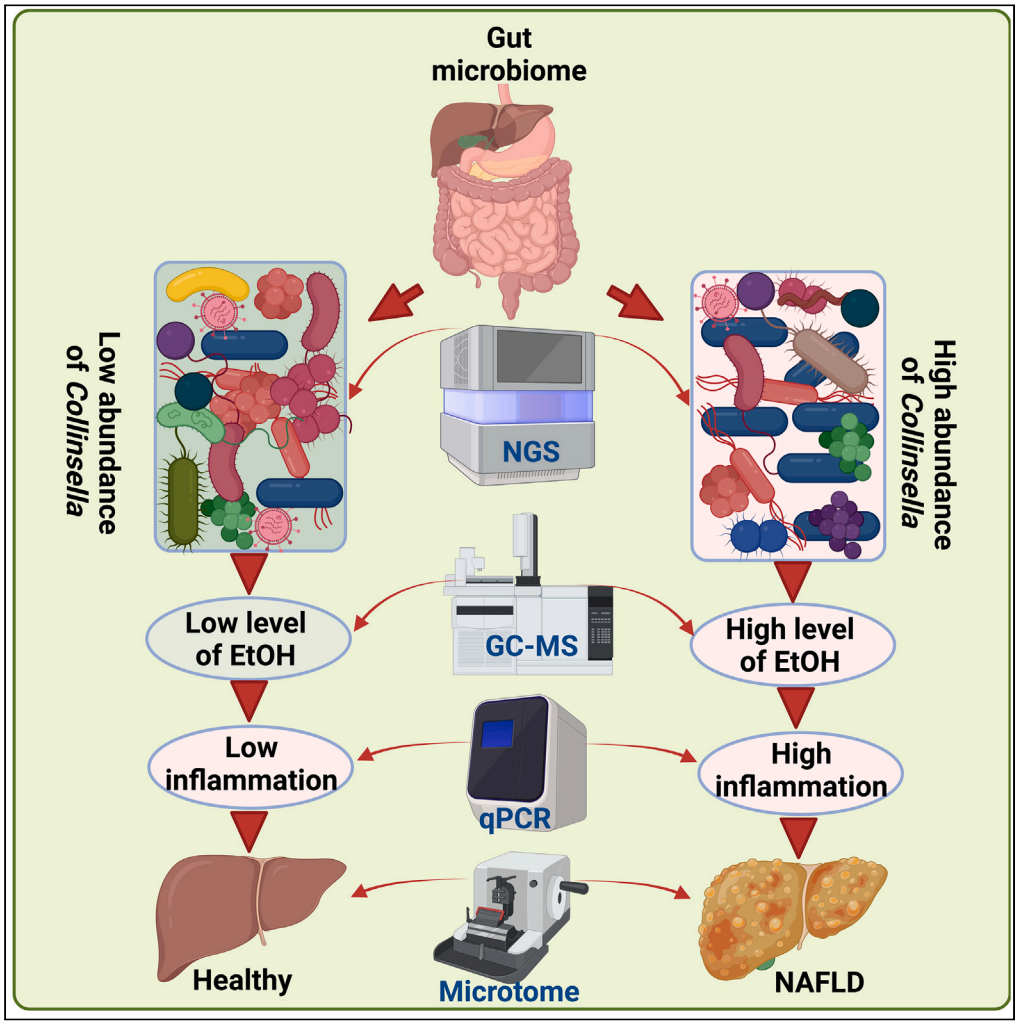
iScience

CellPress
OPEN ACCESS

Article

Collinsella aerofaciens linked with increased ethanol production and liver inflammation contribute to the pathophysiology of NAFLD



Ayushi Purohit, Bharti Kandiyal, Shakti Kumar, ..., Sanjay K. Banerjee, Shalimar, Bhabatosh Das

sanjayk.banerjee@niperguwahati.ac.in (S.K.B.)
drshalimar@gmail.com (S.)
bhabatosh@thsti.res.in (B.D.)

Highlights

We study the gut microbiome of NAFLD, obese, and healthy subjects

We observed a higher abundance of *Collinsella aerofaciens* in obese and NASH patients

C. aerofaciens contributes to an increase in the level of systemic ethanol in mice

Increased ethanol, hydroxyproline, and triglycerides induce hepatic inflammation

Purohit et al., iScience *27*, 108764
February 16, 2024 © 2023 The Author(s).
https://doi.org/10.1016/j.isci.2023.108764

iScience

CellPress
OPEN ACCESS

Article

Collinsella aerofaciens linked with increased ethanol production and liver inflammation contribute to the pathophysiology of NAFLD

Ayushi Purohit,[1] Bharti Kandiyal,[1] Shakti Kumar,[1] Agila Kumari Pragasam,[1] Parul Kamboj,[1] Daizee Talukdar,[1] Jyoti Verma,[1] Vipin Sharma,[1] Soumalya Sarkar,[1] Dinesh Mahajan,[1] Rajni Yadav,[2] Riya Ahmed,[3] Ranjan Nanda,[3] Madhu Dikshit,[1] Sanjay K. Banerjee,[4,*] Shalimar,[5,*] and Bhabatosh Das[1,6,*]

SUMMARY

Non-alcoholic fatty liver disease (NAFLD) is an emerging global health problem and a potential risk factor for metabolic diseases. The bidirectional interactions between liver and gut made dysbiotic gut microbiome one of the key risk factors for NAFLD. In this study, we reported an increased abundance of *Collinsella aerofaciens* in the gut of obese and NASH patients living in India. We isolated *C. aerofaciens* from the fecal samples of biopsy-proven NASH patients and observed that their genome is enriched with carbohydrate metabolism, fatty acid biosynthesis, and pro-inflammatory functions and have the potency to increase ethanol level in blood. An animal study indicated that mice supplemented with *C. aerofaciens* had increased levels of circulatory ethanol, high levels of hepatic hydroxyproline, triglyceride, and inflammation in the liver. The present findings indicate that perturbation in the gut microbiome composition is a key risk factor for NAFLD.

INTRODUCTION

Non-alcoholic fatty liver disease (NAFLD), a metabolic disorder in the liver characterized by increased fat accumulation, inflammation, and the development of cirrhosis, is the most common liver disease globally, accounting for 2% of all fatalities.[1] Several studies have been conducted to better understand the function of the gastrointestinal microbiota in the development and progression of NAFLD.[2] The gastrointestinal tract of healthy humans houses complex and diverse microbial communities consisting of bacteria, archaea, fungi, protists, and viruses.[3] In the last 15 years, the roles of gut microbiota in health and diseases including malnutrition,[4] type 2 diabetes,[5,6] obesity,[7] inflammatory bowel disease[8,9] cancers,[10,11] cardiovascular disease, and liver disease[12] have been demonstrated persuasively. Changes in gut microbiota diversity, abundance, and functional potency increase the risk of metabolic disorders by providing excess metabolic functions or secreting signaling molecules, which may affect host physiology by modulating gut permeability,[13] inflammation,[14] signal transduction,[15] and gene expression.[16] Certain metabolites produced or derivatized by gut microbiota, like ethanol,[17,18] short-chain fatty acids,[19] trimethylamine N-oxide,[20] tryptophan and indole derivatives,[21] branched-chain amino acids,[22] and secondary bile acids,[23] have been associated with the pathophysiology and progression of different metabolic disorders including NAFLD. Several recent studies demonstrated that endogenous ethanol and triglyceride produced by gut fungi like *Pichia kudriavzevii*, *Candida albicans*, and *Candida glabrata* yeasts substantially contribute in the pathophysiology of NASH.[24,25] Several animal experiments have shown that these microbial derived metabolites modulate host signaling system related to the metabolic functions and immune responses and increases gut permeability and hepatotoxicity.[26] Nevertheless, transplantation of gut microbiota into mice from patients can manifest similar metabolic phenotypes, suggesting the gut microbiota may lead to disease development.[27] In addition, the success of use of antibiotics and phage therapy in reducing disease severity further supported the role of the gut microbiota in pathogenesis.[28,29]

In the present study, we adopted multiple approaches, including metagenomics, isolate bacterial genomics, gas chromatography-mass spectrometry (GC-MS), and a combination of *in vivo* experiments coupled with functional genomic analysis, to study the role of *C. aerofaciens* in NAFLD. We compared the diversity and abundance of *C. aerofaciens* in healthy, obese, and NASH subjects and isolated the bacterium from the fecal samples of NASH patients and healthy subjects. Whole genome sequencing and comparative genomics of different isolates help us to identify NASH-specific genomic signatures in *C. aerofaciens*. The predicted functions and their roles in ethanol production,

[1]Translational Health Science and Technology Institute, NCR Biotech Science Cluster, Faridabad 121004, India
[2]Department of Pathology, All India Institute of Medical Sciences, New Delhi 110029, India
[3]Translational Health Group, International Centre for Genetic Engineering and Biotechnology, New Delhi 110067, India
[4]Department of Biotechnology, National Institute of Pharmaceutical Education and Research (NIPER-Guwahati), Changsari, Guwahati, Assam 781101, India
[5]Department of Gastroenterology and Human Nutrition, All India Institute of Medical Sciences, New Delhi 110029, India
[6]Lead contact
*Correspondence: sanjayk.banerjee@niperguwahati.ac.in (S.K.B.), drshalimar@gmail.com (S.), bhabatosh@thsti.res.in (B.D.)
https://doi.org/10.1016/j.isci.2023.108764

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Table 1. Characteristics of the study subjects

Characteristics	Healthy	Obese	NASH
Mean age (sd)	36 ± 8.1 years	34.7 ± 3.8 years	37 ± 3.1 years
Sex (Male/Female)	10 males 10 females	9 males 8 females	5 males 5 females
Diet	Veg/Eggetarian/non-veg	Veg/Eggetarian/non-veg	Veg/Eggetarian/non-veg
BMI (kg/m^2)	Not found	27.9 ± 2.2	28.85 ± 4.5
Living areas	Urban Ballabgarh	NCR Delhi	NCR Delhi
Any other disease pathology	No	Biopsy-proven NAFLD without NASH	Biopsy-proven NASH

inflammation, and hepatic hydroxyproline and triglyceride accumulation were validated in the C5BL6/J mice models. The current findings may provide a lead for diagnostic and therapeutic research to tackle NASH and its disease-related complications.

RESULTS

Patient's recruitment and sample collection

In this study, 47 individuals, including healthy people (n = 20), obese patients (n = 17), and NASH (n = 10) patients of both genders (male and female) between the ages of 30 and 70 years who visited the All India Institute of Medical Sciences, New Delhi, were selected for decoding gut microbiome composition and diversity. These patients are inhabitants of the national capital region (NCR) of India. The NCR area in India is centered on the National Capital Territory (NCT) of Delhi. It includes Delhi and several adjacent districts in Haryana, Uttar Pradesh, and Rajasthan. NASH (NAS score ≥5) was diagnosed based on the liver biopsy. The microbiota in fecal samples was compared between the healthy controls[30] and NASH patients. The details of the included patients are provided in Table 1.

Obese Indians have increased abundance of *Collinsella* in their gut

We implemented 16S rRNA amplicon-based microbiome analysis utilizing QIIME2 across three cohorts [healthy (n = 20), obese (n = 17), and NASH (n = 10)] and found that 3,306; 2,864; and 530 feature IDs were detected in three cohorts, respectively. The distribution of these feature IDs was further investigated at the OTU (Table 2), phylum, and species levels. (Tables S1, S2, and S3). The core OTUs are the ones that are present in 80% of the samples. The number of core OTUs in healthy subjects was 17, in obese 44, and in NASH 56 (Tables S4, S5, and S6). It has been postulated that these frequently occurring organisms that appear in all assemblages associated with a particular habitat are probably essential for the functioning of a particular kind of niche, so identifying the core species (or operational taxonomic units, OTUs) is critical in untangling the ecology of microbial consortia. The taxonomic distribution revealed that NASH subjects had the minimum OTUs yet the maximum genera in the core OTU among the three cohorts. The unique property of the microbiome is characterizing the diversity indices, which are widespread and diverse across three groups. Furthermore, when alpha diversity indices were assessed, substantial changes in all diversity indices were detected (i.e., Shannon, Simpson and Chao1). The most noticeable change occurred in Chao1 diversity, which was significantly increasing in all three groups (Figure 1A). The read summary (Table 3) and this analysis inferred the presence of rare species among the three groups. This inference is supported when the core OTUs of three groups were compared (Figure 1B). The gut microbiome of NASH and obese groups had the highest number of OTUs that have been exclusively found in the two groups (Tables S7, S8, and S9). There were twelve OTUs that were common among core OTUs. Hence, these twelve can be considered true core OTUs, and they belong to twelve different species (Table S10).

Among the twelve core OTUs, *C. aerofaciens* was the most abundant species when median-based distribution was calculated (Figures 1C and 1D). The statistical analysis revealed a significant differential abundance of only two species, i.e., *C. aerofaciens* and *Fusicatenibacter saccharivorans* (Figure 1E). Because *C. aerofaciens* has the highest abundance and it is gradually increased from healthy to NASH, it inspired us to

Table 2. OTU distribution in three cohorts

Group	Sample number	Number of feature ID	Number of OTU^b	P^q: P^q_{80}	C^q: C^q_{80}	O^q: O^q_{80}	F^q: F^q_{80}	G^b: G^b_{80}	S^b: S^b_{80}
Healthy	20	3306	364	11:4	16:5	39:8	77:10	172:15	364:17
Obese	17	2864	389	10:4	15:8	43:13	76:19	175:31	389:44
NASH	10	530	245	9:4	14:7	36:13	59:20	129:38	245:56

P^q, C^q, O^q, F^q = predicted Phylum, Class, Order, and Family by QIIME2 program, respectively.

P^q_{80}, C^q_{80}, O^q_{80}, F^q_{80} = predicted Phylum Class, Order, and Family by QIIME2 program, respectively, that have been observed in 80% of samples of each group.

G^b and S^b = genus and species predicted by nucleotide BLAST program search against 16S rRNA database with query coverage ≥95% and sequence identity ≥99%.

G^b_{80} and S^b_{80} = predicted genus and species by nucleotide BLAST program search against 16S rRNA database with query coverage ≥95% and sequence identity ≥99% that present in 80% of the samples of each group.

