

## Multi-omics analysis reveals Indian gut microbiome variations due to diet and location and its implications on human health --Manuscript Draft--

<b>Manuscript Number:</b>	GIGA-D-18-00212
<b>Full Title:</b>	Multi-omics analysis reveals Indian gut microbiome variations due to diet and location and its implications on human health
<b>Article Type:</b>	Research
<b>Funding Information:</b>	
<b>Abstract:</b>	<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 not well-explored yet. We comprehensively analysed the gut microbiota of 110 healthy individuals from two distinct locations (North-Central and Southern India) using multi-omics approaches, including 16S rRNA marker gene and shotgun metagenomics, and faecal and serum metabolomics.</p> <p>Results</p> <p>The gene catalogue established in this study highlighted the uniqueness of the Indian gut microbiome in comparison to other populations. The North-Central population, which was primarily consuming a plant-based diet, was found to be associated with Prevotella, and thus showed an enrichment of BCAA and lipopolysaccharide biosynthesis pathways. In contrast, the South-Indian population, which was consuming an omnivorous diet, showed associations with Bacteroides, Ruminococcus and Faecalibacterium, and had an enrichment of SCFA biosynthesis pathway and BCAA transporters. This corroborated with the metabolomic results, where the BCAA levels were observed to be higher in the serum metabolome of the North-Central population, apparently regulated by Prevotella. In contrast, BCAAs were found higher in the faecal metabolome of South-Indian population, which was correlated with the enrichment of BCAA transporters.</p> <p>Conclusions</p> <p>The study demonstrates the influence of location and diet on the gut microbiome and its functional consequences on human health, and supplements the current knowledge on the poorly characterized Indian gut microbiome. The integrated approach used provides novel insights on the gut-microbe-metabolic axis, which will be useful for future epidemiological and translational researches.</p>
<b>Corresponding Author:</b>	Vineet Kumar Sharma, Ph.D. Indian Institute of Science Education and Research Bhopal Bhopal, Madhya Pradesh INDIA
<b>Corresponding Author Secondary Information:</b>	
<b>Corresponding Author's Institution:</b>	Indian Institute of Science Education and Research Bhopal
<b>Corresponding Author's Secondary Institution:</b>	
<b>First Author:</b>	Darshan B Dhakan
<b>First Author Secondary Information:</b>	
<b>Order of Authors:</b>	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.
<b>Order of Authors Secondary Information:</b>	
<b>Additional Information:</b>	
<b>Question</b>	<b>Response</b>
Are you submitting this manuscript to a special series or article collection?	No
<p><b>Experimental design and statistics</b></p> <p>Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our <a href="#">Minimum Standards Reporting Checklist</a>. 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><b>Resources</b></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 <a href="#">Research Resource Identifiers</a> (RRIDs) for antibodies, model organisms and tools, where possible.</p> <p>Have you included the information requested as detailed in our <a href="#">Minimum Standards Reporting Checklist</a>?</p>	Yes
<p><b>Availability of data and materials</b></p> <p>All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in <a href="#">publicly available repositories</a> (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?](#)

1 **Title:** Multi-omics analysis reveals Indian gut microbiome variations due to diet and location and  
2 its implications on human health

3 **Authors:** D.B. Dhakan<sup>1†</sup>, A. Maji<sup>1†</sup>, A.K. Sharma<sup>1</sup>, R. Saxena<sup>1</sup>, J. Pulikkan<sup>2</sup>, T. Grace<sup>2,3</sup>, A.  
4 Gomez<sup>4</sup>, J. Scaria<sup>5</sup>, K.R. Amato<sup>6</sup>, V.K. Sharma<sup>1\*</sup>

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

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

15 † These authors contributed equally to this work

16 \*Corresponding author

17 V.K. Sharma: [vineetks@iiserb.ac.in](mailto:vineetks@iiserb.ac.in)

18  
19  
20

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

21 **Abstract**

22 **Background**

23 Metagenomic studies carried out in the past decade have led to an enhanced understanding of the  
24 gut microbiome in human health, however, the Indian gut microbiome is not well-explored yet.  
25 We comprehensively analysed the gut microbiota of 110 healthy individuals from two distinct  
26 locations (North-Central and Southern India) using multi-omics approaches, including 16S rRNA  
27 marker gene and shotgun metagenomics, and faecal and serum metabolomics.

28 **Results**

29 The gene catalogue established in this study highlighted the uniqueness of the Indian gut  
30 microbiome in comparison to other populations. The North-Central population, which was  
31 primarily consuming a plant-based diet, was found to be associated with *Prevotella*, and thus  
32 showed an enrichment of BCAA and lipopolysaccharide biosynthesis pathways. In contrast, the  
33 South-Indian population, which was consuming an omnivorous diet, showed associations with  
34 *Bacteroides*, *Ruminococcus* and *Faecalibacterium*, and had an enrichment of SCFA biosynthesis  
35 pathway and BCAA transporters. This corroborated with the metabolomic results, where the  
36 BCAA levels were observed to be higher in the serum metabolome of the North-Central  
37 population, apparently regulated by *Prevotella*. In contrast, BCAAs were found higher in the faecal  
38 metabolome of South-Indian population, which was correlated with the enrichment of BCAA  
39 transporters.

40 **Conclusions**

41 The study demonstrates the influence of location and diet on the gut microbiome and its functional  
42 consequences on human health, and supplements the current knowledge on the poorly

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

43 characterized Indian gut microbiome. The integrated approach used provides novel insights on the  
44 gut-microbe-metabolic axis, which will be useful for future epidemiological and translational  
45 researches.

46 **Keywords:** Indian Gut Microbiome, Metagenomics, Metabolomics, Enterotypes, Integrated Gene  
47 Catalog, Metagenome-Wide Association Study, Core gut microbiota, Short Chain Fatty Acids,  
48 Branched Chain Amino Acids, Lipopolyschharides.

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

62 **Background**

63 Determining the constitution of a healthy gut microbiota and understanding its variability across  
64 populations is essential for assessing the impact of microbial dysbiosis on human health. Several  
65 large-scale, world-wide microbiome projects have revealed variability in the gut microbial  
66 composition of healthy individuals due to factors such as mode of delivery, age, geographical  
67 location, diet and lifestyle, and have helped in the better understanding of gut microbiome in  
68 human health and disease [1-5]. Most gut microbiome studies have determined microbial  
69 taxonomy and functional diversity using marker gene-based and/or WGS approaches to understand  
70 the functional role of the gut microbiome. However, novel insights on the complex interplay  
71 between diet, gut microbes and human health in the context of key microbial metabolites, such as  
72 short-chain fatty acids (SCFAs) and Branch Chain Amino Acids (BCAAs), derived from the  
73 microbial fermentation of dietary fibres are beginning to emerge from recent gut metabolomics  
74 studies [6, 7]. Moreover, the direct impact of the microbial metabolome on human health is also  
75 becoming apparent from the recent studies focusing on the ‘gut microbiome- host metabolism axis’  
76 [8]. Therefore, an integrative approach using both metagenome and metabolome- based  
77 characterizations of the gut microbiome appears pragmatic for gaining deeper functional and  
78 mechanistic insights into the role of gut microbes on human health.

79 The significant large-scale studies carried out so far represent the gut microbiome of urban  
80 populations majorly from Europe, US and other allegedly named WEIRD countries (i.e., the  
81 Western, Educated, Industrialized, Rich, and Democratic countries) [9]. Only recently, some  
82 initiatives have been taken for the characterization of the human microbiome from diverse ethnic  
83 populations, which have shown significant variations from the major world populations [9-14].  
84 India is the seventh largest country in the world and harbours the second largest population with

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

85 enormous diversity in populations, lifestyles and dietary habits across multiple geographical  
86 locations. India is also a home to the majority of the world's vegetarian population, but is equally  
87 dominated by people consuming both vegetarian and animal-based diets. Moreover, the Indian  
88 population has the highest prevalence of diabetes worldwide [15] and according to the World  
89 Health Organization estimates (WHO, 2011), 53% of deaths in India in the year 2008 were  
90 attributed to non-communicable conditions such as diabetes and cardiovascular diseases, which  
91 are predicted to reach ~75% by 2030 [16]. The gut microbiome has been implicated in many such  
92 diseases in India and in other populations [14, 17, 18]. Describing population-specific variations  
93 in the microbial profile of healthy individuals is critical for identifying population-specific as well  
94 as universal microbiome-based biomarkers for specific diseases [19]. A few studies have  
95 investigated the gut microbiome of the Indian population, but all were focused on small cohort  
96 sizes and have relied only on 16S rRNA gene-based sequencing analysis [10, 20, 21]. Therefore,  
97 investigating the impact of diet and location on the gut microbiome of the Indian population is  
98 crucial for improving our understanding on the role of the gut microbiome in health and disease in  
99 a global context.

100 To cover the enormous gut microbiome diversity inherent in the different sub-populations of India,  
101 extensive sampling and analyses are required. Therefore, as the first large-scale study from India,  
102 we selected two prominent locations in North-Central India, i.e. LOC1: Bhopal city, Madhya  
103 Pradesh, and Southern India, i.e. LOC2: Kerala. The dietary habits between the two locations are  
104 very different, as the South-Indian population (LOC2) diet consists of rice, meat and fish, whereas  
105 the North-Central population (LOC1) consumes a carbohydrate-rich diet including plant-derived  
106 products, wheat and trans-fat food (high-fat dairy, sweets and fried snacks). In addition, the  
107 'Human Development Index Report, UNDP' (United Nations Development Programme), India



1  
2  
3  
4 108 and SRS-based life-table (Sample Registration Survey, 2010-14) has revealed that the citizens  
5  
6  
7 109 from Kerala had the highest life-expectancy rates (>74 years) in India, while those in Madhya  
8  
9 110 Pradesh (capital city 'Bhopal') exhibited the lowest (<65 years) [22]. Further, it is known that there  
10  
11  
12 111 is a higher predisposition of the North-Indian population towards diabetes, cardiovascular diseases  
13  
14 112 and hypertension, which in contrast is much lower in Southern India, perhaps due to the lifestyle  
15  
16 113 differences in the two regions [3]. Thus, to gain deeper functional insights into the microbiome  
17  
18  
19 114 from these two distinct and representative sub-populations of India, a comprehensive multi-omics  
20  
21 115 approach was carried out using amplicon-based profiling of taxonomic composition (16S rRNA  
22  
23  
24 116 sequencing), WGS-based profiling of metagenomic content and GC-MS-based profiling of faecal  
25  
26 117 and serum metabolomic signatures.

## 29 118 **Data Description**

31  
32 119 The two selected locations, Bhopal (LOC1) and Kerala (LOC2) provided a distinct representation  
33  
34 120 of the Indian population in the context of diets and lifestyle from North-Central and Southern parts  
35  
36  
37 121 of India, which are almost 2000 km apart (**Additional File 1**). The 110 (62 females, 58 males)  
38  
39 122 individuals recruited in this study were not suffering from any disease as reported by personal  
40  
41 123 medical history and physical examination, and confirmed no exposure to antibiotics for at least  
42  
43  
44 124 one month prior to sampling, and thus, were considered as 'healthy' (**Additional File 1**). The  
45  
46 125 sequencing of V3 hypervariable region of 16S rRNA gene and shotgun metagenome sequencing  
47  
48  
49 126 from the 110 faecal samples resulted into 54.87 million paired-end reads ( $503,460 \pm 175,547$   
50  
51 127 (mean  $\pm$  sd) reads/sample) and 499.98 million paired-end reads ( $4,545,280 \pm 1,498,663$  (mean  $\pm$   
52  
53 128 sd) reads/sample), respectively (Methods, **Additional file 2** and **Additional file 3**).

## 57 129 **Analyses**

58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 **130 Construction of an Indian gut microbial gene catalogue and updated integrated gene**  
5  
6 **131 catalogue (IGC)**  
7  
8

9 **132** The first step for functional analysis was the construction of an extensive catalogue of gut  
10  
11 **133** microbial genes from the Indian population, which was not yet available. A De Bruijn graph-based  
12  
13 **134** assembly of reads resulted in 1,337,547 contigs of length  $\geq 300$  bp with a total contig length of  
14  
15 **135** 1.78 Gbp representing 43% of the total reads. To obtain assemblies of low coverage genomic  
16  
17 **136** regions or genomes present in the Indian gut microbiome, the singletons from all the samples were  
18  
19 **137** combined and assembled into additional 0.33 million contigs with length  $\geq 300$  bp and a total  
20  
21 **138** assembled length of 232 Mbp. The ORFs predicted from contigs resulted in 1,479,998 non-  
22  
23 **139** redundant genes, which represent the gene catalogue of the Indian gut microbiome. In addition,  
24  
25 **140** the integrated gene catalogue (IGC) represents a cohort of 9,879,896 genes identified from 1,267  
26  
27 **141** gut metagenomes from three populations of the world (HMP, MetaHIT and Chinese dataset), and  
28  
29 **142** was also updated with the Indian gene catalogue to construct an updated IGC [1, 23, 24]. A total  
30  
31 **143** of 718,360 non-redundant genes were added from Indian samples, which increased the size of IGC  
32  
33 **144** to 10,598,256 non-redundant genes (6.7% increase), and was referred to as ‘updated IGC’. A total  
34  
35 **145** of 69.2% ( $\pm 4.01\%$ ) mapping coverage of reads ( $\sim 6\%$  increase in the mapping of reads) was  
36  
37 **146** observed on the updated IGC as compared to 63% ( $\pm 4.61\%$ ) mapping on the previous non-updated  
38  
39 **147** IGC (**Additional File 4**). However, a similar increment in mapping coverage of reads for other  
40  
41 **148** population datasets was not observed, and the mapping coverage of HMP (67.74%), China  
42  
43 **149** (73.38%) and MetaHIT (75.21%) on the updated IGC were comparable to their mapping coverage  
44  
45 **150** to IGC (**Fig. 1A**). This analysis indicates that the genes contributed by the Indian gut microbiome  
46  
47 **151** are unique and not represented in other gut microbiome datasets.  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58

59 **152 Identification of taxonomic signatures of Indian gut microbiome**  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 153 To determine the taxonomic and functional composition of the Indian gut microbiome and to  
5  
6 154 identify Indian specific gut-microbial signatures, a cross-population comparison was carried out  
7  
8  
9 155 using the 16S rRNA and metagenomic data from other populations. In order to derive  
10  
11 156 metagenomic markers for comparison with similar large-scale studies from other populations, a  
12  
13  
14 157 non-reference based metagenome-wide association study (MGWAS) was carried out [25]. The  
15  
16 158 genes from metagenomic samples of four countries (India, China, USA and Denmark) were  
17  
18  
19 159 clustered (see Methods) into 866 clusters based on their co-occurrence and higher Pearson  
20  
21 160 correlations across samples ( $\rho = 0.9$ ) resulting into 224 MGS (metagenomic species) having  $\geq 700$   
22  
23  
24 161 genes in each cluster, and 642 CAGs (co-abundance gene groups) consisting of  $\geq 50$  genes in each  
25  
26 162 cluster. Out of the 866 metagenomic clusters, 197 could be assigned up to species level using the  
27  
28  
29 163 taxonomic assignment strategy described in Methods. Jaccard distances were calculated from  
30  
31 164 MGS/CAG abundance profiles and their PCA analysis was carried out using ‘countries’ as factors  
32  
33  
34 165 for explaining the variance between samples, which showed that the Indian population formed a  
35  
36 166 distinct cluster separated from the other populations at PC1 (**Fig. 1B**). The MGS/CAGs annotated  
37  
38 167 as *Prevotella*, *Mitsuokella*, *Dialister*, *Megasphaera*, and *Lactobacillus* were found to be the drivers  
39  
40  
41 168 of this separation as observed from their factor loading scores, and were associated with and  
42  
43 169 enriched in the Indian population. Further, the identification of enriched MGS showed that the  
44  
45  
46 170 species belonging to the genus *Clostridium*, and phylum Firmicutes and Bacteroides were depleted  
47  
48 171 in the Indian population and were enriched in the other populations (China, Denmark and USA;  
49  
50  
51 172 Log Odds Ratio  $< -7$  and P-value  $< 0.001$ ) (**Additional File 5: Figure S1**). Furthermore, the  
52  
53 173 distribution of microbial families from different populations was also calculated across the globe  
54  
55  
56 174 using 16S rRNA markers. A cross-population comparison revealed Indian gut microbiome to have

1  
2  
3  
4 175 a higher abundance of Prevotellaceae and Veillonellaceae, suggesting them as the marker  
5  
6 176 microbial families associated with the Indian population (**Fig. 1C**).

