

GigaScience

The unique composition of Indian gut microbiome, gene catalogue and associated faecal metabolome deciphered using multi-omics approaches

--Manuscript Draft--

Manuscript Number:	GIGA-D-18-00212R2
Full Title:	The unique composition of Indian gut microbiome, gene catalogue and associated faecal metabolome deciphered using multi-omics approaches
Article Type:	Research
Funding Information:	
Abstract:	<p>Background</p> <p>Metagenomic studies carried out in the past decade have led to an enhanced understanding of the gut microbiome in human health, however, the Indian gut microbiome is still not well explored. We analysed the gut microbiome of 110 healthy individuals from two distinct locations (North-Central and South) in India using multi-omics approaches, including 16S rRNA gene amplicon sequencing, whole genome shotgun metagenomic sequencing, and metabolomic profiling of faecal and serum samples.</p> <p>Results</p> <p>The gene catalogue established in this study emphasizes the uniqueness of the Indian gut microbiome in comparison to other populations. The gut microbiome of the cohort from North Central India, which was primarily consuming a plant-based diet, was found to be associated with Prevotella, and also showed an enrichment of Branched Chain Amino Acid (BCAA) and lipopolysaccharide (LPS) biosynthesis pathways. In contrast, the gut microbiome of the cohort from Southern India, which was consuming an omnivorous diet, showed associations with Bacteroides, Ruminococcus and Faecalibacterium, and had an enrichment of Short Chain Fatty Acid (SCFA) biosynthesis pathway and BCAA transporters. This corroborated well with the metabolomics results, which showed higher concentration of BCAAs in the serum metabolome of the North-Central cohort and an association with Prevotella. In contrast, the concentration of BCAAs were found higher in the faecal metabolome of the South Indian cohort, and showed a positive correlation with higher abundance of BCAA transporters.</p> <p>Conclusions</p> <p>The study revealed the unique composition of Indian gut microbiome, established the Indian gut microbial gene catalogue, and also compared it with the gut microbiomes from other populations. The functional associations revealed using metagenomic and metabolomic approaches provide novel insights on the gut-microbe-metabolic axis, which will be useful for future epidemiological and translational researches</p>
Corresponding Author:	Vineet Kumar Sharma, Ph.D. Indian Institute of Science Education and Research Bhopal Bhopal, Madhya Pradesh INDIA
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Indian Institute of Science Education and Research Bhopal
Corresponding Author's Secondary Institution:	
First Author:	Darshan B Dhakan
First Author Secondary Information:	
Order of Authors:	Darshan B Dhakan

	Abhijit Maji
	Ashok K Sharma
	Rituja Saxena
	Joby Pulikkan
	Tony Grace
	Andres Gomez
	Joy Scaria
	Katherine R Amato
	Vineet Kumar Sharma, Ph.D.
Order of Authors Secondary Information:	
Response to Reviewers:	<p>--Reviewer #1</p> <p>--I commend the authors on their openness and responsiveness to the comments from both reviewers. The additional analyses performed and the clarifications in the manuscript have resulted in a much-improved draft. The availability of both the raw amplicon and WGS data in NCBI's Sequence Read Archive is a great service to the scientific community and should ensure open access to this data and its inclusion in future studies. The origin of the samples and the sequencing data is well-documented and cohort sample collection and storage appears to adhere to ethical and technical standards.</p> <p>--More specifically, the updated manuscript addresses and corrects all points that I raised in the first review. These include a major reanalysis of 16S data with an updated database (SILVA December 2017 versus GreenGenes 2013) and an implementation of statistical analysis with normalization performed using DESeq2. They clarify their use of downsized data for Unifrac analysis. Further, the authors now combine their data and run downstream analysis using an "Updated-IGC." This clearly aids their analysis and broadens the appeal of the manuscript as a whole. The 9% of additional genes appears to be unique to the Indian cohort. The authors also performed the suggested enterotype comparisons with the data from Arumugam et al.</p> <p>--Based on this new version of the manuscript, my recommendation is the manuscript can be accepted without further scientific revision. The authors should, nevertheless, have a careful review of the text to address remaining grammatical errors and awkward phrasing. A few, non-exhaustive, examples are given:</p> <p>The changes suggested by reviewers have been highlighted (orange coloured) in the manuscript.</p> <p>Reply: We thank the reviewer for appreciating our efforts and recommending the manuscript for publication. We have carefully read the manuscript for any grammatical errors and have also made all the suggested changes in the manuscript text.</p> <p>--Line 114: Change "All the recruited individuals" to "Recruited individuals"</p> <p>Reply: As per reviewer's suggestion, we have made this change (Line: 114).</p> <p>--Line 216: Should be "relative abundances"</p> <p>Reply: We have corrected these words (Line: 217).</p> <p>--Line 365: "its inward transport in microbial cells by the BCAA transporters" would be better as "its uptake by microbes via BCAA transporters"</p> <p>Reply: We agree with the rephrasing and have revised this sentence as suggested (Line: 369).</p> <p>--Line 408: Change "Though, the sequencing depth in the study was not too high..." to</p>

"Although sequencing depth was modest....longer paired-ends reads, from the cohort of 110 individuals appears sufficient to provide the first insights on the Indian gut microbiome"

Reply: We have revised the sentence as per the suggestions from both reviewers (Line: 412-415).

--Line 446: "One aspect to this could" could be better written as "One potential explanation could be..."

Reply: We have revised this sentence as per the suggestion (Line: 450)

--Line 486: "has known health benefits..." might be better as "has been reported to be beneficial by preventing...."

Reply: We have revised this sentence as per the reviewer's suggestion (Line: 490).

--Line 505: "are emerging, which results in the increased...." is better as "are emerging, with results showing increased...."

Reply: We have revised this sentence as per the reviewer's suggestion (Line: 509).

In addition, we have also carefully checked the manuscript for any grammatical or phrasing errors and hope that the revised manuscript is much better in reading.

Reviewer #2

--The authors have reasonably addressed the comments I raised in the original submission.

Only one general comment and a few minor comments remain, which should all be readily addressable by the authors.

General comments:

--L286-291: It would be good to test whether the location and diet are correlated and to which extent. In fact, given the information from the authors, I would expect them to be correlated. Hence, the observed results (Fig. S12) are to be expected and this should be qualified. If no such test is performed, I would recommend to at least reemphasize the (strong) influence of location on the diet of the studied Indian populations. This is also important with respect to the results in L333-335.

The changes suggested by reviewers have been highlighted (orange coloured) in the manuscript

Reply: We thank the reviewer for the suggestion. We have now performed a correlation of location and diet across all samples and observed a high correlation ($\rho = 0.708$; FDR Adj. P-value = 2×10^{-16}). We have included these results at both places in the revised manuscript (Line: 293-294, 337-338).

Minor comments:

--Throughout: Frequently, "the"/"a" is missing, e.g., L158 "analysis of microbiome", L159 "reads from other three datasets", L163 "This shows that the addition of subset".

Reply: As suggested by the Reviewer, we have added the/a in the manuscript. We have also carefully checked and corrected the manuscript for any such errors.

--L149-150 - "and unique to IGC": This reads as if the 943,395 genes are unique to the IGC, but aren't his unique to the newly constructed Indian microbial gene catalogue?

Reply: We agree with the reviewer that these 943,395 genes identified from Indian gut microbiome are unique to Indian microbial gene catalogue and not present in IGC. We have rephrased this sentence for clarity (Line 149-150).

--L161 - "did not show a significant ($P < 0.01$)": Not sure if the significance level ($\alpha = 0.01$) is meant here or if the p-value was " < 0.01 ". In the latter case, it would be considered significant at $\alpha = 0.01$. Please clarify and verify throughout.

Reply: We apologize for this confusion. We have now provided the exact P-values for HMP, MetaHIT and China datasets, and the P-values were not significant for all the three datasets as found using the student's t-test, whereas it was significant for Indian dataset. The P-values are mentioned at all places in the manuscript where the results were significant (Line 155-156, 161-162, 166).

--L212-214: Species names should be italicized.

Reply: We have made this correction (Line: 213-215).

--L270: "be" is missing -> "needs to be collected".

Reply: We have corrected this sentence (Line: 271).

--L275: The text suggests a "significance", yet the p-value is listed as 0.6841. Please clarify.

Reply: We apologize for this confusion. We have removed the word "significance" and have rephrased the sentence for clarity. Here, we observed a high concordance in allocation of samples to clusters using both taxonomic (genus abundance) and functional (KEGG abundance) information. Using Fisher's exact test as suggested by reviewer 2 during the earlier revision, no significant difference was observed in cluster allocation (P -value = 0.6841) thus showing similarity in clustering of samples using taxonomic and functional information. We have also provided this information of cluster allocation in Additional File 11. (Line: 274-277).

--Supplements: Fig S11 still contains a reference to "enterotypes" which, as suggested by Reviewer 1 (and I agree) should be generally avoided, unless in combination with the non-Indian populations. Please check this throughout.

Reply: We agree with the reviewer and we have replaced the word 'enterotypes' with "clusters" in Additional File 5.

--L304-305: This is not a necessity for the revision, but rather a question out of curiosity: Was an association with age tested here, in addition to BMI?

Reply: We had examined the association of multiple covariates including age and BMI, with taxonomic and functional data. We have provided the details of these associations in Additional File 13 and Additional File 15, which were also provided with the earlier submitted manuscript.

--L317 + L319: What do "19 MGS/CAG" and "67 MGS/CAG" refer to here? Are these the numbers of MGSs/CAGs that were annotated to likely be *P. copri* populations, i.e., multiple strains/sub-species of *P. copri* were identified? Please clarify this.

Reply: Here, we were referring to the total 19 MGS/CAGs found enriched in LOC1, and 67 MGS/CAGs found enriched in LOC2. We have reframed this sentence for clarity (Line: 317-322).

--L339: Did Cluster-2 show *no* association with location, i.e., was a mixture of samples from LOC1 and LOC2?

Reply: Cluster-2 did show an association with location. Out of a total of 36 samples assigned to Cluster-2 it included 13 samples from LOC-1 and 23 samples from LOC-2. We have mentioned this in the manuscript (Line: 343-344).

--Legend Fig.S17: "OPLD-DA" -> "OPLS-DA "

Reply: We have corrected this word in the legend of Fig. S17 (Additional File 5).

--Fig.S18: Panel A is rather small and the fonts are hard to read. Please increase the

size of the panel.

Reply: We have now increased the font size of Panel A in Figure S18.

--L409-411: I welcome the qualification of the sequencing depth here. Nevertheless, the argument of 2x150bp sequencing is misleading here. Read-length clearly plays a role, so does the overall sequencing depth. While 2x150bp is commonly used currently, and hence the current study is up-to-date, I would suggest the authors to rephrase this slightly. My suggestion would be: "... deviation), the inclusion of 110 individuals from two distinct geographic locations as well as the identification of Indian gut microbiome-specific genes provide a first insight into the Indian gut microbiome and are thus considered important additions to the field."

Reply: We thank the reviewer for this suggestion, and have revised this text as per the suggestion (Line: 412-415).

--L411-413: This sentence reads contradictory in itself. If there is a high diversity, how can (only) two locations be considered representative? I would suggest to rephrase this.

Reply: We have removed the word 'representative' and have rephrased this sentence (Line: 415-417).

--L431: It is not readily clear what "Its" refers to here. I assume it is "Prevotella", yet this should be clarified.

Reply: Yes, the word 'its' was referring to Prevotella. We have reframed this statement for more clarity (Line: 435).

--L439: Please consider removing "driver" unless you can show a causation rather than the association which was presented in the results.

Reply: As per the suggestion, we have removed the word 'driver' from the sentence (Line: 443).

--L442: "bacteria" -> "bacterium"

Reply: We have corrected this word (Line: 446).

--L470-471: The "statistically sound" is not readily clear here. Please consider removing this as I do not find it relevant in this context.

Reply: As per the reviewer's suggestion, we have revised this sentence and have removed the phrase 'statistically sound' (Line: 474).

--L500: "Firmicute" -> "Firmicutes"

Reply: We have corrected this word (Line: 504).

--L515: Please remove "populations", it does not fit in here.

Reply: We have removed this word (Line: 519).

--L578: Please check correct capitalization.

Reply: We thank the reviewer for pointing it out. We have corrected this word to 'UniFrac' and checked it throughout the manuscript (Line: 582).

--L582: Please be consistent in the numbers: "mean = 1.36 Gb" vs. "1.5" (L408).

Reply: Thanks for pointing out this typo, we have corrected this number in line number 412.

--L585: Consider removing "bacterial" unless there was some enrichment step for

	<p>bacterial DNA.</p> <p>Reply: We thank the reviewer for pointing it out. We have removed the word “bacterial” from this sentence (Line: 589).</p>
Additional Information:	
Question	Response
Are you submitting this manuscript to a special series or article collection?	No
<p>Experimental design and statistics</p> <p>Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends.</p> <p>Have you included all the information requested in your manuscript?</p>	Yes
<p>Resources</p> <p>A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible.</p> <p>Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?</p>	Yes
<p>Availability of data and materials</p> <p>All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using</p>	Yes

a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript.

Have you have met the above requirement as detailed in our [Minimum Standards Reporting Checklist](#)?

[Click here to view linked References](#)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1 **Title:** The unique composition of Indian gut microbiome, gene catalogue and associated faecal
2 metabolome deciphered using multi-omics approaches

3 **Authors:** D.B. Dhakan^{1†}, A. Maji^{1†}, A.K. Sharma¹, R. Saxena¹, J. Pulikkan², T. Grace^{2,3}, A.
4 Gomez⁴, J. Scaria⁵, K.R. Amato⁶, V.K. Sharma^{1*}

5 **Affiliations:** ¹Metagenomics and Systems Biology Laboratory, Department of Biological
6 Sciences, Indian Institute of Science Education and Research Bhopal, India, ²Department of
7 Genomic Science, Central University of Kerala, India, ³Division of Biology, Kansas State
8 University USA, ⁴Microbiomics Laboratory, Department of Animal Science, University of
9 Minnesota, USA, ⁵Animal Disease Research & Diagnostic Laboratory, Veterinary and Biomedical
10 Sciences Department, South Dakota State University, USA, ⁶Department of Anthropology,
11 Northwestern University, USA.

12 **Email IDs:** darshan@iiserb.ac.in, abhi71084@gmail.com, ashoks773@gmail.com,
13 ritus@iiserb.ac.in, puljobcmi@gmail.com, tonygrace99@gmail.com, gomez@umn.edu,
14 joy.scaria@sdstate.edu, katherine.amato@northwestern.edu, vineetks@iiserb.ac.in

15 † These authors contributed equally to this work

16 *Corresponding author

17 V.K. Sharma: vineetks@iiserb.ac.in

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

19 Abstract

20 Background

21 Metagenomic studies carried out in the past decade have led to an enhanced understanding of the
22 gut microbiome in human health, however, the Indian gut microbiome is not well explored yet.
23 We analysed the gut microbiome of 110 healthy individuals from two distinct locations (North-
24 Central and Southern) in India using multi-omics approaches, including 16S rRNA gene amplicon
25 sequencing, whole genome shotgun metagenomic sequencing, and metabolomic profiling of faecal
26 and serum samples.

27 Results

28 The gene catalogue established in this study emphasizes the uniqueness of the Indian gut
29 microbiome in comparison to other populations. The gut microbiome of the cohort from North-
30 Central India, which was primarily consuming a plant-based diet, was found to be associated with
31 *Prevotella*, and also showed an enrichment of Branched Chain Amino Acid (BCAA) and
32 lipopolysaccharide (LPS) biosynthesis pathways. In contrast, the gut microbiome of the cohort
33 from Southern India, which was consuming an omnivorous diet, showed associations with
34 *Bacteroides*, *Ruminococcus* and *Faecalibacterium*, and had an enrichment of Short Chain Fatty
35 Acid (SCFA) biosynthesis pathway and BCAA transporters. This corroborated well with the
36 metabolomics results, which showed higher concentration of BCAAs in the serum metabolome of
37 the North-Central cohort and an association with *Prevotella*. In contrast, the concentration of
38 BCAAs were found higher in the faecal metabolome of the Southern-India cohort, and showed a
39 positive correlation with the higher abundance of BCAA transporters.

40 Conclusions

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

41 The study reveals the unique composition of Indian gut microbiome, establishes the Indian gut
42 microbial gene catalogue, and compares it with the gut microbiome of other populations. The
43 functional associations revealed using metagenomic and metabolomic approaches provide novel
44 insights on the gut-microbe-metabolic axis, which will be useful for future epidemiological and
45 translational researches.

46
47 **Keywords:** Indian Gut Microbiome, Whole Genome Shotgun, Metagenomics, Metabolomics,
48 Integrated Gene Catalog, Metagenome-Wide Association Study, Core gut microbiome, Short
49 Chain Fatty Acids, Branched Chain Amino Acids

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

52 **Background**

53 Determining the taxonomic and functional composition of a healthy gut microbiome across
54 different populations is essential for understanding its role in maintaining human health. Several
55 large-scale, world-wide microbiome projects have revealed variability in the gut microbial
56 composition of the healthy individuals due to factors such as mode of delivery, age, geographical
57 location, diet, lifestyle, etc. [1-5]. Majority of the gut microbiome studies have determined
58 microbial taxonomy and functional diversity using 16S rRNA marker gene-based and/or Whole
59 Genome Shotgun (WGS) approaches to understand the functional role of the gut microbiome.
60 However, novel insights on the complex interplay between diet, gut microbes and human health,
61 along with the role of key microbial metabolites, such as Short Chain Fatty Acids and Branched
62 Chain Amino Acids, derived from the microbial fermentation of dietary fibres are beginning to
63 emerge from recent gut metabolomics studies [6, 7]. Moreover, the direct impact of microbial
64 metabolome on human health is also becoming apparent from the recent studies focusing on the
65 ‘gut microbiome- host metabolism axis’ [8]. Therefore, an integrative approach using both
66 metagenome and metabolome-based characterizations of the gut microbiome appears pragmatic
67 for gaining deeper functional and mechanistic insights into the role of gut microbes on human
68 health.

69 The large-scale studies carried out so far mainly represent the gut microbiome of urban populations
70 primarily from Europe, US and other ‘WEIRD’ countries (i.e., the Western, Educated,
71 Industrialized, Rich, and Democratic countries) [9, 10]. Only recently, some studies have
72 characterized the human microbiome from diverse ethnic populations and found significant
73 compositional variations compared to microbiome from other previously studied populations [11-
74 14]. India is the seventh largest country in the world and harbours the second largest population

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

75 spread across multiple geographical locations with enormous diversity in ethnicity, lifestyles and
76 dietary habits. India is a home to the majority of world's vegetarian population but also has an
77 almost equal representation of population consuming animal-based diets. The Indian population
78 also has the highest prevalence of diabetes in the world [15]. According to the World Health
79 Organization estimates (WHO, 2011), 53% of deaths in India in the year 2008 were attributed to
80 metabolic conditions such as diabetes and cardiovascular diseases, which are predicted to reach
81 ~75% by 2030 [16].

82 A few studies have investigated the gut microbiome of the Indian population. A recent study by
83 Maji et al. has shown the functional association of human gut microbiome dysbiosis with
84 tuberculosis through a time-course study carried on six tuberculosis patients in India [17].
85 However, other gut microbiome studies were mainly limited by small cohort sizes and amplicon-
86 based (16S rRNA gene) sequencing and analysis [17-21]. Thus, several large-scale efforts are
87 needed to identify the Indian population-specific microbiome biomarkers, and to understand the
88 impact of gut microbiome on health and disease in the Indian population along with global
89 comparisons.