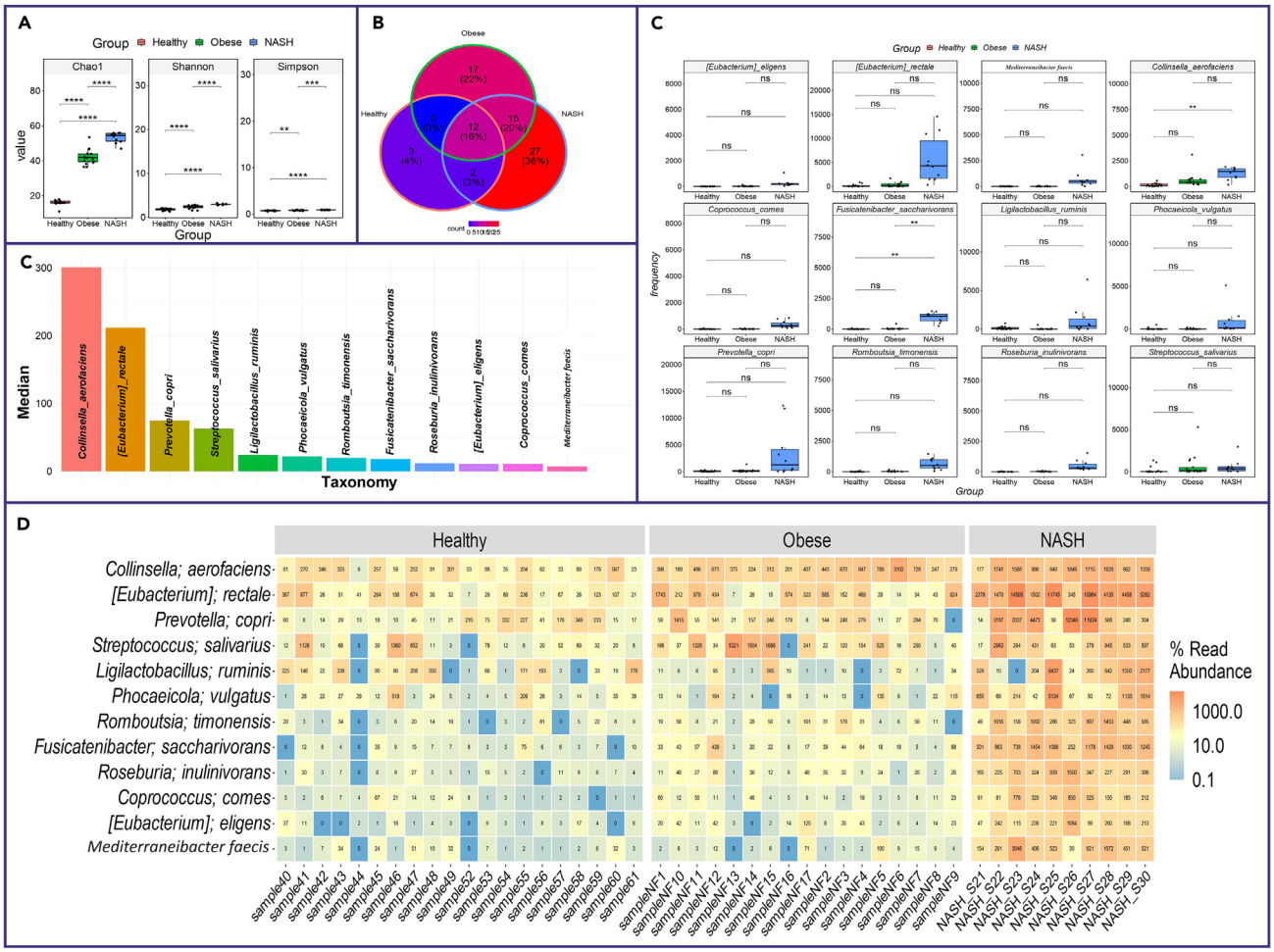


Figure 1. Diversity measurement and statistical analysis of the gut microbiota in healthy, obese, and NASH subjects

(A) Statistical comparison of alpha diversity indices (i.e., Chao1, Shannon, and Simpson). The adjust p values for multiple comparisons have been performed by Hochberg method. The adjust p values have been indicated by ****, ***, **, and * that signify $\leq$0.00005, $\leq$0.0005, $\leq$0.005, and $\leq$0.05.

(B) Venn Diagram depiction of 80% core OTUs in the Healthy, Obese, and NASH groups. Each group is shown by a circle that has been surrounded by fade red, green, and blue circles, respectively. The distribution of OTUs numbers in common and exclusive part of Healthy, Obese, and NASH groups has been shown in blue to red color.

(C) The median distribution of commonly identified twelve species in bar plot. The highest to lowest median values have been arranged in decreasing order from left to right side.

(D) The distribution of OTU numbers of 80% core OTUs in Healthy, Obese, and NASH groups in heatmap plot.

(E) The distribution and statistical comparison of commonly identified 12 species from the 80% core OTUS of each three groups in boxplot. The statistical comparison has been shown by adjust p values. The adjust p values for multiple comparisons has been performed by Hochberg method. The adjust p values have been indicated by ****, ***, **, and * that signify $\leq$0.00005, $\leq$0.0005, $\leq$0.005, and $\leq$0.05.

look further into its genome for functional potency. In addition, the increased abundance of *C. aerofaciens* observed in the 16S rRNA-based metagenomic study was further verified by quantitative PCR (qPCR) analysis using *C. aerofaciens*-specific primers and fecal genomic DNA obtained from the NASH patients (n = 10) and healthy subjects (n = 10). The results of the qPCR also indicated a higher abundance of *C. aerofaciens* in NASH patients than in healthy subjects (Figure 2). In order to identify the association between core bacterial taxa (n = 12), we created a cooccurrence network (Pearson's correlation, with a cutoff r $\geq$ 3, p value<0.05) and compared it with the healthy subjects and NASH patients. In healthy population, five significant bacterial taxa showed positive associations with health (Table S11; Figures S1A and S1B). The microbial intra-taxa interaction was observed between *C. aerofaciens* and *Mediterraneibacter faecis* and *Eubacterium rectale* and *Roseburia inulinivorans*. *Streptococcus salivarius* was positively associated with *Roseburia inulinivorans*, *Eubacterium rectale*, and *Phocaeicola vulgatus*. Seven significant positive interactions in the NASH population were also identified (Table S12). *Eubacterium rectale*, *Coprococcus come*, and *Roseburia inulinivorans* were positively associated with *Prevotella copri*. A strong positive corelation between *Ligilactobacillus ruminis* and *Phocaeicola vulgatus* was also observed. Negative interactions (r ≤ -3) were also identified in both healthy and NASH groups, but they were not significantly associated with each other (p value>0.05) (Figures 3A and 3B).

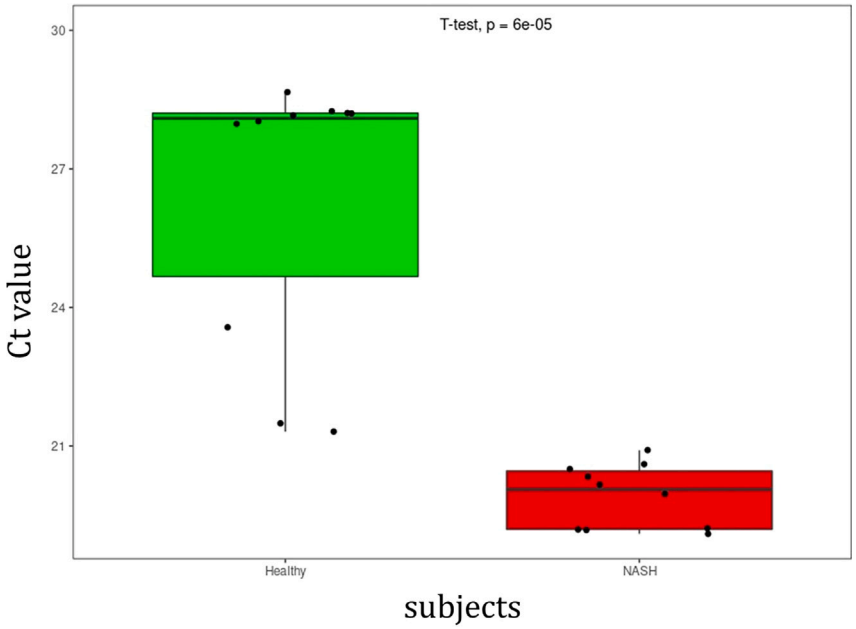


Figure 2. Relative abundance of *C. aerofaciens* in the stool sample of healthy and NASH subjects

The relative abundance was measured by quantitative PCR (qPCR) using specific primers. The abundance of *C. aerofaciens* is significantly high in NASH patients compared with the healthy people.

Morphological depiction of *Collinsella aerofaciens* cultured from the fecal samples of NASH patients

The *C. aerofaciens* colonies isolated from biopsy-proven NASH patients appeared to be smooth, spherical, white in color, and non-motile in nature and showed hemolytic activity. Minimum of three discrete colonies from each NASH patient (n = 10) was used for complete 16S rRNA gene-sequencing-based bacterial species confirmation. Bacterial isolates showed more than 98% sequence identity with the reported 16S rRNA gene sequence of *C. aerofaciens* strains were used for whole genome sequencing (n = 7) and further analysis.

Overview of the genome of *Collinsella aerofaciens* isolated from NASH patients

The genomes of seven strains of *C. aerofaciens* isolated from six different NASH patients were sequenced using an Illumina MiSeq DNA sequencing platform. A total of 39,441,512 reads were generated. An average of 541,655.25 with SD ± 11.39% reads per sample was used for quality assessment. A total of 38,63,327 paired ends were qualified based on Phred quality ≥20 (Q ≥ 20) and minimum read length (bp) ≥36. Finally, an average of 5,30,457.875 reads (SD ± 11.6%) and an average of 186.25 (SD ± 76.5) nucleotide read length per sample were used for *de novo* genome assembly by the Unicycler assembly pipeline.[31] As a result, each assembled genome has an average of 37.625x (SD ± 10.576) of sequencing depth. The checkM analysis shows that all the assembled genomes of seven strains have >98% genome coverage. Sequencing read heterogeneity analysis reveals <0.05% contamination (i.e., mixing of genome parts of other species) in all seven strains reported in this study. The sequence size of the genomes varied between 2.09 Mb and 2.28 Mb (average: 2.143 Mb ± 0.075). No ambiguous base (N) was detected in the assembled genomes. The important features of the genome-relevant information of seven strains of *C. aerofaciens* were reported in Table 4. *In silico* analyses revealed 29–30 genes associated with virulence, disease, or defense. *C. aerofaciens* strain NCA-6 harbors a 34.9-kb prophage in the tRNA-Val locus. Except for structural proteins and integrase-encoding genes, all other genes are hypothetical in nature. Except for strain NCA-6, we could not detect any bacteriophages in the genome of *C. aerofaciens* reported in this study.

Genome of *Collinsella aerofaciens* strains isolated from NASH patients harbor distinct mutations and enriched with metabolic pathways

Comparative genomics of *C. aerofaciens* strains isolated in this study from NASH patients (n = 7) and healthy subjects from a global collection (n = 96) revealed highly diverse genomes with a few spatiotemporally confined sub-clusters. Sub-cluster 1 was named based on the clustering of five NASH isolates (S13, S15, S16, S17, S23) on a large branch as shown in Figure 4 and are phylogenetically similar. "The phylogenetic relationships of the present study isolates showed an interspersed pattern with five isolates forming a sub-cluster (S13, S15, S16, S17, and S23: we refer to it as sub-cluster 1 due to the high genomic similarity), whereas the other two strains are highly divergent (Figure 4). The overall SNP differences within these 103 genomes were found to be 1,18,032, which accounted for a 5.5% variation in comparison with the reference genome (GCA_002736145.1). The phylogenetic relationships of the present study isolates showed an interspersed pattern with five isolates forming a sub-cluster (S13, S15, S16, S17, S23: we refer to it as sub-cluster 1), and the other two strains are highly divergent.

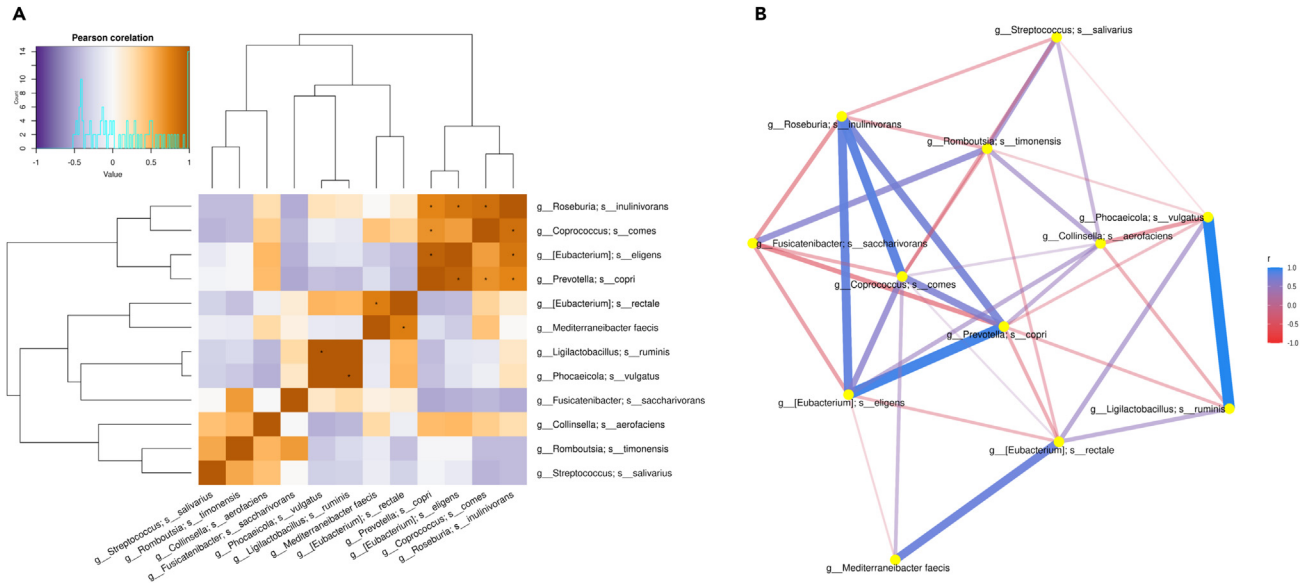


Figure 3. Cooccurrence of gut microbiota in the NASH patients

Cooccurrence among each species in the gut microbiome was measured using Pearson's correlation coefficient (r). The read abundance of the species was used to calculate the correlation network.

(A) Heatmap showing the correlation matrix between different gut bacteria in NASH population. Brown scales represent strong positive interaction, purple scales represent strong negative interaction, and the white ones indicate non-significant association. The stars in the map represent significant correlations (p value < 0.05).

(B) Cooccurrence network with threshold r > 3 or r < −3 between different species in the NASH population. The purple lines (edges) represent strong positive between the species, whereas the pink lines indicate weak negative interaction. The width of the edges indicates the strength of the interaction. The yellow dots represent one individual species.

