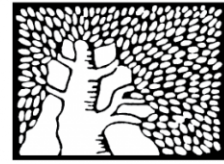


מכון ויצמן למדע

WEIZMANN INSTITUTE OF SCIENCE



## Structural variation in the gut microbiome associates with host health

### Document Version:

Accepted author manuscript (peer-reviewed)

### Citation for published version:

Zeevi, D, Korem, T, Godneva, A, Bar, N, Kurilshikov, A, Lotan-Pompan, M, Weinberger, A, Fu, J, Wijmenga, C, Zhernakova, A & Segal, E 2019, 'Structural variation in the gut microbiome associates with host health', *Nature*, vol. 568, no. 7750, pp. 43-48. <https://doi.org/10.1038/s41586-019-1065-y>

*Total number of authors:*

11

### Digital Object Identifier (DOI):

[10.1038/s41586-019-1065-y](https://doi.org/10.1038/s41586-019-1065-y)

### Published In:

Nature

### License:

Other

### General rights

@ 2020 This manuscript version is made available under the above license via The Weizmann Institute of Science Open Access Collection is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognize and abide by the legal requirements associated with these rights.

### How does open access to this work benefit you?

Let us know @ [library@weizmann.ac.il](mailto:library@weizmann.ac.il)

### Take down policy

The Weizmann Institute of Science has made every reasonable effort to ensure that Weizmann Institute of Science content complies with copyright restrictions. If you believe that the public display of this file breaches copyright please contact [library@weizmann.ac.il](mailto:library@weizmann.ac.il) providing details, and we will remove access to the work immediately and investigate your claim.



1 **Sub-genomic variation in the gut microbiome associates with host metabolic health**

2

3 David Zeevi<sup>1,2,3,+,\*</sup>, Tal Korem<sup>1,2,4,5,+</sup>, Anastasia Godneva<sup>1,2</sup>, Noam Bar<sup>1,2</sup>, Alexander Kurilshikov<sup>6</sup>,  
4 Maya Lotan-Pompan<sup>1,2</sup>, Adina Weinberger<sup>1,2</sup>, Jingyuan Fu<sup>6,7</sup>, Cisca Wijmenga<sup>6,8</sup>, Alexandra  
5 Zhernakova<sup>6</sup>, Eran Segal<sup>1,2,\*</sup>

6

7 **Author affiliations**

8 <sup>1</sup>Department of Computer Science and Applied Mathematics, Weizmann Institute of Science,  
9 Rehovot 7610001, Israel

10 <sup>2</sup>Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot 7610001, Israel

11 <sup>3</sup>Center for Studies in Physics and Biology, The Rockefeller University, New York, NY 10065, USA

12 <sup>4</sup>Department of Systems Biology, Columbia University Irving Medical Center, New York, NY 10032,  
13 USA

14 <sup>5</sup>Department of Obstetrics and Gynecology, Columbia University Irving Medical Center, New York,  
15 NY 10032, USA

16 <sup>6</sup>University of Groningen, University Medical Center Groningen, Department of Genetics, 9713 GZ  
17 Groningen, The Netherlands

18 <sup>7</sup>University of Groningen, University Medical Center Groningen, Department of Pediatrics, 9713 GZ  
19 Groningen, The Netherlands

20 <sup>8</sup>Department of Immunology, K.G. Jebsen Coeliac Disease Research Centre, University of Oslo,  
21 0424 Oslo, Norway

22 <sup>+</sup>These authors contributed equally to this work.

23 <sup>\*</sup>to whom correspondence should be addressed: [eran.segal@weizmann.ac.il](mailto:eran.segal@weizmann.ac.il);

24 [dzeevi@rockefeller.edu](mailto:dzeevi@rockefeller.edu)

25

26

27 **Abstract**

28 Differences in the presence of even a few genes between otherwise identical bacterial strains  
29 may result in critical phenotypic differences, yet exploring variation at this sub-genomic level  
30 across gut microbiomes is challenging, possibly owing to difficulties in correct metagenomic  
31 read assignment. Here, we devised algorithms that improve the assignment accuracy of  
32 metagenomic reads to reference sequences and systematically identify variability in microbial  
33 sub-genomic regions. We find Sub-Genomic Variation (SGV) to be prevalent in the microbiome  
34 across multiple phyla, and that our method produces SGVs that replicate across distinct human  
35 cohorts from different continents. SGVs are associated with bacterial fitness and their member  
36 genes are enriched for CRISPR-associated and antibiotic producing functions and depleted  
37 from housekeeping genes, suggestive of a role in microbial adaptation. We find 124 novel  
38 associations between SGVs and host disease risk factors, of which 40 replicate in an  
39 independent cohort, highlighting the universality of these associations. Finally, by exploring  
40 genes clustered in the same SGV, we uncover several possible mechanistic links between the  
41 microbiome and its host, as in the case of a 31kbp region in *Anaerostipes hadrus* encoding a  
42 composite inositol catabolism-butyrate biosynthesis pathway, whose presence is associated  
43 with significantly lower host body weight and metabolic disease risk. Overall, our results uncover  
44 a nascent layer of variability in the microbiome that is associated with microbial adaptation and  
45 host health.

46

47

## 48 **Introduction**

49  
50 Genes that are deleted or duplicated within different members of a species (also termed copy  
51 number variation; CNV), are a phenomenon common across all kingdoms<sup>1,2</sup>. In humans, CNVs  
52 allowed adaptation to starch consumption by an increase in the copy number of the alpha-  
53 amylase gene<sup>3</sup>, and they are also linked to multiple conditions such as autism spectrum  
54 disorders<sup>4</sup>, psychiatric disorders<sup>5</sup>, obesity<sup>6</sup>, and autoimmune disease<sup>7,8</sup>. In bacteria, even a  
55 small number of genes can underlie phenotypes such as virulence<sup>9,10</sup>, antibiotic resistance<sup>11</sup>,  
56 host metabolic disease<sup>12</sup> and even host longevity<sup>13</sup>, making genetic variation highly important to  
57 both the microbe and its host.

58         Microbes in the human intestines share copious genetic material<sup>14</sup>, resulting in a high  
59 prevalence of CNVs across the gut microbiome<sup>15</sup>. This variability could be critical to human  
60 pathophysiology, as gut microbes were found to be involved in multiple host processes, such as  
61 fiber metabolism<sup>16</sup>, bile acid metabolism<sup>17</sup>, vitamin biosynthesis<sup>18</sup> and immune conditioning<sup>19</sup>,  
62 and are associated with multiple host disorders ranging from obesity and diabetes<sup>20,21</sup>, through  
63 inflammatory bowel diseases<sup>22,23</sup>, to macular degeneration<sup>24</sup> and autism<sup>25</sup>. The mechanisms  
64 underlying these associations are often unclear and could perhaps be elucidated through the  
65 examination of CNVs.

66         The vast majority of microbiome research to date, however, typically studies the  
67 microbiome through the prism of relative abundance of microbial species, with only a small  
68 number of studies focusing on the functional genetic level. Some studies analyzed the genetic  
69 repertoire of the microbiome<sup>21</sup> by mapping metagenomic reads to a collection of microbial genes  
70 (e.g. <sup>26,27</sup>). While useful, this approach is limited as it usually analyzes microbial genes  
71 separately from the microbes in which they are expressed, overlooking their genomic context  
72 and membership in species-specific microbial pathways. Taxonomy-aware methods such as  
73 FishTaco<sup>28</sup> and HUMAnN2<sup>29</sup>, may supply information on microbial membership of genes, but is

74 limited in resolution with regards to within-species variation. Recently, Greenblum et al.<sup>15</sup> have  
75 performed a systematic characterization of intra-species CNVs across the human microbiome.  
76 Both approaches, however, are limited by the scope of the annotation database used (KEGG<sup>26</sup>  
77 in the latter case<sup>15</sup>), and in any case do not account for co-variation of genes encoded in the  
78 same genomic region. Such co-variation is important as it encodes information such as operon  
79 membership, gene regulation, proximal RNA interference and susceptibility for horizontal  
80 transfer that are only evident when analyzing genes in their immediate genomic context.

81 In this study, we focused on sub-genomic regions in the human microbiome that vary  
82 across different hosts. We aimed to detect segments of varying lengths, potentially containing  
83 multiple genes, that are deleted from certain bacteria in some individuals or present in a variable  
84 number of copies in others. We term this phenomenon “sub-genomic variation” to differentiate it  
85 from CNVs at the level of specific genes without genomic context (such as analyzed by  
86 Greenblum et al.<sup>15</sup>).

87 One major difficulty in observing genes in their genomic context stems from the  
88 challenge in correctly assigning metagenomic reads that originate from regions that are similar  
89 between different bacteria. As many sequences are homologous between members of the same  
90 taxonomic clade and others are potentially horizontally transferred between clades<sup>14</sup>, it is often  
91 challenging to discern regions of high copy number within a genome from regions that are  
92 present in multiple members of the metagenome. To overcome these issues, we devised an  
93 Iterative Coverage-based Read Assignment (ICRA) algorithm that resolves ambiguous read  
94 assignments using information on relative abundances of bacterial members of the microbiome,  
95 sequencing-coverage across their genomes, and sequencing and alignment qualities. We show  
96 that our algorithm correctly assigns reads in complex metagenomic settings.

97 We utilize our improved read assignment to develop a novel algorithm, SGV-Finder,  
98 allowing us to detect 7479 microbial SGVs in 56 species from 7 microbial phyla in 887 human  
99 microbiome samples<sup>20,30</sup>, demonstrating that SGVs are widely prevalent in the human

100 microbiome. We show that these SGVs have distinct genetic functions, are associated with  
101 bacterial growth rates, and are stable within the same person even over long periods of time,  
102 altogether implicating SGVs as drivers of adaptation of a microbiome to a specific host  
103 environment. We demonstrate the potential importance of SGVs to the human host by showing  
104 124 cases in which SGVs are significantly associated with multiple disease risk factors. We  
105 replicate our analysis in the Dutch Lifelines DEEP cohort<sup>31,32</sup> and show that SGV positions  
106 replicate in 76% of bacteria present in both cohorts, and that 40 associations with risk factors  
107 also replicate, altogether suggesting that some genomic structural variability is shared between  
108 distinct population, while some is population specific. We further demonstrate that examining  
109 gene clusters in variable regions can reveal potential mechanisms of action, as in the case of an  
110 *A. hadrus* region associated with multiple risk factors and whose genes code for a microbial  
111 pathway which metabolizes sugar-alcohols to butyrate, a short-chain fatty acid (SCFA)  
112 renowned for its advantageous effects on the human host<sup>33-35</sup>. Overall, we show that SGVs  
113 represent a nascent layer of information in the human microbiome that is likely to be of high  
114 relevance to human health.

115

## 116 **Results**

117

### 118 **Accurate metagenomic read assignment using the ICRA algorithm**

119

120 To accurately detect SGVs in the microbiome we sought to obtain a correct assignment of  
121 metagenomic reads to their sequence of origin. Attaining such accurate assignment is  
122 challenging due to the large number of genomic sequences that are shared across different  
123 microbiome members. Here, we analyzed data collected on 887 healthy subjects which includes  
124 microbiome profiling alongside detailed blood glucose measurements over the duration of a  
125 week, anthropometric measurements, blood tests, and medical questionnaires<sup>20,30</sup> (Methods). In

126 these 887 samples, over 15% of the metagenomic reads were assigned ambiguously to multiple  
127 references upon mapping to a reference genome database of 3953 bacterial genomes<sup>36</sup> (Fig.  
128 S1A, Methods).

