Structural variation in the gut microbiome associates with host health

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

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

Author affiliations

[1]Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot 7610001, Israel

[2]Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot 7610001, Israel

[3]Center for Studies in Physics and Biology, The Rockefeller University, New York, NY 10065, USA

[4]Department of Systems Biology, Columbia University Irving Medical Center, New York, NY 10032, USA

[5]Department of Obstetrics and Gynecology, Columbia University Irving Medical Center, New York, NY 10032, USA

[6]University of Groningen, University Medical Center Groningen, Department of Genetics, 9713 GZ Groningen, The Netherlands

[7]University of Groningen, University Medical Center Groningen, Department of Pediatrics, 9713 GZ Groningen, The Netherlands

[8]Department of Immunology, K.G. Jebsen Coeliac Disease Research Centre, University of Oslo, 0424 Oslo, Norway

[+]These authors contributed equally to this work.

[*]to whom correspondence should be addressed: eran.segal@weizmann.ac.il; dzeevi@rockefeller.edu

Abstract

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

Introduction

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

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

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

74 limited in resolution with regards to within-species variation. Recently, Greenblum et al.[15] 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[26]
77 in the latter case[15]), 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.[15]).

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[14], 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[20,30], demonstrating that SGVs are widely prevalent in the human

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

Results

Accurate metagenomic read assignment using the ICRA algorithm

To accurately detect SGVs in the microbiome we sought to obtain a correct assignment of metagenomic reads to their sequence of origin. Attaining such accurate assignment is challenging due to the large number of genomic sequences that are shared across different microbiome members. Here, we analyzed data collected on 887 healthy subjects which includes microbiome profiling alongside detailed blood glucose measurements over the duration of a week, anthropometric measurements, blood tests, and medical questionnaires[20,30] (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[36] (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[37]. 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[38] and MetaPhyler[39]. Notably, we found that ICRA outperforms the

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

As relative abundances are utilized by ICRA for the resolution of ambiguous read assignments, we further validated that ICRA-derived relative abundances are comparable to those derived from state-of-the-art tools created and optimized for this task. We therefore compared microbial relative abundances produced by ICRA, to those derived from the popular tools MetaPhlAn2[40], which uses marker genes to estimate abundances, and Bracken[41], which performs Bayesian reestimation of abundances derived with Kraken[38]. To this end, and to best simulate the genomic phenomena of bacteria growing naturally (rather than sampled *in silico*), we obtained seven different bacterial strains, grew them to stationary phase, and extracted and sequenced DNA from each strain separately (Methods). We then created 100 samples *in silico* by randomly mixing reads sequenced from each of the seven strains at different relative abundances, and applied MetaPhlAn2, Bracken and ICRA to these samples (Methods). We found that while the Bray-Curtis dissimilarities between the relative abundances estimated by these tools and the true relative abundances were lowest in Bracken (Fig. 1C, inset), followed by ICRA and MetaPhlAn2, the abundances estimated by all three tools were comparable and highly correlated with the true abundances ($R^2 > 0.93$ for each microbe across all samples, $p<10^{-10}$; Fig. 1C, S2).

Sub-genomic variation is highly prevalent in the human microbiome

We next sought to systematically characterize the landscape of sub-genomic variation across the healthy human microbiome. To this end, we developed SGV-Finder, which we ran on ICRA-corrected read assignments of 887 metagenomic samples[20,30] to a reference database of 3953 representative microbial genomes derived from progenomes[36] (Methods). SGV-Finder analyzes 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-

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

SGV is prevalent across distinct populations and continents

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

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

SGVs are person specific and are shared with habitat

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

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

Microbiome SGVs are potentially involved in microbial adaptation and function

We sought to systematically characterize the functional landscape of SGV regions by examining 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[43] 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[44] 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

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

To further characterize the potential contribution of SGVs to microbial niche adaptation, we searched for regions that are associated with fitness of their harboring microbe. As a proxy for fitness, we calculated bacterial growth rates of 21 bacterial strains with sufficient coverage and available complete genomes using a method we previously developed that estimates growth through differences in DNA copy number at the origins and terminus locations created during DNA replication[45]. We found 44 highly significant associations (surpassing Bonferroni correction cutoff of $p<3x10^{-5}$; Fig. 2I; Table S2) of these growth rates with deletion-SGVs within the same bacteria (Methods). These significant associations span a total of 8 distinct bacteria, suggesting that certain SGVs may be important for bacterial adaptation and fitness.

