Structural variation in the gut microbiome associates with host health

Zeevi, David; Korem, Tal; Godneva, Anastasia; Bar, Noam; Kurilshikov, Alexander; Lotan-Pompan, Maya; Weinberger, Adina; Fu, Jingyuan; Wijmenga, Cisca; Zhernakova, Alexandra

Published in:
Nature

DOI:
10.1038/s41586-019-1065-y

Publication date:
2019

Structural variation in the gut microbiome associates with host 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. Here, we systematically identify microbial genomic structural variants (SV) and find them to be prevalent in the human gut microbiome across phyla and to replicate in different cohorts. SVs are enriched for CRISPR-associated and antibiotic producing functions and depleted from housekeeping genes, suggestive of a role in microbial adaptation. We find multiple novel associations between SVs and host disease risk factors, many of which replicate in an independent cohort. Exploring genes clustered in the same SV, we uncover several possible mechanistic links between the microbiome and its host, including a region in *Anaerostipes hadrus* encoding a composite inositol catabolism-butyrate biosynthesis pathway, whose presence is associated with lower host metabolic disease risk. Overall, our results uncover a nascent layer of variability in the microbiome 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 of life[1,2]. Even a small number of bacterial genes can underlie phenotypes such as virulence[3], antibiotic resistance[4], host metabolic disease[5], and host longevity[6], making genetic variation highly important to both the microbe and its host.

Greenblum et al.[7] performed a systematic characterization of intra-species CNVs across the human microbiome, and showed that it is highly prevalent[7]. This variability could be critical to human pathophysiology, as gut microbes were found to be involved in multiple host processes[8–10] and associated with multiple disorders[11–15]. However, this and other studies analyzing the genetic repertoire of the microbiome[12,16–18] were potentially limited by the scope of the annotation databases used and by ignoring the co-variation of genes from the same genomic region. Such co-variation is important as it encodes information such as operon membership, gene regulation or susceptibility to horizontal transfer that is only evident when analyzing genes in their neighboring genomic context. Other functional characterization methods[12,16–18] may be limited with regards to within-species variation of genes.

In this study, we aimed to detect segments of varying lengths, potentially containing multiple genes, that are deleted from certain bacteria in some individuals or present in a variable number of copies in others. We term this phenomenon “structural variation”[19] to differentiate it from CNVs at the level of specific genes without genomic context[7].

We devised an Iterative Coverage-based Read Assignment (ICRA) algorithm that resolves ambiguous read assignments to regions that are similar between different bacteria, using information on bacterial relative abundances in the microbiome, their genomic sequencing-coverage, and sequencing and alignment qualities. We show that our algorithm correctly assigns reads in complex metagenomic settings.

We further developed SGV-Finder, allowing us to systematically detect 7,479 SVs in 56 species from 887 human gut microbiome samples[11,20], demonstrating their prevalence. We show that SVs contain distinct genetic functions, are associated with bacterial growth rates, and are stable within the same person. We demonstrate the potential importance of SVs to the human host by discovering 124 significant associations between SVs and multiple disease risk factors. We analyze the Dutch Lifelines DEEP cohort[21] and show replication of 76% of all SVs in bacteria present in both cohorts, and of 40 associations with risk factors, together suggesting that some variability is shared between distinct populations, while other is population-specific. We highlight several cases in which the gene content of a region can reveal potential underlying mechanisms. Overall, we show that SVs represent a nascent layer of information in the microbiome that is likely to be of high relevance to human health.

Results

Accurate metagenomic read assignment using ICRA

To accurately detect SVs we sought a correct assignment of metagenomic reads to their genome of origin, overcoming the large number of regions shared between different microbes. We analyzed microbiome and clinical data collected on 887 healthy subjects[11,20] (Methods), in which over 15% of the metagenomic reads were assigned ambiguously to multiple references upon mapping to a database of 3,953 bacterial genomes (ED Fig. 1a, Methods).

To address this problem, we devised the ICRA algorithm (ED Fig. 1b, Supplementary Methods), which uses read assignments, read and mapping qualities, sequencing coverage depth along microbial entities (e.g., bacterial genomes or genes), and microbial relative abundances, to reassign ambiguously mapped reads. ICRA introduces a demand for sufficient coverage over entities that are to be considered present in a sample, making it robust to genomic regions with extremely high or low coverage that may arise from misassemblies, homology to other microbes, or phage activation. Such regions could otherwise bias the estimated relative abundances, potentially even assigning abundances to genomic entities that are absent from the sample.

To test the performance of ICRA, we validated the two key components of the algorithm: its ability to resolve ambiguous read assignments, and the accuracy of the species relative abundances that it infers (Supplementary Note 1, ED Fig. 1,2).

SVs are highly prevalent in the microbiome

We next sought to systematically characterize structural variation across the healthy human microbiome. We developed SGV-Finder, which we applied to ICRA-corrected read assignments of 887 metagenomic samples[11,20] to a reference database of 3,953 representative microbial genomes (Methods). SGV-Finder analyzes coverage depth across all microbial genomes in all samples to characterize SVs with respect to the standardized coverage of a genome in a given sample (Methods).