The genome of S21 clustered with *C. aerofaciens* isolated from a healthy individual in 2017 (GCA_002736145.1), which differed by 27,350 SNPs. Another strain, S23, although it appeared to be related to the isolates reported from the United States, had phylogenetic distance of 21,288 SNPs. Notably, strains in sub-cluster 1 showed 0 SNP differences within, signifying these being identical. However, the SNP differences of sub-cluster 1 strains and S23 against the *C. aerofaciens* sourced from healthy individuals were 19,150 and 16,662 SNPs, respectively.

A detailed analysis of the subsystem in the genome of *C. aerofaciens* showed the presence of several biosynthetic gene clusters that are essential for central carbohydrate metabolism (pyruvate oxidation, citrate cycle, pentose phosphate pathway, D-galacturonate degradation, galactose degradation, glycogen biosynthesis, nucleotide sugar biosynthesis, and UDP-N-acetyl-D-glucosamine biosynthesis), fatty acid metabolism (fatty acid biosynthesis, sterol biosynthesis), and amino acid metabolism (serine and threonine metabolism, cysteine and methionine metabolism, branched-chain amino acid metabolism, arginine and proline metabolism, and aromatic amino acid metabolism). We observed enriched metabolic repertoires of 251–271 genes in the *C. aerofaciens* genomes that are involved in carbohydrate, fatty acid, and amino acid biosynthesis (Table 4). Such functional potency may help the bacterium compete with other bacterial taxa living in the same ecosystem through efficient energy assimilation from the complex dietary resources. As a consequence, such functional potencies also help *C. aerofaciens* to increase their abundance in a nutritionally rich and microbially crowded gut ecosystem.

Pan-genomic analysis of *Collinsella aerofaciens*

The pan-genome analysis of *C. aerofaciens* strains (n = 103) identified the proportional relationship between the total genes and the number of genomes studied. A total of 17,120 genes were predicted, with 646 categorized as core genes in the default 95% identity cutoff, which accounted for 33.6% of the coding sequence (CDS) (n = 646/1,920) present in the reference genome (CP024160.1). As the genome of *C. aerofaciens* is diverse with high variations, a limited core genome proportion has been captured. Upon performing Roary analysis at different sequence identity thresholds with intervals of 5% each, ranging from 95% to 35%, a significant change in the number of genes being clustered as core, soft core, shell, and cloud genes was observed (Figure S2A). Further, a comparative pan-genomic profile of *C. aerofaciens* isolated from healthy individuals versus NASH patients allowed the identification of differences in the genetic pool (Figure S2B). Notably, on comparing the pan-genome size of the global collections of 97 *C. aerofaciens* with seven NASH isolates and the reference genome, 17,120 genes have been identified, whereas a subset analysis of seven NASH isolates with reference strain depicted 3,428 genes. Such differences in the number of genes explain the open pan-genome nature of *C. aerofaciens*, with an increase

Table 3. Read summary generated by 16S rRNA-based amplicon sequencing

SN	Sample	Group	Type of reads	Sequencing platform	Total reads (quality qualified)	Merged reads	Read length(min.)	Read length (max.)	Read length (avg.)
1	sample_NF1	Obese	SE	Roche454	29,960	29,960	344	948	750.47
2	sample_NF2	Obese	SE	Roche454	19,282	19,282	353	955	754.57
3	sample_NF3	Obese	SE	Roche454	15,163	15,163	354	981	767.82
4	sample_NF4	Obese	SE	Roche454	18,201	18,201	347	997	732.16
5	sample_NF5	Obese	SE	Roche454	14,584	14,584	349	935	766.67
6	sample_NF6	Obese	SE	Roche454	17,037	17,037	361	978	714.22
7	sample_NF7	Obese	SE	Roche454	8,648	8,648	364	979	771.84
8	sample_NF8	Obese	SE	Roche454	7,180	7,180	349	946	750.85
9	sample_NF9	Obese	SE	Roche454	7,543	7,543	351	945	748.19
10	sample_NF10	Obese	SE	Roche454	21,043	21,043	351	987	775.99
11	sample_NF11	Obese	SE	Roche454	22,381	22,381	349	971	754.8
12	sample_NF12	Obese	SE	Roche454	18,186	18,186	349	973	768.91
13	sample_NF13	Obese	SE	Roche454	44,870	44,870	364	980	776.64
14	sample_NF14	Obese	SE	Roche454	16,016	16,016	327	964	747.24
15	sample_NF15	Obese	SE	Roche454	10,921	10,921	356	962	747.14
16	sample_NF16	Obese	SE	Roche454	10,288	10,288	359	938	744.42
17	sample_NF17	Obese	SE	Roche454	9,494	9,494	347	944	754.76
18	NASH_S21	NASH	PE	MiSeq Illumina	2,22,384	29,648	50	440	245
19	NASH_S22	NASH	PE	MiSeq Illumina	3,34,495	1,06,453	50	461	255.5
20	NASH_S29	NASH	PE	MiSeq Illumina	3,06,651	1,03,203	50	459	254.5
21	NASH_S28	NASH	PE	MiSeq Illumina	3,97,262	1,35,513	50	467	258.5
22	NASH_S24	NASH	PE	MiSeq Illumina	3,19,926	1,14,632	50	452	251
23	NASH_S30	NASH	PE	MiSeq Illumina	3,55,671	1,29,954	50	469	259.5
24	NASH_S27	NASH	PE	MiSeq Illumina	4,05,459	1,54,261	50	450	250
25	NASH_S26	NASH	PE	MiSeq Illumina	3,42,104	1,36,216	50	464	257
26	NASH_S25	NASH	PE	MiSeq Illumina	3,46,308	1,44,043	50	468	259
27	NASH_S23	NASH	PE	MiSeq Illumina	3,28,144	1,37,278	50	460	255

SE = single-end reads; PE = paired-end reads.

in the total genes upon including 97 isolates from the global collection, as compared with the NASH vs. healthy group. The increase in the number of genes with a higher number of genomes was attributed to the diversity of this genus, which has subsequently been reduced by optimizing the blast identity percentage. Additional analysis was performed on the results for both the global comparison and the NASH vs. healthy group.

Differences in the biological pathway subsystems in NASH *C. aerofaciens*

Besides analyzing the general characteristics of *C. aerofaciens* genomes, we sought to identify the key biological pathways that impact the host disease progression in the NAFLD, as *C. aerofaciens* abundance was high. To understand the behavior of *C. aerofaciens*, the subsystems present in the NASH isolates were analyzed and compared, with special emphasis on metabolic pathways (Figure 5). Further, differences in the genes gained and lost in the NASH sub-cluster 1 isolates compared with the reference are summarized in Table 5. All the seven isolates from NASH patients have acquired genes encoding for iron transporter protein and genes involved in the inositol catabolic process. Five NASH isolates from sub-cluster 1, were found to have acquired multiple genes coding for sugar transporters such as lichenen, gluconate, cellobiose, mannobiose, and glycerate; secondary bile acid synthesis such as 3-alpha-hydroxycholanate dehydrogenase; as well as genes for toxin-antitoxin systems (*higB2* and *socA*). These isolates also harbored CRISPR-cas system with 33 spacer regions. In contrast, genes absent in the NASH sub-cluster 1 were associated with leucine-/valine-/isoleucine-binding protein, UbiX-like flavin prenyltransferase, and xylulose-5-phosphate/fructose6-phosphate phosphoketolase. More importantly, one of the three copies of alcohol dehydrogenase was absent (Figure 6). Alcohol dehydrogenases catalyze the reversible conversion of aldehyde to ethanol, which has been hampered as one of the three copies of alcohol dehydrogenase was completely absent, whereas the other two copies were present with mutations resulting in the accumulation of ethanol, which is a classic signature in the NAFLD.

Table 4. Overview of the *C. aerofaciens* genomes

Characteristics	NCA-1	NCA-2	NCA-3	NCA-4	NCA-5	NCA-6	NCA-7
Year of isolation	2020	2020	2020	2020	2020	2020	2020
Place of isolation	AIIMS-Delhi	AIIMS-Delhi	AIIMS-Delhi	AIIMS-Delhi	AIIMS-Delhi	AIIMS-Delhi	AIIMS-Delhi
Sample	Fecal	Fecal	Fecal	Fecal	Fecal	Fecal	Fecal
Sequence size (Bp)	2,095,888	2,095,888	2,105,499	2,097,399	2,096,072	2,268,317	2,104,431
GC content (%)	60	60	59.9	60	60	60	60
Number of contigs	42	42	49	31	36	49	39
N50	5	5	6	135553	135553	110476	135553
Number of rRNA	59	59	58	59	60	56	59
Number of subsystems	197	197	197	197	197	194	197
Number of coding sequences	1856	1856	1872	1846	1848	1978	1854
Number of genes in carbohydrate metabolism	121	120	121	120	120	126	120
Number of genes in fatty acid metabolism	22	22	22	22	22	19	22
Number of genes in amino acid metabolism	109	109	112	109	109	123	109
Number of genes in virulence, disease	29	29	29	29	29	29	29
Mobile genetic elements	1	1	1	1	1	0	1

The acronym NCA-1 to NCA-7 stands for isolate number, respectively: e.g., NCA-1 stands for NASH-*Collinsella aerofaciens* 1.

Mutational profile in carbohydrate and lipid metabolism

Genomic comparison of the seven study isolates of *C. aerofaciens* from NASH patients against the reference strain isolated from a healthy individual (CP024160.1) yielded 3,77,141 SNPs among 1,748 genes. This constituted 15% variation between the healthy (2,306,349 bp genome) and NASH isolates. Among these SNPs, synonymous mutations (2,86,058—76%) were predominant, followed by missense (63,571—17%), frameshift (295—0.07%), disruptive-in frame deletion (149–0.05%), and combinations of more than one effect (27,068—7%). SNPs rates of sub-cluster 1 isolates were observed to be similar with S23; however, S21 had higher SNPs, especially synonymous mutations (Figure 7). In addition, we attempted to understand the genomic diversity of *C. aerofaciens* collected across the globe, as this has not been done earlier. Further, the rationale behind comparing the SNPs in NASH isolates against wildtype, which was the reference genome isolated from healthy control, was to capture every SNP variation in the NASH isolates, especially with reference to the key genes involved in the metabolic pathways. Notably, NASH-enriched SNPs have been captured well by comparing them against the *C. aerofaciens* isolated from healthy controls as mentioned in Figures 4 and 8A–8D). As the mutation rates were high, we focused on identifying key SNPs in the genes involved in the metabolic pathways. Using the KEGG pathways annotation of the *C. aerofaciens* genome, a list of genes in the carbohydrate metabolism (n = 91) and fatty acid synthesis (n = 18) pathways was retrieved, and the frequency of synonymous and missense mutational profiles observed is illustrated as a heatmap as shown in the Figures 8A–8D. In the carbohydrate metabolic pathway, it was evident that the locus tag CSV91_03400 encodes for bifunctional acetaldehyde-CoA/alcohol dehydrogenase gene, displaying the highest mutations, 167 SNPs, with 135 synonymous and 32 missense mutations. Notably, 26/32 missense mutations were found to be specific to sub-cluster 1 NASH isolates, possibly impacting the structural and functional level. Similarly, genes that encode for FAD/NAD(P)-binding oxidoreductases in fatty acid metabolism showed the highest SNPs, with 21/35 missense being specific to sub-cluster 1 NASH isolates. The genetically distant sub-cluster 1 (S13, S15, S16 S17, and S22) strains from S21 and S23 have been well correlated with the progression of the disease, attributed to their differences in the genomic repertoires as identified by the comparative genome analysis. Further, the genomic signatures of *C. aerofaciens* isolated from the NASH patients correlated well with the disease progress, as sub-cluster 1 cases biopsy scores ranged from 2 to 4, whereas S21 and S23 were scored as 0. The key differences identified with the two genes related to metabolic pathways could play a major role in the accumulation of ethanol and fatty acid synthesis process, which in turn implicates the adverse progression of the disease. However, it is to be noted that despite the genetic similarities within the sub-cluster 1 cases, these have been isolated from NASH patients of different geographical locations within the Delhi-NCR region.

C. aerofaciens increases circulating ethanol levels, hepatic triglycerides, and hydroxyproline

The gut microbiota is the source of several metabolites for their hosts, which play an important role in host physiology by modulating metabolic and signaling pathways. Because we observed an increase in the abundance of *C. aerofaciens* in the gut of NASH patients compared with the healthy subjects, we intend to test whether the bacterium can contribute to the ethanol, triglyceride, and hydroxyproline accumulation, three important metabolites that directly influence inflammation and hepatic functions associated with NASH development and progression. By gas chromatography-mass spectrometry (GC-MS) analysis, we observed increased level of circulating ethanol, which is a pro-inflammatory metabolite, in the C57BL/6J mice fed with a normal chow or CDHF diet and supplemented with *C. aerofaciens* compared with the control group, which was fed with similar foods but without supplementation with the bacterium (Figures 9A and 9B). We also measured the

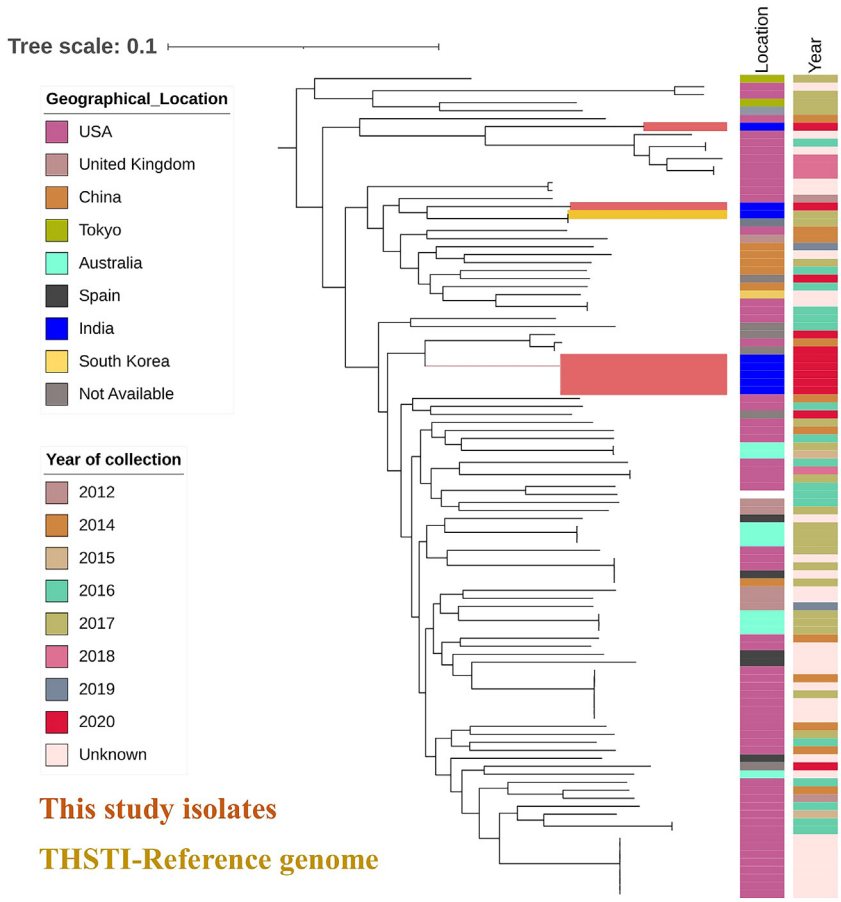


Figure 4. Maximum likelihood phylogenetic tree of 103 genomes of *C. aerofaciens*

Isolates highlighted are the present study genomes (n = 7) and one strain from healthy individual. Color strips indicate the geographical location from where these strains have been isolated and its respective year of isolation. Scale bar indicates the number of SNP substitutions per site per genome. The phylogenetic relationships of the present study isolates showed an interspersed pattern with five isolates forming a sub-cluster (S13, S15, S16, S17, S23: we refer to it as sub-cluster 1 due to the high genomic similarity), whereas the other two strains are highly divergent.

hydroxyproline and triglycerides in the hepatic cells of all the mice groups fed with a normal chow or CDHF diet with or without supplementation of *C. aerofaciens*. Hepatic cells of mice supplemented with *C. aerofaciens* have shown significantly higher levels of hydroxyproline (Figure 10) as well as triglyceride (Figure 11) compared with the normal chow-fed mice groups.

