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 towards the liver. This axis can be disrupted by multiple
48 interconnected factors, affecting the progression of MASLD and fibrosis[6]. An alteration 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]. Also, an alteration of the
54 bile acid composition and circulation affects liver metabolism and function. Because they play a role in
55 the regulation of glucose and lipid metabolism, their dysregulation contributes to gut inflammation and
56 MASLD[10,11].

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

73 **2. Results**

74 ***Agathobacter rectalis* is depleted in MASLD**

75 Certain bacterial clades have been positively correlated with MASLD progression, including
76 *Enterobacteriaceae*[15,17] such as *Escherichia / Shigella spp*.[17,18], *Bacteroidaceae* like *Bacteroides spp.*[18,19],
77 *Veillonellaceae*[15] or *Streptococcus spp.*[20]. Conversely, other clades have been reported to exhibit an
78 inverse correlation with MASLD progression, including *Ruminococcaceae*[15] such as *Faecalibacterium*
79 *prausnitzii*[18,21], *Lachnospiraceae* like *Eubacterium rectale*[18,21] or *Dorea longicatena*[21], *Clostridiaceae* like
80 *Clostridium spp.*[17], or even *Pseudomonas spp.*[20]. However, these studies often yield conflicting results.
81 For example, a reported a positive correlation between *Ruminoccocus spp.* abundance and MASLD
82 progression[19] contradicts one of the main findings in two posterior studies[18,21]. Similarly, *Prevotella spp.*
83 was found to decrease with MASLD development[22] and fibrosis stage[19], yet another study reported an
84 increase in this genus in MASLD patients[20]. There are additional examples of discordant taxonomic
85 signatures in MASLD, such as *Blautia spp.* or *Roseburia spp.*[23].

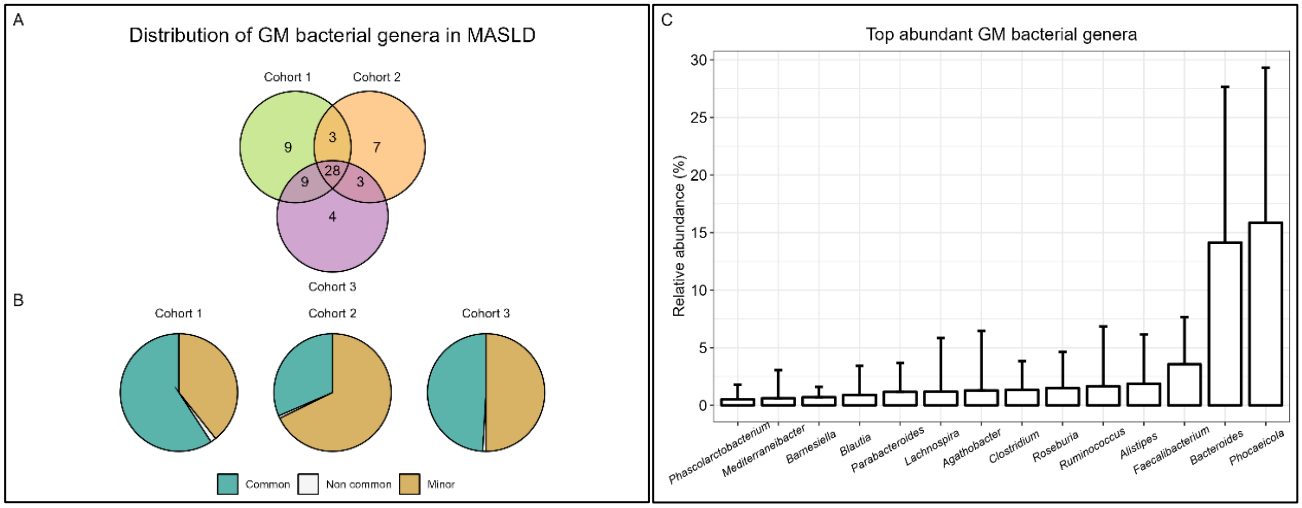
86 To investigate whether any bacterial genera is systematically altered in MASLD, we profiled the
87 taxonomic composition of the GM communities across all samples from the three cohorts at genus and
88 species level using MetaPhlAn[24], as detailed in Materials and Methods. This analysis identified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

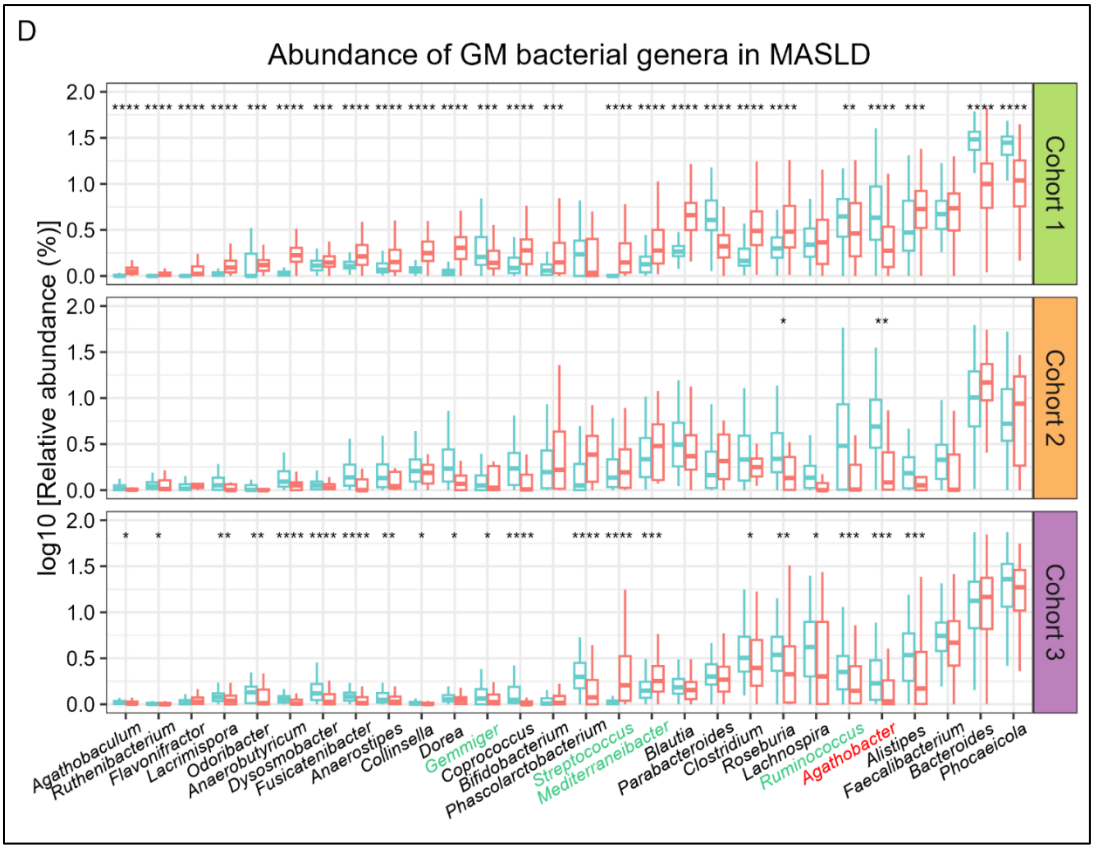
93 Among these shared genera, *Gemmiger*, *Streptococcus*, *Mediterraneibacter*, *Ruminococcus* and
94 *Agathobacter* exhibited a consistent differential abundance pattern between groups across every
95 cohort statistically significant in at least two cohorts (Fig. 1D). This exclusion criterion was applied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

106 Analogously, the highest abundant bacterial species within the GM were interrogated for
107 differences in abundances between groups across the three cohorts (Fig. 1E). Only *Agathobacter*

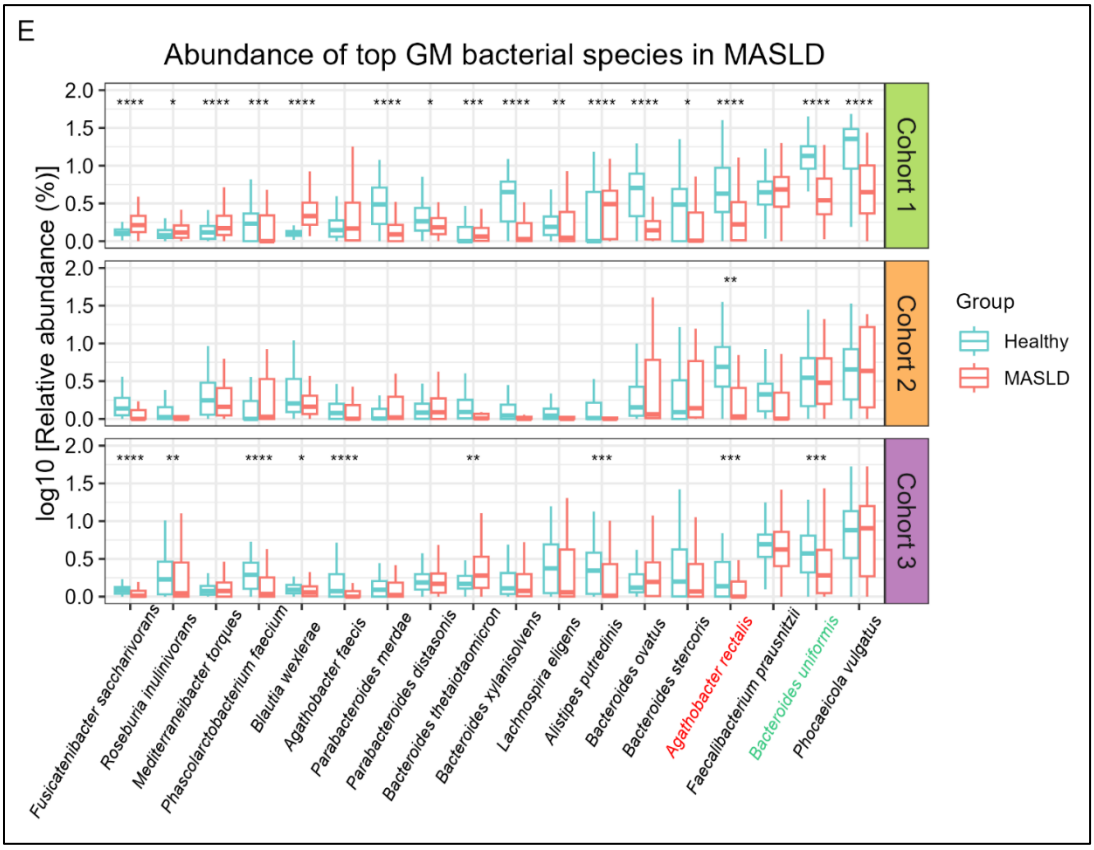
rectalis and *Bacteroides uniformis*, two of the most abundant GM bacterial species (Supplementary Fig. 1), exhibited a consistent pattern of differential abundance across all cohorts, with statistical significance in at least two. However, only *A. rectalis* showed a significant difference between groups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

Figure 1 (cont. next page). 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 in the three cohorts of study. 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 = 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, in logarithmic scale. Each facet represents a cohort of study. Differences in abundances were evaluated using a pairwise Mann-Whitney test with a Benjamini-Hochberg adjustment for multiple testing. Horizontal bars represent median levels, while boxes and whiskers represent the data from first to third quartiles and from the quartiles to the minimum and 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 increase between groups 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

128


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

129

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

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

139 The most well-characterized metabolic family are short chain fatty acids, which are produced
140 by the GM through the colonic fermentation of complex carbohydrates present in dietary fiber that
141 escape digestion in the small intestine. These metabolic pathways have been relatively well understood
142 for nearly thirty years[34] and their impact on GM homeostasis has been reviewed in several articles[35–37].
143 For this reason, we interrogated the metagenomic fecal samples from the patient cohorts for metabolic
144 signatures that could explain the MASLD phenotype. First, we focused on the genes encoding enzymes
145 involved in butyrate formation as a proxy for the bacteria that carry them, independently of their taxa.
146 To validate our hypothesis, we analyzed the genic abundance of the final enzymes in butyrate
147 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 is a
155 dephosphorylation of butyryl-P that yields one ATP through an energetically favorable reaction that is
156 catalyzed by a butyrate kinase encoded by *buk.* In the precursor reaction, butyryl-P is formed through
157 a phosphorylation of butyryl-CoA catalyzed by a phosphate butyryl transferase encoded by *ptb*.
158 Butyrate also accumulates after the transference of a CoA cofactor from butyryl-CoA to acetate through
159 butyryl-CoA:acetate CoA transferase, encoded by *but*. This CoA-transferase route also conserves the
160 energy of the CoA bond in the newly formed CoA-moiety of the co-substrate. Butyryl-CoA is produced
161 from crotonyl-CoA through two different enzymes: butyryl--CoA trans-2-oxidorreductase, encoded by
162 *fabV*, and butyryl--CoA dehydrogenase, encoded by *bcd.* This pathway is further reviewed in the human
163 GM elsewhere[49].

164 We quantified the abundance of these genes in the metagenomic fecal samples of the two
165 groups 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 cirrhosis MASLD patients compared to the healthy group, and they also decreased significantly
168 with MASLD progression (Fig. 2B). Precursor gene *fabV* was nearly an order of magnitude less abundant
169 than *ptb* and *bcd* in both cohorts, underscoring the greater prevalence of *bcd* over *fabV* in the CoA-
170 transferase route. Additionally, *ptb* did not exhibit significant differences in Cohort 2. *but* was
171 significantly depleted in the control group in both cohorts, whereas *buk* showed no differences in
172 Cohort 2 either. Notably, the depletion of butyrate genes did not replicate in MASLD patients from
173 Cohort 1.

