Go beyond “abundance”: cross-cohort single-nucleotide-variant profiling of gut microbiota suggests a novel gut-health assessment approach

Chenchen Ma
Hainan University
Yufeng Zhang
The University of Hong Kong
Shuaiming Jiang
Hainan University
Fei Teng
Qingdao Stomatological Hospital Affiliated to Qingdao University
Shi Huang
The University of Hong Kong
Jiachao Zhang (zhjch321123@163.com)
Hainan University

Research Article

Keywords: Single nucleotide variation, Gut microbiome, Mutation bias, Short-chain fatty acids, Gut microbiome health index, SNV rate

Posted Date: June 17th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1747065/v1

Abstract

Background: The adaptive evolutionary changes can precede the ecological changes in the gut microbial communities under constant host selection pressure, yet their association with host diseased status was underexplored.

Results: Herein, we performed a meta-analysis of 1711 gut metagenomic samples from 16 case-control studies spanning 12 human diseases to systematically explored shared disease-associated single nucleotide variants (SNVs) in gut microbes, and included an additional 446-member cohort for validation. Overall, healthy individuals carried more mutated resident gut microbes, and more SNVs, mainly involving the short-chain fatty acids (SCFAs) producing bacteria. Furthermore, the widespread differences in base mutation bias of gut microbes were observed between health and nonhealthy subjects suggesting divergent gut microbial evolutionary directions under different host medical conditions. We further found that nonsynonymous SNVs can lead to functional inactivation of five SCFA-production genes in non-healthy populations, among of which two genes (i.e., ack, and scpC) from *Faecalibacterium prausnitzii* C, and *Bacteroides stercoris* respectively were externally validated. Subsequently, we developed a novel gut microbiome health index (GMHI) based on the SNV rate of all mutated strains, classifying host health states with 74.23% accuracy, validated with high accuracy (AUROC=69.28%).

Conclusions: Collectively, our study highlights the importance of employing the genetic variability in gut microbiome to characterize the gut microbial adaptation that can also predict human chronic diseases.

Introduction

Gut microbes consistently evolve by gaining the adaptive SNVs, indels and structural variants (SV) in response to the gut selection pressures derived from host physiological changes, dietary changes, antibiotic use, and probiotics interventions [1–5]. In the last decade, many SNVs, indels and SVs of gut microbiota have been reported, which underlie the adaptive evolution of microbial functional genes that can associate with phenotypic changes in both microbes themselves and hosts [6, 7]. However, more studies are still required to understand the population-level genetic processes of gut microbiota under a broad spectrum of host health conditions [8].

Recently, a range of metagenomics studies assessed the genetic variation of host gut microbes and reported characteristic SNVs associated with a few chronic diseases. A set of SNVs in specific gut microbial residents have been associated with host health status, such as *Faecalibacterium prausnitzii* and *Eubacterium rectale* (colorectal cancer [9], liver cirrhosis [10], graves' disease [11]), *Bacteroides vulgatus* (tuberculosis [12], graves' disease [11]). These bacteria usually carried a large number of SNVs, and their SNV profiles were different between healthy and nonhealthy groups. Many studies also implied that a small number of genetic mutations or even a single SNV in the microbial genome can significantly alter the pathogenic behavior of gut bacteria and affect host health. For example, T2D patients often possessed SNVs enriched on the glycosyl hydrolases gene of *Bacteroides coprocola*, which was recognized as an important gut-microbiome-derived therapeutic target for T2D [13]. Furthermore, it is evident that a single SNP (G84E) of *Escherichia coli* can disturb gut lysophospholipid homeostasis and induces host inflammation by epithelial barrier disruption [14]. Nevertheless, the consensus microbial SNV signatures associated with a wide range of human chronic diseases have never been attempted. Furthermore, no studies have systematically assessed if these widespread SNVs in the gut microbiota from multiple diseases link to any core health-related functions to hosts. For example, short-chain fatty acids (SCFAs) produced by gut bacterial fermentation of dietary fibers, are widely considered as key bacterial metabolites regulating host immune response or anti-inflammatory factors [15] [16]. Associations between fecal SCFA level and host health have been widely found in a variety of diseases,

including COVID-19 [17], type 2 diabetes (T2D) [18], colorectal cancer (CRC) [19], Crohn's disease (CD) and ulcerative colitis (UC) [20], Parkinson [21], Polycystic Ovary Syndrome (PCOS) [22], Diabetic Nephropathy [23], Encephalitis [24] etc. We hypothesized that a metagenomic meta-analysis of disease-related datasets would identify cross-disease microbial SNV signatures which can also link to core microbial functional genes modulating the host health.

The large population size in the meta-analysis is required to consolidate the SNV findings and further uncover potential impact of SNV in intestinal microbes on the nonhealthy hosts. Currently, the number of public metagenomes in repositories is growing exponentially [25]. The meta-analysis of publicly accessible shotgun metagenomic data is an economical and powerful way [26] to identify universal gut-microbiota-derived biomarkers for multiple chronic diseases. Importantly, going beyond conventional abundance data, profiling the genetic variability in gut microbes can expand our understanding of the evolutionary and ecological processes in the gut underlying the disease development. However, it is still technically challenging to compare the SNV profiles across samples by the rigorous consideration of variation in sequencing depth and coverage among samples.

To address above challenges, we systematically studied the SNVs in gut resident microbial strains that can associate with host selection pressures under 12 human medical conditions. We attempted to test if the evolutionary directions for individual gut microbes differ between healthy and non-healthy subjects. Most previous studies reported a decrease in the relative abundance of SCFAs-production gut strains in nonhealthy subjects. More importantly, we found that SCFAs production of nonhealthy individual intestinal tract was inactivated due to adaptive mutations in related genes more prone to turn codons into terminators, further suppressing the overall concentration of SCFAs in the GI tract. Finally, we established the "Gut Microbiome Health Index" (GMHI) using SNV profiles and benchmarked its prediction performance against that derived from the species-level abundance profiles conventionally used in most past studies. Notably, SNV profiles of individual gut microbes exhibited a strong predictive power in evaluating the health status of human hosts.

Results

Dataset collection and meta-analysis workflow

In this meta-analysis, we collected 1711 samples from 16 metagenomics studies spanning 12 host phenotypes, including 919 nonhealthy and 792 healthy human individuals (**Figure 1A, Additional file 2: Table S1**). Notably, subjects were categorized into "healthy" or "nonhealthy" according to the original studies, while overweight and obesity alone individual were not classified as nonhealthy group. Expectedly, we observed beta-diversity differences based on species-level fecal microbial abundance between healthy and nonhealthy groups using Bray-Curtis dissimilarity (R=0.002, p<0.001, Adonis, **Figure 1B,** 95% confidence regions), consistent with the past results in most studies.

Subsequently, we sought to explore the disease-associated genetic variation in gut microbiota by mapping metagenomics reads against a comprehensive set of gut microbial reference genomes. If we identified any SNVs from a microbial reference genome, this "mutated genome" or SNV-carrying microbial strain will be further analyzed to characterize its SNV profile and test if it can associate with disease status (**Additional file 2: Table S2**). We identified a total of non-redundant 2740 mutated genomes in 1711 samples. Here, we calculated the distribution of average coverage and breadth of these mutated genomes with prevalence greater than 10% (233, 8.50%), as meaningful comparison of genomic variation between samples requires both breadth and depth of sequencing for each genome [6]. In this study, the sequencing coverage of each SNV was required to be 5, minimally, which could cover most genomes (average coverage of 158 genomes more than 5, 67.81%) (**Additional file 1: Figure S1**).

Interestingly, we next observed the beta-diversity difference based on the composition of SNV-carrying strains (N=2740) between healthy and nonhealthy groups (R=0.003, p<0.001, Adonis, **Figure 1C**, 95% confidence region) based on Euclidean dissimilarity matrix. This suggested that host medical conditions can associate with not only population-level abundance changes but also genetic compositional changes in the gut microbiota. Next, we focused on specific strains, genes and SNVs, showing the association between gut microbial SNVs and host health status.

The healthy gut microbiome harbors a wider range of SNV-carrying strains

According to InStrain pipeline for analysis of co-occurring genome populations ("Methods"), we obtained the comprehensive SNV profiles of gut microbiota for both healthy and nonhealthy cohorts. Of note, in this study, the number of strains and SNV count was normalized based on metagenomic sequencing depth ("Methods", **Additional file 1: Figure S2**).

We first explored whether any difference in the richness of mutated strains, and the total number of SNVs between healthy and diseased states. Firstly, we observed that healthy subjects (19.48±7.38) had more mutated strains than diseased subjects (17.39±6.84) (**Figure 2A**, **Additional file 2: Table S3**, p<0.001), suggesting a higher strain-level diversity in the healthy gut microbiota. Respectively, such a difference pattern was also found in a total of six independent cohorts corresponding to six diseases, including AS (16.92±5.64 vs. 21.93±7.23, p<0.001, average±std), CD (17.33±6.87 vs. 24.76±5.77, p<0.001), GD (14.78±3.75 vs. 17.61±3.37, p<0.001) and LC (19.51±6.01 vs. 23.77±6.51, p<0.001) and SF (17.56±3.81 vs. 25.51±3.0, p=0.016). In contrast, CRC individuals had a higher strain-level diversity than healthy control (22.26±6.80 vs. 20.73±7.11, p=0.025). There were no significant differences in other cohorts, including BC (11.10±2.66 vs. 11.83±2.83), Cov19 (14.29±7.41 vs. 16.28±10.84), PCOS (15.32±3.87 vs.15.6±14.67), SCZ (11.56±3.71 vs.11.29±3.45), T2D (19.13±5.61 vs. 19.78±5.18) and UC (22.91±8.62 vs. 24.76±5.77).

Most studies have reported differences in the species-level alpha diversity of healthy and nonhealthy cohorts. Here, our integrated analysis showed that alpha diversity (using either Shannon or Simpson index) is different between host disease states (**Figure 2B**, p<0.001, Wilcoxon rank-sum test). Intriguingly, the richness of mutated genomes strongly correlates with Shannon index (R=0.652, p<0.001), or Simpson (R=0.596, p<0.001) (**Figure 2C**. One possibility is that lower intestinal microbial diversity leads to a smaller number of mutated genomes. To address these concerns, we matched healthy and nonhealthy patients with alpha diversity within 0.01 (Shannon index) or 0.001 (Simpson index) differences to compare their normalized number of mutated genomes (92.87% and 92.69% samples were included), and Shannon and Simpson index were 0.01 or 0.001 higher in nonhealthy group than in healthy. In such manner, we found that when the alpha diversity is almost the same, the number of mutated genomes is different (Wilcoxon rank-signed test, p<0.001 for samples with equalized Shannon index, and p=0.023 for samples with equalized Simpson index). Therefore, in most cases, we believe that healthy individuals have higher strain-level diversity than nonhealthy individuals (**Figure 2D, Additional file 2: Table S4**). At the same time, we also matched individuals with < 0.01 difference in the normalized number of mutated genomes, and compare their alpha diversity (27.47% samples were included) between host groups. Interestingly, we still observed the disease-related difference in Simpson index at the species level (p=0.032), where healthy group had less microbial species (**Figure 2D, Additional file 2: Table S4**). Therefore, healthy subjects can reach equal strain-level diversity with nonhealthy subjects when they had less microbial species in the gut microbial community. Overall, healthy individuals tend to have higher strain-level diversity than most nonhealthy individuals, expect CRC patients.