C. aerofaciens induces inflammation in hepatocyte but not steatohepatitis and ballooning

To investigate if the development of NAFLD is induced by *C. aerofaciens* in association with a choline-deficient high-fat (CDHF) diet, 8- to 12-week-old C57BL/6J male mice were fed a normal chow and CDHF diet together with *Collinsella* strains isolated from NASH patients. All groups of mice exhibited substantial variations in body weight (Figure S4A) and feed intake (Figure S4B). The variability in liver-to-body mass ratio was also assessed (Figure S4C). However, we did not find any differences in hepatocellular ballooning and steatohepatitis between mice fed with different diets and supplemented with or without *C. aerofaciens*. Disease progression was examined by histological analysis of the liver tissues of mice stained with hematoxylin and eosin (H&E) and Masson's trichrome (MT). We observed that, compared with the normal diet, a CDHF diet leads to the accumulation of a substantial amount of fat in the hepatic cell by gross anatomy and histopathology staining (Figure 12A). The group that was fed *C. aerofaciens* with a CDHF diet had a higher activity score as compared with CDHF alone (Figure 12B).

The highest overall activity score was for the cohort augmented with *C. aerofaciens* on a CDHF diet. Interestingly, increased levels of transcripts of different pro-inflammatory genes (nuclear factor κB [NF-κB], tumor necrosis factor alpha [TNF-α], interleukin-6 [IL-6]) were observed in the hepatocytes at 10 weeks. We also observed a reduced expression level of anti-inflammatory gene encoding IL-10 (Figure 13A) and a larger fold change in the transcript level of fibrotic markers (Figure 13B) and fat uptake (Figure 13C) genes in the hepatocytes of CDHF-diet-fed mice supplemented with *C. aerofaciens* compared with the mice without *C. aerofaciens* supplementation. The findings of the present

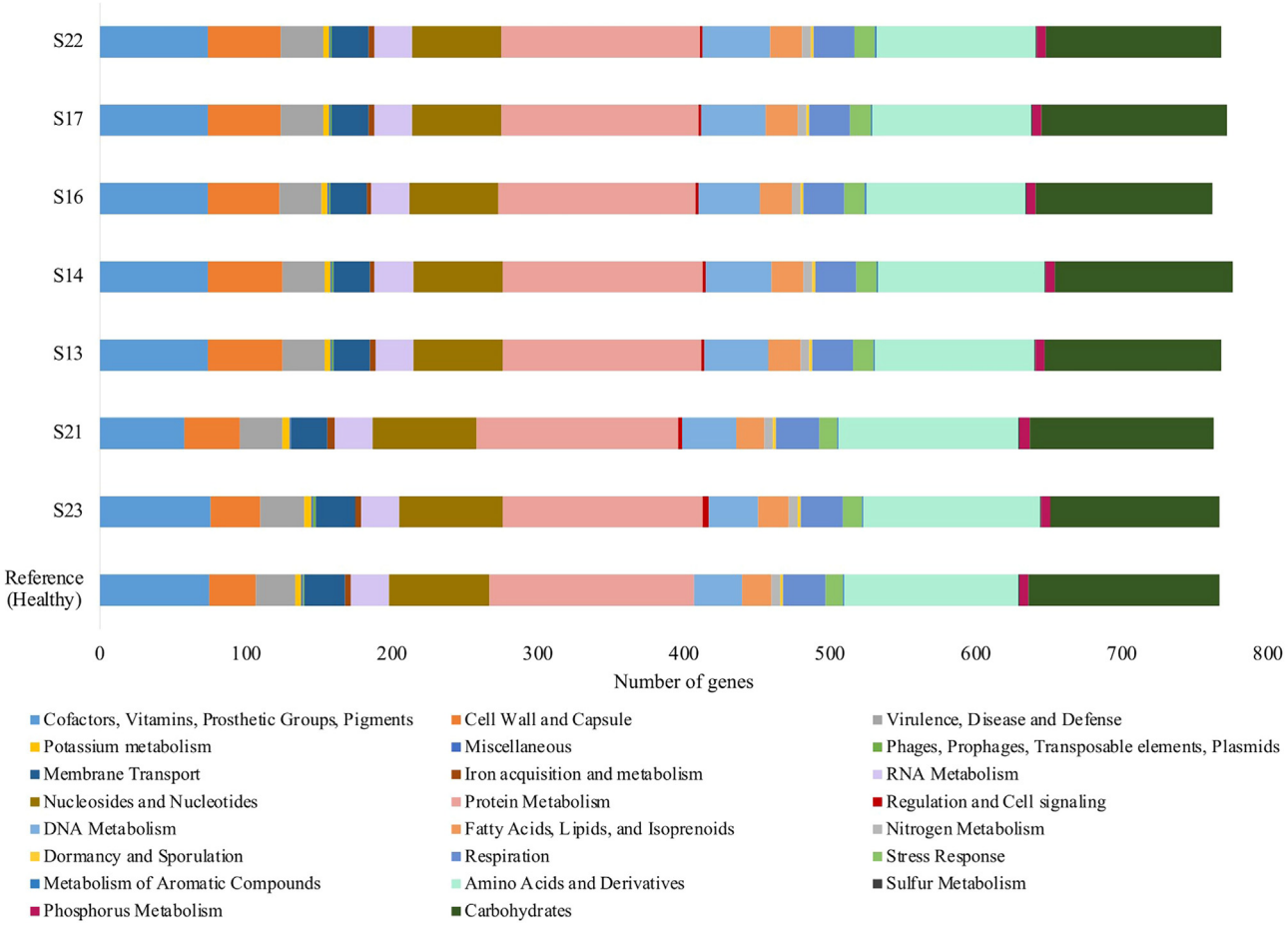


Figure 5. Subsystems analysis of the *Collinsella aerofaciens* genomes

The subsystems present in the *C. aerofaciens* genome isolated from NASH patients were predicted by RAST annotation pipeline.

study indicate that the increased abundance of *C. aerofaciens* in the gut microbiome may stimulate hepatocyte inflammation and induce disease progression along with diets, microbial derived metabolites, and other environmental factors.

DISCUSSION

NAFLD affects populations worldwide, leading to a huge burden on health and the economy. NAFLD can progress to NASH in one-fifth and hepatocellular carcinoma in a fraction of patients.[32] Despite its pervasiveness in developing countries, including India, the human body's associated functions linked with disease development and severity have received little attention from the scientific and clinical community. Our current limited understanding shows that the gut microbiome is one of the crucial risk factors in the pathogenesis and severity of NAFLD. The dynamic interactions between the microbial community and their metabolites with liver cells may directly influence the inflammatory cascade, which in turn affects the prognosis of the disease. The cross-sectional and longitudinal-culture-independent metagenomic studies of the gut microbiome of healthy subjects, obese subjects, and NASH patients have revealed both richness and increased diversity of the microbiome, reducing the risk of NAFLD and other metabolic diseases.[33] The present study also identified lower richness and diversity of gut microbiome in NASH patients compared with healthy subjects. The rarefaction curve of OTUs and read depth showed that most of the microbial taxa present in the fecal samples of the NASH patients were captured. The significant differences in the Chao1 index among the three groups represent the loss of a substantial number of OTUs in the gut of NASH and obese patients. In addition, several bacterial taxa, including *Bifidobacterium*, *Klebsiella*, *Pseudescherichia*, *Howardella*, *Intestinibacter*, and *Enterococcus*, known as pathobionts, have been identified in the fecal samples of NASH patients. These genera may play an important role in chronic inflammation in NASH patients. Bacterial genera that are common in three groups have different abundances, which may also contribute to NASH pathogenesis. The abundance of genus *Collinsella*, a widely conserved gut microbiota with enriched pro-inflammatory functions, has increased from healthy to obese and NASH patients. Higher taxonomic resolution of the gut microbiota at the species level identified *C. aerofaciens* as one of the most dominant species in the gut of obese and NASH patients compared with healthy subjects. Functional insights obtained by decoding whole genome sequences of

Table 5. Differences in the gene content of *C. aerofaciens* isolated from NASH patients compared with healthy individual

Gene	Annotation	Pathway/Involvement
Healthy-Absent; NAFLD_7-Present		
irtA	Iron import ATP-binding/permease protein IrtA	Ion deficiency has been associated with NAFLD
rhaS	HTH-type transcriptional activator RhaS	Rhamnose metabolism, transcription, transcription regulation
iolX	scyllo-Inositol 2-dehydrogenase (NAD(+))	Catalyzes the reversible NAD+-dependent oxidation of scyllo-inositol (SI) to 2,4,6/3,5-pentahydroxycyclohexanone (scyllo-inosose or SIS). Is required for SI catabolism that allows *Bacillus subtilis* to utilize SI as the sole carbon source for growth. Cannot use NADP+ instead of NAD+.
yhjX	Putative MFS-type transporter YhjX	Oxalate/formate antiporter
fldI	(R)-phenyllactate dehydratase activator	This protein is involved in the pathway L-phenylalanine degradation, which is part of amino acid degradation
msbA	Lipid A export ATP-binding/ permease protein MsbA	Sugar transport. Involved in beta-(1→2)glucan export. Transmembrane domains (TMD) form a pore in the inner membrane and the ATP-binding domain (NBD) is responsible for energy generation.
Healthy, NAFLD (2 isolates—S21, S23) Absent; Sub-cluster 1 NAFLD_5 -Present		
baiA	3-Alpha-hydroxycholanate dehydrogenase (NADP(+))	Bile acid biosynthesis
Ugd	UDP-glucose 6-dehydrogenase	—
Fcl	GDP-L-fucose synthase	Cell wall/LPS biosynthesis pathway
deoC1	Deoxyribose-phosphate aldolase 1	Carbohydrate degradation
algA	Alginate biosynthesis protein AlgA	—
purT	Formate-dependent phosphoribosylglycinamide formyltransferase	—
manR	Transcriptional regulator ManR	Mannose metabolism
gmuC	PTS system oligo-beta-mannoside-specific EIIC component	Transport of cellobiose or mannobiose
licB	Lichenan-specific phosphotransferase enzyme IIB component	Transport of Lichenan
gntT	High-affinity gluconate transporter	Transport of gluconate
mngA	PTS system 2-*O*-alpha-mannosyl-D-glycerate-specific EIIABC component	Transport of mannosyl-D-glycerate
epsK	Putative membrane protein EpsK	
sthA	Soluble pyridine nucleotide transhydrogenase	NADP metabolic process
ppm1	Bifunctional apolipoprotein N-acyltransferase/polyprenol monophosphomannose synthase	Mannose metabolism and lipoprotein biosynthesis
glvR	HTH-type transcriptional regulator GlvR	Transcriptional regulator of maltose utilization

the *C. aerofaciens* isolated from healthy subjects and NASH patients showed specific malfunctioning of a bifunctional acetaldehyde-CoA/alcohol dehydrogenase encoding gene, which is involved in alcoholic metabolism in *C. aerofaciens* isolated from the NASH patients. High-alcohol-producing *Klebsiella pneumoniae* living in the gastrointestinal tract of 60% of individuals with NAFLD in a Chinese cohort has previously been reported.[18] A connection between high levels of endogenous ethanol due to the presence Saccharomyces cerevisiae in the gut of NASH patients was also observed.[17] A recent report also mentioned the increased abundance of *Collinsella* in the gut of NASH patients.[34] However, to the best of our knowledge, the genomic functional potency that may contribute to NASH pathogenesis due to the increased abundance of *C. aerofaciens* was missing from previous reports. The comparative genomic study of different *C. aerofaciens* strains isolated from the gut of healthy subjects and NASH patients is a unique addition to the current understanding of the role of the gut microbiome in NAFLD. The unique genomic signatures of *C. aerofaciens* isolated from NASH patients might stimulate chronic inflammation in the hepatocytes if its abundance increases significantly. The findings of the animal experiments with

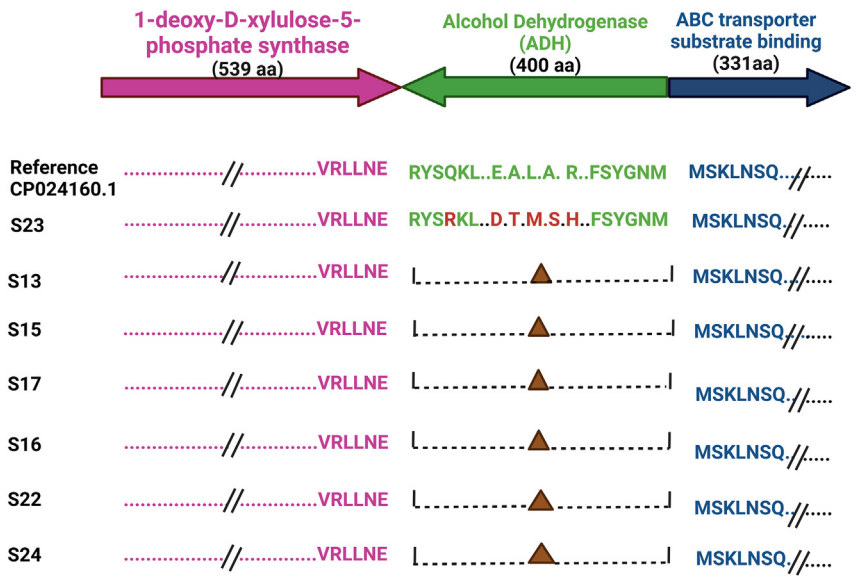


Figure 6. Genetic analysis of the alcohol dehydrogenase encoding functions in *C. aerofaciens*

Amino acid sequence alignment of the alcohol dehydrogenase (ADH) gene along with the upstream and the downstream genes in *C. aerofaciens* chromosome.