We differentiate between deletion-SVs, that are deleted and not covered in 25-75% of samples, and variable-SVs, that have highly variable coverage across samples (Methods). In both SV types, segments are united based on co-occurrence (deletion-SVs) or correlation (variable-SVs; Methods). An online metagenome explorer for all SVs and the genes they encompass is available at http://genie.weizmann.ac.il/SV/ (ED Fig. 3).

Overall, we detected 2,423 variable-SVs and 5,056 deletion-SVs in 56 bacteria that passed our coverage thresholds (ED Fig. 4a). SVs were detected in 6 bacterial phyla and one archaeal phylum, with 5-241 SVs per species in 1.4-18.6 kbp average size per species. Variable- and deletion-SVs make up 0.3-8.4% and 5.0-26.9% of the microbial genome, respectively (ED Fig. 4a). This apparent disparity in size may suggest inherent differences in the formation of the two types of SVs. Out of 887 samples, 769 carried SVs for *Blautia wexlerae*, 727 had 104 deletion-SVs and 33 variable-SVs in *A. hadrus*, and 668 carried SVs for *Bacteroides uniformis*. We detected SVs in every subject and strain analyzed, demonstrating the ubiquity of such variations.

SV is prevalent across distinct populations

To test the universality of these regions and reinforce their biological relevance, we applied ICRA and SGV-Finder independently to 1020 samples from the Dutch Lifelines DEEP cohort[21] (hereinafter “Lifelines”; Methods). We found that in 47 of 56 bacteria present in both cohorts, an average of 72.9% of variable-SVs (0-99.1%) and 78.3% of deletion-SVs (35.3-94.5%) overlapped with SVs found in our cohort (one-sided hypergeometric $p<10^{-10}$; Fig. 1a, ED Fig. 4b). Notably, for 75% of microbes, more than 70% of the regions were replicated despite the different populations examined with different genetic background, lifestyle, and dietary preferences; and the different methods, centers, and staff involved in assembling the two cohorts (Fig. 1a).

Some bacteria, such as *Ruminococcus bicirculanus*, showed very low concordance between the cohorts (27% overlap over 10 variable-SVs totalling 23 kbp; Fig. 1a, ED Fig. 4b), suggestive of geographically-confined variability, or strong population-specific environmental factors. Other bacteria, such as *Parabacteroides merdae*, showed high concordance (95% of 46 variable-SVs totalling 281 kbp; Fig. 1a, ED Fig. 4b).

SVs are person-specific and shared with habitat

We next examined the variability of SVs across people by correlating variable- and deletion-SVs between different subjects. We found that different individuals mostly have different SV profiles, with a median correlation of 0.02 and 0 for variable- and deletion-SVs, respectively (Spearman; Fig. 1b,c). In contrast, SVs were highly stable within the same individual, even over time periods exceeding one year, with median within-person correlations of 0.89 and 0.66 for variable- and deletion-SVs, respectively (Spearman, $p<10^{-20}$ for both; Fig. 1b,c; Methods).

We further analyzed data from co-habiting individuals and for pairs of parents-children / siblings who do not live together[22] ("relatives"; Methods). We found that they share SVs to a significantly higher degree as compared to two random subjects (average Spearman ρ of 0.45 and 0.16 for variable- and deletion-SVs, respectively; $p<10^{-10}$ for both; Fig. 1b,c). Interestingly, relatives have significantly less similar microbiome SV profiles compared to co-habiting subjects (Mann-Whitney *U* $p<0.001$ for both variable- and deletion-SVs, Fig. 1b,c). This result is conservative, as non-cohabiting relatives could still share environmental exposures affecting their microbiome, such as traditional food preferences or shared meals. These results further support our previous findings[22] that the environment dominates over genetics in determining microbiome composition.

SVs are potentially involved in microbial adaptation

We sought to systematically characterize the function of SVs by searching for enriched or depleted genetic functions. We annotated gene functions across variable-SVs, deletion-SVs, and 'conserved' regions (covered in at least 98% of samples containing the bacteria), and sought KEGG[23] modules that were over- and under-represented in these regions (Methods). Using the KEGG BRITE hierarchy, we found that 'housekeeping' modules such as nucleotide and amino acid metabolism or carbohydrate and lipid metabolism were significantly depleted from SVs and significantly enriched in conserved regions ($p<2*10^{-5}$ for all; Methods; Fig. 2a-c; Table S1). Conversely, modules classified as ABC-2 type- and other transport systems were significantly enriched in variable-SVs ($p<2*10^{-5}$; Fig. 2a). In addition, the type-IV secretion system (T4SS) module was enriched in SVs ($p<2*10^{-5}$; Fig. 2a,b) and depleted from conserved regions ($p<2*10^{-5}$; Fig. 2c), suggesting that bacterial conjugation systems, to which the T4SS is related, strongly associate with variability, further implicating SVs as tools of adaptation and speciation.

SVs were additionally enriched with genes with no assigned function ($p<2*10^{-5}$; Fig. 2a,b, red star). We therefore performed a textual enrichment analysis on the Ensembl functional annotation[24] of the genes analyzed (Methods). Bacteriophage-, plasmid- and transposon-related genes, and genes encoding other horizontal gene transfer (HGT) mechanisms were enriched in SVs and depleted from conserved regions (FDR-corrected $q<10^{-4}$ for all), suggesting an important role for these mechanisms in the formation of SVs. Analysis of Pfam[25] motifs pertaining to HGT mechanisms corroborated this finding ($q<10^{-4}$; Methods). In addition, variable-SVs were enriched with antibiotic-producing genes ($q<0.005$) and deletion-SVs 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, accommodates analyses with multiple annotation datasets.

