species above 0.2% (Fig. 1E). Notedly, in this step we faced an issue regarding
557 the status of taxonomic assignments, as several species required manual updates to their nomenclature
558 assigned by MetaPhlAn. Specifically, *Eubacterium rectale* was reassigned to *Agathobacter rectalis*[88],
559 *Ruminococcus torques* to *Mediterraneibacter torques*[105] and *Roseburia faecis* to *Agathobacter faecis*[106].

560 ***E. Isolation of target gene families in the GM***

561 We constructed families of the human GM metabolic genes encoding the enzymes involved in the
562 production of butyrate, ethanol, propanol, TMA, methane and their precursor metabolites, as
563 represented in Figs. 2A, 3A and 4A. The Unified Human Gastrointestinal Genome (UHGG, version
564 2.0.2)[81] served as the reference database for this purpose. This resource, comprising 289,232 bacterial
565 and archaeal genomes and more than 665 million proteins, represents the most comprehensive
566 collection of characterized human GM genes to date. Each gene family was built by retrieving a seed
567 set of sequences based on gene and protein annotation from the UHGG, assessing their similarity with
568 lab-validated enzymatic sequences when available. Validated gene sequences were retrieved from
569 MetaCyc[107] and used as controls to assess their phylogenetic proximity to the retrieved gene seeds.

570 Gene families were aligned using MAFFT (version 7.271)[108] with options *"--localpair"* and *"--*
571 *maxiterate 1000"* and phylogenetic trees were constructed with IQ-TREE (version 2.0.3)[109] with the
572 ultrafast bootstrap option (1000 bootstraps)[110] to assess branch support. The best fitting model for
573 each metabolic gene was estimated using ModelFinder Plus[111], according to the Bayesian information

574 criterion. Low phylogenetic distances between control and retrieved sequences served as a criterion
575 for defining gene families. Duplicated sequences, gene fragments and phylogenetically distant genes
576 were excluded after inspecting the trees. To improve gene identification and classification, we then
577 built profile Hidden Markov Model (HMM) for each target gene family using *HMMER hmpress* (version
578 3.4)[112]. MetaCyc[107] and KEGG[113] were also used to validate gene and enzyme names, KO groups, the
579 biochemical reactions involved in the formation of each metabolite, and their enzyme commission (EC)
580 numbers. The same protocol was applied to isolate five USCGs used as controls. These USCGs and their
581 encoded proteins were: *argS* (arginyl-tRNA synthetase), *dnaA* (chromosomal replication initiator
582 protein DnaA), *rpoA*, *rpoB* and *rpoC* (alpha, beta and gamma subunits of the DNA-directed RNA
583 polymerase, respectively).

584 ***F. Analysis of gene abundance by read-based metagenomic quantification***

585 We quantified the abundance of target genes in the metagenomic samples from the three cohorts
586 by aligning the quality-controlled reads against the isolated gene families using DIAMOND (version
587 2.0.14)[114]. Alignments were filtered to retain only the best match for each read and gene family
588 sequence by using the *"--max-hsps 1"* option, with an E-value threshold of <1E-10. These parameters
589 were based on experimentally validated recommendations for detecting genes in fecal metagenomic
590 samples with DIAMOND[115]. The total number of aligned reads for each gene family was normalized
591 using a variation of the RPKM formula, referred to as RPKSM (reads per kilobase per size per million
592 reads). This measure was used as an indicator of gene abundance (Figs. 2B, 3B-C and 4B-D; and
593 Supplementary Figs. 2 and 3). It is defined as the number of aligned reads per kilobase per gene family
594 size per million reads, to account for variations in gene length, gene family size (i.e., the number of
595 sequences within each gene family) and library size, respectively.