90 However, to uncover the enormous gut microbiome diversity inherent in the different sub-
91 populations of India, extensive sampling and analyses are required. Therefore, as the first large-
92 scale study from India, we selected two prominent locations in North-Central India, i.e. LOC1:
93 Bhopal city, Madhya Pradesh, and Southern India, i.e. LOC2: Kerala. The two locations also had
94 different dietary habits. The Southern-India (LOC2) diet consisted of rice, meat and fish, whereas
95 the North-Central (LOC1) diet consisted of carbohydrate-rich food including plant-derived
96 products, wheat and trans-fat food (high-fat dairy, sweets and fried snacks). The 'Human
97 Development Index Report, UNDP' (United Nations Development Programme), India and SRS-

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

98 based life-table (Sample Registration Survey, 2010-14) has revealed that the citizens from Kerala
99 had the highest life-expectancy rates (>74 years) in India, whereas those in Madhya Pradesh
100 (capital city ‘Bhopal’) exhibited the lowest (<65 years) [22]. Further, a higher predisposition of
101 the North-Indian population towards diabetes, cardiovascular diseases and hypertension is also
102 known, which in contrast is much lower in Southern India, perhaps due to the lifestyle differences
103 in the two regions [15, 23]. Thus, to gain deeper functional insights into the microbiome from these
104 two distinct sub-populations of India, a multi-omics approach was carried out using amplicon-
105 based profiling of taxonomic composition (16S rRNA gene sequencing), whole genome shotgun-
106 based (WGS-based) profiling of metagenome, and GC-MS-based profiling of faecal and serum
107 metabolomic signatures.

108 **Data Description**

109 The two selected locations, Bhopal (LOC1) and Kerala (LOC2) from North-Central and Southern
110 parts of India were about 2,000 kms apart and provided a distinct representation of the Indian
111 population with respect to diet and lifestyle (**Additional File 1**). The 110 (62 females, 58 males)
112 individuals recruited in this study were not suffering from any disease as reported by personal
113 medical history and physical examination, and confirmed no exposure to antibiotics for at least
114 one month prior to sampling. Recruited individuals had an average BMI of 21.16 (± 5.23 standard
115 deviation), an average age of 29.72 (± 17.41 standard deviation) and were not diagnosed with any
116 disease at the time of sample collection, and thus were considered as ‘healthy’ (**Additional File**
117 **1**). Moreover, they did not have a second-degree relative history of T2D. The recruitment of
118 volunteers, sample collection, and other study-related procedures were carried out by following
119 the guidelines and protocols approved by the Institute Ethics Committee of Indian Institute of
120 Science Education and Research (IISER), Bhopal, India. The faecal samples were frozen within

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

121 30 mins of collection and were then used for 16S rRNA gene V3 hypervariable region amplicon
122 sequencing, WGS-based metagenomic sequencing, and metabolomic analysis. Further, the serum
123 samples collected from a subset of volunteers were used for GC-MS based metabolomics analysis.
124 The sequencing of V3 hypervariable region of 16S rRNA gene and shotgun metagenome
125 sequencing from the 110 faecal samples resulted into 54.87 million paired-end reads ($503,460 \pm$
126 $175,547$ (mean \pm standard deviation) reads/sample) and 499.98 million paired-end reads
127 ($4,545,280 \pm 1,498,663$ (mean \pm standard deviation) reads/sample), respectively (Methods,
128 **Additional File 2** and **Additional File 3**). The metabolomic analysis was also performed on all
129 faecal and subset of serum samples collected from the same healthy participants using GC-MS,
130 and the resultant CDF files were used for further analysis. The data description of participants and
131 the data generated from each sample is provided in **Additional File 1** under the Metadata
132 information section.

133 **Analyses**

134 **Construction of an Indian gut microbial gene catalogue and updated integrated gene** 135 **catalogue (IGC)**

136 The first step for functional analysis was the construction of an extensive catalogue of gut
137 microbial genes from the Indian population, which was not available previously. A De Bruijn
138 graph-based assembly of reads resulted in 2,165,507 contigs of length ≥ 500 bp with a total contig
139 size of 3.086 Gbp representing 68.25% of total reads and a mean N50 value of 2,288 bp. To obtain
140 assemblies of low coverage genomic regions or genomes present in the Indian gut microbiome,
141 the reads from each sample were mapped on assembled contigs obtained from their respective
142 sample, and the remaining singletons (unassembled reads) were pooled and re-assembled together

1
2
3
4 143 into additional 45,839 contigs with length ≥ 500 bp and a total assembled length of 34.68 Mbp. A
5
6
7 144 total of 1,551,581 non-redundant genes were predicted from contigs, which represent the gut
8
9 145 microbial gene catalogue of the Indian cohorts.

10
11
12 146 The integrated gene catalogue (IGC) established by Li et al. in a previous multicohort study
13
14
15 147 consisted of 9,879,896 genes identified from 1,267 gut metagenomes representing multiple
16
17 148 populations [24]. A total of 943,395 genes (sharing $< 90\%$ identity with IGC) out of 1,551,581
18
19
20 149 from Indian gut microbial gene catalogue were identified as unique to the Indian microbial gene
21
22 150 catalogue. The IGC was updated to construct an ‘Updated-IGC’ by adding these 943,395 non-
23
24
25 151 redundant genes from the Indian gene catalogue. The updated-IGC consisting of 10,823,291 non-
26
27 152 redundant genes (an 8.8% increase from IGC) was used as the reference gene catalogue for the
28
29
30 153 subsequent analysis performed in this study. A total of 70.74% ($\pm 3.77\%$ standard deviation)
31
32 154 mapping coverage of reads ($\sim 7.5\%$ increase in the mapping of reads) was observed from the 110
33
34
35 155 Indian samples on the updated-IGC as compared to 63% ($\pm 4.61\%$ standard deviation) on IGC,
36
37 156 showing a significant (FDR Adj. P-value = 10^{-16} ; Student’s t-test) increase in mapping of Indian
38
39 157 microbial dataset (**Fig. 1A** and **Additional File 4**). The datasets from populations of USA (HMP),
40
41
42 158 Denmark (MetaHIT) and China (a study from Qin et al.) mentioned in **Table 1** were used for a
43
44 159 comparative analysis of the microbiome of Indian population with other populations [7, 10, 25].
45
46
47 160 The mapping of reads from these three datasets (HMP, MetaHIT and China) on updated-IGC
48
49 161 (mean mapping coverage: HMP = 67.74%, China = 77.44% and MetaHIT = 75.21%) did not show
50
51
52 162 a significant (P-values: HMP=0.5, MetaHIT=0.85 and China = 0.17; Student’s t-test) increment
53
54 163 from their mapping coverage on IGC (mean mapping coverage: HMP (USA) = 66.93%, China =
55
56 164 77.37% and MetaHIT (Denmark) = 75.02%) as observed in **Fig. 1A**. This shows that the addition
57
58
59 165 of a subset of non-redundant genes (sharing $< 90\%$ identity with IGC) from the Indian gut
60
61
62
63
64
65

1
2
3
4 166 microbiome to the IGC significantly increased (FDR Adj. P-value = 10^{-16} ; Student's t-test) the
5
6 167 mapping percentage of reads from Indian microbiome dataset as compared to the other datasets.
7
8
9

10 168 **Identification of taxonomic signatures of Indian gut microbiome**

11
12 169 To determine the taxonomic and functional composition of the Indian gut microbiome and to
13
14
15 170 identify Indian-specific gut-microbial signatures, a cross-population comparison was carried out
16
17
18 171 using the 16S rRNA gene hypervariable region and shotgun metagenomic data from the other
19
20 172 populations. A reference-independent metagenome-wide association study (MGWAS) was carried
21
22 173 out to identify the Indian-specific gut metagenomic markers through a comparison with similar
23
24
25 174 large-scale studies from other populations [26]. The genes from the metagenomic samples of four
26
27 175 countries (India, China, USA and Denmark) were clustered (see Methods) into 924 clusters based
28
29
30 176 on their co-occurrence and Pearson correlations ($\rho \geq 0.9$) across samples resulting into 335 MGS
31
32 177 (metagenomic species) having ≥ 700 genes in each cluster, and 589 CAGs (co-abundance gene
33
34
35 178 groups) consisting of ≥ 100 genes in each cluster. Out of the 924 metagenomic clusters, 195 could
36
37 179 be assigned up to species level using the taxonomic assignment strategy described in Methods.
38
39 180 Canberra distances were calculated from MGS/CAG abundance profiles and their Principal
40
41
42 181 Component Analysis (PCA) was carried out using 'countries' as factors for explaining the variance
43
44 182 between samples, which showed that the Indian population formed a distinct cluster from the other
45
46
47 183 populations in PCA (**Fig. 1B**). It is interesting to note that the samples from the Indian cohort were
48
49 184 more widely spread owing to the higher inter-sample Canberra distances between Indian samples
50
51 185 (mean = 0.689) as compared to other datasets having average inter-sample distances of 0.61, 0.59
52
53
54 186 and 0.54 for USA, China and Denmark populations, respectively (**Additional File 5: Figure S1**).
55
56 187 This could be attributed to the significant (FDR Adj. P-value = 0.00013) differences in MGS
57
58
59
60
61
62
63
64
65

1
2
3
4 188 abundance profiles between LOC1 and LOC2 populations as revealed on comparison of their
5
6 189 principal coordinates (**Additional File 5: Figure S2**).

7
8
9
10 190 Further, the identification of enriched metagenomic species (MGS) from P-values calculated using
11
12 191 negative binomial (NB) model-based Wald test (implemented in DESeq2) and Log Odds Ratio
13
14 192 showed that the species belonging to the genera *Bacteroides*, *Alistipes*, *Clostridium*, and
15
16 193 *Ruminococcus* were depleted in the Indian population (China, Denmark and USA; Log Odds Ratio
17
18 < -2 and Adj. P-value <0.01), whereas the MGS/CAGs annotated as *Prevotella*, *Mitsuokella*,
19
20 194 *Dialister*, *Megasphaera*, and *Lactobacillus* were found to be associated with the Indian population
21
22 195 (Adj. P-value < 0.01; Log Odds Ratio > 2), and were the major drivers for separation of Indian
23
24 196 samples from other populations (**Additional File 5: Figure S3; Additional File 6**). Furthermore,
25
26 197 the distribution of microbial families across ten different populations was also calculated using
27
28 198 16S rRNA gene markers, which revealed Indian gut microbiome to have the highest abundance of
29
30 199 Prevotellaceae (**Fig. 1C**). The feature selection method applied using random forest along with
31
32 200 pairwise Wilcoxon rank-sum test also identified Prevotellaceae to be significantly higher (FDR
33
34 201 Adj. P < 0.05) in gut microbiome of Indian cohort compared to the other population datasets except
35
36 202 Indonesia (P-value = 0.506) (**Additional File 5; Figure S4, S5 and S6**) where a comparable
37
38 203 abundance of Prevotellaceae was present. The high abundance of Prevotellaceae in Indian
39
40 204 population underscores its importance as the marker taxa for the Indian cohort.
41
42
43
44
45
46
47
48
49

50 206 **Microbial functions enriched in the Indian population**

51
52 207 Functional comparison of Indian microbiome with other populations was carried out by mapping
53
54 208 the genes derived from assembled contigs to the EggNOG database. In total 69,386 EggNOG
55
56 209 functions were identified from the Indian gut microbiome, including 2,328 novel functions
57
58 210 obtained from clustering the unmapped genes (see Methods). The core microbial functions that are
59
60
61
62
63
64
65

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

211 essential for microbial survival and present in almost 80% individuals were used for the functional
212 comparison. The core microbiome was derived using a similar strategy as employed in MetaHIT
213 (see Methods) [25]. A set of 1,890 essential genes from six bacterial species namely, *Escherichia*
214 *coli* MG1655I and MG165II, *Bacteroides thetaiotaomicron* VPI-5482, *Pseudomonas* PA01,
215 *Salmonella enteric* serovar *Typhi* and *Staphylococcus aureus* NCTC 8325 were obtained and were
216 assigned with eggNOG annotations. The eggNOG abundance profile generated from relative
217 abundances of genes observed in Indian and other population dataset were ranked based on their
218 mean abundance in descending order. The range of eggNOGs that included 85% of the 1,890
219 essential genes were considered as a part of the core microbial eggNOG set for each population
220 dataset and was used for the analysis. Most of the essential genes were included in the top-ranking
221 clusters suggesting that the essential genes are present in higher abundance than the accessory
222 function genes (**Additional File 5: Figure S7**).

223 The core microbiome of Indian samples was compared with the core microbiome of USA, China
224 and Denmark populations. The proportion of essential genes covered by top-ranking eggNOG
225 clusters showed that 85% of the essential genes could be covered in the least number (15,300) of
226 eggNOGs in the case of Indian population, whereas it was covered by a higher (30,900) number
227 of eggNOGs in the case of USA (20,400), China (19,900) and Denmark populations (**Additional**
228 **File 5: Figure S8**). These observations suggest that the core functional microbiome of Indian
229 population is less diverse than the other populations. This corroborates well with the alpha
230 diversity (mean Shannon index) calculated using gene abundance tables rarefied at 1,000,000
231 seqs/sample (for n=30 random iterations), which also showed that the Indian microbiome is
232 significantly (P-value < 10⁻¹⁶) less diverse than the microbiome of the other populations analysed
233 in this study (**Additional File 5: Figure S9**).

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

234 In total, 5,588 eggNOGs were characterized as core functions commonly present in the core
235 microbiome of all the four population datasets. The co-inertia (Procrustes) analysis and the
236 eigenvalues calculated from PCA using both core and accessory functions also showed that the
237 Indian gut microbiome was significantly (FDR Adj. P-value = 6.4×10^{-10} , 2×10^{-16} and 0.05 with
238 China, Denmark and USA, respectively for PC1) different from the other datasets (**Fig. 2A & B**).
239 These results also show the uniqueness of Indian gut microbial functions in composition and
240 diversity at both core and accessory levels. The Indian gut microbiome was found to be enriched
241 (FDR Adj. $P < 0.05$, Log Odds Ratio > 1.5) in functions for carbohydrate and energy metabolism
242 including degradation of complex polysaccharides and glycogen and was also enriched for
243 enzymes from TCA cycle, which corroborates well with the carbohydrate-rich diet of the Indian
244 population (**Fig. 2C and 2D and Additional File 7: Enriched KO and EggNOG functions**).

245 **Unsupervised clustering of Indian samples and their association with previously identified**
246 **enterotypes**

247 A study by Arumugam et al. classified the samples from multiple populations into clusters based
248 on genus level profiles, and identified three prominent clusters called enterotypes [2]. In order to
249 identify the enterotypes from Indian gut microbiome, a meta-analysis was performed using genus
250 level abundances of samples from the four nations as used by Arumugam et al. along with the
251 Indian cohort. There were three prominent clusters observed with majority (63.6%) of Indian
252 population falling into enterotype-2, which was primarily driven by *Prevotella*. The analysis
253 revealed differences in the distribution of samples from LOC1 and LOC2, where a higher number
254 of samples from LOC1 (73.5%) were associated with enterotype-2 compared to LOC2 (54%). In
255 contrast, LOC2 samples were associated with enterotype-1 (30.3%) and enterotype-3 (16.07%),
256 which were driven by *Bacteroides* and *Ruminococcus*, respectively (**Fig. 3A; Additional File 8**).

1
2
3
4 257 An independent microbial abundance-based clustering of Indian samples using Jensen Shannon
5
6 258 distances revealed two prominent clusters. The clustering was validated using Calinski Harabasz
7
8
9 259 index (CHI) and prediction strength, which uses a cross-validation approach to validate the
10
11 260 robustness of clustering (**Additional File 9**). Cluster 1 was primarily enriched in species from
12
13
14 261 genus *Prevotella* ($P < 10^{-10}$), and Cluster 2 was quite widely spread and was enriched in species
15
16 262 belonging to *Bifidobacterium* ($P = 10^{-13}$), *Ruminococcus* ($P = 0.031$), *Clostridium* ($P = 0.04$) and
17
18
19 263 *Faecalibacterium* ($P = 0.046$) (**Additional File 5: Figure S10, Additional File 10**). The higher
20
21 264 abundance of *Prevotella* in LOC1 and *Bacteroides* in LOC2 in India are perhaps due to the
22
23
24 265 different dietary habits of the two locations. The LOC1 population was mainly consuming a
25
26 266 carbohydrate-rich diet comprising of vegetable-based foods and grains, whereas the LOC2
27
28
29 267 population was consuming a diet consisting of rice, meat and fish. Similar variations in
30
31 268 microbiome diversity due to differences in dietary habits have also been observed in earlier studies
32
33
34 269 [27, 28]. However, to confirm the above observations and to assess the quantitative effect of dietary
35
36 270 habits on microbial variations, further longitudinal studies are necessary where detailed dietary
37
38 271 information needs to be collected through a food-frequency questionnaire.

40
41 272 A similar cluster analysis performed using functional information derived from the abundance of
42
43 273 KEGG Orthologs (KO) also showed the clustering of samples into two distinct clusters, namely
44
45
46 274 C1 and C2 (**Additional File 5: Figure S11**). In comparison to clusters derived from taxonomic
47
48 275 information, only 14 out of 110 samples were placed in different clusters using the functional
49
50
51 276 information showing a similarity ($P\text{-value} = 0.6841$; Fisher's exact test; **Additional File 11**) in
52
53 277 cluster allocation using both taxonomic and functional information. C1 was found enriched in
54
55 278 genes coding for enzymes such as β -glucosidase ($\text{LOR} = 3.364$; $P\text{-value} = 10^{-20}$), and α -fucosidase
56
57
58 279 ($\text{LOR} = 0.73$; $P = 10^{-8}$), which are involved in the breakdown of plant-polysaccharides, whereas the
59
60
61
62
63
64
65

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

280 genes coding for enzymes such as lipase (LOR = -1.34; P=10⁻¹²), carnitine-coA dehydratase (LOR
281 = -1.81; P-value = 0.029) and amino peptidase (LOR = -2.72; P=10⁻¹⁰), which are involved in the
282 metabolism of animal-based diet, were enriched in C2 (FDR Adj. P<0.05) (**Additional File 12**).

283 To identify the covariates explaining the maximum variations in microbial profiles across samples,
284 unweighted UniFrac distances were calculated using phylogenetic distances between OTU
285 reference sequences and OTU table rarefied at 100,000 seqs/sample. The principal component
286 analysis of UniFrac distances and the correlation of loadings for each sample with the covariates
287 using polyserial/biserial correlation identified distinct locations (LOC1 and LOC2) and diet
288 (vegetarian and omnivorous) to be the major covariates explaining the variation in taxonomic
289 diversity between the samples (**Additional File 5: Figure S12, Additional File 13**). An ordination
290 of 110 Indian samples using gene abundance profiles from metagenomic data showed location and
291 diet to be significantly (FDR Adj. P-value < 0.01; Polyserial Correlation) associated with PC1
292 explaining the maximum variation between samples (**Additional File 5: Figure S13, Additional**
293 **File 13**). A significant correlation ($\rho = 0.708$; P-value=2x10⁻¹⁶ Spearman's rank correlation) was
294 also observed between location and diet covariates. A comparison of functional diversity using
295 gene abundance curves with increasing number of samples performed between the two locations
296 showed that the microbiome profiles of LOC2 populations were more diverse in their composition
297 compared to LOC1 populations (**Additional File 5: Figure S14**). The inter-individual Bray-curtis
298 distances calculated on normalized gene abundance profiles between LOC1 and LOC2 populations
299 also showed significant differences (FDR Adj. P<0.05), where LOC2 population displayed higher
300 inter-individual heterogeneity in their microbial community structure as compared to LOC1
301 population (**Additional File5: Figure S15**).