To further characterize the potential contribution of SVs to microbial adaptation, we searched for SVs associated with the fitness of their harboring microbe. As a proxy for fitness, we calculated bacterial growth rates of 21 strains with sufficient coverage and complete reference genomes, using a method that estimates growth from DNA copy number differences created during DNA replication[26]. We found 44 highly significant associations of these growth rates with deletion-SVs within the same bacteria (Methods; Mann-Whitney *U* test, significant with Bonferroni cutoff of $p<3x10^{-5}$; Fig. 2d; Table S2), suggesting that certain SVs may be important for bacterial adaptation and fitness.

To probe the mechanisms potentially underlying this adaptation, we systematically examined the genetic content of growth-associated deletion-SVs, and found similar functional profiles as in all SVs, with a depletion of housekeeping functions and enrichment for genes involved with CRISPR-, transposon- and HGT-associated genes ($q<0.05$; Methods), and a significant enrichment for genes with unknown functions ($p<10^{-5}$, ED Fig. 5).

We further examined two such regions, which were significantly positively and negatively associated ($p<10^{-10}$ for both) with the growth of the harboring species (*Eubacterium eligens*; ED Fig. 6). Notably, the negatively-associated SV (ED Fig. 6a,b) contains, among others, genes for flagellin, flagellar hook-associated protein and lipopolysaccharide (LPS) choline phosphotransferase (Table S3). Flagellin and the flagellar hook protein were shown to elicit strong immune responses[27,28]; LPS choline phosphotransferase attaches choline phosphate to the bacterial LPS molecule, which was shown to increase C-reactive protein-mediated innate immune clearing[29]. Both of these could potentially inhibit microbial growth, and increase growth rates in bacteria missing them may point to loss-of-function adaptation to the host gut and immune system. In contrast, the SV that was positively associated with the microbe’s growth rates (ED Fig. 6c,d) contained several hypothetical genes, and also a gene for antibiotic transport system ATP-binding protein, whose presence could endow a selective advantage in certain human hosts by conferring resistance to antibiotics[30] (Table S3). These results demonstrate the ability of our methodology to suggest underlying mechanisms using the genomic content of SVs.

Overall, SVs associate with common mechanisms of conjugation, transposition and phage lysogeny, and may thus be powerful tools of adaptation. Microbial evolution in densely populated ecosystems such as the human microbiome may thus be driven strongly by SVs, affecting both microbes and host.

SVs associate with risk factors across cohorts

To explore the potential relevance of microbiome SVs to human health, we associated the abundance of variable-SVs and the presence of deletion-SVs with metrics of health and risk factors: mean arterial blood pressure (MAP); total and HDL cholesterol; waist circumference; weight; body mass index (BMI); median glucose levels over one week; percent glycated hemoglobin (HbA1c%); and age. We found 81 (Spearman; Fig. 3a, ED Fig. 7) and 43 (Mann-Whitney *U*; Fig. 3b) significant associations FDR corrected at 0.1 for variable- and deletion-SVs, respectively, demonstrating the potential importance of microbial SVs to the human host.

In several cases, the relative abundances of a microbe harboring risk-factor-associated SVs were correlated with the same risk factors. For example, we found five deletion-SVs in *A. hadrus* to be associated with lower BMI, body weight and waist circumference, and with higher HDL cholesterol levels (Fig. 3b), and indeed *A. hadrus* was negatively correlated with 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 ($p<10^{-7}$). Even so, the associations of specific SVs with risk factors allows us to pinpoint specific regions and mechanism that may underlie the association.

Notwithstanding, the relative abundances of some bacteria have opposite associations with host phenotypes compared to the SVs they contain. For example, three variable-SVs in *Ruminococcus torques* were negatively associated with multiple risk factors (Fig. 3a) but *R. torques* abundance was positively associated with weight ($p<10^{-3}$) and BMI ($p<0.05$), similar to results from a different cohort[31]. Several variable-SVs 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-SV in *Faecalibacterium cf. prausnitzii* KLE1255 was positively associated with the weekly median blood glucose level (Fig. 3B), and though *F. prausnitzii* was not significantly associated with it in our cohort, two studies found it was negatively associated with type II diabetes, characterized by high blood glucose levels[12,32]. These seemingly paradoxical associations between SVs and risk factors further suggest that

SVs represent a different layer of information compared to the taxonomic level, one which may assist in obtaining mechanistic insights into host-microbe interactions.

To test the replicability of these associations, we ran ICRA on samples from the Lifelines cohort, and calculated the coverage of the SVs defined from the 887-person cohort. We then calculated the association of these regions with host risk factors measured in the Lifelines cohort, and compared those to the associations found in our cohort (Methods). Notably, despite presumed inter-cohort differences in genetics, dietary preferences and lifestyles, more than a third (40 out of 117) of the associations found in microbes present in both cohorts were replicated, while only 4 of the remaining 77 were significantly associated in the opposite direction (Fig. 3; ED Fig. 7).