596 ***G. Whole metagenomic analysis***

597 We applied a whole-metagenome analysis workflow to validate our results at single-gene level. To
598 this end, SqueezeMeta (version 1.6.0)[116] was used to predict both the taxonomic and functional
599 composition of the metagenomic samples from the three cohorts. Due to computational limitations, it
600 was not feasible to co-assemble all available samples, so approximately 24-30 samples per cohort were
601 randomly selected, with equal representation from each phenotypic group. Briefly, SqueezeMeta
602 assembled the metagenomes from each sub-cohort into a single co-assembly using MEGAHIT (version
603 1.2.9)[117] by pooling the sequencing reads from every sample. ORFs were then predicted from contigs[118],
604 and reads from individual samples were mapped back against the co-assembly to quantify the
605 abundance of each KO group per metagenome. We further processed these results with SQMtools
606 (version 1.6.0)[119] to identify the most differentially abundant KOs between phenotypic groups within
607 each cohort through DESeq2 (version 1.34.0)[120], applying a log fold-change >2 and a p-adjusted value
608 <0.05 as thresholds (Supplementary Fig. 4).

609 ***H. Presence of candidate metabolic genes across the human GM genomes and annotated plasmids***

610 To determine whether candidate metabolic genes involved in MASLD were core or accessory across
611 the human GM, their presence was quantified in both UHGG genomes and curated RefSeq200
612 plasmids[121]. Curation involved eliminating sequences corresponding to partial plasmid DNA sequences,
613 unassignable hosts, or PacBio internal control sequences, as done previously[16]. Since many UHGG
614 genomes originate from metagenome-assembled or single-cell-amplified genomes, they may be
615 incomplete, potentially leading to false positive annotations. To minimize this risk, we included only
616 UHGG genomes with >95% completeness and containing the five USCGs described above. ORFs were
617 predicted using Prodigal (version 2.6.3)[122], retaining only those covering >90% of the HMM size. We

618 then used HMMER *hmmscan* with the following parameters: protein-pHMM coverage threshold >90%,
619 E-value <0.01 and independent E-value (i-E value) <0.01, strictly enforcing the 90% threshold to avoid
620 capturing ORFs with similar protein domains. This approach yielded 31,227 genomes classified as
621 complete and containing the five USCGs.

622 To classify genes as core or accessory, we defined their presence as the percentage of strains within
623 a given genus and species that encode the gene. Genes were categorized as core if present in >80% of
624 congeneric (Fig. 5A) or conspecific (Fig. 5B) genomes, as accessory if present in 20-80%, and as highly
625 accessory if present in <20% genomes. Only the most abundant clades in the GM with >100 genomes
626 meeting the completeness criteria were included. Additionally, we assessed whether these genes could
627 be encoded in annotated plasmids. Using the same constraints, we predicted ORFs from the RefSeq200
628 plasmid collection, which includes 23,309 plasmids encoding over 2 million ORFs. ORFS were retained
629 if they met the >90% HMM size threshold, with *hmmscan* parameters identical to those applied in the
630 UHGG genome analysis (Fig. 5C).

631 ***I. R packages and statistical analysis***

632 Data manipulation was performed using R (version 4.1.3) and the *tidyverse* package (version
633 1.3.2)[123]. Figures were generated with the R packages *ggplot2* (version 3.4.2)[124], *ggpubr* (version
634 0.6.0)[125] and *pheatmap* (version 1.0.12)[126]. Mann-Whitney tests with a Benjamini-Hochberg adjustment
635 for multiple testing were applied on the relative abundance of bacterial clades (Figs. 1D and 1E) and
636 gene abundance (Figs. 2B, 3B-C, 4B-D, and Supplementary Figs. 2 and 3) between phenotypic groups in
637 the metagenomic fecal samples of each cohort to assess the significance of the differences. Statistical
638 analysis and graphical representation on the plots were conducted using the R package *rstatix* (version
639 0.7.2)[127].