C. aerofaciens supplementation and the CDHF diet strongly support the possible role of the bacterium in chronic inflammation in hepatocytes. A similar observation was also reported in rheumatoid arthritis, where the *Prevotella copri* colonized the patient population but not the healthy subjects and had a unique genomic signature associated with the disease.[35] The *C. aerofaciens* residing in the gut of healthy subjects may not have the similar functional potency to increase the level of hepatic or circulatory ethanol, thus not induced inflammation in the hepatocytes. Findings of the animal experiments with normal chow or CDHF diet supplemented with *C. aerofaciens* strongly support this hypothesis. We furthermore suspect that *C. aerofaciens* inflammatory potency is strongly influenced by the host's dietary patterns, as demonstrated by higher inflammation cascade after CDHF and *C. aerofaciens* supplementation. Through bacterial adaptation, the bacterial genomic repertoire changes with gene gain or loss or by the accumulation of mutations. In this study, we have identified several genomic changes in the *C. aerofaciens* isolated from NASH patients compared with the strain isolated from a healthy individual, both in terms of mutations and gene gain/loss.

This is the first study to correlate the genomic signatures of *C. aerofaciens* and its association with NAFLD disease progression. As discussed in the results, sub-cluster 1 isolated from the gut of NASH patients was phylogenetically distinct from the other isolates obtained from the healthy subjects. The nature of the bacterial pan-genome being closed or open always depends on its lifestyle and environmental niches. In that context, the genomes of *C. aerofaciens* were found to have an open genome attributed to its genome diversity. A greater number of genomes are needed to validate this finding. Carbohydrates are a significant stimulus for hepatic *de novo* lipogenesis (DNL) and are more likely to lead directly to NAFLD than dietary fat. DNL uses glucose, fructose, and amino acids as substrates for synthesizing newly

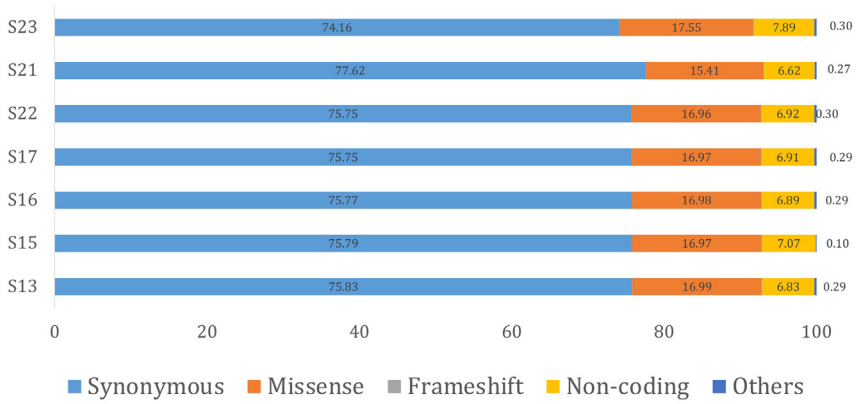


Figure 7. Analysis of the single nucleotide polymorphisms (SNPs) in the *Collinsella aerofaciens* genome

Distribution of SNPs with effects as synonymous, missense, frameshift, non-coding, and others in the NASH *C. aerofaciens* predicted in comparison with the *C. aerofaciens* isolated from healthy individual.

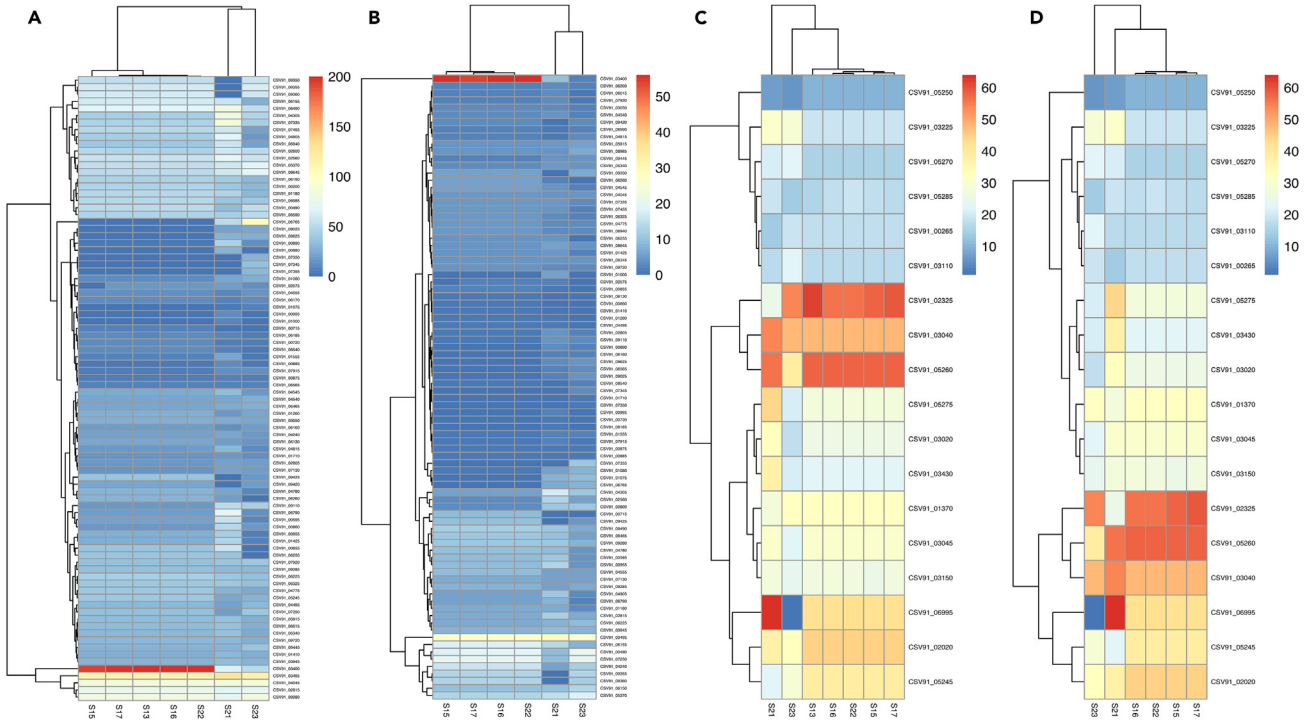


Figure 8. Analysis of the gene mutations in the *C. aerofaciens* genome

Heatmap represents the number of gene mutations, (A) synonymous (B) missense mutation in the genes involved in the carbohydrate metabolism (n = 91), (C) synonymous (D) missense mutations in the genes involved in the fatty acid biosynthesis (n = 18).

produced fatty acids. The NASH-inducing features of the CDHF diet may be augmented by a greater dietary fat intake (60% fat kcal, CDHF model), which has been shown to cause steatosis, inflammation, and mild pericellular fibrosis after eight weeks of dieting.[36] CDHF is a significant modifiable risk factor for obesity, type 2 diabetes, and NAFLD.[37] Several remarkable observations documented in this study discern the role of genes in the carbohydrate metabolism and fatty acid biosynthesis pathways, implicating them in adverse host disease progression. The findings suggest that multiple high-quality genome sequences of the pathogen from different hosts from healthy and NASH subjects should be used to better understand the virulence and genetic factors that allow the *C. aerofaciens* to adapt in a broad niche and rapidly

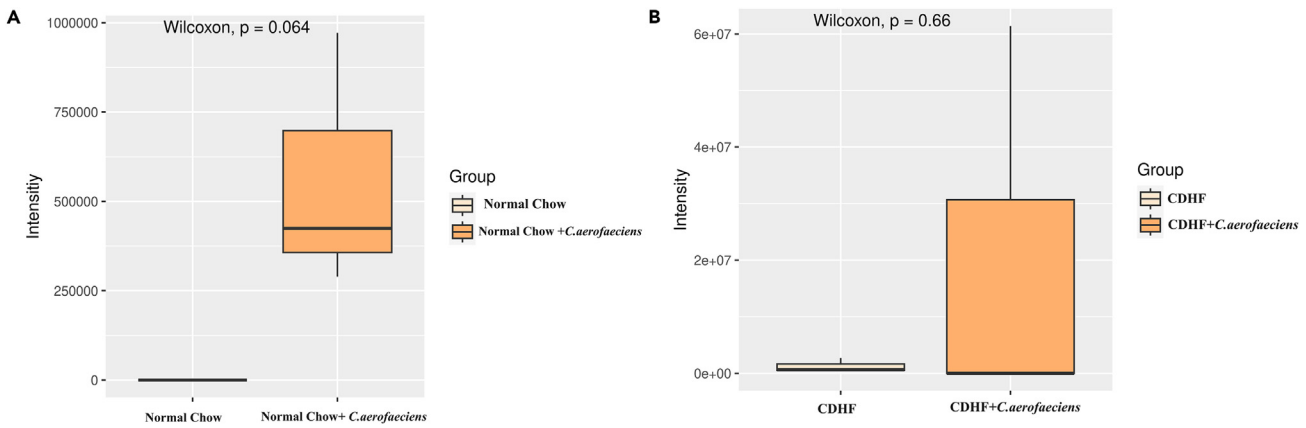


Figure 9. Analysis of the ethanol level in the blood samples by GC-MS

The boxplot depicted the intensity of ethanol production.

(A) The production of ethanol was found to be high when normal chow plus *C. aerofaciens* diets was given to mice models, whereas no ethanol production was observed in case of normal chow diet (Wilcoxon, p value = 0.064.

(B) The production of ethanol was found to be high when CDHF (choline-deficient high-fat) plus *C. aerofaciens* diets were given to mice models, whereas less ethanol production was observed in case of CDHF diet. (Wilcoxon, p value = 0.066). The number of mice models used for both experiments are n = 4 for each group.

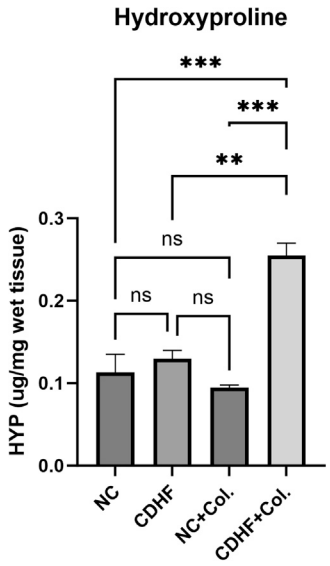


Figure 10. Measurement of the hydroxyproline content in the hepatic tissue of mice

Bar plot shows the concentration of hydroxyproline in mice hepatic tissue of distinct diet induces NAFLD model. In comparison to the normal chow, CDHF + *Collinsella* is showing significant higher abundance of hydroxyproline. **p ≤ 0.01, ***p ≤ 0.001.

swipe from commensal to pathobiont. Our findings may help develop therapeutics to tackle NAFLD by targeting specific functions or growth of a particular gut microbiome population.

The gut microbiome composition of the NASH patients and mouse studies of the current report has provided evidence that the increased abundance of *C. aerofaciens* in the gastrointestinal tract with unique genomic repertoires may contribute to NAFLD development by changing the hepatic and/or circulatory metabolites derived from gut microbiota. The inflammatory functions, ethanol-producing capacity, and enriched metabolic functions for efficient use of simple sugars and fatty acids are the potential functions of *C. aerofaciens* that can directly increase the risk of NAFLD development.

Limitations of the study

There are certain limitations of the study. In this study, host-bacterial interactions were not explored. In addition, the study findings are primarily focused on the role of bacterial taxa in the gut microbiome based on the limited sample size. The fungal component of the gut microbiome and possible microbial invasion into the liver from the gut has also not been incorporated. Cross-kingdom interactions with additional samples will strengthen the current findings. The results need to be validated in larger sample size studies and across different ethnicities.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Patient's recruitment and sample collection
  - Mouse housing conditions and sampling
- METHOD DETAILS
  - DNA extraction and metagenomic sequencing
  - Processing 16S rRNA raw reads
  - Analysis of 16S rRNA reads
  - Gut microbial interactions analysis
  - qRT PCR assay
  - Isolation, identification and cultivation of *Collinsella aerofaciens*
  - Whole genome sequencing

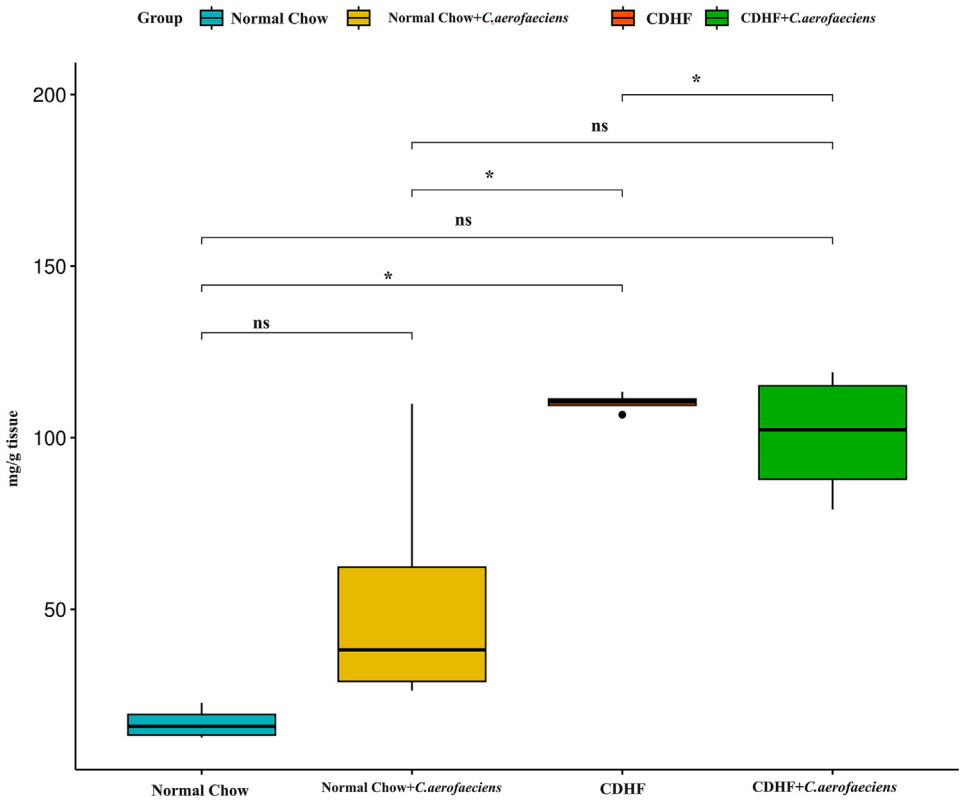


Figure 11. Measurement of the triglycerides content in the hepatic tissue of mice

Boxplot depicting the level of triglycerides in hepatic tissue of mice, as compared with normal-chow-fed mice group, CDHF groups has higher level of triglycerides. *p ≤ 0.05.