### 10 177 **Microbial functions enriched in the Indian population**

11  
12 178 Functional comparison of Indian microbiome with other populations was carried out by mapping  
13  
14  
15 179 the genes derived from assembled contigs to the EggNOG database. In total 68,693 EggNOG  
16  
17 180 functions were identified from the Indian gut microbiome, including 1,726 novel functions  
18  
19  
20 181 obtained from clustering the unmapped genes (see Methods). The core microbial functions which  
21  
22 182 are essential for microbial survival and are present in almost 80% individuals were used for the  
23  
24  
25 183 functional comparison. The core microbiome was derived using a similar strategy as employed in  
26  
27 184 MetaHIT (see Methods) [26]. A core microbial EggNOG profile was generated using a gene cohort  
28  
29  
30 185 comprising of 1,890 essential genes from six bacterial species namely, *Escherichia coli* MG1655I  
31  
32 186 and MG165II, *Bacteroides thetaiotaomicron* VPI-5482, *Pseudomonas* PA01, *Salmonella enteric*  
33  
34 187 serovar Typhi and *Staphylococcus aureus* NCTC 8325. The eggNOGs were ranked based on their  
35  
36  
37 188 mean abundance in descending order, and the range that included 85% of essential genes were  
38  
39 189 considered for building the core microbial eggNOG set and were used for the analysis. Most of  
40  
41  
42 190 the essential genes were included in the top-ranking clusters suggesting that the essential genes are  
43  
44 191 present in higher abundance than the accessory function genes (**Additional File 5: Figure S2**).

45  
46 192 The core microbiome of Indian samples was compared with the core microbiome of USA, China  
47  
48  
49 193 and Denmark populations. The proportion of essential genes covered by top-ranking nine eggNOG  
50  
51 194 clusters showed that 85% of the essential genes could be covered in the least number (15,000) of  
52  
53  
54 195 eggNOGs in the case of Indian population, while in the case of Denmark it was covered by twice  
55  
56 196 the number (30,000) of eggNOGs (**Additional File 5: Figure S3**). These observations suggest that  
57  
58  
59 197 the core functional microbiome of Indian population is less diverse than other populations. This

60  
61  
62  
63  
64  
65

1  
2  
3  
4 198 also corroborates with the alpha diversity (Shannon) calculations using gene abundances, which  
5  
6 199 showed that the Indian microbiome is less diverse than the microbiome of other world populations  
7  
8  
9 200 (**Additional File 5: Figure S4**). In total, 5,296 eggNOGs were characterized as core functions  
10  
11 201 commonly present in the core microbiome of all the four population datasets. The co-inertia  
12  
13  
14 202 (Procrustes) analysis and the Eigen values, and their scores calculated from PCA, using both core  
15  
16 203 and accessory functions also showed that the Indian gut microbiome was significantly different  
17  
18  
19 204 from other datasets (**Fig. 2A and 2B**). This data also shows the uniqueness of Indian microbial  
20  
21 205 functions in composition and diversity at both core and accessory levels. The Indian microbiome  
22  
23  
24 206 was found to be enriched (FDR Adj.  $P < 0.05$ , Log Odds Ratio  $> 1.5$ ) in functions for carbohydrate  
25  
26 207 and energy metabolism including degradation of complex polysaccharides, which corroborates  
27  
28  
29 208 well with the carbohydrate-rich diet of the Indian population (**Fig. 2C and Additional File 6**).

### 30 31 32 209 **Detection of enterotypes and variations in Indian gut microbiome between locations**

33  
34 210 To determine the diversity of gut microbial communities present in the Indian population,  
35  
36  
37 211 detection of enterotypes (groups of samples having similar profiles and lesser variance) was  
38  
39 212 performed using an unsupervised clustering approach [2]. The Jensen Shannon distance matrices  
40  
41  
42 213 were used and principal component analysis identified two prominent enterotypes. ET-1 was  
43  
44 214 primarily driven by *Prevotella* ( $P < 0.001$ ), and ET-2 was driven by other microbes belonging to  
45  
46 215 *Bacteroides* ( $P < 0.02$ ), *Ruminococcus* ( $P < 0.001$ ) and *Faecalibacterium* ( $P < 0.02$ ) (**Additional File**  
47  
48  
49 216 **5: Figure S5, Additional File 7**). The abundances of *Prevotella* in LOC1 and *Bacteroides* in LOC2  
50  
51 217 in India are perhaps due to the dietary habits of the two locations. The LOC1 population was  
52  
53  
54 218 mainly consuming a carbohydrate-rich diet comprising of vegetable-based foods and grains,  
55  
56 219 whereas the LOC2 population was consuming a diet consisting of rice, meat and fish. These  
57  
58  
59 220 patterns seem to align with the patterns reported in other populations [27, 28].  
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

221 The robustness of clusters was demonstrated using Calinski Harabasz index (CHI) and prediction  
222 strength, which uses a cross validation approach (**Additional File 8**). A similar cluster analysis  
223 performed using the functional information derived from the abundance of KEGG Orthologs (KO)  
224 also showed the clustering of samples into two enterotypes, named as C1 and C2 (**Additional File**  
225 **5: Figure S6**). In comparison to enterotypes derived from taxonomic information, only 16 out of  
226 110 samples were placed in different clusters using the functional information revealing significant  
227 concordance (FDR Adj.  $P < 0.0001$ , Spearman's correlation Coefficient = 0.69). C1 was found  
228 enriched in genes coding for enzymes such as Phytase (Log Odds Ratio (LOR) = 2.96),  $\beta$ -  
229 glucosidase (LOR = 1.75), and  $\alpha$ -fucosidase (LOR = 1.32), which are involved in the breakdown  
230 of plant-polysaccharides, whereas the genes coding for enzymes such as lipase (LOR = -5.34),  
231 carnitine-coA dehydratase (LOR = -2.59) and amino peptidase (LOR = -2.66), which are involved  
232 in the metabolism of animal-based diet, were enriched in C2 (FDR Adj.  $P < 0.05$ ) (**Additional File**  
233 **9**).

234 To identify the components explaining the variations in microbial profiles across samples,  
235 unweighted UniFrac distances were calculated using 16S rRNA sequences rarefied at 100,000  
236 sequences per sample. The principal component analysis (PCA) of Unifrac distances and the scores  
237 for each sample correlated with the covariates using polyserial correlation, and distinct locations  
238 (LOC1 and LOC2) and diets (vegetarian and omnivorous) were identified to be the major variables  
239 explaining the variation between samples at PC2 (**Fig. 3A, Additional File 10**). A comparison of  
240 taxonomic and functional diversity performed between the two locations using Shannon diversity  
241 index and rarefactions of genes from each sample, also showed that the microbiome profiles of  
242 LOC2 populations were more diverse in their composition compared to LOC1 populations (**Fig.**  
243 **3B and Additional File 5: Figure S7**). The inter-individual Bray Curtis distances of gene profiles

1  
2  
3  
4 244 between LOC1 and LOC2 populations also showed significant differences (FDR Adj.  $P < 0.05$ ),  
5  
6 245 where LOC2 population displayed higher inter-individual heterogeneity in their microbial  
7  
8  
9 246 community structure as compared to LOC1 population (**Fig. 3C**).

10  
11  
12 247 Major differences in the microbiome profiles (using the 16S rRNA dataset) at the phylum level  
13  
14  
15 248 were apparent from the higher Bacteroidetes to Firmicutes ratios ( $P < 0.002$ ) in LOC1 (1.93)  
16  
17 249 compared to LOC2 (0.86), which have been previously reported as a result of differences in dietary  
18  
19  
20 250 habits, i.e. vegetarian or plant-based (carbohydrate-rich) vs. omnivore or animal-based (protein-  
21  
22 251 rich) diets (**Additional File 5: Figure S8**) [29, 30]. Notably, these variations were not attributable  
23  
24  
25 252 to BMI (Spearman's Rank correlation, FDR Adj.  $P = 0.78$ ). At the genus level also *Prevotella*,  
26  
27 253 *Megasphaera*, *Mitsuokella*, and *Lactobacillus* were observed to be higher in LOC1, whereas  
28  
29  
30 254 *Ruminococcus*, *Clostridium*, *Faecalibacterium* and *Roseburia* were higher in LOC2 (FDR Adj.  
31  
32 255  $P < 0.05$ , Wilcoxon rank sum test); (**Fig. 3D & E**). Similarly, out of 107 marker MGS/CAG  
33  
34 256 obtained from MGWAS, those annotated to *Prevotella copri* were found enriched in LOC1 (Log  
35  
36  
37 257 Odds Ratio  $> 2$ ; FDR Adj.  $P < 0.05$ ; 41 MGS/CAG), whereas MGS/CAGs annotated to SCFA  
38  
39 258 producing species such as *Faecalibacterium prausnitzii* and *Roseburia inulinivorans*, were  
40  
41  
42 259 enriched in LOC2 (FDR Adj.  $P < 0.05$ ; Log Odds Ratio  $< -2$ ; 66 MGS/CAG) (**Additional File 11**).  
43  
44 260 Interestingly, the two species found higher in LOC2 are known SCFA producers and have also  
45  
46  
47 261 been regarded as commensals with anti-inflammatory properties [31]. In contrast, *Prevotella*,  
48  
49 262 which was abundant in the LOC1, is known to be associated with high fibre-rich diet [32].  
50  
51

### 52 263 **Defining the Indian gut metabolome**

53  
54  
55 264 The analysis of microbial community structure and functions from the two locations having  
56  
57 265 different lifestyle and diet revealed significant insights. Previous studies have shown a direct role  
58  
59  
60 266 of diet in the selection of differential gut microbiomes [33]. Thus, to gain deeper insights into the  
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

metabolic activity of microbiomes from LOC1 and LOC2 as driven by different diets, faecal metabolites were analysed using a GC-MS-based metabolomics approach. An unsupervised between class analysis of metabolomic profiles separated the samples into three separate clusters, and the robustness was confirmed using prediction strength and Silhouette index (**Fig. 4A and 4B**). Polyserial correlation of covariates showed location to be the major factor explaining the variation at PC1 (FDR Adj.  $P < 0.01$ ) separating Metabotype-1 from Metabotype-2 and 3. In contrast, vegetarian and omnivorous diet groups emerged as other factors explaining the variation at PC2 (FDR Adj.  $P < 0.01$ ), and separating Metabotype-2 from 3 (**Additional File 12**). The OPLS-DA model derived from normalized peak intensities also showed differential clustering of samples from the two locations (**Fig. 4C, Table 2**). Metabotype-1 was associated with LOC1 and showed higher abundances of saturated fatty acids including palmitic acid, stearic acid, and valeric acid. Metabotype-3 was associated with LOC2 and showed higher abundances of BCAAs valine, leucine and isoleucine, and SCFAs propionate and butyrate. Metabotype-2 was enriched in D-glucose, galactose, mannose, lauric acid and cadaverine (a polyamine that denotes meat consumption) [34].

**Positive correlation of BCAA transporters with BCAA levels in faecal metabolome**

We also identified the marker metabolites, which showed significant (Spearman's correlation, FDR Adj.  $P < 0.05$ ) associations with LOC1 or LOC2. In total, 17 metabolite clusters were identified, of which nine were associated with LOC1, and eight were associated with LOC2 (**Additional File 13**). These marker metabolites showed a positive association with MGS/CAGs. For instance, *Prevotella* annotated clusters correlated significantly with valeric acid and sedoheptulose metabolite markers, which showed a higher relative abundance in LOC1. In contrast, MGS/CAGs belonging to *Faecalibacterium*, *Clostridium*, *Ruminococcus*, and *Alistipes*



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

were positively associated with BCAAs, cadaverine, propanoate and lauric acid in LOC2 (**Fig. 5A**). In addition to the positive association of BCAAs with species enriched in LOC2, a correlation analysis of significantly different (FDR Adj.  $P < 0.05$ , Wilcoxon rank sum test; **Additional File 14**) functional modules revealed that faecal BCAA abundances were positively correlated with BCAA transporter abundance in LOC2. In contrast, BCAA abundance in the faecal metabolome showed a negative correlation ( $P < 0.05$ ) with BCAA biosynthesis pathways (**Fig. 5B**).

The above observations are significant given that BCAAs are important metabolites involved in glucose homeostasis, by stimulating insulin secretion [35]. Higher BCAA levels in the faecal matter could be a result of its inward transport in microbial cells by the BCAA transporters, thus leading to their accumulation in the colon lumen. This is concordant with higher relative abundance of *Bacteroides vulgatus* and *Eubacterium sireaeum* in LOC2 compared to LOC1, which are known to harbour higher abundance of BCAA transporters [36]. Further support for this hypothesis emerged from the correlation of circulating BCAA levels (valine and isoleucine) in serum with the corresponding levels in feces. Interestingly, serum BCAA levels were significantly higher in LOC1 individuals as compared to LOC2 individuals, which contrasted with the BCAA levels in the faecal metabolome (**Fig. 6A**). Thus, it is likely that the accumulation of BCAA in the feces of individuals of LOC2 was mediated by their gut microbiome. In contrast, due to the lower BCAA accumulation in feces and a higher BCAA biosynthesis by microbial species in LOC1, BCAA levels were observed to be in higher concentration in serum of LOC1 population, and hence higher BCAA absorption.

***Prevotella copri* regulates BCAA levels through threonine-independent isoleucine biosynthesis pathway**

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

312 To explore the differences in association of functional pathway modules between the two  
313 locations, KOs within each module were correlated with KOs from other modules using  
314 Spearman's correlation coefficient. The KOs showing significant differences in correlations  
315 between LOC1 and LOC2 were identified. This differential correlation analysis of BCAA  
316 biosynthetic modules with other pathways in LOC1 and LOC2 revealed that BCAA modules were  
317 independently driven in LOC1 and LOC2 (Spearman's rank correlation, FDR Adj. P<0.01) (**Fig.**  
318 **6B and 6C**). To identify the species and the metabolic pathways that contributed most to the  
319 BCAA abundance in faecal and serum metabolome profiles, a correlation analysis with iterations  
320 leaving each species out was performed for each metabolic module (**Additional File 5: Figure**  
321 **S9**). The species whose removal leads to a maximum change in the correlation of metabolic  
322 pathway with metabolite was identified, and was considered as an important contributor of that  
323 metabolite [8]. Notably, a single species *Prevotella copri* was found driving the 'threonine-  
324 independent isoleucine biosynthesis' functional module. Among the other BCAA biosynthesis  
325 pathways, valine biosynthesis was also driven by species from *Prevotella*.

326 The correlation network analysis with different MGS/CAGs also revealed threonine-independent  
327 isoleucine biosynthesis pathway to be highly correlated with *Prevotella copri* in LOC1, and was  
328 the major pathway utilized by this species for BCAA biosynthesis (**Fig. 6D**). The first enzyme, D-  
329 citramalate synthase, catalysing the threonine-independent isoleucine biosynthesis pathway was  
330 also observed as highly enriched (LOR = 1.7) in LOC1. Further, BCAA biosynthesis was observed  
331 to be higher in LOC1 as compared to LOC2, and BCAA transporters were found higher in LOC2  
332 as compared to LOC1 (**Fig. 6E**).

333 **Discussion**

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

334 Compositional and functional human gut microbiome studies in different populations have been  
335 instrumental in establishing the role of gut microbiome in human health [28, 37-39]. However,  
336 such population-specific signatures and functional insights for the Indian gut microbiome are yet  
337 unknown. Thus, the present work provides the first comprehensive survey of the Indian gut  
338 microbiome represented through a cohort of 110 individuals from two prominent locations. Several  
339 insights into the taxonomic and functional diversity emerged from the 16S rRNA and metagenomic  
340 analysis and were validated through metabolomic profiling, which is a prominent highlight of this  
341 study. Given the high diversity of diet and lifestyle in India, the selection of two distinct locations  
342 (Bhopal – LOC1, and Kerala – LOC2) as the representative sub-populations was an important  
343 consideration. The inclusion of LOC1 provided a representation of the population from North-  
344 Central India mainly consuming a carbohydrate and fat rich diet, whereas LOC2 represented a  
345 population from Southern India consuming an omnivorous diet with rice and animal-based  
346 products as the primary components.