640 **References**

641 1. Riazi, K. *et al.* The prevalence and incidence of NAFLD worldwide: a systematic review and meta-analysis. *Lancet*
642 *Gastroenterol. Hepatol.* **7**, 851–861 (2022).
643 2. EASL. EASL–EASD–EASO Clinical practice guidelines for the management of non-alcoholic fatty liver disease. *J. Hepatol.*
644 **64**, 1388–1402 (2016).
645 3. Allen, A. M. *et al.* Clinical course of non-alcoholic fatty liver disease and the implications for clinical trial design. *J.*
646 *Hepatol.* **77**, 1237–1245 (2022).
647 4. Powell, E. E., Wong, V. W.-S. & Rinella, M. Non-alcoholic fatty liver disease. *The Lancet* **397**, 2212–2224 (2021).
648 5. Yki-Järvinen, H. Non-alcoholic fatty liver disease as a cause and a consequence of metabolic syndrome. *Lancet Diabetes*
649 *Endocrinol.* **2**, 901–910 (2014).
650 6. Martín-Mateos, R. & Albillos, A. The role of the gut-liver axis in metabolic dysfunction-associated fatty liver disease.
651 *Front. Immunol.* **12**, 660179 (2021).
652 7. Carpino, G. *et al.* Increased liver localization of lipopolysaccharides in human and experimental NAFLD. *Hepatology* **72**,
653 470–485 (2020).
654 8. Li, Q., Rempel, J. D., Yang, J. & Minuk, G. Y. The effects of pathogen-associated molecular patterns on peripheral blood
655 monocytes in patients with non-alcoholic fatty liver disease. *J. Clin. Exp. Hepatol.* **12**, 808–817 (2022).
656 9. Miele, L. *et al.* Increased intestinal permeability and tight junction alterations in non-alcoholic fatty liver disease.
657 *Hepatology* **49**, 1877–1887 (2009).
658 10. Wang, W. *et al.* Tauroursodeoxycholic acid inhibits intestinal inflammation and barrier disruption in mice with non-
659 alcoholic fatty liver disease. *Br. J. Pharmacol.* **175**, 469–484 (2018).
660 11. Kuang, J. *et al.* Hyodeoxycholic acid alleviates non-alcoholic fatty liver disease through modulating the gut-liver axis.
661 *Cell Metab.* **35**, 1752-1766.e8 (2023).
662 12. Yang, C. *et al.* Altered gut microbial profile accompanied by abnormal short chain fatty acid metabolism exacerbates
663 nonalcoholic fatty liver disease progression. *Sci. Rep.* **14**, 22385 (2024).
664 13. Zhang, R. *et al.* Gut microbial metabolites in MASLD: implications of mitochondrial dysfunction in the pathogenesis and
665 treatment. *Hepatol. Commun.* **8**, 1–21 (2024).
666 14. Del Chierico, F. *et al.* Gut microbiota profiling of pediatric non-alcoholic fatty liver disease and obese patients unveiled
667 by an integrated meta-omics-based approach. *Hepatology* **65**, 451–464 (2017).
668 15. Lee, G. *et al.* Distinct signatures of gut microbiome and metabolites associated with significant fibrosis in non-obese
669 NAFLD. *Nat. Commun.* **11**, 4982 (2020).
670 16. Redondo-Salvo, S. *et al.* Pathways for horizontal gene transfer in bacteria revealed by a global map of their plasmids.
671 *Nat. Commun.* **11**, 3602 (2020).
672 17. Lanthier, N. *et al.* Microbiota analysis and transient elastography reveal new extra-hepatic components of liver steatosis
673 and fibrosis in obese patients. *Sci. Rep.* **11**, 659 (2021).
674 18. Loomba, R. *et al.* Gut microbiome-based metagenomic signature for non-invasive detection of advanced fibrosis in
675 human non-alcoholic fatty liver disease. *Cell Metab.* **25**, 1054-1062.e5 (2017).
676 19. Boursier, J. *et al.* The severity of non-alcoholic fatty liver disease is associated with gut dysbiosis and shift in the
677 metabolic function of the gut microbiota. *Hepatology* **63**, 764–775 (2016).
678 20. Caussy, C. *et al.* A gut microbiome signature for cirrhosis due to non-alcoholic fatty liver disease. *Nat. Commun.* **10**,
679 1406 (2019).
680 21. Oh, T. G. *et al.* A universal gut-microbiome-derived signature predicts cirrhosis. *Cell Metab.* **32**, 878-888.e6 (2020).
681 22. Shen, F. *et al.* Gut microbiota dysbiosis in patients with non-alcoholic fatty liver disease. *Hepatobiliary Pancreat. Dis.*
682 *Int.* **16**, 375–381 (2017).
683 23. Aron-Wisnewsky, J. *et al.* Gut microbiota and human NAFLD: disentangling microbial signatures from metabolic
684 disorders. *Nat. Rev. Gastroenterol. Hepatol.* **17**, 279–297 (2020).