Next, we estimated the SNV number per Mkb sequencing data for all mutated genomes in both healthy and nonhealthy gut microbiota, and demonstrated healthy individuals have more SNVs in total than nonhealthy

individuals (23.44±9.93 vs. 25.20±10.11, **Figure 2E,** p <0.001). Respectively, we validated such a difference in a total of five diseases/cohorts, including AS (212.85±67.13 vs. 264.80±65.15, p<0.001), CD (211.48±81.11 vs. 301.02±61.23, p<0.001), GD (213.46±54.13 vs. 241.71±56.14, p=0.012) and UC (257.86±107.20 vs. 301.02±61.23, p=0.002). By contrast, CRC individuals have more SNVs than healthy control (260.39±75.34 vs. 241.52±71.09, p=0.045). No significant differences were found in other cohorts, including BC (148.22±40.47 vs. 156.71±49.76), Cov19 (190.14±125.20 vs. 260.43±238.00), LC (270.74±77.99 vs. 289.17±64.32), PCOS (214.99±73.86 vs. 216.58±65.64), SCZ (148.39±62.71 vs. 135.49±58.73), SF (604.81±280.69 vs. 781.80±83.49) and T2D (371.49±78.86 vs. 339.52±79.23). Given the genetic compositional difference found at such a higher level between disease states, SNV profiles on specific strains/mutated genomes need to be revealed further.

Strain-level diversity associated with multiple host health status

To further explore if disease states can also associate with the SNV profiles for individual microbial strains and other key evolutionary patterns for gut microbiota, we comprehensively analyzed all 75 strains with a mutation frequency of more than 30% in the 1711-member population (**Figure 3),** and the corresponding genome ids of these strains can be found in **Additional file 2: Table S2**. Firstly, we compared the prevalence of these strains (**Additional file 2: Table S2**), and the top-four most prevalent strains in both healthy and unhealthy groups were: *Bacteroides dorei* (non-healthy: 87.16%, healthy: 91.92%, all: 89.36%), *Bacteroides uniformis* (non-healthy: 76.50%, healthy: 86.87%, all: 81.30%), *Faecalibacterium prausnitzii* G (non-healthy: 65.18%, healthy: 77.90%, all: 71.07%) and *Parabacteroides distasonis* (non-healthy: 69.42%, healthy: 71.21%, all: 70.25%). 46 strains were more prevalent (46/75, 61.33%) in the healthy group, while 5 strains (5/75, 6.67%) were more prevalent in the nonhealthy group. At the species level, many strains from *Bacteroides* consistently have a high frequency of SNVs occurred in the healthy gut. At the strain level, *Faecalibacterium prausnitzii* G, *Faecalibacterium prausnitzii* D, *Faecalibacterium prausnitzii* C, *Faecalibacterium prausnitzii* K and *Faecalibacterium prausnitzii* E also presented diverse patterns between healthy and nonhealthy population. Overall, the above results also confirm that the healthy group will have a wide range of mutations in the genome (**Figure 2A**).

We next investigated if SNV rate (“Methods”, Eq 1) and microdiversity of these gut microbial strains can associate with disease states. “SNV rate” is defined as, the relative frequency of SNVs on a given genome that can compare with other genomes. Microdiversity is a measurement of intra-population genetic diversity for a microbial strain in the gut microbiota (**Additional file 1: Figure S3)**. Clearly, the SNV rate and microdiveristy substantially varied among microbial species even strains (e.g., *Faecalibacterium prausnitzii*). We identified 15 strains (out of 75 strains, 20%) that had a SNV-rate difference between disease states. Among these 15 strains, 8 strains (8/75, 10.67%) were higher in the nonhealthy group and 7 (7/75, 9.33%) strains were higher in the healthy group. Remarkably, eight out of 15 strains (8/15, 53.33%) were mainly from Lachnospiraceae, including *Blautia wexlerae*, *Agathobacter rectalis*, Anaerostipes hadrus, *Lachnospira eligens*, *Blautia* sp900066165, *Faecalicatena torques*, *Roseburia intestinalis*, KLE1615 sp900066985, a typical group of bacteria that produce SCFA (i.e., acetate and butyrate). In contrast, *Bacteroides* were the main genus (6 strains, 6/20, 30%) with microdiversity difference between disease states, including *Bacteroides dorei*, *Bacteroides uniformis*, *Bacteroides xylanisolvens*, *Bacteroides sartorii*, *Bacteroides thetaiotaomicron* and *Bacteroides massiliensis*, and only three strains (3/75, 4%, 2 *Bifidobacterium*) were enriched in the nonhealthy group while 17 (17/75, 22.67%) were enriched in the healthy group. Overall, these results confirmed that healthy individuals have more SNV in their gut microbial genomes (**Figure 2E**).

Universal base mutation bias associated with host health states

Next, we narrowed down to the specific pattern of mutation types at the single SNV level (i.e., nucleotide diversity) since the SNV rate and microdiversity of 75 gut strains were associated with disease states. Firstly, we profiled six mutation types of all 1771 samples (**Additional file 2: Table S5**), and compared alpha diversity and beta diversity based on mutation-type profiles across samples. Intriguingly, Shannon diversity for the mutation types (p=0.03) had significant differences between two groups (**Figure 4A**). In addition, the mutation-type composition of 75 gut microbial strains was primarily clustered by host disease states (R^2=0.004, p<0.001, PCoA1=50.72%, PCoA2=12.36%) (**Figure 4B**). Specifically, we found that two mutation types (A>G|T>C, p=0.006 and A|T>T>A, p=0.011, Wilcoxon rank-sum test) were enriched in the healthy group while one in the nonhealthy group (C>G|G>C, p<0.001, Wilcoxon rank-sum test) (**Figure 4C, Additional file 2: Table S5**). We also showed and compared the mutation types between host states for each strain (**Additional file 2: Table S6**). Overall, we observed the diverse distribution patterns of mutation types among microbial strains (**Figure 4D**). C>G|G>C in many strains tend to be significantly higher in nonhealthy people (31/75, 41.33%), while such mutation type in only three strains were higher in the healthy group (3/75, 4%). In contrast, the A>G|T>C mutant type was more common in multiple strains in the healthy group (16/75, 21.33%).

We next compared the overall percentage of adaptive mutations occurring in the gene coding region for all 75 strains, and no differences between disease states were found (89.84% vs. 89.75%, p=0.1, **Figure 4E**). However, at the strain level, we observed mutations in coding regions of 32 strains were significantly higher in the nonhealthy group than in the healthy group (32/75, 42.67%), while only three strains (*Bacteroides* sp.) were higher in the healthy group (3/75, 4%). This suggested that disease states can change SNV patterns in coding regions of gut microbial strains.

Strain-level codon mutation bias in SCFA-production-involved genes

We next focused on key functional genes of specific gut strains and assessed potential functional changes in gut microbiota induced by microbial adaptive evolution under disease conditions. SCFA is widely considered to be closely related to human health, and SCFAs abundance deficiency has been observed in a variety of diseases, and the genetic evolution process of SCFAs related genes has not been reported so far. We then first counted the enzymes encoded by these strains that are related to SCFAs production, including acetate kinase (ack, E.C. 2.7.2.1), Propionyl-CoA:succinate CoA transferase (scpC, E.C. 2.8.3.-) and Butyrate kinase (buk, E.C. 2.7.7.7), acetate CoA-transferase YdiF (ydiF, E.C. 2.8.3.8), acetate CoA-transferase subunit alpha/beta (atoD/A, E.C. 2.8.3.8), butyrate–acetoacetate CoA-transferase subunit A/B (ctfA/B, E.C.2.8.3.9) and butanoate coenzyme A-transferase (E.C. 2.8.3.-, BcAt) (**Figure 5A**).

We next presented the SNV profile of 75 strains encoding the above SCFAs-related enzymes (**Figure 5B**). We observed that almost all strains encode ack gene (74/75, 98.67%), involved in the acetate production. 74.67% strains have at least one metabolic pathway associated with butyrate production, and six strains have at least two pathways. In addition, *Bacteroides* spp. are the main producers of propionate (16/21, 76.19%). Then, we compared whether the pN/pS values of these genes were different between healthy and nonhealthy cohorts to assess the size of intestinal selection pressure. Specifically, the pN/pS value of ack gene of 11 strains was different between host groups, suggesting this gene can be under positive selection due to medical conditions. Its pN/pS ratio in four strains was higher in the healthy group, while seven strains were higher in the nonhealthy group. Differential intestinal selection pressures related to diseases drove the production of butyrate by related genes in 10 strains. Only scpC gene of *Bacteroides stercoris* had a difference in pN/pS, and the nonhealthy group was higher than the healthy group. Overall, the adaptive evolutionary patterns of SCFAs-associated genes in different strains are diverse due to host health status.

Next, we explored if and how these SNVs specifically modulated the production of SCFAs. Here, we compared the frequency that a SNV on SCFAs-related gene that causes codon to become termination codon or the start codon was inactivated (**Figure 5C**). Intriguingly, those SNVs were mainly found in four strains (*Faecalibacterium prausnitzii* C, *Faecalibacterium prausnitzii* D, *Bifidobacterium pseudocatenulatum* and *Bacteroides stercoris*). The frequency of SNV-derived codon changes in the nonhealthy group was significantly higher than that in the healthy group due to the disease-related selection pressure (**Additional file 2: Table S7**). Respectively, two genes of *Faecalibacterium prausnitzii* C possessed codon mutation bias affecting initiation and termination codon, including ack (8.98% vs. 4.37%, p=0.007) and BcAt (2.20% vs. 0.69%, p=0.04). In addition, codon mutation bias also appeared in ack gene of *Faecalibacterium prausnitzii* D (1.08% vs. 0%, p=0.023), ack gene of *Bifidobacterium pseudocatenulatum* (1.68% vs. 0%, p=0.039) and scpC gene of *Bacteroides stercoris* (5.03% vs. 1.86%, p=0.007). Therefore, the adaptive mutations occurred in the above strains in the nonhealthy group would block in the expression of SCFAs-production genes. Although the SCFAs genes of some 6 strains were subjected to greater intestinal selection pressure in the healthy group, it did not cause an increase in the above codon mutation types (**Additional file 1: Figure S4**). We also compared the relative abundance of the four strains aforementioned between host states and found that the three strains were more abundant in the healthy group, including *Faecalibacterium prausnitzii* C (0.56% vs. 0.61%, p<0.001), *Faecalibacterium prausnitzii* D (0.51% vs. 0.43%, p<0.001) and *Bacteroides stercoris* (2.43% vs. 2.56%, p<0.007). Therefore, we concluded that not only did the nonhealthy individual have fewer SCFAs producers in the gut microbiota, but also these producers' gene expression was also blocked due to adaptive mutations. Finally, we reported potentially common codon mutations in gut microbial strains (prevalent in > 3 host individuals, **Figure 5D, Additional file 2: Table S7**) affecting SCFA production, including three amino acids of *Faecalibacterium prausnitzii* C ack gene, one amino acid of *Faecalibacterium prausnitzii* C BcAt gene and two amino acids of *Bacteroides stercoris* scpC gene.