129 To address this problem, we devised an Iterative Coverage-based Read Assignment  
130 (ICRA) algorithm (Fig. 1A, Methods). In its first step, ICRA uses read assignments and mapping  
131 qualities to calculate the sequencing coverage depth along microbial entities (e.g., bacterial  
132 genomes or genes), and then uses this sequencing coverage to estimate microbial relative  
133 abundances, while demanding sufficient coverage over entities that are to be considered  
134 present in a sample (Methods). In the next step, ICRA reassigns reads using the updated  
135 relative abundances, and repeats the process to convergence. The use of sequencing coverage  
136 makes our method robust to genomic regions with extremely high or low coverage that may  
137 arise from misassemblies, homology to other microbes, or phage activation. Such regions could  
138 otherwise bias the estimated relative abundances, potentially even assigning abundances to  
139 genomic entities that are not present in the sample, but contain a region homologous to other  
140 entities present in reference databases.

141 To test the performance of ICRA, we validated the two key components of the algorithm:  
142 its ability to resolve ambiguous read assignments, and the accuracy of the relative abundances  
143 that it assigns to each bacterial species. To this end, we analyzed the assignment of reads from  
144 simulated metagenomes provided by the CAMI challenge dataset along with their correct read  
145 assignments<sup>37</sup>. The CAMI dataset contains three sets of samples ranging from 30 to 450  
146 genomes that account for varying microbiome complexities. We mapped each of these samples  
147 to a reference of 482 bacteria derived from this dataset and compared the fraction of  
148 metagenomic reads incorrectly or ambiguously assigned to reference genomes between a  
149 baseline setting (uncorrected read assignment; Methods), the output of our algorithm, and two  
150 state-of-the-art tools - Kraken<sup>38</sup> and MetaPhyler<sup>39</sup>. Notably, we found that ICRA outperforms the



151 alternatives in assigning reads to reference genomes in both the species and sub-species  
152 taxonomic levels in all complexity levels available from CAMI ( $p < 0.01$ ; Fig. 1B, S1B,C)

153 As relative abundances are utilized by ICRA for the resolution of ambiguous read  
154 assignments, we further validated that ICRA-derived relative abundances are comparable to  
155 those derived from state-of-the-art tools created and optimized for this task. We therefore  
156 compared microbial relative abundances produced by ICRA, to those derived from the popular  
157 tools MetaPhlan2<sup>40</sup>, which uses marker genes to estimate abundances, and Bracken<sup>41</sup>, which  
158 performs Bayesian reestimation of abundances derived with Kraken<sup>38</sup>. To this end, and to best  
159 simulate the genomic phenomena of bacteria growing naturally (rather than sampled *in silico*),  
160 we obtained seven different bacterial strains, grew them to stationary phase, and extracted and  
161 sequenced DNA from each strain separately (Methods). We then created 100 samples *in silico*  
162 by randomly mixing reads sequenced from each of the seven strains at different relative  
163 abundances, and applied MetaPhlan2, Bracken and ICRA to these samples (Methods). We  
164 found that while the Bray-Curtis dissimilarities between the relative abundances estimated by  
165 these tools and the true relative abundances were lowest in Bracken (Fig. 1C, inset), followed  
166 by ICRA and MetaPhlan2, the abundances estimated by all three tools were comparable and  
167 highly correlated with the true abundances ( $R^2 > 0.93$  for each microbe across all samples,  
168  $p < 10^{-10}$ ; Fig. 1C, S2).

169

### 170 **Sub-genomic variation is highly prevalent in the human microbiome**

171

172 We next sought to systematically characterize the landscape of sub-genomic variation across  
173 the healthy human microbiome. To this end, we developed SGV-Finder, which we ran on ICRA-  
174 corrected read assignments of 887 metagenomic samples<sup>20,30</sup> to a reference database of 3953  
175 representative microbial genomes derived from progenomes<sup>36</sup> (Methods). SGV-Finder analyzes  
176 coverage-depth across all microbial genomes in all samples by dividing each genome to 1000

177 basepair bins and counting the number of reads mapped to each bin. To ensure proper  
178 statistical support for copy number analyses, we discard genomes in samples whose median  
179 bin coverage is lower than 10 reads (corresponding to a genome coverage of 1x, with ten 100bp  
180 reads in each 1kbp bin; Methods), and microbial genomes present in less than 75 subjects. The  
181 coverage depth of each genome in a given sample is then standardized by subtracting the  
182 mean sample coverage and dividing by its standard deviation (Methods).

183 For detecting SGVs, we further differentiate between two SGV types. Deletion-SGVs are  
184 sub-genomic areas that are deleted in enough subjects yet are present in others, and are  
185 detected by searching for bins that are deleted in 25-75% of samples, with the read coverage  
186 cutoff for deleted bins selected according to the distribution of read coverages (Methods).  
187 Variable-SGVs are sub-genomic areas which have highly variable coverage across samples,  
188 and are detected by fitting a beta-prime distribution on the standardized coverage of all samples  
189 in a single bin, for bins that are not deleted in more than 5% of samples, and selecting bins with  
190 abundance higher than 95% of values in the fitted distribution. In both variable- and deletion-  
191 SGVs, detected bins are subsequently united based on cooccurrence (deletion-SGVs) or  
192 correlation (variable-SGVs) (Methods). An online metagenome explorer for all SGVs and the  
193 genes they encompass is available at <http://genie.weizmann.ac.il/SGV/> (Fig. S3).

194 Overall, we detected 2423 variable-SGVs and 5056 deletion-SGVs in 56 bacteria found  
195 with sufficient coverage in at least 75 out of 887 samples (Fig. 2A). Sub-genomic variability was  
196 detected in all 6 bacterial phyla and one archaeal phylum, with the number of variable or  
197 deletion SGVs ranging from 5 to 241 SGVs per species in average sizes ranging between 1.4  
198 and 18.6 kbp per species. Variable-SGVs make up between 0.3% and 8.4% of the microbial  
199 genome while deletion SGVs exist in 5.0% to 26.9% of the genome (Fig. 2A). This apparent  
200 disparity in size may suggest inherent differences in the formation of the two types of SGVs. Out  
201 of 887 samples, 769 carried deletion- and variable-SGVs for *Blautia wexlerae*, 727 subjects had  
202 104 deletion-SGVs and 33 variable-SGVs in *A. hadrus*, and 668 carried deletion- and variable-

203 SGVs for *Bacteroides uniformis*. Notably, we detected SGVs in all microbial strains that had  
204 sufficient coverage, and in every subject analyzed, demonstrating the ubiquity of such  
205 variations.

206

## 207 **SGV is prevalent across distinct populations and continents**

208

209 To test the universality of these regions and reinforce their biological relevance, we applied  
210 ICRA and SGV-Finder independently to 1020 out of 1135 samples from the Dutch Lifelines  
211 DEEP cohort<sup>31,32</sup> which had sufficient sequencing depth (Methods). We found that in 47 out of  
212 56 bacteria present in both cohorts, an average of 72.9% of variable-SGVs (0% to 99.1%) and  
213 78.3% of deletion-SGVs (35.3% to 94.5%) overlapped with SGVs found in our cohort (one-sided  
214 hypergeometric  $p < 10^{-10}$ ; Fig. 2B,C). Notably, for 75% of microbes, more than 70% of the regions  
215 were replicated despite the different populations examined with different genetic background,  
216 cultural setting, and dietary preferences (Fig. 2C).

217 Some bacteria, such as *Ruminococcus bicirculanus*, showed very low concordance  
218 between the two cohorts (27% overlap over 10 variable-SGV regions totalling 23kbp; Fig. 2B,C),  
219 suggestive of geographical confinement of the variability, or a strong influence of population-  
220 specific environmental factors. Conversely, other bacteria, such as *Parabacteroides merdae*,  
221 showed high concordance (95% of 46 variable-SGVs totalling 281 kbp; Fig. 2B,C). Given the  
222 different methods, centers, and staff involved in assembling the two cohorts, the replication of  
223 the variable regions suggest that the variability detected here is not artifact but rather a  
224 widespread phenomena in the gut microbiome across distinct geographical regions.

225

## 226 **SGVs are person specific and are shared with habitat**

227

228 We next examined the variability of SGVs across people by correlating the abundance of  
229 variable- and deletion-SGVs between different subjects. We found that different individuals  
230 mostly have different SGVs, with a median correlation of 0.02 and 0 for variable- and deletion-  
231 SGVs, respectively (Fig. 2D,E). In contrast, SGVs were highly stable within the same individuals  
232 even over time periods exceeding one year, with median within-person correlations of 0.89 and  
233 0.66 for variable- and deletion-SGVs, respectively (Spearman correlation  $p < 10^{-20}$  for both; Fig.  
234 2D,E; Methods).

235 To estimate the effect of the environment and host genetics on SGVs, we analyzed data  
236 from cohabiting individuals and for pairs of parents-children / siblings who do not live together<sup>42</sup>  
237 (Methods). We found that cohabiting individuals and parent-children / sibling pairs share both  
238 deletion- and variable-SGVs to a significantly higher degree as compared to two randomly  
239 chosen subjects from our cohort (average Spearman  $\rho$  of 0.45 and 0.16 for variable- and  
240 deletion-SGVs, respectively;  $p < 10^{-10}$  for both; Fig. 2D,E). Interestingly, siblings / parents-  
241 children have a significantly less similar SGV profile in their microbiome as compared to  
242 cohabiting subjects ( $p < 0.001$  for both variable- and deletion-SGVs, Fig. 2D,E). This result is  
243 conservative, as such similarity in the SGV profiles of genetically-related individuals cannot be  
244 efficiently decoupled from confounders such as traditional food preferences or instances in  
245 which these individuals share meals or experiences that may affect their microbiome as part of  
246 their family get-togethers. These results replicate and strengthen our previous findings<sup>42</sup>  
247 showing that environment dominate over genetics in determining microbiome composition.

248

### 249 **Microbiome SGVs are potentially involved in microbial adaptation and function**

250

251 We sought to systematically characterize the functional landscape of SGV regions by examining  
252 genetic functions that are enriched or depleted from SGVs. We annotated gene function across

253 variable- and deletion-SGVs, as well as in regions of microbial genomes that were covered  
254 consistently in at least 98% samples that contained the bacteria (hereinafter termed 'conserved'  
255 regions; Methods). We then performed enrichment analysis to seek for KEGG modules that  
256 were over- and under-represented in these regions (Methods). Using the KEGG BRITE  
257 hierarchy, we found that modules categorized into 'housekeeping' functions such as nucleotide  
258 and amino acid metabolism or carbohydrate and lipid metabolism were significantly depleted  
259 from variable- ( $p < 10^{-5}$  for both groups; Fig. 2F; Table S1) and deletion- ( $p < 10^{-5}$ ; Fig. 2G; Table  
260 S1) SGVs and significantly enriched in conserved regions ( $p < 10^{-5}$ ; Fig. 2H; Table S1).  
261 Conversely, modules classified as ABC-2 type- and other transport systems were significantly  
262 enriched in SGVs ( $p < 10^{-5}$ ), possibly driven by the KEGG module pertaining to putative ABC  
263 transporters ( $p < 10^{-5}$ ; Fig. 2F). In addition, SGVs were enriched with the type-IV secretion  
264 system (T4SS) KEGG module ( $p < 10^{-5}$ ; Fig. 2F,G) suggesting that bacterial conjugation  
265 systems, to which the T4SS is related, are strong drivers of variability. These systems were  
266 strongly depleted from conserved regions ( $p < 10^{-5}$ ; Fig. 2H) suggesting that they are much more  
267 prevalent in the accessory genome compared to the core genome, and once more implicating  
268 SGVs as tools of adaptation and speciation.