24. Blanco-Míguez, A. *et al.* Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat. Biotechnol.* **41**, 1633–1644 (2023).
25. McPherson, S. *et al.* Age as a confounding factor for the accurate non-invasive diagnosis of advanced NAFLD fibrosis. *Am. J. Gastroenterol.* **112**, 740–751 (2017).
26. Ponziani, F. R., Gasbarrini, A. & Pompili, M. NAFLD or comorbidities, that is the question. *J. Hepatol.* **73**, 723 (2020).
27. Chu, H., Duan, Y., Yang, L. & Schnabl, B. Small metabolites, possible big changes: a microbiota-centered view of non-alcoholic fatty liver disease. *Gut* **68**, 359–370 (2019).
28. Dai, X. *et al.* Microbial metabolites: critical regulators in NAFLD. *Front. Microbiol.* **11**, 567654 (2020).
29. Martin-Grau, M. & Monleón, D. The role of microbiota-related co-metabolites in MASLD progression: a narrative review. *Curr. Issues Mol. Biol.* **46**, 6377–6389 (2024).
30. Oliphant, K. & Allen-Vercoe, E. Macronutrient metabolism by the human gut microbiome: major fermentation by-products and their impact on host health. *Microbiome* **7**, 91 (2019).
31. Donaldson, G. P., Lee, S. M. & Mazmanian, S. K. Gut biogeography of the bacterial microbiota. *Nat. Rev. Microbiol.* **14**, 20–32 (2016).
32. Albillos, A., De Gottardi, A. & Rescigno, M. The gut-liver axis in liver disease: pathophysiological basis for therapy. *J. Hepatol.* **72**, 558–577 (2020).
33. Iruzubieta, P., Medina, J. M., Fernández-López, R., Crespo, J. & De La Cruz, F. A role for gut microbiome fermentative pathways in fatty liver disease progression. *J. Clin. Med.* **9**, 1369 (2020).
34. Miller, T. L. & Wolin, M. J. Pathways of acetate, propionate, and butyrate formation by the human fecal microbial flora. *Appl. Environ. Microbiol.* **62**, 1589–1592 (1996).
35. Koh, A., De Vadder, F., Kovatcheva-Datchary, P. & Bäckhed, F. From dietary fiber to host physiology: short-chain fatty acids as key bacterial metabolites. *Cell* **165**, 1332–1345 (2016).
36. Morrison, D. J. & Preston, T. Formation of short chain fatty acids by the gut microbiota and their impact on human metabolism. *Gut Microbes* **7**, 189–200 (2016).
37. Blaak, E. E. *et al.* Short chain fatty acids in human gut and metabolic health. *Benef. Microbes* **11**, 411–455 (2020).
38. Donohoe, D. R. *et al.* The microbiome and butyrate regulate energy metabolism and autophagy in the mammalian colon. *Cell Metab.* **13**, 517–526 (2011).
39. Wang, H.-B., Wang, P.-Y., Wang, X., Wan, Y.-L. & Liu, Y.-C. Butyrate enhances Intestinal epithelial barrier function via up-regulation of tight junction protein claudin-1 transcription. *Dig. Dis. Sci.* **57**, 3126–3135 (2012).
40. Gao, Z. *et al.* Butyrate improves insulin sensitivity and increases energy expenditure in mice. *Diabetes* **58**, 1509–1517 (2009).
41. Li, Z. *et al.* Butyrate reduces appetite and activates brown adipose tissue via the gut-brain neural circuit. *Gut* **67**, 1269–1279 (2018).
42. De Vadder, F. *et al.* Microbiota-generated metabolites promote metabolic benefits via gut-brain neural circuits. *Cell* **156**, 84–96 (2014).
43. Sanna, S. *et al.* Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. *Nat. Genet.* **51**, 600–605 (2019).
44. Endo, H., Niioka, M., Kobayashi, N., Tanaka, M. & Watanabe, T. Butyrate-producing probiotics reduce non-alcoholic fatty liver disease progression in rats: new insight into the probiotics for the gut-liver axis. *PLoS ONE* **8**, e63388 (2013).
45. Jin, C. J., Sellmann, C., Engstler, A. J., Ziegenhardt, D. & Bergheim, I. Supplementation of sodium butyrate protects mice from the development of non-alcoholic steatohepatitis (NASH). *Br. J. Nutr.* **114**, 1745–1755 (2015).
46. Zhou, D. *et al.* Sodium butyrate attenuates high-fat diet-induced steatohepatitis in mice by improving gut microbiota and gastrointestinal barrier. *World J. Gastroenterol.* **23**, 60 (2017).
47. Zhou, D. *et al.* Sodium butyrate reduces high-fat diet-induced non-alcoholic steatohepatitis through upregulation of hepatic GLP-1R expression. *Exp. Mol. Med.* **50**, 1–12 (2018).