GMHI based on SNVs profiles indicates host health status

Gut Microbiome Health Index (GMHI) was proposed to develop for predicting disease presence (or absence) using species-level taxonomic abundance profiles derived from metagenomic sequencing data. Overall, microbial abundance essentially reflects the ecological changes related to disease states. Given our findings above, we hypothesized that SNV profile can also predict the host states, reflecting the degree and extent by which gut microbes adapt under multiple disease conditions.

Firstly, to test and validate how well the abundance-based GMHI can predict the host states, sequence abundance profile of the mutated genomes was used to construct GMHI (abundance-based GMHI, "Methods" and **Additional file 3: Table S8**). The relative sequence abundance of a total of non-redundant 2740 strains was calculated for 1711 metagenomics samples in the discovery cohorts and 446 samples in the validation cohort. Notably, we selected microbial strains differentially abundant in DNA sequence between diseased states as the biomarkers for calculating GMHI, which was slightly different from previous study [27] ("Methods" and "Code availability"). Among 471 selected strain-level markers, 222 were enriched in the nonhealthy group, while 249 were enriched in the healthy group. We found that the overall prediction accuracy of the abundance-based GMHI was 73.97%, p=9.98e-66, while this GMHI showed a significant difference between healthy and nonhealthy groups in nine cohorts (**Figure 6A**). The result was similar to the reported accuracy of the original GMHI method (~70%) based on the taxonomic abundance of gut microbes at the species level [27].

We next explored if SNV rate, representing the scale of genetic variation for each microbial strain (Eq 1), can be used as an independent indicator to predict host disease states. Therefore, we calculated the SNV rate of each of non-redundant 2740 strains in each sample, which formed a new SNV-rate feature table as compared to conventionally

used abundance-based feature table. Next, we identified microbial strains that have a significant difference in the SNV rate between host states and used their SNV rate to construct a new GMHI for classifying disease states ("Methods", **Additional file 4: Table S9**). Among 470 selected markers, 214 were enriched in the nonhealthy group, while 256 were enriched in the healthy group. With the receiver operating characteristic curve (ROC), we further showed high prediction accuracy of this SNV-rate GMHI (AUROC=74.23%, Wilcoxon rank-sum test, p=4.12e-67). Its performance was slightly better than the strain-level abundance-based GMHI. Furthermore, we found this GMHI can distinguish host states in 10 independent cohorts (**Figure 6B**).

We've shown that the mutation frequency/bias of SCFAs genes in gut microbes can potentially distinguish health states. Therefore, we next explored the feasibility of constructing GMHI based on the SNV profile in SCFAs genes of 75 microbial gut strains. We collected 185 genes from nine gene families (ack, buk, ydiF, atoD, atoA, ctfA, ctfB, BcAt, scpC) from 75 strains that carried adaptive SNVs. We first formed a count table for the SNVs (presence and absence) identified in the 185 genes. Then we aggregated the SNV count to the genome level, and calculate the relative frequency of SNV for SCFAs genes in each genome in each sample ("Methods", **Additional file 5: Table S10**). Among 48 selected markers, 11 were enriched in the nonhealthy group, while 37 were enriched in the healthy group. With the selected microbial SNV biomarkers, this SCFA GMHI showed that fairly good prediction accuracy (AUROC=65.20%, Wilcoxon rank-sum test, p=1.07e-26). Among 12 cohorts in this meta-analysis, the SCFA-based GMHI can be validated in six cohorts (**Figure 6C**). Although this accuracy is not perfect, it strongly suggested that genetic variability in the SCFAs genes alone can explain, at least partially, the microbiome difference between host health conditions.

External validation of functional changes in SCFA production induced by SNVs and the predictive power of GMHIs

To validate the above results, we further included three independent cohorts with a total of 446 samples, including 244 non-healthy and 202 healthy individuals, related to Atherosclerotic cardiovascular disease (ACVD) and Tuberculosis (TB) (**Additional file 6: Table S11**). The differences in relative abundance of the four strains changed compared with Figure 5 due to the different cohorts (**Figure 7A**). We observed that the relative abundances of *Faecalibacterium prausnitzii* C (p=0.37) was no longer significantly higher in the healthy group than in the non-healthy group, while the abundances of *Faecalibacterium prausnitzii* D (p<0.001), *Bacteroides stercoris* (p=0.042) and *Bifidobacterium pseudocatenulatum* (p<0.001) were higher in the healthy group. Interestingly, we still observed differential patterns of codon mutation bias in ack gene of *Faecalibacterium prausnitzii* C (p=0.023) and scpC gene of *Bacteroides stercoris* (p=0.025), however, other codon biases have not been proven (**Additional file 1: Figure S5**). We speculated that the other three genes were not verified due to the low probability of codon mutation bias (initiator and terminator), which can also be confirmed from Figure 5. Overall, we demonstrated that SCFAs metabolic genes on specific gut microbes exhibit codon mutation bias (initiator and terminator) in hosts with different health states.

Next, the ACVD and TB cohorts were further validated for GMHI based on SNV profiles (**Additional file 6: Table S12-14)**. Overall, the three different GMHIs showed acceptable prediction accuracy, respectively. The validation accuracy (i.e., AUROC) of abundance-based GMHI was 71.50% (Wilcoxon rank-sum test, p=5.2e-15), that of SNV-rate GMHI was 69.28% (Wilcoxon rank-sum test, p=2.3e-12) and that of SCFAs-SNV GMHI was 63.57% (Wilcoxon rank-sum test, p=8e-7). Furthermore, these GMHI can distinguish ACVD or TB from health. Collectively, we demonstrated that health status changes can be distinguished by genetic changes in gut microbiome, and highlighted the importance of SNV-induced functional changes related to SCFA production which underlie the pathogenesis of inflammatory bowel diseases and many diseases that were not interrogated before.

Discussion

The genetic composition of gut microbes was believed to be ever changing for microbial adaptation under different host intestinal environments, where SNV is one of main representative forms. [27-29]. Here, we described for the first time the universal evolutionary patterns of gut microbes under different health status in multiple human cohorts [3]. Previous study has provided insights that SNV makeup did not correlate with change in abundance [30] even in a shallow sequencing data. Therefore, the population-level genetic processes can be a new information layer of microbiome data for predictive modeling of diseases. Next, the highly diverse evolutionary patterns have been found across microbial strains [5, 6, 11, 30], which highlighted the importance of in-depth understanding the relationship between the scale and speed of gut microbial evolution and host disease development.

Gut microbes can modulate the host gut health through SCFA production, which contribute to intestinal homeostasis and the regulation of energy metabolism [31]. Certain major SCFAs-producing bacteria in the gut are of concern due to their remarkable genetic or microbial evolutionary diversity difference between healthy and nonhealthy groups. Lachnospiraceae is a typical group of acetate and butyrate producing bacteria, which provides beneficial effects for the host [32]. In our study, more than half of the gut microbes with different SNV rate belong to the Lachnospiraceae. About 30% of the strains that are different in microdiversity belong to *Bacteroides*, which are common producers of acetate and propionate. These results suggest that the evolution driven by intestinal selection pressure related to health status may play a role in the activity of SCFAs-related genes. We next demonstrated adaptive evolution affected the metabolic activity of SCFAs genes under multiple medical conditions. The suppressed production of SCFAs can be related to the unfavorable mutations occurred in the related gene coding regions. We supposed that larger pN/pS may contribute to the tendency of codon mutations, and mutations on initiators and terminators would link to potential functional deficits in the gut microbiota leading to multiple chronic diseases. Interestingly, we did observe that a large number of codons tend to mutate to terminators in nonhealthy group compared with healthy group in four strains. Furthermore, this tendency does not occur in healthy groups, even if the genes of these microbes are subjected to greater intestinal selection pressure. Therefore, we infer that the intestinal selection pressure driven by different host health status promotes the obvious codon mutation bias of SCFAs gene, affecting the expression of microbial functional genes.

Biomarker identification is one key goal towards the establishment of gut-microbiome-based prediction model for chronic diseases. On the one hand, it can be applied to non-invasive diagnosis, and on the other hand, it also provides a vision for constructing the causal relationship between host health and gut microbes. The composition of gut microbes has become the most basic biomarker, followed by functional genes, metabolite and clinical characteristics [33, 34]. Recently, we have tried to build diagnosis model based on SNVs or combined profiles and performed well [9, 11]. GMHI was a powerful index to distinguish healthy from non-healthy groups based on species relative abundance [27]. We use GMHI based on SNV rate to compare healthy and nonhealthy individuals, and we demonstrated the SNV rate was an independent indicator to evaluate the health status of hosts. At the SNV level of SCFAs gene, it may help to understand the deep mechanism of the relationship between microbes and host health. At present, it may be difficult to achieve a unified model for assessing host health based on gut microbes, but the genetic evolution of gut microbes can provide additional insights.

A few limitations of our study should be noted. Firstly, the sample size of our meta-analysis is not quite large, even we have included as many case samples as we can from the public database. In the discovery cohort, we included 1711 samples of 12 phenotypes from 16 studies and while 446 samples were collected in the validation cohort. A few studies that also attempted to explore the universal gut microbial markers for multiple chronic diseases usually include over 3000 metagenomic samples in the meta-analysis. In this big-data era, more metagenomic data should be included into meta-analysis to validate our findings related to gut microbial genetic variation affecting host health.

Secondly, we supposed to construct an association between the SNV and more host phenotypes, including blood pressure, high density lipoprotein, low density lipoprotein, etc. However, our attempt was failed due to the lack of sample metadata. Accordingly, it is urgent to call for open-data science that let scientists can employ more publicly accessible data to explore the consensus SNV signatures associated with more host phenotypes [7].

Conclusions

Collectively, natural selection is the main driving force for ecological changes in the gut microbiota. The constant and rapid adaptive evolutionary processes under gut selection pressures can either favor or deteriorate bacterial colonization and growth in the gut, resulting in ecological consequences (i.e., the abundance profile) in the microbial communities. However, most studies focused much on the change in abundance, and often failed to explain the microbiome variation related to disease conditions, Herein, we argue that microbial genetic changes can precede the ecological changes associated the host physiological changes, and thus would offer a new information layer from metagenomic data for predictive modeling of diseases. Interestingly, we preliminarily found a few genetic biomarkers on SCFA production can cover most chronic diseases involved in the meta-analysis. In the future, it is of both scientific and clinical significance to further explore the dynamic interactions between adaptive evolution and ecology of gut microbiota associated with host health status.