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

302 Major differences in the microbiome profiles were apparent at the Phylum level (using 16S rRNA
303 gene amplicon sequencing) from the higher Bacteroidetes to Firmicutes ratio (P=0.002) in LOC1
304 (1.93) compared to LOC2 (0.86), which have been previously reported as a result of differences
305 in dietary habits, i.e. vegetarian or plant-based (carbohydrate-rich) vs. omnivore or animal-based
306 (protein-rich) diets (**Additional File 5: Figure S16**) [29, 30]. Notably, these variations were not
307 attributable to BMI (Spearman's Rank correlation, FDR Adj. P=0.78). Taxonomic profiles
308 generated from metagenomic datasets through reads mapped to reference genomes were compared
309 between the two locations at genus and species level using NB model-based Wald test
310 implemented in DESeq2. *Prevotella* and *Megasphaera* were observed to be higher in LOC1,
311 whereas *Ruminococcus* and *Faecalibacterium* were higher in LOC2 (FDR Adj. P<0.05, Wilcoxon
312 rank-sum test); (**Fig. 3B**). Within these genera, *P. copri*, *P. stercorea* species were significantly
313 higher in LOC1, whereas *F. prausnitzii* and *R. bromii* belonging to genus *Faecalibacterium* and
314 *Ruminococcus*, respectively were higher in LOC2. In addition, *Akkermansia muciniphila*,
315 *Eubacterium siraeum* and *Roseburia hominis* were observed higher in LOC2, and *M. funiformis*
316 and *M. hypermegale* from genus *Megamonas* were higher in LOC1 (**Fig. 3C**). Moreover, the
317 metagenomic species derived from clustering of gene profiles showed that a total of 19
318 MGS/CAGs were enriched in LOC1 (Log Odds Ratio > 2; Adj. P<0.05), of which 7 MGS/CAGs
319 were annotated to *Prevotella copri*. Similarly, 67 MGS/CAGs were found enriched in LOC2 (Adj.
320 P<0.05; Log Odds Ratio < -2) and included 8 and 3 MGS/CAGs annotated to SCFA producing
321 species *Faecalibacterium prausnitzii* and *Roseburia inulinivorans*, respectively (**Additional File**
322 **14**). Interestingly, both, *F. prausnitzii* and *R. inulinivorans*, species enriched in LOC2 are known
323 SCFA producers, and are regarded as commensals with anti-inflammatory properties [31]. In

1
2
3
4 324 contrast, *Prevotella*, which was abundant in the LOC1, is known to be associated with fibre-rich
5
6 325 diet [32].
7
8
9

10 326 **Defining the Indian gut metabolome**

11
12
13

14 327 The analysis of microbial community structure and functions from the two locations having
15
16 328 different lifestyle and diet revealed significant insights. Previous studies have shown a direct role
17
18 329 of diet in shaping the different gut microbiomes [33]. Thus, to gain deeper insights into the
19
20 330 metabolic activity of microbiomes from LOC1 and LOC2 as driven by different diets, faecal
21
22 331 metabolites were analysed using a GC-MS-based metabolomics approach. An unsupervised
23
24 332 between class analysis of metabolomic profiles separated the samples into three separate clusters,
25
26 333 and the robustness was confirmed using prediction strength and Silhouette index (**Fig. 4A and**
27
28 334 **4B**). Polyserial correlation of covariates showed location to be the major factor explaining the
29
30 335 variation at PC1 (FDR Adj. $P < 0.01$) separating Cluster 1 from Cluster 2 and 3. In contrast,
31
32 336 vegetarian and omnivorous diet groups emerged as other factors explaining the variation at PC2
33
34 337 (FDR Adj. $P < 0.01$), and separating Cluster-2 from 3 (**Additional File 15**). The covariates location
35
36 338 and diet were also observed to be highly correlated variables showing their strong influence on gut
37
38 339 microbiome. Cluster-1 was associated with LOC1 and showed higher concentration of saturated
39
40 340 fatty acids including palmitic acid, stearic acid, and valeric acid. Cluster-3 was associated with
41
42 341 LOC2 and showed higher abundances of BCAAs, valine, leucine and isoleucine, and SCFAs,
43
44 342 propionate and butyrate concentrations. Cluster-2 was enriched in D-glucose, galactose, mannose,
45
46 343 lauric acid and cadaverine (a polyamine associated with meat consumption) and was also observed
47
48 344 to be associated with LOC2 [34]. To assess the effect of different covariates on the separation of
49
50 345 samples, PERMANOVA was performed (**Table 2**). The location was found to explain maximum
51
52 346 variation for separation of samples, whereas diet was the second most important variable in
53
54
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

347 explaining the variance. The OPLS-DA model was used to expose the class separation for each of
348 the covariates using Q^2 values which assesses the quality measurement (**Table 3**). The OPLS-DA
349 models validated by random permutation (n=200) of class labels showed Q^2 values for location
350 and diet to be higher than Q^2 values produced from random permutations with location showing
351 highest Q^2 values (**Additional File 5: Figure S17**). The OPLS-DA model also showed clear
352 separation of samples between locations as class of separation (**Fig. 4C**).

Positive correlation of BCAA transporters with BCAA levels in faecal metabolome

354 We also identified the marker metabolites, which showed significant (Spearman's correlation,
355 FDR Adj. $P < 0.05$) associations with LOC1 or LOC2. In total, 17 metabolite clusters were
356 identified, of which nine were associated with LOC1, and eight were associated with LOC2
357 (**Additional File 16**). These marker metabolites showed a positive association with MGS/CAGs.
358 For instance, *Prevotella* annotated clusters correlated significantly with valeric acid and
359 sedoheptulose metabolite markers, which showed a higher relative abundance in LOC1. In
360 contrast, MGS/CAGs belonging to *Faecalibacterium*, *Clostridium*, *Ruminococcus*, and *Alistipes*
361 were positively associated with BCAAs, cadaverine, propanoate and lauric acid in LOC2 (**Fig.**
362 **5A**). In addition to the positive association of BCAAs with species enriched in LOC2, a correlation
363 analysis of significantly different (FDR Adj. $P < 0.05$, DESeq2-based Wald test; **Additional File**
364 **17**) functional modules revealed that faecal BCAA abundances were positively correlated with
365 BCAA transporter abundance in LOC2. In contrast, BCAA abundance in the faecal metabolome
366 showed a negative correlation ($P < 0.05$) with BCAA biosynthesis pathways (**Fig. 5B**).

367 The above observations are significant given that BCAAs are important metabolites involved in
368 glucose homeostasis by stimulating insulin secretion [35]. Higher BCAA levels in the faecal
369 samples could be a result of its uptake by microbes via BCAA transporters, leading to their

1
2
3
4 370 accumulation in the microbial cells detected in faecal metabolome. This is concordant with higher
5
6 371 relative abundance of *Bacteroides vulgatus* and *Eubacterium sireaeum* in LOC2 compared to
7
8
9 372 LOC1, which are known to harbour higher abundance of BCAA transporters (**Fig.3C**) [8]. Further
10
11 373 support for this hypothesis emerged from the correlation of circulating BCAA levels (valine and
12
13 374 isoleucine) in serum with the corresponding concentrations in faeces. Interestingly, serum BCAA
14
15
16 375 concentrations were significantly higher in LOC1 individuals as compared to LOC2 individuals,
17
18
19 376 which is in contrast with their BCAA levels in the faecal metabolome (**Fig. 6A**). Thus, one
20
21 377 possibility is that the accumulation of BCAA in the faeces of individuals of LOC2 was mediated
22
23
24 378 by the inward transport of BCAA by the gut bacteria. In contrast, the lower BCAA accumulation
25
26 379 in gut microbes and a higher BCAA biosynthesis by microbial species and its eventual absorption
27
28
29 380 in serum appears to be a plausible reason for the higher BCAA concentrations in serum of LOC1
30
31 381 population.

32 33 34 382 **Role of *Prevotella copri* in the regulation of BCAA levels**

35
36
37 383 To explore the differences in association of functional pathway modules between the two
38
39 384 locations, KOs within each module were correlated with KOs from other modules using
40
41
42 385 Spearman's correlation coefficient. The KOs showing significant differences in correlations
43
44 386 between LOC1 and LOC2 were identified. This differential correlation analysis of BCAA
45
46
47 387 biosynthetic modules with other pathways in LOC1 and LOC2 revealed that BCAA modules were
48
49 388 independently driven in LOC1 and LOC2 (Spearman's rank correlation, FDR Adj. $P < 0.01$)
50
51 389 (**Additional File 5: Figure S18A & B**). To identify the species and the metabolic pathways that
52
53
54 390 contributed most to the BCAA abundance in faecal and serum metabolome profiles, a correlation
55
56 391 analysis with iterations leaving each species out was performed for each metabolic module (**Figure**
57
58
59 392 **6B**). The species whose removal leads to a maximum change in the correlation of metabolic
60
61
62
63
64
65

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

393 pathway with metabolite was identified, and was considered as an important contributor of that
394 metabolite.

395 Notably, the BCAA biosynthesis-dependent changes in BCAA levels were largely driven by
396 *Prevotella* species through threonine-dependent and independent biosynthesis pathways as
397 observed from Delta SCC_{bg} values when genes from this species were removed (see Methods).
398 The correlation network analysis with differential MGS/CAGs revealed threonine-independent
399 isoleucine biosynthesis pathway to be highly correlated with *Prevotella copri* in LOC1 (**Fig. 6C**).
400 The first enzyme, D-citramalate synthase, catalysing the first step of threonine-independent
401 isoleucine biosynthesis pathway was also observed as highly enriched (LOR = 1.7) in LOC1 [36].
402 Further, BCAA biosynthesis pathways was found higher in LOC1, whereas BCAA transporters
403 were higher in LOC2 (**Fig. 6D**) leading to the dynamic changes in BCAA concentrations in faecal
404 and serum metabolome in LOC1 and LOC2 as observed in Fig. 6A.

405 **Discussion**

406 Compositional and functional human gut microbiome studies in different populations have been
407 instrumental in establishing the role of gut microbiome in human health [2, 28, 37, 38]. However,
408 such population-specific signatures and their functional roles are yet unknown for the Indian gut
409 microbiome. This study provides the first insights into the Indian gut microbiome represented
410 through a cohort of 110 individuals from two prominent locations to reveal the taxonomic and
411 functional diversity using 16S rRNA gene, metagenomic analysis, and metabolomic profiling.
412 Although the sequencing depth was modest (1.36 ± 0.5 Gbp per sample, mean \pm standard
413 deviation), the inclusion of 110 individuals from two distinct geographic locations as well as the
414 identification of Indian gut microbiome-specific genes provide a first insight into the Indian gut
415 microbiome and are thus considered important additions to the field. The selection of two distinct

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

416 sub-populations (Bhopal-LOC1, and Kerala-LOC2) was an important consideration to capture the
417 microbiome variations resulting from different diet and lifestyle of these two cohorts. LOC1
418 provided a representation of the population from North-Central India mainly consuming a
419 carbohydrate and fat rich diet, whereas LOC2 represented a population from Southern India
420 consuming an omnivorous diet with rice and animal-based products as the primary components.

421 This study established the gene catalogue of the Indian gut microbiome, which provides the first
422 insights into the yet unknown functional gut microbiome of the Indian population. The genes
423 encoding several transposons, peptidase, glucosidase, and plant polysaccharide degradation
424 enzymes were unique to the Indian population and not represented in other microbiome datasets.
425 The Updated-IGC (IGC+India) constructed by the addition of unique non-redundant genes from
426 the Indian population to the Integrated gene catalogue is likely to act as a reference dataset for gut
427 microbiome studies for global comparative studies, and particularly for studies involving South-
428 Asian populations that have similar dietary habits and lifestyle.

429 In addition to the basic housekeeping functions of the gut microbiome, which were also found
430 abundant in other datasets, the Indian gut microbiome was enriched in functions for carbohydrate
431 and energy metabolism including degradation of complex polysaccharides, which corroborates
432 well with the typical carbohydrate-rich diet of the Indian population [39]. The distant clustering of
433 Indian samples from other populations revealed the unique composition of the Indian gut
434 microbiota (**Fig. 1B**). *Prevotella* emerged as the most discriminatory genus associated with the
435 Indian population as revealed by both amplicon and MGWAS. The abundance of *Prevotella* was
436 also indicated in the previous 16S rRNA gene-based microbiome studies of the Indian population
437 carried out in small to medium-sized cohorts [18, 19]. Recently, *Prevotella* has been commonly
438 observed in different non-Western communities that consume a plant-rich diet, such as in the

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

439 Papua New Guineans, native Africans, rural Malawians, BaAka pygmies, etc [11, 40]. and has
440 also been associated with vegetarianism in the Western populations [41, 42]. However, it has not
441 been observed at such high abundance in the western countries so far. The MGWAS approach in
442 this study showed the presence of *Megasphaera*, *Lactobacillus* and *Mitsuokella* as the other major
443 genera associated with the Indian gut microbiome.

444 Several recent studies have shown a relationship between the abundance of specific strains of
445 *Prevotella* with inflammatory diseases, since it has a higher intrinsic capacity to stimulate Th17-
446 mediated inflammation, which is generally not expected in a strict commensal bacterium [41, 43,
447 44]. However, the high abundance of *Prevotella* in the healthy gut microbiome of the Indian
448 population does not corroborate with its potential inflammatory role reported so far. Since this
449 study was only focussed on the gut microbiome of healthy individuals, it is difficult to draw
450 conclusions on the potential inflammatory role of this species. One potential explanation could be
451 the complex set of interactions between host genetic risk factors and environment in which the
452 presence of *Prevotella* may be only one of the factors [45]. Further, strain-level variations are
453 known in the inflammatory responses and not all species of *Prevotella* could be potentially
454 inflammatory, as also evident from the known high genetic diversity within and between the
455 species of *Prevotella* [43]. Thus, the high abundance of *Prevotella* in the healthy microbiota
456 emphasizes the requirement for larger cohort studies in different populations to gain deeper
457 insights into the potential inflammatory roles of gut microbes.

458 The abundance of *Prevotella* has been associated with plant-based diets, and the typical
459 carbohydrate-rich diet of the Indian population could be one of the reasons for the over-
460 representation of this genus in the Indian gut microbiome [46]. Likewise, the predominance of
461 other microbial species from genus *Lactobacillus*, *Megasphaera* and *Mitsuokella* could be due to

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

462 the higher intake of fermented food and dairy products along with the carbohydrate-rich diet in
463 LOC1 [46, 47]. Similarly, *Bacteroides* and *Clostridium*, which were abundant in LOC2, are
464 associated with diets rich in animal-based products, consistent with the omnivorous diet of LOC2
465 [42]. Interestingly, taxonomy-based clusters 1 and 2 showed associations with the two locations
466 LOC1 and LOC2, and also with the two KO-based clusters (C1 and C2) (**Additional File 5: Figure**
467 **S10 and S11**). It is to be noted that C1 was enriched in enzymes involved in the degradation of
468 carbohydrate and plant polysaccharides, which correlates well with the carbohydrate-rich diet in
469 LOC1. In contrast, C2 was enriched in enzymes involved in lipid and protein degradation, which
470 relate to the constituents of an omnivorous diet in LOC2. These observations further support the
471 correlation between location, diet and enterotype. Although, the concept of enterotype
472 classification is sometimes criticised due to statistical weakness in some studies, however, a meta-
473 analysis of Indian samples with samples from Arumugam et al. revealed three robust clusters with
474 Indian samples mostly associated with enterotype-2 driven by *Prevotella* [2]. This classification
475 of samples from multiple population/studies into enterotypes has the potential to be clinically
476 relevant in various aspects such as disease diagnosis, early-detection of disease, biomarker
477 development, personalised treatments and xenobiotic metabolism [48]. It is a representation of the
478 major microbial species in the gut microbiome, and thus appears useful for microbiome-based
479 population stratification. A robust statistical analysis with increased sample sizes, direct clinical
480 associations, and detailed molecular interventions are essential for further strengthening its
481 potential.

482 The study also established the previously unknown faecal metabolome of the Indian population,
483 which showed strong clustering into three metabolomic clusters differentiated by location and diet.
484 The metabolomic clusters also correlated well with the respective dietary habits of the two

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

485 locations, where metabolomic Cluster-1 showed an association with LOC1 and was enriched in
486 saturated fatty acids such as palmitic acid and stearic acid, whereas metabolomic Cluster-3 showed
487 an association with LOC2, and was enriched in BCAAs such as isoleucine, valine and leucine, and
488 SCFAs such as propionic acid, and butyric acid. A medium chain fatty acid (MCFA) ‘lauric acid’
489 was also found abundant in LOC2 perhaps due to the high dietary consumption of coconut oil in
490 this location [49, 50]. Lauric acid has been reported to be beneficial by preventing fat deposition
491 in blood vessels and acting as an anti-inflammatory and anti-oxidative agent [51].

492 The major BCAA ‘isoleucine’ being produced through a less common threonine-independent
493 pathway for isoleucine biosynthesis, and the higher enrichment of the key enzyme, D-citramalate
494 synthase of the above pathway confirmed its higher abundance in LOC1 as compared to LOC2.
495 Further, this pathway was found to be associated with a single species, *Prevotella copri* as reported
496 earlier [8]. Taken together, it appears that the higher abundance of BCAA biosynthesis genes and
497 a lower abundance of BCAA inward transporters in gut microbiome resulted in the lower BCAA
498 accumulation in the fecal metabolome, and higher BCAA concentration in serum as observed in
499 LOC1 (Fig. 7) [8]. However, a contrasting pattern was observed in the case of LOC2, where the
500 lower abundance of BCAA biosynthesis genes and the higher abundance of BCAA inward
501 transporters correlated well with the higher and lower BCAA concentrations in faeces and serum,
502 respectively.

503 The higher levels of SCFAs in LOC2 could be a consequence of the consumption of omnivorous
504 diet, which is associated with a Firmicutes-rich gut microbiome [52]. SCFAs have well-established
505 roles in human health as an energy source, an anti-inflammatory agent, and for improving intestinal
506 homeostasis by increasing IL-18 production [53]. In contrast, higher serum BCAA levels have
507 well-known roles in promoting insulin resistance and Type-2 Diabetes (T2D), and were found

1
2
3
4 508 higher in the serum in LOC1. Several reports on the role of a high-fat diet in the modulation of
5
6 509 microbiota and alteration in intestinal barrier are emerging with results showing increased
7
8
9 510 absorption and circulating levels of branched-chain amino acid (BCAA) and reduction of SCFAs
10
11 511 such as butyrate, acetate, propionate, and secondary bile acids, as also noted in the case of LOC1
12
13
14 512 [54, 55]. High-fat and carbohydrate-rich diets have also been associated with an increase in
15
16 513 abundance of Bacteroidetes (gram-negative bacteria) leading to a skewed Bacteroidetes:
17
18
19 514 Firmicutes ratio towards the former phylum [32]. Such a ratio was also apparent in this study in
20
21 515 LOC1 dominated by *Prevotella* from the phylum Bacteroidetes. Further, a higher serum
22
23
24 516 concentrations of circulating BCAA were also observed in LOC1. These results provide hints on
25
26 517 the role of dietary habits in shaping the gut microbiome and its plausible impact on the BCAA and
27
28
29 518 SCFA dynamics observed in these populations.

30
31
32 519 To conclude, this multi-omics based gut microbiome study of a healthy cohort from two different
33
34 520 parts of India provides novel insights into the Indian gut microbiome and metabolome, and reveals
35
36 521 the unique gene catalogue from the poorly characterized Indian population. Further studies using
37
38
39 522 higher sequencing depths, and including both healthy and diseased individuals will help in
40
41 523 obtaining more comprehensive functional and taxonomic information of gut microbiome from
42
43
44 524 Indian population and its impact on human health.

45 46 47 525 **Methods**

48 49 50 526 **Study design and subject enrolment**

51
52
53 527 The study cohort consisted of 110 healthy individuals belonging to different age groups from
54
55
56 528 infants (<1 year) to aged (>50 years), with an average subject age of 29.72 ± 17.4 years (mean \pm
57
58 529 sd) from two different locations across India i.e., Bhopal (LOC1, n=53) and Kerala (LOC2, n=57),
59
60
61
62
63
64
65

1
2
3
4 530 which are separated by ~1000 miles. LOC1 was located in North-Central India with the majority
5
6
7 531 of population being vegetarian, whereas LOC2 was located in Southern India where the population
8
9 532 with dietary habits mostly consisting of rice, seafood and red meat (Diet description section in
10
11 **Additional File 1**). According to the 'Indian Food Composition Table', the primary Indian diet is
12
13
14 534 rich in carbohydrates such as rice, wheat and potato, and in fat and proteins from milk and dairy
15
16 535 products [56]. In addition, several accompaniments to the primary diet also exist including a
17
18
19 536 variety of grains, vegetables, fruits, and usage of oil, spices and animal products.