730 48. Zhou, D. *et al.* Total fecal microbiota transplantation alleviates high-fat diet-induced steatohepatitis in mice via
731 beneficial regulation of gut microbiota. *Sci. Rep.* **7**, 1529 (2017).
732 49. Louis, P. & Flint, H. J. Formation of propionate and butyrate by the human colonic microbiota. *Environ. Microbiol.* **19**,
733 29–41 (2017).
734 50. Zhu, L. *et al.* Characterization of gut microbiomes in non-alcoholic steatohepatitis (NASH) patients: a connection
735 between endogenous alcohol and NASH. *Hepatology* **57**, 601–609 (2013).
736 51. Michail, S. *et al.* Altered gut microbial energy and metabolism in children with non-alcoholic fatty liver disease. *FEMS*
737 *Microbiol. Ecol.* **91**, 1–9 (2015).
738 52. Meijnikman, A. S. *et al.* Microbiome-derived ethanol in non-alcoholic fatty liver disease. *Nat. Med.* **28**, 2100–2106
739 (2022).
740 53. Yuan, J. *et al.* Fatty liver disease caused by high-alcohol-producing *Klebsiella pneumoniae*. *Cell Metab.* **30**, 675-688.e7
741 (2019).
742 54. Chen, X. *et al.* Endogenous ethanol produced by intestinal bacteria induces mitochondrial dysfunction in non-alcoholic
743 fatty liver disease. *J. Gastroenterol. Hepatol.* **35**, 2009–2019 (2020).
744 55. Fromenty, B. & Roden, M. Mitochondrial alterations in fatty liver diseases. *J. Hepatol.* **78**, 415–429 (2023).
745 56. Ruuskanen, M. O. *et al.* Links between gut microbiome composition and fatty liver disease in a large population sample.
746 *Gut Microbes* **13**, 1888673 (2021).
747 57. Malaguarnera, G. Gut microbiota in alcoholic liver disease: pathogenetic role and therapeutic perspectives. *World J.*
748 *Gastroenterol.* **20**, 16639 (2014).
749 58. Schut, G. J. *et al.* Tungsten enzymes play a role in detoxifying food and antimicrobial aldehydes in the human gut
750 microbiome. *Proc. Natl. Acad. Sci.* **118**, e2109008118 (2021).
751 59. Pony, P., Rapisarda, C., Terradot, L., Marza, E. & Fronzes, R. Filamentation of the bacterial bi-functional
752 alcohol/aldehyde dehydrogenase AdhE is essential for substrate channeling and enzymatic regulation. *Nat. Commun.*
753 **11**, 1426 (2020).
754 60. Hitschler, L., Nissen, L. S., Kuntz, M. & Basen, M. Alcohol dehydrogenases AdhE and AdhB with broad substrate ranges
755 are important enzymes for organic acid reduction in *Thermoanaerobacter sp.* strain X514. *Biotechnol. Biofuels* **14**, 187
756 (2021).
757 61. Pick, A., Rühmann, B., Schmid, J. & Sieber, V. Novel CAD-like enzymes from *Escherichia coli* K-12 as additional tools in
758 chemical production. *Appl. Microbiol. Biotechnol.* **97**, 5815–5824 (2013).
759 62. Merchel Piovesan Pereira, B., Adil Salim, M., Rai, N. & Tagkopoulos, I. Tolerance to glutaraldehyde in *Escherichia coli*
760 mediated by overexpression of the aldehyde reductase YqhD by YqhC. *Front. Microbiol.* **12**, 680553 (2021).
761 63. Moreira De Gouveia, M. I., Daniel, J., Garrivier, A., Bernalier-Donadille, A. & Jubelin, G. Diversity of ethanolamine
762 utilization by human commensal *Escherichia coli*. *Res. Microbiol.* **174**, 103989 (2023).
763 64. Dellomonaco, C., Clomburg, J. M., Miller, E. N. & Gonzalez, R. Engineered reversal of the β-oxidation cycle for the
764 synthesis of fuels and chemicals. *Nature* **476**, 355–359 (2011).
765 65. Eram, M. & Ma, K. Decarboxylation of pyruvate to acetaldehyde for ethanol production by hyperthermophiles.
766 *Biomolecules* **3**, 578–596 (2013).
767 66. Ho, K. K. & Weiner, H. Isolation and characterization of an aldehyde dehydrogenase encoded by the *aldB* gene of
768 *Escherichia coli*. *J. Bacteriol.* **187**, 1067–1073 (2005).
769 67. Ravcheev, D. A., Moussu, L., Smajic, S. & Thiele, I. Comparative genomic analysis reveals novel microcompartment-
770 associated metabolic pathways in the human gut microbiome. *Front. Genet.* **10**, 636 (2019).
771 68. Dank, A. *et al.* Bacterial microcompartment-dependent 1,2-propanediol utilization of *Propionibacterium freudenreichii*.
772 *Front. Microbiol.* **12**, 679827 (2021).
773 69. Chen, Y. *et al.* Associations of gut-flora-dependent metabolite trimethylamine-N-oxide, betaine and choline with non-
774 alcoholic fatty liver disease in adults. *Sci. Rep.* **6**, 19076 (2016).