SV-risk associations facilitate mechanistic insights

As with bacterial adaptation, examining the genetic content of disease-risk-associated SVs facilitated a potentially mechanistic view into these phenomena. While many SVs harbor genes with unknown function, we observed several intriguing functions coded in risk-associated SVs. For example, the presence of a 11-kbp deletion-SV from *E. rectale* is associated with higher HbA1c% ($p<10^{-4}$; n=630, 377 retaining; ED Fig. 8a). An examination of this region reveals a class 1 CRISPR-Cas system and three genes of unknown function (ED Fig. 8b). Interestingly, subjects 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 inducing high postprandial glucose responses[11]. A 6-kbp variable-SV from *R. torques* is inversely associated with median blood glucose levels (Spearman ρ=-0.237, $p<10^{-5}$; Fig. 4a) and features genes encoding phage-associated proteins and additional genes of unknown function, suggesting that this SV is a prophage, and that it may carry additional functionality (Fig. 4b). These genes of unknown function are therefore putatively related to host glucose metabolism, demonstrating the utility of our methods for generating mechanistic hypotheses.

Other examples include a 4-kb deletion-SV in *A. hadrus* that is inversely associated with BMI (median 1.15kg/m^2 lower for retention; $p<10^{-4}$; n=681, 405 retaining; ED Fig. 8c) and weight (median 3.5kg lower; $p<10^{-4}$). This SV contains genes coding for the enzymes ADC synthase (EC2.6.1.85) and 4-amino-4-deoxychorismate lyase (EC4.1.3.38), both instrumental in folate biosynthesis (ED Fig. 8d,e). An 18-kb deletion-SV in *Roseburia intestinalis* that is significantly associated with total cholesterol (median 12.5mmHg lower for retention; $p<10^{-4}$; n=262, 68 retaining; ED Fig. 8f) contained multiple beta- and other glucosidases (ED Fig. 8g), potentially suggesting microbial adaptation to a fiber-rich host diet. An 8-kb deletion-SV in *Coprococcus comes* which is significantly associated with BMI (median 2.4kg/m^2 higher for retention; n=450; 292 retaining; $p<10^{-5}$; ED Fig. 8h) and weight (median 5kg higher; $p<10^{-4}$) contains several ABC transporters of possible future interest with undetermined substrates (ED Fig. 8i). Notably, all the above regions were also detected in the Lifelines cohort (ED Fig. 9) and replicate the patterns detected in our cohort.

As one particularly intriguing example, a 31-kbp deletion-SV in *A. hadrus* was significantly associated with lower weight (median 6kg lower for retention; $p<10^{-6}$; n=681, 468 retaining; Fig. 4c), waist circumference (median 4cm lower; $p<10^{-4}$; ED Fig. 10a), BMI (median 1.17kg/m^2 lower; $p<0.001$; ED Fig. 10b), and higher HDL cholesterol (median 5.7mg/dL higher; $p<10^{-4}$; ED Fig. 10c), and was well annotated, allowing us to demonstrate the potential of SGV-Finder detected regions to expose potential underlying mechanisms.

This SV encodes a metabolic module for inositol catabolism[33] metabolizing myo-inositol or D-chiro inositol to (a) glycerone phosphate, a precursor for glyceraldehyde-3-phosphate; and (b) 3-oxopropanoate, a precursor for acetyl-CoA. The region also encodes a metabolic module that metabolizes 3-hydroxybutanoyl-CoA to butyrate, a short-chain fatty acid (SCFA), while oxidizing an electron-transferring flavoprotein, also encoded in this SV. These two pathways are connected through reactions encoded elsewhere in the *A. hadrus* genome (Fig. 4d,e, Table S4).

This SV additionally encodes seven sugar transporters, one of which specific to sorbitol and six are unassigned to a target; two transcriptional regulators; and several other genes.

Altogether, we hypothesize that this SV is unifunctional, providing the capability to ferment sugar alcohols to SCFAs in an energetically-favorable procedure. The myo-inositol catabolism module combined with glycolysis and acetyl-CoA synthesis has a positive energetic effect, and the butyrate synthesis module consumes energy for butyrate production, with a positive effect on the energy metabolism of *A. hadrus*, earning a net gain of 2 ATP- and 2 NADH-equivalent molecules.

This SV is replicated in the Lifelines cohort (ED Fig. 9), and so are several of its associations with host phenotypes: Dutch individuals retaining the region exhibiting lower BMI (median 0.9kg/m^2 lower for retention; n=797, 547 retaining; p<0.005; ED Fig. 10d), weight (median 4kg. lower; p<0.01), and waist-to-hip ratio (median 0.017 lower; p<0.001) potentially pointing to a generalized mechanistic association between SV and disease-risk. This region is also associated with a significantly different predicted metabolic profile at the entire microbiome level (Supplementary Note 2).

Discussion

In this work we systematically detect SVs across metagenomic samples, and show that they are highly abundant in the human microbiome and largely conserved across different cohorts. We found that SVs harbor genes of distinct functions, and are associated with bacterial growth rates, indicating a potential utility in bacterial adaptation. Finally, we found they are associated with numerous host disease risk-factors, many of which replicated in an independent cohort, and that they facilitate exploration of genes varying together, exposing a new layer of putative mechanistic information regarding host-microbiome interactions.