- Genome assembly and annotations
- Phylogenetic and pan-genome analyses
- Animal experiment

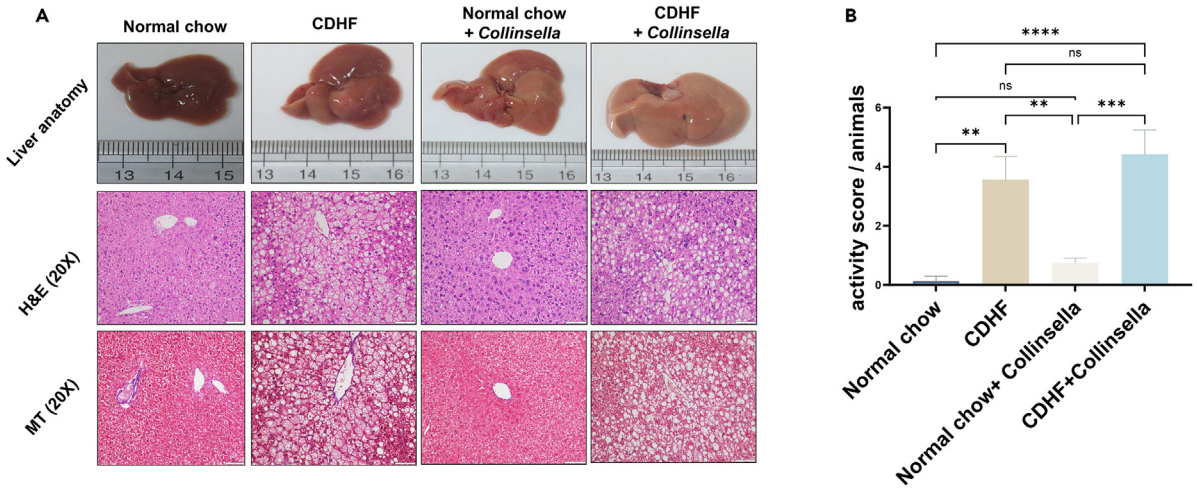


Figure 12. Histology of the liver tissue of mice supplemented with *C. aerofaciens*

(A) Representative staining images of liver sections for H&E (scale bar, 100 mm) and Masson trichome (scale bar, 50 mm). (n = 8) Mice per group of livers after 10 weeks (70 days) with oral administration of $1*10^8$ CFU of *C. aerofaciens*.

(B) The same histological scoring is represented as total scoring in the form of boxplot on the basis of inflammation, ballooning, and steatosis scoring in concordance with NAS score. No fibrosis was seen in any group.

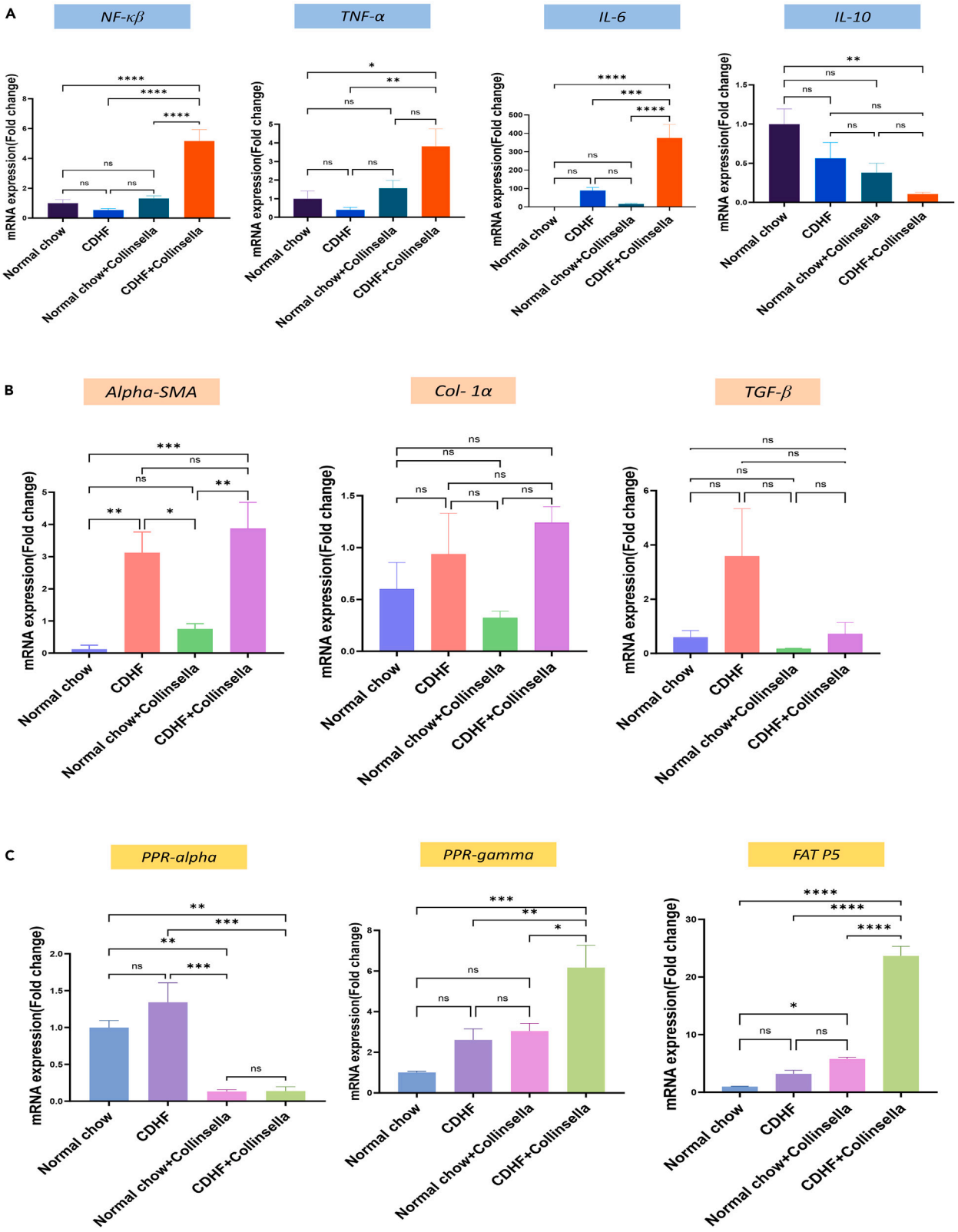

A
NF-κβ
TNF-α
IL-6
IL-10
mRNA expression(Fold change)
Normal chow
CDHF
Normal chow+Collinsella
CDHF+Collinsella
B
Alpha-SMA
Col- 1α
TGF-β
C
PPR-alpha
PPR-gamma
FAT P5

Figure 13. Analysis of the expression level of inflammation-associated genes

Gene expressions of (A) inflammation, (B) fibrosis, and (C) fat uptake genes were measured by qPCR in all four groups of mice supplemented with *C. aerofaciens*. The mRNA expression levels of target genes were normalized to that of 18S rRNA gene. n = 5 replicates. Data are represented as mean ± SEM, n = 5, and statistical analysis consisted of one-way ANOVA followed by Bonferroni's test (the p value represents significant level by * < 0.1, ** < 0.01, *** < 0.001, and **** <0.0001).

- Development of animal models for non-alcoholic steatohepatitis (NASH)
- Liver histology
- Measurement of gene expression related to inflammation, fibrosis and fat accumulation in liver
- Measurement of hepatic hydroxyproline and triglyceride
- Measurement of serum ethanol level
- Ethical approvals
- Statistical analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.108764.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of Prof. Pramod K Garg, Executive Director, THSTI. A.P. received research fellowship from the Dept. of Science and Technology, Govt. of India (No. DST/INSPIRE Fellowship/SPRVSR/2019/IF170872). B.K. received research fellowship from the Council of Scientific and Industrial Research (CSIR), Govt. of India. P.K. received research fellowship from the Indian Council of Medical Research (ICMR), Govt. of India (No. 3/1/2(18)/OBS/2022-NCD-II).

Funding: This work was supported by the THSTI to Dr. Bhabatosh Das (AIIMS-THSTI Inter-Institutional Collaborative Research Grant, Year 2023); Science and Engineering Research Board, Govt. of India (SPR/2020/000315, Date 12th March 2021) to Dr Shalimar; AIIMS-THSTI Inter-Institutional Collaborative Research Grant (Year 2019). A.P. received research fellowship from Dept. of Science and Technology (DST), Govt. of India (No. DST/INSPIRE Fellowship/SPRVSR/2019/IF170872). B.K. received research fellowship from Council of Scientific and Industrial Research (CSIR), Govt. of India.

AUTHOR CONTRIBUTIONS

B.D. conceived the study. B.D., Shalimar, and S.K.B. designed research plan. A.P., B.K., S.K., A.K.P., S.S., P.K., D.T., J.V., V.S., R.Y., R.N., B.D. performed research. M.D. and D.M. contributed new reagents/analytical tools. A.P., S.K., A.K.P., and B.D. analyzed data. A.P., S.K., A.K.P., and B.D. wrote the paper. S.K.B., Shalimar, and B.D. reviewed and finalized the manuscript.

DECLARATION OF INTERESTS

No potential conflict of interest was reported by any of the author(s) listed in this manuscript.

Received: July 18, 2023
Revised: October 21, 2023
Accepted: December 15, 2023
Published: December 26, 2023

REFERENCES

1. Huang, D.Q., El-Serag, H.B., and Loomba, R. (2021). Global epidemiology of NAFLD-related HCC: trends, predictions, risk factors and prevention. Nat. Rev. Gastroenterol. Hepatol. *18*, 223–238.
2. Tripathi, A., Debelius, J., Brenner, D.A., Karin, M., Loomba, R., Schnabl, B., and Knight, R. (2018). The gut–liver axis and the intersection with the microbiome. Nat. Rev. Gastroenterol. Hepatol. *15*, 397–411.
3. Ahrodia, T., Das, S., Bakshi, S., and Das, B. (2022). Structure, functions, and diversity of the healthy human microbiome. Prog. Mol. Biol. Transl. Sci. *191*, 53–82.
4. Ghosh, T.S., Gupta, S.S., Bhattacharya, T., Yadav, D., Barik, A., Chowdhury, A., Das, B., Mande, S.S., and Nair, G.B. (2014). Gut microbiomes of Indian children of varying nutritional status. PLoS One *9*, e95547.
5. Martínez-Ramírez, O.C., Salazar-Piña, D.A., de Lorena, R.G.M., Castro-Hernández, C., Casas-Ávila, L., Portillo-Jacobo, J.A., Rubio, J., and Association of NFκβ, T.N.F. (2021). IL-6, IL-1β, and LPL Polymorphisms with Type 2 Diabetes Mellitus and Biochemical Parameters in a Mexican Population. Biochem. Genet. *59*, 940–965.
6. Alvarez-Silva, C., Kashani, A., Hansen, T.H., Pinna, N.K., Anjana, R.M., Dutta, A., Saxena, S., Støy, J., Kampmann, U., Nielsen, T., et al. (2021). Trans-ethnic gut microbiota signatures of type 2 diabetes in Denmark and India. Genome Med. *37*, 1–13.
7. Flint, H.J. (2011). Obesity and the gut microbiota. J. Clin. Gastroenterol. *45*, S128–S132.
8. Morgan, X.C., Tickle, T.L., Sokol, H., Gevers, D., Devaney, K.L., Ward, D.V., Reyes, J.A., Shah, S.A., LeLeiko, N., Snapper, S.B., et al. (2012). Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biol. *13*, 1–18.
9. Kedia, S., Mouli, V.P., Kamat, N., Sankar, J., Ananthakrishnan, A., Makharia, G., and Ahuja, V. (2020). Risk of tuberculosis in patients with inflammatory bowel disease on infliximab or adalimumab is dependent on the local disease burden of tuberculosis: a systematic review and meta-analysis. Am. J. Gastroenterol. *115*, 340–349.