269 SGVs were additionally enriched with genes to which no function was assigned by  
270 KEGG ( $p < 10^{-5}$ ; Fig. 2F,G marked by a red star). To overcome this obstacle, we performed  
271 enrichment analysis on word categories from the Ensembl functional annotation<sup>43</sup> of 167,389  
272 genes in the 56 bacteria analyzed (Methods). Bacteriophage- and plasmid-related genes, genes  
273 associated with transposable elements, and genes encoding other horizontal gene transfer  
274 (HGT) mechanisms were strongly enriched in variable- (FDR-corrected  $q < 10^{-4}$ ) and deletion-  
275 SGVs ( $q < 10^{-4}$ ) and strongly depleted from conserved regions ( $q < 10^{-4}$ ), suggesting an important  
276 role for these mechanisms in the formation of these regions. Analysis of Pfam<sup>44</sup> motifs  
277 pertaining to HGT mechanisms (Methods) corroborated this finding and showed an enrichment  
278 of phage-, prophage-, transposon and conjugated-transposon-related motifs in variable- and

279 deletion-SGVs and their depletion from conserved regions ( $q < 10^{-4}$ ). In addition, variable-SGVs  
280 were enriched with antibiotic-producing genes ( $q < 0.005$ ) and deletion-SGVs were enriched with  
281 CRISPR-associated genes ( $q < 0.05$ ) suggesting that these regions function as attainable  
282 microbial tools for interacting with their environment. This analysis also demonstrates how SGV-  
283 Finder, which operates directly at the genomic level, can accommodate analyses with multiple  
284 annotation datasets.

285 To further characterize the potential contribution of SGVs to microbial niche adaptation,  
286 we searched for regions that are associated with fitness of their harboring microbe. As a proxy  
287 for fitness, we calculated bacterial growth rates of 21 bacterial strains with sufficient coverage  
288 and available complete genomes using a method we previously developed that estimates  
289 growth through differences in DNA copy number at the origins and terminus locations created  
290 during DNA replication<sup>45</sup>. We found 44 highly significant associations (surpassing Bonferroni  
291 correction cutoff of  $p < 3 \times 10^{-5}$ ; Fig. 2I; Table S2) of these growth rates with deletion-SGVs within  
292 the same bacteria (Methods). These significant associations span a total of 8 distinct bacteria,  
293 suggesting that certain SGVs may be important for bacterial adaptation and fitness.

294 To better probe the mechanisms potentially underlying this adaptation, we systematically  
295 examined the genetic content of the deletion-SGVs that were significantly associated with  
296 growth, and found a similar pattern to that seen when analyzing all SGVs, with a depletion of  
297 housekeeping functions and enrichment for genes involved with CRISPR-, transposon- and  
298 HGT-associated genes ( $q < 0.05$ ; gene categories based analysis; Methods), as well as a  
299 significant enrichment for genes with unknown functions ( $p < 10^{-5}$ , Fig. S4).

300 We further examined two such regions, which were significantly positively and negatively  
301 associated ( $p < 10^{-10}$  for both) with the growth of the same harboring species (*Eubacterium*  
302 *eligens*; Fig. S5A-D). Notably, the SGV whose presence is negatively associated with the  
303 growth dynamics of the microbial host (Fig. S5A,B) contain genes for flagellin, flagellar hook-  
304 associated protein and lipopolysaccharide (LPS) choline phosphotransferase among a few

305 metabolic genes and response regulators (Table S3). Flagellin and the flagellar hook protein  
306 were shown to elicit strong immune responses in mammals<sup>46,47</sup>, possibly inhibiting bacterial  
307 growth. LPS choline phosphotransferase attaches choline phosphate to the bacterial LPS  
308 molecule, which was shown to increase C-reactive protein-mediated innate immune clearing<sup>48</sup>,  
309 again suggesting possible inhibition of microbial growth. Thus, increased growth rates in  
310 bacteria missing these subgenomic regions may point to loss-of-function adaptation of these  
311 bacteria to the host gut and its immune system. In contrast, the SGVs whose presence was  
312 positively associated with their microbial host growth dynamics (Fig. S5C,D) contained mostly  
313 hypothetical coding genes, but also a gene for antibiotic transport system ATP-binding protein,  
314 whose presence could have a selective advantage in the human host by conferring resistance  
315 to antibiotics<sup>49</sup> (Table S3). These results demonstrate the ability of our methodology to suggest  
316 underlying mechanisms using the genomic context of SGVs.

317 Overall, our results show that SGVs associate with common mechanisms of conjugation,  
318 transposition and phage lysogeny, and may thus be powerful tools of niche adaptation. The  
319 acquisition of bulk genetic material not present in a microbial genome, and changes in copy  
320 number of regions that are, may be much stronger drivers of adaptation than rarely occurring  
321 point mutations. Microbial evolution in densely populated ecosystems such as the human  
322 microbiome may thus be driven strongly by SGVs, which allow incorporation of functional  
323 genetic material conferring higher fitness, and affecting both microbes and host.

324

### 325 **Microbiome subgenomic variation is associated with host disease risk factors**

326

327 To explore the potential relevance of microbiome SGVs to human health, we used data  
328 collected on 887 subjects which includes microbiome profiling alongside detailed blood glucose  
329 measurements over the duration of a week, anthropometric measurements, blood tests, and  
330 medical questionnaires<sup>20,30</sup>. We associated the abundance of variable-SGVs and the presence

331 or absence of deletion-SGVs with multiple metrics of health and metabolic risk factors: mean  
332 arterial blood pressure (MAP); total and HDL cholesterol; waist circumference; body weight;  
333 body mass index (BMI); median glucose levels over the measured week; percent glycated  
334 hemoglobin (HbA1C%); and age. We found 81 (Fig. 3A, S6) and 43 (Fig. 3B) significant  
335 associations at a false discovery rate (FDR)<sup>50</sup> of 0.1 for variable- and deletion-SGVs,  
336 respectively, potentially demonstrating the importance of SGVs not only to the microbe, but also  
337 to the host.

338         Several of the associations of risk factors and SGVs found in this study are in line with  
339 the associations of the harboring microbe. For example, we found five deletion-SGVs in *A.*  
340 *hadrus* to be associated with lower BMI, body weight and waist circumference, and with higher  
341 HDL cholesterol levels (Fig. 3B), and we indeed found this bacteria to be negatively correlated  
342 with body weight ( $p < 10^{-5}$ ), waist circumference ( $p < 10^{-5}$ ), median blood glucose levels ( $p < 10^{-4}$ )  
343 and BMI ( $p < 0.005$ ) and positively correlated with HDL cholesterol levels ( $p < 10^{-7}$ ). Additionally,  
344 this bacteria was previously shown to increase in abundance following a very low calorie diet<sup>51</sup>.  
345 Despite being both correlated with similar risk factors, the association of the highlighted SGV  
346 with risk factors allows us to pinpoint specific regions and mechanism that may underlie the  
347 association.

348         In some cases, we potentially expose novel associations between the microbiome and  
349 disease as some associations between host phenotypes and SGVs do not take the same  
350 direction as the associations of the same phenotypes with the abundances of the harboring  
351 bacteria. For example, three variable-SGVs in *Ruminococcus torques* were negatively  
352 associated with multiple risk factors for the metabolic syndrome (Fig. 3A) but we found *R.*  
353 *torques* abundance to be positively associated with body weight ( $p < 10^{-3}$ ) and BMI ( $p < 0.05$ ), and  
354 it was also positively associated with the metabolic syndrome in a different cohort<sup>52</sup>. Similarly,  
355 several variable-SGVs in *Eubacterium rectale* were positively associated with age (Fig. 3A),  
356 while the relative abundances of *E. rectale* were negatively associated with it ( $p < 10^{-6}$ ). A 2-kbp



357 deletion-SGV in *Faecalibacterium cf. prausnitzii* KLE1255 was positively associated with the  
358 weekly median glucose level (Fig. 3B), and even though *F. prausnitzii* was not significantly  
359 associated with median blood glucose levels in our cohort, two independent studies found it to  
360 be negatively associated with type II diabetes mellitus, a disease for which blood glucose levels  
361 are a major risk factor<sup>21,53</sup>. These seemingly paradoxical associations between SGVs and  
362 disease-risk factors further suggest that SGVs represent a different layer of information  
363 compared to the taxonomic level, one which may assist in obtaining mechanistic insights into  
364 the etiology of gut microbiota-associated metabolic disease.

365

#### 366 **Disease risk-associated SGVs replicate in the Dutch Lifelines DEEP cohort**

367

368 To test the replicability of these associations, we ran ICRA on read assignments from the  
369 Lifelines DEEP cohort, and used the corrected assignments to calculate the coverage and  
370 presence/absence of variable- and deletion-SGVs as defined from the 887-person cohort. We  
371 then calculated the association of these regions with similar host disease risk factors measured  
372 in the Lifelines DEEP cohort, and compared those to the associations with metabolic risk factors  
373 found in our cohort (Methods). Notably, despite presumed inter-cohort differences in genetics,  
374 dietary preferences and lifestyles, potentially also leading to differences in the etiology of  
375 metabolic disease between the two cohorts, more than a third (40 out of 117) of the  
376 associations found in our cohort in microbes also present in the Lifelines cohort were replicated,  
377 while only 4 out of the remaining 77 were significantly associated in the opposite direction (Fig.  
378 3A,B; Fig. S6).

379

380

381 **Disease risk-associated SGVs facilitate an investigation of putative mechanisms**

382

383 As in the case of bacterial adaptation, examining the genetic content of SGVs facilitated a  
384 potentially mechanistic view into the observed phenomena, and we therefore next looked into  
385 the functions encoded in disease risk-associated SGVs. While many SGVs harbor genes that  
386 are of unknown function, we did observe several intriguing functions coded in SGVs associated  
387 with disease risk factors. For example, the existence of a 11-kbp deletion-SGV from *E. rectale* is  
388 associated with higher HbA1C% ( $p < 10^{-4}$ ; total 630 subjects, 377 retaining; Fig. 3C). A close  
389 examination of this region reveals a class 1 CRISPR-Cas system (Fig. 3D). While it is unclear  
390 how a CRISPR system could be directly related to host disease risk factor, we note the  
391 existence of additional three genes of unknown function in this region. Interestingly, subjects  
392 with *E. rectale* harboring this region had a higher abundance of the microbe (Mann-Whitney U  
393  $p < 0.02$ ), which we had previously shown to increase in abundance following a diet designed to  
394 induce high postprandial glucose responses<sup>20</sup>. A 6-kbp variable-SGV from *R. torques* is  
395 inversely associated with weekly median glucose levels ( $R = -0.237$ ,  $p < 10^{-5}$ ; Fig. 3E) and features  
396 several genes encoding phage-associated proteins and additional genes of unknown function,  
397 suggesting that this SGV is a prophage, and that it may carry additional functionality (Fig. 3F).  
398 These genes of unknown function are therefore putatively related to host glucose metabolism,  
399 demonstrating the utility of our methods for generating mechanistic hypotheses.