Methods

Metagenomic dataset collection and curation

To perform a comprehensive metagenomics meta-analysis for identifying the consensus in genetic variations in gut microbes associated with human diseases, we extensively searched for publicly accessible metagenomic datasets using a range of key words (e.g., “shotgun”, “gut microbiome”, “intestinal microbiome”, “metagenomic”, “metagenome”, “whole genome sequence”) in PubMed and ISI Web of Science (as of October 2021). Metagenomic studies related to dietary, drug or antibiotic interventions were excluded. We also require that the cohort must include both healthy and nonhealthy individuals, which ensures that the cohorts are comparable with each other. Importantly, studies on overweight and obesity alone were not included and classified as nonhealthy group. Specially, if a study collected non-healthy samples from multiple time points, we included only the first or baseline samples. Finally, we pinpointed fifteen studies for our meta-analysis, a total of 1711 metagenomic samples from 919 nonhealthy and 792 healthy individuals were included, spanning 12 host phenotypes (Cov19: Covid-19, SCZ: Schizophrenia, SF: Stone formers, AS: Atherosclerosis, PCOS: Polycystic ovary syndrome, GD: Graves' disease, T2D: Type 2 diabetes, CRC: Colorectal cancer, BC: Breast cancer, UC: Ulcerative colitis, CD: Crohn's disease, LC: Liver cirrhosis), representing a comprehensive set of human health conditions. Publicly available raw sequences (.fastq) and corresponding metadata were downloaded from NCBI SRA database. The detailed technical information on each study, such as shotgun metagenomics sequencing platform, average sequencing depth, target read lengths and other profiles was shown in **Table 1**. Notably, UC and CD cohorts share the same healthy control group. Two validation cohorts (AVCD: Atherosclerotic cardiovascular disease, TB: Tuberculosis) were collected using the consistent technical standards as we used for discovery cohorts.

Quality control of the raw metagenomic data

Sratoolkit 2.10.7 software (https://github.com/ncbi/sra-tools) was performed to separate raw sra files into paired or single fastq files. The raw reads were trimmed using Sickle (https://github.com/najoshi/sickle) and subsequently aligned to the host genome（GRCh38) to remove the host DNA fragments using Bowtie2 [35] with default settings.

Species taxonomic profiling

Firstly, we identified microbial species and estimated their relative abundances in each stool sample using MetaPhlAn 2.8 using default parameters based on the database (mpa_v29_CHOCOPhlAn_201901) [36]. Next, alpha diversity was calculated based on the species abundance profiles.

SNV calling for resident gut microbiota

We mainly conducted this analysis using inStrain. Firstly, Bowtie2 was applied to map our reads with a default reference genome of inStrain including 204938 genomes, 4644 representative strain genomes to create bam files. Then, each sam file was converted to a sorted and indexed .bam file using samtools [37]. Lastly, InStrain was used to call SNVs, analyze the sequencing coverage and breadth of mutated genomes, scaffolds, genes, estimate the microdiversity of mutated genomes, etc. with default parameters [38]. Notably, the minimum read-to-genome ANI (--min_read_ani) was set to be 92% by default, where all reads are expected to have an actual 95% ANI to species representative genomes (i.e., at the strain level, https://instrain.readthedocs.io/en/latest/index.html). The pN/pS of a gene and mutation bias of genome were calculated at positions where at least two alleles present rather than in relation to the reference genome. We next attempted to calculate the relative frequency of SNVs on a given genome (i.e., SNV rate) that can compare with other genomes. The raw SNV count is not an ideal measurement as the sequencing depth and breadth can vary substantially among mutated genomes. Therefore, we derived a new normalized SNV count metric, SNV rate, as follows

$$SNV\ rate = \frac{\text{SNV number}}{\text{g\_len}\times\text{breadth\_minCov}} \quad (1)$$

SNV number indicates the number of SNVs called from a mutated genome in a metagenomic sample. g_len indicates the genome length of this mutated genome. breadth_minCov, a typical inStrain output, indicates the percentage of bases in a scaffold/genome that have at least min_cov coverage. Particularly, it refers to the percentage of bases that have a nucleotide diversity value and meet the minimum sequencing depth that allowed us to call SNVs. Therefore, SNV rate was calculated as the percentage of bases that can call single-nucleotide variants by all SNV-callable bases in a microbial genome.

We next estimate the relative sequence abundance for mutated genomes using the relative number of individual reads that pass the selecting pairing filter which can be found in the inStrain output “mapping_info.tsv”. Gff profiles for each genome are required to call SNVs and can be accessed at https://doi.org/10.5281/zenodo.4441269.

The normalized SNV count by sequencing depth

We investigated the sequencing depth for all 1711 samples using seqkit 2.1.0 [39]. Of note, our previous study has shown the strong correlation between raw number of SNVs and sequencing depth [3]. Certainly, we want to reconfirm this relationship with simulated data and determine the relationship between other profiles (e.g., the number of mutated genomes) and the sequencing depth. To solve above issue, firstly, we selected six samples from different cohorts (three healthy and three nonhealthy) from 1711 samples, and their sequencing depths ware slightly higher than 10G. Next, we apply seqtk 1.3 to extract sequencing data from fastq files randomly (https://github.com/lh3/seqtk). And then, we acquired simulated data from 1G to 10G (step size: 1G) with three times (seed =11, 12 and 13). Our results suggested the particularly strong positive correlation between the number of

mutated genomes and SNV and the sequencing depth (**Additional file 1:** Figure S1, R =0.988-0.999, p<0.001). The simulation results in the range of 10G sequencing depth are applicable to most samples (**Additional file 1: Figure S2**). So, in this study, number of mutated genomes and SNV was normalized based on sequencing depth.

Establish a GMHI based on SNV profiles and genome-level sequence abundances

GMHI based on species relative abundance was previously proposed to distinguish healthy from non-healthy groups [27], which can be potentially developed as new diagnostic tool for host health. Here, we sought to test if any improvement in prediction accuracy using the variability in genetic composition within each species to assess host health status as compared to genome-level abundance profiles. Therefore, we generate three feature tables for building up a predictive index for health status (i.e., GMHI): (1) the sequence abundance profiles for all mutated genomes, (2) SNV rate for all mutated genomes (Eq 1); (3) SNV count for all mutant genes producing SCFAs for each sample.

GMHI based on SNV rate of mutated genomes.

1. The health- and nonhealthy-enriched markers were identified for each feature table using multiple Wilcoxon rank-sum tests. The compositional data was central-log-ratio transformed prior to statistical tests. Each feature table was further filtered by healthy-enriched markers (MH) and healthy-depleted markers (MN) respectively.

2. In either MH or MN sub-table, we calculate Shannon diversity (Hs or Ns) and richness (Hr or Nr) based on markers for each sample.

3. We further calculated the median Hr (or Nr) from 1% of the top (bottom)-ranked samples from (Hp or Np).

4. We next calculated the “collective” sequence abundance, SNV rate of mutated genomes or SNV frequency of genes producing SCFAs for health-enriched (i.e., psi_H) or nonhealthy-enriched markers (i.e., psi_N) for each metagenomic sample.

$$psi\_H = \frac{Hr}{\mathrm{Hp}} \times Hs \ (2)$$

$$psi\_N = \frac{Nr}{\mathrm{Np}} \times Ns \ (3)$$

5. Calculating a GMHI for each sample.

$$GMHI = log_{10}(\frac{\mathrm{psi\_H}}{psi\_N}) \ (4)$$

The necessary scripts and markers of three profiles can be found by https://github.com/HNUmcc/Meta_SNV_2157.

Statistics analysis

The statistical analysis was performed using R software. The differential abundances of various profiles were tested with the Wilcoxon rank-sum test with fdr adjust if needed, and the significant difference was considered at a nominal level of $p<0.05$. Alpha diversity analysis was performed “picante” and “veagn” package. Beta diversity analysis was conducted using “vegan”, “plyr” and “ggExtra” package, and PCoA based on Bray-Curtis and Euclidean dissimilarity matrix was used to visualize the sample clustering based on gut microbial or gut mutated genomes composition, and Adonis analysis was conducted using the vegan package, and the permuted P value was obtained by 999 permutations. The package “ggplot” and “gghalves” were used to generate boxplot, violin plot, density plot and fitted curve. The heatmap was constructed using the “pheatmap” package.

Abbreviations

SNVs: single nucleotide variants; SCFAs: short-chain fatty acids; GMHI: gut microbiome health index

Declarations

Ethics approval and Consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets supporting the conclusions of this article are included within the article and its additional files (Table 1 and supplementary table 12), and no additional sequencing data are generated in this study. The script can be found by https://github.com/HNUmcc/ Meta_SNV_2157, and the corresponding author (Jiachao Zhang) can be contacted for further additional information.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the specific research fund of "The Innovation Platform for Academicians of Hainan Province (YSPTZX202121)" and Qingdao Clinical Research Center for Oral Diseases (22-3-7-lczx-7-nsh).

Authors' contributions

C.M, Y.Z, S.J, F.T, S.H and J.Z performed data collection and analysis. C.M, S.H and J.Z write the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank all the teams that have made metagenomic data available.

References

1. Huang S, Jiang S, Huo D, Allaband C, Estaki M, Cantu V, et al. Candidate probiotic Lactiplantibacillus plantarum HNU082 rapidly and convergently evolves within human, mice, and zebrafish gut but differentially influences the resident microbiome. Microbiome. 2021;9(1):151; doi: 10.1186/s40168-021-01102-0.
2. Lee J, Mir A, Edraki A, Garcia B, Amrani N, Lou HE, et al. Potent Cas9 Inhibition in Bacterial and Human Cells by AcrIIC4 and AcrIIC5 Anti-CRISPR Proteins. mBio. 2018;9(6); doi: 10.1128/mBio.02321-18.
3. Ma C, Zhang C, Chen D, Jiang S, Shen S, Huo D, et al. Probiotic consumption influences universal adaptive mutations in indigenous human and mouse gut microbiota. Commun Biol. 2021;4(1):1198; doi: 10.1038/s42003-021-02724-8.
4. Ferreiro A, Crook N, Gasparrini AJ, Dantas G. Multiscale Evolutionary Dynamics of Host-Associated Microbiomes. Cell. 2018;172(6):1216-27; doi: 10.1016/j.cell.2018.02.015.
5. Roodgar M, Good BH, Garud NR, Martis S, Avula M, Zhou W, et al. Longitudinal linked-read sequencing reveals ecological and evolutionary responses of a human gut microbiome during antibiotic treatment. Genome Res. 2021;31(8):1433-46; doi: 10.1101/gr.265058.120.
6. Schloissnig S, Arumugam M, Sunagawa S, Mitreva M, Tap J, Zhu A, et al. Genomic variation landscape of the human gut microbiome. Nature. 2013;493(7430):45-50; doi: 10.1038/nature11711.
7. Wang D, Doestzada M, Chen L, Andreu-Sanchez S, van den Munckhof ICL, Augustijn HE, et al. Characterization of gut microbial structural variations as determinants of human bile acid metabolism. Cell Host Microbe. 2021;29(12):1802-14 e5; doi: 10.1016/j.chom.2021.11.003.
8. Garud NR, Pollard KS. Population Genetics in the Human Microbiome. Trends Genet. 2020;36(1):53-67; doi: 10.1016/j.tig.2019.10.010.
9. Ma C, Chen K, Wang Y, Cen C, Zhai Q, Zhang J. Establishing a novel colorectal cancer predictive model based on unique gut microbial single nucleotide variant markers. Gut Microbes. 2021;13(1):1-6; doi: 10.1080/19490976.2020.1869505.
10. Chen Y, Liu P, Liu R, Hu S, He Z, Dong G, et al. Comprehensive Strain-Level Analysis of the Gut Microbe Faecalibacterium prausnitzii in Patients with Liver Cirrhosis. mSystems. 2021;6(4):e0077521; doi: 10.1128/mSystems.00775-21.
11. Zhu Q, Hou Q, Huang S, Ou Q, Huo D, Vazquez-Baeza Y, et al. Compositional and genetic alterations in Graves' disease gut microbiome reveal specific diagnostic biomarkers. Isme J. 2021;15(11):3399-411; doi: 10.1038/s41396-021-01016-7.
12. Hu Y, Feng Y, Wu J, Liu F, Zhang Z, Hao Y, et al. The Gut Microbiome Signatures Discriminate Healthy From Pulmonary Tuberculosis Patients. Front Cell Infect Microbiol. 2019;9:90; doi: 10.3389/fcimb.2019.00090.
13. Chen Y, Li Z, Hu S, Zhang J, Wu J, Shao N, et al. Gut metagenomes of type 2 diabetic patients have characteristic single-nucleotide polymorphism distribution in Bacteroides coprocola. Microbiome. 2017;5(1):15; doi: 10.1186/s40168-017-0232-3.
14. Zou D, Pei J, Lan J, Sang H, Chen H, Yuan H, et al. A SNP of bacterial blc disturbs gut lysophospholipid homeostasis and induces inflammation through epithelial barrier disruption. EBioMedicine. 2020;52:102652; doi: 10.1016/j.ebiom.2020.102652.
15. Koh A, De Vadder F, Kovatcheva-Datchary P, Backhed F. From Dietary Fiber to Host Physiology: Short-Chain Fatty Acids as Key Bacterial Metabolites. Cell. 2016;165(6):1332-45; doi: 10.1016/j.cell.2016.05.041.
16. Parada Venegas D, De la Fuente MK, Landskron G, Gonzalez MJ, Quera R, Dijkstra G, et al. Short Chain Fatty Acids (SCFAs)-Mediated Gut Epithelial and Immune Regulation and Its Relevance for Inflammatory Bowel Diseases. Front Immunol. 2019;10:277; doi: 10.3389/fimmu.2019.00277.