Following a functional analysis of genes in those regions, we hypothesize that the main forces driving SVs are mechanisms of HGT as evident from the enrichment of genes performing these functions in SVs. Many genes found in SVs, 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 environments and in communication with the host.

Our current methodology depends on a reference dataset, typically sufficient for human microbiome analyses. We note that this is a practical rather than a conceptual limitation, as any type of database would suffice, even that created through ad-hoc assembly. Creation and validation of such methodologies could be pursued in future work.

Detecting SVs directly from sequencing coverage facilitates an independent analysis of their encoded functions. We demonstrated the utility of such examination with several SVs whose genes were well annotated. This includes an *A. hadrus* SV containing genes hypothesized to enable the transport and metabolism of sugar alcohols to butyrate, and which was strongly associated with lower metabolic risk. SCFAs, and specifically butyrate, have been shown to nourish host intestinal cells[34] and mitigate inflammatory disease[35]. In mice, SCFAs were shown to improve insulin sensitivity and increase energy expenditure[36]. We therefore hypothesize that by possessing this SV, 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 butyrate.

The associations described here between SVs and host health are not directional or causal, and could also be confounded. While further research is needed to fully understand the interactions between the host, its microbiome, and disease, we demonstrate the wealth of mechanistic hypotheses obtained through examining genes with variable copy number 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.

Our methodology is highly adaptable to any metagenomic scenario and could be used, for example, to detect SVs in the soil microbiome. It is especially useful for raising mechanistic hypotheses, and could therefore be useful in case-control microbiome studies. Taken together, our study exposes a new facet of the microbiome that brings us closer to mechanistically understanding host-microbe interactions.

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* **329**, 533–8 (2010).
3. 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).
4. 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).
5. Koeth, R. a *et al.* Intestinal microbiota metabolism of L-carnitine, a nutrient in red meat, promotes atherosclerosis. *Nat. Med.* **19**, 576–85 (2013).
6. Han, B. *et al.* Microbial Genetic Composition Tunes Host Longevity. *Cell* **169**, 1249–1262.e13 (2017).
7. Greenblum, S., Carr, R. & Borenstein, E. Extensive strain-level copy-number variation across human gut microbiome species. *Cell* **160**, 583–94 (2015).
8. 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).
9. LeBlanc, J. G. *et al.* Bacteria as vitamin suppliers to their host: a gut microbiota perspective. *Curr. Opin. Biotechnol.* **24**, 160–8 (2013).
10. Levy, M. *et al.* Microbiota-Modulated Metabolites Shape the Intestinal Microenvironment by Regulating NLRP6 Inflammasome Signaling. *Cell* **163**, 1428–1443 (2015).
11. Zeevi, D. *et al.* Personalized Nutrition by Prediction of Glycemic Responses. *Cell* **163**, 1079–94 (2015).
12. Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**, 55–60 (2012).
13. Halfvarson, J. *et al.* Dynamics of the human gut microbiome in inflammatory bowel disease. *Nat. Microbiol.* **2**, 17004 (2017).
14. Pascal, V. *et al.* A microbial signature for Crohn's disease. *Gut* **66**, 813–822 (2017).
15. 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).
16. Li, J. *et al.* An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32**, 834–841 (2014).
17. 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).
18. Franzosa, E. A. *et al.* Species-level functional profiling of metagenomes and metatranscriptomes. *Nat. Methods* **15**, 962–968 (2018).
19. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* **12**, 363–376 (2011).
20. Korem, T. *et al.* Bread Affects Clinical Parameters and Induces Gut Microbiome-Associated Personal Glycemic Responses. *Cell Metab.* **25**, 1243–1253.e5 (2017).
21. Zhernakova, A. *et al.* Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* **352**, 565–569 (2016).
22. Rothschild, D. *et al.* Environment dominates over host genetics in shaping human gut microbiota. *Nature* **555**, 210–215 (2018).
23. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic*

Acids Res. **28**, 27–30 (2000).
24. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2018).
25. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res.* (2018). doi:10.1093/nar/gky995
26. Korem, T. *et al.* Growth dynamics of gut microbiota in health and disease inferred from single metagenomic samples. *Science* **349**, 1101–6 (2015).
27. Hayashi, F. *et al.* The innate immune response to bacterial flagellin is mediated by Toll-like receptor 5. *Nature* **410**, 1099–1103 (2001).
28. Shen, Y. *et al.* Flagellar Hooks and Hook Protein FlgE Participate in Host Microbe Interactions at Immunological Level. *Sci. Rep.* **7**, 1433 (2017).
29. 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).
30. 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).
31. 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).
32. Karlsson, F. H. *et al.* Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* **498**, 99–103 (2013).
33. Yoshida, K. *et al.* myo-Inositol catabolism in Bacillus subtilis. *J. Biol. Chem.* **283**, 10415–24 (2008).
34. Bergman, E. N. Energy contributions of volatile fatty acids from the gastrointestinal tract in various species. *Physiol. Rev.* **70**, 567–90 (1990).
35. 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).
36. Gao, Z. *et al.* Butyrate improves insulin sensitivity and increases energy expenditure in mice. *Diabetes* **58**, 1509–17 (2009).