400 Other intriguing examples for putative mechanisms include a 4-kb deletion-SGV in *A.*  
401 *hadrus* that is significantly associated with lower BMI (median lower by 1.15 kg/m<sup>2</sup> in subjects  
402 retaining the region;  $p < 10^{-4}$ ; total n=681, 405 retaining; Fig. S7A) and body weight (median  
403 lower by 3.5 kg;  $p < 10^{-4}$ ). This SGV contains genes coding for the enzymes ADC synthase (EC  
404 2.6.1.85) and 4-amino-4-deoxychorismate lyase (EC 4.1.3.38), both instrumental in folate  
405 biosynthesis in *A. hadrus* (Fig. S7B, C). An 18-kb deletion-SGV in *Roseburia intestinalis* that is  
406 significantly associated with total cholesterol (median lower by 12.5mmHg for subjects retaining

407 the region;  $p < 10^{-4}$ ;  $n = 262$ , 68 retaining; Fig. S7D) contained multiple beta- and other  
408 glucosidases (Fig. S7E), potentially suggesting microbial adaptation to a fiber-rich host diet. An  
409 8-kb deletion-SGV in *Coprococcus comes* which is significantly associated with BMI (median  
410 higher by 2.4 kg/m<sup>2</sup> for subjects retaining this region;  $n = 450$ ; 292 retaining;  $p < 10^{-5}$ ; Fig. S7F)  
411 and body weight (median higher by 5 kg;  $p < 10^{-4}$ ) contains several ABC transporters with  
412 undetermined substrates of possible future interest (Fig. S7G).

413 Notably, all of the above regions of interest were also detected as SGVs in the Lifelines  
414 DEEP cohort (Fig. S8) and replicate the patterns of deletion or variation across the region that  
415 were detected in our cohort.

416

#### 417 **Carbohydrate metabolism and SCFA biosynthesis gene clusters encoded in a disease** 418 **risk-associated region**

419

420 As one particularly intriguing example, a 31-kbp deletion-SGV in *A. hadrus* was significantly  
421 associated with lower body weight (median 6kg lower for subjects retaining the region;  $p < 10^{-6}$ ;  
422  $n = 681$ , 468 retaining; Fig. 4A), waist circumference (median lower by 4 cm;  $p < 10^{-4}$ ; Fig. S9A)  
423 BMI (median lower by 1.17 kg/m<sup>2</sup>;  $p < 0.001$ ; Fig. S9B), and higher HDL cholesterol (median  
424 higher by 5.7 mg/dL;  $p < 10^{-4}$ ; Fig. S9C), and was well annotated, allowing us to speculate about  
425 its possible role in the microbiome, and demonstrating the potential of SGV-finder detected  
426 regions to expose potential underlying mechanisms.

427 This genomic region encodes two full metabolic modules, seven sugar transporters and  
428 two transcriptional regulators, among several unrelated genes (Fig. 4B). Of the two metabolic  
429 modules, one performs inositol catabolism<sup>54</sup> metabolizing myo-inositol or D-chiro inositol to (a)  
430 glycerone phosphate, a precursor for glyceraldehyde-3-phosphate, a constituent of the  
431 Embden–Meyerhof–Parnas glycolysis pathway<sup>26</sup>; and (b) 3-oxopropanoate, a precursor for  
432 acetyl-CoA. The second metabolic module encoded in this SGV metabolizes 3-

433 hydroxybutanoyl-CoA to butyrate, a short-chain fatty acid (SCFA), while oxidizing an electron-  
434 transferring flavoprotein encoded in the same SGV. The two pathways are connected through a  
435 series of reactions encoded elsewhere in the *A. hadrus* genome (Fig. 4C, Table S4). Of the  
436 sugar transporters, one is specific to the sugar alcohol sorbitol and six were not assigned a  
437 specific target.

438 Combining the information regarding the two metabolic modules and the glucose  
439 transporters in this SGV, we hypothesize that this region is unifunctional, providing the  
440 bacterium with the capability to ferment sugar alcohol such as inositol to SCFAs in an  
441 energetically-favorable procedure. The combined effect of the two metabolic pathways on the  
442 energy metabolism of *A. hadrus* is positive, earning a net gain of 2 ATP- and 2 NADH-  
443 equivalent molecules, where the myo-inositol catabolism module combined with glycolysis and  
444 acetyl-CoA synthesis have a positive energetic effect and the butyrate synthesis module  
445 consumes energy for butyrate production.

446 This 31-kbp deletion-SGV in *A. hadrus* was replicated with the Dutch cohort (Fig. S8),  
447 and so were several of its association with host phenotypes: Dutch individuals harboring the  
448 region exhibiting lower BMI (median lower by 0.9kg/m<sup>2</sup> for individuals retaining the region;  
449 p<0.005; Fig. S9D), body weight (median lower by 4kg.; n=797, 547 retaining; p<0.01), and  
450 waist-to-hip ratio (median lower by 0.017; p<0.001) potentially pointing to a generalized  
451 mechanistic association between SGV and disease-risk.

452 In order to study the metabolic context of this adaptation in a broader ecological context,  
453 we applied mimosa<sup>55</sup> to obtain the metabolic potential of the metagenomes of different subjects  
454 and compared the differences between the community metabolic potential (CMP) of compounds  
455 in subjects for whom the SGV is deleted and for subjects in which it is retained. We found that  
456 free (unphosphorylated) sorbose, mannitol, galactitol and sorbitol are decreased in individuals  
457 retaining the region (FDR adjusted two-sided Mann-Whitney  $U$   $q < 10^{-4}$ ,  $q < 0.01$ ,  $q < 0.05$  and  
458  $q < 0.1$ , respectively; Table S5), whereas sorbose-1-phosphate, mannitol-1-phosphate and

459 sorbitol-6-phosphate are increased ( $q < 10^{-4}$ ,  $q < 0.01$  and  $q < 0.05$ , respectively; Table S5),  
460 altogether demonstrating an association between adaptation in a specific bacteria to the  
461 metabolic state of the microbiome, in the context of metabolic disease risk. As phosphorylation  
462 is used in the phosphotransferase system to prevent sugar diffusion out of the cell, these  
463 predictions support our observed increase in sugar-alcohol transport. Thus, we hypothesize that  
464 the contribution of this SGV to the overall metabolic function of the microbiome is such that it  
465 increases SCFA production from sugars and consequently exerts beneficial effects on the host.

466

## 467 **Discussion**

468

469 In this work we uncover a new facet of host-microbiome interactions in the context of health and  
470 risk of disease. We present ICRA, a metagenomic read assignment algorithm, which we  
471 validate by showing superior read-assignment and comparable bacterial abundance estimation  
472 with respect to state-of-the-art algorithms. We also present SGV-Finder, a genomic coverage-  
473 based algorithm for the detection of SGVs across metagenomic samples. Using this algorithm,  
474 we show that SGVs are highly abundant in the human microbiome, and are largely conserved  
475 across cohorts that differ in their genetic, cultural and dietary backgrounds. SGVs are host-  
476 specific, conserved in the same individual over time and are more conserved in cohabiting  
477 vs. genetically-related individuals. We found that SGVs harbor genes of distinct functions, and  
478 are associated with bacterial growth rates, indicating a potential utility in bacterial adaptation.  
479 Finally, we found that SGVs are associated with numerous host disease risk-factors, many of  
480 which replicated across two independent cohorts, and that they facilitate exploration of genes  
481 varying together, exposing a new layer of putative mechanistic information regarding host-  
482 microbiome interactions, which we highlight by the discovery of a potentially butyrate-producing  
483 SGV in *A. hadrus*.

484 To our knowledge, ICRA is the first metagenomic read assignment algorithm to  
485 introduce the demand that for a genetic element, whether bacteria, genomic region, or gene, to  
486 be considered present in the sample, its genomic sequence should be sufficiently covered by  
487 metagenomic reads. This precondition increases robustness to shared genomic regions,  
488 assembly errors, and phage activation. We note that a challenging problem which ICRA does  
489 not address is the lack of accurate reference genomes for many of the microbial members of the  
490 gut microbiome. De novo long-read approaches to generate reference genomes from  
491 metagenomes such as Molecu<sup>56</sup> and the 10x platform<sup>57</sup> could prove useful in this context.  
492 Combined with ICRA and SGV-Finder these approaches would successfully delineate additional  
493 interpersonal differences in sub-genomic regions of the microbiome.

494 Using SGV-Finder, we show that SGVs are highly abundant in the human microbiome,  
495 with variable regions present in all 56 microbes from 7 different microbial phyla which had  
496 sufficient coverage, 46 of which replicate to a high degree in an independent cohort. Following a  
497 functional analysis of genes in those regions, we hypothesize that the main forces driving SGVs  
498 are bacteriophage infections and microbial mechanisms of conjugation and transposable  
499 elements, as evident from the high abundance of genes performing such functions in SGV  
500 regions. However, many genes found in SGVs, such as antibiotic biosynthesis genes, can  
501 possibly be characterized as passengers to this process of transposition and may have  
502 important roles in the adaptation of microbes to their ecological niche and in communication with  
503 the host. We show many SGVs are strongly linked to microbial growth, a proxy for fitness,  
504 demonstrating the potential functional importance of SGVs in their harboring microbe.

505 Our results show that SGVs also associate with host disease risk. We found more than  
506 120 significant associations between SGVs and multiple metrics of metabolic disease,  
507 highlighting their potential relevance to host health. Notably, more than one third of the  
508 associations testable in an independent cohort were replicated, demonstrating the conserved  
509 association of these SGVs to disease risk. Many of these regions demonstrate associations with

510 host health that are in opposite direction to the associations found between their harboring  
511 microbe and disease risk, indicating that this is a complimentary layer of information to that of  
512 taxonomical abundances.

513         We have closely examined these regions and the genes that they harbor, and  
514 demonstrated the utility of such examination with several SGVs whose genes were well  
515 annotated, including a 31-kbp SGV that was strongly associated with lower metabolic risk  
516 across multiple biomarkers and which we also found to encode a bacterial pathway pertaining to  
517 the transport and fermentation of sugar alcohols to the short chain fatty acid butyrate. SCFAs,  
518 and specifically butyrate, have been previously shown to nourish host intestinal cells<sup>58,59</sup> and  
519 mitigate inflammatory disease<sup>60</sup>. In mice, SCFAs were shown to improve insulin sensitivity and  
520 increase energy expenditure<sup>61</sup>, suggesting that the inclusion of this SGV in the bacterial genome  
521 and thereby the potential boosting of SCFA production may be advantageous for both the  
522 bacteria and host metabolism. We hypothesize that by possessing this SGV, bacteria  
523 demonstrate increased symbiosis with the host, as fermenting sugar alcohols to butyrate  
524 benefits the microbe by producing additional energy and benefits the host with the  
525 advantageous effects of intestinal butyrate.

526         Despite the visible links between this SGV and host metabolism, and between this SGV  
527 and bacterial metabolism, we do not know whether the SGV leads to the observed lean  
528 phenotype or whether the diet, lifestyle and other factors in the host lead to the incorporation or  
529 loss of this SGV. While further research is needed to fully understand the links between host  
530 diet and lifestyle, the microbiome and metabolic disease, this SGV demonstrates the wealth of  
531 mechanistic knowledge obtained through examining genes with variable copy number in their  
532 genomic context and along with neighboring variable genes. This type of analysis, connecting  
533 genomic variation with genetic function, could be instrumental for raising multiple mechanistic  
534 hypotheses about the pathophysiological role of the microbiome. We therefore made our

535 algorithms available for the scientific community and developed an online metagenomic SGV  
536 explorer that will enable further exploration (all available at <http://genie.weizmann.ac.il/SGV/>).