17. Zhang F, Wan Y, Zuo T, Yeoh YK, Liu Q, Zhang L, et al. Prolonged Impairment of Short-Chain Fatty Acid and L-Isoleucine Biosynthesis in Gut Microbiome in Patients With COVID-19. Gastroenterology. 2022;162(2):548-61 e4; doi: 10.1053/j.gastro.2021.10.013.
18. Zhao L, Zhang F, Ding X, Wu G, Lam YY, Wang X, et al. Gut bacteria selectively promoted by dietary fibers alleviate type 2 diabetes. Science. 2018;359(6380):1151-6; doi: 10.1126/science.aao5774.
19. Yang J, Yu J. The association of diet, gut microbiota and colorectal cancer: what we eat may imply what we get. Protein Cell. 2018;9(5):474-87; doi: 10.1007/s13238-018-0543-6.
20. Goncalves P, Araujo JR, Di Santo JP. A Cross-Talk Between Microbiota-Derived Short-Chain Fatty Acids and the Host Mucosal Immune System Regulates Intestinal Homeostasis and Inflammatory Bowel Disease. Inflamm Bowel Dis. 2018;24(3):558-72; doi: 10.1093/ibd/izx029.
21. Chen SJ, Chen CC, Liao HY, Lin YT, Wu YW, Liou JM, et al. Association of Fecal and Plasma Levels of Short-Chain Fatty Acids With Gut Microbiota and Clinical Severity in Patients With Parkinson Disease. Neurology. 2022;98(8):e848-e58; doi: 10.1212/WNL.0000000000013225.
22. Zhang J, Sun Z, Jiang S, Bai X, Ma C, Peng Q, et al. Probiotic Bifidobacterium lactis V9 Regulates the Secretion of Sex Hormones in Polycystic Ovary Syndrome Patients through the Gut-Brain Axis. mSystems. 2019;4(2); doi: 10.1128/mSystems.00017-19.
23. Tang G, Du Y, Guan H, Jia J, Zhu N, Shi Y, et al. Butyrate ameliorates skeletal muscle atrophy in diabetic nephropathy by enhancing gut barrier function and FFA2-mediated PI3K/Akt/mTOR signals. Br J Pharmacol. 2022;179(1):159-78; doi: 10.1111/bph.15693.
24. Xu R, Tan C, He Y, Wu Q, Wang H, Yin J. Dysbiosis of Gut Microbiota and Short-Chain Fatty Acids in Encephalitis: A Chinese Pilot Study. Front Immunol. 2020;11:1994; doi: 10.3389/fimmu.2020.01994.
25. Kasmanas JC, Bartholomaus A, Correa FB, Tal T, Jehmlich N, Herberth G, et al. HumanMetagenomeDB: a public repository of curated and standardized metadata for human metagenomes. Nucleic Acids Res. 2021;49(D1):D743-D50; doi: 10.1093/nar/gkaa1031.
26. Amann RI, Baichoo S, Blencowe BJ, Bork P, Borodovsky M, Brooksbank C, et al. Toward unrestricted use of public genomic data. Science. 2019;363(6425):350-2; doi: 10.1126/science.aaw1280.
27. Gupta VK, Kim M, Bakshi U, Cunningham KY, Davis JM, 3rd, Lazaridis KN, et al. A predictive index for health status using species-level gut microbiome profiling. Nat Commun. 2020;11(1):4635; doi: 10.1038/s41467-020-18476-8.
28. Jiang P, Wu S, Luo Q, Zhao XM, Chen WH. Metagenomic Analysis of Common Intestinal Diseases Reveals Relationships among Microbial Signatures and Powers Multidisease Diagnostic Models. mSystems. 2021;6(3); doi: 10.1128/mSystems.00112-21.
29. Wirbel J, Pyl PT, Kartal E, Zych K, Kashani A, Milanese A, et al. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. Nat Med. 2019;25(4):679-89; doi: 10.1038/s41591-019-0406-6.
30. Chen L, Wang D, Garmaeva S, Kurilshikov A, Vich Vila A, Gacesa R, et al. The long-term genetic stability and individual specificity of the human gut microbiome. Cell. 2021;184(9):2302-15 e12; doi: 10.1016/j.cell.2021.03.024.
31. van der Hee B, Wells JM. Microbial Regulation of Host Physiology by Short-chain Fatty Acids. Trends Microbiol. 2021;29(8):700-12; doi: 10.1016/j.tim.2021.02.001.
32. Sorbara MT, Littmann ER, Fontana E, Moody TU, Kohout CE, Gjonbalaj M, et al. Functional and Genomic Variation between Human-Derived Isolates of Lachnospiraceae Reveals Inter- and Intra-Species Diversity. Cell

Host Microbe. 2020;28(1):134-46 e4; doi: 10.1016/j.chom.2020.05.005.
33. Coker OO, Liu C, Wu WKK, Wong SH, Jia W, Sung JJY, et al. Altered gut metabolites and microbiota interactions are implicated in colorectal carcinogenesis and can be non-invasive diagnostic biomarkers. Microbiome. 2022;10(1):35; doi: 10.1186/s40168-021-01208-5.
34. Verstockt B, Parkes M, Lee JC. How Do We Predict a Patient's Disease Course and Whether They Will Respond to Specific Treatments? Gastroenterology. 2022;162(5):1383-95; doi: 10.1053/j.gastro.2021.12.245.
35. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357-9; doi: 10.1038/nmeth.1923.
36. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods. 2015;12(10):902-3; doi: 10.1038/nmeth.3589.
37. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9; doi: 10.1093/bioinformatics/btp352.
38. Olm MR, Crits-Christoph A, Bouma-Gregson K, Firek BA, Morowitz MJ, Banfield JF. inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. Nat Biotechnol. 2021;39(6):727-36; doi: 10.1038/s41587-020-00797-0.
39. Shen W, Le S, Li Y, Hu F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. PLoS One. 2016;11(10):e0163962; doi: 10.1371/journal.pone.0163962.

Table

Table 1

Fecal metagenomic studies included in this meta-analysis

Study[1]	Country	nonhealthy sample size	healthy sample size	total	sequencing platform	average sequencing target depth	target read length	sequencing data
Cov19	China	15	15	30	Illumina NextSeq 550	4.39 GB	150 bp	PRJNA624223
SCZ	China	90	81	171	Illumina HiSeq X	10.94 GB	150 bp	PRJEB29127
SF	Italy	5	5	10	Illumina NextSeq 500	1.57 GB	150 bp	PRJNA418941
AS	China	97	114	211	Illumina HiSeq 2000	3.73 GB	100 bp	PRJNA375935
PCOS	China	14	14	28	Illumina HiSeq 4000	6.22 GB	150 bp	PRJNA549764
	China	50	43	93	Illumina HiSeq 2500	8.78 GB	150 bp	PRJNA530971
GD	China	102	62	164	Illumina HiSeq 2500	8.07 GB	100 bp	PRJNA602729, PRJNA602731, PRJNA602732, PRJNA638403, PRJNA638404, PRJNA638405
T2D	China	71	74	145	Illumina GAIIx and HiSeq 2000	2.51 GB	175 bp	PRJNA422434
CRC	China	8	12	20	Illumina HiSeq 2500	6.45 GB	150 bp	PRJNA663646
	Japan	40	40	80	Illumina HiSeq 2500	6.42 GB	150 bp	DRA006684
	Italy	32	28	60	Illumina HiSeq 2500	3.89 GB	100 bp	SRP136711
	Austria	46	63	109	Illumina HiSeq 2000	4.84 GB	100 bp	ERP008729
BC	China	62	71	133	ION_TORRENT	10.94 GB	150 bp	PRJNA718520

Table 1. Fecal metagenomic studies included in this meta-analysis (Continued)

Study	country	nonhealthy sample size	healthy sample size	total	sequencing platform	average sequencing target depth	target read length	sequencing data
UC	American, Holland	76	56	220	Illumina HiSeq 2500	4.01 GB	101 bp	PRJNA400072
CD	American, Holland	88				4.34 GB		
LC	China	123	114	237	Illumina HiSeq 2000	1.74 GB	100 bp	PRJEB6337
		919	792	1711				

[1.] Cov19: Covid-19, SCZ: Schizophrenia, SF: Stone formers, AS: Atherosclerosis, PCOS: Polycystic ovary syndrome, GD: Graves' disease, T2D: Type 2 diabetes, CRC: Colorectal cancer, BC: Breast cancer, UC: Ulcerative colitis, CD: Crohn's disease, LC: Liver cirrhosis.