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

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

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

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

Microbiome subgenomic variation is associated with host disease risk factors

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

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

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

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

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

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

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

Disease risk-associated SGVs facilitate an investigation of putative mechanisms

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

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

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

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

Carbohydrate metabolism and SCFA biosynthesis gene clusters encoded in a disease risk-associated region

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

This genomic region encodes two full metabolic modules, seven sugar transporters and two transcriptional regulators, among several unrelated genes (Fig. 4B). Of the two metabolic modules, one performs inositol catabolism[54] metabolizing myo-inositol or D-chiro inositol to (a) glycerone phosphate, a precursor for glyceraldehyde-3-phosphate, a constituent of the Embden–Meyerhof–Parnas glycolysis pathway[26]; and (b) 3-oxopropanoate, a precursor for 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^2 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[55] 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

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

Discussion

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

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

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

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

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

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

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

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

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

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

References

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

24. Rowan, S. *et al.* Involvement of a gut-retina axis in protection against dietary glycemia-induced age-related macular degeneration. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E4472–E4481 (2017).
25. Hsiao, E. Y. *et al.* Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders. *Cell* **155**, 1451–63 (2013).
26. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
27. Li, J. *et al.* An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32**, 834–841 (2014).
28. Manor, O. & Borenstein, E. Systematic Characterization and Analysis of the Taxonomic Drivers of Functional Shifts in the Human Microbiome. *Cell Host Microbe* **21**, 254–267 (2017).
29. Franzosa, E. A. *et al.* Species-level functional profiling of metagenomes and metatranscriptomes. *Nat. Methods* **15**, 962–968 (2018).
30. Korem, T. *et al.* Bread Affects Clinical Parameters and Induces Gut Microbiome-Associated Personal Glycemic Responses. *Cell Metab.* **25**, 1243–1253.e5 (2017).
31. Tigchelaar, E. F. *et al.* Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. *BMJ Open* **5**, e006772 (2015).
32. Zhernakova, A. *et al.* Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science (80-.).* **352**, 565–569 (2016).
33. Kelly, C. J. *et al.* Crosstalk between Microbiota-Derived Short-Chain Fatty Acids and Intestinal Epithelial HIF Augments Tissue Barrier Function. *Cell Host Microbe* **17**, 662–71 (2015).
34. Donohoe, D. R. *et al.* The Microbiome and Butyrate Regulate Energy Metabolism and Autophagy in the Mammalian Colon. *Cell Metab.* **13**, 517–526 (2011).
35. Blacher, E., Levy, M., Tatirovsky, E. & Elinav, E. Microbiome-Modulated Metabolites at the Interface of Host Immunity. *J. Immunol.* **198**, 572–580 (2017).
36. Mende, D. R. *et al.* proGenomes: a resource for consistent functional and taxonomic annotations of prokaryotic genomes. *Nucleic Acids Res.* **45**, D529–D534 (2017).
37. Sczyrba, A. *et al.* Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software. *Nat. Methods* **14**, 1063–1071 (2017).
38. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).
39. Liu, B., Gibbons, T., Ghodsi, M. & Pop, M. MetaPhyler: Taxonomic profiling for metagenomic sequences. in *2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 95–100 (IEEE, 2010). doi:10.1109/BIBM.2010.5706544
40. Truong, D. T. *et al.* MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).
41. Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: estimating species abundance in metagenomics data. *PeerJ Comput. Sci.* **3**, e104 (2017).
42. Rothschild, D. *et al.* Environment dominates over host genetics in shaping human gut microbiota. *Nature* **555**, 210–215 (2018).
43. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2018).
44. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res.* (2018). doi:10.1093/nar/gky995
45. Korem, T. *et al.* Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. *Science* **349**, 1101–6 (2015).
46. Hayashi, F. *et al.* The innate immune response to bacterial flagellin is mediated by Toll-like receptor 5. *Nature* **410**, 1099–1103 (2001).
47. Shen, Y. *et al.* Flagellar Hooks and Hook Protein FlgE Participate in Host Microbe

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

Acknowledgements

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

Author contributions

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

Methods

Reference database preprocessing

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

Metagenomic samples - Israeli cohort

We obtained metagenomic samples from two studies[20,30] (accession numbers ENA: PRJEB11532, ENA: PRJEB17643). In the latter study[30], only baseline samples were used (before the intervention took place).