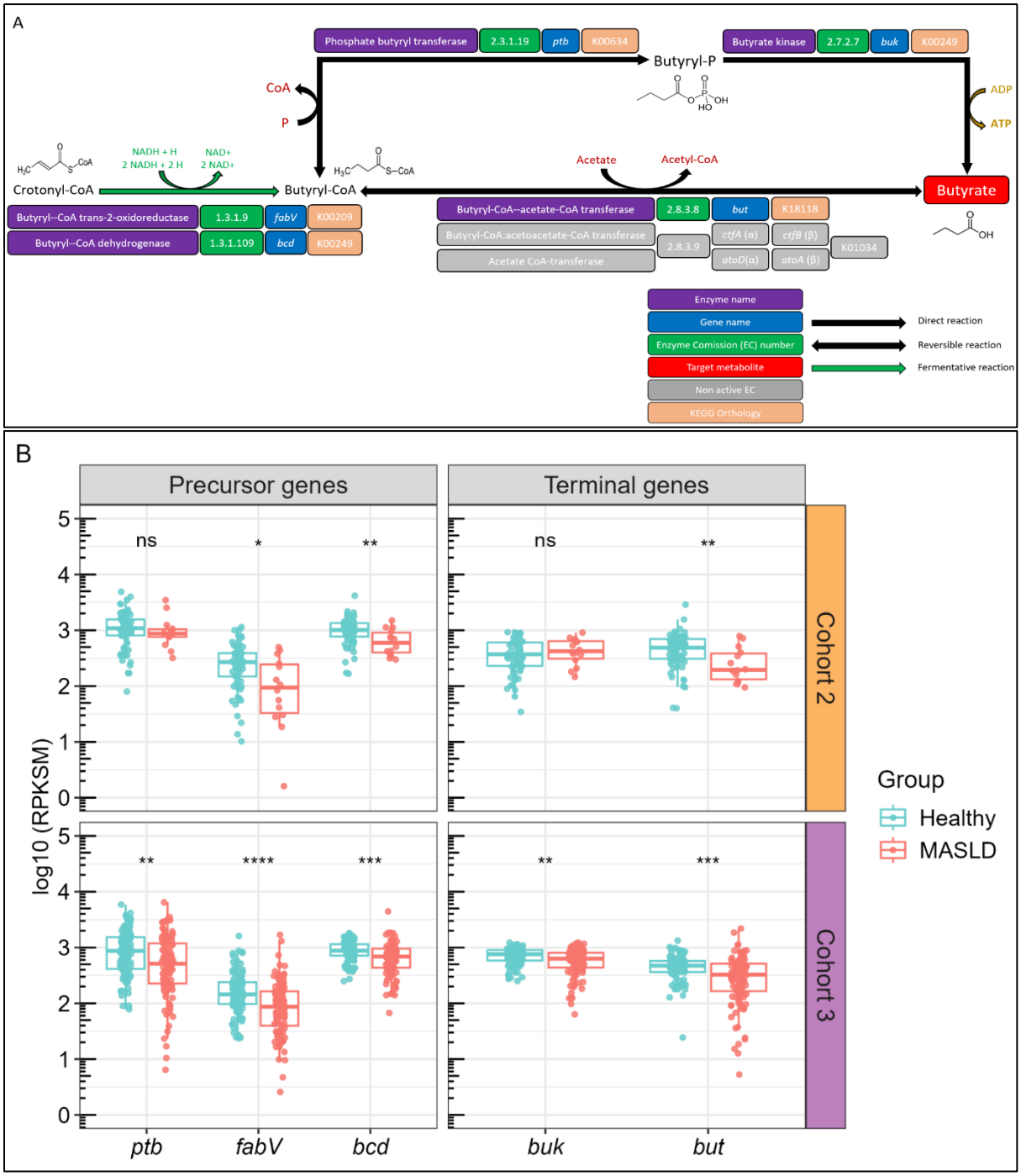


Figure 2. Abundance of butyrate-producing genes in patient cohorts. (A) Characterization of the terminal reactions involved in the formation of bacterial butyrate from crotonyl-CoA. Enzymes, EC numbers, coding genes and KOs, as well as the reaction type, are indicated for each metabolic step according to the annexed legend. **(B)** Abundance of butyrate-producing genes in Cohorts 2 and 3, in logarithmic scale. Boxplots represent the genic abundance from the metagenomic measures of the fecal samples, and differences in abundances were evaluated using a pairwise Mann-Whitney test with a 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 SCAs are a metabolic family also suggested to regulate GLA integrity, with several studies linking
187 GM endogenous alcohol production to MASLD. The first, published over a decade ago, reported
188 elevated peripheral blood ethanol and fecal *Escherichia* spp. 16S rRNA levels in MASH patients[50]. Similar
189 findings were reported in a pediatric cohort, where children with MASLD exhibited higher levels of fecal
190 ethanol, *Prevotella* spp., and Gammaproteobacteria 16S rRNA[51]. A more recent study showed that
191 postprandial peripheral blood ethanol and fecal *Lactobacillaceae* 16S rRNA levels increased in MASH
192 patients after selective inhibition of host ADH, an effect that was reversed with broad-spectrum
193 antibiotic treatment[52]. In 2019, Yuan and colleagues identified a high-alcohol-producing strain of
194 *Klebsiella pneumoniae* capable of causing MASLD through ethanol production[53]. Specifically, they found
195 that FMT from a MASH patient containing this strain induced MASLD in healthy mice, whereas removing
196 the strain before FMT prevented the disease. Furthermore, ethanol produced by this *K. pneumoniae*
197 strain was shown to disrupt fat homeostasis in hepatic cells and cause mitochondrial dysfunction,
198 oxidative stress, and the accumulation of reactive oxygen species[54], which are critical to MASH
199 development as they drive lipid peroxidation and inflammation[55]. Additionally, a recent computational
200 study identified a positive correlation between the abundance of two fermentative pathways in certain
201 families from class Clostridia and fatty-liver disease in patients. These pathways include the production
202 of ethanol from pyruvate and the heterolactic fermentative pathway[56].

203 With this ample evidence, we hypothesized that genes encoding the enzymes involved in the
204 formation of endogenous SCAs could be another metabolic mechanism involved in MASLD pathology.
205 We focused on ethanol and propanol, which are produced by the reduction of their respective
206 aldehydes through ethanol and propanol dehydrogenases. These enzymes help to reduce the levels of

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

226 We quantified the abundance of genes involved in bacterial SCA formation in the metagenomic
227 samples of the three cohorts as described above. *mhpF* and *aldB*, involved in ethanol production, and
228 *yahK* and *ahr*, involved in acetaldehyde production, were significantly enriched in the GM of MASLD

229 patients from Cohort 1 and Cohort 3 compared to healthy groups. Their abundance also increased with

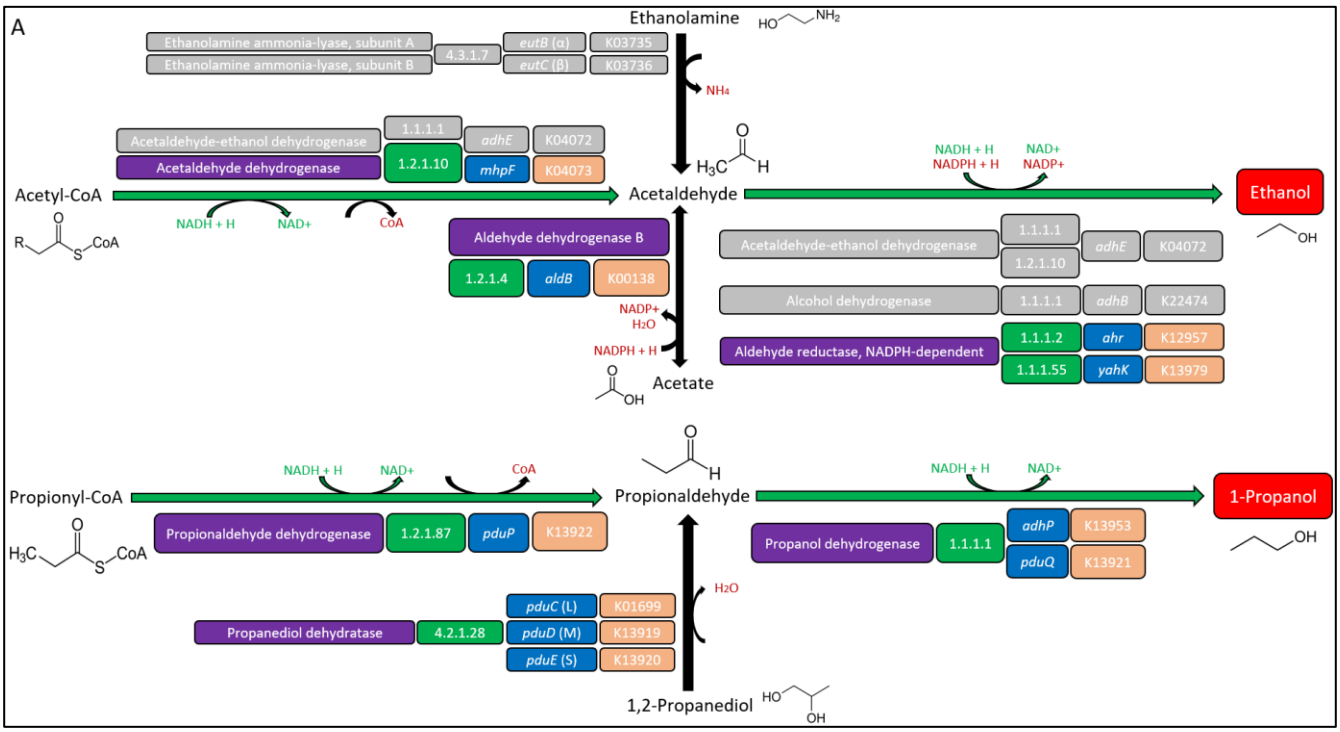
230 MASLD progression (Fig. 3B). *mhpF* and *ahr* were depleted compared to *aldB* and *yahK* respectively in

231 all three cohorts. All genes involved in the formation of 1-propanol and propionaldehyde were also

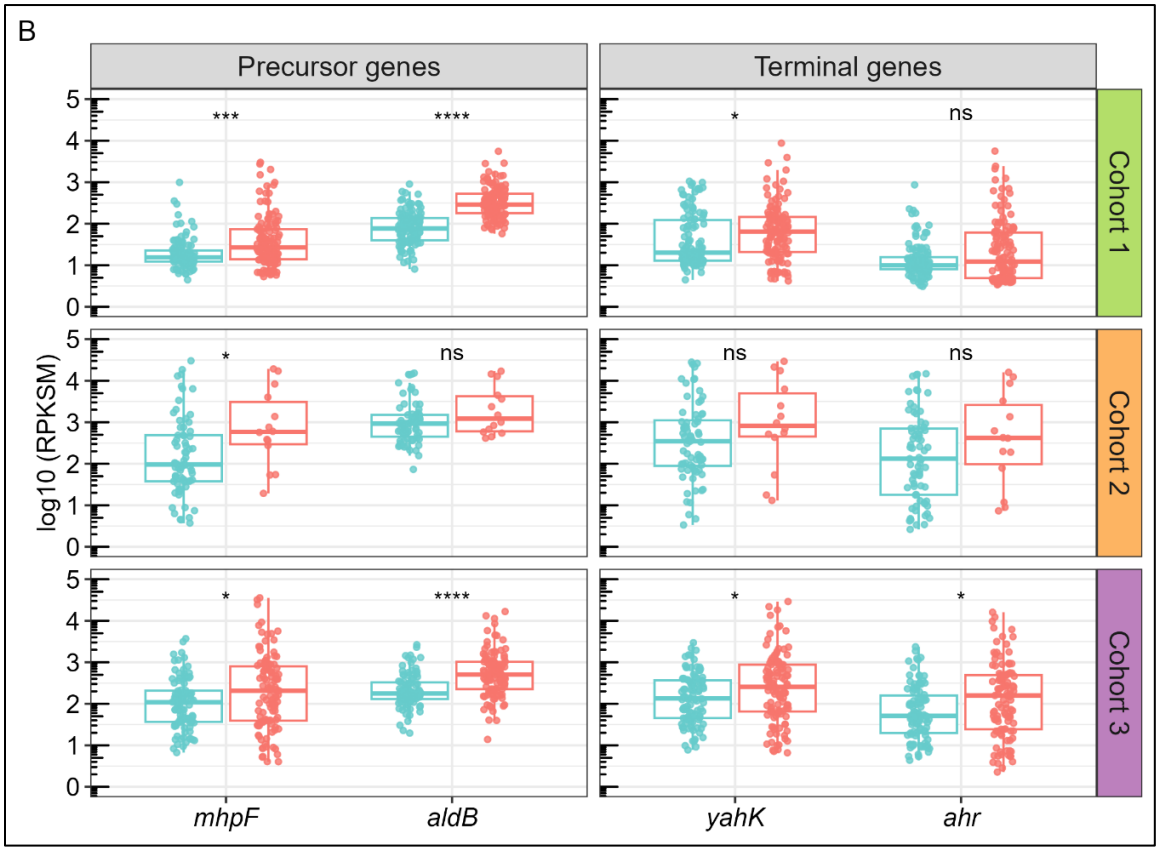
232 significantly enriched in MASLD patients by more than half order of magnitude, increasing with MASLD

233 progression (Fig. 3C, Supplementary Fig. 4B). These findings reveal a systematic increase of SCA-

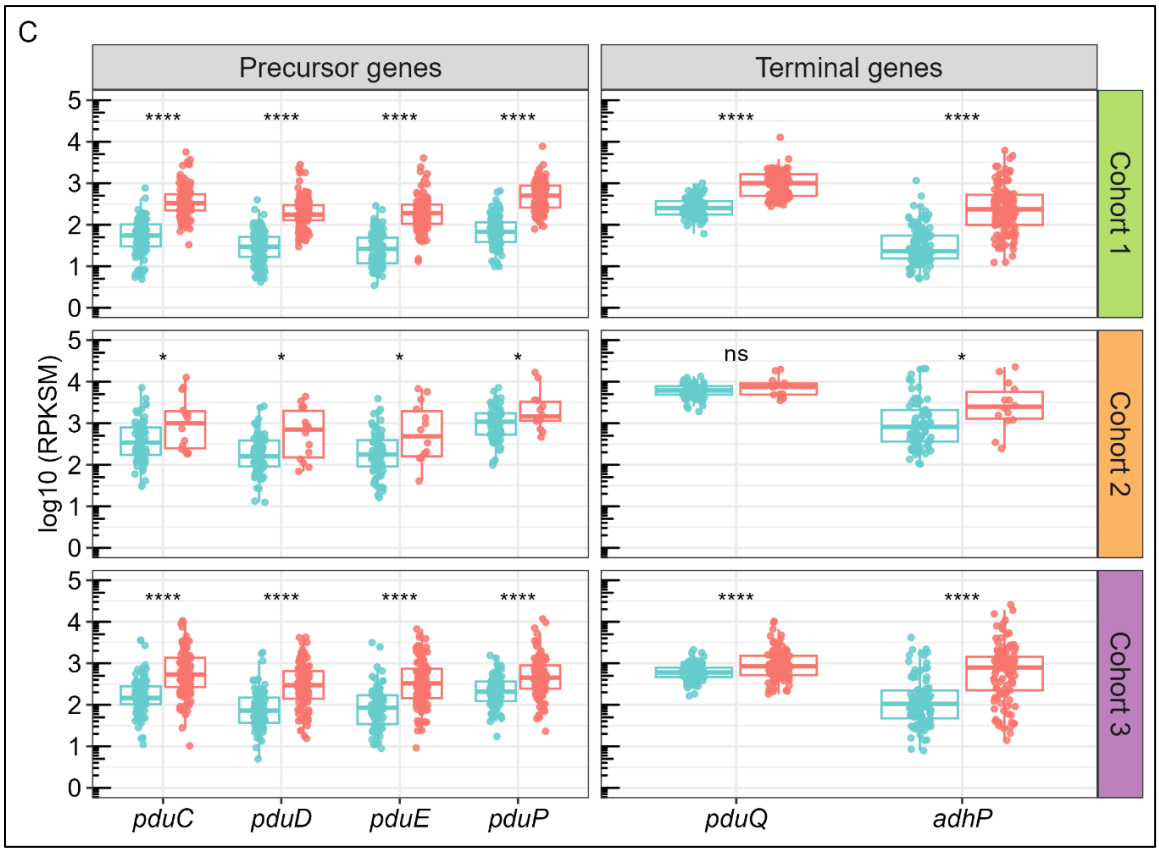
234 producing genes in MASLD.