Figures

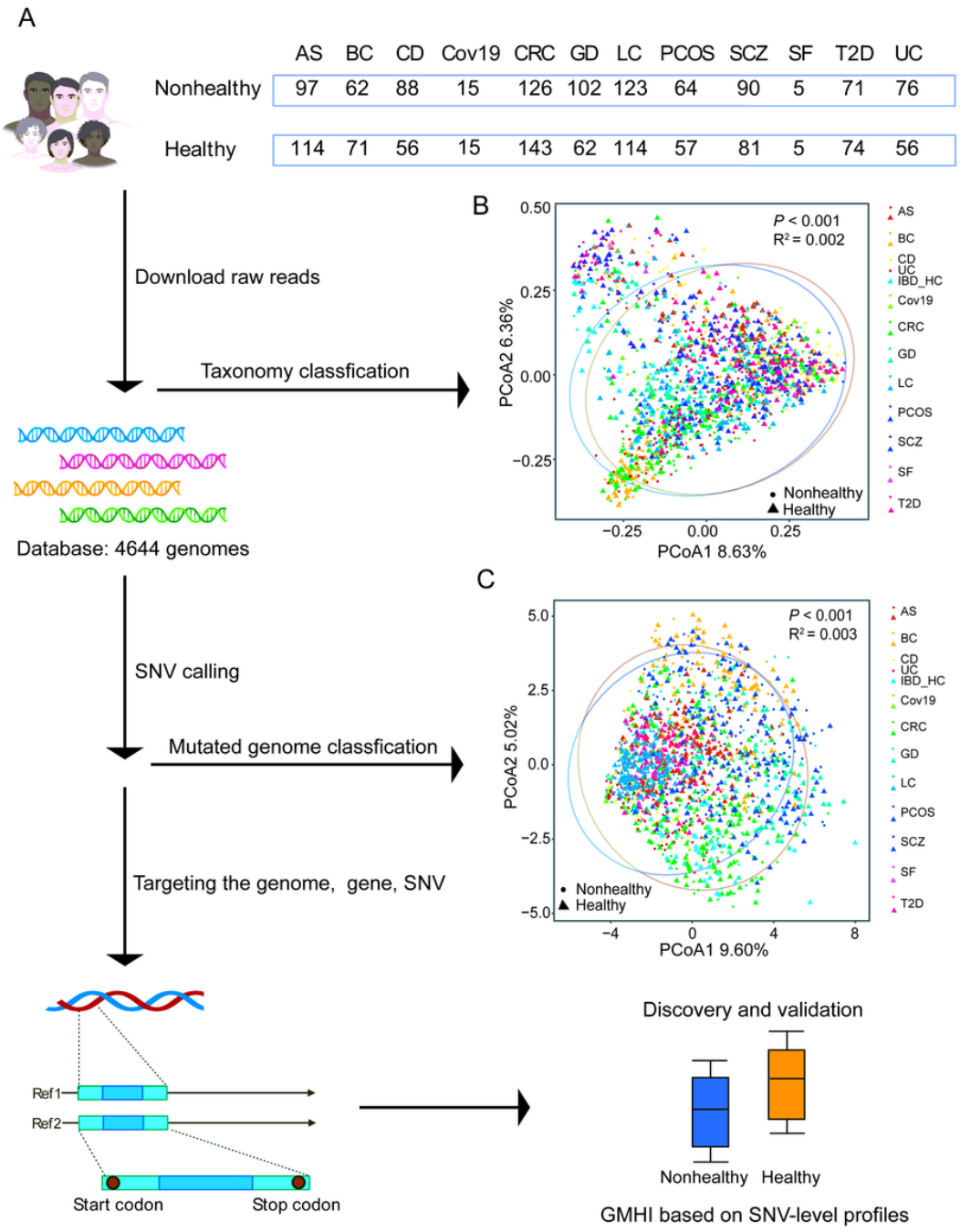

Figure 1

Dataset integration and meta-analysis workflow. A) In this study, 1711 metagenomic samples from 16 microbiome studies (including 919 healthy and 792 nonhealthy individuals) were integrated into our meta-analysis for calling, detection and profiling of genetic variation in the gut microbiome. Each of 16 cohorts in this study has both healthy and nonhealthy individuals. First, comparisons between different health status were limited to one nonhealth phenotype, and more importantly, global comparisons. Next, specific mutated genomes, genes and SNVs were mentioned to assess the impact of SNVs on health status. **B)** Principal coordinate analysis (PCoA) plot based on Bray-Curtis dissimilarity showed the between-sample difference in gut microbial composition between nonhealthy (circle, n=919) and healthy (triangle, n=792) groups. **C)** PCoA based on Euclidean distance showed a significant difference in the diversity of mutated genomes in the gut microbiota between nonhealthy and healthy groups. For

each PCoA plot, each dot corresponds to a sample, and dot's shape corresponds to the health status, and dot's color corresponds to the disease phenotype. The ellipse corresponds to 95% confidence region.

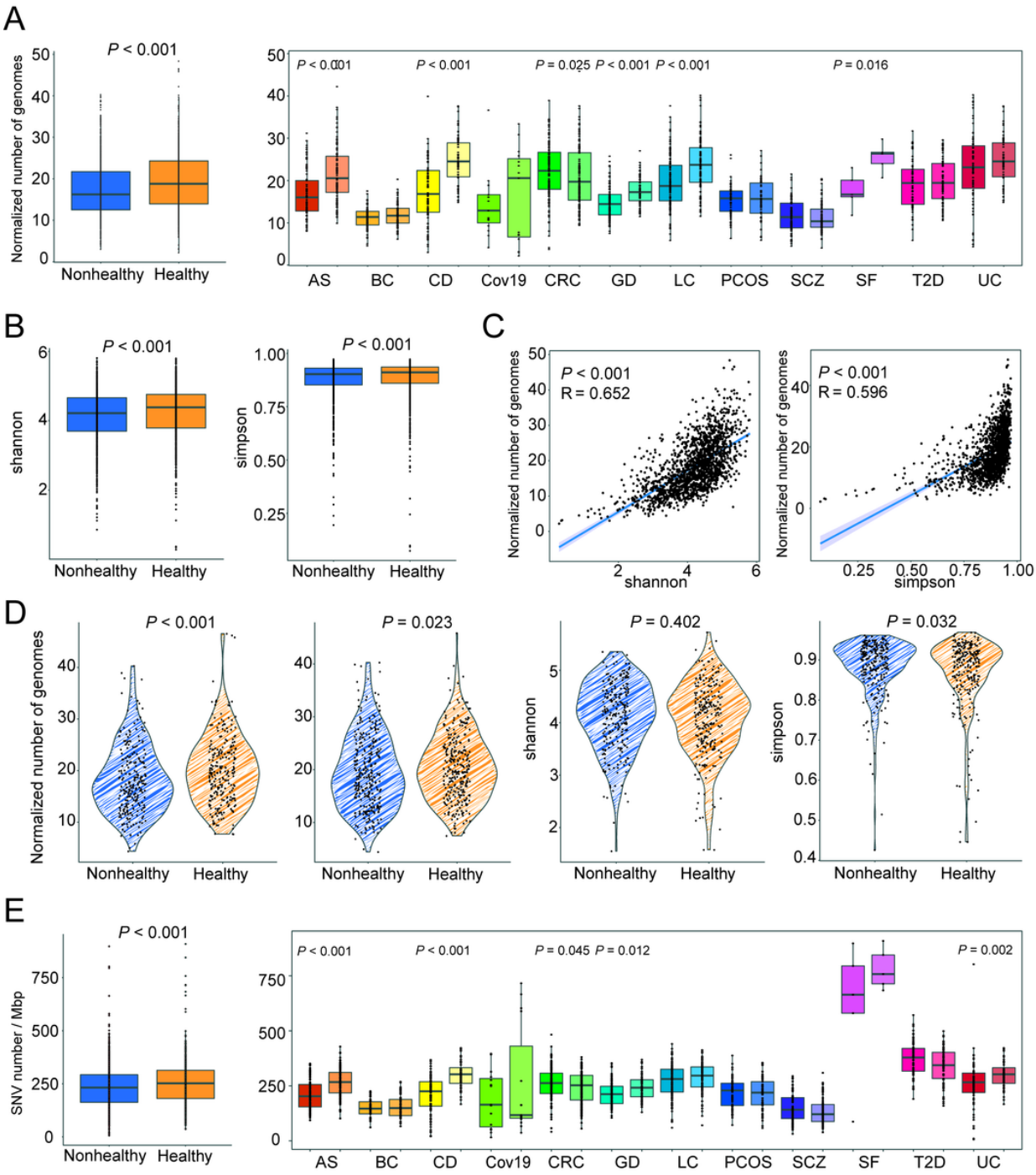

Figure 2

Normalized evaluation and comparisons of number of mutated genomes and SNVs. A) The boxplot shows the comparison of normalized number of mutated genomes in the whole dataset (N=1711) and for each disease phenotype (Wilcoxon rank sum test). **B)** Comparison of alpha diversity (Shannon and Simpson index) based on gut microbial profiles at the species level. **C)** The strong correlations between number of mutated genomes and Shannon index (Spearman correlation). **D)** Matched individuals with the same number of mutated genomes from healthy and nonhealthy groups were used to compare alpha diversity, and individuals with the same alpha diversity were also compared with the number of mutated genomes (Wilcoxon sign rank test). **E)** Comparison of the normalized number

of SNV of gut microbes between healthy and nonhealthy individuals in whole dataset (N=1711) and each disease phenotype by Boxplot (Wilcoxon rank sum test). The p values were corrected with fdr method, and the significant difference was considered at a nominal level of *p<0.05, **p<0.01 and ***p<0.001. For all boxplots, the nonhealthy group was on the left and the healthy group was on the right.