347 This study established the gene catalogue of the Indian gut microbiome, which also exemplified  
348 its uniqueness. The genes encoding several transposons, peptidase, glucosidase, and plant  
349 polysaccharide degradation enzymes were unique to the Indian population and not represented in  
350 other microbiome datasets. This catalogue is likely to act as a reference dataset for gut microbiome  
351 studies in South-Asian populations, which have similar dietary habits and lifestyle, and for global  
352 comparative studies. Apart from the basic housekeeping functions of the microbiome, which were  
353 also found abundant in other datasets, the Indian gut microbiome was enriched in functions for  
354 carbohydrate and energy metabolism including degradation of complex polysaccharides, which  
355 corroborates well with the typical carbohydrate-rich diet of the Indian population [24]. The distant  
356 clustering of Indian samples from other populations revealed the unique composition of the Indian

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

357 gut microbiota (**Fig. 1B**). *Prevotella* emerged as the most discriminatory genus associated with the  
358 Indian population, as revealed by both amplicon and MGWAS. Its abundance was also indicated  
359 in the previous 16S rRNA-based microbiome studies of the Indian population, from small to  
360 medium-sized cohorts [21, 40]. Recently, *Prevotella* has been commonly observed in different  
361 non-Western communities who consume a plant-rich diet, such as in the Papua New Guineans,  
362 native Africans, rural Malawians, BaAka pygmies, etc. and has also been associated with  
363 vegetarianism in the Western populations [41-43]. However, it has not been observed at such high  
364 abundance in the western countries so far. The MGWAS approach in this study showed the  
365 presence of *Megasphaera*, *Lactobacillus* and *Mitsuokella* as the other major driver genera  
366 associated with the Indian microbiome.

367 Interestingly, the most abundant genus *Prevotella* in the Indian gut microbiome is a gram-negative  
368 bacterium from the phylum Bacteroidetes that typically releases lipopolysaccharides (LPS), a  
369 constituent of the bacterial outer membrane, from the dead bacterial cells, which can enter the  
370 circulation to elicit an inflammatory response through endotoxemia [44]. Several recent studies  
371 have shown a relationship between the abundance of specific strains of *Prevotella* with  
372 inflammatory diseases, since it has a higher intrinsic capacity to stimulate Th17-mediated  
373 inflammation, which is generally not expected in the strict commensal bacteria [41, 45]. However,  
374 the high abundance of *Prevotella* in the healthy gut microbiome of the Indian population does not  
375 corroborate with its potential inflammatory role reported so far. Further, the species *P. copri*,  
376 which is observed to be the most abundant in this study has been constantly reported to promote  
377 rheumatoid arthritis in different populations, which yet again is inconsistent with its high  
378 abundance in the healthy Indian population [46]. A probable explanation for this emerges from the  
379 understanding that the elicitation of an inflammatory response is mediated by a complex set of

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

380 interactions between host genetic risk factors and environment in which the presence of *Prevotella*  
381 may only be one of the factors [47]. Further, strain-level variations are known in the inflammatory  
382 responses and not all species of *Prevotella* could be potentially inflammatory, as also evident from  
383 the known high genetic diversity within and between the species of *Prevotella* [45]. Taken  
384 together, this description seemingly explains the high abundance of *Prevotella* in the healthy  
385 microbiota despite of its potential inflammatory properties, and emphasizes the requirement for  
386 larger cohort studies in different populations to gain deeper insights into the potential inflammatory  
387 roles of gut microbiome.

388 The abundance of *Prevotella* has been associated with plant-based diets, and the typical  
389 carbohydrate-rich diet of the Indian population could be one of the reasons for the over-  
390 representation of this genus in the Indian gut microbiome [48]. Likewise, the predominance of  
391 other microbial species from genus *Lactobacillus*, *Megasphaera* and *Mitsuokella* could be due to  
392 the higher intake of fermented food and dairy products along with the carbohydrate-rich diet in  
393 LOC1 [33, 48]. Similarly, *Bacteroides* and *Clostridium*, which were abundant in LOC2, are  
394 associated with diets rich in animal-based products, consistent with the omnivorous diet of LOC2  
395 [37]. Interestingly, ET-1 and ET-2 enterotypes showed associations with the two locations LOC1  
396 and LOC2, and also with the two KO-based clusters (C1 and C2) (**Additional File 5: Figure S5**  
397 **and S6**). It is to be noted that C1 was enriched in enzymes involved in the degradation of  
398 carbohydrate and plant polysaccharides, which correlates well with the carbohydrate-rich diet in  
399 LOC1. In contrast, C2 was enriched in enzymes involved in lipid and protein degradation, which  
400 relate to the constituents of an omnivorous diet in LOC2. These observations further support the  
401 correlation between location, diet, and enterotype. Although, the concept of enterotype  
402 classification is sometimes criticised due to statistical weakness in some studies, however, a

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

403 statistically sound classification has the potential to be clinically relevant in various aspects such  
404 as disease diagnosis, early-detection of disease, biomarker development, personalised treatments  
405 and xenobiotic metabolism [19]. It is a representation of the major microbial species in the gut  
406 microbiome, and thus appears useful for microbiome-based population stratification. A robust  
407 statistical analysis with increased sample sizes, direct clinical associations, and detailed molecular  
408 interventions are essential for further strengthening its potential [38].

409 The study also established the previously unknown faecal metabolome of the Indian population,  
410 which showed strong clustering into three metatypes differentiated by location and diet. The  
411 metatypes also correlated well with the respective dietary habits of the two locations, where  
412 Metatype-1 showed an association with LOC1 and was enriched in saturated fatty acids such as  
413 palmitic acid and stearic acid, whereas Metatype-3 showed an association with LOC2, and was  
414 enriched in BCAAs such as isoleucine, valine and leucine, and SCFAs such as propionic acid, and  
415 butyric acid. A medium chain fatty acid (MCFA) 'lauric acid' was also found abundant in LOC2  
416 perhaps due to the high dietary consumption of coconut oil in this location [49, 50]. Lauric acid  
417 has known health benefits such as preventing fat deposition in blood vessels and acting as an anti-  
418 inflammatory and anti-oxidative agent [51].

419 The major BCAA 'isoleucine' being produced through a less common threonine-independent  
420 pathway for isoleucine biosynthesis, and the higher enrichment of the key enzyme, D-citramalate  
421 synthase of the above pathway confirmed its higher abundance in LOC1 as compared to LOC2.  
422 Further, this pathway was found to be associated with a single species, *Prevotella copri* as reported  
423 earlier [36]. Taken together, it appears that at LOC1, the higher abundance of BCAA biosynthesis  
424 genes and a lower abundance of BCAA inward transporters in gut microbiome resulted in the  
425 lower BCAA accumulation in the gut microbiome, leading to a higher absorption and a higher

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

426 BCAA levels in serum, which was also supported by lower abundance of BCAA in faeces (**Fig.**  
427 **7**). However, a contrasting pattern was observed in the case of LOC2, where the lower abundance  
428 of BCAA biosynthesis genes and the higher abundance of BCAA inward transporters correlated  
429 well with the higher and lower BCAA abundances in feces and serum, respectively.

430 The higher levels of SCFAs in LOC2 could be a consequence of the consumption of omnivorous  
431 diet, which is associated with a Firmicute-rich gut microbiome [31]. SCFAs now have well-  
432 established roles in human health as an energy source, an anti-inflammatory agent, and for  
433 improving intestinal homeostasis by increasing IL-18 production [52]. In contrast, higher serum  
434 BCAA levels have well-known roles in promoting insulin resistance and Type-2 Diabetes (T2D),  
435 and were found higher in the serum in LOC1. Several reports on the role of a high-fat diet in the  
436 modulation of microbiota and alteration in intestinal barrier are emerging, which results in the  
437 increased absorption and circulating levels of LPS and branched-chain amino acid (BCAA) and in  
438 the reduction of SCFAs such as butyrate, acetate, propionate, and secondary bile acids, as also  
439 noted in the case of LOC1 [44]. A high-fat and carbohydrate-rich diet have also been associated  
440 with an increase in abundance of Bacteroidetes (gram-negative bacteria), which reduces the  
441 abundance of Firmicutes leading to a skewed Bacteroidetes: Firmicutes ratio towards the former  
442 phylum [33]. Such a ratio was also apparent in this study in LOC1 dominated by *Prevotella* from  
443 the phylum Bacteroidetes [53].

444 Further, a several-fold increased risk of developing T2D has been found with the increase in  
445 circulating BCAA, which were also observed to be higher in LOC1 [36]. In contrast, secondary  
446 bile acids, which can activate glucagon-like peptide-1 (GLP1) secretion and help in protection  
447 against insulin resistance, were high in LOC2 [53]. These results correlate well with the known  
448 higher predisposition of the North-Indian population towards diabetes, cardiovascular diseases and

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

449 hypertension, as compared to Southern India. These observations also provide clues for the  
450 differential metabolic risks in the two populations due to the differences in dietary habits, which  
451 drive their characteristic microbiome. Many of the high-risk components such as trans-fat food  
452 (high-fat dairy, sweets and fried snacks) in North-Indian diets appear to be a reason for the higher  
453 prevalence of cardio-metabolic risk factors such as abdominal adiposity and hypertension, which  
454 are linked to the higher incidences of diabetes and cardiovascular diseases, and could be among  
455 one of the reasons for the shorter life-expectancy as compared to the South-Indian population [54,  
456 55]. These metabolic diseases impose a drastic social, economic and health burden making India  
457 the World's diabetes capital and needs imperative measures for its control. In this scenario, the  
458 data and results from this study provides significant insights on the impact of diet on gut  
459 microbiome, which appears promising in reducing the metabolic risk factors originating through  
460 the interactions between diet and gut microbes to maintain a healthy gut flora, and necessitates the  
461 need for further studies to provide confirmatory evidences for the diet-gut microbiome mediated  
462 metabolic risks between the two populations.

463 This multi-omics based gut microbiome study of a healthy Indian population provides novel  
464 insights into the ecology and biogeography of the human gut microbiome from the poorly  
465 characterized Indian population, and their functional potential as determined by metagenomics and  
466 metabolomics. The comparison of the Indian gut microbiome with other available large-scale gut  
467 microbiome studies reveals the unique microbial community structures in the Indian population  
468 and demonstrates variations in the gut microbiome of Indians due to variation in location and  
469 dietary habits. The study also provides further evidence on the 'diet-gut microbiome-host  
470 metabolism axis' and confirms the notion that the gut microbiome is not just a passive substrate-  
471 degrading system but is actively involved in the host-microbiome crosstalk [56]. Further, the study



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

472 shows that an integrated approach using metabolomics and metagenomics is crucial for the  
473 identification of the repertoire of signals between microbiome and host, and in establishing the  
474 confounding factors for the gut-microbe-metabolic axis. The results from this study are also  
475 prospective to serve as a reference point for future epidemiological studies and translational  
476 applications.

477 **Methods**

478 **Study design and subject enrolment**

479 The study cohort consisted of 110 healthy individuals belonging to different age groups from  
480 infants (<1 year) to aged (>50 years), with an average subject age of  $29.72 \pm 17.4$  years (mean  $\pm$   
481 sd) from two different locations across India i.e., Bhopal (LOC1, n=53) and Kerala (LOC2, n=57),  
482 which are separated by ~1000 miles. LOC1 was located in North-Central India with the majority  
483 of population being vegetarian, whereas LOC2 was located in Southern India where the population  
484 with dietary habits mostly consisting of rice, seafood and red meat (Diet description section in  
485 **Supplementary Table 1**). According to the ‘Indian Food Composition Table’, the primary Indian  
486 diet is rich in carbohydrates such as rice, wheat and potato, and in fat and proteins from milk and  
487 dairy products [55]. In addition, several accompaniments to the primary diet also exist including a  
488 variety of grains, vegetables, fruits, and usage of oil, spices and animal products.

489 The faecal samples for metagenomics and blood samples for serum metabolomics were collected  
490 from healthy participants and their metadata is provided in **Supplementary Data** under the  
491 Metadata information section. The recruitment of volunteers, sample collection, and other study-  
492 related procedures were carried out by following the guidelines and protocols approved by the  
493 Institute Ethics Committee of Indian Institute of Science Education and Research (IISER), Bhopal,

1  
2  
3  
4 494 India. Each faecal sample was frozen within 30 mins of the collection. A written informed consent  
5  
6  
7 495 was obtained from all subjects prior to any study-related procedures, along with information on  
8  
9 496 gender, age, and diet for a period of one month prior to the collection of faecal samples. The  
10  
11  
12 497 recruited individuals did not undergo any medication at least one month prior to the sample  
13  
14 498 collection. All the recruited individuals had an average BMI of 21.16 ( $\pm 5.23$ ), and were not  
15  
16 499 diagnosed with T2D at the time of sample collection, and did not have a second-degree relative  
17  
18  
19 500 history of T2D. The above samples were then used for 16S rRNA V3 hypervariable region  
20  
21 501 amplicon sequencing, shotgun metagenomic sequencing, and metabolomic analysis.  
22  
23

#### 24 25 502 **Faecal metagenomic DNA extraction**

26  
27 503 Metagenomic DNA was isolated from all the faecal samples using QIAamp Stool Mini Kit  
28  
29  
30 504 (Qiagen, CA, USA) according to the manufacturer's instructions. DNA concentration was  
31  
32 505 estimated by Qubit HS dsDNA assay kit (Invitrogen, CA, USA), and quality was estimated by  
33  
34  
35 506 agarose gel electrophoresis. All the DNA samples were stored at  $-80^{\circ}\text{C}$  until sequencing.  
36  
37

#### 38 507 **16S rRNA amplicon and shotgun metagenome sequencing**

39  
40 508 The extracted DNA (5ng) was PCR amplified with seven different custom modified 5'-end  
41  
42  
43 509 adaptor-ligated 341F and 534R primers (See the primer details section in **Supplementary Data**)  
44  
45 510 targeting the V3 hypervariable region of 16S rRNA gene. After evaluating the amplified products  
46  
47  
48 511 on 2% w/v agarose gel, the products were purified using Ampure XP kit (Beckman Coulter, Brea,  
49  
50 512 CA USA). Amplicon libraries were prepared by following the Illumina 16S metagenomic library  
51  
52  
53 513 preparation guide. Metagenomic libraries were prepared using Illumina Nextera XT sample  
54  
55 514 preparation kit (Illumina Inc., USA) by following the manufacturer's protocol. Library size of all  
56  
57  
58 515 the libraries was assessed using Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara,  
59  
60 516 USA.), and quantified on a Qubit 2.0 fluorometer using Qubit dsDNA HS kit (Life technologies,  
61  
62  
63  
64  
65

1  
2  
3  
4 517 USA) and by qPCR using KAPA SYBR FAST qPCR Master mix and Illumina standards and  
5  
6 518 primer premix (KAPA Biosystems, Wilmington, MA, USA) following the Illumina suggested  
7  
8  
9 519 protocol. Both the amplicon and metagenomic libraries were loaded on Illumina NextSeq 500  
10  
11 520 platform using NextSeq 500/550 v2 sequencing reagent kit (Illumina Inc., USA), and 150 bp  
12  
13  
14 521 paired-end sequencing was performed at the Next-Generation Sequencing (NGS) Facility, IISER  
15  
16 522 Bhopal, India.

### 20 523 **Amplicon-based taxonomic analysis**

22 524 A total of 24 Gbps of data were retrieved on de-multiplexing of paired-end reads with an average  
23  
24  
25 525 of 210 Mbp per sample. The paired-end reads were assembled using FLASH and were quality  
26  
27 526 filtered at Q20 (80% bases) Phred quality score, and the primer sequences were trimmed from the  
28  
29  
30 527 High Quality (HQ) reads [57]. The reads were further clustered into OTUs using closed-reference  
31  
32 528 OTU picking protocol of QIIME at  $\geq 97\%$  identity against Greengenes Database v 13\_5 [58, 59].  
33  
34  
35 529 The most abundant read was selected as the representative sequence for each OTU and was  
36  
37 530 assigned with taxonomy using the Greengenes database. OTU table containing the abundance of  
38  
39  
40 531 each OTU for each sample was generated and used for further analysis. For phylogenetic analysis,  
41  
42 532 representative 16S rRNA of phylotypes were aligned against a core set of 16S rRNA gene  
43  
44 533 sequences in Greengenes database using align\_seqs.py with the PyNAST algorithm [60]. The  
45  
46  
47 534 phylogenetic distances between reads were calculated using aligned dataset and were used for the  
48  
49 535 calculation of unweighted UniFrac distances.