235


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

236


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

237

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

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

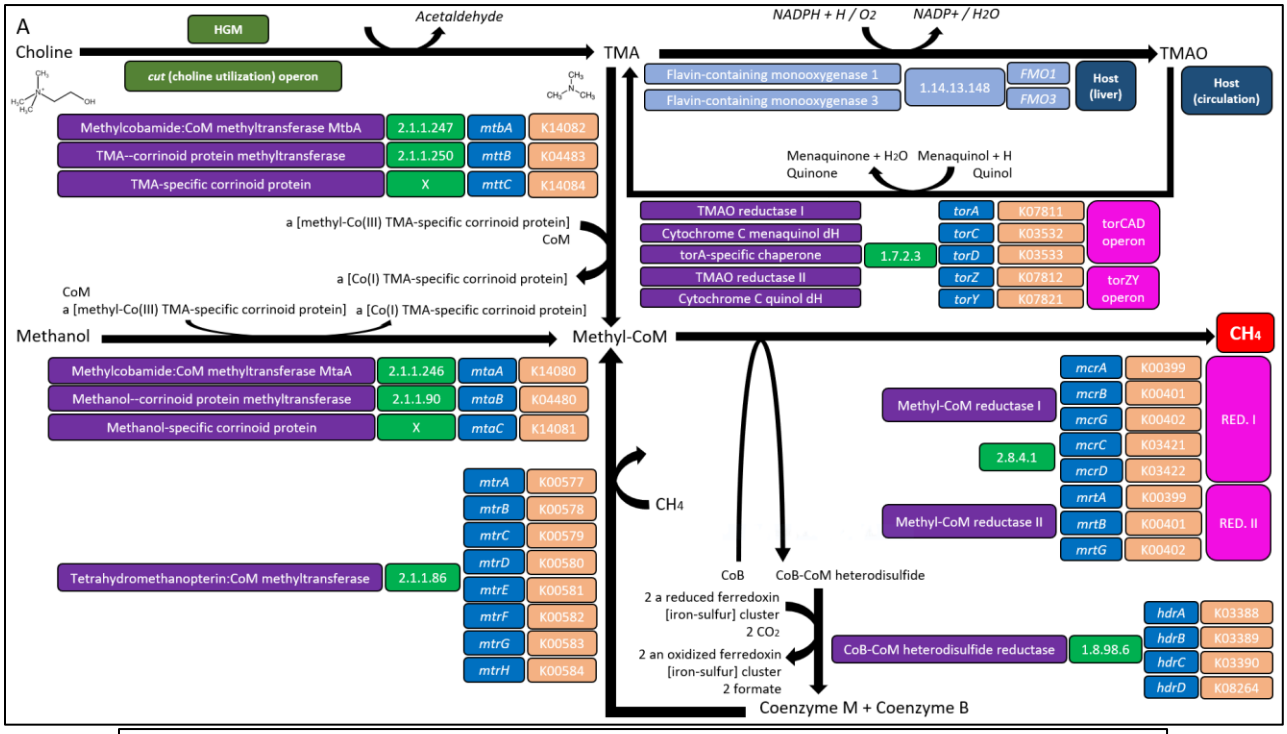
247 Trimethylamine N-oxide (TMAO) is a metabolite derived from the oxidation of trimethylamine
248 (TMA), which is produced by the microbial degradation of dietary compounds such as choline and L-
249 carnitine. Circulating TMAO levels have also been linked to MASLD severity[69] and all-cause mortality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

253 Methane production in the GM occurs through a two-step process: a simultaneous TMA
254 demethylation and CoM methylation, followed by the reduction of methyl-CoM (Fig. 4A). TMA
255 originates from choline degradation via the *cut* operon[73] and is oxidized in the liver to TMAO[74] or
256 reduced back to TMA under anaerobic conditions via the *torCAD* and *torZY* operons. Specifically, *torA*
257 and *torZ* encode two TMAO reductases, whose molecular mechanisms are reviewed elsewhere[75]. CoM
258 methylation occurs through two well-characterized enzymatic systems: a TMA-specific (composed of
259 *mttC*, *mtbA* and *mttB*) and a methanol-specific pathway (composed of *mtaC, mtaA* and *mtaB*)[76]. The
260 final reduction of methyl-CoM to methane is catalyzed by methyl-CoM reductase (MCR) I, which is
261 encoded by the *mcrBDCGA* operon[77]. MCR I has an alternative isozyme termed MCR II, encoded by
262 *mrtBGA* operon[78]. The resulting heterodisulfide (CoB-S-S-CoM) is then reduced to CoM and CoB by a
263 heterodisulfide reductase encoded by operon *hdrABC*, which regenerate both cofactors[79]. Additionally,

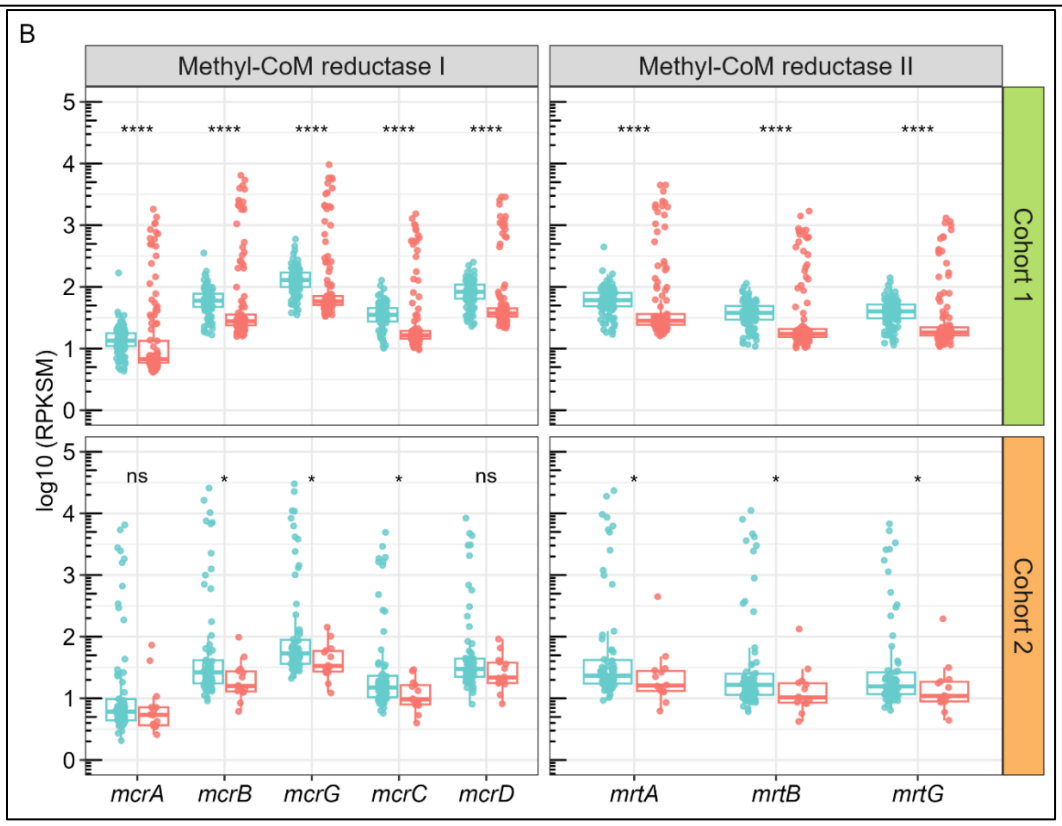
264 methyl-CoM is replenished through tetrahydromethanopterin CoM-methyltransferase (encoded by
265 operon *mtrABCDEFGH*), which methylates CoM[80].

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

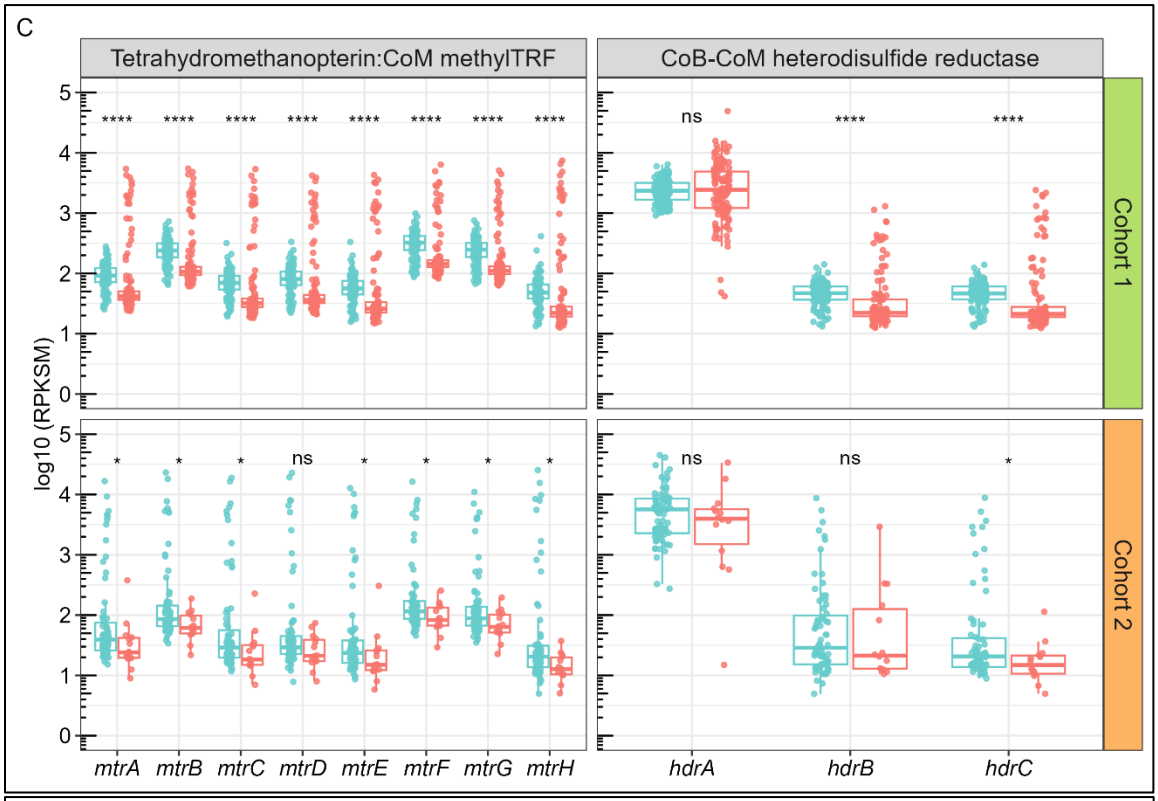
281 Finally, we measured the abundance of five universal single-copy genes (USCGs), as described
282 in Material and Methods, to confirm that observed differences in genic abundance between sample
283 groups were truly due to MASLD phenotype and not influenced by other biological factors nor spurious
284 technical artifacts derived from the quantification method. We found no differences in gene abundance
285 between groups for any USCGs in any of the cohorts (Supplementary Fig. 1).