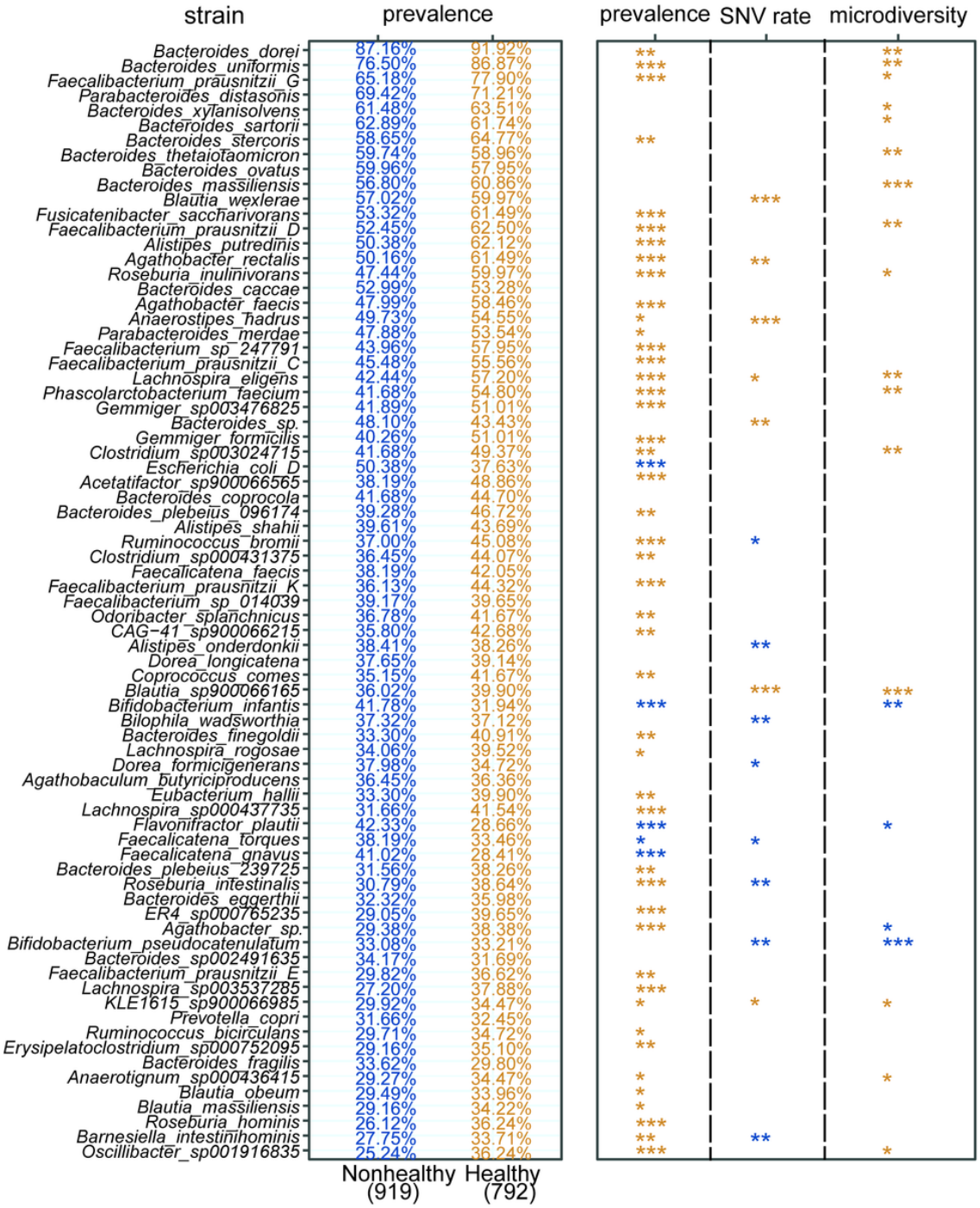

Figure 3

The overall strain-level diversity associated with host health status. A total of 75 mutated genomes were identified in > 30% of 1711 human individuals. The overall features (i.e., mutated genome frequency, SNV rate ("Methods"), microdiversity) for each of these 75 mutated strains were compared between healthy and nonhealthy cohorts using Wilcoxon rank-sum test. The p values were corrected with fdr method, and the significant difference was considered at a nominal level of *p<0.05, **p<0.01 and ***p<0.001.

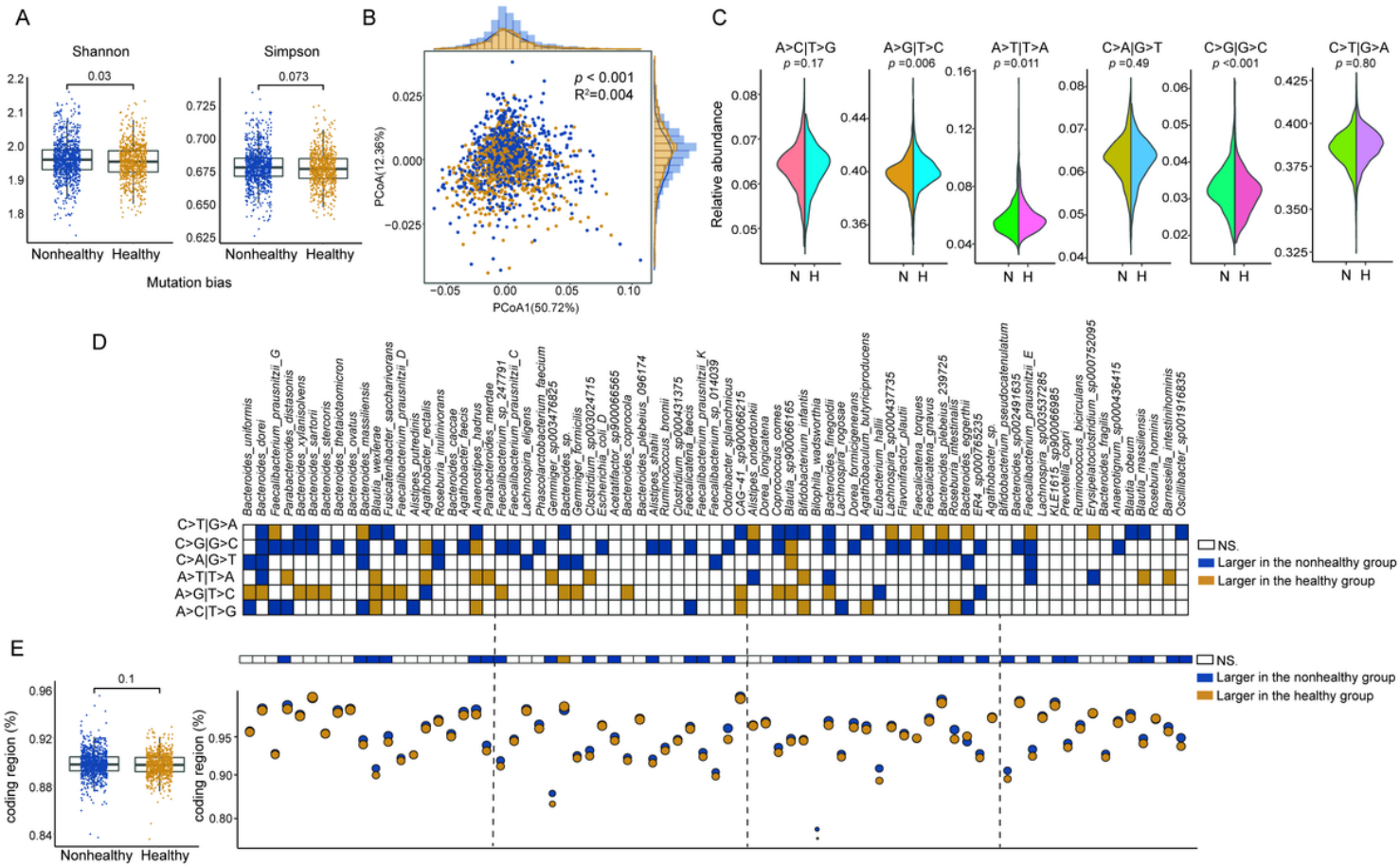

Figure 4

Universal base mutation bias associated with host health status. A) The comparison of Shannon index and Simpson index of mutation-type profiles between healthy and nonhealthy groups (Wilcoxon rank-sum test). **B)** The PCoA plot shows the between-sample difference based on Bray-Curtis dissimilarity of six-base mutation-type profiles. Adonis was used to estimate the effect size of host health status on the mutation types. The histograms show the distribution of healthy (orange) and nonhealthy (blue) samples along each axis. **C)** The proportion of six-base mutation types between healthy and nonhealthy groups was showed and compared in the half violin diagrams. N means nonhealthy group, H means healthy group. **D)** The relative abundance of six mutation types of 75 mutated genomes was compared, and white mean no difference in base mutation bias, orange mean higher in the healthy group, and blue mean higher in the nonhealthy group. **E)** The proportion of SNVs located in the coding region was evaluated for the whole dataset and for each of 75 strains. The p values were obtained from Wilcoxon rank-sum tests with fdr correction, and the significant difference was considered at a nominal level of p<0.05.

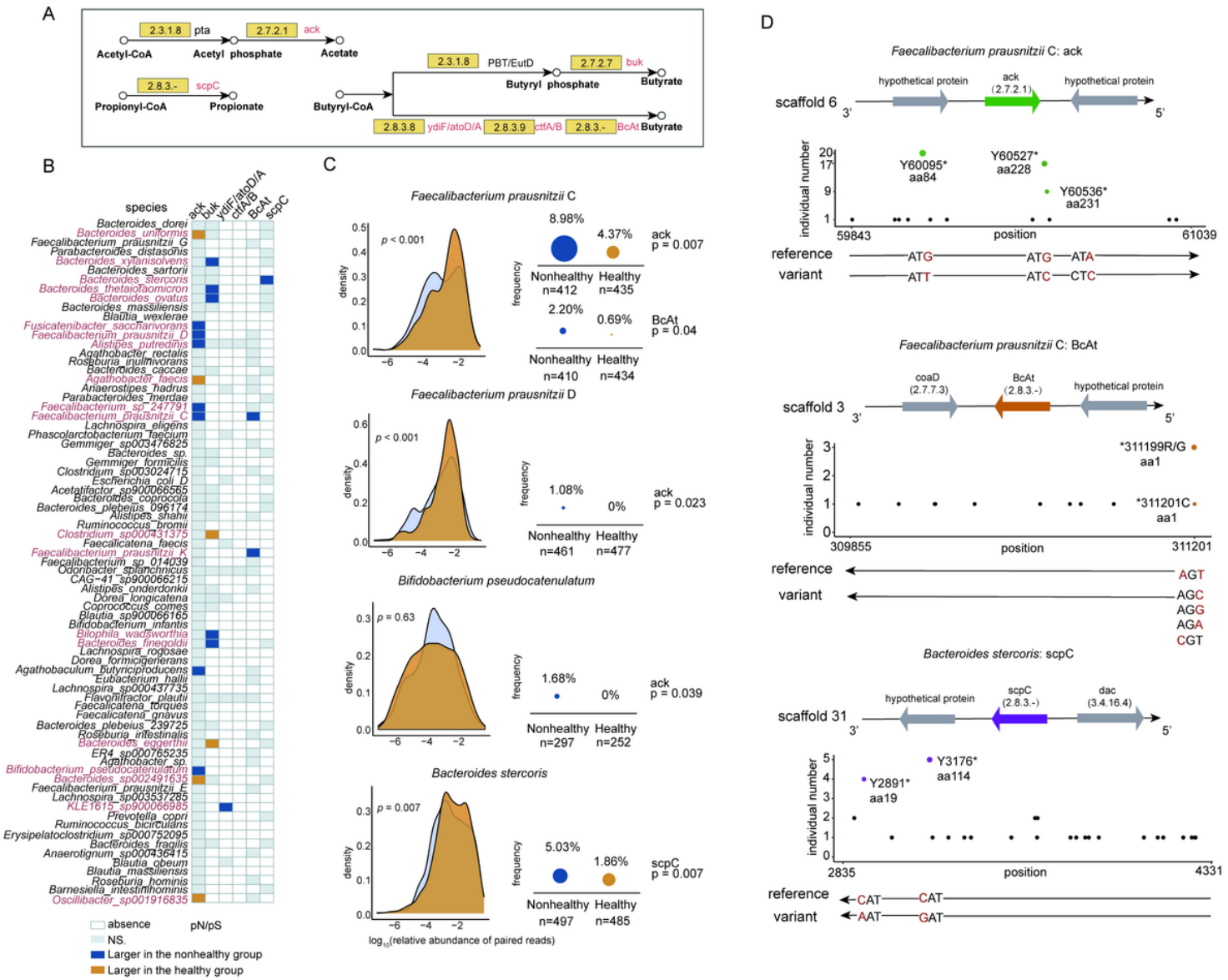

Figure 5

Strain-level codon mutation bias in SCFA-production-involved genes. A) The enzymes related in the production of three short-chain fatty acids (acetate, propionate and butyrate) are shown. Ack: Acetate kinase, scpC: Propionyl-CoA:succinate CoA transferase, buk: Butyrate kinase, ydiF: Acetate CoA-transferase YdiF, atoD/A: Acetate CoA-transferase subunit alpha/beta, ctfA/B: Butyrate–acetoacetate CoA-transferase subunit A/B, BcAt: Butanoate coenzyme A-transferase. **B)** The heatmap shows enrichment directionality of pN/pS (i.e., natural selection forces) for each of six SCFA-related gene families (including a total of 185 genes) encoded by 75 strains between healthy and nonhealthy groups. The pN/pS is the ratio of 2 rates: the rates of non-synonymous (pN) and synonymous (pS) SNVs. White cells indicate gene absence in a strain, gray cells mean no difference of pN/pS, orange cells mean larger pN/pS in the healthy group, and blue mean larger pN/pS in the nonhealthy group. **C)** The SCFA genes of the four strains had codon mutation bias, suggesting that SNV causes codon to become termination codon or the start codon is inactivated. The density plot displays the sequence relative abundance of the four strains in both healthy and unhealthy groups. The between-group comparison was performed by using Wilcoxon rank-sum test. N represented the number of individuals with SNV on these genes. When the prevalence was calculated, the denominator was N, and the numerator was the number of individuals with codon mutation bias. **D)** Some SNVs were detected in at least

three individuals, which causes codon to become termination codon or the start codon is inactivated. The horizontal axis represented the position of SNV, the vertical axis represented the number of individuals, and in addition, the scaffold, codon, reference genome, variant genome, and adjacent genes were also shown. The p values from Wilcoxon rank-sum test were fdr corrected, and the significant difference was considered at a nominal level of $p<0.05$.