### 52 536 **Pre-processing of the Metagenomic reads**

55 537 A total of 150 Gbp of metagenomic sequence data (mean = 1.36 Gb) was generated from 110  
56  
57 538 faecal samples. The metagenomic reads were filtered using NGSQC toolkit with a cutoff  $\geq Q20$   
58  
59  
60 539 [61]. The high-quality reads were further filtered to remove the host-origin reads (human  
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

contamination) from bacterial metagenomic reads, which resulted in the removal of an average of 1% reads. The reads from each sample were assembled into contigs at a k-mer size of 63 bp using SOAPdenovo [62]. The singletons resulting from each sample were pooled together and denovo assembly was repeated on the combined set of singleton reads from all samples. The ORFs from each contig (length  $\geq 300$ bp) were predicted using MetaGeneMark [63]. Pair-wise alignment of genes was performed using BLAT, and the genes which had an identity  $\geq 95\%$  and alignment coverage  $\geq 90\%$  were clustered into a single set of non-redundant genes, from which the longest gene was selected as the representative ORF to construct the non-redundant gene catalog.

Integrated Gene Catalog (IGC), which represents 1,297 human gut metagenomic samples comprising of HMP, MetaHIT and Chinese datasets, was retrieved [23]. The gene catalog constructed from Indian samples was combined with the IGC to construct a non-redundant gene catalog (using identity  $\geq 95\%$  and alignment coverage  $\geq 90\%$ ) and is referred to as ‘updated IGC’ in the subsequent analysis.

**Quantification of gene content**

The quantification of gene content was carried out using the strategy performed by Qin et al., [7] where the high-quality reads were aligned against the updated IGC using SOAP2 in SOAP aligner with an identity cut off  $\geq 90\%$  [64]. Two types of alignments were considered for sequence-based profiling:

- (1) The entire paired-end read mapped to the gene.
- (2) One end of paired-end read mapped to a gene and other end remained unmapped.

In both cases, the mapped read was counted as one copy. Further, the read count was normalized based on length of the gene as:  $b_i = \frac{x_i}{L_i}$

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

The relative abundance of a gene within the sample was calculated as:  $a_i = \frac{b_i}{\sum_j b_j} = \frac{x_i}{\sum_j x_j}$

$a_i$ : relative abundance of gene in sample S;  $x_i$ : The times in which gene i was detected in sample S (the number of mapped reads);  $L_i$ : length of gene i;  $b_i$ : copy number of gene i in sequenced data from sample S.

### Phylogenetic assignment of reads

A total of 4,097 reference microbial genomes were obtained from Human Microbiome Project (HMP) and National Centre for Biotechnology Information (NCBI) on 5<sup>th</sup> December 2015. The databases were independently indexed into two Bowtie indexes using Bowtie-2 [65]. The metagenomic reads were aligned to the reference microbial genomes using Bowtie-2. The mapped reads from both indexes were merged by selecting the alignment having the higher identity ( $\geq 90\%$  identity). The percent identity was calculated using the formula: %identity =  $100 * (\text{matches} / \text{total aligned length})$ . The normalized abundance of a microbial genome was calculated by summing the total number of reads aligned to its reference genome, normalized by the genome length and the total number of reads in the dataset. For reads showing hits to both indexed databases with equal identity, each genome was assigned 0.5 read count. The relative abundance of each genome was calculated by adding the normalized abundance of each genome divided by the total abundance. The Calinski Harabasz index (CHI) was used to calculate the variance between the clusters compared to the variance within clusters [2].

### Construction of common core microbial functions

To identify the core microbial functions in the gut microbiome of Indian populations and to understand their abundance compared to the other populations, the core microbiome was constructed using a similar strategy as mentioned in MetaHIT [2]. However, to construct a

1  
2  
3  
4 584 comprehensive core functional microbiome, the information of essential functions from six  
5  
6 585 different microbes including two strains of *Escherichia coli*, *Bacteroides thetaiotaomicron*,  
7  
8  
9 586 *Pseudomonas aeruginosa*, *Salmonella enteric* and *Staphylococcus aureus*, was used instead of  
10  
11 587 considering a single microorganism. The list of essential genes was collected from DEG database  
12  
13  
14 588 v [66]. 1,890 genes were identified as essential genes in all the six microorganisms. The core gut  
15  
16 589 microbiome functions were also calculated using the above strategy for the USA, Denmark and  
17  
18  
19 590 Chinese population gut microbial samples to remove the variations arising due to differences in  
20  
21 591 data analysis procedures. Apart from identifying the clusters that represented  $\geq 85\%$  genes within  
22  
23  
24 592 the range of essential gene functions, the low prevalent eggNOG functions, which were present in  
25  
26 593  $\geq 0.0001\%$  abundance in  $\geq 80\%$  of samples in that population, were further filtered out. This added  
27  
28  
29 594 filtration step helped in removing all the low abundant functions. To represent the core, the  
30  
31 595 variance of these functions was also calculated between the two Indian locations. The eggNOGs  
32  
33  
34 596 which showed significant deviations in variations ( $P\text{-value} \leq 0.05$ ; Levene's test) were further  
35  
36 597 filtered out from the analysis.

### 39 598 **Construction of Metagenomic Species for MGWAS**

40  
41  
42 599 To identify metagenomic markers using a non-reference based approach on metagenomic samples,  
43  
44 600 a metagenome-wide association study was performed for 340 samples (age and gender matched)  
45  
46  
47 601 including India (both locations), USA, China and Denmark populations. The genes present in at  
48  
49 602 least  $\geq 10\%$  of samples were considered and clustered using the canopy-mgs algorithm as described  
50  
51  
52 603 [7]. The genes having Pearson's correlation coefficient ( $\geq 0.9$ ) were clustered into CAGs.  
53  
54 604 Furthermore, the genes for which  $\geq 90\%$  abundance was obtained from a single sample were  
55  
56 605 discarded.

1  
2  
3  
4 606 To determine the taxonomic origin of each MGS/CAG (metagenomic cluster), all the genes were  
5  
6 607 aligned against reference microbial genomes of 4,097 genomes from HMP and NCBI at nucleotide  
7  
8  
9 608 level using BLASTN. The alignment hits were filtered using an E-value  $\leq 10^{-6}$  and alignment  
10  
11 609 coverage  $\geq 80\%$  of the gene length, and 2,687,688 genes showed alignments against the reference  
12  
13  
14 610 genomes. The remaining genes were aligned against UNIREF database (UniRef 50) at protein  
15  
16 611 sequences [67]. The multiple best hits with equal identity and scores were further assigned  
17  
18  
19 612 taxonomy based on LCA (Lowest Common Ancestor) method. The genes were finally assigned to  
20  
21 613 taxa based on comprehensive parameters of sequence similarity across phylogenetic ranks as  
22  
23  
24 614 described earlier [68]. The identity threshold of  $\geq 95\%$  was used for assignment up to species level,  
25  
26 615  $\geq 85\%$  identity threshold for assignment up to genus level, and  $\geq 65\%$  identity was used for phylum  
27  
28  
29 616 level assignment using BLASTN. The taxonomic assignments of MGS/CAGs were performed  
30  
31 617 with the criteria that  $\geq 50\%$  genes in each MGS should map to the same lowest phylogenetic group.  
32  
33  
34 618 So if a particular species is assigned  $\geq 50\%$  genes out of total the assignment will be carried out at  
35  
36 619 species level rather than at genus or higher orders. The relative abundance of MGS/CAGs in each  
37  
38 620 sample was estimated by using relative abundance values of all genes from that MGS/CAG. A  
39  
40  
41 621 Poisson distribution was fitted to the relative abundance values of the data. The mean estimated  
42  
43 622 from Poisson distribution was assigned as the relative abundance of that MGS. The profile of  
44  
45  
46 623 MGS/CAGs were generated and used for further analysis.

#### 49 624 **Faecal and Serum metabolomic sample preparation and derivatization**

50  
51 625 Lyophilized faecal samples were used to achieve better metabolite coverage, as described  
52  
53  
54 626 previously [69]. Metabolites were extracted with 1 mL of ice-cold methanol: water (8:2) from 80  
55  
56 627 mg of lyophilized samples in a bath ultrasonicator (Bioruptor<sup>TM</sup> UCD-200, Diagenode, USA) at  
57  
58  
59 628 4°C for 30 min followed by 2 min of vortexing. The supernatant was extracted by centrifugation  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 629 at 18,000 g for 15 min at 4°C and dried at 50°C under a gentle stream of nitrogen gas. To remove  
5  
6 630 the residual water molecules from the samples, 100uL of toluene was added to the dry residue and  
7  
8  
9 631 evaporated completely at 50°C under nitrogen gas. Dry extracted metabolites were first derivatized  
10  
11 632 with 50 uL of methoxyamine hydrochloride (MOX) in pyridine (20 mg/mL) at 60°C for 2 hours,  
12  
13  
14 633 and the second derivatization was performed with 100 uL of MSTFA in 1% TMCS at 60°C for 45  
15  
16 634 min to form trimethylsilyl (TMS) derivatives. Finally, 150 uL of the TMS derivatives was  
17  
18  
19 635 transferred into a GC glass vial inserts and subjected to GC/TOFMS analysis. Serum samples were  
20  
21 636 prepared (polar metabolites only) and derivatized as described by Psychogium et al., 2011 [70].  
22  
23

#### 24 637 **Method development and validation**

25  
26  
27 638 Matrix dilution approach was used for validating the linearity and range of dilution [69]. Pooled  
28  
29  
30 639 faecal samples were used to create the reference peaks to validate the peaks coming from  
31  
32 640 individual samples, which were needed due to the presence of a relatively high abundance of faecal  
33  
34  
35 641 metabolites in the pooled samples. The supernatant of feces after extraction was serially diluted 2,  
36  
37 642 5, 10, 50, 100, 200 and 500 times with methanol: water (8:2). At dilution 2, the maximum numbers  
38  
39 643 of peaks were seen and were processed with the same dilution factor for all the samples. A total of  
40  
41  
42 644 30 chemical standards mixture and the pooled faecal samples were used to validate the method.  
43  
44 645 Each stock solution of test standard was carefully prepared in deionized water or with pure ethanol  
45  
46  
47 646 (50,150 350, 500 um) for the determination of linear range, regression coefficient (R<sup>2</sup>), limit of  
48  
49 647 detection (LOD), and repeatability. L-norvaline (1, 2.5, 5, 10, 20 mg/ml in ethanol) was used as a  
50  
51  
52 648 spiked external standard for the optimized derivatization of the method.  
53

#### 54 649 **GC-MS analysis**

55  
56  
57 650 GC-MS was performed on an in-house Agilent 7890A gas chromatograph with 5975C MS system.  
58  
59  
60 651 An HP-5 (25 m × 320 um × 0.25 um i.d.) fused silica capillary column (Agilent J&W Scientific,  
61  
62  
63  
64  
65



1  
2  
3  
4 652 Folsom, CA), was used with the open split interface. The injector, transfer line and ion source  
5  
6 653 temperatures were maintained at 220, 220 and 250 °C, respectively. Oven temperature was  
7  
8  
9 654 programmed at 70°C for 0.2 min, and increased at 10°C/min to 270°C where it was sustained for  
10  
11 655 5 min, and further increased at 40°C/min to 310°C where it was held for 11 minutes. The MS was  
12  
13  
14 656 operated in the electron impact ionization mode at 70eV. Mass data were acquired in full scan  
15  
16 657 mode from m/z 40 to 600 with an acquisition rate of 20 spectra per second. To detect retention  
17  
18  
19 658 time shifts and enable Kovats retention index (RI) calculation, a standard Alkane series mixture  
20  
21 659 (C10–C40) was injected periodically during the sample analysis. RIs are relative retention times  
22  
23  
24 660 normalized to n-alkanes eluted adjacently. For serum samples, we used 2uL aliquot with a split  
25  
26 661 ratio of 4:1 on the same column as described above. The injector port temperature was held at  
27  
28  
29 662 250°C, and the helium gas flow rate was set to 1mL/min at an initial oven temperature of 50°C.  
30  
31 663 The oven temperature was increased at 10°C/min to 310°C for 11min and mass data were acquired  
32  
33 664 in full scan mode from m/z 40 to 600 with an acquisition rate of 20 spectra per second.

### 36 665 **Metabolomic analysis and metabolite profile generation**

37  
38  
39 666 Raw CDF files were used for peak identification and filtering and the XCMS package in R were  
40  
41  
42 667 used for pre-processing of the peaks. First, the parameters used for pre-processing of the reads  
43  
44 668 were optimized by calculating the reliability index using the formula given below:

45  
46  
47 669 Reliability index = (number of reliable peaks)<sup>2</sup>/number of unreliable peaks.

48  
49  
50 670 The reliable peaks were identified for each of the settings such as fwhm, S/N and bw, with a  
51  
52 671 predefined range of values and regression coefficient was calculated for dilutions of QC samples.  
53  
54  
55 672 The number of peaks with a high coefficient of determination ( $R^2 \geq 0.9$ ) were considered reliable,  
56  
57 673 whereas the peaks with very low  $R^2 (\leq 0.05)$  were considered unreliable peaks[71]. The finally  
58  
59 674 optimized parameters were: profmethod = bin, method = matched Filter, fwhm =8 and 5 for  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 675 faecal and serum samples, respectively, and S/N = 12 and 3 for faecal and serum samples,  
5  
6  
7 676 respectively, bw =5 (for first grouping), smooth = linear, family = gaussian, extra = 1, plot type  
8  
9 677 = mdevden, missing =8, bw = 3 (for second grouping). Further, in order to compare across  
10  
11  
12 678 multiple samples, the peak intensities were normalized (root transformed) and scaled using z-  
13  
14 679 transformation. These normalized and scaled peak intensities were used for further statistical  
15  
16 680 analysis.

17  
18  
19 681 A multivariate statistical method, Orthogonal Projections to Latent Structures Discriminant  
20  
21 682 Analysis (OPLS-DA), was used to identify differences between LOC-1 samples (n=53) and  
22  
23  
24 683 LOC-2 (n=55) samples. Metabolites driving the differences were identified in metabolic  
25  
26 684 profiles of LOC-1 and LOC-2 samples using correlations coefficients. The clusters of co-  
27  
28  
29 685 abundant metabolite profiles were identified using R package "WGCNA". Signed weighted  
30  
31 686 metabolite co-abundance correlation after scaling and centering was calculated across all  
32  
33  
34 687 samples. The soft threshold of  $\beta = 15$  was chosen for scale-free topology. The dynamic hybrid  
35  
36 688 tree cutting algorithm was used to identify the clusters with a deepsplit = 4 and minimum cluster  
37  
38 689 size = 4. The profile of each faecal metabolite cluster was summarized using eigenvector. The  
39  
40  
41 690 abundance profile of each cluster of metabolites (MES) was calculated using the same  
42  
43 691 methodology as used for MGS cluster abundance profiles.