A
Choline
HGM
Acetaldehyde
TMA
NADPH + H / O2
NADP+ / H2O
TMAO
cut (choline utilization) operon
Flavin-containing monooxygenase 1
Flavin-containing monooxygenase 3
1.14.13.148
FMO1
FMO3
Host (liver)
Host (circulation)
Methylcobamide:CoM methyltransferase MtbA
2.1.1.247
mtbA
K14082
TMA--corrinoid protein methyltransferase
2.1.1.250
mttB
K04483
TMA-specific corrinoid protein
X
mttC
K14084
Menaquinone + H2O
Quinone
Menaquinol + H
Quinol
a [methyl-Co(III) TMA-specific corrinoid protein]
CoM
a [Co(I) TMA-specific corrinoid protein]
TMAO reductase I
Cytochrome C menaquinol dH
torA-specific chaperone
TMAO reductase II
Cytochrome C quinol dH
1.7.2.3
torA K07811
torC K03532
torD K03533
torZ K07812
torY K07821
torCAD operon
torZY operon
CoM
a [methyl-Co(III) TMA-specific corrinoid protein]
a [Co(I) TMA-specific corrinoid protein]
Methanol
Methyl-CoM
CH4
Methylcobamide:CoM methyltransferase MtaA
2.1.1.246
mtaA
K14080
Methanol--corrinoid protein methyltransferase
2.1.1.90
mtaB
K04480
Methanol-specific corrinoid protein
X
mtaC
K14081
Methyl-CoM reductase I
2.8.4.1
mcrA K00399
mcrB K00401
mcrG K00402
mcrC K03421
mcrD K03422
RED. I
Methyl-CoM reductase II
mrtA K00399
mrtB K00401
mrtG K00402
RED. II
CH4
mtrA K00577
mtrB K00578
mtrC K00579
mtrD K00580
mtrE K00581
mtrF K00582
mtrG K00583
mtrH K00584
Tetrahydromethanopterin:CoM methyltransferase
2.1.1.86
CoB
CoB-CoM heterodisulfide
2 a reduced ferredoxin [iron-sulfur] cluster
2 CO2
2 an oxidized ferredoxin [iron-sulfur] cluster
2 formate
CoB-CoM heterodisulfide reductase
1.8.98.6
hdrA K03388
hdrB K03389
hdrC K03390
hdrD K08264
Coenzyme M + Coenzyme B

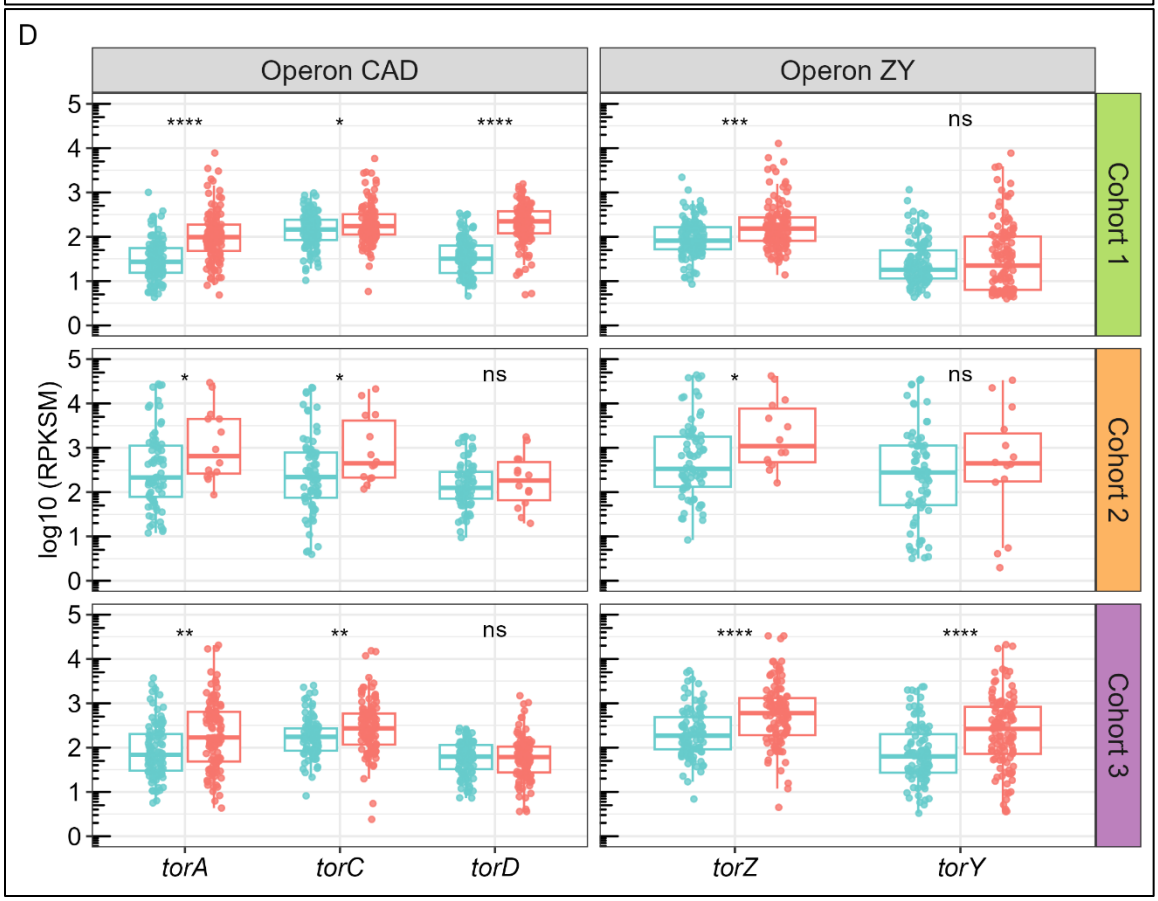
286


B
Methyl-CoM reductase I
Methyl-CoM reductase II
Cohort 1
Cohort 2
log10 (RPKSM)
****
ns
*
mcrA
mcrB
mcrG
mcrC
mcrD
mrtA
mrtB
mrtG

287


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

288


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

289
290

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

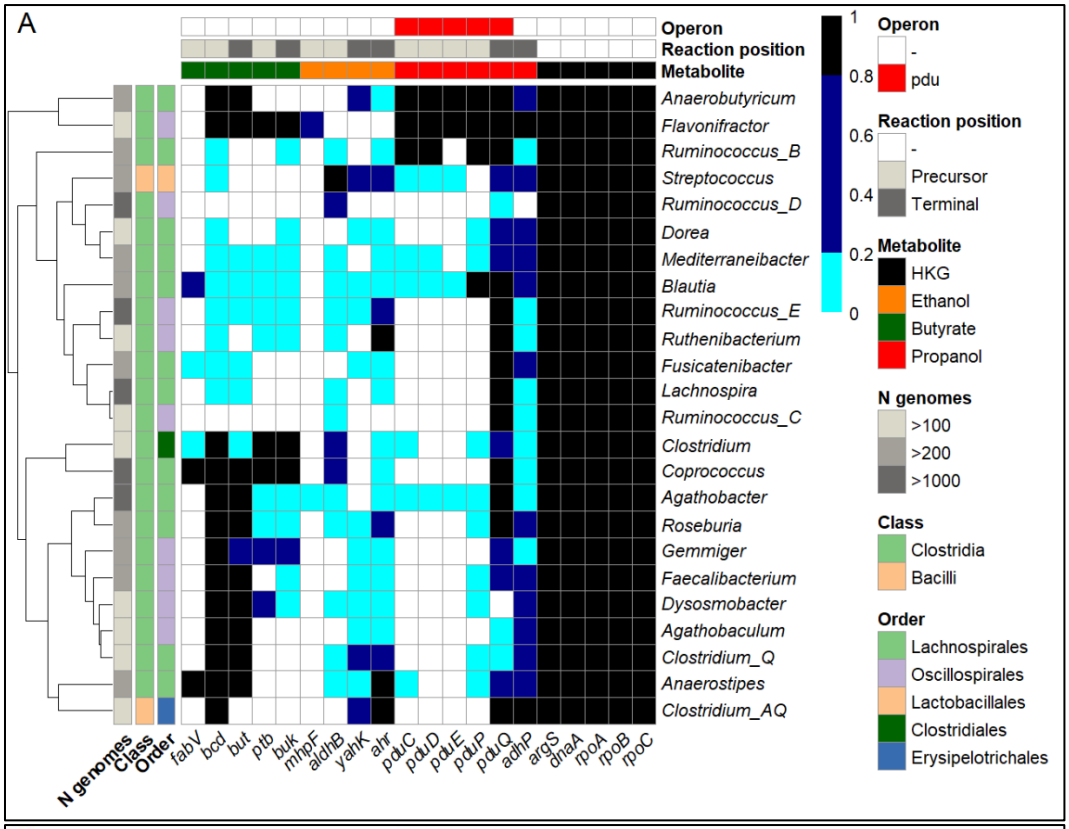
297 **Candidate metabolic genes are accessory in the GM**

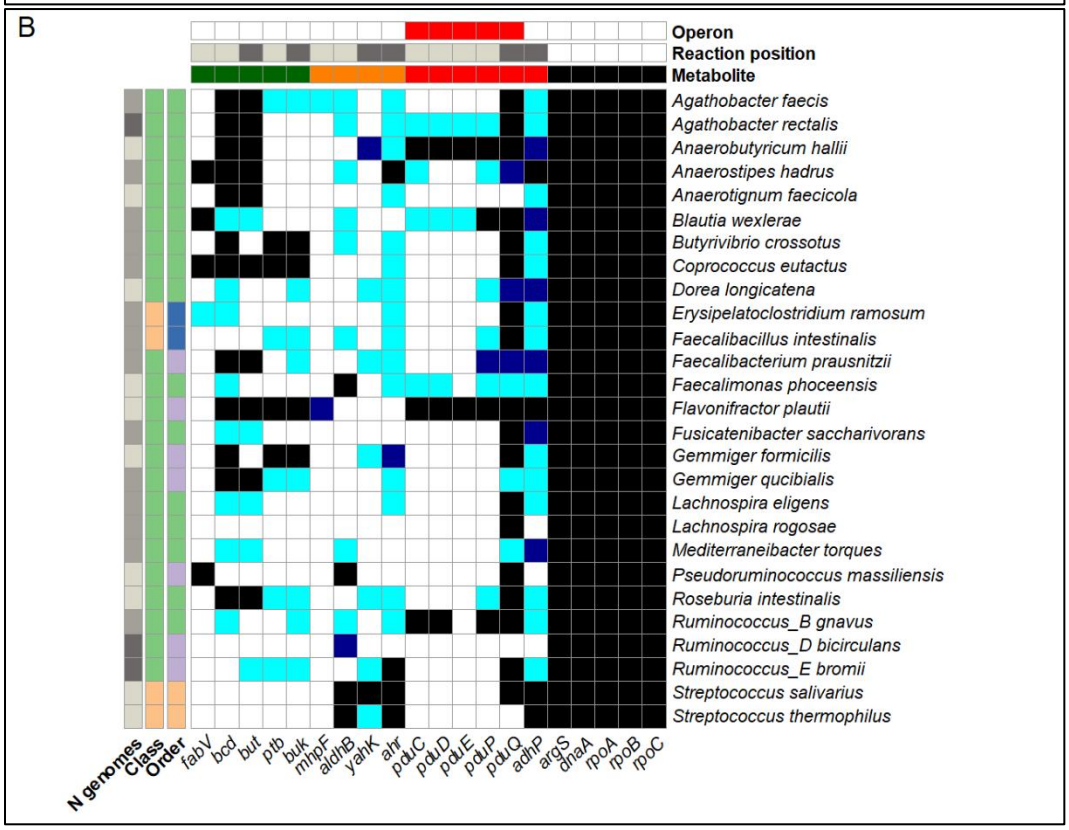
298 Although this analysis identifies *A. rectalis* and multiple metabolic genes involved in the
299 formation of butyrate, SCAs, TMAO and methane as consistent taxonomic and functional signatures
300 associated with MASLD, these markers alone are not sufficiently robust to establish a causal link
301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

306 To test this hypothesis, we assessed whether candidate metabolic genes involved in MASLD
307 were core or accessory across some of the most abundant human GM genomes. Selection of genomes
308 and genic presence were determined as described in Materials and Methods. Briefly, we defined the
309 presence of a target gene in a clade as the percentage of strains encoding it, classifying it as core if
310 present in >80% of congeneric or conspecific genomes, as accessory if present in 20-80%, and as highly
311 accessory if present in <20% genomes. Most of the selected GM genera (n=24) belong to class Clostridia,
312 specifically to orders Lachnospirales and Oscillospirales (Fig. 5A). Butyrate-producing genes from the
313 butyryl-CoA pathway (*bcd* and *but*) were core in nearly half of these genera, while those from the
314 butyryl-P pathway (*ptb* and *buk*) were core only in *Flavonifractor, Clostridium* and *Coprococcus.*
315 Notably, butyrate-producing genes were highly accessory across several genera, including *Lachnospira,*

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

327 Beyond their accessory nature within GM pangenomes, candidate metabolic genes were also
328 encoded in plasmids (Fig. 5C). Genes involved in ethanol formation such as *aldhB, yahK* and *ahr* were
329 among the most widely distributed, being present in hundreds of different plasmids. In contrast,
330 butyrate-producing genes (*fabV, but, ptb,* and *buk*) and propanol-related genes (*pduC, pduD, pduE,*
331 *pduP,* and *pduQ*) were detected in fewer plasmids, further supporting their variable presence across
332 GM genomes. A notable exception was *bcd*, which exhibited a broader presence across both
333 chromosomal and plasmid-encoded sequences. These findings reinforce the idea that non-essential
334 metabolic functions, especially ethanol production, are encoded in MGEs, contributing to their
335 accessory nature within the GM.