20
21
22 537 The faecal samples for metagenomics and blood samples for serum metabolomics were collected
23
24 538 from healthy participants and their metadata is provided in **Additional File 1** under the Metadata
25
26
27 539 information section. The recruitment of volunteers, sample collection, and other study-related
28
29
30 540 procedures were carried out by following the guidelines and protocols approved by the Institute
31
32 541 Ethics Committee of Indian Institute of Science Education and Research (IISER), Bhopal, India.
33
34 542 Each faecal sample was frozen within 30 mins of collection. A written informed consent was
35
36
37 543 obtained from all subjects prior to any study-related procedures, along with information on gender,
38
39 544 age, and diet for a period of one month prior to the collection of faecal samples. The recruited
40
41
42 545 individuals did not undergo any medication at least one month prior to the sample collection. All
43
44 546 the recruited individuals had an average BMI of 21.16 (± 5.23), and were not diagnosed with T2D
45
46
47 547 at the time of sample collection, and did not have a second-degree relative history of T2D. The
48
49 548 above samples were then used for 16S rRNA gene V3 hypervariable region amplicon sequencing,
50
51 549 shotgun metagenomic sequencing, and metabolomic analysis.

52 53 54 55 550 **Faecal metagenomic DNA extraction**

56
57 551 Metagenomic DNA was isolated from all the faecal samples using QIAamp Stool Mini Kit
58
59
60 552 (Qiagen, CA, USA) according to the manufacturer's instructions. DNA concentration was
61
62
63
64
65

1
2
3
4 553 estimated by Qubit HS dsDNA assay kit (Invitrogen, CA, USA), and quality was estimated by
5
6 554 agarose gel electrophoresis. All the DNA samples were stored at -80 °C until sequencing.
7
8
9

10 555 **16S rRNA gene amplicon and shotgun metagenome sequencing**

11
12 556 The extracted DNA (5ng) was PCR amplified with seven different custom modified 5'-end
13
14
15 557 adaptor-ligated 341F and 534R primers (See the primer details section in Additional File 1)
16
17 558 targeting the V3 hypervariable region of 16S rRNA gene. After evaluating the amplified products
18
19
20 559 on 2% w/v agarose gel, the products were purified using Ampure XP kit (Beckman Coulter, Brea,
21
22 560 CA USA). Amplicon libraries were prepared by following the Illumina 16S rRNA gene
23
24 561 metagenomic library preparation guide. Metagenomic libraries were prepared using Illumina
25
26
27 562 Nextera XT sample preparation kit (Illumina Inc., USA) by following the manufacturer's protocol.
28
29
30 563 Library size of all the libraries was assessed using Agilent 2100 Bioanalyzer (Agilent
31
32 564 Technologies, Santa Clara, USA.), and quantified on a Qubit 2.0 fluorometer using Qubit dsDNA
33
34 565 HS kit (Life technologies, USA) and by qPCR using KAPA SYBR FAST qPCR Master mix and
35
36
37 566 Illumina standards and primer premix (KAPA Biosystems, Wilmington, MA, USA) following the
38
39 567 Illumina suggested protocol. Both the amplicon and metagenomic libraries were loaded on
40
41
42 568 Illumina NextSeq 500 platform using NextSeq 500/550 v2 sequencing reagent kit (Illumina Inc.,
43
44 569 USA), and 150 bp paired-end sequencing was performed at the Next-Generation Sequencing
45
46
47 570 (NGS) Facility, IISER Bhopal, India.
48
49

50 571 **Amplicon-based taxonomic analysis**

51
52 572 A total of 24 Gbps of data were retrieved on de-multiplexing of paired-end reads with an average
53
54
55 573 of 210 Mbp per sample. The paired-end reads were assembled using FLASH and were quality
56
57 574 filtered at Q20 (80% bases) Phred quality score using NGSQC Toolkit v 2.3.3 [57, 58]. The primer
58
59
60 575 sequences were trimmed from the High Quality (HQ) reads. The reads were further clustered into
61
62
63
64
65

1
2
3
4 576 OTUs using closed-reference OTU picking protocol of QIIME at $\geq 97\%$ identity against ARB
5
6 577 SILVA database release 132 (13th December 2017) [59, 60]. The most abundant read was selected
7
8
9 578 as the representative sequence for each OTU and was assigned with taxonomy using the SILVA
10
11
12 579 database. OTU table containing the abundance of each OTU for each sample was generated and
13
14 580 used for further analysis. For phylogenetic analysis, representative 16S rRNA genes of phylotypes
15
16 581 were aligned against a core set of 16S rRNA gene sequences using align_seqs.py with the PyNAST
17
18
19 582 v.1.2.2 algorithm [61]. The unweighted UniFrac distances between samples were calculated using
20
21 583 rarefied OTU abundance (100,000 seqs/sample) table and phylogenetic distances between
22
23
24 584 representative sequences from each OTUs [62].
25
26

27 585 **Pre-processing of the Metagenomic reads**

28
29 586 A total of 150 Gbp of metagenomic sequence data (mean = 1.36 Gb) was generated from 110
30
31
32 587 faecal samples. The metagenomic reads were filtered using NGSQC toolkit v2.3.3 with a cutoff
33
34 588 $\geq Q20$ [57]. The high-quality reads were further filtered to remove the host-origin reads (human
35
36 589 contamination) from metagenomic reads using 18mer matches parameter in Best Match Tagger
37
38
39 590 BMTagger v3.101 (BMTagger, RRID:SCR_014619; [http://casbioinfo.cas.unt.edu/sop/mediawiki/](http://casbioinfo.cas.unt.edu/sop/mediawiki/index.php/Bmtagger)
40
41 591 [index.php/Bmtagger](http://casbioinfo.cas.unt.edu/sop/mediawiki/index.php/Bmtagger)), which resulted in the removal of an average of 1% reads. The reads from
42
43
44 592 each sample were assembled separately into contigs using IDBA ud version 1.1.0 [63] with
45
46 593 parameters “-mink 31 -maxk 87 -step 5”. The reads from each samples were mapped to contigs
47
48
49 594 to estimate read recruitment using FR-HIT version 0.7 [64]. The unmapped reads resulting from
50
51 595 each sample were pooled together and denovo assembly was performed on the combined set of
52
53
54 596 singleton (unmapped) reads from all samples. The ORFs from each contig (length ≥ 500 bp) were
55
56 597 predicted using MetaGeneMark v.3.38 [65]. Pair-wise alignment of genes was performed using
57
58
59 598 BLAT version 2.7.6 [66], and the genes which had an identity $\geq 95\%$ and alignment coverage \geq
60
61
62
63
64
65

1
2
3
4 599 90% were clustered into a single set of non-redundant genes, from which the longest gene was
5
6
7 600 selected as the representative ORF to construct the non-redundant gene catalog.

8
9
10 601 Integrated Gene Catalog (IGC), which represents 1,297 human gut metagenomic samples
11
12 602 comprising of HMP, MetaHIT and Chinese datasets, was retrieved [24]. The gene catalogue
13
14
15 603 constructed from Indian samples was combined with the IGC to construct a non-redundant gene
16
17 604 catalog (using identity $\geq 95\%$ and alignment coverage $\geq 90\%$) and is referred to as ‘Updated-IGC’
18
19
20 605 in the subsequent analysis.

21 22 23 606 **Quantification of gene content**

24
25 607 The quantification of gene content was carried out using the strategy performed by Qin et al., [7]
26
27
28 608 where the high-quality reads were aligned against the updated IGC using SOAP2 in SOAP aligner
29
30 609 version 2.21 with an identity cut off $\geq 90\%$ [67]. Two types of alignments were considered for
31
32
33 610 sequence-based profiling:

34
35
36 611 (1) The entire paired-end read mapped to the gene.

37
38
39 612 (2) One end of paired-end read mapped to a gene and other end outside genic region.

40
41
42 613 In both cases, the mapped read was counted as one copy.

43
44
45
46 614 The relative abundance of a gene within the sample was calculated as: $a_i = \frac{b_i}{\sum_j b_j}$

47
48
49
50 615 a_i : relative abundance of gene in sample S; x_i : The times in which gene i was detected in sample S
51
52
53 616 (the number of mapped reads); b_i : copy number of gene i in sequenced data from sample S.

54 55 56 617 **Phylogenetic assignment of reads**

57
58
59
60
61
62
63
64
65

1
2
3
4 618 A total of 4,097 reference microbial genomes were obtained from Human Microbiome Project
5
6 619 (HMP) and National Centre for Biotechnology Information (NCBI) on 5th December 2015
7
8
9 620 (**Additional File 18**). The databases were independently indexed into two Bowtie indexes using
10
11 621 Bowtie-2 version 2.2.9 (Bowtie 2, RRID:SCR_016368) [68]. The metagenomic reads were aligned
12
13
14 622 to the reference microbial genomes using Bowtie-2. The mapped reads from both indexes were
15
16 623 merged by selecting the alignment having the higher identity ($\geq 90\%$ identity). The percent identity
17
18
19 624 was calculated using the formula: $\%identity = 100 * (\text{matches} / \text{total aligned length})$. The normalized
20
21 625 abundance of a microbial genome was calculated by summing the total number of reads aligned to
22
23
24 626 its reference genome. For reads showing hits to both indexed databases with equal identity, each
25
26 627 genome was assigned 0.5 read count. The relative abundance of each genome was calculated by
27
28
29 628 adding the normalized abundance of each genome divided by the total abundance. The Calinski
30
31 629 Harabasz index (CHI) was used to calculate the variance between the clusters compared to the
32
33
34 630 variance within clusters [2].
35

36 631 **Construction of common core microbial functions**

37
38
39 632 To identify the core microbial functions in the gut microbiome of Indian populations and to
40
41 633 understand their abundance compared to the other populations, the core microbiome was
42
43
44 634 constructed using a similar strategy as mentioned in MetaHIT [25]. However, to construct a
45
46 635 comprehensive core functional microbiome, the information of essential functions from six
47
48
49 636 different microbes including two strains of *Escherichia coli*, *Bacteroides thetaiotaomicron*,
50
51 637 *Pseudomonas aeruginosa*, *Salmonella enteric* and *Staphylococcus aureus*, was used instead of
52
53
54 638 considering a single microorganism. The list of essential genes was collected from DEG database
55
56 639 v5.0 [69]. 1,890 genes were identified as essential genes in all the six microorganisms. These genes
57
58
59 640 were aligned against eggnog v4.1 database using diamond and were annotated with eggNOG ID
60
61
62
63
64
65

1
2
3
4 641 [70, 71]. The core gut microbiome functions were also calculated using the above strategy for the
5
6 642 USA, Denmark and Chinese population gut microbial samples to remove the variations arising
7
8
9 643 due to differences in data analysis procedures. Apart from identifying the clusters that represented
10
11 644 $\geq 85\%$ genes within the range of essential gene functions, the low prevalent eggNOG functions,
12
13
14 645 which were present in $\geq 0.0001\%$ abundance in $\geq 80\%$ of samples in that population, were further
15
16 646 filtered out. This added filtration step helped in removing all the low abundant functions. To
17
18
19 647 represent the core, the variance of these functions was also calculated between the two Indian
20
21 648 locations. The eggNOGs showing significant deviations in variations ($P\text{-value} \leq 0.05$; Levene's
22
23
24 649 test) [72] were further filtered out from the analysis.

27 650 **Construction of Metagenomic Species for MGWAS**

28
29
30 651 To identify metagenomic markers using a reference-independent approach on metagenomic
31
32 652 samples, a metagenome-wide association study was performed for 340 samples (age and gender
33
34 653 matched) including India (both locations), USA, China and Denmark populations. The genes
35
36
37 654 present in at least $\geq 10\%$ of samples were considered and clustered using the canopy-mgs algorithm
38
39 655 as described [73]. The genes having Pearson's correlation coefficient (≥ 0.9) were clustered into
40
41
42 656 CAGs. Furthermore, the genes for which $\geq 90\%$ abundance was obtained from a single sample
43
44 657 were discarded.

45
46
47
48 658 To determine the taxonomic origin of each MGS/CAG (metagenomic cluster), all the genes were
49
50 659 aligned against reference microbial genomes of 4,097 genomes from HMP and NCBI at nucleotide
51
52 660 level using BLASTN [74]. The alignment hits were filtered using an E-value $\leq 10^{-6}$ and alignment
53
54
55 661 coverage $\geq 80\%$ of the gene length, and 2,773,591 (25.6%) genes showed alignments against the
56
57
58 662 reference genomes. The remaining 8,049,540 unassigned genes were aligned against UNIREF
59
60 663 database (UniRef 50) at protein sequences [75], of which 4,553,299 genes (56.56%) could be

1
2
3
4 664 assigned with taxonomic annotations. The sequences that found multiple top hits with equal %
5
6 665 sequence identity and scores were further assigned taxonomy based on LCA (Lowest Common
7
8
9 666 Ancestor) method. The genes were finally assigned to taxa based on comprehensive parameters of
10
11 667 sequence similarity across phylogenetic ranks as described earlier [76]. The identity threshold of
12
13
14 668 $\geq 95\%$ was used for assignment up to species level, $\geq 85\%$ identity threshold for assignment up to
15
16 669 genus level, and $\geq 65\%$ identity was used for phylum level assignment using BLASTN. The
17
18
19 670 taxonomic assignments of MGS/CAGs were performed with the criteria that $\geq 50\%$ genes in each
20
21 671 MGS should map to the same lowest phylogenetic group. Thus, if a particular species is assigned
22
23
24 672 $\geq 50\%$ genes out of the total genes, the assignment will be carried out at species level rather than
25
26 673 at genus or higher orders. The relative abundance of MGS/CAGs in each sample was estimated by
27
28
29 674 using relative abundance values of all genes from that MGS/CAG. A Poisson distribution was
30
31 675 fitted to the relative abundance values of the data. The mean estimated from Poisson distribution
32
33
34 676 was assigned as the relative abundance of that MGS. The profile of MGS/CAGs were generated
35
36 677 and used for further analysis.

678 **Faecal and Serum metabolomic sample preparation and derivatization**

679 Lyophilized faecal samples were used to achieve better metabolite coverage as described
680 previously [77]. Metabolites were extracted with 1 mL of ice-cold methanol: water (8:2) from 80
681 mg of lyophilized samples in a bath ultrasonicator (BioruptorTM UCD-200, Diagenode, USA) at
682 4°C for 30 min followed by 2 min of vortexing. The supernatant was extracted by centrifugation
683 at 18,000 g for 15 min at 4°C and dried at 50°C under a gentle stream of nitrogen gas. To remove
684 the residual water molecules from the samples, 100uL of toluene was added to the dry residue and
685 evaporated completely at 50°C under nitrogen gas. Dry extracted metabolites were first derivatized
686 with 50 uL of methoxyamine hydrochloride (MOX) in pyridine (20 mg/mL) at 60°C for 2 hours,

1
2
3
4 687 and the second derivatization was performed with 100 uL of MSTFA in 1% TMCS at 60°C for 45
5
6 688 min to form trimethylsilyl (TMS) derivatives. Finally, 150 uL of the TMS derivatives was
7
8
9 689 transferred into a GC glass vial inserts and subjected to GC/TOFMS analysis. Serum samples were
10
11
12 690 prepared (polar metabolites only) and derivatized as described by Psychogium et al., 2011 [78].
13

14 691 **Method development and validation**

15
16
17 692 Matrix dilution approach was used for validating the linearity and range of dilution [77]. Pooled
18
19
20 693 faecal samples were used to create the reference peaks to validate the peaks coming from
21
22 694 individual samples, which were needed due to the presence of a relatively high abundance of faecal
23
24
25 695 metabolites in the pooled samples. The supernatant of feces after extraction was serially diluted 2,
26
27 696 5, 10, 50, 100, 200 and 500 times with methanol: water (8:2). At dilution 2, the maximum numbers
28
29
30 697 of peaks were seen and were processed with the same dilution factor for all the samples. A total of
31
32 698 30 chemical standards mixture and the pooled faecal samples were used to validate the method.
33
34
35 699 Each stock solution of test standard was carefully prepared in deionized water or with pure ethanol
36
37 700 (50,150 350, 500 um) for the determination of linear range, regression coefficient (R²), limit of
38
39 701 detection (LOD), and repeatability. L-norvaline (1, 2.5, 5, 10, 20 mg/ml in ethanol) was used as a
40
41
42 702 spiked external standard for the optimized derivatization of the method.
43

44 703 **GC-MS analysis**

45
46
47
48 704 GC-MS was performed on an in-house Agilent 7890A gas chromatograph with 5975C MS system.
49
50 705 An HP-5 (25 m × 320 um × 0.25 um i.d.) fused silica capillary column (Agilent J&W Scientific,
51
52
53 706 Folsom, CA) was used with the open split interface. The injector, transfer line and ion source
54
55 707 temperatures were maintained at 220, 220 and 250 °C, respectively. Oven temperature was
56
57
58 708 programmed at 70°C for 0.2 min, and increased at 10°C/min to 270°C where it was sustained for
59
60 709 5 min, and further increased at 40°C/min to 310°C where it was held for 11 minutes. The MS was
61
62
63
64
65

1
2
3
4 710 operated in the electron impact ionization mode at 70eV. Mass data were acquired in full scan
5
6 711 mode from m/z 40 to 600 with an acquisition rate of 20 spectra per second. To detect retention
7
8 712 time shifts and enable Kovats retention index (RI) calculation, a standard Alkane series mixture
9
10 713 (C10–C40) was injected periodically during the sample analysis. RIs are relative retention times
11
12 714 normalized to n-alkanes eluted adjacently. For serum samples, we used 2uL aliquot with a split
13
14 715 ratio of 4:1 on the same column as described above. The injector port temperature was held at
15
16 716 250°C, and the helium gas flow rate was set to 1mL/min at an initial oven temperature of 50°C.
17
18 717 The oven temperature was increased at 10°C/min to 310°C for 11min and mass data were acquired
19
20 718 in full scan mode from m/z 40 to 600 with an acquisition rate of 20 spectra per second.
21
22
23
24

25 26 719 **Metabolomic analysis and metabolite profile generation**

27
28
29 720 Raw CDF files were used for peak identification and filtering, and the XCMS package in R were
30
31 721 used for pre-processing of the peaks. First, the parameters used for pre-processing of the reads
32
33 722 were optimized by calculating the reliability index using the formula given below:
34
35
36

37 723 Reliability index = (number of reliable peaks)²/number of unreliable peaks.
38
39

40 724 The reliable peaks were identified for each of the settings such as fwhm, S/N and bw, with a
41
42 725 predefined range of values and regression coefficient was calculated for dilutions of QC samples.
43
44

45 726 The number of peaks with a high coefficient of determination ($R^2 \geq 0.9$) were considered reliable,
46
47 727 whereas the peaks with very low $R^2 (\leq 0.05)$ were considered unreliable peaks [79]. The finally
48
49 728 optimized parameters were: profmethod = bin, method = matched Filter, fwhm =8 and 5 for
50
51 729 faecal and serum samples, respectively, and S/N = 12 and 3 for faecal and serum samples,
52
53 730 respectively, bw =5 (for first grouping), smooth = linear, family = gaussian, extra = 1, plot type
54
55 731 = mdevden, missing =8, bw = 3 (for second grouping). Further, to compare across multiple
56
57 732 samples, the peak intensities were normalized (root transformed) and scaled using z-
58
59
60
61
62
63
64
65

1
2
3
4 733 transformation. These normalized and scaled peak intensities were used for further statistical
5
6 734 analysis.

7
8
9 735 A multivariate statistical method, Orthogonal Projections to Latent Structures Discriminant
10
11 736 Analysis (OPLS-DA) [80], was used to identify differences between LOC1 samples (n=53) and
12
13
14 737 LOC2 (n=55) samples. Metabolites driving the differences were identified in metabolic profiles
15
16 738 of LOC1 and LOC2 samples using correlations coefficients. The clusters of co-abundant
17
18
19 739 metabolite profiles were identified using R package "WGCNA" [81]. Signed weighted
20
21 740 metabolite co-abundance correlation after scaling and centering was calculated across all
22
23
24 741 samples. The soft threshold of $\beta = 15$ was chosen for scale-free topology. The dynamic hybrid
25
26 742 tree cutting algorithm was used to identify the clusters with a deepsplit = 4 and minimum cluster
27
28
29 743 size = 4. The profile of each faecal metabolite cluster was summarized using eigenvector. The
30
31 744 abundance profile of each cluster of metabolites (MES) was calculated using the same
32
33
34 745 methodology as used for MGS cluster abundance profiles.

35 36 746 **Retention index (RI) calculation**

37
38
39 747 GC-MS data obtained from the alkene series run was used to calculate the RI for each peak in
40
41
42 748 the samples, and the obtained RI values were further used at the time of library search for the
43
44 749 identification of individual metabolite.

$$45
46
47 750 \quad I = 100 X [n + (\log tx - \log tn) / (\log tn + 1 - \log tn)]$$

48
49
50 751 Where, tx = retention time of the peak, tn = retention time of preceding alkane, and tn+1 =
51
52 752 retention time of the following alkane.