#### 44 45 46 692 **Retention index (RI) calculation**

47  
48  
49 693 GC-MS data obtained from the alkene series run was used to calculate the RI for each peak in  
50  
51  
52 694 the samples, and the obtained RI values were further used at the time of library search for the  
53  
54 695 identification of individual metabolite.

$$55  
56  
57 696 \quad I = 100 X [n + (\log tx - \log tn) / (\log tn + 1 - \log tn)]$$

58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 697 Where,  $t_x$  = retention time of the peak,  $t_n$  = retention time of preceding alkane, and  $t_{n+1}$  =  
5  
6 698 retention time of the following alkane.  
7  
8  
9

### 10 699 **Enterotype Analysis**

11  
12 700 Enterotypes in the dataset were identified from the relative abundance profiles of Genus or  
13  
14  
15 701 Orthologous groups (OG) in the samples. The Jensen-Shannon distances (which estimates the  
16  
17 702 probability distributions) between the samples were calculated and the abundance profiles were  
18  
19  
20 703 clustered using PAM (partitioning around medoids) clustering algorithm as mentioned previously  
21  
22 704 [72]. The optimal number of clusters was assessed using CHI that has shown good performance in  
23  
24  
25 705 recovering the optimal number of clusters. Similarly, the prediction strength was also employed  
26  
27 706 as another metric for cluster validation. Both the CHI and prediction strength showed quite  
28  
29  
30 707 significantly correlated results. For clustering, CHI and prediction strength gave non-identical  
31  
32 708 values, silhouette index was calculated to estimate the robustness of clusters.  
33  
34

### 35 709 **Between class analysis**

36  
37  
38 710 Between class analysis was performed to identify the drivers and support the clustering of the  
39  
40 711 genus/species/OG abundance profiles into enterotypes. The instrumental variables were the  
41  
42 712 enterotype classification and the top species, which contributed the maximum to the principal  
43  
44  
45 713 components obtained from between class analysis, and were identified as driver species/genus/OG  
46  
47 714 based on their factor scores.  
48  
49  
50

### 51 715 **Diversity Analysis**

52  
53 716 The within-sample diversity metrics such as a number of observed species, Shannon index, and  
54  
55  
56 717 Phylogenetic distance were calculated for each rarefied sample (at fixed or varying depths) and  
57  
58 718 were compared to different types of samples. The beta diversity (between the samples) was  
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

719 calculated using unweighted UniFrac distances between the samples for rarefied OTU tables. The  
720 effect of covariates such as age, enterotypes, diet, geography and gender were compared for  
721 correlation with principal components identified from principal component analysis using UniFrac  
722 distances. The polyserial correlations with P-values were calculated for categorical variables and  
723 the significance of the covariates for explaining the variation was estimated at each principal  
724 component.

**725 Network Analysis**

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

**735 Supervised learning**

736 Predictive models were built using supervised machine learning algorithm Random Forest (RF).  
737 The models were optimized using 10,000 trees and default settings of mtry (number for variables  
738 used to build the model). The mean three-fold cross-validation error rates were calculated for each  
739 of the binary tree and the ensemble of trees. The mean decrease in accuracy, which is the increase  
740 in error rates on leaving the variable out, was calculated for each prediction and tree and was used  
741 to estimate the importance score. The variables showing a higher mean decrease in accuracy of

1  
2  
3  
4 742 prediction were considered important for the segregation of the datasets into groups based on the  
5  
6 743 categorical variable.  
7  
8  
9

## 10 744 **Statistical Analysis**

11  
12 745 All the statistical comparisons between groups were performed using non-parametric Wilcoxon  
13  
14  
15 746 Rank Sum Test with FDR Adjusted P-Values to control for multiple comparisons. The correlations  
16  
17 747 between two variables and the correlations within were calculated using Spearman's Correlation  
18  
19  
20 748 Coefficient with Adjusted P-Values. The correlations between categorical and numeric variables  
21  
22 749 were performed using Polyserial correlation/biserial correlations. To identify the enrichment of  
23  
24  
25 750 enzymes/species associated with a host, Odds Ratio was used as a measure of the enrichment of  
26  
27 751 an enzyme in a host. The Odds Ratio was calculated as  $OR(k) = \frac{[\sum_{s=LOC1} A_{sk} / \sum_{s=LOC1} (\sum_{i \neq k} A_{si})]}{[\sum_{s=LOC2} A_{sk} / \sum_{s=LOC2} (\sum_{i \neq k} A_{si})]}$ , where  $A_{sk}$  denotes abundance of enzyme  $k$  in sample  $S$ . Apart  
28  
29  
30 752 from that, Reporter features algorithm was used for gene-set analysis of significant pathways  
31  
32 753 associated with different groups of samples. The algorithm takes the adjusted P-values and folds  
33  
34  
35 754 changes (log odds ratio) as input for each KO. The gene statistic is calculated based on the  
36  
37 755 significant association of KO and its direction of change through which the pathway is scored by  
38  
39 756 calculating the global P-value. All the graphs and plots were generated using the ggplot2 package  
40  
41  
42 757 in R.  
43  
44 758

## 47 759 **Correlation analysis between functional modules and metabolite clusters**

48  
49  
50 760 To calculate the association of microbial functional modules with faecal metabolite clusters, the  
51  
52 761 Spearman's correlation coefficients were calculated to rank KOs for association with metabolite  
53  
54  
55 762 clusters and Metabotypes. To quantify the shift in Spearman correlation between given KEGG  
56  
57 763 module and the metabolite cluster compared to the background distribution, the background  
58  
59  
60 764 adjusted median Spearman's correlation was calculated for a given KEGG module  $m$  as:  
61  
62  
63  
64  
65

1  
2  
3  
4 765  $SCC_{bg,adj} = \text{median}(SCC_{KOs \in \text{KEGG Module } m}) - \text{median}(SCC_{KOs \text{ KEGG Module } m})$

7 766 Where  $SCC_{KO}$  is the partial Spearman's correlation coefficient between KO and the metabolite  
8  
9 767 cluster.

13 768 Identification of microbial species driving the association between KEGG Module and metabolite  
14  
15 769 abundance was done by iterating the correlation between KO belonging to the KEGG module and  
16  
17  
18 770 the metabolite after excluding the genes annotated to that KO from each species. The change in  
19  
20 771 median Spearman's correlation coefficient between the KOs and the metabolite, when genes from  
21  
22 772 that species are excluded from the analysis, was calculated as described previously [36]. The  
23  
24  
25 773 species showing the maximum change in the overall correlation of module with metatype was  
26  
27 774 plotted.

#### 31 775 **List of abbreviations**

33 776 Indian Gut Microbiome (IGM), Enterotypes (ET), Integrated Gene Catalog (IGC), Metagenome-  
34  
35 777 Wide Association Study (MGWAS), Short Chain Fatty Acids (SCFAs), Branched Chain Amino  
36  
37  
38 778 Acids (BCAAs).

#### 41 779 **Declarations**

#### 44 780 **Collection of Datasets for Comparative analysis**

47 781 The 74 HMP metagenomes were collected from <http://hmpdacc.org/HMASM> or NCBI SRA  
48  
49 782 (accession SRR059347). The 85 Danish fecal metagenomes from METAHIT were obtained from  
50  
51 783 European Nucleotide Archive (<http://www.ebi.ac.uk/ena>, study accession number ERP000108).  
52  
53  
54 784 The 71 Chinese metagenome samples were obtained from NCBI SRA (accession number –  
55  
56 785 SRR341581).

#### 59 786 **Ethics approval and consent to participate**

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

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

**791 Consent for publication**

792 Not applicable

**793 Availability of data and materials**

794 The datasets generated and/or analysed during the current study have been deposited in the  
795 National Centre for Biotechnology Information (NCBI) BioProject database under the project  
796 number PRJNA397112 and will be made publicly available on publication or on request at the  
797 time of peer review.

**798 Competing interests**

799 The authors declare that they have no competing interests.

**800 Funding**

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

**803 Author's contributions**

804 VKS and AM conceived the work and participated in the design of the study. AM designed the  
805 study protocol. AM and JP collected all the samples in collaboration with TG. AM performed the  
806 all sample processing, DNA extraction, metabolite extraction and profiling from faecal and blood  
807 samples. RS and AM carried out the library preparation and sequencing work. DBD carried out all  
808 metagenomic data and statistical analysis. AKS and DBD analyzed the metabolomics data. AM

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

809 and DBD did the primary data interpretation of analytical outcomes under the supervision of VKS.  
810 AM, DBD, RS, AG, JS, KRA and VKS drafted the manuscript. All authors read and approved the  
811 final manuscript.

**812 Acknowledgments**

813 We thank MHRD, Govt of India and Centre for Research on Environment and Sustainable  
814 Technologies (CREST) at IISER Bhopal for providing financial support. However, the views  
815 expressed in this manuscript are that of the authors alone and no approval of the same, explicit or  
816 implicit, by MHRD should be assumed. The sequencing and computational analysis were  
817 performed at the NGS Facility and HPC and computing facility, respectively, at IISER Bhopal.  
818 DBD, AM, RS and JP received fellowships from the UGC (University Grants Commission),  
819 Centre for Research on Environment and Sustainable Technologies (CREST, IISER Bhopal),  
820 DST-INSPIRE and Central University of Kerala, respectively.

821



1  
2  
3  
4 **822**    **References:**  
5  
6

- 7 823    1.    Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R and Gordon JI. The human  
8 824    microbiome project. *Nature*. 2007;449 7164:804.  
9 825    2.    Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, et al. Enterotypes of the  
10 826    human gut microbiome. *nature*. 2011;473 7346:174.  
11 827    3.    Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, et al. Human gut  
12 828    microbiome viewed across age and geography. *nature*. 2012;486 7402:222.  
13 829    4.    Dominguez-Bello MG, Costello EK, Contreras M, Magris M, Hidalgo G, Fierer N, et al. Delivery  
14 830    mode shapes the acquisition and structure of the initial microbiota across multiple body habitats  
15 831    in newborns. *Proceedings of the National Academy of Sciences*. 2010;107 26:11971-5.  
16 832    5.    Saxena R and Sharma VK. A Metagenomic Insight Into the Human Microbiome: Its Implications in  
17 833    Health and Disease. In: Kumar D and Antonarakis S, editors. *Medical and Health Genomics*. Mica  
18 834    Haley, Academic Press, Elsevier; 2016. p. 107-19.  
19 835    6.    Schwiertz A, Taras D, Schäfer K, Beijer S, Bos NA, Donus C, et al. Microbiota and SCFA in lean and  
20 836    overweight healthy subjects. *Obesity*. 2010;18 1:190-5.  
21 837    7.    Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, et al. A metagenome-wide association study of gut microbiota  
22 838    in type 2 diabetes. *Nature*. 2012;490 7418:55.  
23 839    8.    Pedersen HK, Gudmundsdottir V, Nielsen HB, Hyotylainen T, Nielsen T, Jensen BA, et al. Human  
24 840    gut microbes impact host serum metabolome and insulin sensitivity. *Nature*. 2016;535 7612:376-  
25 841    81.  
26 842    9.    Gupta VK, Paul S and Dutta C. Geography, ethnicity or subsistence-specific variations in human  
27 843    microbiome composition and diversity. *Frontiers in microbiology*. 2017;8:1162.  
28 844    10.    Bhute S, Pande P, Shetty SA, Shelar R, Mane S, Kumbhare SV, et al. Molecular characterization and  
29 845    meta-analysis of gut microbial communities illustrate enrichment of *Prevotella* and *Megasphaera*  
30 846    in Indian subjects. *Frontiers in microbiology*. 2016;7:660.  
31 847    11.    Gomez A, Petrzelkova KJ, Burns MB, Yeoman CJ, Amato KR, Vlckova K, et al. Gut Microbiome of  
32 848    Coexisting BaAka Pygmies and Bantu Reflects Gradients of Traditional Subsistence Patterns. *Cell*  
33 849    reports. 2016;14 9:2142-53. doi:10.1016/j.celrep.2016.02.013.  
34 850    12.    Obregon-Tito AJ, Tito RY, Metcalf J, Sankaranarayanan K, Clemente JC, Ursell LK, et al. Subsistence  
35 851    strategies in traditional societies distinguish gut microbiomes. *Nature communications*.  
36 852    2015;6:6505. doi:10.1038/ncomms7505.  
37 853    13.    Häsler R, Kautz C, Rehman A, Podschun R, Gassling V, Brzoska P, et al. The antibiotic resistome  
38 854    and microbiota landscape of refugees from Syria, Iraq and Afghanistan in Germany. *Microbiome*.  
39 855    2018;6 1:37.  
40 856    14.    Pulikkan J, Maji A, Dhakan DB, Saxena R, Mohan B, Anto MM, et al. Gut Microbial Dysbiosis in  
41 857    Indian Children with Autism Spectrum Disorders. *Microbial ecology*.1-13.  
42 858    15.    Mohan V, Sandeep S, Deepa R, Shah B and Varghese C. Epidemiology of type 2 diabetes: Indian  
43 859    scenario. *Indian journal of medical research*. 2007;125 3:217.  
44 860    16.    Organization WH. Waist circumference and waist-hip ratio: report of a WHO expert consultation,  
45 861    Geneva, 8-11 December 2008. 2011.  
46 862    17.    Schmidt TS, Raes J and Bork P. The human gut microbiome: from association to modulation. *Cell*.  
47 863    2018;172 6:1198-215.  
48 864    18.    Maji A, Misra R, Dhakan DB, Gupta V, Mahato NK, Saxena R, et al. Gut microbiome contributes to  
49 865    impairment of immunity in pulmonary tuberculosis patients by alteration of butyrate and  
50 866    propionate producers. *Environmental Microbiology*. 2017.  
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

867 19. Costea PI, Hildebrand F, Arumugam M, Backhed F, Blaser MJ, Bushman FD, et al. Enterotypes in  
868 the landscape of gut microbial community composition. *Nature microbiology*. 2018;3 1:8-16.  
869 doi:10.1038/s41564-017-0072-8.

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

872 21. Tandon D, Haque MM, R S, Shaikh S, P S, Dubey AK, et al. A snapshot of gut microbiota of an adult  
873 urban population from Western region of India. *PLoS One*. 2018;13 4:e0195643.  
874 doi:10.1371/journal.pone.0195643.

875 22. Suryanarayana M, Agrawal A and Prabhu KS. Inequality-adjusted human development index for  
876 India's states. New Delhi, India: United Nations Development Programme. 2011.

877 23. Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, et al. An integrated catalog of reference genes in  
878 the human gut microbiome. *Nature biotechnology*. 2014;32 8:834.

879 24. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene  
880 catalogue established by metagenomic sequencing. *Nature*. 2010;464 7285:59-65.  
881 doi:10.1038/nature08821.

882 25. Wang J and Jia H. Metagenome-wide association studies: fine-mining the microbiome. *Nature*  
883 *Reviews Microbiology*. 2016;14 8:508.

884 26. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene  
885 catalogue established by metagenomic sequencing. *nature*. 2010;464 7285:59.

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

889 28. Liang C, Tseng H-C, Chen H-M, Wang W-C, Chiu C-M, Chang J-Y, et al. Diversity and enterotype in  
890 gut bacterial community of adults in Taiwan. *BMC genomics*. 2017;18 1:932.

891 29. Aleksandrowicz L, Tak M, Green R, Kinra S and Haines A. Comparison of food consumption in  
892 Indian adults between national and sub-national dietary data sources. *British Journal of Nutrition*.  
893 2017;117 7:1013-9.

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

897 31. Ríos-Covián D, Ruas-Madiedo P, Margolles A, Gueimonde M, de los Reyes-Gavilán CG and Salazar  
898 N. Intestinal short chain fatty acids and their link with diet and human health. *Frontiers in*  
899 *microbiology*. 2016;7:185.

900 32. Kovatcheva-Datchary P, Nilsson A, Akrami R, Lee YS, De Vadder F, Arora T, et al. Dietary fiber-  
901 induced improvement in glucose metabolism is associated with increased abundance of  
902 *Prevotella*. *Cell metabolism*. 2015;22 6:971-82.

903 33. De Filippo C, Cavalieri D, Di Paola M, Ramazzotti M, Poullet JB, Massart S, et al. Impact of diet in  
904 shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa.  
905 *Proceedings of the National Academy of Sciences*. 2010;107 33:14691-6.

906 34. Ruiz-Capillas C and Jimenez-Colmenero F. Biogenic amines in meat and meat products. *Critical*  
907 *reviews in food science and nutrition*. 2004;44 7-8:489-99.

908 35. Layman DK. The role of leucine in weight loss diets and glucose homeostasis. *The Journal of*  
909 *nutrition*. 2003;133 1:261S-7S.

910 36. Pedersen HK, Gudmundsdottir V, Nielsen HB, Hyotylainen T, Nielsen T, Jensen BA, et al. Human  
911 gut microbes impact host serum metabolome and insulin sensitivity. *Nature*. 2016;535 7612:376.

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

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

38. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, et al. Enterotypes of the human gut microbiome. *Nature*. 2011;473 7346:174-80. doi:10.1038/nature09944.

39. Cho I and Blaser MJ. The human microbiome: at the interface of health and disease. *Nat Rev Genet*. 2012;13 4:260-70. doi:10.1038/nrg3182.