A
Operon
Reaction position
Metabolite
Anaerobutyricum
Flavonifractor
Ruminococcus_B
Streptococcus
Ruminococcus_D
Dorea
Mediterraneibacter
Blautia
Ruminococcus_E
Ruthenibacterium
Fusicatenibacter
Lachnospira
Ruminococcus_C
Clostridium
Coprococcus
Agathobacter
Roseburia
Gemmiger
Faecalibacterium
Dysosmobacter
Agathobaculum
Clostridium_Q
Anaerostipes
Clostridium_AQ
N genomes
Class
Order
fabV
bcd
but
ptb
buk
mhpF
aldhB
yahK
ahr
pduC
pduD
pduE
pduP
pduQ
adhP
argS
dnaA
rpoA
rpoB
rpoC
1
0.8
0.6
0.4
0.2
0
Operon
-
pdu
Reaction position
-
Precursor
Terminal
Metabolite
HKG
Ethanol
Butyrate
Propanol
N genomes
>100
>200
>1000
Class
Clostridia
Bacilli
Order
Lachnospirales
Oscillospirales
Lactobacillales
Clostridiales
Erysipelotrichales


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

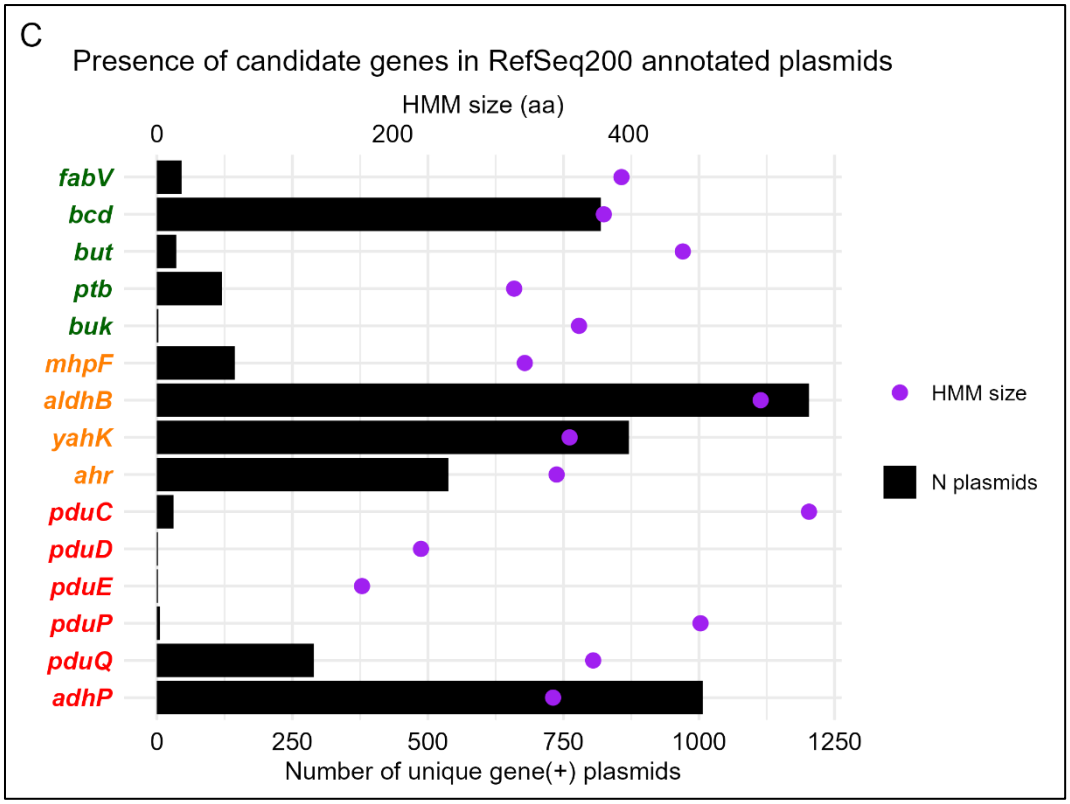


Figure 5. Presence profile of candidate metabolic genes in the human GM. (A-B) Heatmap showing the presence of candidate genes involved in MASLD across some of the most abundant clades in the human GM. The x-axis represents genes associated with butyrate and SCAs production, as well as USCGs. The y-axis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

3. Discussion

MASLD is a liver pathology influenced by multiple contributing factors, and the GM has been long suspected to play a role in its progression. Numerous studies have tried to ascribe a causality between GM composition and MASLD, but taxonomic signatures in key GM clades, such as *Blautia* and *Roseburia* spp., have been inconsistently associated with MASLD progression[23]. These issues may be linked to the use of 16S rRNA-based profiling in metataxonomic studies, which often lacks resolution at species and strain level[84]. However, the heterogeneity concomitant to bacterial genomes poses another

356 possible cause for the lack of association between disease progression and the taxonomic distribution.
357 Bacterial genomes are highly plastic, and strains from the same species often harbor different genes,
358 many of which are involved in different metabolic pathways[81]. Fermentative pathways in lactobacilli,
359 for example, are often strain-specific[85], and the products of the enterobacterial mixed acid
360 fermentation also exhibit wide variability between strains[86]. If MASLD progression is linked to the
361 metabolic functions of the GM, then the association between the disease and the taxonomic profile
362 may be obscured by the inherent metabolic diversity of bacterial species. To study these potential
363 alternatives, we followed two distinct strategies.

364 To identify species associated with MASLD, overcoming the limitations of 16S rRNA analysis, we
365 utilized multiple marker genes for taxonomic classification, in an attempt to provide a more
366 comprehensive view of microbial abundances in health and disease states. Using MetaPhlan, we
367 profiled the taxonomic composition of the GM in metagenomic fecal samples of healthy and MASLD
368 donors across three independent cohorts. We identified 28 bacterial genera shared across all cohorts
369 (Fig. 1A), collectively accounting for 30-60% of the total GM abundance (Fig. 1B). Among these, five
370 candidate genera -*Gemmiger*, *Streptococcus*, *Mediterraneibacter*, *Ruminococcus* and *Agathobacter*
371 (Fig. 1C)- and two species, *Agathobacter rectalis* and *Bacteroides uniformis* (Fig. 1E), exhibited
372 consistent differences in their relative abundance between groups across all the cohorts. These taxa
373 rank among the most abundant clades in the GM (Fig. 1C, Supplementary Fig. 1). However, only
374 *Agathobacter rectalis* showed statistically significant differences. Given its high abundance in the GM,
375 its coherent abundance trend at both genus and species level, and its consistent depletion in MASLD,
376 *A. rectalis* emerges as a robust taxonomic marker for the disease.

377 Although this observation has been previously reported[18,21], this result should be interpreted
378 with caution due to ongoing debates in bacterial nomenclature, as *A. rectalis* is a newer designation
379 that has replaced *Eubacterium rectale*[87]. Updates in taxonomic assignments pose challenges, as newer
380 species or genera may not be consistently reflected across different databases or studies[88,89]. Also,
381 taxonomic classifications themselves, whether based solely on 16S rRNA or on multiple markers, limit
382 our ability to reliably infer GM changes. Bacterial genome plasticity, driven by MGEs, can blur the lines
383 of taxonomy-phenotype associations, as they contribute to a vast strain-level diversity[81] that remains
384 undetected by taxonomic profiling methods. In fact, we have previously reported that geno-metabolic
385 shifts in GM propionate production associated with inflammatory bowel disease do not align with
386 species-level variations[90].

387 GM associated metabolism contributes to MASLD by disrupting the integrity of the GLA. For this
388 reason, we focused on identifying abundance differences in the genes responsible for producing
389 metabolites potentially implicated in MASLD across several patient cohorts, i.e. butyrate, SCAs, TMA
390 and methane. We isolated the GM gene families encoding the enzymes involved in the biosynthesis of
391 these metabolites and aligned them against the metagenomic samples to quantify their abundance.
392 We found that butyrate-producing genes are depleted in the GM of MASLD patients (Fig. 2B),
393 particularly those involved in the crotonyl-butyryl CoA axis: *fabV* (butyryl-CoA oxidoreductase), *bcd*
394 (butyryl-CoA dehydrogenase) and *but* (butyryl-CoA:acetate-CoA transferase) (Fig. 2A). While *ptb* and
395 *buk* were also higher in both healthy subcohorts, their levels did not differ significantly between initial
396 and advanced MASLD groups within Cohort 2. This suggests that CoA transfer may be the predominant
397 pathway for butyrate production in the GM, as previously reported[91]. The smaller differences in *ptb*
398 and *buk* in Cohort 2 may reflect a common MASLD phenotype across different fibrosis stages, along

399 with the limited sample size of this cohort compared to Cohort 3. Additionally, two other enzymes -
400 structured as two-subunit systems-, can transfer CoA from different cofactors to butyrate to produce
401 butyryl-CoA: butyryl-CoA:acetoacetate-CoA transferase (encoded by *atoD* and *atoA,* EC 2.8.3.9) and
402 acetate CoA-transferase (encoded by *ctfA* and *ctfB,* EC 2.8.3.9) (Fig. 2A). These were not included in this
403 study as they catalyze the reverse reaction, consuming butyrate to form butyryl-CoA[92]. Notably, the
404 depletion of butyrate genes was not observed in MASLD patients from Cohort 1.

405 Contrarily, genes involved in the formation of ethanol and propanol through connected
406 fermentative reactions are increased in MASLD (Fig. 3). Genes *mhpF* and *aldB* -encoding two aldehyde
407 dehydrogenases that produce ethanol-, and *yahK* and *ahr* -encoding two NADPH-dependent aldehyde
408 reductases that form acetaldehyde- are significantly increased in the GM of MASLD patients compared
409 to the healthy groups (Fig. 3B). Genes involved in the production of 1-propanol structured in operon
410 *pdu* are also increased in MASLD (Fig. 3C). Specifically, *pduCDE* -encoding a propanediol dehydratase-,
411 *pduP* -encoding a propionaldehyde dehydrogenase- and *pduQ* and *adhP* -encoding two distinct
412 propanol dehydrogenases-. The marked differential abundances of propanol-producing genes,
413 organized in the *pdu* operon and *adhP*, which have been largely unexplored in the context of liver
414 disease, suggests that 1-propanol produced in bacterial microcompartments may play an even more
415 significant role than ethanol in MASLD onset and progression.

416 The abundance of genes in the TMA-methane metabolic axis is also disrupted in MASLD. Genes
417 involved in methane formation (Fig. 4A) are depleted in the GM of individuals with the pathology. This
418 reduction includes both MCR I and II coding-genes (Fig. 4B), which reduce methyl-CoM to methane and
419 *mtr*, *hdr* (Fig. 4C), and Mtb / Mta-coding genes (Supplementary Fig. 2), dedicated to regenerating
420 methyl-CoM. In contrast, TMAO-reductive genes, organized in *torCAD* and *torZY* operons and

421 responsible for regenerating TMA, are increased in MASLD (Fig. 3D). These results suggest that methane
422 production is impaired in MASLD. The opposing trend in the abundance of *tor* operons further indicates
423 that this impairment stems from a metabolic bottleneck that leads to TMA accumulation instead of its
424 conversion to methyl-CoM, effect that is driven by TMAO reductases I and II.

425 Taken together, these results reveal a geno-metabolic alteration associated with MASLD,
426 characterized by an increase in genes involved in the production of SCAs and TMA, and a concurrent
427 depletion of genes responsible for butyrate and methane production (Fig. 6). We confirmed that genic
428 abundance differences were driven by the MASLD phenotype by showing no differences in USCG
429 abundance between groups (Supplementary Fig. 3). These findings support the hypothesis that the
430 accumulation of endogenous alcohols may represent a key metabolic feature of MASLD, occurring
431 alongside the displacement of butyrate-producing taxa, as previously proposed[33]. Moreover, the
432 observed decrease in *A. rectalis* abundance partially explain the reduction in butyrate levels, as this
433 organism is a known GM butyrate producer[87].

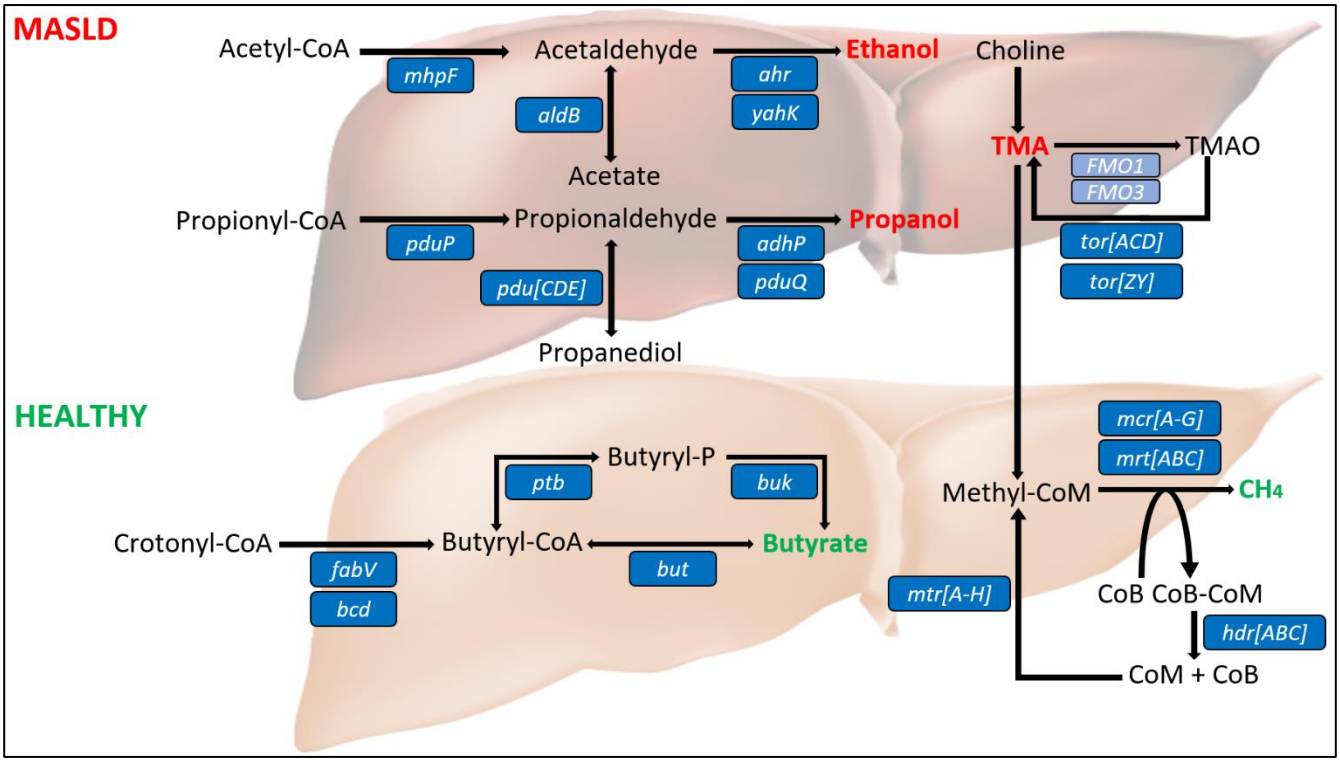


Figure 6. Geno-metabolic scheme involved in MASLD. Proposed scheme showing the GM genic alterations related to butyrate, SCAs, TMA and methane production in MASLD. Ethanol, propanol and TMA-producing genes are increased in the pathology (top), whereas butyrate and methane-producing genes are decreased (bottom). Blue boxes indicate the catalyzing genes whose abundance is altered in each metabolic step.