Acknowledgements

We thank members of the Segal group and the Center for Studies in Physics and Biology for discussions. E.S. is supported by grants from the European Research Council and the Israel Science Foundation. D.Z. is supported by the James S. McDonnell Foundation and the Dan David Prize Scholarship. D.Z. and T.K were partly supported by the Israeli Ministry of Science and Technology. Lifelines DEEP was funded by: ERC-2012-322698 and NWO-SPI-92-266 to C.W.; ERC-715772 and NWO-178.056 to A.Z; NWO-864.13.013 and CVON-2012-03 to JF.

Author contributions

T.K. and D.Z. conceived and 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. analyzed the Dutch cohort. M.L.-P and A.W. did experimental work. A.W. designed the study. E.S. conceived, directed and designed the project and analyses, interpreted the results and wrote the manuscript.

Author information

Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing interests. Correspondence and requests for materials should be addressed to eran.segal@weizmann.ac.il or dzeevi@rockefeller.edu.

Figure Legends

Figure 1. SVs replicate across cohorts and are stable within individuals over time. (a) The genomic length overlap of SVs replicated in the Lifelines cohort for all microbes analyzed. (b-c) Boxplots (centre, median; box, IQR; whiskers, 1.5*IQR) of the correlations between variable- (b) or deletion-SV (c) across different subjects (green, n=704), within the same subject (blue, n=21), among co-habiting subjects (yellow, n=39) and among pairs of siblings or parents/children ('1st deg. relatives', red, n=38). **- two-sided Mann Whitney *U p<0.01* ****p<0.001* *****p<10*$^{-5}$.

Figure 2. SVs associate with microbial growth rates and specific functions. (a-c) Statistical significance (Methods) vs. fold change of KEGG[23] modules in variable-SVs (a), deletion-SVs (b) and conserved regions (Methods; c). (d) Statistical significance in a two-sided Mann-Whitney *U* test vs. difference in median value, comparing bacterial growth rates (PTR[26]) under deletion

versus retention of 1,756 SVs in 21 microbes. See Tables S1 (panels a-c), S2 (d) for group sizes.

Figure 3. SVs associate with disease risk, replicated in another cohort. (a-b) Heatmap of statistically significant correlations (Methods) between disease risk factors and variable- (a) or deletion-SVs (b). Stars, associations replicated (yellow), replicated using a different variable (orange) or reversed (gray) in the Lifelines cohort[21]. Striped stars, rows from the same bacteria that were collapsed for display purposes (see ED Fig. 7 for full heatmap).

Figure 4. Risk-associated SVs harbor functionally diverse genes (a) Scatterplot showing the Spearman correlation between the abundance of a 6-kbp variable-SV in *R. torques* and weekly median glucose levels (n=373); *p* - Methods. (b) (top) 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). (c) Boxplot (centre, median; box, IQR; whiskers, IQR*1.5) of weight in individuals harboring a 31-kbp deletion in the *A. hadrus* genome (blue, n=213) and individuals with no deletion (maroon, n=468). *p* - two-sided Mann-Whitney *U* test. (d) (top) Deletion rate across the cohort (y-axis) along a genomic region of *A. hadrus* (x-axis). (bottom) gene locations (arrows) colored according to function (legend). (e) The metabolic pathways encoded in this SV, which turns inositol to butyrate. Note correspondence of enzyme commission (EC) numbers with panel d.

Methods

Reference database preprocessing

We downloaded the EMBL progenomes[37] 5,306 representatives dataset and used dRep[38] 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 3,953 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 S5.

Metagenomic samples - Israeli cohort

We obtained metagenomic samples from two studies[11,20] (accession numbers ENA: PRJEB11532, ENA: PRJEB17643). In the latter study[20], 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[39] 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[41] comparisons, we built a custom Kraken database using our preprocessed database and subsequently classified with default parameters and generated a Kraken report. For Bracken[42] 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.

SV detection - preprocessing

We mapped metagenomic reads to the reference database of 3,953 representative microbial genomes detailed above and corrected read assignments using ICRA. All scaffolds from each microbial genome were concatenated and subsequently divided into 1 kbp bins. For each genome in each microbial sample, we counted the number of reads mapped to each of the bins. In the rare case in which ICRA produces a distribution of probabilities of different read assignment for a specific read rather than a deterministic assignment, we determined the read count that was added to each bin using the probability of assignment calculated by ICRA. To ensure proper statistical support for coverage analyses, we discard genomes in samples whose median bin coverage is lower than 10 reads (corresponding to a genome coverage of 1x, with ten 100bp reads in each 1kbp bin), and microbial genomes present in less than 75 subjects. In addition, we removed microbes in which the median bin coverage across samples was lower than one read for more than 30% of the bins.

Detection of deletion SVs

We examined the coverage in each metagenomic bin across all samples to detect regions that were deleted from some individuals and retained in others. To this end, for each microbe in each sample, we calculated a histogram of coverage across all metagenomic bins. We then 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 deletion-SV, we demanded that it be deleted in 25-75% of samples. We concatenated adjacent deletion-SV 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 SV 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 SVs

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 SV. We concatenated adjacent variable SVs 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 SV stretches similarly to deletion SV 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 a threshold of 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 SVs 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 SV region in the Israeli cohort, its presence / absence (deletion SV) or standardized coverage (variable SV) 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-SVs) or the Spearman correlation (variable-SVs).