40. Bhute S, Pande P, Shetty SA, Shelar R, Mane S, Kumbhare SV, et al. Molecular characterization and meta-analysis of gut microbial communities illustrate enrichment of *Prevotella* and *Megasphaera* in Indian Subjects. *Frontiers in microbiology*. 2016;7.

41. Ley RE. Gut microbiota in 2015: *Prevotella* in the gut: choose carefully. *Nature reviews Gastroenterology & hepatology*. 2016;13 2:69-70. doi:10.1038/nrgastro.2016.4.

42. 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. doi:10.1016/j.cmet.2015.10.001.

43. Martinez I, Stegen JC, Maldonado-Gomez MX, Eren AM, Siba PM, Greenhill AR, et al. The gut microbiota of rural papua new guineans: composition, diversity patterns, and ecological processes. *Cell Rep*. 2015;11 4:527-38. doi:10.1016/j.celrep.2015.03.049.

44. Hawkesworth S, Moore S, Fulford A, Barclay G, Darboe A, Mark H, et al. Evidence for metabolic endotoxemia in obese and diabetic Gambian women. *Nutrition & diabetes*. 2013;3 8:e83.

45. Larsen JM. The immune response to *Prevotella* bacteria in chronic inflammatory disease. *Immunology*. 2017;151 4:363-74. doi:10.1111/imm.12760.

46. Scher JU, Sczesnak A, Longman RS, Segata N, Ubeda C, Bielski C, et al. Expansion of intestinal *Prevotella copri* correlates with enhanced susceptibility to arthritis. *eLife*. 2013;2:e01202. doi:10.7554/eLife.01202.

47. Renz H, von Mutius E, Brandtzaeg P, Cookson WO, Autenrieth IB and Haller D. Gene-environment interactions in chronic inflammatory disease. *Nat Immunol*. 2011;12 4:273-7. doi:10.1038/ni0411-273.

48. Tremaroli V and Bäckhed F. Functional interactions between the gut microbiota and host metabolism. *Nature*. 2012;489 7415:242.

49. Jaarin K, Norliana M, Kamisah Y, Nursyafiza M and Qodriyah HMS. Potential role of virgin coconut oil in reducing cardiovascular risk factors. *Exp Clin Cardiol*. 2014;20 8:3399-410.

50. Boemeke L, Marcadenti A, Busnello FM and Gottschall CBA. Effects of coconut oil on human health. *Open Journal of Endocrine and Metabolic Diseases*. 2015;5 07:84.

51. Intahphuak S, Khonsung P and Panthong A. Anti-inflammatory, analgesic, and antipyretic activities of virgin coconut oil. *Pharmaceutical biology*. 2010;48 2:151-7.

52. Nicholson JK, Holmes E, Kinross J, Burcelin R, Gibson G, Jia W, et al. Host-gut microbiota metabolic interactions. *Science*. 2012:1223813.

53. Boulangé CL, Neves AL, Chilloux J, Nicholson JK and Dumas M-E. Impact of the gut microbiota on inflammation, obesity, and metabolic disease. *Genome medicine*. 2016;8 1:42.

54. Joshi SR and Parikh RM. India; the diabetes capital of the world: Now heading towards hypertension. *Journal-Association of Physicians of India*. 2007;55 Y:323.

55. Longvah T, Ananthan R, Bhaskarachary K and Venkaiah K. Indian food composition tables. National Institute of Nutrition, Indian Council of Medical Research, Department of Health Research, Ministry of Health and Family Welfare, Government of India (505 pp). 2017.

56. Li X, Shimizu Y and Kimura I. Gut microbial metabolite short-chain fatty acids and obesity. *Bioscience of microbiota, food and health*. 2017;36 4:135-40. doi:10.12938/bmfh.17-010.

57. Magoč T and Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*. 2011;27 21:2957-63.

58. Ritari J, Salojärvi J, Lahti L and de Vos WM. Improved taxonomic assignment of human intestinal 16S rRNA sequences by a dedicated reference database. *BMC genomics*. 2015;16 1:1056.

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

59. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. *Nature methods*. 2010;7 5:335.

60. Caporaso JG, Bittinger K, Bushman FD, DeSantis TZ, Andersen GL and Knight R. PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics*. 2009;26 2:266-7.

61. Patel RK and Jain M. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PloS one*. 2012;7 2:e30619.

62. Li R, Li Y, Kristiansen K and Wang J. SOAP: short oligonucleotide alignment program. *Bioinformatics*. 2008;24 5:713-4.

63. Zhu W, Lomsadze A and Borodovsky M. Ab initio gene identification in metagenomic sequences. *Nucleic acids research*. 2010;38 12:e132-e.

64. Li R, Yu C, Li Y, Lam T-W, Yiu S-M, Kristiansen K, et al. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*. 2009;25 15:1966-7.

65. Langmead B and Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nature methods*. 2012;9 4:357.

66. Zhang R and Lin Y. DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. *Nucleic acids research*. 2008;37 suppl\_1:D455-D8.

67. Suzek BE, Huang H, McGarvey P, Mazumder R and Wu CH. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*. 2007;23 10:1282-8.

68. Huson DH, Auch AF, Qi J and Schuster SC. MEGAN analysis of metagenomic data. *Genome research*. 2007;17 3:377-86.

69. Phua LC, Koh PK, Cheah PY, Ho HK and Chan ECY. Global gas chromatography/time-of-flight mass spectrometry (GC/TOFMS)-based metabonomic profiling of lyophilized human feces. *Journal of Chromatography B*. 2013;937:103-13.

70. Psychogios N, Hau DD, Peng J, Guo AC, Mandal R, Bouatra S, et al. The human serum metabolome. *PloS one*. 2011;6 2:e16957.

71. Gao X, Pujos-Guillot E, Martin J-F, Galan P, Juste C, Jia W, et al. Metabolite analysis of human fecal water by gas chromatography/mass spectrometry with ethyl chloroformate derivatization. *Analytical biochemistry*. 2009;393 2:163-75.

72. Kaufman L and Rousseeuw PJ. Partitioning around medoids (program pam). *Finding groups in data: an introduction to cluster analysis*. 1990:68-125.

992  
993

1  
2  
3  
4 994 **Table 1. Metagenomic datasets used for comparative analysis (Meta-analysis) of the**  
5  
6  
7 995 **microbiome and MGWAS.**  
8  
9

<b>Dataset</b>	<b>No. of samples</b>	<b>Amount of data</b>	<b>No. of genes</b>
<b>INDIA</b>	110	110Gb	4,565,784
<b>USA</b>	74	441 Gb	5,813,403
<b>DENMARK</b>	85	103.87 Gb	5,502,045
<b>CHINA</b>	71	180.78Gb	7,198,512

10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22 996  
23  
24  
25  
26 997 **Table 2. OPLS-DA analysis of Metabolomic datasets with fraction of variation explained by**  
27  
28 998 **X and Y axis with their P-values.**  
29  
30

<b>R<sup>2</sup>X</b>	<b>R<sup>2</sup>Y</b>	<b>Q<sup>2</sup></b>	<b>RMSE</b>	<b>Pre</b>	<b>Ort</b>	<b>pR<sup>2</sup></b>	<b>pQ<sup>2</sup></b>
<b>0.174</b>	0.597	0.391	0.322	2	0	0.05	0.05

31  
32  
33  
34  
35  
36  
37  
38 999  
39  
40  
41 1000  
42  
43  
44 1001  
45  
46  
47 1002  
48  
49  
50 1003  
51  
52  
53 1004  
54  
55  
56  
57 1005  
58  
59  
60 1006  
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

**Figure title and legends**

**Fig. 1. Comparison of Indian gut microbiome with other major populations using 16S rRNA 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 MWAS. The samples are plotted along with the MGS/CAGs having taxonomic annotations. The MGS/CAGs are coloured according to their phylum. Variations across populations is 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 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 showing 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) Eigen values and their scores calculated from PCA of samples using core EggNOGs and accessory EggNOGs are plotted. The boxplots showing for core and accessory factor scores for all samples in different populations are shown. Each box

1  
2  
3  
4 1030 plot represents the median shown as white line between the boxes, the upper and lower ends of the  
5  
6  
7 1031 boxes representing upper quartile (75<sup>th</sup> percentile) and lower quartile (25<sup>th</sup> percentile). The  
8  
9 1032 whiskers extending on both the ends represent 2.5\* IQR (Inter Quartile Range). The different  
10  
11 1033 coloured dots overlaid for each sample are plotted over the box. (C) The enrichment or depletion  
12  
13  
14 1034 of functions in India compared to other populations are shown as volcano plots. The log-  
15  
16 1035 transformed FDR adj P-values calculated from Wilcoxon rank sum test are plotted on the x-axis.  
17  
18  
19 1036 The log odds ratio calculated for India vs Other datasets are plotted on the y-axis. The EggNOGs  
20  
21 1037 with P-value<0.05 are shown in blue while those were having P-values>0.05 are shown in red.  
22  
23  
24 1038 The EggNOGs extending on right and left side and with P-value>0.05 are labelled as highly  
25  
26 1039 enriched in India and other datasets, respectively.

27  
28  
29 1040 **Fig. 3. Variations in gut microbiome at the two locations.** (A) PCA analysis of unweighted  
30  
31 1041 UniFrac distances of OTUs from Indian population and their differentiation due to locations and  
32  
33  
34 1042 diet. Here, the samples are grouped based on their locations (LOC1 and LOC2). The top six  
35  
36  
37 1043 principal components tested for correlations with known factors showed location and diet to  
38  
39 1044 represent the most significant correlations. (B) The within-sample Shannon diversity calculated  
40  
41 1045 for LOC1 and LOC2 are plotted as box plots showing the difference in within-samples diversity  
42  
43  
44 1046 between the two locations (\*: P<0.05). (C) Inter sample Bray Curtis distances calculated for  
45  
46 1047 samples in LOC1 and LOC2 are shown as boxplots (\*: P<0.05). (D) Heatmap showing the  
47  
48  
49 1048 abundance of OTUs as z-transformed scores. The x-axis represents the OTUs and the genera  
50  
51 1049 assigned to the three prominent OTU clusters. (E) Significantly different genera between the two  
52  
53  
54 1050 locations are shown as boxplots with boxes representing *interquartile range* (IQR), dark lines  
55  
56 1051 between the boxes representing median values and whiskers representing the 1.5 x IQR on each  
57  
58  
59 1052 side.

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

**Fig. 4. Between class analysis to identify metabolotypes and their associated metabolites. (A)**

Metabolite clusters (MES) abundance profiles of samples were generated and their clustering was performed using PAM (partition around medoids) clustering. The between class and PCA of JSD distances and PAM clustering identified 3 metabolotypes to be optimum for their segregation using (B) Silhouette index. The metabolites valeric acids, and saturated fatty acids such as palmitic acid and stearic acid, were found higher in Metabotype1. The carbohydrates such as glucose and galactose, were found higher in Metabotype2. The branched chain amino acids, lauric acid and butyric acid were found higher in Metabotype3. (C) OPLS-DA analysis using locations as classes shows locations as differentiating factors in separating the samples based on their metabolomic profiles.

**Fig. 5. Spearman's Rank correlations of metabolites with species and metabolic modules. (A)**

Spearman's Rank Correlation coefficients were calculated between significantly different metagenomic species and significantly different metabolites between LOC1 and LOC2 populations. The correlations showing significant FDR Adj. P <0.05 are plotted. The bars on the right show the Log Odds Ratio of the abundance of MGS with positive values indicating enrichment in LOC1, and the negative values indicating enrichment in LOC2. (B) Spearman's Rank correlations between significantly different (FDR Adj. P <0.05, Wilcoxon test) pathway modules and significantly different metabolite abundances in all samples. The significant (P <0.05) correlations are plotted and the colour intensities depict the correlation coefficients. The correlation of metabolites with locations is shown with labels in dark red colours showing association with LOC2, and the labels in green colours showing correlation with LOC1.

**Fig. 6. BCAA abundance and their differential correlation with LOC1 and LOC2. (A) Bar**

plot showing z-normalized values of serum BCAA levels in LOC1 and LOC2. Differential



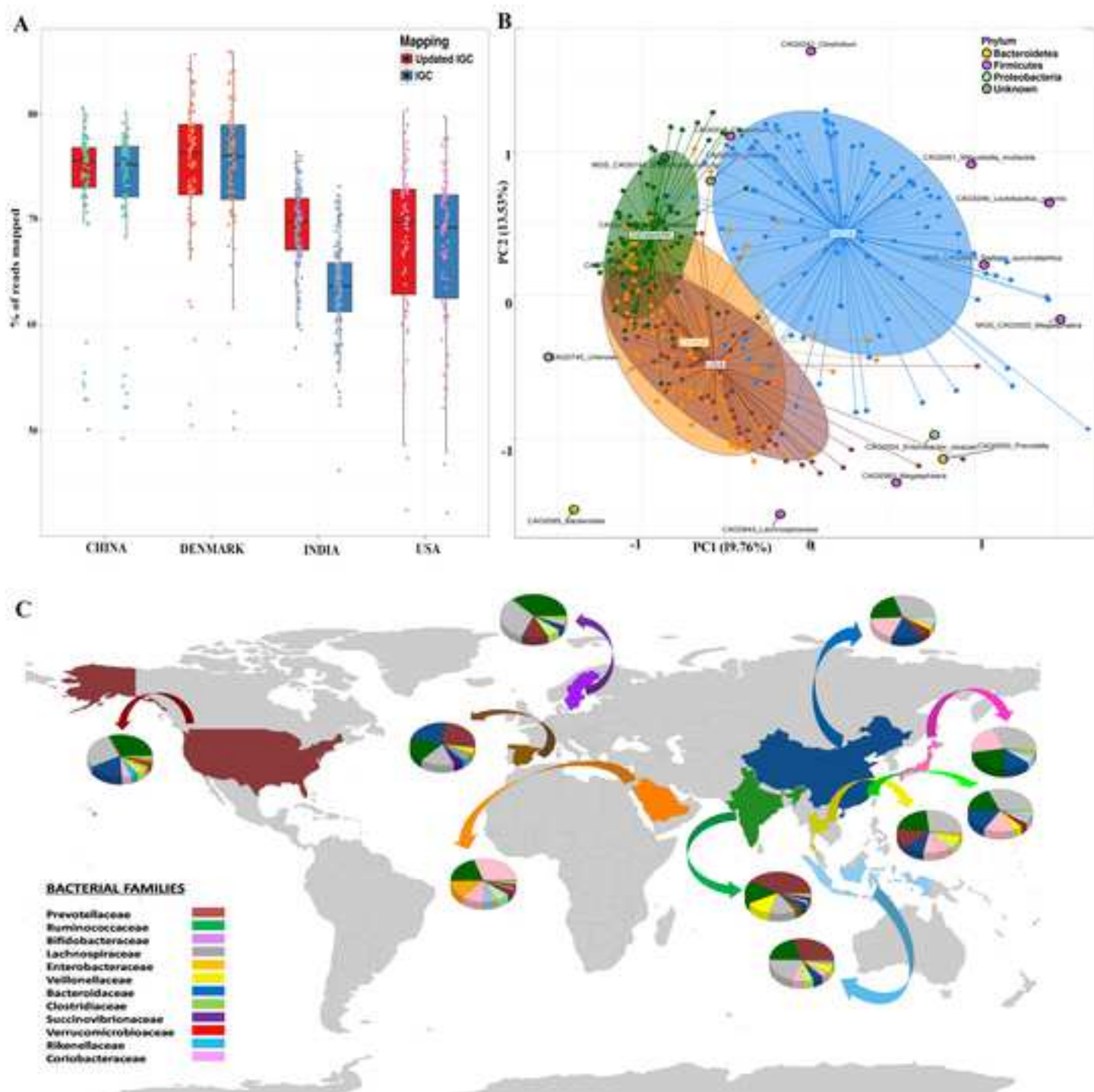
1  
2  
3  
4 1076 correlations between KO modules in **(B)** LOC1 and **(C)** LOC2 which showed significant  
5  
6  
7 1077 differences (FDR Adj. P-value <0.05) in Spearman's correlations are plotted. The KOs within each  
8  
9 1078 module are associated with KOs from other modules. The KOs belonging to BCAA metabolism  
10  
11  
12 1079 and their correlations with other KOs showed significant (FDR Adj. P-value <0.05) difference  
13  
14 1080 between LOC1 and LOC2. The network shows each KO as node and their associations with other  
15  
16 1081 KOs as edges. Only significant correlations (Correlation P-value < 0.05) are plotted. The KOs  
17  
18  
19 1082 which had positive correlations with other KOs are connected by edges and the network analysis  
20  
21 1083 identifies important associations between modules from KO correlations. **(D)** Network analysis of  
22  
23  
24 1084 Spearman's correlations between the branched chain amino acids biosynthesis, degradation and  
25  
26 1085 transport KEGG modules with MGS abundance in both LOC1 and LOC2 populations. The node  
27  
28  
29 1086 size is proportional to the degree of interactions and the links between module and MGS show  
30  
31 1087 interactions or significant correlations (FDR Adj. P < 0.05) with negative (in Red) and positive (in  
32  
33  
34 1088 Blue) correlation coefficients. **(E)** Plot showing z-normalized abundance of KOs associated with  
35  
36 1089 different modules of BCAA biosynthesis and transporters between LOC1 and LOC2.