53 54 55 56 753 **Clustering and enterotype Analysis**

57
58
59
60
61
62
63
64
65

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

754 Cluster of samples in the dataset were identified from the relative abundance profiles of Genus or
755 Orthologous groups (OG) in the samples. The Jensen-Shannon distances (which estimates the
756 probability distributions between the samples) were calculated and the abundance profiles were
757 clustered using PAM (partitioning around medoids) clustering algorithm as mentioned previously
758 [82]. The optimal number of clusters was assessed using Calinski Harabasz index (CHI) that has
759 shown good performance in recovering the optimal number of clusters [83]. Similarly, the
760 prediction strength from ‘fpc’ package in R which used cross-validation approach was also
761 employed as another metric for cluster validation. Both the CHI and prediction strength showed
762 quite significantly correlated results. For clustering, CHI and prediction strength gave non-
763 identical values, silhouette index was calculated to estimate the robustness of clusters.

764 Between class analysis

765 The between class analysis was performed to identify the drivers and support the clustering of the
766 genus/species/OG abundance profiles into clusters. The between class analysis is a type of
767 principal component analysis with instrumental variables which maximizes the separation between
768 classes of this variable. The instrumental variables here is the cluster classification using PAM
769 clustering and the top species, which contributed the maximum to the principal components
770 obtained from between class analysis were identified as driver species/genus/OG based on their
771 eigenvalues. The analysis was performed using ade4 package in R.

772 Diversity Analysis

773 The inter-sample Canberra distances were also calculated using MGS Abundance between
774 populations. The richness of microbiome samples across populations was obtained from Shannon
775 index calculated using raw gene abundance table rarefied at equal depth (1,000,000 seqs/sample)
776 over n=30 random samplings. The beta diversity for 16S rRNA genes (between the samples) was

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

777 calculated as unweighted UniFrac distances using OTU tables rarefied at 100,000 seqs/sample
778 and phylogenetic distance between representative sequences from each OTU [84]. The effect of
779 covariates such as age, diet, location (LOC1 and LOC2) and gender were compared for correlation
780 with principal components identified from principal component analysis using UniFrac distances.
781 The polyserial correlations with P-values were calculated for categorical variables and the
782 significance of the covariates for explaining the variation was estimated at each principal
783 component.

784 **Network Analysis**

785 Spearman's rank correlations were computed between each of the species/MGS and the between
786 MGS and functional modules/metabolites. The correlations with significant P-values were selected
787 and were used for the network analysis. The undirected links were generated between correlated
788 nodes (species/KOs/modules) and the strength of the links were given weights based on their
789 correlation coefficients. The network structure was generated using "igraph" package in R. The
790 modularity of the network for KOs association was generated with each module representing the
791 functional modules defined in KEGG database. The negative correlation was not considered in
792 generating the network modules. Moreover, the positive correlations were filtered ($\rho \geq 0.6$) for
793 most of the network analysis.

794 **Supervised learning**

795 Predictive models were built using supervised machine learning algorithm Random Forest
796 (RF)[85]. The models were optimized using 10,000 trees and default settings of mtry (number for
797 variables used to build the model). The mean three-fold cross-validation error rates were calculated
798 for each of the binary tree and the ensemble of trees. The mean decrease in accuracy, which is the
799 increase in error rates on leaving the variable out, was calculated for each prediction and tree and

1
2
3
4 800 was used to estimate the importance score. The variables showing a higher mean decrease in
5
6 801 accuracy of prediction were considered important for the segregation of the datasets into groups
7
8
9 802 based on the categorical variable.

10 11 12 803 **Statistical Analysis**

13
14
15 804 All the statistical comparisons between groups were performed using Negative Binomial model-
16
17 805 based Wald test implemented in DESeq2 and non-parametric Wilcoxon Rank-Sum Test with FDR
18
19
20 806 Adjusted P-Values to control for multiple comparisons [86-88]. The correlations between two
21
22 807 variables and the correlations within were calculated using Spearman's Correlation Coefficient
23
24
25 808 with Adjusted P-Values [89]. The correlations between categorical and numeric variables were
26
27 809 performed using Polyserial correlation/biserial correlations [90]. To identify the enrichment of
28
29
30 810 enzymes/species associated with a host, Odds Ratio was used as a measure of the enrichment of a
31
32 811 feature in a group. The Odds Ratio was calculated as $OR(k) = \frac{[\sum_{s=LOC1} A_{sk} / \sum_{s=LOC1} (\sum_{i \neq k} A_{si})]}{[\sum_{s=LOC2} A_{sk} / \sum_{s=LOC2} (\sum_{i \neq k} A_{si})]}$ for enrichment of genes/species between two locations, where
33
34 812 A_{sk} denotes abundance of species/gene k in sample S. Also the enrichment of species/genes
35
36
37 813 between Indian microbiome compared to other datasets consisting of USA, Denmark and China
38
39 814 referred as "OTHERS" were computed as $OR(k) = \frac{([\sum_{s=INDIA} A_{sk} / \sum_{s=INDIA} (\sum_{i \neq k} A_{si})]}{([\sum_{s=OTHERS} A_{sk} / \sum_{s=OTHERS} (\sum_{i \neq k} A_{si})]}$. All the graphs and plots were generated using the ggplot2 package in R.
40
41
42 815
43
44 816
45
46

47 817 **Correlation analysis between functional modules and metabolite clusters**

48
49
50 818 To calculate the association of microbial functional modules with faecal metabolite clusters, the
51
52 819 Spearman's correlation coefficients were calculated to rank KOs for association with metabolite
53
54
55 820 clusters and Metabotypes. To quantify the shift in Spearman correlation between given KEGG
56
57 821 module and the metabolite cluster compared to the background distribution, the background
58
59
60 822 adjusted median Spearman's correlation was calculated for a given KEGG module m as:

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

823 $SCC_{bg,adj} = \text{median}(SCC_{KOs \in \text{KEGG Module } m}) - \text{median}(SCC_{KOs \text{ KEGG Module } m})$

824 Where SCC_{KO} is the partial Spearman's correlation coefficient between KO and the metabolite
825 cluster.

826 Identification of microbial species driving the association between KEGG Module and metabolite
827 abundance was done by iterating the correlation between KO belonging to the KEGG module and
828 the metabolite after excluding the genes annotated to that KO from each species. The change in
829 median Spearman's correlation coefficient between the KOs and the metabolite, when genes from
830 that species are excluded from the analysis, was calculated as described previously [8]. The species
831 showing the maximum change in the overall correlation of module with metabotype was plotted.

832 **List of abbreviations**

833 Indian Gut Microbiome (IGM), Enterotypes (ET), Integrated Gene Catalog (IGC), Metagenome-
834 Wide Association Study (MGWAS), Short Chain Fatty Acids (SCFAs), Branched Chain Amino
835 Acids (BCAAs).

836 **Declarations**

837 **Collection of Datasets for Comparative analysis**

838 The 74 HMP metagenomes were collected from <http://hmpdacc.org/HMASM> or NCBI SRA
839 (accession SRR059347). The 85 Danish fecal metagenomes from METAHIT were obtained from
840 European Nucleotide Archive (<http://www.ebi.ac.uk/ena>, study accession number ERP000108).
841 The 71 Chinese metagenome samples were obtained from NCBI SRA (accession number –
842 SRR341581).

843 **Ethics approval and consent to participate**

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

844 The recruitment of volunteers, sample collection, and other study-related procedures were carried
845 out by following the guidelines and protocols approved by the Institute Ethics Committee of Indian
846 Institute of Science Education and Research (IISER), Bhopal, India. Written informed consent was
847 obtained from all subjects prior to any study-related procedures.

848 Consent for publication

849 Not applicable

850 Availability of data

851 The datasets generated and/or analysed during the current study have been deposited in the
852 National Centre for Biotechnology Information (NCBI) BioProject database under the project
853 number PRJNA397112. Metabolomics data are available via the MetaboLights database
854 (accession number MTBLS803). Supporting data are also available via the *GigaScience*
855 repository, GigaDB [91].

856 Competing interests

857 The authors declare that they have no competing interests.

858 Funding

859 This work was supported by the intramural funding received from IISER Bhopal, Madhya Pradesh,
860 India.

861 Author's contributions

862 VKS and AM conceived the work and participated in the design of the study. AM and JP collected
863 all the samples in collaboration with TG. AM designed the study protocols and performed sample
864 processing, DNA extraction, metabolite extraction and profiling from faecal and blood samples.
865 RS and AM carried out the library preparation and sequencing work. DBD carried out all

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

866 metagenomic data and statistical analysis. AKS and DBD analyzed the metabolomics data. AM
867 and DBD did the primary data interpretation of analytical outcomes under the supervision of VKS.
868 AM, DBD, AKS, RS, AG, JS, KRA and VKS drafted the manuscript. All authors read and
869 approved the final manuscript.

870 Acknowledgments

871 The sequencing and computational analysis were performed at the NGS Facility and HPC and
872 computing facility, respectively, at IISER Bhopal. DBD, AM, RS and JP received fellowships
873 from the UGC (University Grants Commission), Centre for Research on Environment and
874 Sustainable Technologies (CREST, IISER Bhopal), DST-INSPIRE and Central University of
875 Kerala, respectively.

876

1
2
3
4 877 **References:**
5
6

- 7 878 1. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R and Gordon JI. The human
8 879 microbiome project. *Nature*. 2007;449 7164:804-10.
9 880 2. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, et al. Enterotypes of the
10 881 human gut microbiome. *Nature*. 2011;473 7346:174-80.
11 882 3. Yatsunencko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, et al. Human gut
12 883 microbiome viewed across age and geography. *Nature*. 2012;486 7402:222-7.
13 884 4. Dominguez-Bello MG, Costello EK, Contreras M, Magris M, Hidalgo G, Fierer N, et al. Delivery
14 885 mode shapes the acquisition and structure of the initial microbiota across multiple body habitats
15 886 in newborns. *Proc Natl Acad Sci U S A*. 2010;107 26:11971-5.
16 887 5. Saxena R and Sharma V. A metagenomic insight into the human microbiome: Its implications in
17 888 health and disease. *Medical and Health Genomics*. Elsevier; 2015. p. 107-19.
18 889 6. Schwartz A, Taras D, Schafer K, Beijer S, Bos NA, Donus C, et al. Microbiota and SCFA in lean and
19 890 overweight healthy subjects. *Obesity (Silver Spring)*. 2010;18 1:190-5.
20 891 7. Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, et al. A metagenome-wide association study of gut microbiota
21 892 in type 2 diabetes. *Nature*. 2012;490 7418:55-60.
22 893 8. Pedersen HK, Gudmundsdottir V, Nielsen HB, Hyotylainen T, Nielsen T, Jensen BA, et al. Human
23 894 gut microbes impact host serum metabolome and insulin sensitivity. *Nature*. 2016;535 7612:376-
24 895 81.
25 896 9. Gupta VK, Paul S and Dutta C. Geography, Ethnicity or Subsistence-Specific Variations in Human
26 897 Microbiome Composition and Diversity. *Front Microbiol*. 2017;8:1162.
27 898 10. Human Microbiome Project C. Structure, function and diversity of the healthy human
28 899 microbiome. *Nature*. 2012;486 7402:207-14.
29 900 11. Gomez A, Petrzelkova KJ, Burns MB, Yeoman CJ, Amato KR, Vlckova K, et al. Gut Microbiome of
30 901 Coexisting BaAka Pygmies and Bantu Reflects Gradients of Traditional Subsistence Patterns. *Cell*
31 902 *Rep*. 2016;14 9:2142-53.
32 903 12. Obregon-Tito AJ, Tito RY, Metcalf J, Sankaranarayanan K, Clemente JC, Ursell LK, et al. Subsistence
33 904 strategies in traditional societies distinguish gut microbiomes. *Nat Commun*. 2015;6:6505.
34 905 13. Schnorr SL, Candela M, Rampelli S, Centanni M, Consolandi C, Basaglia G, et al. Gut microbiome
35 906 of the Hadza hunter-gatherers. *Nat Commun*. 2014;5:3654.
36 907 14. Hasler R, Kautz C, Rehman A, Podschun R, Gassling V, Brzoska P, et al. The antibiotic resistome
37 908 and microbiota landscape of refugees from Syria, Iraq and Afghanistan in Germany. *Microbiome*.
38 909 2018;6 1:37.
39 910 15. Mohan V, Sandeep S, Deepa R, Shah B and Varghese C. Epidemiology of type 2 diabetes: Indian
40 911 scenario. *Indian J Med Res*. 2007;125 3:217-30.
41 912 16. Mushtaq MU, Gull S, Abdullah HM, Shahid U, Shad MA and Akram J. Waist circumference, waist-
42 913 hip ratio and waist-height ratio percentiles and central obesity among Pakistani children aged five
43 914 to twelve years. *BMC Pediatr*. 2011;11:105.
44 915 17. Maji A, Misra R, Dhakan DB, Gupta V, Mahato NK, Saxena R, et al. Gut microbiome contributes to
45 916 impairment of immunity in pulmonary tuberculosis patients by alteration of butyrate and
46 917 propionate producers. *Environ Microbiol*. 2018;20 1:402-19.
47 918 18. Pulikkan J, Maji A, Dhakan DB, Saxena R, Mohan B, Anto MM, et al. Gut Microbial Dysbiosis in
48 919 Indian Children with Autism Spectrum Disorders. *Microb Ecol*. 2018;76 4:1102-14.
49 920 19. Bhute S, Pande P, Shetty SA, Shelar R, Mane S, Kumbhare SV, et al. Molecular Characterization
50 921 and Meta-Analysis of Gut Microbial Communities Illustrate Enrichment of *Prevotella* and
51 922 *Megasphaera* in Indian Subjects. *Front Microbiol*. 2016;7:660.
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

923 20. Shetty SA, Marathe NP and Shouche YS. Opportunities and challenges for gut microbiome studies in the Indian population. *Microbiome*. 2013;1 1:24.

924

925 21. Tandon D, Haque MM, R S, Shaikh S, P S, Dubey AK, et al. A snapshot of gut microbiota of an adult urban population from Western region of India. *PLoS One*. 2018;13 4:e0195643.

926

927 22. Suryanarayana M, Agrawal A and Prabhu KS. Inequality-adjusted human development index for India's states. United Nations Development Programme (UNDP) India. http://www.undp.org/content/dam/india/docs/inequality_adjusted_human_development_index_for_indias_state_1.pdf [NS], 2011.

928

929

930

931 23. Misra A, Pandey RM, Devi JR, Sharma R, Vikram NK and Khanna N. High prevalence of diabetes, obesity and dyslipidaemia in urban slum population in northern India. *Int J Obes Relat Metab Disord*. 2001;25 11:1722-9.

932

933

934 24. Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, et al. An integrated catalog of reference genes in the human gut microbiome. *Nat Biotechnol*. 2014;32 8:834-41.

935

936 25. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*. 2010;464 7285:59-65.

937

938 26. Wang J and Jia H. Metagenome-wide association studies: fine-mining the microbiome. *Nat Rev Microbiol*. 2016;14 8:508-22.

939

940 27. Tyakht AV, Kostyukova ES, Popenko AS, Belenikin MS, Pavlenko AV, Larin AK, et al. Human gut microbiota community structures in urban and rural populations in Russia. *Nat Commun*. 2013;4:2469.

941

942

943 28. Liang C, Tseng HC, Chen HM, Wang WC, Chiu CM, Chang JY, et al. Diversity and enterotype in gut bacterial community of adults in Taiwan. *BMC Genomics*. 2017;18 Suppl 1:932.

944

945 29. Aleksandrowicz L, Tak M, Green R, Kinra S and Haines A. Comparison of food consumption in Indian adults between national and sub-national dietary data sources. *Br J Nutr*. 2017;117 7:1013-9.

946

947

948 30. Joy EJ, Green R, Agrawal S, Aleksandrowicz L, Bowen L, Kinra S, et al. Dietary patterns and non-communicable disease risk in Indian adults: secondary analysis of Indian Migration Study data. *Public Health Nutr*. 2017;20 11:1963-72.

949

950

951 31. Rios-Covian D, Ruas-Madiedo P, Margolles A, Gueimonde M, de Los Reyes-Gavilan CG and Salazar N. Intestinal Short Chain Fatty Acids and their Link with Diet and Human Health. *Front Microbiol*. 2016;7:185.

952

953

954 32. Kovatcheva-Datchary P, Nilsson A, Akrami R, Lee YS, De Vadder F, Arora T, et al. Dietary Fiber-Induced Improvement in Glucose Metabolism Is Associated with Increased Abundance of *Prevotella*. *Cell Metab*. 2015;22 6:971-82.

955

956

957 33. De Filippo C, Cavalieri D, Di Paola M, Ramazzotti M, Poullet JB, Massart S, et al. Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proc Natl Acad Sci U S A*. 2010;107 33:14691-6.

958

959

960 34. Ruiz-Capillas C and Jimenez-Colmenero F. Biogenic amines in meat and meat products. *Crit Rev Food Sci Nutr*. 2004;44 7-8:489-99.

961

962 35. Layman DK. The role of leucine in weight loss diets and glucose homeostasis. *J Nutr*. 2003;133 1:261S-7S.

963

964 36. Drevland RM, Waheed A and Graham DE. Enzymology and evolution of the pyruvate pathway to 2-oxobutyrate in *Methanocaldococcus jannaschii*. *J Bacteriol*. 2007;189 12:4391-400.

965

966 37. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen YY, Keilbaugh SA, et al. Linking long-term dietary patterns with gut microbial enterotypes. *Science*. 2011;334 6052:105-8.

967

968 38. Cho I and Blaser MJ. The human microbiome: at the interface of health and disease. *Nat Rev Genet*. 2012;13 4:260-70.