Importantly, we found that metabolic genes associated with MASLD exhibit a highly accessory profile within the GM (Fig. 5). Butyrate-producing genes are highly accessory in nearly half of the most abundant GM genera, while SCA-producing genes display an even more pronounced accessory pattern. Among the two main butyrate biosynthesis routes, genes from the butyryl-CoA pathway are more pervasive in the GM than those from the butyryl-P pathway. Notably, *Phocaeicola* and *Bacteroides* -the two most abundant GM genera- were excluded from this analysis due to a limited number of high-quality genomes passing our completeness filters, despite their extensive genomic representation in the UHGG. Interestingly, the two key genes in the butyryl-CoA pathway (*bcd* and *but*) are core in *A. rectalis*. Nonetheless, some strains of this known butyrate-producing species also harbor genes for ethanol and propanol production, which our findings suggest may contribute to MASLD onset. This

449 highlights a critical point: taxonomic signatures alone are insufficient for associating bacterial clades
450 with disease. Genome plasticity introduces substantial strain-level metabolic variability in the GM,
451 leading to distinct phenotypic traits even within the same species: an effect well-documented in both
452 pathogenic and commensal *E. coli* strains[93,94].

453 Further supporting their mobility, candidate genes are frequently encoded in plasmids (Fig. 5C).
454 Among these, ethanol formation genes (*aldhB*, *yahK*, and *ahr*) are the most widely distributed, while
455 butyrate and propanol-related genes are found in plasmids less frequently. Since non-essential,
456 adaptive metabolic pathways are often encoded in plasmids across many bacterial species[94–96], their
457 presence likely drives phenotypic differences in MASLD without altering the overall species composition
458 detected by conventional metataxonomic marker genes. If MASLD is actually driven by bacterial
459 metabolism, strains from different species, but harboring similar metabolic pathways, may produce
460 similar liver-impacting metabolites, potentially contributing to MASLD pathogenesis regardless of their
461 specific taxonomic position.

462 **4. Conclusions**

463 Our study reveals that MASLD is marked by distinct geno-metabolic alterations in the GM that
464 are not captured by conventional taxonomic profiling. Specifically, we observed an increase in genes
465 involved in the production of SCAs and TMA, coupled with a depletion of genes responsible for butyrate
466 and methane production. These results confirm multiple previous findings and suggest a broader
467 metabolic shift that may contribute to the pathogenesis of MASLD, with the displacement of beneficial
468 butyrate-producing taxa such as *Agathobacter rectalis*. Importantly, our work underscores the value of
469 analyzing metagenomes at gene level to extract additional phenotypic signatures beyond what 16S
470 rRNA or marker gene-based approaches can offer. By integrating metabolic pathway analysis with the

isolation of specific gene families, and by examining gene abundance, accessory gene profiles and plasmid-mediated gene mobilization, we have significantly advanced our understanding of bacterial metabolism in MASLD. This comprehensive approach not only enhances our vision of the microbial etiology of MASLD, but also paves the way for improved bacterial analysis in clinical settings, offering new avenues for diagnosis and therapeutic intervention in liver pathologies.

Materials and Methods

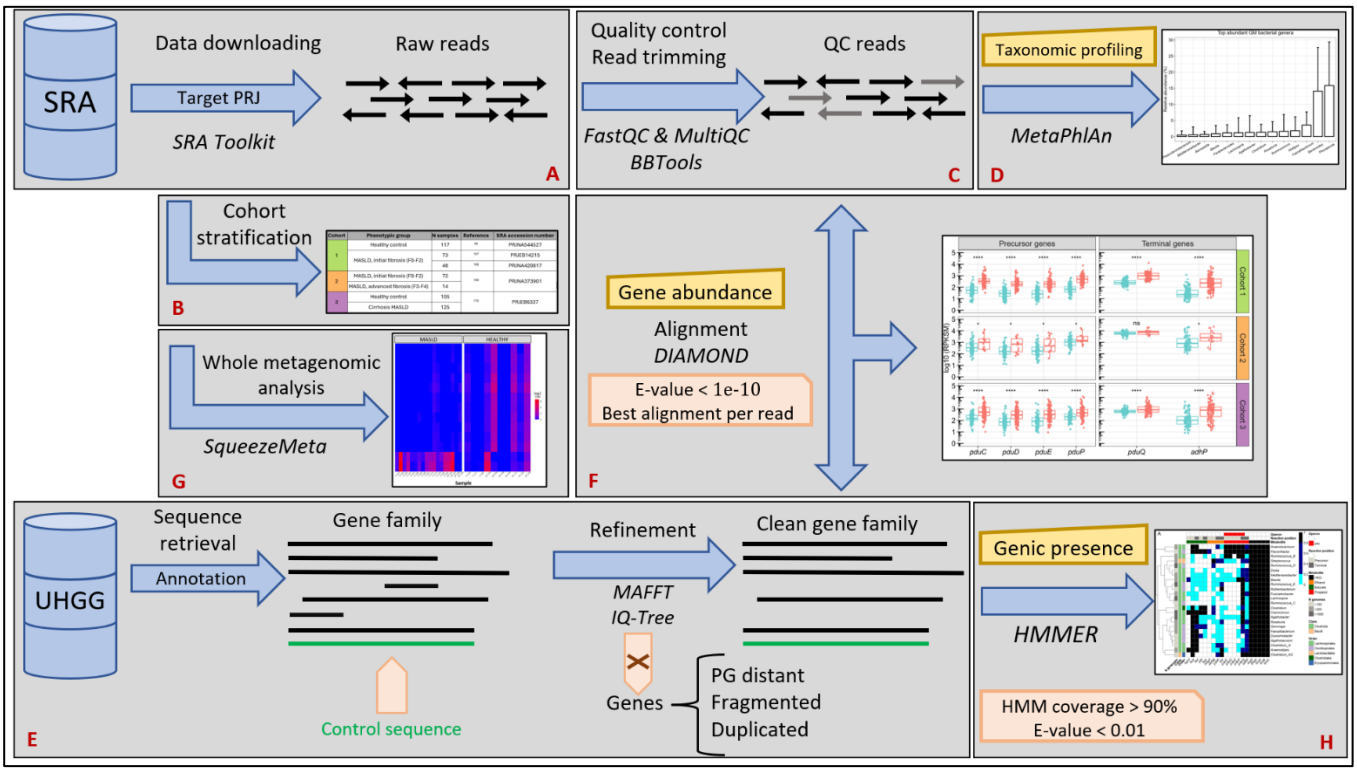


Figure 7. Computational workflow used in this study. This diagram summarizes the full computational methodology for processing whole paired-end metagenomic samples, including data retrieval from the Short Read Archive (SRA), read quality control and taxonomic profiling. Target gene families were isolated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

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

488 A total of 554 metagenomic libraries generated from the fecal samples of MASLD and healthy
489 subjects were downloaded from the SRA with SRA Toolkit´s *fastq-dump*[97]. These samples were
490 retrieved using the accession numbers corresponding to five different studies, and they were stratified
491 in three MASLD cohorts, as detailed in Table 1. Information on the collection and processing of fecal
492 samples, bacterial DNA extraction and sequencing protocols can be found in the original studies.
493 Additionally, we retrieved the medical metadata associated with each sample donor to confirm their
494 MASLD status and classify samples into two phenotypic groups of comparison within each cohort.
495 Samples lacking a clearly defined MASLD status were excluded from this study. All data was downloaded
496 between March 2022 and March 2024

Cohort	Phenotypic group	N samples	Reference	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		

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

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

503

504

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

506 Cohort 1 consists of two sets of European, obese MASLD patients aged 20-64 years with initial
507 fibrosis (F0-F2) and varying degrees of steatosis: 73 patients[100] and 10 patients[100]. In the latter group,
508 48 fecal samples were collected at five time points during the study (two samples from day 1 were not
509 obtained). Diagnoses were confirmed through hepatic biopsy performed during obesity surgery, and
510 none of the subjects had alcoholic liver disease, hepatitis or diabetes. The control group for comparison
511 includes 117 healthy patients[98]. Cohort 2 comprises 86 American MASLD patients[18], including 72 with
512 initial (F0-F2) and 14 with advanced fibrosis (F3-F4). Diagnoses were confirmed through hepatic biopsy
513 and magnetic resonance imaging, and none of the patients had alcoholic liver disease, hepatitis or
514 diabetes. Note that in this study, the initial fibrosis group is referred to as "healthy" for group
515 comparison purposes, despite including patients with early-stage MASLD. Cohort 3 includes 230 Han
516 Chinese patients[101]: 125 diagnosed with cirrhotic MASLD and 105 with no hepatic pathology. None of
517 these subjects had hepatitis. All cohorts consist of one metagenomic stool sample per patient, except
518 for the 48 samples collected at multiple time points from 10 patients[100].

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

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

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

524 To determine whether any bacterial clade is systematically increased or decreased in MASLD
525 across the three cohorts, we profiled the taxonomic composition of GM communities in fecal samples
526 using MetaPhlAn (version 4.0.6), which uses a wide-range of clade-specific maker genes to estimate

527 taxonomic abundances[24]. The diversity of these markers allows MetaPhlAn to provide highly accurate
528 taxonomic classification across a wide spectrum of bacterial clades from the species level to broader
529 taxonomic ranks, without being limited to 16S rRNA alone.

530 For this study, we used the database version *mpa_vOct22_CHOCOPhlAnSGB_202212*,
531 downloaded on August 22, 2023, and added the *--ignore_eukaryotes* flag to skip profiling eukaryotic
532 organisms. Post-processing of the MetaPhlAn output involved a) removing low-abundant clades (i.e.,
533 those with an average relative abundance of 0 in at least one cohort), b) excluding taxa predicted as
534 "unclassified", and c) verifying that excluded clades were not solely present in only one of the
535 comparison groups, in order to avoid what could be very clear taxonomic markers. To identify the most
536 abundant clades present in the GM, we selected genera with an average relative abundance above
537 0.5% (Fig. 1C) and species above 0.2% (Fig. 1E). Notedly, in this step we faced an issue regarding the
538 status of taxonomic assignments, as several species required manual updates to their nomenclature
539 assigned by MetaPhlAn. Specifically, *Eubacterium rectale* was reassigned to *Agathobacter rectalis*[87],
540 *Ruminococcus torques* to *Mediterraneibacter torques*[105] and *Roseburia faecis* to *Agathobacter faecis*[106].

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

542 We constructed families of the human GM-metabolic genes encoding the enzymes involved in the
543 production of butyrate, ethanol, propanol, TMA, methane and their precursor metabolites, as
544 represented in Figs. 2A, 3A and 4A. The Unified Human Gastrointestinal Genome (UHGG, version
545 2.0.2)[81] served as the reference database for this purpose. This resource, comprising 289,232 bacterial
546 and archaeal genomes and more than 665 million proteins, represents the most comprehensive
547 collection of characterized human GM genes to date. Each gene family was built by retrieving a seed of
548 sequences based on gene and protein annotation from the UHGG, assessing their similarity with lab-

549 validated enzymatic sequences when available. Validated gene sequences were retrieved from
550 MetaCyc[107] and used as controls to assess their phylogenetic proximity to the retrieved gene seeds.
551 Gene families were aligned using MAFFT (version 7.271)[108] and phylogenetic trees were constructed
552 with IQ-TREE (version 2.0.3)[109]. Low phylogenetic distances between control and retrieved sequences
553 served as a criterion for defining gene families. Duplicated sequences, gene fragments and
554 phylogenetically distant genes were excluded after inspecting the trees. To improve gene identification
555 and classification, we then built Hidden Markov Model (HMM) profiles for each target gene family using
556 *HMMER hmpress* (version 3.4)[110]. MetaCyc[107] and KEGG[111] were also used to validate gene and enzyme
557 names, KEGG Orthology (KOs) groups, the biochemical reactions involved in the formation of each
558 metabolite, and their enzyme commission (EC) numbers. The same protocol was applied to isolate five
559 USCGs, used as controls. These USCGs and their encoded proteins were: *argS* (arginyl-tRNA synthetase),
560 *dnaA* (chromosomal replication initiator protein DnaA), *rpoA*, *rpoB* and *rpoC* (alpha, beta and gamma
561 subunits of the DNA-directed RNA polymerase, respectively).

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

563 We quantified the abundance of target genes in the metagenomic samples from the three cohorts
564 by aligning the quality-controlled reads against the isolated gene families using DIAMOND (version
565 2.0.14)[112]. Alignments were filtered to retain only the best match for each read and gene family
566 sequence by using the *"--max-hsps 1"* option, with an E-value threshold <1E-10. These parameters were
567 based on experimentally validated recommendations for detecting genes in fecal metagenomic
568 samples with DIAMOND[113]. The total number of aligned reads for each gene family was normalized
569 using a modified RPKM formula, referred to as RPKSM (reads per kilobase per size per million reads).
570 This measure was used as an indicator of gene abundance (Figs. 2B, 3B-C and 4B-D; and Supplementary

571 Figs. 2 and 3), and it is defined as the number of aligned reads per kilobase per gene family size per
572 million reads, to account for variations in gene length, gene family size (i.e., the number of sequences
573 within each gene family) and library size, respectively.

574 ***G. Whole metagenomic analysis***

575 We applied a whole-metagenome analysis workflow to validate our results at single-gene level. To
576 this end, we used SqueezeMeta (version 1.6.0)[114] to predict both the taxonomic and functional
577 composition of the metagenomic samples from the three cohorts. Due to computational limitations, it
578 was not feasible to co-assemble all available samples, so we randomly selected approximately 24-30
579 random samples per cohort, with equal representation from each phenotypic group. Briefly,
580 SqueezeMeta assembled the metagenomes from each sub-cohort into a single co-assembly using
581 MEGAHIT (version 1.2.9)[115] by pooling the sequencing reads from every sample. ORFs were then
582 predicted from contigs[116] , and reads from individual samples were mapped back against the co-
583 assembly to quantify the abundance of each KO group per metagenome. We further processed these
584 results with SQMtools (version 1.6.0)[117] to identify the most differentially abundant KOs between
585 phenotypic groups within each cohort through DESeq2 (version 1.34.0)[118], applying a log fold-change
586 >2 and a p-adjusted value <0.05 as thresholds (Supplementary Fig. 4).