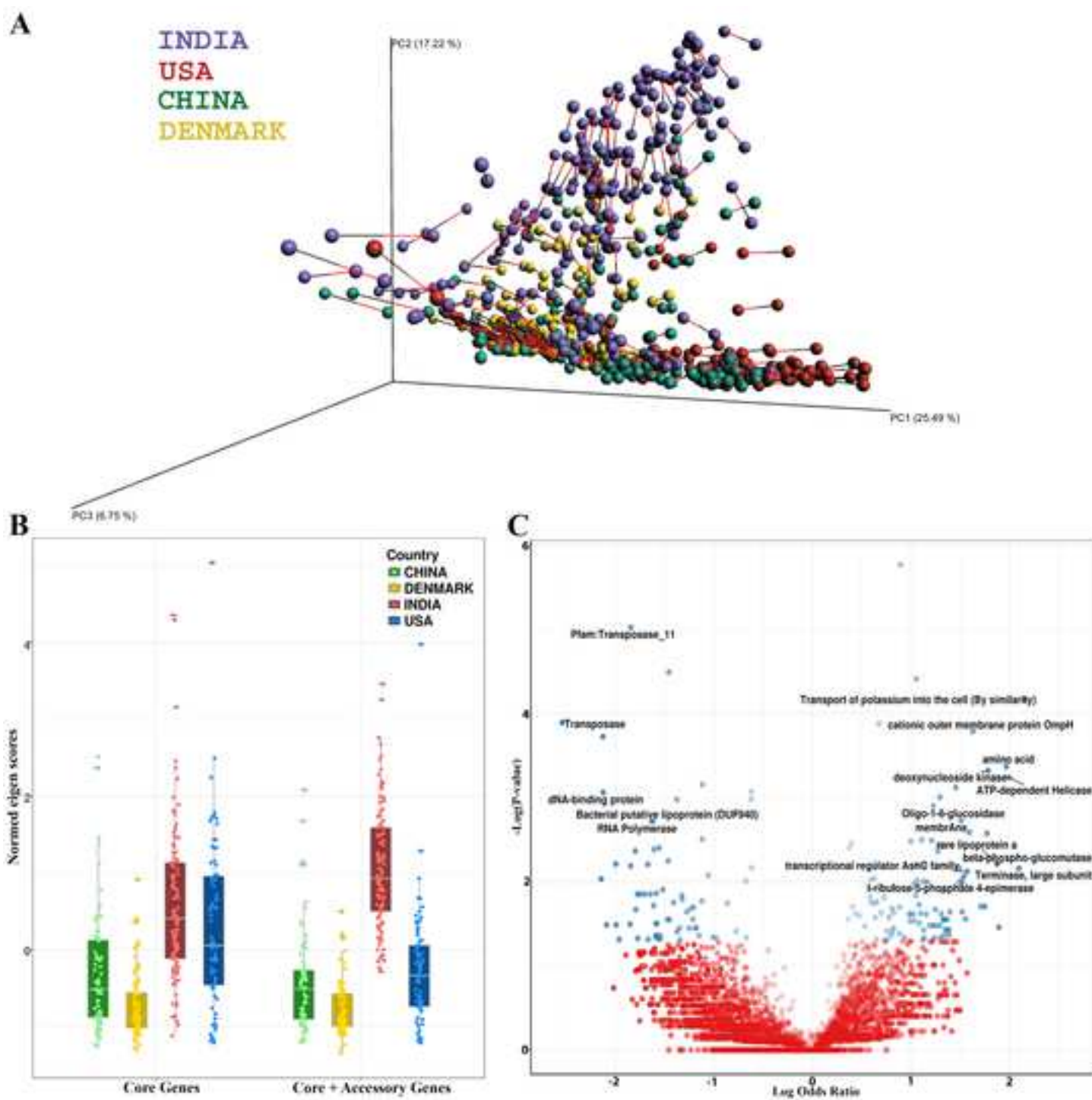
37  
38  
39 1090 **Fig. 7. BCAA transporters playing a key role in maintaining the levels of BCAAs in feces and**  
40  
41 1091 **serum.** The BCAA levels were observed to be significantly high in the serum samples of LOC1  
42  
43  
44 1092 and in the faecal samples of LOC2. The higher abundance of BCAA biosynthesis genes and the  
45  
46 1093 lower abundance of BCAA inward transporters in gut bacteria of LOC1 results in a higher  
47  
48  
49 1094 availability of BCAAs for absorption in the blood stream through the gut lumen, and thus were  
50  
51 1095 observed in high abundance in the serum samples. In contrast, the high abundance of BCAA  
52  
53  
54 1096 inward transporters in the gut bacteria of LOC2, results in a lower availability of BCAAs for  
55  
56 1097 absorption in the gut lumen, and thus were observed in lower abundance in the serum samples.

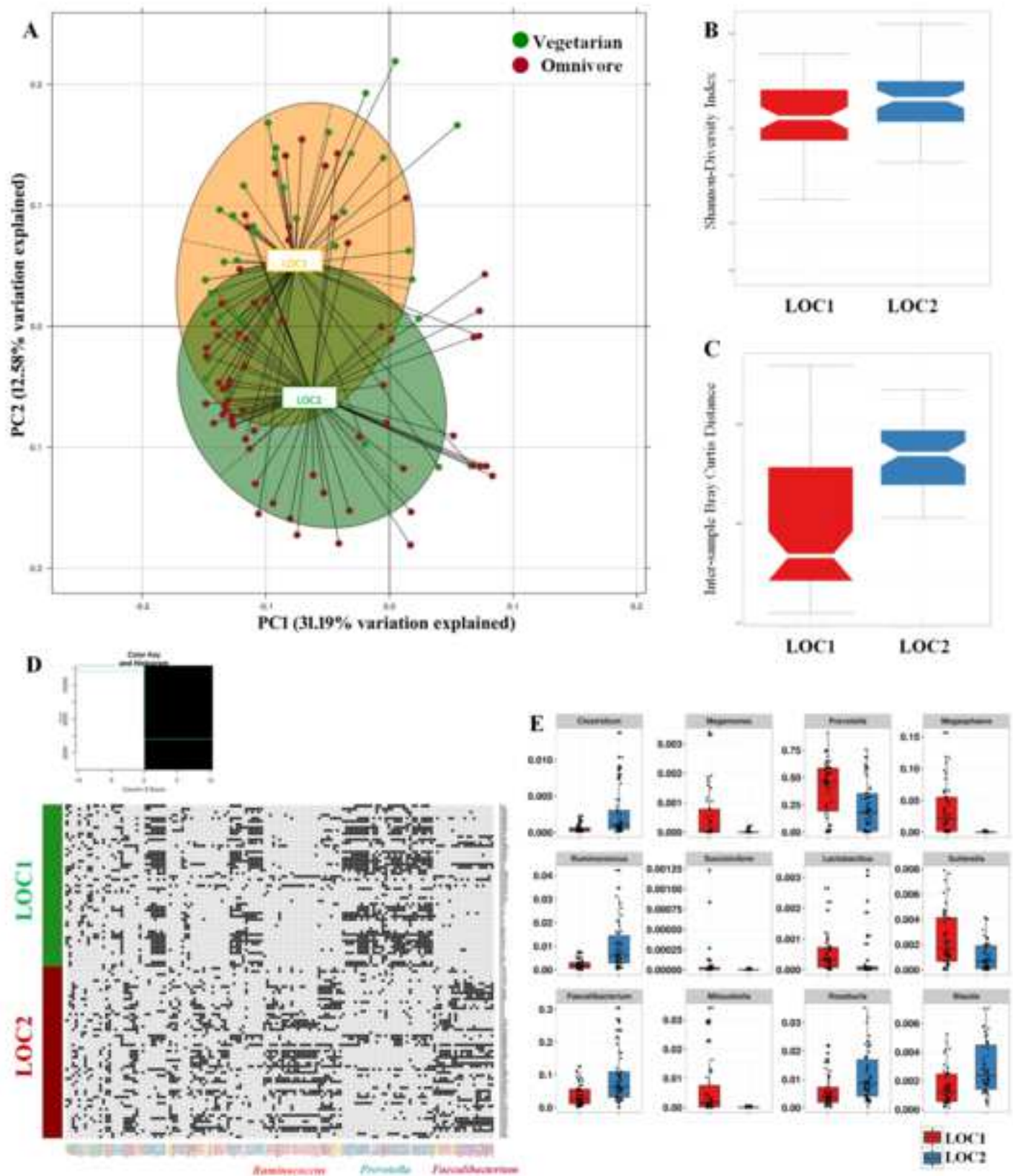
57  
58  
59 1098

1  
2  
3  
4 1099 **Additional Files**  
5  
6  
7 1100 **Additional File 1:** Supplementary data  
8  
9  
10 1101 **Additional File 2:** Summary of sequencing statistics showing the number of reads per sample for  
11  
12 16S rRNA amplicon dataset  
13 1102  
14  
15  
16 1103 **Additional File 3:** Summary of sequencing statistics showing the number of reads per sample for  
17  
18 1104 Whole Genome Shotgun metagenomic dataset  
19  
20  
21 1105 **Additional File 4:** Summary of the reads mapped to Integrated Gene Catalogue and Indian  
22  
23 catalogue combined with IGC.  
24 1106  
25  
26  
27 1107 **Additional File 5: Figures S1 to S9**  
28  
29  
30 1108 **Additional File 6:** Enriched core microbial functions in Indian gut microbiome compared to other  
31  
32 populations  
33 1109  
34  
35  
36 1110 **Additional File 7:** Genus level differences between Enterotype-1 and Enterotype-2 with FDR  
37  
38 1111 Adjusted P-values determined by Wilcoxon rank sum test  
39  
40  
41 1112 **Additional File 8:** Tables showing Calinski Harabasz index and prediction strength determined  
42  
43 for each cluster  
44 1113  
45  
46  
47 1114 **Additional File 9:** Enriched KOs identified using Wilcoxon rank sum test and Log Odds Ratios  
48  
49 between ET-1 and ET-2  
50 1115  
51  
52  
53 1116 **Additional File 10:** Table showing Polyserial correlation of covariates with the principal  
54  
55 1117 components with FDR Adj. P-values  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65

1  
2  
3  
4 1118 **Additional File 11:** Table showing enrichment of MGS/CAGs obtained from MWAS with their  
5  
6  
7 1119 taxonomic annotations in LOC-1 and LOC-2  
8  
9  
10 1120 **Additional File 12:** Polyserial correlation of covariates with principal components explaining  
11  
12 1121 variations across samples using metabolomic dataset.  
13  
14  
15 1122 **Additional File 13:** Table showing Spearman's rank correlation coefficient values of metabolites  
16  
17  
18 1123 with Metabotypes  
19  
20  
21 1124 **Additional File 14:** Table showing differential abundance of KEGG Modules between LOC-1  
22  
23  
24 1125 and LOC-2  
25  
26  
27 1126 **Additional File 15:** List of reference Genomes from NCBI and HMP databases for reference  
28  
29 1127 mapping  
30  
31  
32 1128  
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

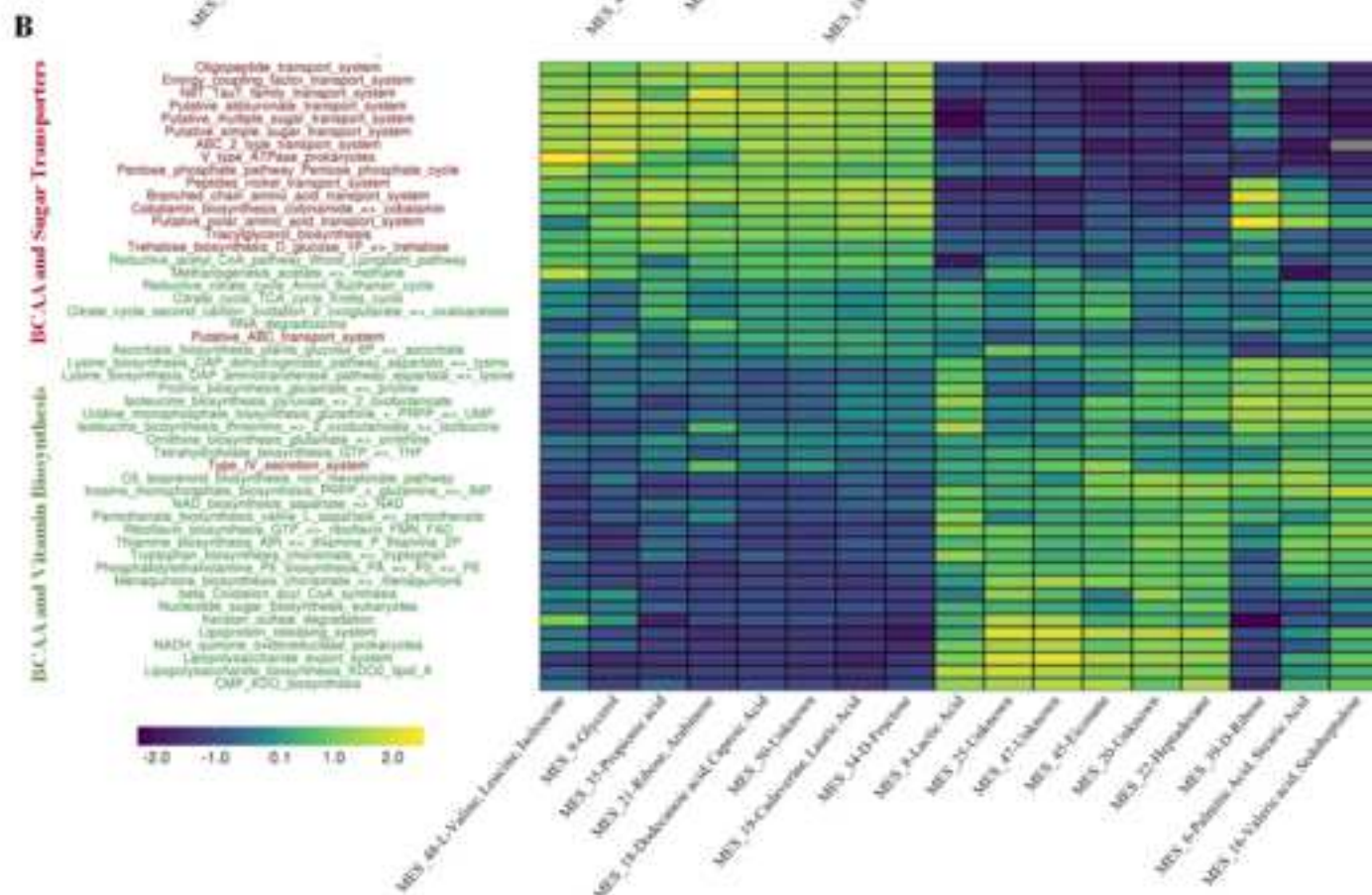
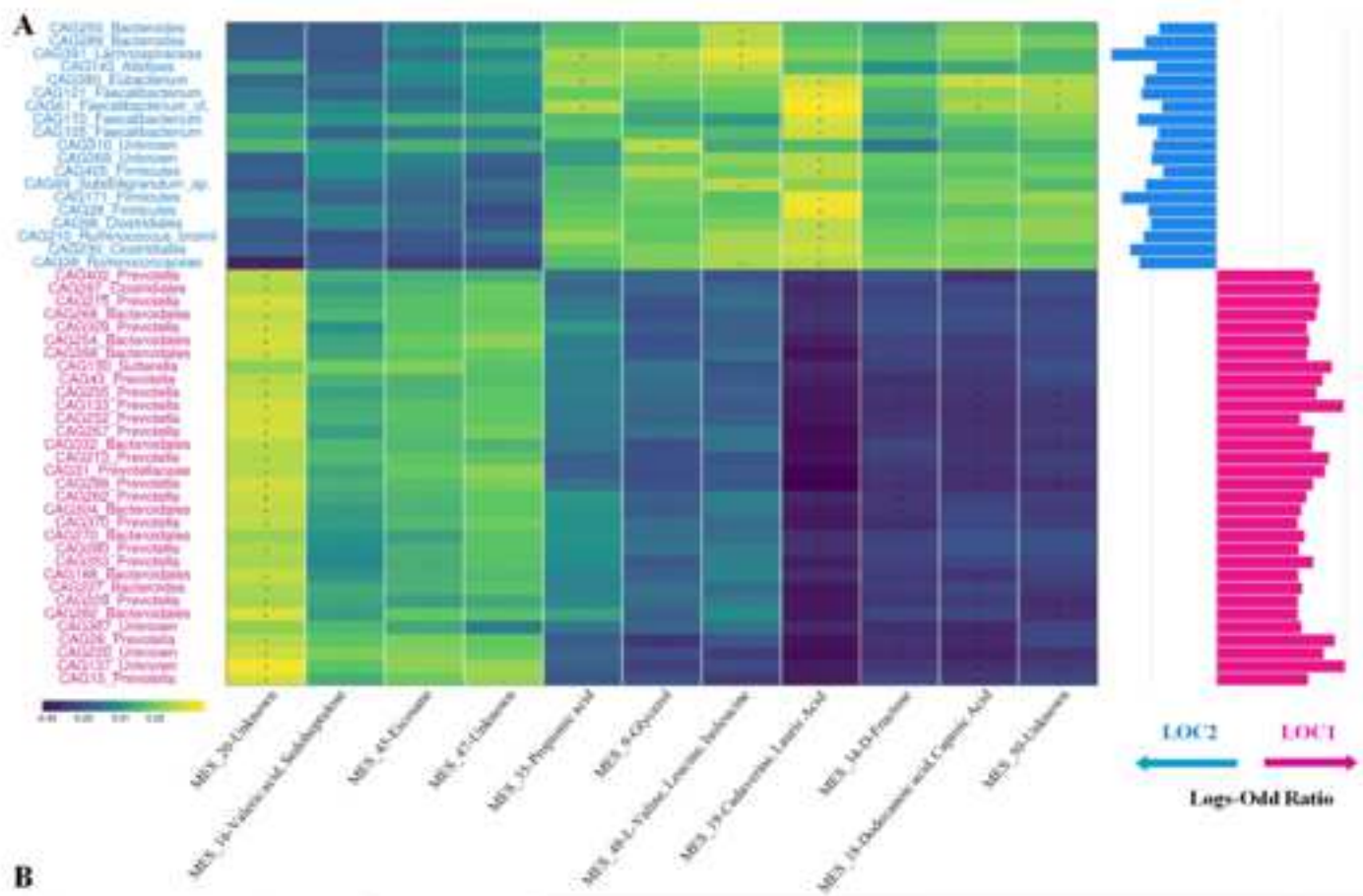