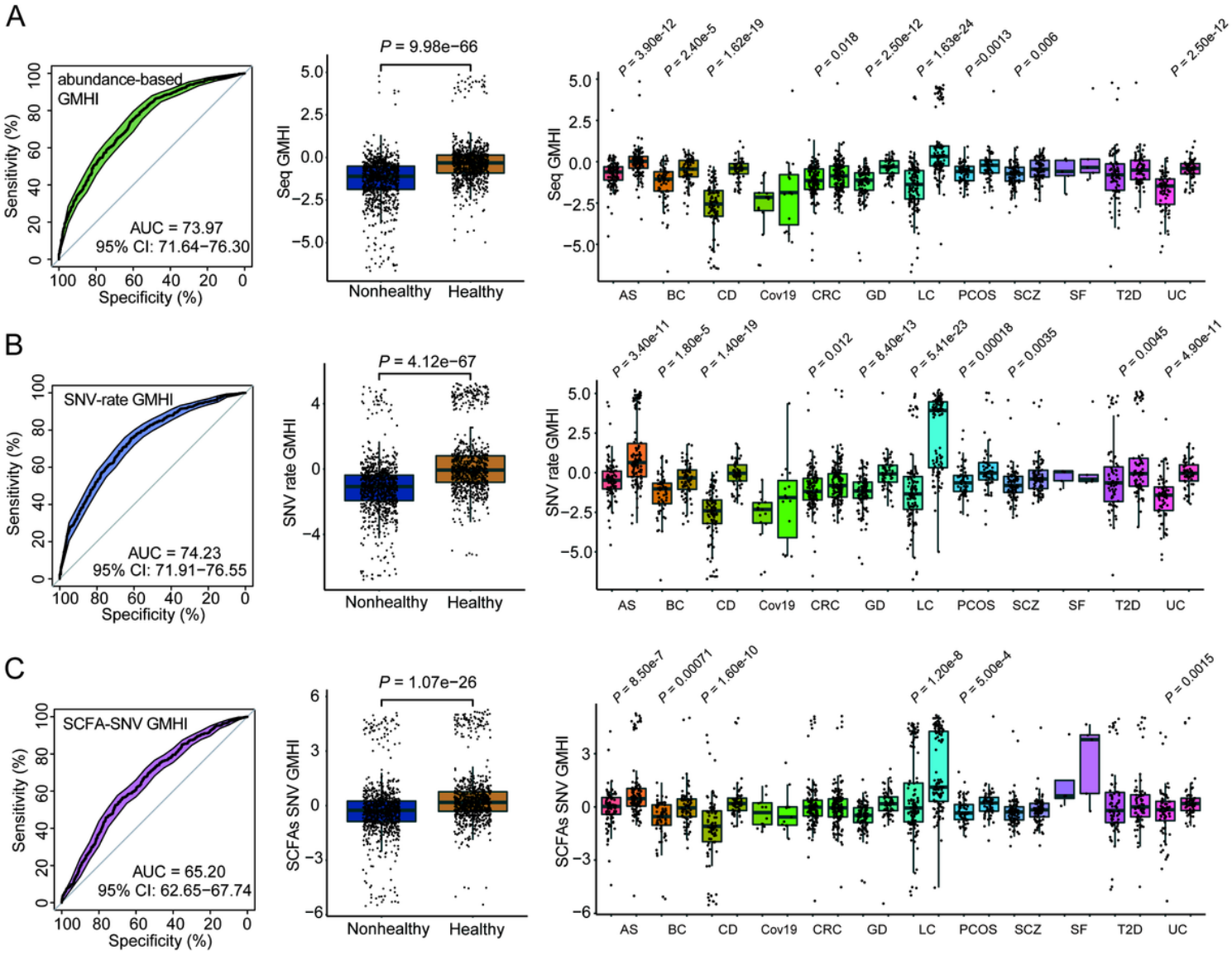

Figure 6

GMHI based on SNVs profiles indicates host health status. GMHI was calculated based on sequence abundance profiles of mutated genomes (i.e., abundance-based GMHI), SNV-rate profiles of mutated genomes (i.e., SNV-rate GMHI) and SNV profile of SCFA genes (i.e., SCFA-SNV GMHI), respectively. **A)** The prediction performance of abundance-based GMHI (AUROC: 73.97%). **B)** The prediction performance of the GMHI based on SNV rate profiles (AUROC: 74.23%) p=4.12e-67, and ten cohorts are applicable. **C)** The predictive performance of the GMHI based on SNV profiles of SCFA genes (AUROC: 65.20%). The GMHI values were compared between healthy and nonhealthy groups using Wilcoxon rank-sum test. The significant difference was considered at a nominal level of $p<0.05$. For all boxplots, the nonhealthy group was on the left and the healthy group was on the right.

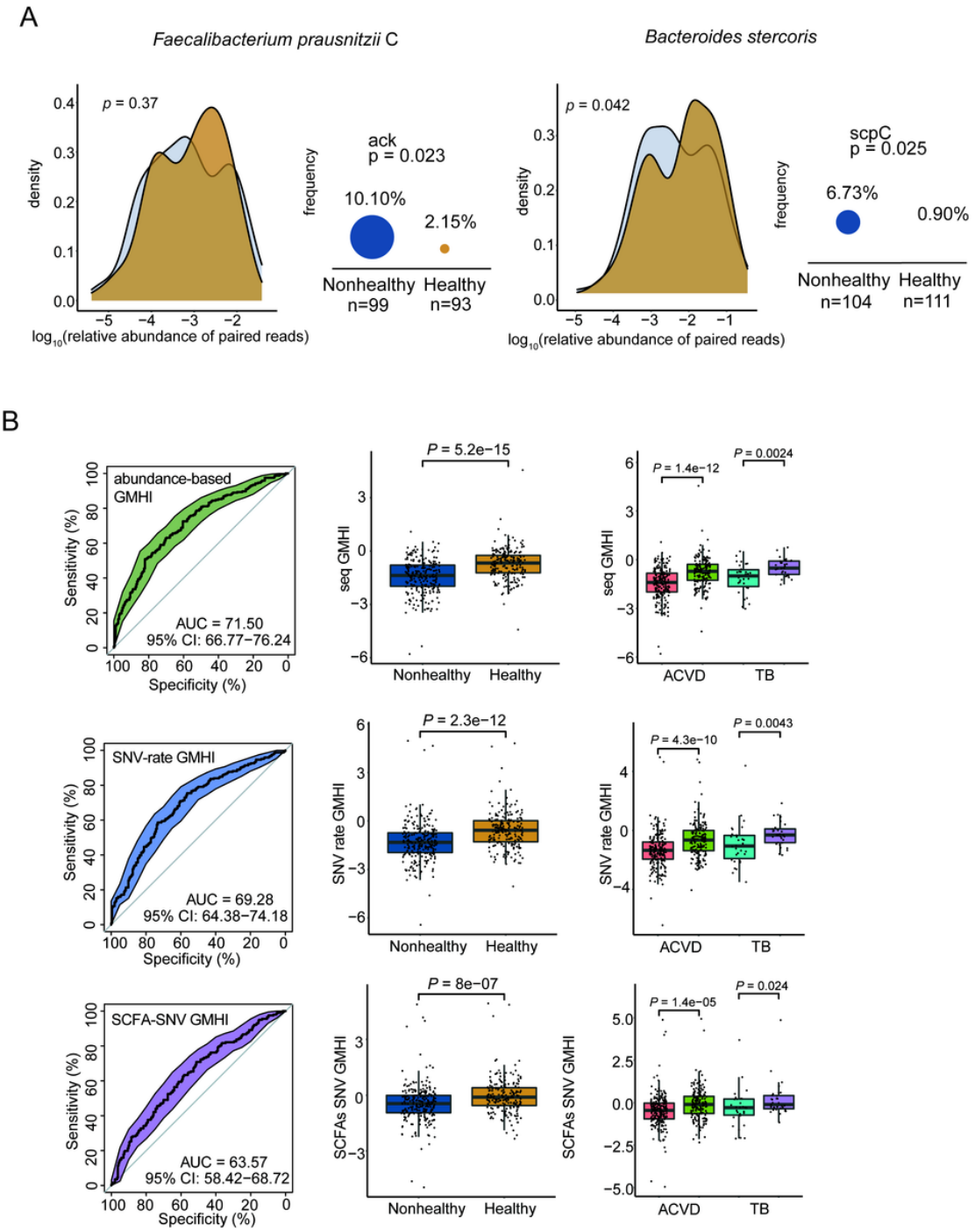

Figure 7

External validation of functional changes in SCFA production induced by SNVs and the predictive power of GMHIs

A) Codon mutation bias of two key SCFA-producing genes from two gut microbes (i.e., ack of *Faecalibacterium prausnitzii* C and scpC of *Bacteroides stercoris*) was validated using 446 additional metagenomic samples. The density plot shows the comparison of the sequence relative abundance of the two strains between healthy and nonhealthy groups using Wilcoxon rank-sum test. The mutation ratios of ack and scpC genes were compared between host groups. N indicates the number of individuals with SNVs on each of these genes. When the prevalence was calculated, the denominator was N, and the numerator was the number of individuals having this codon mutation bias. **B)** The prediction performance of three GMHIs in the validation cohort. ROC analysis revealed that AUROC for three GMHI were 71.50%, 69.28% and 63.57%, respectively. Each of these GMHIs can distinguish ACVD or

TB from health. The GMHI values were compared between healthy and unhealthy groups with Wilcoxon rank-sum test. The significant difference was considered at a nominal level of $p<0.05$. For all boxplots, the nonhealthy group was on the left and the healthy group was on the right.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.pdf
- Additionalfile2.xlsx
- Additionalfile3.xlsx
- Additionalfile4.xlsx
- Additionalfile5.xlsx
- Additionalfile6.xlsx