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

588 To determine whether candidate metabolic genes involved in MASLD were core or accessory across
589 the human GM, we analyzed their presence in both UHGG genomes and curated RefSeq200 plasmids[119].
590 Curation consisted in eliminating sequences that corresponded to partial plasmid DNA sequences,
591 unassignable hosts, or PacBio internal control sequences, as detailed previously[16]. Since many UHGG
592 genomes originate from metagenome-assembled or single-cell-amplified genomes, they may be

593 incomplete, potentially leading to false positive annotations. To minimize this risk, we included only
594 UHGG genomes with >95% completeness and containing the five USCGs described above. ORFs were
595 predicted using Prodigal (version 2.6.3)[120], retaining only those >90% of the HMM size. We then used
596 HMMER *hmmscan* with the following parameters: protein-pHMM coverage threshold >90%, E-value
597 <0.01 and independent E-value (i-E value) <0.01, strictly enforcing the 90% threshold to avoid capturing
598 ORFs with similar protein domains. This approach yielded 31,227 genomes classified as complete and
599 containing the five USCGs.

600 To classify genes as core or accessory, we defined their presence as the percentage of strains within
601 a given genus and species that encode the gene, considering only the most abundant clades in the GM
602 with >100 genomes meeting the completeness criteria. Genes were categorized as core if present in
603 >80% of congeneric (Fig. 5A) or conspecific (Fig. 5B) genomes, as accessory if present in 20-80%, and as
604 highly accessory if present in <20% genomes. Additionally, we assessed whether these genes could be
605 encoded in annotated plasmids. Using the same constraints, we predicted ORFs from the RefSeq200
606 plasmid collection, which includes 23,309 plasmids encoding over 2 million ORFs. ORFS were retained
607 if they met the >90% HMM size threshold, with *hmmscan* parameters identical to those applied in the
608 UHGG genome analysis (Fig. 5C).

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

610 Data manipulation was performed using R (version 4.1.3) and the *tidyverse* package (version
611 1.3.2)[121]. Figures were generated with the R packages *ggplot2* (version 3.4.2)[122], *ggpubr* (version
612 0.6.0)[123] and *pheatmap* (version 1.0.12)[124]. Mann-Whitney tests with a Benjamini-Hochberg adjustment
613 for multiple testing were applied on the relative abundance of bacterial clades (Figs. 1D and 1E) and
614 genic abundance (Figs. 2B, 3B-C, 4B-D, and Supplementary Figs. 2 and 3) between phenotypic groups

615 in the metagenomic fecal samples of each cohort to assess the significance of the differences. The
616 statistical analysis and its graphic representation was conducted using the R package *rstatix* (version
617 0.7.2)[125].

618 **References**

619 1. Riazi, K. *et al.* The prevalence and incidence of NAFLD worldwide: a systematic review and meta-analysis.
620 *Lancet Gastroenterol. Hepatol.* **7**, 851–861 (2022).
621 2. EASL. EASL–EASD–EASO Clinical practice guidelines for the management of non-alcoholic fatty liver
622 disease. *J. Hepatol.* **64**, 1388–1402 (2016).
623 3. Allen, A. M. *et al.* Clinical course of non-alcoholic fatty liver disease and the implications for clinical trial
624 design. *J. Hepatol.* **77**, 1237–1245 (2022).
625 4. Powell, E. E., Wong, V. W.-S. & Rinella, M. Non-alcoholic fatty liver disease. *The Lancet* **397**, 2212–2224
626 (2021).
627 5. Yki-Järvinen, H. Non-alcoholic fatty liver disease as a cause and a consequence of metabolic syndrome.
628 *Lancet Diabetes Endocrinol.* **2**, 901–910 (2014).
629 6. Martín-Mateos, R. & Albillos, A. The role of the gut-liver axis in metabolic dysfunction-associated fatty liver
630 disease. *Front. Immunol.* **12**, 660179 (2021).
631 7. Carpino, G. *et al.* Increased liver localization of lipopolysaccharides in human and experimental NAFLD.
632 *Hepatology* **72**, 470–485 (2020).
633 8. Li, Q., Rempel, J. D., Yang, J. & Minuk, G. Y. The effects of pathogen-associated molecular patterns on
634 peripheral blood monocytes in patients with non-alcoholic fatty liver disease. *J. Clin. Exp. Hepatol.* **12**, 808–
635 817 (2022).
636 9. Miele, L. *et al.* Increased intestinal permeability and tight junction alterations in non-alcoholic fatty liver
637 disease†. *Hepatology* **49**, 1877–1887 (2009).

638 10. Wang, W. *et al.* Tauroursodeoxycholic acid inhibits intestinal inflammation and barrier disruption in mice
639 with non-alcoholic fatty liver disease. *Br. J. Pharmacol.* **175**, 469–484 (2018).
640 11. Kuang, J. *et al.* Hyodeoxycholic acid alleviates non-alcoholic fatty liver disease through modulating the gut-
641 liver axis. *Cell Metab.* **35**, 1752-1766.e8 (2023).
642 12. Yang, C. *et al.* Altered gut microbial profile accompanied by abnormal short chain fatty acid metabolism
643 exacerbates nonalcoholic fatty liver disease progression. *Sci. Rep.* **14**, 22385 (2024).
644 13. Zhang, R. *et al.* Gut microbial metabolites in MASLD: implications of mitochondrial dysfunction in the
645 pathogenesis and treatment. *Hepatol. Commun.* **8**, 1–21 (2024).
646 14. Del Chierico, F. *et al.* Gut microbiota profiling of pediatric non-alcoholic fatty liver disease and obese
647 patients unveiled by an integrated meta-omics-based approach. *Hepatology* **65**, 451–464 (2017).
648 15. Lee, G. *et al.* Distinct signatures of gut microbiome and metabolites associated with significant fibrosis in
649 non-obese NAFLD. *Nat. Commun.* **11**, 4982 (2020).
650 16. Redondo-Salvo, S. *et al.* Pathways for horizontal gene transfer in bacteria revealed by a global map of their
651 plasmids. *Nat. Commun.* **11**, 3602 (2020).
652 17. Lanthier, N. *et al.* Microbiota analysis and transient elastography reveal new extra-hepatic components of
653 liver steatosis and fibrosis in obese patients. *Sci. Rep.* **11**, 659 (2021).
654 18. Loomba, R. *et al.* Gut microbiome-based metagenomic signature for non-invasive detection of advanced
655 fibrosis in human non-alcoholic fatty liver disease. *Cell Metab.* **25**, 1054-1062.e5 (2017).
656 19. Boursier, J. *et al.* The severity of non-alcoholic fatty liver disease is associated with gut dysbiosis and shift in
657 the metabolic function of the gut microbiota. *Hepatology* **63**, 764–775 (2016).
658 20. Caussy, C. *et al.* A gut microbiome signature for cirrhosis due to non-alcoholic fatty liver disease. *Nat.*
659 *Commun.* **10**, 1406 (2019).
660 21. Oh, T. G. *et al.* A universal gut-microbiome-derived signature predicts cirrhosis. *Cell Metab.* **32**, 878-888.e6
661 (2020).

662 22. Shen, F. *et al.* Gut microbiota dysbiosis in patients with non-alcoholic fatty liver disease. *Hepatobiliary*
663 *Pancreat. Dis. Int.* **16**, 375–381 (2017).
664 23. Aron-Wisnewsky, J. *et al.* Gut microbiota and human NAFLD: disentangling microbial signatures from
665 metabolic disorders. *Nat. Rev. Gastroenterol. Hepatol.* **17**, 279–297 (2020).
666 24. Blanco-Míguez, A. *et al.* Extending and improving metagenomic taxonomic profiling with uncharacterized
667 species using MetaPhlAn 4. *Nat. Biotechnol.* **41**, 1633–1644 (2023).
668 25. McPherson, S. *et al.* Age as a confounding factor for the accurate non-invasive diagnosis of advanced
669 NAFLD fibrosis. *Am. J. Gastroenterol.* **112**, 740–751 (2017).
670 26. Ponziani, F. R., Gasbarrini, A. & Pompili, M. NAFLD or comorbidities, that is the question. *J. Hepatol.* **73**, 723
671 (2020).
672 27. Chu, H., Duan, Y., Yang, L. & Schnabl, B. Small metabolites, possible big changes: a microbiota-centered
673 view of non-alcoholic fatty liver disease. *Gut* **68**, 359–370 (2019).
674 28. Dai, X. *et al.* Microbial metabolites: critical regulators in NAFLD. *Front. Microbiol.* **11**, 567654 (2020).
675 29. Martin-Grau, M. & Monleón, D. The role of microbiota-related co-metabolites in MASLD progression: a
676 narrative review. *Curr. Issues Mol. Biol.* **46**, 6377–6389 (2024).
677 30. Oliphant, K. & Allen-Vercoe, E. Macronutrient metabolism by the human gut microbiome: major
678 fermentation by-products and their impact on host health. *Microbiome* **7**, 91 (2019).
679 31. Donaldson, G. P., Lee, S. M. & Mazmanian, S. K. Gut biogeography of the bacterial microbiota. *Nat. Rev.*
680 *Microbiol.* **14**, 20–32 (2016).
681 32. Albillos, A., De Gottardi, A. & Rescigno, M. The gut-liver axis in liver disease: pathophysiological basis for
682 therapy. *J. Hepatol.* **72**, 558–577 (2020).
683 33. Iruzubieta, P., Medina, J. M., Fernández-López, R., Crespo, J. & De La Cruz, F. A role for gut microbiome
684 fermentative pathways in fatty liver disease progression. *J. Clin. Med.* **9**, 1369 (2020).

685 34. Miller, T. L. & Wolin, M. J. Pathways of acetate, propionate, and butyrate formation by the human fecal
686 microbial flora. *Appl. Environ. Microbiol.* **62**, 1589–1592 (1996).
687 35. Koh, A., De Vadder, F., Kovatcheva-Datchary, P. & Bäckhed, F. From dietary fiber to host physiology: short-
688 chain fatty acids as key bacterial metabolites. *Cell* **165**, 1332–1345 (2016).
689 36. Morrison, D. J. & Preston, T. Formation of short chain fatty acids by the gut microbiota and their impact on
690 human metabolism. *Gut Microbes* **7**, 189–200 (2016).
691 37. Blaak, E. E. *et al.* Short chain fatty acids in human gut and metabolic health. *Benef. Microbes* **11**, 411–455
692 (2020).
693 38. Donohoe, D. R. *et al.* The microbiome and butyrate regulate energy metabolism and autophagy in the
694 mammalian colon. *Cell Metab.* **13**, 517–526 (2011).
695 39. Wang, H.-B., Wang, P.-Y., Wang, X., Wan, Y.-L. & Liu, Y.-C. Butyrate enhances Intestinal epithelial barrier
696 function via up-regulation of tight junction protein claudin-1 transcription. *Dig. Dis. Sci.* **57**, 3126–3135
697 (2012).
698 40. Gao, Z. *et al.* Butyrate improves insulin sensitivity and increases energy expenditure in mice. *Diabetes* **58**,
699 1509–1517 (2009).
700 41. Li, Z. *et al.* Butyrate reduces appetite and activates brown adipose tissue via the gut-brain neural circuit.
701 *Gut* **67**, 1269–1279 (2018).
702 42. De Vadder, F. *et al.* Microbiota-generated metabolites promote metabolic benefits via gut-brain neural
703 circuits. *Cell* **156**, 84–96 (2014).
704 43. Sanna, S. *et al.* Causal relationships among the gut microbiome, short-chain fatty acids and metabolic
705 diseases. *Nat. Genet.* **51**, 600–605 (2019).
706 44. Endo, H., Niioka, M., Kobayashi, N., Tanaka, M. & Watanabe, T. Butyrate-producing probiotics reduce non-
707 alcoholic fatty liver disease progression in rats: new insight into the probiotics for the gut-liver axis. *PLoS*
708 *ONE* **8**, e63388 (2013).

709 45. Jin, C. J., Sellmann, C., Engstler, A. J., Ziegenhardt, D. & Bergheim, I. Supplementation of sodium butyrate
710 protects mice from the development of non-alcoholic steatohepatitis (NASH). *Br. J. Nutr.* **114**, 1745–1755
711 (2015).
712 46. Zhou, D. *et al.* Sodium butyrate attenuates high-fat diet-induced steatohepatitis in mice by improving gut
713 microbiota and gastrointestinal barrier. *World J. Gastroenterol.* **23**, 60 (2017).
714 47. Zhou, D. *et al.* Sodium butyrate reduces high-fat diet-induced non-alcoholic steatohepatitis through
715 upregulation of hepatic GLP-1R expression. *Exp. Mol. Med.* **50**, 1–12 (2018).
716 48. Zhou, D. *et al.* Total fecal microbiota transplantation alleviates high-fat diet-induced steatohepatitis in mice
717 via beneficial regulation of gut microbiota. *Sci. Rep.* **7**, 1529 (2017).
718 49. Louis, P. & Flint, H. J. Formation of propionate and butyrate by the human colonic microbiota. *Environ.*
719 *Microbiol.* **19**, 29–41 (2017).
720 50. Zhu, L. *et al.* Characterization of gut microbiomes in non-alcoholic steatohepatitis (NASH) patients: a
721 connection between endogenous alcohol and NASH. *Hepatology* **57**, 601–609 (2013).
722 51. Michail, S. *et al.* Altered gut microbial energy and metabolism in children with non-alcoholic fatty liver
723 disease. *FEMS Microbiol. Ecol.* **91**, 1–9 (2015).
724 52. Meijnikman, A. S. *et al.* Microbiome-derived ethanol in non-alcoholic fatty liver disease. *Nat. Med.* **28**,
725 2100–2106 (2022).
726 53. Yuan, J. *et al.* Fatty liver disease caused by high-alcohol-producing *Klebsiella pneumoniae*. *Cell Metab.* **30**,
727 675-688.e7 (2019).
728 54. Chen, X. *et al.* Endogenous ethanol produced by intestinal bacteria induces mitochondrial dysfunction in
729 non-alcoholic fatty liver disease. *J. Gastroenterol. Hepatol.* **35**, 2009–2019 (2020).
730 55. Fromenty, B. & Roden, M. Mitochondrial alterations in fatty liver diseases. *J. Hepatol.* **78**, 415–429 (2023).
731 56. Ruuskanen, M. O. *et al.* Links between gut microbiome composition and fatty liver disease in a large
732 population sample. *Gut Microbes* **13**, 1888673 (2021).