10. Bhatt, A.P., Redinbo, M.R., and Bultman, S.J. (2017). The role of the microbiome in cancer development and therapy. CA. Cancer J. Clin. *67*, 326–344.
11. Kumari, A.P., Maurya, S., Jain, K., Pal, S., Raja, C., Kumar, S., Purohit, A., Pradhan, D., Kajal, K., Talukdar, D., et al. (2023). Chronic salmonella infection contributes to gallbladder carcinogenesis. https://ssrn.com/abstract=4540291.
12. Shreiner, A.B., Kao, J.Y., and Young, V.B. (2015). The gut microbiome in health and in disease. Curr. Opin. Gastroenterol. *31*, 69–75.
13. Lin, R., Zhou, L., Zhang, J., and Wang, B. (2015). Abnormal intestinal permeability and microbiota in patients with autoimmune hepatitis. Int. J. Clin. Exp. Pathol. *8*, 5153–5160.
14. Al Bander, Z., Nitert, M.D., Mousa, A., and Naderpoor, N. (2020). The gut microbiota and inflammation: an overview. Int. J. Environ. Res. Public Health *17*, 7618.
15. Jones, R.M. (2016). Focus: Microbiome: The influence of the gut microbiota on host physiology: In pursuit of mechanisms. Yale J. Biol. Med. *89*, 285–297.
16. Chen, J., Wright, K., Davis, J.M., Jeraldo, P., Marietta, E.V., Murray, J., Nelson, H., Matteson, E.L., and Taneja, V. (2016). An expansion of rare lineage intestinal microbes characterizes rheumatoid arthritis. Genome Med. *8*, 43–44.
17. Zhu, L., Baker, S.S., Gill, C., Liu, W., Alkhouri, R., Baker, R.D., and Gill, S.R. (2013). Characterization of gut microbiomes in nonalcoholic steatohepatitis (NASH) patients: a connection between endogenous alcohol and NASH. Hepatology *57*, 601–609.
18. Yuan, J., Chen, C., Cui, J., Lu, J., Yan, C., Wei, X., Zhao, X., Li, N., Li, S., Xue, G., et al. (2019). Fatty Liver Disease Caused by High-Alcohol-Producing *Klebsiella pneumoniae*. Cell Metab. *30*, 675–688.e7.
19. Xiong, J., Chen, X., Zhao, Z., Liao, Y., Zhou, T., and Xiang, Q. (2022). A potential link between plasma short-chain fatty acids, TNF-α level and disease progression in non-alcoholic fatty liver disease: A retrospective study. Exp. Ther. Med. *24*, 598.
20. Wang, Z., Klipfell, E., Bennett, B.J., Koeth, R., Levison, B.S., Dugar, B., Feldstein, A.E., Britt, E.B., Fu, X., Chung, Y.M., et al. (2011). Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. Nature *472*, 57–63.
21. Agus, A., Planchais, J., and Sokol, H. (2018). Gut Microbiota Regulation of Tryptophan Metabolism in Health and Disease. Cell Host Microbe *23*, 716–724.
22. McCann, J.R., and Rawls, J.F. (2023). Essential Amino Acid Metabolites as Chemical Mediators of Host-Microbe Interaction in the Gut. Annu. Rev. Microbiol. *77*, 479–497.
23. Guzior, D.V., and Quinn, R.A. (2021). Review: microbial transformations of human bile acids. Microbiome *9*, 140.
24. Zhu, L., Baker, R.D., Zhu, R., and Baker, S.S. (2016). Gut microbiota produce alcohol and contribute to NAFLD. Gut *65*, 1232.
25. Mbaye, B., Borentain, P., Magdy Wasfy, R., Alou, M.T., Armstrong, N., Mottola, G., Meddeb, L., Ranque, S., Gérolami, R., Million, M., and Raoult, D. (2022). Endogenous Ethanol and Triglyceride Production by Gut *Pichia kudriavzevii, Candida albicans* and *Candida glabrata* Yeasts in Non-Alcoholic Steatohepatitis. Cells *11*, 3390.
26. Carotti, S., Guarino, M.P.L., Vespasiani-Gentilucci, U., and Morini, S. (2015). Starring role of toll-like receptor-4 activation in the gut-liver axis. World J. Gastrointest. Pathophysiol. *6*, 99–109.
27. Ridaura, V.K., Faith, J.J., Rey, F.E., Cheng, J., Duncan, A.E., Kau, A.L., Griffin, N.W., Lombard, V., Henrissat, B., Bain, J.R., et al. (2013). Gut microbiota from twins discordant for obesity modulate metabolism in mice. Science *341*, 1241214.
28. Ramirez, J., Guarner, F., Bustos Fernandez, L., Maruy, A., Sdepanian, V.L., and Cohen, H. (2020). Antibiotics as Major Disruptors of Gut Microbiota. Front. Cell. Infect. Microbiol. *10*, 572912.
29. Gan, L., Feng, Y., Du, B., Fu, H., Tian, Z., Xue, G., Yan, C., Cui, X., Zhang, R., Cui, J., et al. (2023). Bacteriophage targeting microbiota alleviates non-alcoholic fatty liver disease induced by high alcohol-producing Klebsiella pneumoniae. Nat. Commun. *14*, 3215.
30. Das, B., Ghosh, T.S., Kedia, S., Rampal, R., Saxena, S., Bag, S., Mitra, R., Dayal, M., Mehta, O., Surendranath, A., et al. (2018). Analysis of the gut microbiome of rural and urban healthy Indians living in sea level and high-altitude areas. Sci. Rep. *8*, 10104–10105.
31. Wick, R.R., Judd, L.M., Gorrie, C.L., and Holt, K.E. (2017). Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput. Biol. *13*, e1005595.
32. Kanwal, F., Kramer, J.R., Mapakshi, S., Natarajan, Y., Chayanupatkul, M., Richardson, P.A., Li, L., Desiderio, R., Thrift, A.P., Asch, S.M., et al. (2018). Risk of Hepatocellular Cancer in Patients With Non-Alcoholic Fatty Liver Disease. Gastroenterology *155*, 1828–1837.e2.
33. Aron-Wisnewsky, J., Vigliotti, C., Witjes, J., Le, P., Holleboom, A.G., Verheij, J., Nieuwdorp, M., and Clément, K. (2020). Gut microbiota and human NAFLD: disentangling microbial signatures from metabolic disorders. Nat. Rev. Gastroenterol. Hepatol. *17*, 279–297.
34. Astbury, S., Atallah, E., Vijay, A., Aithal, G.P., Grove, J.I., and Valdes, A.M. (2020). Lower gut microbiome diversity and higher abundance of proinflammatory genus Collinsella are associated with biopsy-proven nonalcoholic steatohepatitis. Gut Microb. *11*, 569–580.
35. Nii, T., Maeda, Y., Motooka, D., Naito, M., Matsumoto, Y., Ogawa, T., Oguro-Igashira, E., Kishikawa, T., Yamashita, M., Koizumi, S., et al. (2023). Genomic repertoires linked with pathogenic potency of arthritogenic *Prevotella copri* isolated from the gut of patients with rheumatoid arthritis. Ann. Rheum. Dis. *82*, 621–629.
36. Honda, T., Ishigami, M., Luo, F., Lingyun, M., Ishizu, Y., Kuzuya, T., Hayashi, K., Nakano, I., Ishikawa, T., Feng, G.G., et al. (2017). Branched-chain amino acids alleviate hepatic steatosis and liver injury in choline-deficient high-fat diet induced NASH mice. Metab *69*, 177–187.
37. Basaranoglu, M., Basaranoglu, G., and Bugianesi, E. (2015). Carbohydrate intake and nonalcoholic fatty liver disease: fructose as a weapon of mass destruction. Hepatobiliary Surg. Nutr. *4*, 109–116.
38. Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., Formsma, K., Gerdes, S., Glass, E.M., Kubal, M., et al. (2008). The RAST Server: rapid annotations using subsystems technology. BMC Genom. *9*, 1–5.
39. Bag, S., Ghosh, T.S., and Das, B. (2017). Complete genome sequence of Collinsella aerofaciens isolated from the gut of a healthy Indian subject. Genome Announc. *5*, e013611-17.
40. Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., Alexander, H., Alm, E.J., Arumugam, M., Asnicar, F., et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat. Biotechnol. *37*, 852–857.
41. Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics *34*, i884–i890.
42. Enright, A.J., Van Dongen, S., and Ouzounis, C.A. (2002). An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. *30*, 1575–1584.
43. Letunic, I., and Bork, P. (2019). Interactive Tree of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. *47*, W256–W259.
44. Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. (2016). VSEARCH: a versatile open-source tool for metagenomics. PeerJ *4*, e2584.
45. Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics *30*, 2068–2069.
46. Seemann, T. (2015). Snippy: Rapid Haploid Variant Calling and Core SNP Phylogeny (GitHub).
47. Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and virus strains		
Collinsella aerofaciens (Healthy)	Fecal sample of healthy subjects	Bag et al., 2017[39]; https://doi.org/10.1128/genomeA.01361-17
Collinsella aerofaciens (NASH)	Fecal sample of NASH patients	Functional Genomics Lab., THSTI
Experimental models: Organisms/strains		
C57BL/6J mice	Small Animal facility of institute THSTI	IEAC/THSTI/120
Biological samples		
Human stool samples	NASH Patients at AIIMS Delhi	*IEC/NP-28/09.01.2015, OP-2/01.04.2016).*
Mice stool samples	Diet induced C57BL/6J mice	Small animal facility of institute
Mice blood samples	C57BL/6J male mice	Small animal facility of institute
Mice liver samples	C57BL/6J male mice	Small animal facility of institute
Metagenomic DNA	Fecal sample of NASH patient	Functional Genomics Lab., THSTI
Bacterial genomic DNA	C. *aerofaciens* bacteria	Functional Genomics Lab., THSTI
Chemicals, peptides, and recombinant proteins		
Mutanolysin	Sigma Aldrich	Cat#M9901
Lysostaphin	Sigma Aldrich	Cat#L7386
Lysozyme	Sigma Aldrich	Cat# L6876
Guanidine thiocyanate	Sigma Aldrich	Cat#G9277
N-lauryl sarcosine	Sigma Aldrich	Cat#L5777
TRIzol reagent	Takara	Cat#9109
Formalin	Sigma Aldrich	Cat#HT501128
Superscript III First-Strand Kit	Thermo Fisher	Cat#12574026
Trypticase soy broth	Sigma Aldrich	Cat#22092
Trypticase soy agar	Sigma Aldrich	Cat#22091
Defibrinated sheep blood	R-Biopharm Neugen Pvt. Ltd	Lot number - 260623
SYBR green master mix	Thermo Fisher	Cat#A25742
QIAquick gel extraction Kit	Quigen	Cat#28706
KAPA HiFi Hot Start Ready-mix	Roche	Cat#KK2602
AMPure XP beads	Backman coulter	Cat#A63881
Qiagen RNA Kit from tissue	Quigen	Cat#74104
RNA ladder	Sigma Aldrich	Cat#R7020
RNA ZAP	Thermo Fisher	Cat#AM9782
Nextera XT DNA Library preparation kit	Illumina	Cat#FC-131-1096
Oligonucleotides		
Deoxynucleotides	New England biolabs	Cat#N0447L
1. Mice NF-κB F,	Sigma Aldrich	*GAAATTCCTGATCCAGACAAAAAC*
2. Mice NF-κB R	Sigma Aldrich	*ATCACTTCAATGGCCTCTGTGTAG*
3. Mice TNF-β F	Sigma Aldrich	*GGTGCCTATGTCTCAGCCTCTT*
4. Mice TNF-β R	Sigma Aldrich	CCATAGAACTGATGAGAGGGAG
5. 1410F (Mice IL-6)	Sigma Aldrich	*GTCCTTCCTACCCCAATTTCC*
6. 1411R (Mice IL-6)	Sigma Aldrich	*GTCCTTGGAAACCCCAATTTCC*

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
7. Mice IL-10 F	IDT	*AAGGCAGTGGAGCAGGTGAA*
8. Mice IL-10 R	IDT	*CCAGCAGACTCAATACACAC*
9. 1418F (Mice Col-1-alpha)	IDT	*TAAGGGTCCCCAATGGTGAGA*
10. 1419R (Mice Col-1-alpha)	IDT	*GGGTCCCTCGACTCCTACAT*
11. 1406F (Mice alpha-SMA)	IDT	*GCGGGCATCCACGAAACC*
12. 1407R ((Mice alpha-SMA)	IDT	*GATCTTCATGGTGCTGGGTGC*
13. 1412F (Mice TGF-beta)	IDT	*CTTCAATACGTCAGACATTCGGG*
14. 1413R ((Mice TGF-beta)	IDT	*GTAACGCCAGGAATTGTTGCTA*
15. Mice FATP5 F	IDT	*CGGGTCATACAAGTGAGCAA*
16. Mice FATP5 R	IDT	*ATCACTGTTACGCCATGCTG*
17. 1408F(Mice IL-1beta)	*IDT*	*GCACGATGCACCTGTACGAT*
18. 1409R (Mice IL-1beta)	*IDT*	*CACCAAGCTTTTTTGCTGTGAGT*
19. Mice PPRA α F	*Sigma Aldrich*	*GTCCTCAGTGCTTCCAGAGG*
20. Mice PPRA α R	*Sigma Aldrich*	*GGTCACCTACGAGTGGCATT*
21. Mice PPRG γ F	*Sigma Alrich*	*AGTGGAGACCGCCCAGG*
22. Mice PPRG γ R	*Sigma Alrich*	*GCAGCAGGTTGTCTTGGATGT*
Software and algorithms		
1. VSEARCH program	Rognes et al., 2016[44]	https://github.com/torognes/vsearch
2. Qiime2-2021.11	Bolyen et al., 2019[40]	QIIME 2 user documentation — QIIME 2 2021.11.0 documentation
3. Snippy v4.6.0	Seeman, 2015[46]	(https://github.com/tseemann/snippy)
4. iTOL	Letunic and Bork, 2019).[43]	iTOL: Interactive Tree Of Life (embl.de)
5. Prokka v. 1.14	(Seemann et al., 2014[45]	(https://github.com/tseemann/prokka)
6. RAST	(Aziz et al., 2008)[38]	http://rast.nmpdr.org/
7. RaxML	(Stamatakis A et al., 2014)[47]	https://github.com/stamatak/standard-RAxML.
Other		
Choline deficient high-fat (CDHF) diet	Research Diet industry	Cat#D05010403

RESOURCE AVAILABILITY

Lead contact

Further information related to the manuscript, requests for any resources including bacterial isolates, and non-commercial reagents should be directed to the lead contact, Bhabatosh Das (bhabatosh@thsti.res.in).

Materials availability

Bacterial strains isolated from the healthy subjects and NASH patients are available with the lead contact, Bhabatosh Das (bhabatosh@thsti.res.in).

Data and code availability

- The next generation DNA sequencing data generated in the context of present study have been submitted to the National Center for Biotechnology Information (NCBI) databases. 16S rRNA gene sequences will be available in the BioProject: PRJNA882621 (Submission ID: SUB12083195). Whole genome sequences of *Collinsella aerofaciens* will be available in the BioProject: PRJNA882578.
- All original code is available in this paper supplemental information (Data S3)
- Any additional information required to reanalyse the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Patient's recruitment and sample collection

A cohort of biopsy-proved NASH patients (n = 10) enrolled in the All India Institute of Medical Sciences (AIIMS), New Delhi, was used in the current study to collect fecal samples for gut microbiome study. The 5 males and 5 females NASH patients with a mean age of 37 (sd ± 3.1)

years were selected for the current study. The subjects were vegetarians and non-vegetarians (Table 1). Power calculation indicated that empirical comparisons between obese and healthy subjects on the given sample sizes could still identify markers with an effect size of greater than 0.7 and a p-value (alpha) < 0.05. The AIIMS New Delhi ethics committee approved the study (IEC/NP-28/09.01.2015, OP-2/01.04.2016). Self-collected fecal samples were used for gut microbiota analysis following the methodology of our recently published study.[30] The fecal samples of obese patients (n = 17) were collected by the AIIMS, New Delhi, from the people living in the NCR, India. Healthy subjects were selected from our recent study.[30]

Mouse housing conditions and sampling

Male mice (C57BL/6; 8–12 Weeks) were housed in a cage (4/cage) kept at a temperature of 23°C ± 2°C, with a 12-h light/dark cycle and habituated in the room for 1 week before experiments. After 1 week of adaption, animals weighing 20–30 g (n = 32) were divided into 4 groups with unrestricted access to food and water. All animal experiments were performed in accordance with the Institutional Animal Ethics Committee (IAEC) of Translational Health Science and Technology Institute, India, for the Care and Use of Laboratory Animals (IAEC No.120). For the induction of experimental NAFLD, the mice were divided into four groups: Normal Chow, Normal Chow with *Collinsella aerofaciens*, CDHF, and CDHF with *Collinsella aerofaciens*. They were fed either a chow diet or a choline-deficient high fat (CDHF) diet and received 200 μL of ~10^8 CFU of *Collinsella aerofaciens* on the 4th and 8th weeks. After being observed for 10 weeks for NAFLD induction, the mice were euthanized after being fasted overnight. Liver and stool specimens were collected in 10% formalin (pH 7.2) and NAPG buffer for histopathological analysis (H&E staining; MT staining) and metagenome sequencing, respectively.