Calculation of SV conservation in co-habiting and related individuals

We calculated Spearman correlations between the deletion- and variable-SV vectors of 39 pairs of individuals registered in our cohort as living in the same house. To calculate SV retention in first degree relatives, we calculated these correlations in 38 pairs of individuals whose genomic SNP-based similarity[22] was between 40 and 60% and whose self-reported residential addresses were different.

Functional enrichment analysis

This analysis was performed similarly yet separately to variable-SVs, deletion-SVs, 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[24] available through progenomes[37], and annotated orphan ORFs by mapping the protein sequence to all KEGG[23] protein sequences using DIAMOND[43] 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[25] modules with the keywords ‘phage’, ‘prophage’, ‘transposon’, ‘conjugative transposon’ using hmmscan (HMMER v3.1[44]) 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 100,000 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. 26. PTRs were calculated for all the strains that were found to contain at least one deletion-SV and that whose reference genome sequence was complete (i.e., not fragmented to contigs, as required by the PTR method[26]), 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-SV 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.

Association of SVs with disease-risk factors

Variable-SVs were Spearman correlated to disease risk factors (MAP, waist circumference, weight, BMI, week-long median glucose levels, glycated hemoglobin, age, total and HDL cholesterol) and p-values were calculated, ensuring a minimum of 20 subjects in each comparison. Two-sided Mann-Whitney *U* test was used to calculate significance of associations between deletion-SVs and the same disease risk factors, demanding at least 5 subjects in each comparison and at least 5 unique values in each group. FDR correction was performed on variable- and deletion-SV associations separately.

Statistical analyses

Unless otherwise mentioned, all relevant statistical tests used were two-sided. p-values for Spearman correlations where calculated using the t-distribution, through the implementation in the python scipy.stats.spearmanr module (http://www.scipy.org/). FDR was performed using the Benjamini and Hochberg method[45] implemented in the python mne package.

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 the 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.

Bacterial strain culture and sequencing

Seven strains were obtained and grown to stationary phase as listed in Table S6. DNA was extracted using QIAgen DNAeasy Blood & Tissue kit (Cat# 69504) by the protocol using pretreatment of Gram-positive or Negative bacteria following purification of total DNA from animal tissues.

Following that, 100 ng of DNA was sonicated using Covaris E220X and and Illumina library was prepared for each strain as previously described[46]. The seven strains were sequenced to a minimum depth of 3M reads by a NextSeq® 500 machine with Illumina NS 500/550 High Output V2 75 cycle kit.

CAMI dataset comparison

We downloaded all 180bp-spaced toy datasets for the 1st CAMI challenge[47] 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[39] 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[48] 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[41] 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.

SV explorer

SV explorer, presented in ED Fig. 3 and accessible through https://genie.weizmann.ac.il/SV/, was created using bokeh for Python (http://bokeh.pydata.org)

Code availability

ICRA, SGV-Finder, and the SV 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: PRJEB25194. The 887 samples are publicly available through ENA, accession numbers ENA: PRJEB11532, ENA: PRJEB17643. The raw metagenomic sequencing data for the LifeLines DEEP cohort, and age and sex information per sample are available from the European genome-phenome archive (https://www.ebi.ac.uk/ega/) at accession number EGAS00001001704. Other phenotypic data can be requested from the LifeLines cohort study (https://lifelines.nl/lifelines-research/access-to-lifelines) following the standard protocol for data access.

Methods references:

37. Mende, D. R. *et al.* proGenomes: a resource for consistent functional and taxonomic annotations of prokaryotic genomes. *Nucleic Acids Res.* **45**, D529–D534 (2017).

38. 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).

39. Marco-Sola, S., Sammeth, M., Guigó, R. & Ribeca, P. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat. Methods* **9**, 1185–1188 (2012).

40. Truong, D. T. *et al.* MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).

41. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).

42. Lu, J., Breitwieser, F. P., Thielen, P. & Salzberg, S. L. Bracken: estimating species abundance in metagenomics data. *PeerJ Comput. Sci.* **3**, e104 (2017).

43. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).

44. Potter, S. C. *et al.* HMMER web server: 2018 update. *Nucleic Acids Res.* **46**, W200–W204 (2018).

45. 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).

46. Suez, J. *et al.* Artificial sweeteners induce glucose intolerance by altering the gut microbiota. *Nature* **514**, 181–6 (2014).

47. Sczyrba, A. *et al.* Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software. *Nat. Methods* **14**, 1063–1071 (2017).

48. 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

Extended Data Figure 1. Superior assignment of metagenomic reads using the Iterative Coverage-based Read-Assignment (ICRA) algorithm. (a) Boxplot (centre, median; box, IQR; whiskers, 10th and 90th percentiles) of ambiguous read assignment ratios of 887 samples[11,20] mapped to a reference database of 3,953 representative microbial genomes (Methods) before (blue) and after (yellow) ICRA correction. (b) Illustration of our computational pipeline. (c-e) Swarm-plots of the ratio of correct read assignment per taxonomy level with no assignment correction (blue) or following assignment correction with ICRA (yellow), Kraken[41] (red) or MetaPhyler[48] (green) for CAMI[47] high complexity (c; n=5), medium complexity (d; n=2) and low complexity (e; n=1) datasets. Note that MetaPhyler did not provide sub-species level read assignments. * two-sided Mann-Whitney *U* $p<0.05$, **$p<0.01$.