733 57. Malaguarnera, G. Gut microbiota in alcoholic liver disease: pathogenetic role and therapeutic perspectives.
734 *World J. Gastroenterol.* **20**, 16639 (2014).
735 58. Schut, G. J. *et al.* Tungsten enzymes play a role in detoxifying food and antimicrobial aldehydes in the
736 human gut microbiome. *Proc. Natl. Acad. Sci.* **118**, e2109008118 (2021).
737 59. Pony, P., Rapisarda, C., Terradot, L., Marza, E. & Fronzes, R. Filamentation of the bacterial bi-functional
738 alcohol/aldehyde dehydrogenase AdhE is essential for substrate channeling and enzymatic regulation. *Nat.*
739 *Commun.* **11**, 1426 (2020).
740 60. Hitschler, L., Nissen, L. S., Kuntz, M. & Basen, M. Alcohol dehydrogenases AdhE and AdhB with broad
741 substrate ranges are important enzymes for organic acid reduction in *Thermoanaerobacter sp.* strain
742 X514. *Biotechnol. Biofuels* **14**, 187 (2021).
743 61. Pick, A., Rühmann, B., Schmid, J. & Sieber, V. Novel CAD-like enzymes from *Escherichia coli* K-12 as
744 additional tools in chemical production. *Appl. Microbiol. Biotechnol.* **97**, 5815–5824 (2013).
745 62. Merchel Piovesan Pereira, B., Adil Salim, M., Rai, N. & Tagkopoulos, I. Tolerance to glutaraldehyde in
746 *Escherichia coli* mediated by overexpression of the aldehyde reductase YqhD by YqhC. *Front. Microbiol.* **12**,
747 680553 (2021).
748 63. Moreira De Gouveia, M. I., Daniel, J., Garrivier, A., Bernalier-Donadille, A. & Jubelin, G. Diversity of
749 ethanolamine utilization by human commensal *Escherichia coli*. *Res. Microbiol.* **174**, 103989 (2023).
750 64. Dellomonaco, C., Clomburg, J. M., Miller, E. N. & Gonzalez, R. Engineered reversal of the β-oxidation cycle
751 for the synthesis of fuels and chemicals. *Nature* **476**, 355–359 (2011).
752 65. Eram, M. & Ma, K. Decarboxylation of pyruvate to acetaldehyde for ethanol production by
753 hyperthermophiles. *Biomolecules* **3**, 578–596 (2013).
754 66. Ho, K. K. & Weiner, H. Isolation and characterization of an aldehyde dehydrogenase encoded by the *aldB*
755 gene of *Escherichia coli*. *J. Bacteriol.* **187**, 1067–1073 (2005).

756 67. Ravcheev, D. A., Moussu, L., Smajic, S. & Thiele, I. Comparative genomic analysis reveals novel
757 microcompartment-associated metabolic pathways in the human gut microbiome. *Front. Genet.* **10**, 636
758 (2019).
759 68. Dank, A. *et al.* Bacterial microcompartment-dependent 1,2-propanediol utilization of *Propionibacterium*
760 *freudenreichii* . *Front. Microbiol.* **12**, 679827 (2021).
761 69. Chen, Y. *et al.* Associations of gut-flora-dependent metabolite trimethylamine-N-oxide, betaine and choline
762 with non-alcoholic fatty liver disease in adults. *Sci. Rep.* **6**, 19076 (2016).
763 70. Flores-Guerrero, J. L. *et al.* Circulating trimethylamine-*N*-oxide is associated with all-cause mortality in
764 subjects with non-alcoholic fatty liver disease. *Liver Int.* **41**, 2371–2382 (2021).
765 71. Wang, Z. *et al.* Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. *Nature* **472**,
766 57–63 (2011).
767 72. Roncal, C. *et al.* Trimethylamine-N-oxide (TMAO) predicts cardiovascular mortality in peripheral artery
768 disease. *Sci. Rep.* **9**, 15580 (2019).
769 73. Martínez-del Campo, A. *et al.* Characterization and detection of a widely distributed gene cluster that
770 predicts anaerobic choline utilization by human gut bacteria. *mBio* **6**, e00042-15 (2015).
771 74. Goris, M. *et al.* Increased thermostability of an engineered flavin-containing monooxygenase to remediate
772 trimethylamine in fish protein hydrolysates. *Appl. Environ. Microbiol.* **89**, e00390-23 (2023).
773 75. Leimkühler, S. The biosynthesis of the molybdenum cofactors in *Escherichia coli*. *Environ. Microbiol.* **22**,
774 2007–2026 (2020).
775 76. Kurth, J. M. *et al.* Methanogenic archaea use a bacteria-like methyltransferase system to demethoxylate
776 aromatic compounds. *ISME J.* **15**, 3549–3565 (2021).
777 77. Gendron, A. & Allen, K. D. Overview of diverse methyl/alkyl-coenzyme M reductases and considerations for
778 their potential heterologous expression. *Front. Microbiol.* **13**, 867342 (2022).

779 78. Thauer, R. K. Methyl (alkyl)-coenzyme M reductases: nickel F-430-containing enzymes involved in
780 anaerobic methane formation and in anaerobic oxidation of methane or of short chain alkanes.
781 *Biochemistry* **58**, 5198–5220 (2019).
782 79. Wagner, T., Koch, J., Ermler, U. & Shima, S. Methanogenic heterodisulfide reductase (HdrABC-MvhAGD)
783 uses two noncubane [4Fe-4S] clusters for reduction. *Science* **357**, 699–703 (2017).
784 80. Aziz, I. *et al.* Structural and mechanistic basis of the central energy-converting methyltransferase complex
785 of methanogenesis. *Proc. Natl. Acad. Sci.* **121**, e2315568121 (2024).
786 81. Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat.*
787 *Biotechnol.* **39**, 105–114 (2021).
788 82. Rosconi, F. *et al.* A bacterial pan-genome makes gene essentiality strain-dependent and evolvable. *Nat.*
789 *Microbiol.* **7**, 1580–1592 (2022).
790 83. Yu, M. K., Fogarty, E. C. & Eren, A. M. Diverse plasmid systems and their ecology across human gut
791 metagenomes revealed by PlasX and MobMess. *Nat. Microbiol.* **9**, 830–847 (2024).
792 84. O'Callaghan, J. L., Willner, D., Buttini, M., Huygens, F. & Pelzer, E. S. Limitations of 16S rRNA gene
793 sequencing to characterize *Lactobacillus* species in the upper genital tract. *Front. Cell Dev. Biol.* **9**, 641921
794 (2021).
795 85. Tomita, S., Saito, K., Nakamura, T., Sekiyama, Y. & Kikuchi, J. Rapid discrimination of strain-dependent
796 fermentation characteristics among *Lactobacillus* strains by NMR-based metabolomics of fermented
797 vegetable juice. *PLOS ONE* **12**, e0182229 (2017).
798 86. Forster, A. H. & Gescher, J. Metabolic engineering of *Escherichia coli* for production of mixed-acid
799 fermentation end products. *Front. Bioeng. Biotechnol.* **2**, (2014).
800 87. Rosero, J. A. *et al.* Reclassification of *Eubacterium rectale* (Hauduroy *et al.* 1937) in a new genus
801 *Agathobacter* gen. nov. as *Agathobacter rectalis* comb. nov., and description of *Agathobacter ruminis* sp.
802 nov., isolated from the rumen contents of sheep and cows. *Int. J. Syst. Evol. Microbiol.* **66**, 768–773 (2016).

803 88. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the
804 tree of life. *Nat. Biotechnol.* **36**, 996–1004 (2018).
805 89. Parks, D. H. *et al.* GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically
806 consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* **50**, D785–D794
807 (2022).
808 90. Medina, J. M., Fernández-López, R., Crespo, J. & Cruz, F. D. L. Propionate fermentative genes of the gut
809 microbiome decrease in inflammatory bowel disease. *J. Clin. Med.* **10**, 2176 (2021).
810 91. Louis, P., McCrae, S. I., Charrier, C. & Flint, H. J. Organization of butyrate synthetic genes in human colonic
811 bacteria: phylogenetic conservation and horizontal gene transfer. *FEMS Microbiol. Lett.* **269**, 240–247
812 (2007).
813 92. Cui, Y. *et al.* Production of butyl butyrate from lignocellulosic biomass through *Escherichia coli-Clostridium*
814 *beijerinckii* G117 co-culture. *Process Biochem.* **128**, 58–67 (2023).
815 93. Gonzalez-Alba, J. M., Baquero, F., Cantón, R. & Galán, J. C. Stratified reconstruction of ancestral *Escherichia*
816 *coli* diversification. *BMC Genomics* **20**, 936 (2019).
817 94. Palomino, A. *et al.* Metabolic genes on conjugative plasmids are highly prevalent in *Escherichia coli* and can
818 protect against antibiotic treatment. *ISME J.* **17**, 151–162 (2023).
819 95. Abriouel, H. *et al.* New insights into the role of plasmids from probiotic *Lactobacillus pentosus* MP-10 in
820 Aloreña table olive brine fermentation. *Sci. Rep.* **9**, 10938 (2019).
821 96. Pudlo, N. A. *et al.* Phenotypic and genomic diversification in complex carbohydrate-degrading human gut
822 bacteria. *mSystems* **7**, e00947-21 (2022).
823 97. Leinonen, R., Sugawara, H., Shumway, M., & on behalf of the International Nucleotide Sequence Database
824 Collaboration. The Sequence Read Archive. *Nucleic Acids Res.* **39**, D19–D21 (2011).
825 98. Poyet, M. *et al.* A library of human gut bacterial isolates paired with longitudinal multiomics data enables
826 mechanistic microbiome research. *Nat. Med.* **25**, 1442–1452 (2019).

827 99. Hoyles, L. *et al.* Molecular phenomics and metagenomics of hepatic steatosis in non-diabetic obese
828 women. *Nat. Med.* **24**, 1070–1080 (2018).
829 100. Mardinoglu, A. *et al.* An integrated understanding of the rapid metabolic benefits of a carbohydrate-
830 restricted diet on hepatic steatosis in humans. *Cell Metab.* **27**, 559-571.e5 (2018).
831 101. Qin, N. *et al.* Alterations of the human gut microbiome in liver cirrhosis. *Nature* **513**, 59–64 (2014).
832 102. Bushnell, B. BBMap. (2014).
833 103. Andrews, S. FastQC: A quality control tool for high throughput sequence data. (2010).
834 104. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools
835 and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
836 105. Togo, A. H. *et al.* Description of *Mediterraneibacter massiliensis*, gen. nov., sp. nov., a new genus
837 isolated from the gut microbiota of an obese patient and reclassification of *Ruminococcus faecis*,
838 *Ruminococcus lactaris*, *Ruminococcus torques*, *Ruminococcus gnavus* and *Clostridium glycyrrhizinilyticum*
839 as *Mediterraneibacter faecis* comb. nov., *Mediterraneibacter lactaris* comb. nov., *Mediterraneibacter*
840 *torques* comb. nov., *Mediterraneibacter gnavus* comb. nov. and *Mediterraneibacter glycyrrhizinilyticus*
841 comb. nov. *Antonie Van Leeuwenhoek* **111**, 2107–2128 (2018).
842 106. Murotomi, K., Tourlousse, D. M., Hamajima, M. & Sekiguchi, Y. Complete genome sequence of
843 *Roseburia faecis* M72/1^{T}. *Microbiol. Resour. Announc.* **14**, e01253-24 (2025).
844 107. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes - a 2019 update. *Nucleic*
845 *Acids Res.* **48**, D445–D453 (2020).
846 108. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in
847 performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
848 109. Minh, B. Q. *et al.* IQ-TREE 2: new models and efficient methods for phylogenetic inference in the
849 genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
850 110. Eddy, S. HMMER. (2019).

851 111. Kanehisa, M., Furumichi, M., Sato, Y., Matsuura, Y. & Ishiguro-Watanabe, M. KEGG: biological systems
852 database as a model of the real world. *Nucleic Acids Res.* **53**, D672–D677 (2025).
853 112. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND.
854 *Nat. Methods* **18**, 366–368 (2021).
855 113. Treiber, M. L., Taft, D. H., Korf, I., Mills, D. A. & Lemay, D. G. Pre- and post-sequencing
856 recommendations for functional annotation of human fecal metagenomes. *BMC Bioinformatics* **21**, 74
857 (2020).
858 114. Tamames, J. & Puente-Sánchez, F. SqueezeMeta, a highly portable, fully automatic metagenomic
859 analysis pipeline. *Front. Microbiol.* **9**, 3349 (2019).
860 115. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for
861 large and complex metagenomics assembly via succinct *de Bruijn* graph. *Bioinformatics* **31**, 1674–1676
862 (2015).
863 116. Hyatt, D., LoCascio, P. F., Hauser, L. J. & Uberbacher, E. C. Gene and translation initiation site prediction
864 in metagenomic sequences. *Bioinformatics* **28**, 2223–2230 (2012).
865 117. Puente-Sánchez, F., García-García, N. & Tamames, J. SQMtools: automated processing and visual
866 analysis of ’omics data with R and anvi’o. *BMC Bioinformatics* **21**, 358 (2020).
867 118. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq
868 data with DESeq2. *Genome Biol.* **15**, 550 (2014).
869 119. Haft, D. H. *et al.* RefSeq and the prokaryotic genome annotation pipeline in the age of metagenomes.
870 *Nucleic Acids Res.* **52**, D762–D769 (2024).
871 120. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC*
872 *Bioinformatics* **11**, 119 (2010).
873 121. Wickham, H. *et al.* Welcome to the Tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).
874 122. Wickham, H. Ggplot2*: Elegant Graphics for Data Analysis*. (Springer-Verlag New York, 2016).

123. Kassambara, A. *ggpubr*: *ggplot2* based publication ready plots. (2023).

124. Kolde, R. *pheatmap*: pretty heatmaps. (2019).

125. Kassambara, A. *rstatix*: pipe-friendly framework for basic statistical tests. (2023).