775 70. Flores-Guerrero, J. L. *et al.* Circulating trimethylamine-*N*-oxide is associated with all-cause mortality in subjects with
776 non-alcoholic fatty liver disease. *Liver Int.* **41**, 2371–2382 (2021).
777 71. Wang, Z. *et al.* Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. *Nature* **472**, 57–63
778 (2011).
779 72. Roncal, C. *et al.* Trimethylamine-N-oxide (TMAO) predicts cardiovascular mortality in peripheral artery disease. *Sci. Rep.*
780 **9**, 15580 (2019).
781 73. Martínez-del Campo, A. *et al.* Characterization and detection of a widely distributed gene cluster that predicts
782 anaerobic choline utilization by human gut bacteria. *mBio* **6**, e00042-15 (2015).
783 74. Goris, M. *et al.* Increased thermostability of an engineered flavin-containing monooxygenase to remediate
784 trimethylamine in fish protein hydrolysates. *Appl. Environ. Microbiol.* **89**, e00390-23 (2023).
785 75. Leimkühler, S. The biosynthesis of the molybdenum cofactors in *Escherichia coli*. *Environ. Microbiol.* **22**, 2007–2026
786 (2020).
787 76. Kurth, J. M. *et al.* Methanogenic archaea use a bacteria-like methyltransferase system to demethoxylate aromatic
788 compounds. *ISME J.* **15**, 3549–3565 (2021).
789 77. Gendron, A. & Allen, K. D. Overview of diverse methyl/alkyl-coenzyme M reductases and considerations for their
790 potential heterologous expression. *Front. Microbiol.* **13**, 867342 (2022).
791 78. Thauer, R. K. Methyl (alkyl)-coenzyme M reductases: nickel F-430-containing enzymes involved in anaerobic methane
792 formation and in anaerobic oxidation of methane or of short chain alkanes. *Biochemistry* **58**, 5198–5220 (2019).
793 79. Wagner, T., Koch, J., Ermler, U. & Shima, S. Methanogenic heterodisulfide reductase (HdrABC-MvhAGD) uses two
794 noncubane [4Fe-4S] clusters for reduction. *Science* **357**, 699–703 (2017).
795 80. Aziz, I. *et al.* Structural and mechanistic basis of the central energy-converting methyltransferase complex of
796 methanogenesis. *Proc. Natl. Acad. Sci.* **121**, e2315568121 (2024).
797 81. Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat. Biotechnol.*
798 **39**, 105–114 (2021).
799 82. Rosconi, F. *et al.* A bacterial pan-genome makes gene essentiality strain-dependent and evolvable. *Nat. Microbiol.* **7**,
800 1580–1592 (2022).
801 83. Yu, M. K., Fogarty, E. C. & Eren, A. M. Diverse plasmid systems and their ecology across human gut metagenomes
802 revealed by PlasX and MobMess. *Nat. Microbiol.* **9**, 830–847 (2024).
803 84. Parks, D. H. *et al.* GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent,
804 rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* **50**, D785–D794 (2022).
805 85. O'Callaghan, J. L., Willner, D., Buttini, M., Huygens, F. & Pelzer, E. S. Limitations of 16S rRNA gene sequencing to
806 characterize *Lactobacillus* species in the upper genital tract. *Front. Cell Dev. Biol.* **9**, 641921 (2021).
807 86. Tomita, S., Saito, K., Nakamura, T., Sekiyama, Y. & Kikuchi, J. Rapid discrimination of strain-dependent fermentation
808 characteristics among *Lactobacillus* strains by NMR-based metabolomics of fermented vegetable juice. *PLOS ONE* **12**,
809 e0182229 (2017).
810 87. Forster, A. H. & Gescher, J. Metabolic engineering of *Escherichia coli* for production of mixed-acid fermentation end
811 products. *Front. Bioeng. Biotechnol.* **2**, (2014).
812 88. Rosero, J. A. *et al.* Reclassification of *Eubacterium rectale* (Hauduroy *et al.* 1937) in a new genus *Agathobacter* gen. nov.
813 as *Agathobacter rectalis* comb. nov., and description of *Agathobacter ruminis* sp. nov., isolated from the rumen
814 contents of sheep and cows. *Int. J. Syst. Evol. Microbiol.* **66**, 768–773 (2016).
815 89. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life.
816 *Nat. Biotechnol.* **36**, 996–1004 (2018).
817 90. Medina, J. M., Fernández-López, R., Crespo, J. & Cruz, F. D. L. Propionate fermentative genes of the gut microbiome
818 decrease in inflammatory bowel disease. *J. Clin. Med.* **10**, 2176 (2021).
819 91. Louis, P., McCrae, S. I., Charrier, C. & Flint, H. J. Organization of butyrate synthetic genes in human colonic bacteria:
820 phylogenetic conservation and horizontal gene transfer. *FEMS Microbiol. Lett.* **269**, 240–247 (2007).