969

- 1
2
3
4 970 39. Green R, Milner J, Joy EJ, Agrawal S and Dangour AD. Dietary patterns in India: a systematic review.
5 971 Br J Nutr. 2016;116 1:142-8.
6 972 40. Martinez I, Stegen JC, Maldonado-Gomez MX, Eren AM, Siba PM, Greenhill AR, et al. The gut
7 973 microbiota of rural papua new guineans: composition, diversity patterns, and ecological
8 974 processes. Cell Rep. 2015;11 4:527-38.
9 975 41. Ley RE. Gut microbiota in 2015: Prevotella in the gut: choose carefully. Nat Rev Gastroenterol
10 976 Hepatol. 2016;13 2:69-70.
11 977 42. Losasso C, Eckert EM, Mastroianni E, Villiger J, Mancin M, Patuzzi I, et al. Assessing the Influence of
12 978 Vegan, Vegetarian and Omnivore Oriented Westernized Dietary Styles on Human Gut Microbiota:
13 979 A Cross Sectional Study. Front Microbiol. 2018;9:317.
14 980 43. Larsen JM. The immune response to Prevotella bacteria in chronic inflammatory disease.
15 981 Immunology. 2017;151 4:363-74.
16 982 44. Scher JU, Szczesnak A, Longman RS, Segata N, Ubeda C, Bielski C, et al. Expansion of intestinal
17 983 Prevotella copri correlates with enhanced susceptibility to arthritis. Elife. 2013;2:e01202.
18 984 45. Renz H, von Mutius E, Brandtzaeg P, Cookson WO, Autenrieth IB and Haller D. Gene-environment
19 985 interactions in chronic inflammatory disease. Nat Immunol. 2011;12 4:273-7.
20 986 46. Tremaroli V and Backhed F. Functional interactions between the gut microbiota and host
21 987 metabolism. Nature. 2012;489 7415:242-9.
22 988 47. Selhub EM, Logan AC and Bested AC. Fermented foods, microbiota, and mental health: ancient
23 989 practice meets nutritional psychiatry. J Physiol Anthropol. 2014;33:2.
24 990 48. Costea PI, Hildebrand F, Arumugam M, Backhed F, Blaser MJ, Bushman FD, et al. Enterotypes in
25 991 the landscape of gut microbial community composition. Nat Microbiol. 2018;3 1:8-16.
26 992 49. Jaarin K, Norliana M, Kamisah Y, Nursyafiza M and Qodriyah HMSJECC. Potential role of virgin
27 993 coconut oil in reducing cardiovascular risk factors. 2014;20 8:3399-410.
28 994 50. Boemeke L, Marcadenti A, Busnello FM, Gottschall CBAJOJoE and Diseases M. Effects of coconut
29 995 oil on human health. 2015;5 07:84.
30 996 51. Intahphuak S, Khonsung P and Panthong AJPb. Anti-inflammatory, analgesic, and antipyretic
31 997 activities of virgin coconut oil. 2010;48 2:151-7.
32 998 52. Nicholson JK, Holmes E, Kinross J, Burcelin R, Gibson G, Jia W, et al. Host-gut microbiota metabolic
33 999 interactions. Science. 2012;336 6086:1262-7.
34 1000 53. Boulange CL, Neves AL, Chilloux J, Nicholson JK and Dumas ME. Impact of the gut microbiota on
35 1001 inflammation, obesity, and metabolic disease. Genome Med. 2016;8 1:42.
36 1002 54. Neis EP, Dejong CH and Rensen SS. The role of microbial amino acid metabolism in host
37 1003 metabolism. Nutrients. 2015;7 4:2930-46.
38 1004 55. Li X, Shimizu Y and Kimura I. Gut microbial metabolite short-chain fatty acids and obesity. Biosci
39 1005 Microbiota Food Health. 2017;36 4:135-40.
40 1006 56. Longvah T, Ananta I, Bhaskarachary K and Venkaiah K. Indian food composition tables. National
41 1007 Institute of Nutrition, Indian Council of Medical Research; 2017.
42 1008 57. Patel RK and Jain M. NGS QC Toolkit: a toolkit for quality control of next generation sequencing
43 1009 data. PloS one. 2012;7 2:e30619.
44 1010 58. Magoc T and Salzberg SL. FLASH: fast length adjustment of short reads to improve genome
45 1011 assemblies. Bioinformatics. 2011;27 21:2957-63.
46 1012 59. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene
47 1013 database project: improved data processing and web-based tools. Nucleic Acids Res. 2013;41
48 1014 Database issue:D590-6.
49 1015 60. Kuczynski J, Stombaugh J, Walters WA, Gonzalez A, Caporaso JG and Knight R. Using QIIME to
50 1016 analyze 16S rRNA gene sequences from microbial communities. Curr Protoc Bioinformatics.
51 1017 2011;Chapter 10:Unit 10 7.

61
62
63
64
65

1
2
3
4 1018 61. Caporaso JG, Bittinger K, Bushman FD, DeSantis TZ, Andersen GL and Knight R. PyNAST: a flexible
5 1019 tool for aligning sequences to a template alignment. *Bioinformatics*. 2010;26 2:266-7.
6
7 1020 62. Lozupone C, Lladser ME, Knights D, Stombaugh J and Knight R. UniFrac: an effective distance
8 1021 metric for microbial community comparison. *ISME J*. 2011;5 2:169-72.
9 1022 63. Peng Y, Leung HC, Yiu SM and Chin FY. IDBA-UD: a de novo assembler for single-cell and
10 1023 metagenomic sequencing data with highly uneven depth. *Bioinformatics*. 2012;28 11:1420-8.
11 1024 64. Niu B, Zhu Z, Fu L, Wu S and Li W. FR-HIT, a very fast program to recruit metagenomic reads to
12 1025 homologous reference genomes. *Bioinformatics*. 2011;27 12:1704-5.
13 1026 65. Zhu W, Lomsadze A and Borodovsky M. Ab initio gene identification in metagenomic sequences.
14 1027 *Nucleic acids research*. 2010;38 12:e132-e.
15 1028 66. Kent WJ. BLAT--the BLAST-like alignment tool. *Genome Res*. 2002;12 4:656-64.
16 1029 67. Li R, Yu C, Li Y, Lam T-W, Yiu S-M, Kristiansen K, et al. SOAP2: an improved ultrafast tool for short
17 1030 read alignment. *Bioinformatics*. 2009;25 15:1966-7.
18 1031 68. Langmead B and Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nature methods*. 2012;9
19 1032 4:357.
20 1033 69. Zhang R and Lin Y. DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes.
21 1034 *Nucleic acids research*. 2008;37 suppl_1:D455-D8.
22 1035 70. Buchfink B, Xie C and Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nature*
23 1036 *methods*. 2014;12 1:59.
24 1037 71. Powell S, Forslund K, Szklarczyk D, Trachana K, Roth A, Huerta-Cepas J, et al. eggNOG v4. 0: nested
25 1038 orthology inference across 3686 organisms. *Nucleic acids research*. 2014;42 D1:D231-D9.
26 1039 72. Lim T-S and Loh W-Y. A comparison of tests of equality of variances. *Computational Statistics &*
27 1040 *Data Analysis*. 1996;22 3:287-301.
28 1041 73. Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, et al. Identification and
29 1042 assembly of genomes and genetic elements in complex metagenomic samples without using
30 1043 reference genomes. *Nat Biotechnol*. 2014;32 8:822-8.
31 1044 74. Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ. Basic local alignment search tool. *Journal*
32 1045 *of molecular biology*. 1990;215 3:403-10.
33 1046 75. Suzek BE, Huang H, McGarvey P, Mazumder R and Wu CH. UniRef: comprehensive and non-
34 1047 redundant UniProt reference clusters. *Bioinformatics*. 2007;23 10:1282-8.
35 1048 76. Huson DH, Auch AF, Qi J and Schuster SC. MEGAN analysis of metagenomic data. *Genome*
36 1049 *research*. 2007;17 3:377-86.
37 1050 77. Phua LC, Koh PK, Cheah PY, Ho HK and Chan ECY. Global gas chromatography/time-of-flight mass
38 1051 spectrometry (GC/TOFMS)-based metabonomic profiling of lyophilized human feces. *Journal of*
39 1052 *Chromatography B*. 2013;937:103-13.
40 1053 78. Psychogios N, Hau DD, Peng J, Guo AC, Mandal R, Bouatra S, et al. The human serum metabolome.
41 1054 *PloS one*. 2011;6 2:e16957.
42 1055 79. Gao X, Pujos-Guillot E, Martin J-F, Galan P, Juste C, Jia W, et al. Metabolite analysis of human fecal
43 1056 water by gas chromatography/mass spectrometry with ethyl chloroformate derivatization.
44 1057 *Analytical biochemistry*. 2009;393 2:163-75.
45 1058 80. Worley B and Powers R. Multivariate Analysis in Metabolomics. *Curr Metabolomics*. 2013;1 1:92-
46 1059 107.
47 1060 81. Langfelder P and Horvath S. WGCNA: an R package for weighted correlation network analysis.
48 1061 *BMC Bioinformatics*. 2008;9:559.
49 1062 82. Kaufman L and Rousseeuw PJ. Partitioning around medoids (program pam). Finding groups in
50 1063 data: an introduction to cluster analysis. 1990:68-125.
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

1064 83. Liao M, Li Y, Kianifard F, Obi E and Arcona S. Cluster analysis and its application to healthcare
1065 claims data: a study of end-stage renal disease patients who initiated hemodialysis. *BMC Nephrol.*
1066 2016;17:25.

1067 84. McMurdie PJ and Holmes S. Waste not, want not: why rarefying microbiome data is inadmissible.
1068 *PLoS Comput Biol.* 2014;10 4:e1003531.

1069 85. Liaw A and Wiener MJRn. Classification and regression by randomForest. 2002;2 3:18-22.

1070 86. Wilcoxon F. Individual comparisons of grouped data by ranking methods. *Journal of economic*
1071 *entomology.* 1946;39 2:269-70.

1072 87. Benjamini Y, Drai D, Elmer G, Kafkafi N and Golani I. Controlling the false discovery rate in behavior
1073 genetics research. *Behavioural brain research.* 2001;125 1-2:279-84.

1074 88. Love M, Anders S and Huber W. Differential analysis of count data—the DESeq2 package. *Genome*
1075 *Biol.* 2014;15 550:10.1186.

1076 89. Kendall MG. Rank correlation methods. 1955.

1077 90. Olsson U, Drasgow F and Dorans NJ. The polyserial correlation coefficient. *Psychometrika.*
1078 1982;47 3:337-47.

1079 91. Dhakan DB, Maji A, Sharma AK, Saxena R, Pulikkan J, Grace T, et al. Supporting data for "The
1080 unique composition of Indian gut microbiome, gene catalogue and associated faecal metabolome
1081 deciphered using multi-omics approaches" *GigaScience Database* 2019.
1082 <http://dx.doi.org/10.5524/100548>

1083

Table 1. Metagenomic datasets used for comparative analysis (Meta-analysis) of the microbiome and MGWAS

Dataset	No. of samples	Sequence data (GB)	No. of genes
INDIA	110	110	4,809,378
USA	74	441	6,521,885
DENMARK	85	103.87	7,141,214
CHINA	71	180.78	5,464,702

Table2. PERMANOVA to assess the effect of Covariates on metabolomics profiles of samples

Variable	Sum of Sq	Mean Sq	F-Model	R ²	P-value
Location	0.05841	0.058406	4.9423	0.04455	0.0009
Diet	0.04701	0.04701	4.2132	0.03586	0.0009
Age	0.01618	0.01618	1.4505	0.0123	0.161
Gender	0.00488	0.00488	0.4370	0.00373	0.927

Table3. OPLS-DA model and its validation for different covariates as class of separation

Variable	R ² X	Q ² (cumulative)	pR ²	pQ ²
Location	0.165	0.205	0.005***	0.005***
Diet	0.168	0.123	0.005***	0.005***
Age	0.155	-0.00067	0.075	0.065
Gender	0.106	-0.247	0.145	0.96
Cluster (Genus based)	0.16	0.15	0.005***	0.005***

pR² and pQ² show p-values for validation of OPLS-DA model with p value < 0.01 shown as significant (*)

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

Figure title and legends

Fig. 1. Comparison of Indian gut microbiome with other major populations using 16S rRNA gene and metagenomic datasets. (A) Percentage of total reads that could be mapped to IGC and updated IGC containing Indian gene catalogue. Plotted are interquartile ranges (IQR in boxes), median (as dark lines in the boxes), lowest and highest values within 1.5 times the IQR (shown as whiskers extending from boxes) and outliers as points beyond these whiskers. The blue and red boxes showed percentage of reads mapped to IGC and updated IGC (containing the Indian microbial genes). (B) Principal Component Analysis using MGS/CAG proportion derived from MGWAS. The samples are plotted along with the MGS/CAGs having taxonomic annotations. The MGS/CAGs are coloured according to their phylum. Variations across populations are shown using PC1 and PC2 along with factor loadings of major MGS/CAGs as biplots. (C) Illustration of proportions of bacterial families in different populations and their composition as determined from 16S rRNA gene datasets (adult population only). The mean family compositions of abundant families ($\geq 1\%$) are represented in separate pie plots from 10 different country-wise datasets, showing their overall microbial composition compared to Indian population.

Fig. 2. Functional variations and differences between Indian populations and other populations determined from core & accessory microbial functions. (A) Procrustes analysis was performed on Bray Curtis distances calculated from core EggNOG and accessory EggNOG abundance tables in all populations. PCA analysis shows the concordance of core and accessory functions in India, Denmark, USA and China populations. The red and black lines are associated with core and accessory datasets, respectively. (B) Eigenvalues calculated from PCA of samples using core EggNOGs and accessory EggNOGs are plotted. The boxplots showing for core and accessory eigenvalues for all samples in different populations are shown. Each box plot represents

1
2
3
4 1115 the median shown as white line between the boxes, the upper and lower ends of the boxes
5
6 1116 representing upper quartile (75th percentile) and lower quartile (25th percentile). The whiskers
7
8
9 1117 extending on both the ends represent 2.5* IQR (Inter Quartile Range). The different coloured dots
10
11
12 1118 overlaid for each sample are plotted over the box. The enrichment or depletion of (C) Egnog,
13
14 1119 and (D) Kegg functions in India compared to other populations are shown as volcano plots. The
15
16 1120 log-transformed FDR Adj. P-values calculated from negative binomial-based Wald test from
17
18
19 1121 DESeq2 are plotted on the x-axis. The log odds ratio calculated for India vs Other datasets are
20
21 1122 plotted on the y-axis. The EggNOGs/KOs with P-value<0.05 are shown in Blue whereas those
22
23
24 1123 having P-values>0.05 are shown in Red. The EggNOGs/KOs extending on right and left side and
25
26 1124 with P-value>0.05 are labelled as highly enriched in India and other datasets, respectively.

27
28
29 1125 **Fig. 3. Variations in gut microbiome at the two locations. (A)** Between class Analysis, which
30
31
32 1126 visualizes results from PCA and clustering, using genus level abundance from 37 cross national
33
34 1127 dataset and genus abundance of 110 Indian samples obtained from mapping of reads to reference
35
36
37 1128 genomes. The samples from LOC1 (cyan), LOC2 (pink) and 37 cross national samples from
38
39 1129 Arumugam et al. (grey and labelled) are placed into three distinct enterotypes based on clustering.
40
41 1130 **(B)** Significantly different genera (FDR Adj. P-value < 0.05; NB model-based Wald test) between
42
43
44 1131 the two locations are shown as boxplots with boxes representing interquartile range (IQR), dark
45
46 1132 lines between the boxes representing median values and whiskers representing the 1.5 x IQR on
47
48
49 1133 each side. **(C)** Scatterplot of log-transformed mean values of species abundance in LOC1 (n=53)
50
51 1134 and LOC2 (n = 57) individuals. Red colour gradient points represent differentially abundant (FDR
52
53
54 1135 Adj. P< 0.05; NB model-based Wald test) species with lower p-values from Red to Blue.

55
56
57 1136 **Fig. 4. Between class analysis to identify metatypes and their associated metabolites. (A)**
58
59 1137 Metabolite clusters (MES) abundance profiles of samples were generated and their clustering was

60
61
62
63
64
65

1
2
3
4 1138 performed using PAM (partition around medoids) clustering. The between class and PCA of JSD
5
6
7 1139 distances and PAM clustering identified 3 clusters to be optimum for their segregation using (B)
8
9 1140 Silhouette index. The metabolites valeric acids, and saturated fatty acids such as palmitic acid and
10
11 1141 stearic acid, were found higher in Cluster1. The carbohydrates such as glucose and galactose were
12
13
14 1142 found higher in Cluster2. The branched chain amino acids, lauric acid and butyric acid were found
15
16 1143 higher in Cluster3. (C) OPLS-DA analysis using locations as classes shows locations as
17
18
19 1144 differentiating factors in separating the samples based on their metabolomic profiles.
20
21

22 1145 **Fig. 5. Spearman's Rank correlations of metabolites with species and metabolic modules. (A)**
23
24 1146 Spearman's Rank Correlation coefficients were calculated between significantly different
25
26
27 1147 metagenomic species and significantly different metabolites between LOC1 and LOC2
28
29 1148 populations. The correlations showing significant FDR Adj. P <0.05 are plotted. The bars on the
30
31
32 1149 right show the Log Odds Ratio of the abundance of MGS with positive values indicating
33
34 1150 enrichment in LOC1, and the negative values indicating enrichment in LOC2. (B) Spearman's
35
36
37 1151 Rank correlations between significantly different (FDR Adj. P<0.05, NB model-based Wald test)
38
39 1152 pathway modules and significantly different metabolite abundances in all samples. The significant
40
41
42 1153 (P<0.05) correlations are plotted and the colour intensities depict the correlation coefficients. The
43
44 1154 correlation of metabolites with locations is shown with labels in dark red colours showing
45
46 1155 association with LOC2, and the labels in green colours showing correlation with LOC1.
47
48

49 1156 **Fig. 6. BCAA abundance and their differential correlation with LOC1 and LOC2. (A)** Bar
50
51
52 1157 plot showing z-normalized values of serum and faecal BCAA (Valine and Isoleucine) relative
53
54 1158 concentration in LOC1 and LOC2. (B) The effect of specific microbial species on associations
55
56
57 1159 between BCAA biosynthesis pathways and BCAA levels in faecal metabolome, illustrated by
58
59 1160 change in background adjusted Spearman's correlation coefficient when a given species has been
60
61
62
63
64
65

1
2
3
4 1161 excluded from analysis is shown (see Methods). The density plot shows the distribution of
5
6
7 1162 correlation for species and the changes caused by specific species as marked by lines below. (C)
8
9 1163 Network analysis of Spearman's correlations between the branched chain amino acids
10
11
12 1164 biosynthesis, degradation and transport KEGG modules with MGS abundance in both LOC1 and
13
14 1165 LOC2 populations. The node size is proportional to the degree of interactions and the links between
15
16
17 1166 module and MGS show interactions or significant correlations (FDR Adj. $P < 0.05$) with negative
18
19 1167 (in Red) and positive (in Blue) correlation coefficients. (D) Plot showing relative abundance of
20
21 1168 KOs associated with different modules of BCAA biosynthesis and transporters in LOC1 and
22
23
24 1169 LOC2.

25
26
27 1170 **Fig. 7. BCAA transporters playing a key role in maintaining the levels of BCAAs in faeces**
28
29 1171 **and serum**

30
31
32
33 1172 The dynamics of BCAA concentration levels in faecal and serum metabolome influenced by
34
35 1173 microbial BCAA biosynthesis and transport pathways and their differential abundance in LOC1
36
37 1174 and LOC2 is shown

38
39
40
41 1175

42
43
44 1176 **Additional Files**

45
46
47 1177 **Additional File 1:** Supplementary data containing the metadata and sample information

48
49
50 1178 **Additional File 2:** Summary of sequencing statistics showing the number of reads per sample for
51
52 1179 16S rRNA gene amplicon dataset

53
54
55 1180 **Additional File 3:** Summary of sequencing statistics showing the number of reads per sample for
56
57
58 1181 Whole Genome Shotgun metagenomic dataset

59
60
61
62
63
64
65

1
2
3
4 1182 **Additional File 4:** Summary of the reads mapped to Integrated Gene Catalogue and Indian
5
6 1183 catalogue combined with IGC.
7
8
9
10 1184 **Additional File 5: Figures S1 to S18**
11
12
13 1185 **Additional File 6:** Differentially abundant MGS between India and other populations
14
15
16 1186 **Additional File 7:** Differentially abundant functions (Kegg Orthologues (KOs) and EggNOGs)
17
18 1187 between India and other populations.
19
20
21 1188 **Additional File 8:** Sample-wise representation of Indian samples into Enterotypes identified from
22
23 1189 Meta-analysis with 37 samples from four nations used in Arumugam et al.
24
25
26 1190 **Additional File 9:** Calinski Harabasz index and prediction strength calculated for clusters derived
27
28 1191 from 16S rRNA gene based genus abundance, metagenome based species abundance and
29
30 1192 metagenome based KO abundance profiles.
31
32
33 1193 **Additional File 10:** Mean relative abundance of genus in Cluster-1 and Cluster-2 and their
34
35 1194 associated P-values of difference calculated using NB model based Wald test.
36
37
38 1195 **Additional File 11:** The sample-wise association into clusters using Genus based and KO based
39
40 1196 clustering and their differences.
41
42
43 1197 **Additional File 12:** Differentially abundant KEGG orthologue functions between Cluster-1 and
44
45 1198 Cluster-2.
46
47
48
49
50
51 1199 **Additional File 13:** Polyserial correlation of covariates with principal components explaining
52
53 1200 variations across samples using unweighted UniFrac distances.
54
55
56
57 1201 **Additional File 14:** Differentially abundant MGS observed between two locations and their
58
59 1202 enrichment calculated using Log Odds ratio and NB model based P-values.
60
61
62
63
64
65