537         The current implementation of both ICRA and SGV-Finder depends on a genomic  
538 reference dataset, which are typically sufficient for human microbiome analyses. Even so, we  
539 note that this is a practical rather than a conceptual approach, as the algorithms are capable of  
540 running on any type of database of genetic elements. Future work could validate and use these  
541 methods following metagenome assembly, ORF prediction and functional prediction stages,  
542 which would allow their application to different host-associated environments and different  
543 realms of microbiology and cellular biology, such as to soil or extreme microbiomes.

544         Our methodology is highly adaptable to any metagenomic scenario and could be used,  
545 for example, to detect SGVs in the soil microbiome and associate them with the presence of  
546 specific nutrients and metabolites to detect candidate biosynthetic gene clusters. Taken  
547 together, our study exposes a new facet of the microbiome that brings us closer to  
548 mechanistically understanding links between microbe and host.



549 **References**

- 550
- 551 1. McCarroll, S. A. & Altshuler, D. M. Copy-number variation and association studies of  
552 human disease. *Nat. Genet.* **39**, S37-42 (2007).
- 553 2. Taniguchi, Y. *et al.* Quantifying *E. coli* proteome and transcriptome with single-molecule  
554 sensitivity in single cells. *Science (80-. )*. **329**, 533–8 (2010).
- 555 3. Perry, G. H. *et al.* Diet and the evolution of human amylase gene copy number variation.  
556 *Nat. Genet.* **39**, 1256–60 (2007).
- 557 4. Pinto, D. *et al.* Functional impact of global rare copy number variation in autism spectrum  
558 disorders. *Nature* **466**, 368–72 (2010).
- 559 5. Malhotra, D. & Sebat, J. CNVs: harbingers of a rare variant revolution in psychiatric  
560 genetics. *Cell* **148**, 1223–41 (2012).
- 561 6. Jacquemont, S. *et al.* Mirror extreme BMI phenotypes associated with gene dosage at the  
562 chromosome 16p11.2 locus. *Nature* **478**, 97–102 (2011).
- 563 7. McCarroll, S. A. *et al.* Deletion polymorphism upstream of IRGM associated with altered  
564 IRGM expression and Crohn’s disease. *Nat. Genet.* **40**, 1107–12 (2008).
- 565 8. Wellcome Trust Case Control Consortium *et al.* Genome-wide association study of CNVs  
566 in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature* **464**, 713–  
567 20 (2010).
- 568 9. Jones, T. A., Hernandez, D. Z., Wong, Z. C., Wandler, A. M. & Guillemin, K. The bacterial  
569 virulence factor CagA induces microbial dysbiosis that contributes to excessive epithelial  
570 cell proliferation in the *Drosophila* gut. *PLOS Pathog.* **13**, e1006631 (2017).
- 571 10. Sokurenko, E. V *et al.* Pathogenic adaptation of *Escherichia coli* by natural variation of  
572 the FimH adhesin. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 8922–6 (1998).
- 573 11. Gill, S. R. *et al.* Insights on Evolution of Virulence and Resistance from the Complete  
574 Genome Analysis of an Early Methicillin-Resistant *Staphylococcus aureus* Strain and a  
575 Biofilm-Producing Methicillin-Resistant *Staphylococcus epidermidis* Strain. *J. Bacteriol.*  
576 **187**, 2426–2438 (2005).
- 577 12. Koeth, R. a *et al.* Intestinal microbiota metabolism of L-carnitine, a nutrient in red meat,  
578 promotes atherosclerosis. *Nat. Med.* **19**, 576–85 (2013).
- 579 13. Han, B. *et al.* Microbial Genetic Composition Tunes Host Longevity. *Cell* **169**, 1249–  
580 1262.e13 (2017).
- 581 14. Smillie, C. S. *et al.* Ecology drives a global network of gene exchange connecting the  
582 human microbiome. *Nature* **480**, 241–4 (2011).
- 583 15. Greenblum, S., Carr, R. & Borenstein, E. Extensive strain-level copy-number variation  
584 across human gut microbiome species. *Cell* **160**, 583–94 (2015).
- 585 16. Tremaroli, V. & Bäckhed, F. Functional interactions between the gut microbiota and host  
586 metabolism. *Nature* **489**, 242–9 (2012).
- 587 17. Swann, J. R. *et al.* Systemic gut microbial modulation of bile acid metabolism in host  
588 tissue compartments. *Proc. Natl. Acad. Sci. U. S. A.* **108 Suppl**, 4523–30 (2011).
- 589 18. LeBlanc, J. G. *et al.* Bacteria as vitamin suppliers to their host: a gut microbiota  
590 perspective. *Curr. Opin. Biotechnol.* **24**, 160–8 (2013).
- 591 19. Levy, M. *et al.* Microbiota-Modulated Metabolites Shape the Intestinal Microenvironment  
592 by Regulating NLRP6 Inflammasome Signaling. *Cell* **163**, 1428–1443 (2015).
- 593 20. Zeevi, D. *et al.* Personalized Nutrition by Prediction of Glycemic Responses. *Cell* **163**,  
594 1079–94 (2015).
- 595 21. Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes.  
596 *Nature* **490**, 55–60 (2012).
- 597 22. Halfvarson, J. *et al.* Dynamics of the human gut microbiome in inflammatory bowel  
598 disease. *Nat. Microbiol.* **2**, 17004 (2017).
- 599 23. Pascal, V. *et al.* A microbial signature for Crohn’s disease. *Gut* **66**, 813–822 (2017).

- 600 24. Rowan, S. *et al.* Involvement of a gut-retina axis in protection against dietary glycemia-  
601 induced age-related macular degeneration. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E4472–  
602 E4481 (2017).
- 603 25. Hsiao, E. Y. *et al.* Microbiota modulate behavioral and physiological abnormalities  
604 associated with neurodevelopmental disorders. *Cell* **155**, 1451–63 (2013).
- 605 26. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic*  
606 *Acids Res.* **28**, 27–30 (2000).
- 607 27. Li, J. *et al.* An integrated catalog of reference genes in the human gut microbiome. *Nat.*  
608 *Biotechnol.* **32**, 834–841 (2014).
- 609 28. Manor, O. & Borenstein, E. Systematic Characterization and Analysis of the Taxonomic  
610 Drivers of Functional Shifts in the Human Microbiome. *Cell Host Microbe* **21**, 254–267  
611 (2017).
- 612 29. Franzosa, E. A. *et al.* Species-level functional profiling of metagenomes and  
613 metatranscriptomes. *Nat. Methods* **15**, 962–968 (2018).
- 614 30. Korem, T. *et al.* Bread Affects Clinical Parameters and Induces Gut Microbiome-  
615 Associated Personal Glycemic Responses. *Cell Metab.* **25**, 1243–1253.e5 (2017).
- 616 31. Tigchelaar, E. F. *et al.* Cohort profile: LifeLines DEEP, a prospective, general population  
617 cohort study in the northern Netherlands: study design and baseline characteristics. *BMJ*  
618 *Open* **5**, e006772 (2015).
- 619 32. Zhernakova, A. *et al.* Population-based metagenomics analysis reveals markers for gut  
620 microbiome composition and diversity. *Science (80-. )*. **352**, 565–569 (2016).
- 621 33. Kelly, C. J. *et al.* Crosstalk between Microbiota-Derived Short-Chain Fatty Acids and  
622 Intestinal Epithelial HIF Augments Tissue Barrier Function. *Cell Host Microbe* **17**, 662–71  
623 (2015).
- 624 34. Donohoe, D. R. *et al.* The Microbiome and Butyrate Regulate Energy Metabolism and  
625 Autophagy in the Mammalian Colon. *Cell Metab.* **13**, 517–526 (2011).
- 626 35. Blacher, E., Levy, M., Tatirovsky, E. & Elinav, E. Microbiome-Modulated Metabolites at  
627 the Interface of Host Immunity. *J. Immunol.* **198**, 572–580 (2017).
- 628 36. Mende, D. R. *et al.* proGenomes: a resource for consistent functional and taxonomic  
629 annotations of prokaryotic genomes. *Nucleic Acids Res.* **45**, D529–D534 (2017).
- 630 37. Sczyrba, A. *et al.* Critical Assessment of Metagenome Interpretation—a benchmark of  
631 metagenomics software. *Nat. Methods* **14**, 1063–1071 (2017).
- 632 38. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification  
633 using exact alignments. *Genome Biol.* **15**, R46 (2014).
- 634 39. Liu, B., Gibbons, T., Ghodsi, M. & Pop, M. MetaPhyler: Taxonomic profiling for  
635 metagenomic sequences. in *2010 IEEE International Conference on Bioinformatics and*  
636 *Biomedicine (BIBM)* 95–100 (IEEE, 2010). doi:10.1109/BIBM.2010.5706544
- 637 40. Truong, D. T. *et al.* MetaPhlan2 for enhanced metagenomic taxonomic profiling. *Nat.*  
638 *Methods* **12**, 902–903 (2015).
- 639 41. Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: estimating species  
640 abundance in metagenomics data. *PeerJ Comput. Sci.* **3**, e104 (2017).
- 641 42. Rothschild, D. *et al.* Environment dominates over host genetics in shaping human gut  
642 microbiota. *Nature* **555**, 210–215 (2018).
- 643 43. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2018).
- 644 44. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res.*  
645 (2018). doi:10.1093/nar/gky995
- 646 45. Korem, T. *et al.* Growth dynamics of gut microbiota in health and disease inferred from  
647 single metagenomic samples. *Science* **349**, 1101–6 (2015).
- 648 46. Hayashi, F. *et al.* The innate immune response to bacterial flagellin is mediated by Toll-  
649 like receptor 5. *Nature* **410**, 1099–1103 (2001).
- 650 47. Shen, Y. *et al.* Flagellar Hooks and Hook Protein FlgE Participate in Host Microbe