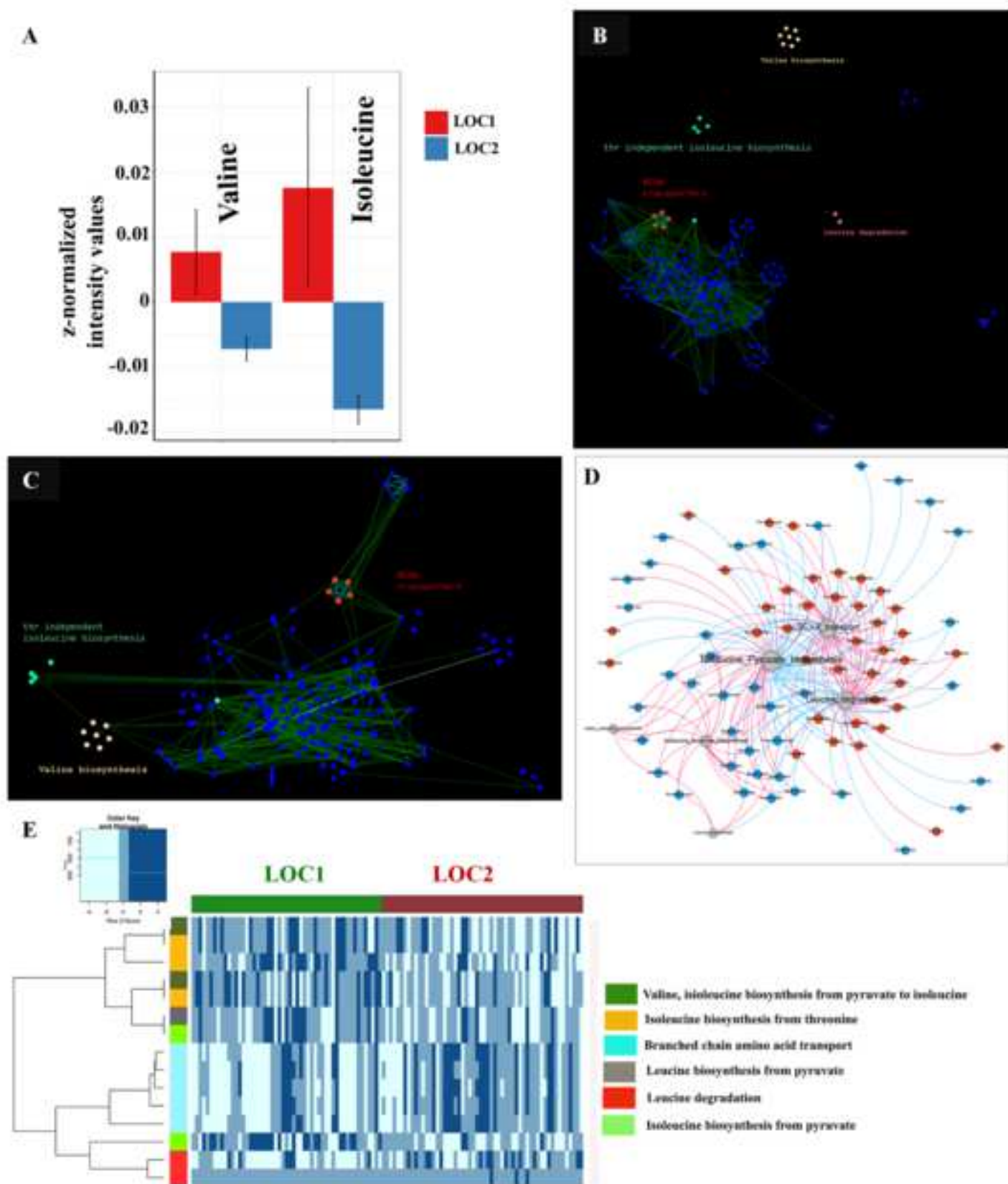


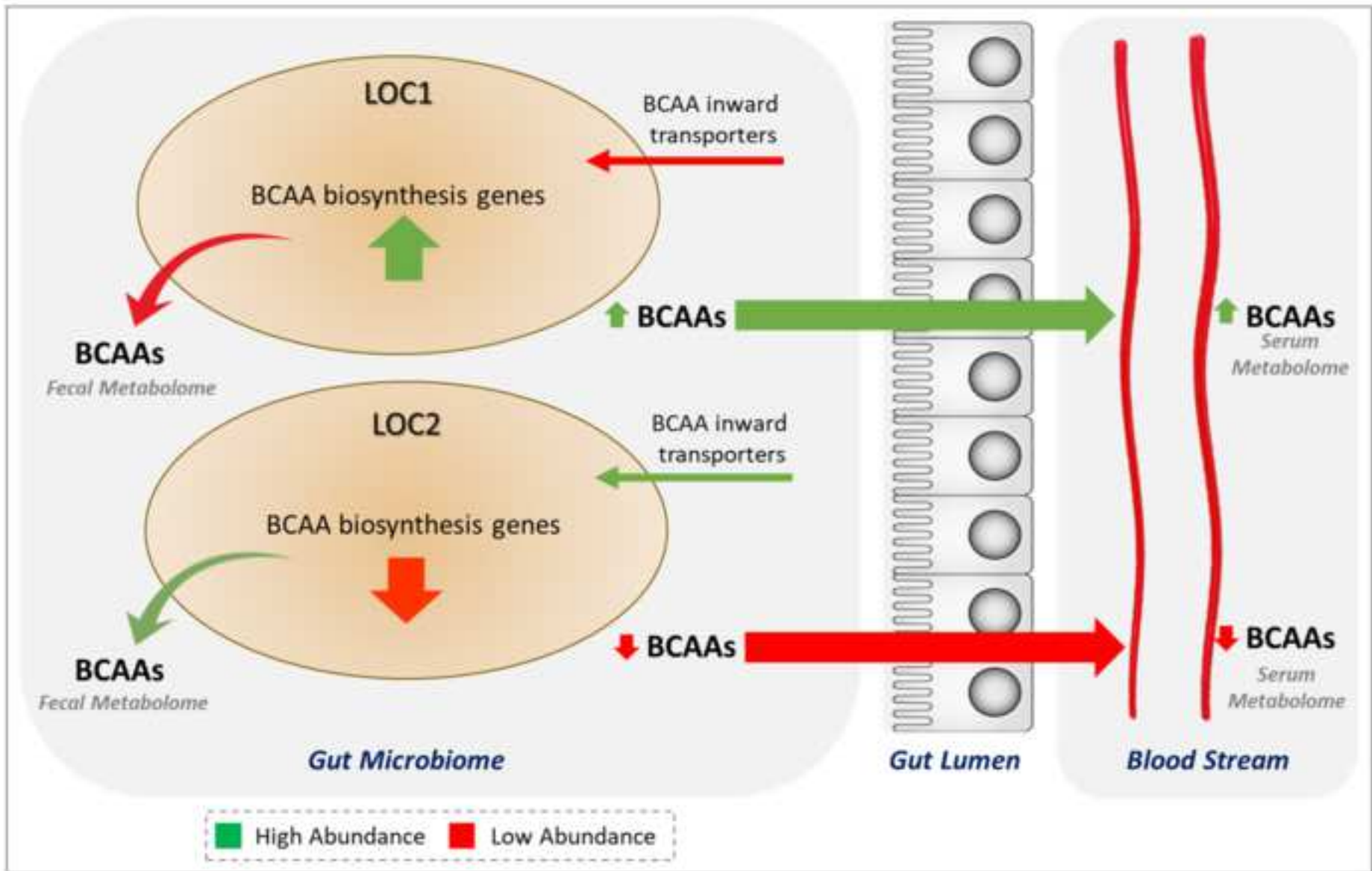


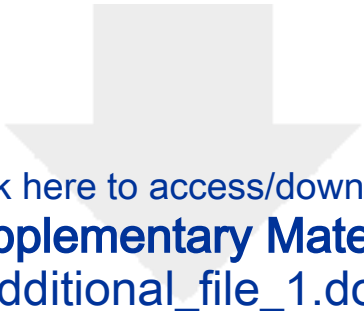




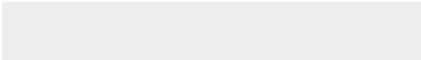



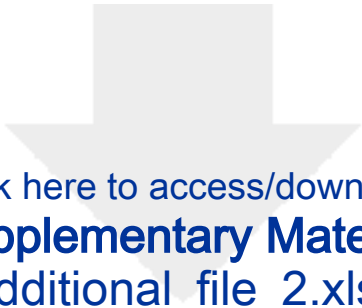







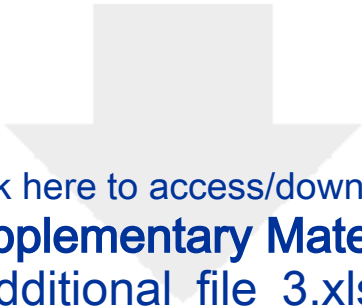
Click here to access/download  
**Supplementary Material**  
Additional\_file\_1.doc







Click here to access/download  
**Supplementary Material**  
Additional\_file\_2.xlsx






Click here to access/download  
**Supplementary Material**  
Additional\_file\_3.xlsx

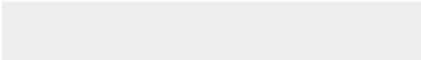






Click here to access/download  
**Supplementary Material**  
Additional\_file4.xlsx



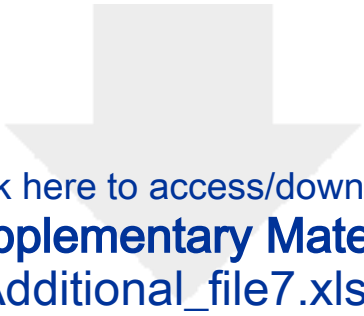
Click here to access/download  
**Supplementary Material**  
Additional\_file5.docx







Click here to access/download  
**Supplementary Material**  
Additional\_file6.xlsx






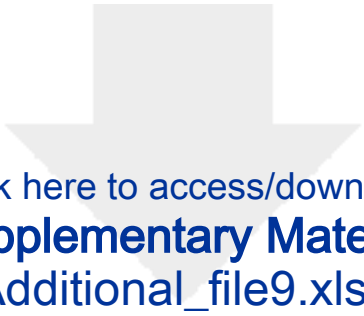
Click here to access/download  
**Supplementary Material**  
Additional\_file7.xlsx






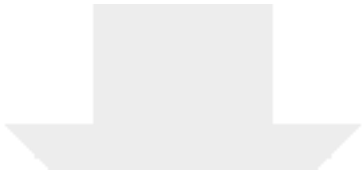
Click here to access/download  
**Supplementary Material**  
Additional\_file8.xlsx






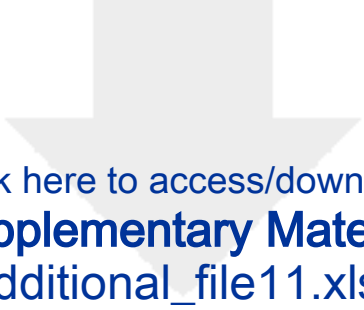
Click here to access/download  
**Supplementary Material**  
Additional\_file9.xlsx



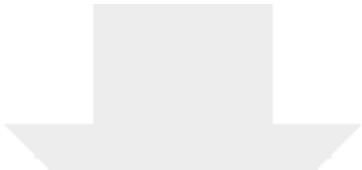


Click here to access/download  
**Supplementary Material**  
Additional\_file10.xlsx

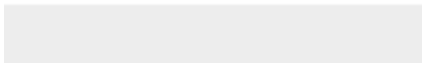
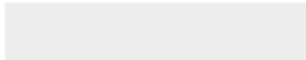


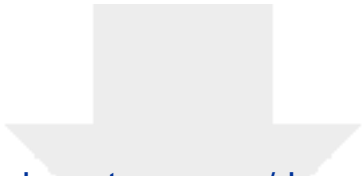


Click here to access/download  
**Supplementary Material**  
Additional\_file11.xlsx




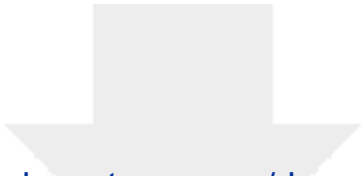
Click here to access/download  
**Supplementary Material**  
Additional\_file12.xlsx






Click here to access/download  
**Supplementary Material**  
Additional\_file13.xlsx

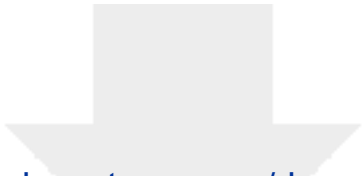




Click here to access/download  
**Supplementary Material**  
Additional\_file14.xlsx







Click here to access/download  
**Supplementary Material**  
Additional\_file15.xlsx