Gut microbiome analysis

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

For MetaPhlAn2 comparisons, we obtained relative abundances (RA) from metagenomic sequencing via MetaPhlAn2[40] with default parameters. For Kraken[38] comparisons, we built a custom Kraken database using our preprocessed database and subsequently classified with default parameters and generated a Kraken report. For Bracken[41] abundance estimation, we generated a Bracken-database file using bracken-build on the above Kraken database with a kmer length of 31 and read length of 100bp and used it to estimate abundance using the aforementioned Kraken report.

ICRA - Iterative Coverage-based Read Assignment algorithm

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

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

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

$$\lambda_{i,j,k} = 1 \; iff \; read \; i \; originated \; from \; genomic \; element \; j \; in \; position \; p(i,j)_k$$

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

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

Where:

- $\pi_j = f(\{\delta_{i,j,k} \; \forall i,k\})$

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

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

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

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

- $q_{i,j,k} = \prod_{pos=0}^{\rho} qual(pos)^{\mu(i,j,p(i,j)_k+pos)}(1-qual(pos))^{1-\mu(i,j,p(i,j)_k+pos)}$ is the probability of a correct mapping, given the mismatches in the read and the sequencing qualities. Where qual(*pos*) is the probability of correct sequencing in position *pos* calculated from fastq qualities and $\mu(i,j,p(i,j)_k+pos) = 1$if there is a match between nucleotide in position *pos* in read *i* to the one in position $p(i,j)_k$+*pos* in genomic element *j* and 0 otherwise.
- 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 assignments of read *i* equals 1, where *l* and *m* refer to all possible genomic elements and positions thereof to which read *i* is mapped.

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

CAMI dataset comparison

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

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

Bacterial strain culture and sequencing

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

Species	Strain ID	Growth condition – Medium	Growth condition - Temp	Growth to saturation
Lactobacillus gasseri	ATCC 33323	Lactobacillus MRS agar	37^0C	24 hrs
Enterococcus faecalis	ATCC 29212	ATCC Medium 44	37^0C	overnight
Streptococcus cristatus	ATCC 51100	ATCC Medium 44	37^0C	<24 hrs
Akkermansia muciniphila	ATCC BAA-835, DSM 22959	DSM medium 104 + 0.05% mucin or ATCC medium 44	37^0C	72 hrs
Cellulomonas flavigena	ATCC 482, DSM 20109	DSM 53 or ATCC Medium: 3 Nutrient Agar/Broth	30^0C	72 hrs
Brachybacterium faecium	ATCC 43885, DSM 4810	DSM 92 or ATCC Medium: 3 Nutrient Agar/Broth	30^0C	72 hrs
Alistipes finegoldii	DSM 17242	DSM medium 104 + vitamin solution (see medium 131) or 693	37^0C	> 24 hrs

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[64]. 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

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

Detection of variable SGVs

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

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

Detection of conserved regions

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

Analysis of replication in Dutch Lifelines DEEP cohort

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

Calculation of SGV conservation in cohabiting and related individuals

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

Functional enrichment analysis

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

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

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

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

Other HGT mechanisms:

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

Carbohydrate active: glycosyltransferase|glycoside hydrolase|xylan|monooxygenase|rhamnos\S*|cellulose|sialidase|\S*ose($|\s|\-)|acetylglucosaminidase|cellobiose|galact\S*|fructose|aldose|starch|mannose|mannan\S*|glucan|lyase|glycosyltransferase|glycosidase|pectin|SusD|SusC|fructokinase|galacto\S*|arabino\S*

Antibiotic resistance: azole resistance|antibiotic resistance|TetR|tetracycline resistance|VanZ|betalactam\S*|beta-lactam|antimicrob\S*|lantibio\S*

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

Calculation of microbial growth rates

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

SGV explorer

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

Code availability

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

Data availability

The 7 strains samples used in Fig. 1C are available through ENA, accession ENA: PRJEB2519. The 887 samples are publicly available through ENA, accession numbers ENA: PRJEB11532, ENA: PRJEB17643.