651 Interactions at Immunological Level. *Sci. Rep.* **7**, 1433 (2017).  
652 48. Weiser, J. N. *et al.* Phosphorylcholine on the lipopolysaccharide of *Haemophilus*  
653 *influenzae* contributes to persistence in the respiratory tract and sensitivity to serum  
654 killing mediated by C-reactive protein. *J. Exp. Med.* **187**, 631–40 (1998).  
655 49. Ross, J. I. *et al.* Inducible erythromycin resistance in staphylococci is encoded by a  
656 member of the ATP-binding transport super-gene family. *Mol. Microbiol.* **4**, 1207–1214  
657 (1990).  
658 50. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and  
659 Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B*  
660 *(Methodological)* **57**, 289–300 (1995).  
661 51. Ott, B. *et al.* Effect of caloric restriction on gut permeability, inflammation markers, and  
662 fecal microbiota in obese women. *Sci. Rep.* **7**, 11955 (2017).  
663 52. Zupancic, M. L. *et al.* Analysis of the gut microbiota in the old order Amish and its relation  
664 to the metabolic syndrome. *PLoS One* **7**, e43052 (2012).  
665 53. Karlsson, F. H. *et al.* Gut metagenome in European women with normal, impaired and  
666 diabetic glucose control. *Nature* **498**, 99–103 (2013).  
667 54. Yoshida, K. *et al.* myo-Inositol catabolism in *Bacillus subtilis*. *J. Biol. Chem.* **283**, 10415–  
668 24 (2008).  
669 55. Noecker, C. *et al.* Metabolic Model-Based Integration of Microbiome Taxonomic and  
670 Metabolomic Profiles Elucidates Mechanistic Links between Ecological and Metabolic  
671 Variation. *mSystems* **1**, (2016).  
672 56. White, R. A. *et al.* MolecuLo Long-Read Sequencing Facilitates Assembly and Genomic  
673 Binning from Complex Soil Metagenomes. *mSystems* **1**, (2016).  
674 57. Eisenstein, M. Startups use short-read data to expand long-read sequencing market. *Nat.*  
675 *Biotechnol.* **33**, 433–435 (2015).  
676 58. McNeil, N. I., Cummings, J. H. & James, W. P. Short chain fatty acid absorption by the  
677 human large intestine. *Gut* **19**, 819–22 (1978).  
678 59. Bergman, E. N. Energy contributions of volatile fatty acids from the gastrointestinal tract  
679 in various species. *Physiol. Rev.* **70**, 567–90 (1990).  
680 60. Harig, J. M., Soergel, K. H., Komorowski, R. A. & Wood, C. M. Treatment of diversion  
681 colitis with short-chain-fatty acid irrigation. *N. Engl. J. Med.* **320**, 23–8 (1989).  
682 61. Gao, Z. *et al.* Butyrate improves insulin sensitivity and increases energy expenditure in  
683 mice. *Diabetes* **58**, 1509–17 (2009).  
684 62. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate  
685 genomic comparisons that enables improved genome recovery from metagenomes  
686 through de-replication. *ISME J.* **11**, 2864–2868 (2017).  
687 63. Marco-Sola, S., Sammeth, M., Guigó, R. & Ribeca, P. The GEM mapper: fast, accurate  
688 and versatile alignment by filtration. *Nat. Methods* **9**, 1185–8 (2012).  
689 64. Suez, J. *et al.* Artificial sweeteners induce glucose intolerance by altering the gut  
690 microbiota. *Nature* **514**, 181–6 (2014).  
691 65. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using  
692 DIAMOND. *Nat. Methods* **12**, 59–60 (2015).  
693 66. Potter, S. C. *et al.* HMMER web server: 2018 update. *Nucleic Acids Res.* **46**, W200–  
694 W204 (2018).  
695

696 **Acknowledgements**

697 We thank the Segal group members and members of the Center for Studies in Physics and  
698 Biology for discussions; and participants and staff of the Lifelines DEEP cohort for their  
699 collaboration. E.S. is supported by the Crown Human Genome Center; the Else Kroener  
700 Fresenius Foundation; D. L. Schwarz; J. N. Halpern; L. Steinberg; and grants funded by the  
701 European Research Council and the Israel Science Foundation. D.Z. is supported by the James  
702 S. McDonnell Foundation. D.Z. and T.K were partly supported by the Israeli Ministry of Science  
703 and Tehcnology. Lifelines DEEP was made possible by grants from the Top Institute Food and  
704 Nutrition (GH001) to C.W. C.W. is funded by a European Research Council (ERC) advanced  
705 grant (FP/2007-2013/ERC grant 2012-322698), a Netherlands Organization for Scientific  
706 Research (NWO) Spinoza prize (NWO SPI 92-266) and the Stiftelsen Kristian Gerhard Jebsen  
707 foundation (Norway). A.Z. holds a Rosalind Franklin Fellowship (University of Groningen), ERC  
708 starting grant (715772) and NWO Vidi grant (178.056). J.F. is funded by an NWO Vidi grant  
709 (NWO-VIDI 864.13.013). A.Z. and J.F. are also funded by CardioVasculair Onderzoek  
710 Nederland (CVON 2012-03).

711

712 **Author contributions**

713 T.K. and D.Z. conceived the project, designed the study, designed and conducted all analyses,  
714 interpreted the results, and wrote the manuscript. T.K. and D.Z. equally contributed to this work  
715 and are listed in random order. A.G. and N.B. developed methods. A.K., J.F., C.W. and A.Z.  
716 performed the analyses of the Dutch cohort. M.L.-P and A.W. did experimental work on the 7  
717 strains. A.W. designed the study. E.S. conceived and directed the project and analyses,  
718 designed the analyses, interpreted the results and wrote the manuscript.

719

720

721

722 **Methods**

723

724 Reference database preprocessing

725 We downloaded the EMBL progenomes<sup>36</sup> 5306 representatives dataset and used dRep<sup>62</sup> to  
726 calculate distances between genomes. Next, we applied ward hierarchical clustering with a  
727 Euclidean distance metric to the dRep distance matrix, calculated a dendrogram and retrieved  
728 the cut tree at a height of 0.15 (corresponding to approximately 15% dissimilarity in genome  
729 sequence) resulting in 3953 clusters. As a representative species for each cluster we chose the  
730 genome with the minimal distance to all other genomes in the cluster. In clusters with only two  
731 members, we chose one randomly. Database taxa and assembly accession numbers are listed  
732 in Table S6.

733

734 Metagenomic samples - Israeli cohort

735 We obtained metagenomic samples from two studies<sup>20,30</sup> (accession numbers ENA:  
736 PRJEB11532, ENA: PRJEB17643). In the latter study<sup>30</sup>, only baseline samples were used  
737 (before the intervention took place).

738

739 Gut microbiome analysis

740 To prevent bias generated by analyzing single- and paired-end sequenced samples together,  
741 we took the first end of all samples, and trimmed each read to a maximal length of 75bp (100bp  
742 for Lifelines DEEP cohort). We filtered metagenomic reads containing Illumina adapters, filtered  
743 low quality reads and trimmed low quality read edges. We detected host DNA by mapping with  
744 GEM<sup>50</sup> to the Human genome with inclusive parameters, and removed those reads. We  
745 randomly subsampled all samples to 10M reads, and removed samples with less than 10M  
746 reads from subsequent analyses.

747 For MetaPhlan2 comparisons, we obtained relative abundances (RA) from metagenomic  
748 sequencing via MetaPhlan2<sup>40</sup> with default parameters. For Kraken<sup>38</sup> comparisons, we built a  
749 custom Kraken database using our preprocessed database and subsequently classified with  
750 default parameters and generated a Kraken report. For Bracken<sup>41</sup> abundance estimation, we  
751 generated a Bracken-database file using bracken-build on the above Kraken database with a  
752 kmer length of 31 and read length of 100bp and used it to estimate abundance using the  
753 aforementioned Kraken report.

754

#### 755 ICRA - Iterative Coverage-based Read Assignment algorithm

756 We devised an iterative read assignment algorithm which uses read assignments and  
757 sequencing qualities to calculate the sequencing coverage depth along genomic elements (i.e.,  
758 bacterial genomes or gene sequences) in the microbiome. Sequencing coverage is then used to  
759 both qualitatively assess the presence or absence of each microbe by demanding a minimum  
760 coverage across each genomic element, as well as to quantitatively estimate the relative  
761 abundance of each microbe disregarding outlier genomic positions where extremely high or low  
762 coverage exists. Microbial relative abundances are subsequently used to estimate read  
763 assignments, repeating the process to convergence.

764 For a more formal description of our algorithm, let  $i = 1, 2, \dots, R$  be the index of  
765 metagenomic reads in a sample; let  $j = 1, 2, \dots, G$  be the index of genomic elements in a  
766 database of such elements; and  $p(i, j)_k = p(i, j)_1, p(i, j)_2, \dots, p(i, j)_{N(i, j)}$  be all the possible  
767 alignment positions for read  $i$  in genomic element  $j$  ( $N(i, j)$  is the total number of possible  
768 alignments of  $i$  to element  $j$ , in most cases only one) such that if metagenomic read  $i$  is assigned  
769 to position  $p(i, j)_k$ , it spans an alignment from  $p(i, j)_k$  to approximately  $p(i, j)_k + \rho_i$ , where  $\rho_i$  is  
770 the length of read  $i$ .

771 Our goal is, therefore to find, for each i, j and k,  $\lambda_{i,j,k}$ , an indicator variable for the origin  
772 of read i:

773  $\lambda_{i,j,k} = 1$  iff read i originated from genomic element j in position  $p(i,j)_k$   
774 To approximate  $\lambda_{i,j,k}$ , we calculate, for each read the probability  $\delta_{i,j,k}$  that read i  
775 originated from the genomic element j at position  $p(i,j)_k$ , as:

$$776 \quad \delta_{i,j,k} = \frac{\pi_j \theta_j q_{i,j,k}}{\sum_{l,m} \pi_l \theta_l q_{i,l,m}}$$

777 Where:

778 •  $\pi_j = f(\{\delta_{i,j,k} \forall i, k\})$

779  $\pi_j$  is the estimated relative abundance of the genomic element j. In the initial iteration of  
780 the algorithm,  $\pi_j$  is calculated by counting all reads mapped to genomic element j and  
781 then dividing the result by the total number of reads. Reads mapped to multiple genomic  
782 elements are initially distributed according to quality of mapping (see q below).  
783 Function f divides the genomic element j to bins of a size defined by the user (1kbp by  
784 default), calculates bin coverage by summing all  $\delta_{i,j,k}$  (from previous iteration) in each  
785 genomic bin, and calculates  $\pi_j$  as the median of the n% most closely covered bins in the  
786 genomic element, with n defined by the user. For the default n of 60, we calculate the  
787 difference between the most covered bin and the least covered bin for every subset  
788 spanning 60% of the bins, find the subset in which the difference is minimal, and take its  
789 median coverage. This median is then multiplied by the number of reads to reach an  
790 estimation of the true number of reads originating from the genomic element j. This  
791 number is then divided by the total number of reads assigned to all genomic elements to  
792 calculate  $\pi_j$ .  $\pi_j$  is then normalized by the length of the genomic element (or its harboring  
793 microbe), but this could be turned off by the user.

794 •  $\theta_j = \sum_{i,k} I_{i,j,k}$

795 Where  $I_{i,j,k} = 1$  iff  $\delta_{i,j,k} > \delta_{i,l,m} \forall l, m$

796 i.e., the sum of reads preferentially mapped to this genomic element. This parameter  
797 facilitates faster convergence but results in reduced accuracy, and is suggested for use  
798 in case of very large reference datasets. With default ICRA parameters, it will be set to 1  
799 (and therefore ignored).

800 •  $q_{i,j,k} = \prod_{pos=0}^{\rho} qual(pos)^{\mu(i,j,p(i,j)_k+pos)} (1 - qual(pos))^{1-\mu(i,j,p(i,j)_k+pos)}$

801 is the probability of a correct mapping, given the mismatches in the read and the  
802 sequencing qualities. Where  $qual(pos)$  is the probability of correct sequencing in position  
803  $pos$  calculated from fastq qualities and  $\mu(i,j,p(i,j)_k + pos) = 1$  if there is a match  
804 between nucleotide in position  $pos$  in read  $i$  to the one in position  $p(i,j)_k+pos$  in genomic  
805 element  $j$  and 0 otherwise.

806 • The term  $\sum_{l,m} \pi_l \theta_l q_{i,l,m}$  is used to normalize  $\delta_{i,j,k}$  such that the sum of all possible  
807 assignments of read  $i$  equals 1, where  $l$  and  $m$  refer to all possible genomic elements  
808 and positions thereof to which read  $i$  is mapped.