Extended Data Figure 2. ICRA estimates relative abundances with accuracy comparable to other tools. (a) Dot-plot of the calculated relative abundance of 7 bacterial species in 100 samples, using either ICRA (yellow), MetaPhlAn2[40] (blue), or Bracken[42] (red), as compared to the true relative abundances. Inset shows a violin plot (white dot, median; black rectangle, IQR, whiskers, 1.5*IQR) of Bray-Curtis dissimilarities between the estimates (n=100) of each method and the true abundances. ** two-sided Wilcoxon signed-rank $p=1.3\text{x}10^{-4}$ **** $p=3.0\text{x}10^{-18}$. (b-h) Dot-plot of the calculated relative abundance (y-axis) of *A. finegoldii* (b), *B. faecium* (c), *C. flavigena* (d), *E. faecalis* (e), *L. gasseri* (f), *S. cristatus* (g) and *A. muciniphila* (h) 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.

Extended Data Figure 3. SV Explorer enables investigation of co-varying genes. (a-b) Illustration of the online SV explorer available at http://genie.weizmann.ac.il/SV/, spanning the entire *R. torques* genome (a) and spanning a 26-kbp region of this genome (b).

Extended Data Figure 4. SVs are prevalent in the human microbiome across two cohorts. (a) Heatmap showing the number of subjects with SVs (yellow color scale), the number of SVs (green color scale), the mean SV 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) Heatmap showing the genomic length percentage of variable- and deletion-SVs replicated in the Lifelines cohort for each microbe analyzed.

Extended Data Figure 5. Growth rates-associated SVs harbor specific functions. Fold difference (x-axis) and statistical significance (Methods; y-axis) of the enrichment of functional KEGG modules in SVs present in regions significantly associated with microbial growth dynamics. A total of 56,088 genes were considered, 3,805 of them in growth rates-associated SVs.

Extended Data Figure 6. SVs are associated with microbial growth rates. (a) Boxplot (centre, median; box, IQR; whiskers, IQR*1.5) of microbial growth rates calculated using PTR[26] in individuals harboring a 7-segment deletion in the *E. eligens* genome (blue, n=281) and individuals with no deletion (maroon, n=166); (b) Genomic map of *E. eligens* with the 7 segments marked in yellow. (c) As in (a) for a 9-segment deletion-SV in the *E. eligens* genome (blue, n=57) and individuals with no deletion (maroon, n=390); (d) As in (b) with the 9 segments marked in orange. *p* - two-sided Mann-Whitney *U* test.

Extended Data Figure 7. SVs are associated with disease risk, replicated in a second cohort. Full heatmap of statistically significant correlations (Methods) between disease risk factors and variable-SVs, depicting associations replicated (yellow star), replicated using a different variable (orange star) or reversed (gray star) in the Lifelines cohort.

Extended Data Figure 8. Gene content of SVs associated with host risk factors. (a) Boxplot (centre, median; Box, IQR; whiskers, IQR*1.5) of glycated hemoglobin in individuals harboring an 11-kbp deletion in the *E. rectale* genome (blue, n=253) and individuals with no deletion (maroon, n=377); *p* - two-sided Mann-Whitney *U* test. (b) Same as Fig. 4d for this 11-kbp genomic region of *E. rectale*. (c) Boxplot of BMI in individuals harboring a 4-kbp deletion in the *A. hadrus* genome (blue, n=276) and individuals with no deletion (maroon, n=403). (d) Same as Fig. 4d for this 4-kbp genomic region of *A. hadrus.* (e) 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 d. (f) Boxplot of total cholesterol in individuals harboring an 18-kbp deletion in the *R. intestinalis* genome (blue, n=194) and individuals with no deletion (maroon, n=68). (g) same as Fig. 4d for a 10-kbp stretch of the 18-kbp region in *R. intestinalis*. (h) Boxplot of BMI in individuals harboring an 8-kbp deletion in the *C. comes* genome (blue, n=158) and individuals with no deletion (maroon, n=294). (i) Same as Fig. 4d for this 8-kbp genomic region of *C. comes. p* - two-sided Mann-Whitney *U* test. Boxplots - centre, median; box, IQR; whiskers, IQR*1.5.

Extended Data Figure 9. Detailed examples of SV replication. Replication of deletion and variable regions depicted in Fig. 4 and ED Fig. 8 between the Israeli (yellow) and Dutch Lifelines DEEP (blue) cohorts.

Extended Data Figure 10. A SV of *A. hadrus* associated with host risk factors. (a-c) Boxplot 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, n=213) and individuals with no deletion (maroon, n=468). (d) Boxplot of BMI in individuals of the Dutch Lifelines DEEP cohort harboring the same 31-kbp deletion in the *A. hadrus* genome (blue, n=249) and individuals with no deletion (maroon, n=547). *p* - two-sided Mann-Whitney *U* test. Boxplots - centre, median; box, IQR; whiskers, IQR*1.5.