Figure Legends

Figure 1. Superior assignment of metagenomic reads using the Iterative Coverage-based Read-Assignment (ICRA) algorithm. (A) Illustration of our computational pipeline. (B) Bar-plots (bar, mean; whiskers, standard deviation) of the ratio of correct read assignment per taxonomy level with no assignment correction (blue) or following assignment correction with ICRA (yellow), Kraken[38] (red) or MetaPhyler[39] (green). * two-sided Mann-Whitney *U* $p<0.05$, ** $p<0.01$ (C) Dot-plot of the calculated relative abundance of 7 bacterial species in 100 samples, using either ICRA (yellow), MetaPhlAn2[40] (blue), or Bracken[41] (red), as compared to the true relative abundances. Inset shows a violin plot (white dot, median; black box, IQR) of Bray-Curtis dissimilarities between the estimates of each method and the true abundances. ** two-sided Wilcoxon signed-rank $p<0.01$ **** $p<10^{-4}$

Figure 2. Sub-Genomic Variation (SGV) is prevalent in the human microbiome, replicable across cohorts and associated with specific functions. (A) Heatmap showing the number of subjects with SGVs (yellow color scale), the number of SGV regions (green color scale), the mean SGV size (blue color scale) and the fraction of the genome that is variable (red color scale), for each microbe analyzed, along with their phylogenetic tree. (B-C) Heatmap (B) and swarm plot (C) showing the genomic length percentage of variable and deletions SGVs replicated in the Lifelines DEEP cohort for each microbe analyzed. (D-E) Boxplot (box, IQR; whiskers, 1.5*IQR) of the distribution of the correlations between variable- (D) or deletion-SGV (E) across different subjects (green), within the same subject (blue), among cohabiting subjects (yellow) and among pairs of siblings or parents/children (red). **- two-sided Mann Whitney *U* $p<0.01$ ***$p<0.001$ ****$p<10^{-5}$. (F-H) Fold change (x-axis) and statistical significance (Methods) of the enrichment of functional KEGG modules in variable-SGVs (F), deletion-SGVs (G) and conserved regions (Methods; H). (I) Difference in median value (x-axis) and statistical

significance in a Mann-Whitney *U* test (y-axis) comparing calculated bacterial growth rates (PTR[45]) under deletion versus retention of SGV.

Figure 3. SGVs are associated with disease risk and these associations replicate across cohorts. (A-B) Heatmap of statistically significant correlations (Spearman p<0.001, FDR adjusted at 0.1) between disease risk factors and variable-(A) or deletion-SGVs (B). Stars singnify associations replicated (yellow), replicated using a different variable (orange) or reversed (gray) in the Lifelines DEEP cohort. Striped stars denote associations from the same bacteria that were collapsed for display purposes (see Figure S6 for full heatmap). (C) Boxplot (Box, IQR; whiskers, IQR*1.5) of glycated hemoglobin (HbA1C%) in individuals harboring an 11-kbp deletion in the *E. rectale* genome (blue) and individuals with no deletion (maroon); *p* - Two-sided Mann-Whitney *U* test. (D) (top) Deletion rate across the cohort (y-axis) along a genomic region of *E. rectale* (x-axis). (bottom) gene locations (arrows) colored according to function (legend). (E) Scatterplot showing the correlation between the abundance of a 6-kbp variable-SGV in *R. torques* and weekly median glucose levels; *p* - Spearman correlation *p*-value. (F) (top) depiction of standardized variability (y-axis; plotted lines, percentiles 1, 25, 50, 75 and 99) along a genomic region of *R. torques* (x-axis). (bottom) gene locations (arrows) colored according to function (legend).

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

(A) Boxplot (Box, IQR; whiskers, IQR*1.5) of body weight in individuals harboring a 31-kbp deletion in the *A. hadrus* genome (blue) and individuals with no deletion (maroon). *p* - Two-sided Mann-Whitney *U* test. (B) Same as Fig. 3D for this genomic region of *A. hadrus.* (C) Depiction of the metabolic pathways encoded in the region, which turns inositol to the short-

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

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

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

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

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

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

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

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

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

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