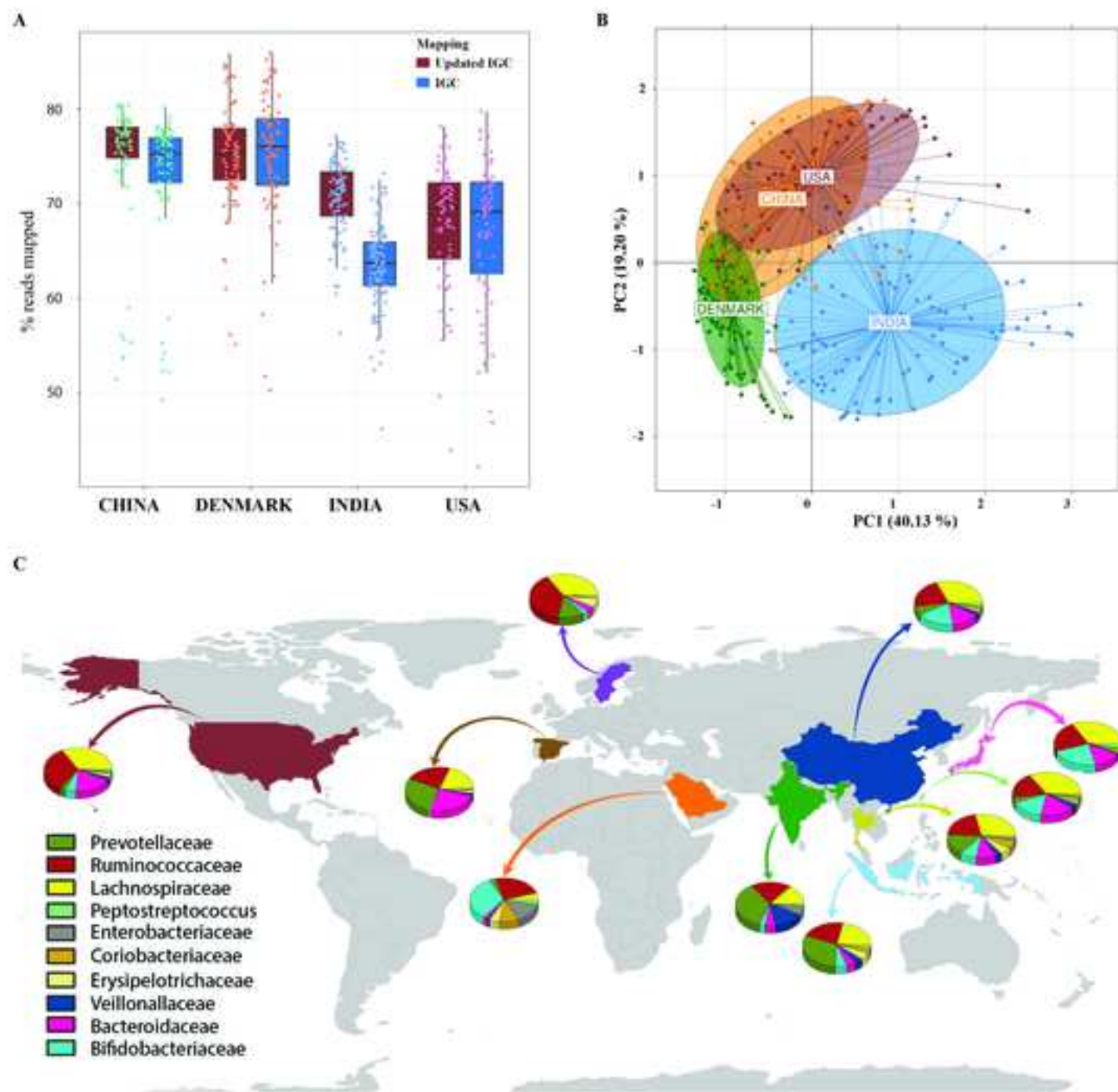
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65

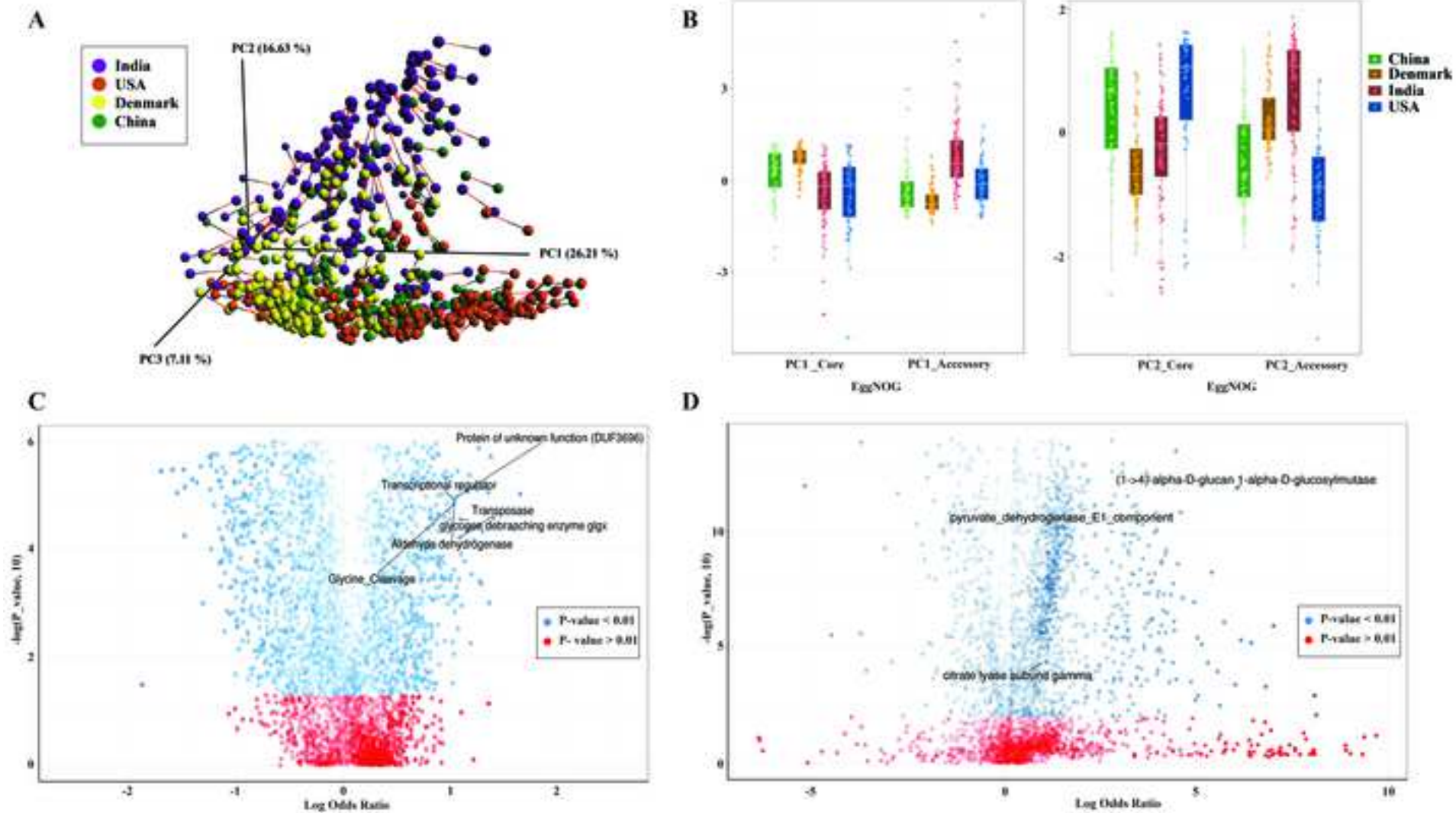
Additional File 15: Polyserial correlation of covariates with principal components explaining variations across samples using metabolomics data.

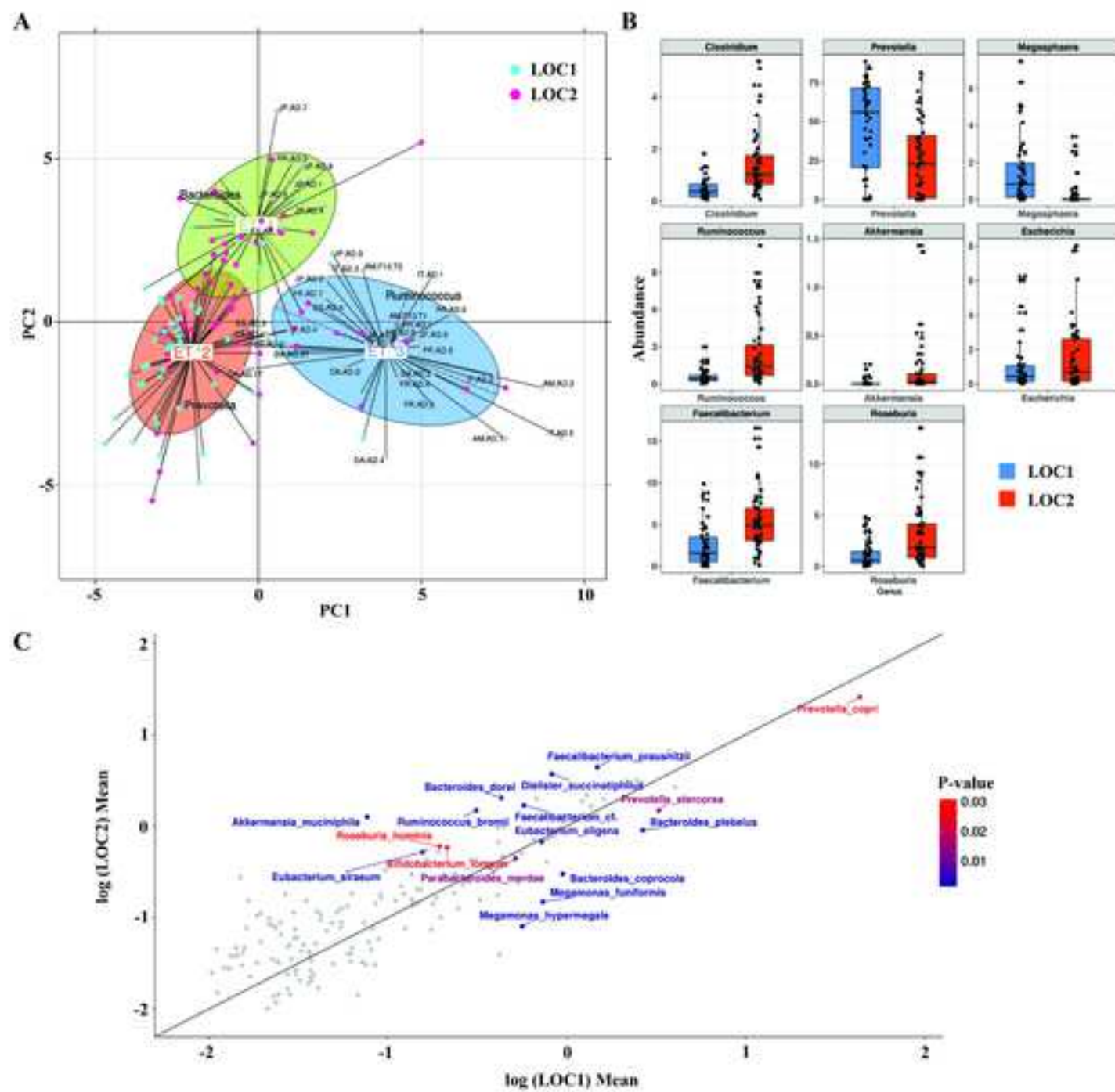
Additional File 16: Table shows the Spearman's rank correlation coefficient values of metabolites with Metabotypes.

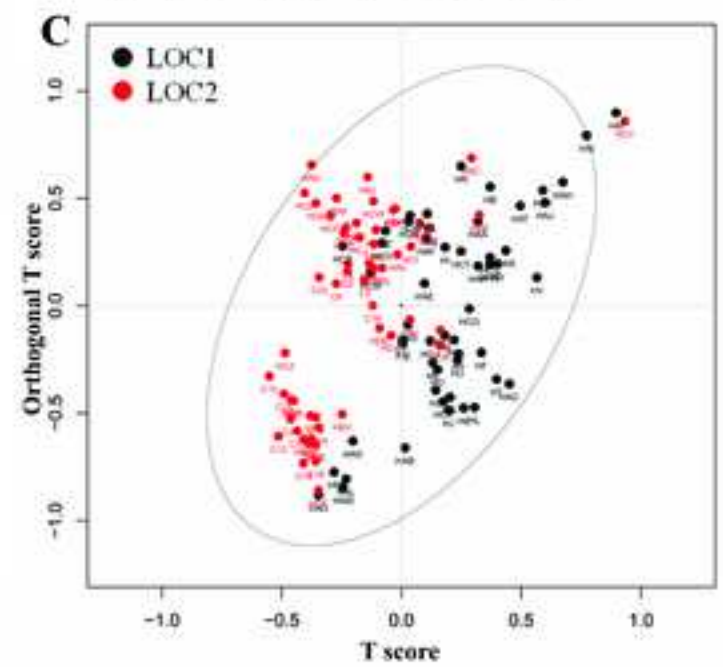
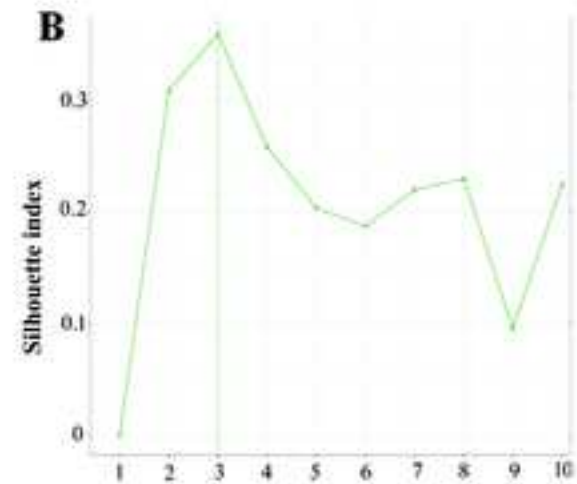
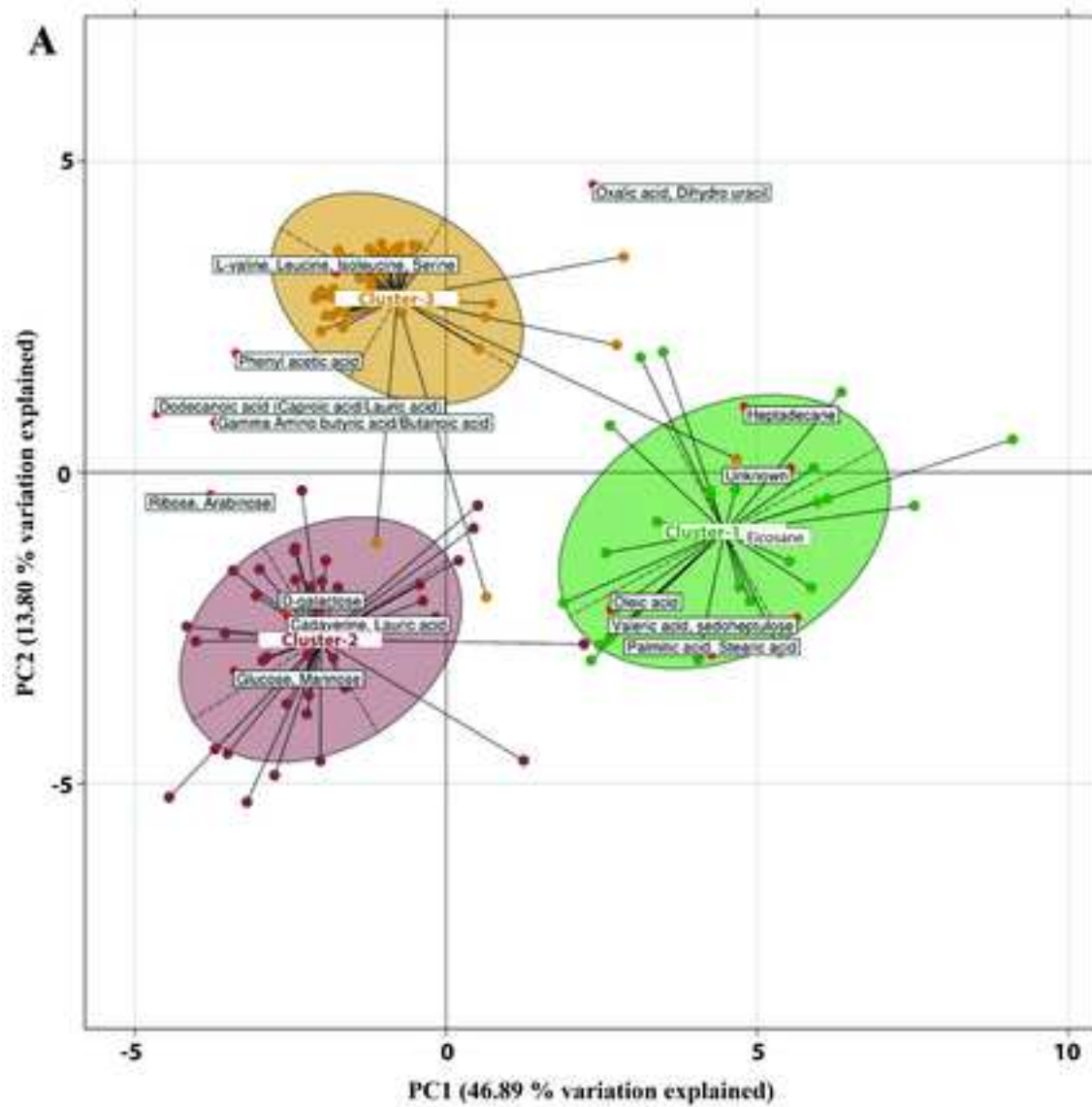
Additional File 17: Table shows the differential abundance of KEGG Modules between LOC1 and LOC2