809 If  $\delta_{i,j,k}$  is lower than a user-set parameter  $\epsilon$ , with a default of  $10^{-6}$ , this specific mapping is  
810 removed from subsequent analysis thereby reducing noise typically originating by highly  
811 homologous regions from in subsequent iterations.

812

### 813 CAMI dataset comparison

814 We downloaded all 180bp-spaced toy datasets for the 1st CAMI challenge<sup>37</sup> from the CAMI  
815 challenge website (<https://data.cami-challenge.org/participate>). We created a database of all  
816 taxonomic entities in CAMI using NCBI taxon IDs provided for all gold-standard abundances.  
817 We indexed this database using GEM indexer<sup>63</sup> and mapped all metagenomic reads to the  
818 indexed database using GEM mapper. In the baseline setting, read assignment was not  
819 corrected using ICRA, and the assignment of reads that were mapped to more than one  
820 genome was a uniform division between these genomes. In the ICRA-corrected setting, read  
821 assignment was given by applying ICRA to GEM mapper output. For MetaPhyler<sup>39</sup> read  
822 classification, we created a MetaPhyler classifier based on the same CAMI reference database



823 using the *buildMetaphyler.pl* command with a sequence length of 100bp and classified CAMI  
 824 reads using the *runClassifier.pl* command with default parameters. For Kraken<sup>38</sup> comparison,  
 825 we built a custom Kraken database based on the same CAMI reference database and ran  
 826 Kraken as above. The four resulting assignment sets were compared to the gold standard  
 827 provided by CAMI to derive correct assignment ratios.

828

829 Bacterial strain culture and sequencing

830 The following strains were obtained and grown in the following conditions:

Species	Strain ID	Growth condition – Medium	Growth condition - Temp	Growth to saturation
<i>Lactobacillus gasseri</i>	ATCC 33323	Lactobacillus MRS agar	37°C	24 hrs
<i>Enterococcus faecalis</i>	ATCC 29212	ATCC Medium 44	37°C	overnight
<i>Streptococcus cristatus</i>	ATCC 51100	ATCC Medium 44	37°C	<24 hrs
<i>Akkermansia muciniphila</i>	ATCC BAA-835, DSM 22959	DSM medium 104 + 0.05% mucin or ATCC medium 44	37°C	72 hrs
<i>Cellulomonas flavigena</i>	ATCC 482, DSM 20109	DSM 53 or ATCC Medium: 3 Nutrient Agar/Broth	30°C	72 hrs
<i>Brachybacterium faecium</i>	ATCC 43885, DSM 4810	DSM 92 or ATCC Medium: 3 Nutrient Agar/Broth	30°C	72 hrs
<i>Alistipes finegoldii</i>	DSM 17242	DSM medium 104 + vitamin solution (see medium 131) or 693	37°C	> 24 hrs

831

832 Strains were grown to stationary phase as listed in the table. DNA was extracted using  
833 QIAgen DNAeasy Blood & Tissue kit (Cat# 69504) by the protocol using pretreatment of Gram-  
834 positive or Negative bacteria following purification of total DNA from animal tissues.

835 Following that, 100 ng of DNA was sonicated using Covaris E220X and and Illumina  
836 library was prepared for each strain as previously described<sup>64</sup>. The seven strains were  
837 sequenced to a minimum depth of 3M reads by a NextSeq® 500 machine with Illumina NS  
838 500/550 High Output V2 75 cycle kit. Data was deposited to ENA, accession ENA:  
839 PRJEB25194.

840

#### 841 SGV detection - preprocessing

842 We mapped metagenomic reads to the reference database of 3953 representative microbial  
843 genomes detailed above and corrected read assignments using ICRA. All scaffolds from each  
844 microbial genome were concatenated and subsequently divided into 1 kbp bins. For each  
845 genome in each microbial sample, we counted the number of reads mapped to each of the bins.  
846 In the rare case in which ICRA produces a distribution of probabilities of different read  
847 assignment for a specific read rather than a deterministic assignment, we determined the read  
848 count that was added to each bin using the probability of assignment calculated by ICRA.  
849 Microbes with a median coverage smaller than 10 reads per bin were discarded from  
850 subsequent analyses. In addition, we removed microbes in which the median bin coverage  
851 across samples was lower than one read for more than 30% of the bins.

852

#### 853 Detection of deletion SGVs

854 We examined the coverage in each metagenomic bin across all samples to detect regions that  
855 were deleted from some individuals and retained in others. To this end, for each microbe in  
856 each sample, we calculated a histogram of coverage across all metagenomic bins. We then  
857 searched for a trough, separating bins whose coverage is close to 0 from bins whose coverage

858 is close to the median across the microbe, which we previously demanded to be greater than 10  
859 reads. The position of the trough separates the two modes of the distribution, between bins  
860 which were deleted (number of reads per bin smaller than the trough position) and retained  
861 (number of bins greater than the trough position). To mark a bin as a potential deletion-SGV, we  
862 demanded that it be deleted in 25-75% of samples. We concatenated adjacent deletion-SGV  
863 bins into stretches based on bin cooccurrence dissimilarity, defined as the proportion of samples  
864 which are in disagreement on the deletion-state of the two bins being compared (wherein one  
865 bin is deleted and one is retained for the same sample) out of all samples that harbor the  
866 microbe. Bins were concatenated to an existing stretch if they had an average cooccurrence  
867 dissimilarity lower than 0.25 with all the bins in the stretch, and that the newly created stretch is  
868 deleted in 25-75% of samples. We then clustered deletion SGV stretches belonging to the same  
869 microbe based on cooccurrence. First, we calculated a cooccurrence dissimilarity matrix for any  
870 two bins within the microbe (calculated as 1 minus the cooccurrence metric defined above).  
871 Next, using this bin-dissimilarity matrix we calculated a region dissimilarity matrix by calculating  
872 the average distance between all bins of one region to all bins of the other region. We next  
873 calculated linkage over the bin-dissimilarity matrix using the 'average' method of the  
874 cluster.hierarchy.linkage function in scipy v1.1.0 and divided into clusters with maximal  
875 cooccurrence dissimilarity of 0.25.

876

#### 877 Detection of variable SGVs

878 For each microbe, we first removed all bins that were deleted in more than 95% of subjects. We  
879 examined the coverage in each remaining metagenomic bin across all samples to detect  
880 regions with variable coverage. To this end, we standardized the coverage across all non-  
881 deleted bins of a single microbe in each sample by subtracting the mean coverage and dividing  
882 by the standard deviation. Next, for each bin, we fit a beta-prime distribution over all samples  
883 and marked bins whose value is in the top 5th percentile of the fit distribution as variable SGV.

884 We concatenated adjacent variable SGVs into stretches if their average correlation (Spearman)  
885 with all bins in the stretch was higher than 0.75 and the resulting stretch was in the top 5th  
886 percentile of the beta-prime fit distribution of the resulting bin size. We then clustered variable  
887 SGV stretches similarly to deletion SGV stretches, with a dissimilarity metric calculated as  $1 -$   
888  $((\rho(u,v)+1)/2)$ , where  $\rho$  is the Spearman correlation and  $u, v$  are the bin vectors being compared;  
889 and threshold 0.125. This roughly corresponds to an average Spearman correlation threshold of  
890 0.75.

891

#### 892 Detection of conserved regions

893 For each microbe in each sample, we detected retained / deleted bins as above and defined  
894 conserved regions to be stretches of bins that were deleted in less than 1% of samples.

895

#### 896 Analysis of replication in Dutch Lifelines DEEP cohort

897 To analyze the overlap between SGVs detected in the Israeli cohort to those detected in the  
898 Lifelines DEEP cohort, we ran ICRA and SGV-Finder independently on 1020 out of 1135  
899 samples from the Lifelines DEEP cohort (EGA: EGAS00001001704) that had more than 10M  
900 reads, and computed the percent of overlap between regions in both cohorts. To analyze  
901 replication of associations between cohorts, we calculated for each SGV region in the Israeli  
902 cohort, its presence / absence (deletion SGV) or abundance (variable SGV) in the Lifelines  
903 DEEP cohort. We then tested the association of these regions with mean arterial pressure,  
904 waist-to-hip ratio (stand in for the Israeli cohort waist circumference), body weight, BMI, fasting  
905 glucose (stand in for the Israeli cohort median glucose), glycated hemoglobin, age, total and  
906 HDL cholesterol measured in the Lifelines DEEP cohort, using a Mann-Whitney  $U$  test (deletion  
907 SGVs) or the Spearman correlation (variable SGV).

908

909

910 Calculation of SGV conservation in cohabiting and related individuals

911 We calculated Spearman correlations between the deletion- and variable-SGV vectors of 39  
912 pairs of individuals registered in our cohort as living in the same house. To calculate SGV  
913 retention in first degree relatives, we calculated these correlations in 38 pairs of individuals  
914 whose genomic SNP-based similarity<sup>42</sup> was between 40 and 60%.

915

916 Functional enrichment analysis

917 This analysis was performed similarly yet separately to variable-SGVs, deletion-SGVs,  
918 conserved regions, and regions significantly associated with the PTR of their harboring microbe.  
919 For brevity, we collectively term them “regions”. We examined all gene annotations for all  
920 microbial genomes analyzed using Ensembl functional annotation<sup>43</sup> available through  
921 progenomes<sup>36</sup>, and annotated orphan ORFs by mapping the protein sequence to all KEGG<sup>26</sup>  
922 protein sequences using DIAMOND<sup>65</sup> and selecting the top result with e-value<10<sup>-6</sup> and at least  
923 50% identity. We then used KEGG annotations to assign genes to modules, and calculated the  
924 following textual categories by searching the progenomes gene function annotation using the  
925 following regular expressions:

926 **Transposon:** transpos\S\*|insertion|Tra[A-Z]|Tra[0-9]|IS[0-9]|conjugate transposon

927 **Plasmid:** relax\S\*|conjug\S\*|mob\S\*|plasmid|type IV|chromosome partitioning|chromosome segregation

928 **Phage:** capsid|phage|tail|head|tape measure|antiterminatio

929 **Other HGT mechanisms:**

930 integrase|excision\S\*|exonuclease|recomb|toxin|restrict\S\*|resolv\S\*|topoisomerase|reverse transcrip

931 **Carbohydrate active:** glycosyltransferase|glycoside

932 hydrolase|xylan|monooxygenase|rhamnos\S\*|cellulose|sialidase|\S\*ose(\s|-

933 )|acetylglucosaminidase|cellobiose|galact\S\*|fructose|aldose|starch|mannose|mannan\S\*|glucan|lyase|glycosyltransfe

934 rase|glycosidase|pectin|SusD|SusC|fructokinase|galacto\S\*|arabino\S\*

935 **Antibiotic resistance:** azole resistance|antibiotic resistance|TetR|tetracycline resistance|VanZ|betalactam\S\*|beta-

936 lactam|antimicrob\S\*|antibio\S\*

937 We searched for genes containing Pfam<sup>44</sup> modules with the keywords 'phage', 'prophage',  
938 'transposon', 'conjugative transposon' using hmmscan (HMMER v3.1<sup>66</sup>) with cutoff 1e-5. We  
939 next counted, for each KEGG module, KEGG brite functional category, progenomes textual  
940 gene category and Pfam keyword category the number of genes included and excluded in all  
941 regions combined across all microbes. As the location of genes along microbial genomes is not  
942 random p-values were calculated by permutations. In each permutation the sizes of both the  
943 regions and the gaps between them were preserved but their ordering was randomly shuffled,  
944 followed by examinations of genes in these regions and comparison of the number of included  
945 and excluded gene in each KEGG module, brite functional category, etc., to the number found  
946 without randomization. This was performed 1000 times.