METHOD DETAILS

DNA extraction and metagenomic sequencing

Prior to extracting genomic DNA from the stools, the samples were stored at −80°C. THSTI method[41] was used to extract DNA from approximately 200 mg of frozen fecal samples. The purity and content of DNA were evaluated using a Biospectrometer (Eppendorf, Germany) and 0.8% agarose gel electrophoresis. Variable regions V1-V5 of the 16S rRNA genes were amplified in 50 μL of reaction volume using 0.1 ng of template DNA with 27F (C1) and 926R (C5) primers. The 950-bp PCR products were gel purified using the QIAquick gel elution Kit (Qiagen, Germany). At THSTI in India, equimolar concentration amplicon libraries were assembled and sequenced using a Roche 454 GS FLX+ pyrosequencer. For 16S rRNA gene sequencing through Miseq Illumina, extracted DNA was quantified with Qubit fluorometer to obtain a correct amount of the dsDNA. Genomic DNA with concentration of 5 ng/μl was selected for library preparation for 16S rRNA gene sequencing in the Illumina MiSeq NGS platform. For PCR amplification, 25μL reaction contained 2x KAPA HiFi Hot Start ReadyMix, both the primers and 2.5μL of the template DNA were used. Amplified products were cleaned using AMPure XP beads, tagged with Illumina sequencing adapters and created the sequencing library.

Processing 16S rRNA raw reads

The next generation amplicon sequencing produced a total of 534797 processed reads (>Q20), with an average of 17105 reads per subject (range varying between minimum: 7180 to maximum: 44870 reads per subject). The average read length obtained was 754 bps, which covers V3-V5 regions of the 16S rRNA gene (cohort averages ranging from 327 to 997 bps) (Table 3). In the case of NASH samples, 16S rRNA (V3-V4 region) reads were generated based on next-generation targeted amplicon sequencing by Illumina MiSeq. A total of 3358404 quality passed pair-end reads were generated, and an average of 335840 pair-end reads per sample were used for merging. The merging of pair-end reads per samples was between ~31% and ~41%, except for a single sample with only ~13% merged reads. After merging the paired-end reads, the average read length is ~254 ranging from a minimum of 50 to maximum of ~469 base pairs. The reads of obese and NASH samples were generated by two different platforms. To avoid the discrepancy in performing the OTU generation, a representative merged read having minimum length (i.e., 440 base pairs) was used to trim Roche454 reads. Further, the trimmed reads were used for OTUs generation.

Analysis of 16S rRNA reads

The composition and diversity of gut microbiota was analyzed based on the 16S rRNA sequences. The read quality was checked, cleaned and qualified paired-end reads were merged with fastp program (Chen et al., 2018).[41] These parameters filter all reads with a Phred score ≥20 and a minimum read length ≥50. The chimeric sequences were analyzed in merged reads to find the chimeric component by the VSEARCH program[44] Since 454 Roche produces multiplexed two separate files (i) read quality file and (ii) read sequence file (in fasta format) in which all samples are pooled. These two files were merged to generate fastq file. Now, the pooled reads were de-multiplexed sample-wise based on index sequences by in-house developed Perl script. Next, the quality of all reads was examined by fastQC program (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and preprocessed by fastp program for read quality assessment to generate adapter-free and high-quality reads. Qiime2-2021.11 pipeline[40] and SILVA 138 SSU rRNA reference database were used to generate operational taxonomy units (OTUs). Hence, the nucleotide sequence of each OTUs were again searched in NCBI (16S rRNA database) by megaBLAST program. Only the best hit was selected based on ≥99% query sequence identity percentage and 95% coverage with the highest bit-score. To perform all the above-mentioned steps, in-house Perl scripts were written for processing of megaBLAST results, splitting and merging of the OTU tables. The statistical and microbial alpha diversity analyses

among three groups were analyzed by ampvis2 R-package and ggplot2. Standard strain *C. aerofaciens* submitted at NCBI was used as control (CP024160.1).

Gut microbial interactions analysis

The Pearson correlation coefficient (r) was used to calculate the pairwise correlations between the read abundance of the core taxa (n = 12) present in each of the three groups independently. To perform the analysis the cut off was set to $r \geq 3$. The significant interaction was indicated by the p value <0.05. Pearson correlation was calculated using the functions cor () in the R environment 4.1.2. The heatmap was generated using the packages Hmisc v 5.1–1, tidyverse, ggplot2 and gplots and RColoRBrewer. The corr package, igraph, and ggraph were used to build the correlation network in R studio.

qRT PCR assay

A Bio-Rad CFX100 sequence detection equipment and a Bio-Rad SYBR Green kit were used to conduct quantitative PCR (Bio-Rad). The specificity of the amplification products was confirmed by melting-point determination analysis. Total DNA was extracted from fecal samples of both cohorts adopting THSTI method (Bag et al. 2017[39]). The copy counts of target genes in each sample were estimated by comparing the Ct values obtained from the standard curves with the control gene. The *recA*-specific primers recA-F/recA-R was employed as an internal control in each experiment. The results are given as the averages of qPCR data performed in triplicate. The fold change and p value derived with the t-test were then used to identify differentially expressed mRNA levels.

Isolation, identification and cultivation of *Collinsella aerofaciens*

All the strains of C. *aerofaciens* were isolated from the fecal sample of biopsy-proven NASH patients. A 200-mg fecal sample was homogenized in 1 mL of phosphate-buffered saline (PBS), sequentially diluted in the same buffer, and plated on a pre-reduced Trypticase soy agar plate (pH 7.3) enriched with 5% (v/v) defibrinated sheep blood. Utilizing 4 to 5 glass beads (3.00 mm), bacterial cells were distributed throughout the surface of the plates. Plates were incubated for 42 h at 37°C in an anaerobic workstation (Whitley DG250) filled with 80% N_2, 10% CO_2, and 10% H_2. Discrete colonies of *C. aerofaciens* were cultivated in 5 mL of tryptic soy broth (TSB). The confirmation of *C. aerofaciens* was done using Sanger sequencing of 16S rRNA that sowed greater than 98% sequence identify, and screening of colonies using *C. aerofaciens* specific sets of primers. A spectrophotometer was used to monitor the growth of the cells in TSB.

Whole genome sequencing

The genomic DNA of *C. aerofaciens* was extracted using our laboratory-optimized DNA extraction method. The concentration and purity of genomic DNA were determined using Nanodrop 2000 (Thermo Fisher) and agarose gel electrophoresis, respectively. Genomic DNA with an OD260/OD280 ratio between 1.8 and 2.0 and no significant RNA or protein contamination was selected for whole genome sequencing in Illumina MiSeq high-throughput sequencing platform (Illumina, Inc., USA). Approximately 100 ng of genomic DNA was used for sequencing library preparation. The Nextera XT DNA Library preparation kit was used for pair-end sequencing (Illumina, Inc., USA).

Genome assembly and annotations

Trimmomatic sequence analysis tool version 0.39 (http://www.usadellab.org/cms/?page=trimmomatic) was used to assess the quality of sequencing reads generated in our lab using the next-generation DNA sequencing platform Illumina MiSeq. Using GS *De Novo* Assembler, genome assembly was performed on filtered sequencing reads with a quality score >20 (Version 3.0). The scaffolds were verified using the Basic Local Alignment Search Tool (BLAST) tool. The genomic scaffolds were input into the Rapid Annotation using Subsystem Technology (RAST) server (version 2.0).[38] Using the RAST database, subsystem of metabolic pathways of all the genomes was analyzed.

Phylogenetic and pan-genome analyses

Whole genome sequences of *C. aerofaciens* (n = 96) available in the NCBI genomic database (n = 96, Data S1) were downloaded and included in this study for phylogenetic analysis of the current study isolates (n = 7). The assemblies were mapped against the reference genome of *C. aerofaciens* isolated from a healthy individual (Accession No. NZ_CP024160.1) using Snippy v4.6.0.[46] Genomic polymorphisms of core genome sequences were identified based on the genome sequence of a reference strain (Accession No. NZ_CP024160.1), followed by constructing a maximum likelihood (ML) phylogenetic tree by RaxML with the GTRGAMMA model.[47] The phylogeny was then rooted against the reference genome and further annotated using iTOL.[43] Any changes in single nucleotide polymorphisms (SNPs), insertions or deletions (in-dels) were captured using Snippy v4.6.0. pipeline with variant calling features. These variations have been quantified with respect to their effect measured as synonymous, missense or frameshift mutations. For the pan genome analysis, the genomes of *C. aerofaciens* were annotated using Prokka v. 1.14.[45] The annotated GFF file obtained from prokka was used to assess the pan genome of the 103 *C aerofaciens* using roary pipeline.[42] In addition to the default protein identify level of 95%, a gradual decrease of 5% identity from 95% to 35% has been carried out to evaluate the difference in the distribution of core genes.

Animal experiment

Total 32 healthy C57BL/6J male mice (8–12 weeks old with body weight 20–30 g) were obtained from a breeding colony at the animal facility of Translational Health Science and Technology Institute (THSTI), in accordance with Institutional Animal Care and Use Committee (IACUC) guidelines. There was total four groups with (n = 8) mice in each. Mice were housed at 4 animals per cage in a 12-h light/12-h dark cycle with *ad libitum* access to food and water at a controlled temperature (23 C±2°C). All C57BL/6J mice were bred and maintained in a germ-free mouse facility in sterile isolators and screened for bacterial, fungal, and viral contamination. Natural ingredient NIH #31M Rodent Diet as a standard Chow diet, the composition can be found on the website (Taconic Biosciences).

Development of animal models for non-alcoholic steatohepatitis (NASH)

For the induction of NAFLD, choline deficient high-fat diet (CDHF) (Research Diet, USA) along with *C. aerofaciens* biomass were administered in mice for 10 weeks. Stool samples of the biopsy proven NASH patient (NAS score = 1–4) were collected and *C. aerofaciens* was isolated. Mice were gavage with single doses of *C. aerofaciens* strains suspended in 1X PBS (10^8 CFU, 200 mL) twice subsequently for every two weeks as shown in the diagrams (Figures-S3A and S3B). Furthermore, mice fed with normal chow were also administered with *C. aerofaciens* similarly as described before. Two groups of mice were fed with normal chow and CDHF diets for 10 weeks to compare the effect of *C. aerofaciens*. All mice were survived till 10 weeks. The body weights and feed intake of mice were monitored on a weekly basis. At the end of the study period (70 days), all mice were sacrificed and different organs were stored at −80°C or in formalin for biochemical, molecular biology, and histopathology study.

Liver histology

Livers were obtained from different experimental groups at time point i.e., 10 weeks, was subjected to histopathology for observing lipid accumulation, tissue inflammation and fibrosis. Liver tissue was fixed in 10% formalin, routinely processed and embedded in paraffin. Paraffin sections were cut and mounted on glass slides and stained with hematoxylin and eosin (H&E), Masson's trichrome stain, which were examined under a light microscope. Histological liver assessments were done by a Pathologist from AIIMS, based on the NAS score.

Measurement of gene expression related to inflammation, fibrosis and fat accumulation in liver

Real-time qPCR was employed to detect hepatic expression of some genes related to inflammatory parameters (*TNF-alpha, IL-6, NF-κβ, IL-10*), fibrosis markers (*Col-1-alpha, alpha-SMA, TGF-beta*) and fat accumulation genes (*FATP5, PPRA, PPRG*) using specific primers targeting the interest genes. Briefly, the total RNA was isolated from control and NAFLD livers with TRIzol reagent (Takara). Reverse transcriptase reactions were performed using the Superscript III First-Strand Kit (Invitrogen) for first-strand cDNA synthesis. Primers for real-time quantitative PCR (qPCR) analysis were designed using published sequence information, avoiding regions of homology with other genes. qPCR of above-mentioned genes was performed from the same cDNA. For each gene, 10 ng of cDNA was analyzed on an CFX100 (Bio-Rad) using SYBR Green Master mix (Bio-Rad). Fold-change analysis was normalized to 18S transcript levels for each sample.

Measurement of hepatic hydroxyproline and triglyceride

The estimation of triglycerides in mice hepatic tissue were performed using Randox kit (REF:TR1697). A total amount of 30mg hepatic tissue were required to perform the assay. Tissue samples were resuspended and homogenize in 500 µl of 5% NP-40 solution using hand homogenizer. A uniform mixture of hepatic tissue was heated at 80–100° C in a water bath for 2–5 min. The heated mixture was allowed to cool down at room temperature in ice-bath. A cycle of heating and cooling were performed three times to solubilize all the triglycerides in solution. After centrifugation supernatant of solution were collected and 10ul of sample was used in 96 well plate along with calibrator(standard). Further the absorbance of sample were taken at 500nm as per kit protocol. The estimation of hydroxyproline was performed using the kit (cat#no- E-BC-K062-M). 100mg of wet tissue were used to carried out the procedure. Initially the tissue sample were resuspended in 1mL solution of 6M HCL to hydrolyze completely at 95°C for 6 h. After heating, solution was allowed to cool down under running water. Further, the pH were adjusted at 6.5 to 7 by using kit reagent 7 and 8. All the further steps were followed based on kit instruction.

Measurement of serum ethanol level

The detection of ethanol in mice serum sample was performed through GC/MS technology. The samples incubation temperature was kept at 120°C with the equilibration period of 5 min. Injection time was 0.15 min at a pressure of 135 kPa, with the injector needle held at 125°C. Pressure was built-up during 0.9 min with a dwell time of 0.8 min. The temperature of the transfer line was maintained at 225°C, the injector port's temperature in the split-less injection mode at 95 kPa, with the maintenance of septum purge flow rate 20 mL/min. Separation of the compounds was achieved using an RTX-5 capillary column (Agilent, California, USA). Helium was used as a carrier gas at a constant pressure of 95 kPa. After maintaining the initial oven temperature of 60°C for 0.2 min, the temperature increased by 10°C/min to a final temperature of 60°C, that remained for 1 min, thereafter increasing by 2.5°C/min, 10°C/min, and 120°C, which had been sustained for 1 min. Each sample's total runtime was 732 s. Electron impact ionization (EI) was applied at 70 eV. Temperatures of the ion source were set to 230°C. Chromatograms of the samples were captured in full scan mode from 25 to 500 u at a acquisition rate 50 spectra/second. Data evaluation was carried out using the ChromaTOF Software (4.51.6.0).

Ethical approvals

The Human Ethics Committees of All India Institute of Medical Sciences (AIIMS) New Delhi has approved this study (Ref.# IEC/NP-28/09.01.2015, OP-2/01.04.2016).

Statistical analysis

Co-occurrence network of gut microbiota was analyze and tested with Pearson's correlation with a cut off $r \geq 3$, p-value<0.05.