821 92. Cui, Y. *et al.* Production of butyl butyrate from lignocellulosic biomass through *Escherichia coli-Clostridium beijerinckii*
822 G117 co-culture. *Process Biochem.* **128**, 58–67 (2023).
823 93. Gonzalez-Alba, J. M., Baquero, F., Cantón, R. & Galán, J. C. Stratified reconstruction of ancestral *Escherichia coli*
824 diversification. *BMC Genomics* **20**, 936 (2019).
825 94. Palomino, A. *et al.* Metabolic genes on conjugative plasmids are highly prevalent in *Escherichia coli* and can protect
826 against antibiotic treatment. *ISME J.* **17**, 151–162 (2023).
827 95. Abriouel, H. *et al.* New insights into the role of plasmids from probiotic *Lactobacillus pentosus* MP-10 in Aloreña table
828 olive brine fermentation. *Sci. Rep.* **9**, 10938 (2019).
829 96. Pudlo, N. A. *et al.* Phenotypic and genomic diversification in complex carbohydrate-degrading human gut bacteria.
830 *mSystems* **7**, e00947-21 (2022).
831 97. Leinonen, R., Sugawara, H., Shumway, M., & on behalf of the International Nucleotide Sequence Database
832 Collaboration. The Sequence Read Archive. *Nucleic Acids Res.* **39**, D19–D21 (2011).
833 98. Poyet, M. *et al.* A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic
834 microbiome research. *Nat. Med.* **25**, 1442–1452 (2019).
835 99. Hoyles, L. *et al.* Molecular phenomics and metagenomics of hepatic steatosis in non-diabetic obese women. *Nat. Med.*
836 **24**, 1070–1080 (2018).
837 100. Mardinoglu, A. *et al.* An integrated understanding of the rapid metabolic benefits of a carbohydrate-restricted diet on
838 hepatic steatosis in humans. *Cell Metab.* **27**, 559-571.e5 (2018).
839 101. Qin, N. *et al.* Alterations of the human gut microbiome in liver cirrhosis. *Nature* **513**, 59–64 (2014).
840 102. Bushnell, B. BBMap. (2014).
841 103. Andrews, S. FastQC: A quality control tool for high throughput sequence data. (2010).
842 104. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in
843 a single report. *Bioinformatics* **32**, 3047–3048 (2016).
844 105. Togo, A. H. *et al.* Description of *Mediterraneibacter massiliensis*, gen. nov., sp. nov., a new genus isolated from the gut
845 microbiota of an obese patient and reclassification of *Ruminococcus faecis*, *Ruminococcus lactaris*, *Ruminococcus*
846 *torques*, *Ruminococcus gnavus* and *Clostridium glycyrrhizinilyticum* as *Mediterraneibacter faecis* comb. nov.,
847 *Mediterraneibacter lactaris* comb. nov., *Mediterraneibacter torques* comb. nov., *Mediterraneibacter gnavus* comb. nov.
848 and *Mediterraneibacter glycyrrhizinilyticus* comb. nov. *Antonie Van Leeuwenhoek* **111**, 2107–2128 (2018).
849 106. Murotomi, K., Tourlousse, D. M., Hamajima, M. & Sekiguchi, Y. Complete genome sequence of *Roseburia faecis* M72/1^{T}.
850 *Microbiol. Resour. Announc.* **14**, e01253-24 (2025).
851 107. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes - a 2019 update. *Nucleic Acids Res.* **48**, D445–
852 D453 (2020).
853 108. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and
854 usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
855 109. Minh, B. Q. *et al.* IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol.*
856 *Biol. Evol.* **37**, 1530–1534 (2020).
857 110. Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: improving the ultrafast bootstrap
858 approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).
859 111. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for
860 accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
861 112. Eddy, S. HMMER. (2019).
862 113. Kanehisa, M., Furumichi, M., Sato, Y., Matsuura, Y. & Ishiguro-Watanabe, M. KEGG: biological systems database as a
863 model of the real world. *Nucleic Acids Res.* **53**, D672–D677 (2025).
864 114. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods*
865 **18**, 366–368 (2021).