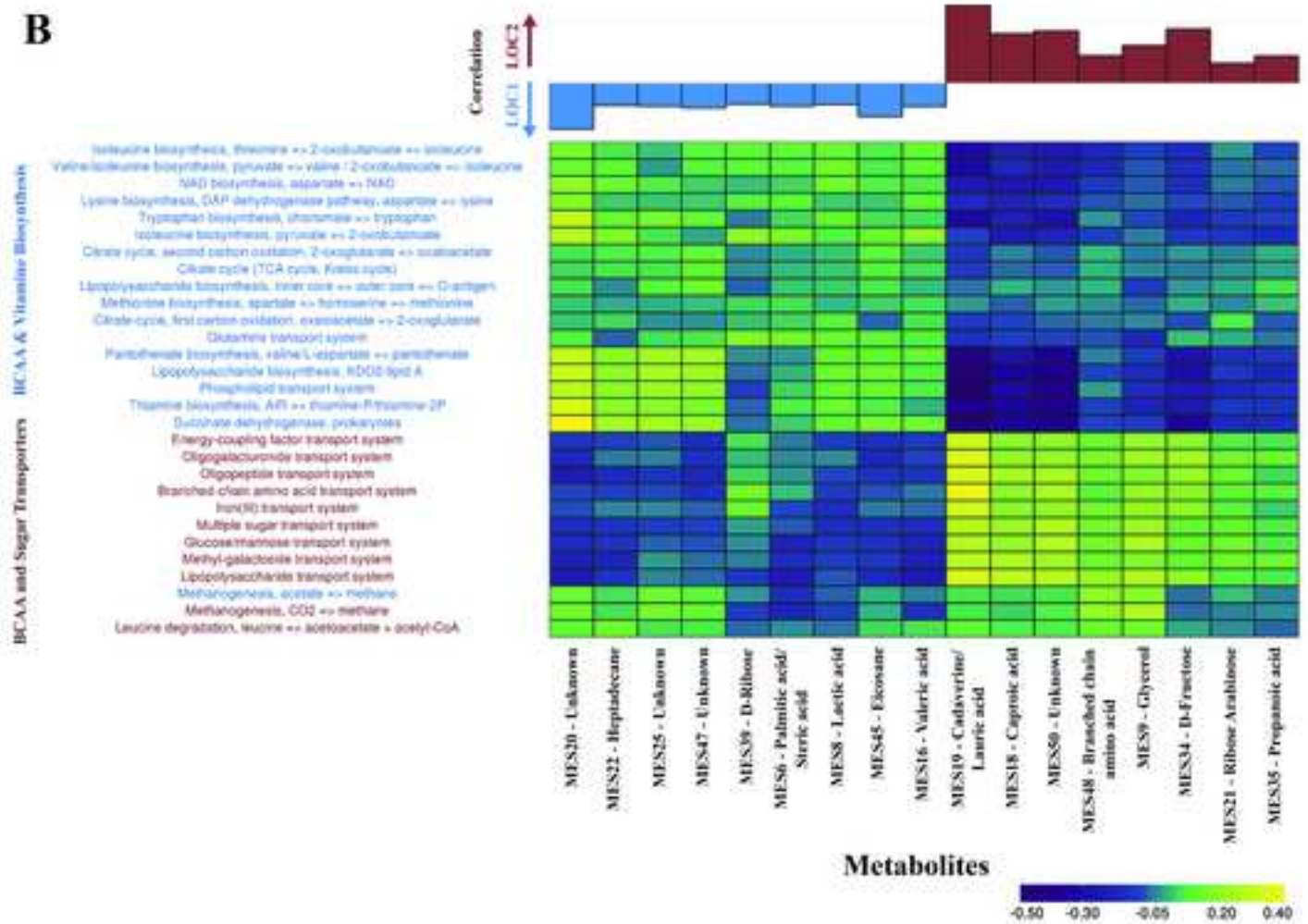
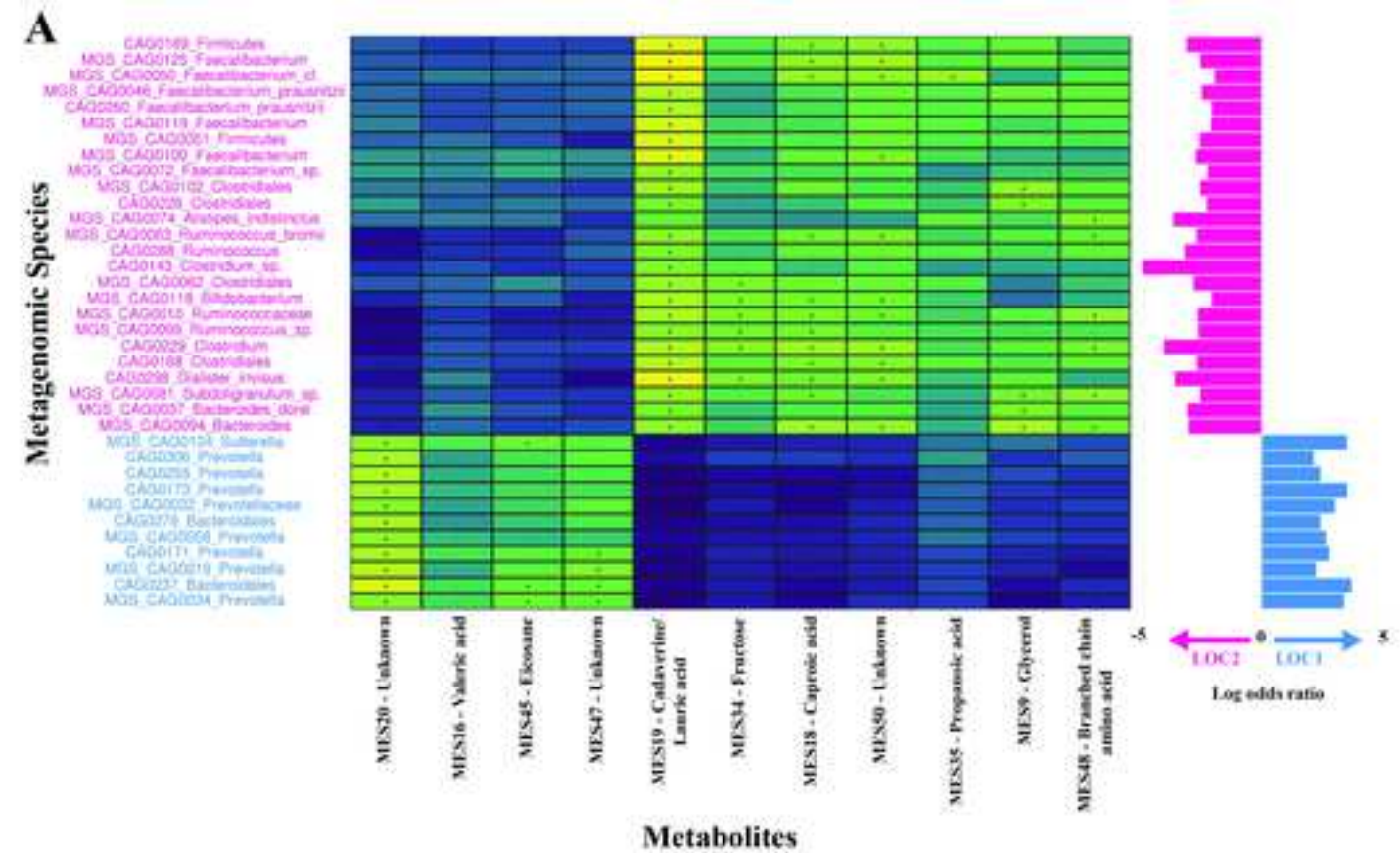
Additional File 18: List of reference genomes from NCBI and HMP databases for reference mapping

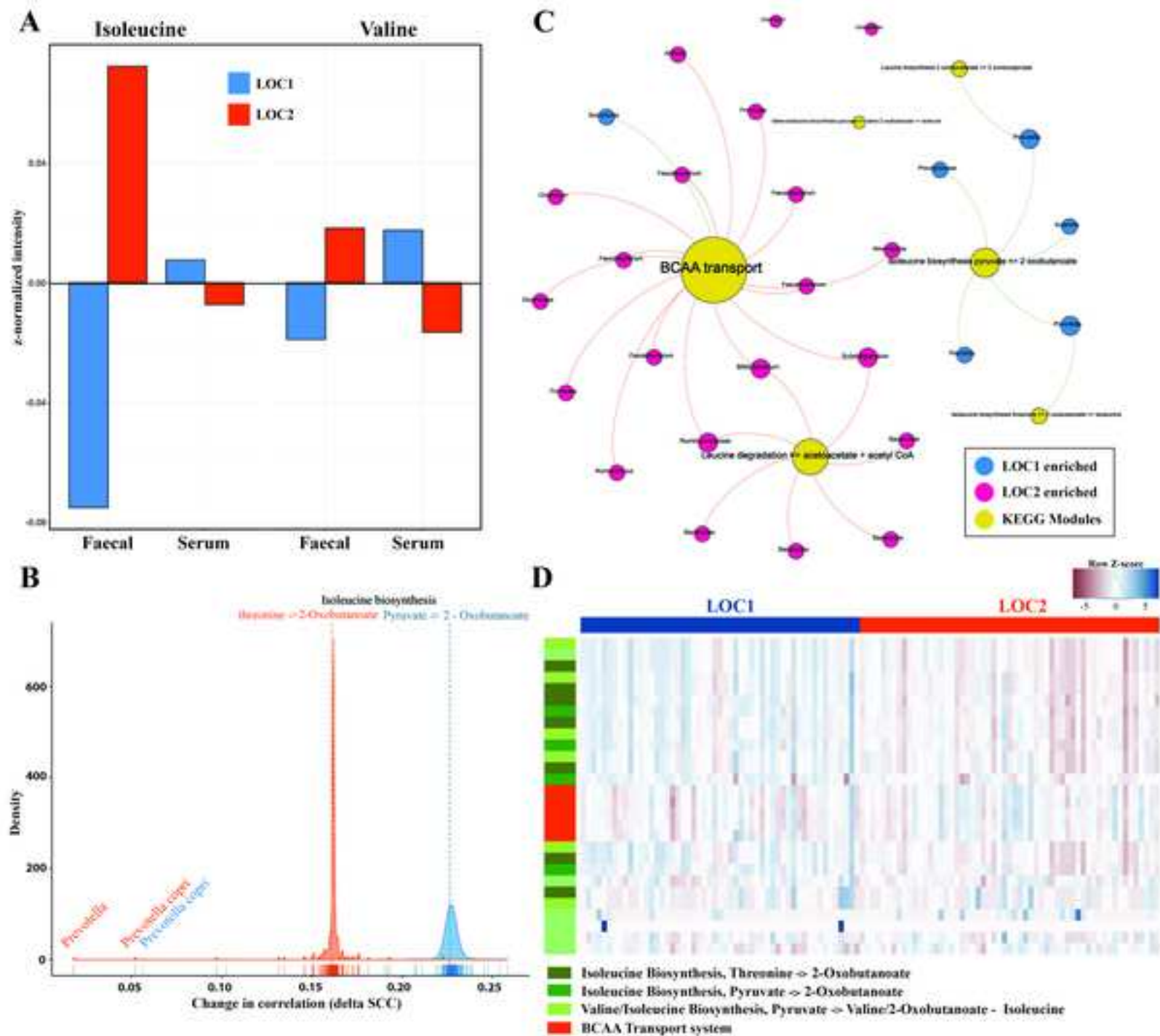


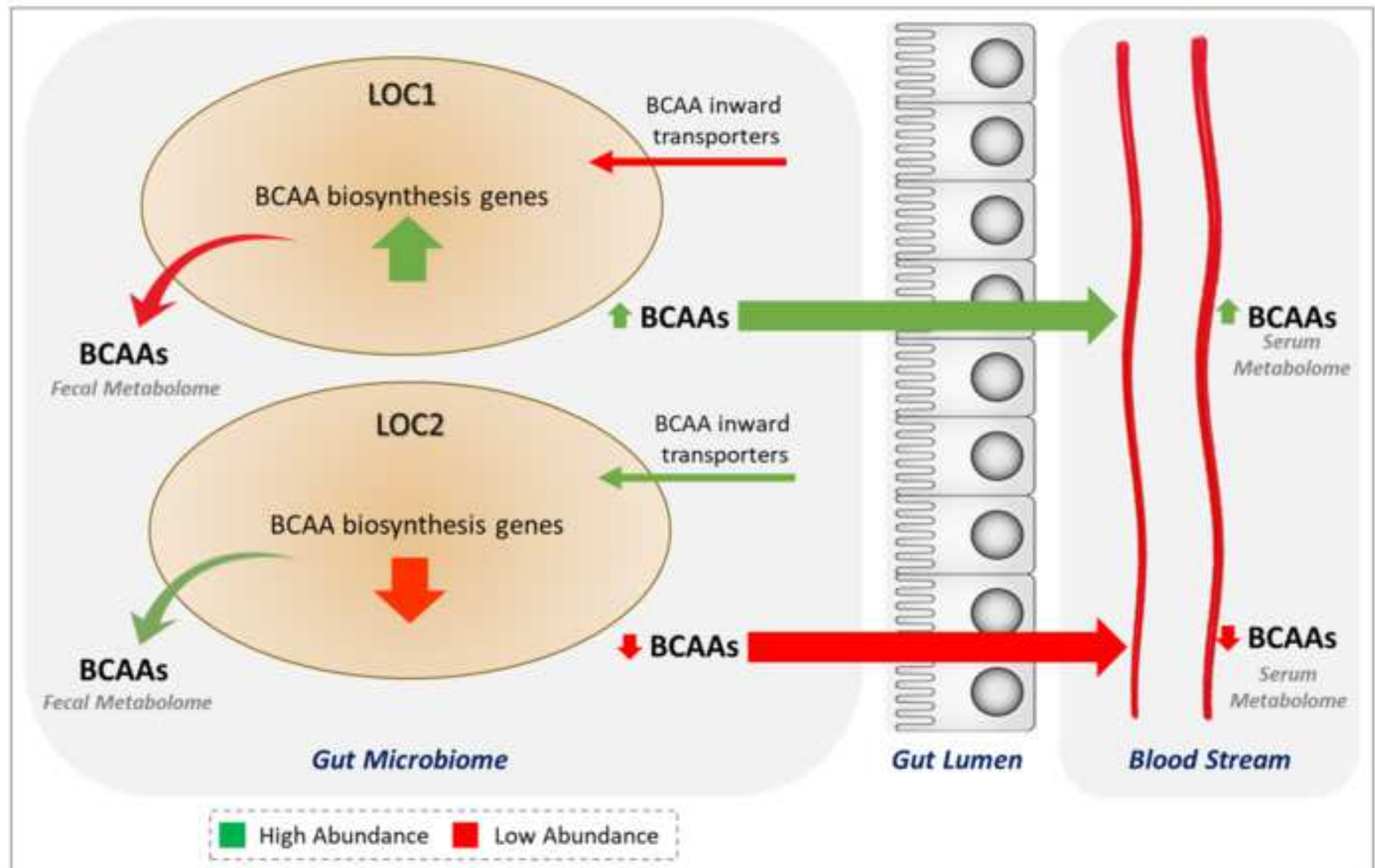


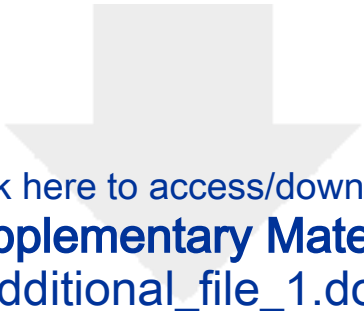







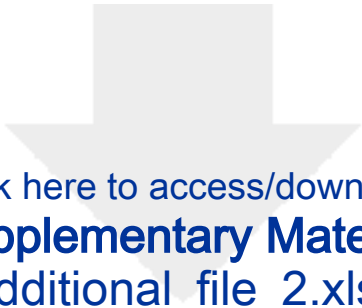







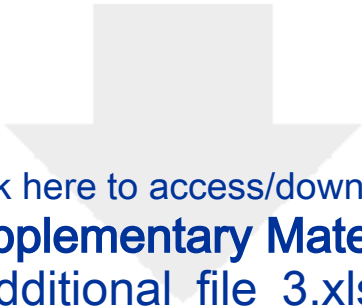
Click here to access/download
Supplementary Material
Additional_file_1.doc






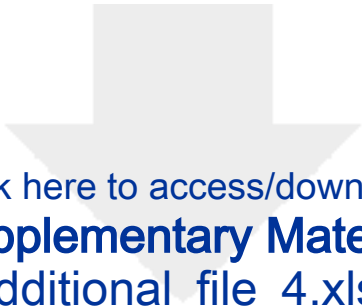
Click here to access/download
Supplementary Material
Additional_file_2.xlsx






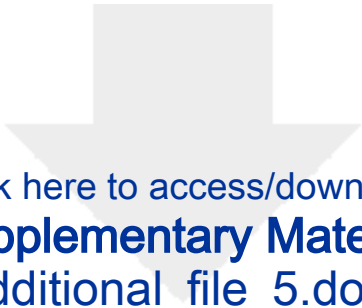
Click here to access/download
Supplementary Material
Additional_file_3.xlsx






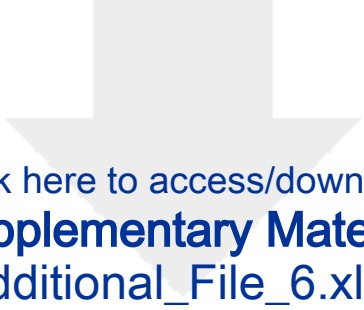
Click here to access/download
Supplementary Material
Additional_file_4.xlsx



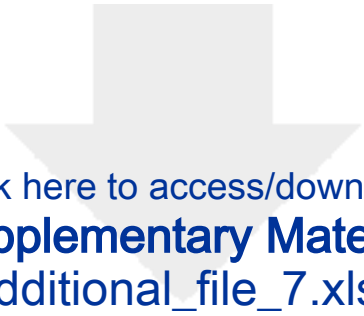


Click here to access/download
Supplementary Material
Additional_file_5.docx




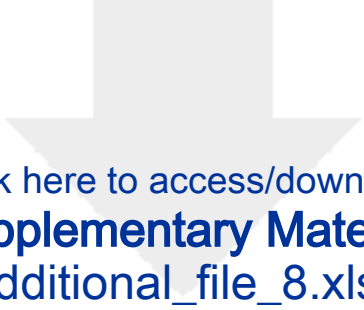


Click here to access/download
Supplementary Material
Additional_File_6.xlsx

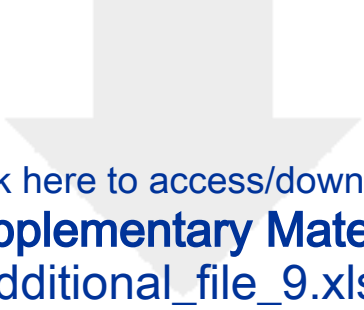


Click here to access/download
Supplementary Material
Additional_file_7.xlsx

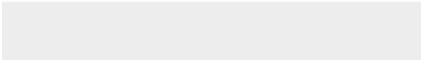



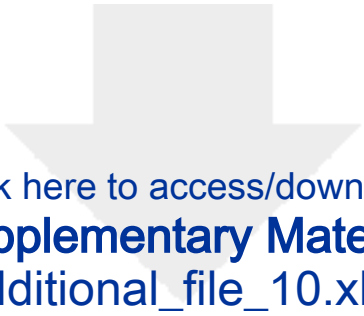


Click here to access/download
Supplementary Material
Additional_file_8.xlsx




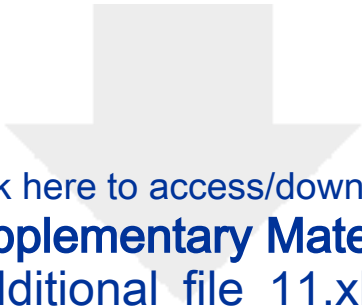
Click here to access/download
Supplementary Material
Additional_file_9.xlsx






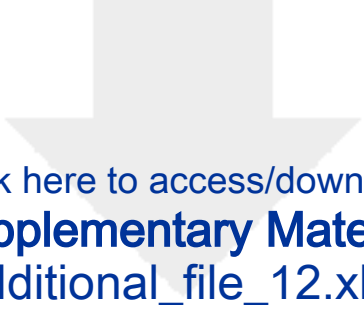
Click here to access/download
Supplementary Material
Additional_file_10.xlsx



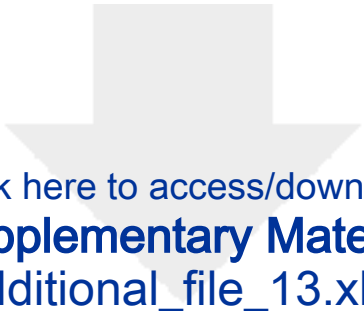


Click here to access/download
Supplementary Material
Additional_file_11.xlsx







Click here to access/download
Supplementary Material
Additional_file_12.xlsx




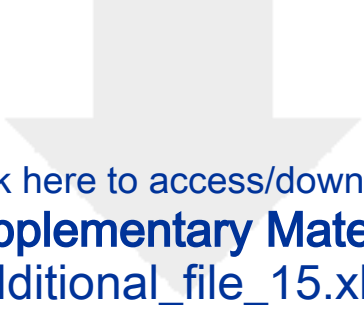
Click here to access/download
Supplementary Material
Additional_file_13.xlsx



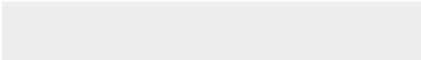



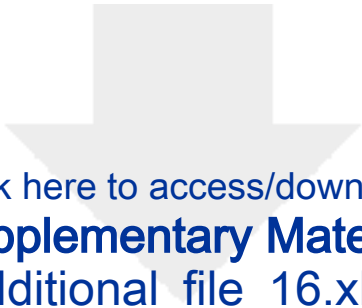
Click here to access/download
Supplementary Material
Additional_file_14.xlsx






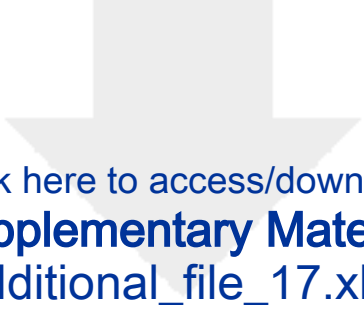
Click here to access/download
Supplementary Material
Additional_file_15.xlsx



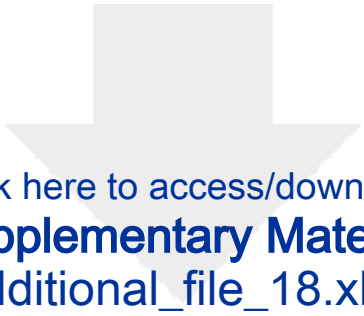


Click here to access/download
Supplementary Material
Additional_file_16.xlsx





Click here to access/download
Supplementary Material
Additional_file_17.xlsx



Click here to access/download
Supplementary Material
Additional_file_18.xlsx