947

#### 948 Calculation of microbial growth rates

949 Microbial growth rates were quantified as peak-to-trough ratio (PTR) using the method and  
950 software provided in ref.<sup>45</sup>. PTRs were calculated for all the strains that were found to contain at  
951 least one deletion-SGV and that whose reference genome sequence was complete (i.e., not  
952 fragmented to contigs, as required by the PTR method<sup>45</sup>), skipping the step of selecting a  
953 representative strain per species. Mann-Whitney *U*-test was ran between PTRs of a bacteria in  
954 samples in which it contained a certain deletion-SGV and PTRs of the same bacteria in samples  
955 in which the same region was deleted, provided that at least 25 samples of each kind were  
956 present.

957

#### 958 SGV explorer

959 SGV explorer, presented in Figure S3 and accessible through  
960 <https://genie.weizmann.ac.il/SGV/>, was created using bokeh for Python  
961 (<http://bokeh.pydata.org>)

962

963 Code availability

964 ICRA, SGV-Finder, and the SGV Browser are available through github at  
965 <https://github.com/segalab/SGVFinder>.

966

967 Data availability

968 The 7 strains samples used in Fig. 1C are available through ENA, accession ENA: PRJEB2519.

969 The 887 samples are publicly available through ENA, accession numbers ENA: PRJEB11532,

970 ENA: PRJEB17643.

971 **Figure Legends**

972

973 **Figure 1. Superior assignment of metagenomic reads using the Iterative Coverage-based**

974 **Read-Assignment (ICRA) algorithm.** (A) Illustration of our computational pipeline. (B) Bar-

975 plots (bar, mean; whiskers, standard deviation) of the ratio of correct read assignment per

976 taxonomy level with no assignment correction (blue) or following assignment correction with

977 ICRA (yellow), Kraken<sup>38</sup> (red) or MetaPhyler<sup>39</sup> (green). \* two-sided Mann-Whitney  $U$   $p < 0.05$ ,

978  $**p < 0.01$  (C) Dot-plot of the calculated relative abundance of 7 bacterial species in 100

979 samples, using either ICRA (yellow), MetaPhlAn2<sup>40</sup> (blue), or Bracken<sup>41</sup> (red), as compared to

980 the true relative abundances. Inset shows a violin plot (white dot, median; black box, IQR) of

981 Bray-Curtis dissimilarities between the estimates of each method and the true abundances. \*\*

982 two-sided Wilcoxon signed-rank  $p < 0.01$  \*\*\*\*  $p < 10^{-4}$

983

984 **Figure 2. Sub-Genomic Variation (SGV) is prevalent in the human microbiome, replicable**

985 **across cohorts and associated with specific functions.** (A) Heatmap showing the number of

986 subjects with SGVs (yellow color scale), the number of SGV regions (green color scale), the

987 mean SGV size (blue color scale) and the fraction of the genome that is variable (red color

988 scale), for each microbe analyzed, along with their phylogenetic tree. (B-C) Heatmap (B) and

989 swarm plot (C) showing the genomic length percentage of variable and deletions SGVs

990 replicated in the Lifelines DEEP cohort for each microbe analyzed. (D-E) Boxplot (box, IQR;

991 whiskers,  $1.5 \times \text{IQR}$ ) of the distribution of the correlations between variable- (D) or deletion-SGV

992 (E) across different subjects (green), within the same subject (blue), among cohabiting subjects

993 (yellow) and among pairs of siblings or parents/children (red). \*\*- two-sided Mann Whitney  $U$

994  $p < 0.01$  \*\*\* $p < 0.001$  \*\*\*\* $p < 10^{-5}$ . (F-H) Fold change (x-axis) and statistical significance (Methods)

995 of the enrichment of functional KEGG modules in variable-SGVs (F), deletion-SGVs (G) and

996 conserved regions (Methods; H). (I) Difference in median value (x-axis) and statistical



997 significance in a Mann-Whitney  $U$  test (y-axis) comparing calculated bacterial growth rates  
998 (PTR<sup>45</sup>) under deletion versus retention of SGV.

999

1000 **Figure 3. SGVs are associated with disease risk and these associations replicate across**

1001 **cohorts.** (A-B) Heatmap of statistically significant correlations (Spearman  $p < 0.001$ , FDR

1002 adjusted at 0.1) between disease risk factors and variable-(A) or deletion-SGVs (B). Stars

1003 signify associations replicated (yellow), replicated using a different variable (orange) or

1004 reversed (gray) in the Lifelines DEEP cohort. Striped stars denote associations from the same

1005 bacteria that were collapsed for display purposes (see Figure S6 for full heatmap). (C) Boxplot

1006 (Box, IQR; whiskers, IQR\*1.5) of glycated hemoglobin (HbA1C%) in individuals harboring an 11-

1007 kbp deletion in the *E. rectale* genome (blue) and individuals with no deletion (maroon);  $p$  - Two-

1008 sided Mann-Whitney  $U$  test. (D) (top) Deletion rate across the cohort (y-axis) along a genomic

1009 region of *E. rectale* (x-axis). (bottom) gene locations (arrows) colored according to function

1010 (legend). (E) Scatterplot showing the correlation between the abundance of a 6-kbp variable-

1011 SGV in *R. torques* and weekly median glucose levels;  $p$  - Spearman correlation  $p$ -value. (F)

1012 (top) depiction of standardized variability (y-axis; plotted lines, percentiles 1, 25, 50, 75 and 99)

1013 along a genomic region of *R. torques* (x-axis). (bottom) gene locations (arrows) colored

1014 according to function (legend).

1015

1016 **Figure 4. A 31kbp deletion-SGV in *Anaerostipes hadrus* is associated with reduced**

1017 **weight.**

1018 (A) Boxplot (Box, IQR; whiskers, IQR\*1.5) of body weight in individuals harboring a 31-kbp

1019 deletion in the *A. hadrus* genome (blue) and individuals with no deletion (maroon).  $p$  - Two-

1020 sided Mann-Whitney  $U$  test. (B) Same as Fig. 3D for this genomic region of *A. hadrus*. (C)

1021 Depiction of the metabolic pathways encoded in the region, which turns inositol to the short-

1022 chain fatty acid butyrate. Note correspondence of enzyme commission (EC) numbers with panel  
1023 B.

1024  
1025 **Figure S1. ICRA reduces ambiguous assignments and noise.** (A) Boxplot (Box, IQR;  
1026 whiskers, 10th and 90th percentiles) of ambiguous read assignment ratios of 887 samples<sup>20,30</sup>  
1027 mapped to a reference database of 3953 representative microbial genomes (Methods) before  
1028 (blue) and after (yellow) ICRA correction. (B,C) Bar-plots (bar, mean; whiskers, standard  
1029 deviation) of the ratio of incorrect read assignment per taxonomy level with no correction (blue)  
1030 or following assignment correction with ICRA (yellow), Kraken (red) or MetaPhyler (green) for  
1031 CAMI medium complexity (B; n=3) and low complexity (C; n=1) datasets. Note that MetaPhyler  
1032 did not provide sub-species level read assignments.

1033  
1034 **Figure S2.** (A-G) Dot-plot of the calculated relative abundance (y-axis) of *A. muciniphila* (A), *A.*  
1035 *finnegoldii* (B), *B. faecium* (C), *C. flavigena* (D), *E. faecalis* (E), *L. gasseri* (F) and *S. cristatus* (G)  
1036 in 100 samples, using either ICRA (yellow), MetaPhlAn (blue), or Bracken (red), as compared to  
1037 the true relative abundances (x-axis). R<sup>2</sup> was calculated using Pearson correlation.

1038  
1039 **Figure S3.** (A-B) Illustration of the online SGV explorer available at  
1040 <http://genie.weizmann.ac.il/SGV/>, spanning the entire *R. torques* genome (A) and spanning a  
1041 26-kbp region of the genome (B).

1042  
1043 **Figure S4.** Fold difference (x-axis) and statistical significance (Methods) of the enrichment of  
1044 functional KEGG modules in SGVs present in regions significantly associated with microbial  
1045 growth dynamics.

1046

1047 **Figure S5. SGVs are associated with microbial growth rates.** (A) Boxplot (Box, IQR;  
1048 whiskers, IQR\*1.5) of microbial growth rates calculated using PTR<sup>45</sup> in individuals harboring a 7-  
1049 segment deletion in the *E. eligens* genome (blue) and individuals with no deletion (maroon); (B)  
1050 Genomic map of *E. eligens* with the 7 segments marked in yellow. (C) As in A for a 9-segment  
1051 deletion-SGV in the *E. eligens* genome; (D) As in B with the 9 segments marked in orange.

1052

1053 **Figure S6.** Full heatmap of statistically significant correlations (Spearman  $p < 0.001$ , FDR  
1054 adjusted at 0.1) between disease risk factors and variable-SGVs, depicting associations  
1055 replicated (yellow star), replicated using a different variable (orange star) or reversed (gray star)  
1056 in the Lifelines DEEP cohort.

1057

1058 **Figure S7.** (A) Boxplot (Box, IQR; whiskers, IQR\*1.5) of BMI in individuals harboring a 4-kbp  
1059 deletion in the *A. hadrus* genome (blue) and individuals with no deletion (maroon). (B) Same as  
1060 Fig. 3D for this 4-kbp genomic region of *A. hadrus*. (C) Depiction of the genes encoded in the  
1061 region, which encode key enzymes in the folate biosynthesis pathway. Note correspondence of  
1062 enzyme commission (EC) numbers with panel B. (D) Boxplot (Box, IQR; whiskers, IQR\*1.5) of  
1063 total cholesterol in individuals harboring an 18-kbp deletion in the *R. intestinalis* genome (blue)  
1064 and individuals with no deletion (maroon). (E) same as Fig. 3D for a 10-kbp stretch of the 18-  
1065 kbp region in *R. intestinalis*. (F) Boxplot (Box, IQR; whiskers, IQR\*1.5) of BMI in individuals  
1066 harboring an 8-kbp deletion in the *C. comes* genome (blue) and individuals with no deletion  
1067 (maroon). (G) Same as Fig. 3D for this 8-kbp genomic region of *C. comes*.  $p$  - Two-sided Mann-  
1068 Whitney  $U$  test.

1069

1070 **Figure S8.** Replication of deletion and variable regions depicted in Fig. 3, 4 and S7 between the  
1071 Israeli (yellow) and Dutch Lifelines DEEP (blue) cohorts.

1072

1073 **Figure S9.** (A-C) Boxplot (Box, IQR; whiskers, IQR\*1.5) of waist circumference (A), BMI (B) and  
1074 HDL cholesterol (C) in individuals of the Israeli cohort harboring the 31-kbp deletion in the *A.*  
1075 *hadrus* genome depicted in Fig. 4 (blue) and individuals with no deletion (maroon). (D) Boxplot  
1076 (Box, IQR; whiskers, IQR\*1.5) of BMI in individuals of the Dutch Lifelines DEEP cohort  
1077 harboring the same 31-kbp deletion in the *A. hadrus* genome (blue) and individuals with no  
1078 deletion (maroon).  $p$  - Two-sided Mann-Whitney  $U$  test.