115. Treiber, M. L., Taft, D. H., Korf, I., Mills, D. A. & Lemay, D. G. Pre- and post-sequencing recommendations for functional annotation of human fecal metagenomes. *BMC Bioinformatics* **21**, 74 (2020).
116. Tamames, J. & Puente-Sánchez, F. SqueezeMeta, a highly portable, fully automatic metagenomic analysis pipeline. *Front. Microbiol.* **9**, 3349 (2019).
117. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct *de Bruijn* graph. *Bioinformatics* **31**, 1674–1676 (2015).
118. Hyatt, D., LoCascio, P. F., Hauser, L. J. & Uberbacher, E. C. Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics* **28**, 2223–2230 (2012).
119. Puente-Sánchez, F., García-García, N. & Tamames, J. SQMtools: automated processing and visual analysis of 'omics data with R and anvi'o. *BMC Bioinformatics* **21**, 358 (2020).
120. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
121. Haft, D. H. *et al.* RefSeq and the prokaryotic genome annotation pipeline in the age of metagenomes. *Nucleic Acids Res.* **52**, D762–D769 (2024).
122. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
123. Wickham, H. *et al.* Welcome to the Tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).
124. Wickham, H. Ggplot2*: Elegant Graphics for Data Analysis*. (Springer-Verlag New York, 2016).
125. Kassambara, A. *ggpubr*: *ggplot2* based publication ready plots. (2023).
126. Kolde, R. *pheatmap*: pretty heatmaps. (2019).
127. Kassambara, A. *rstatix*: pipe-friendly framework for basic statistical tests. (2023).