LETTER

OPEN

https://doi.org/10.1038/s41586-018-0620-2

# The human gut microbiome in early-onset type 1 diabetes from the TEDDY study

Tommi Vatanen[1]*, Eric A. Franzosa[1,2], Randall Schwager[2], Surya Tripathi[1], Timothy D. Arthur[1], Kendra Vehik[3], Åke Lernmark[4], William A. Hagopian[5], Marian J. Rewers[6], Jin-Xiong She[7], Jorma Toppari[8,9], Anette-G. Ziegler[10,11,12], Beena Akolkar[13], Jeffrey P. Krischer[3], Christopher J. Stewart[14,15], Nadim J. Ajami[14], Joseph F. Petrosino[14], Dirk Gevers[1,19], Harri Lähdesmäki[16], Hera Vlamakis[1], Curtis Huttenhower[1,2,20]* & Ramnik J. Xavier[1,17,18,20]*

**Type 1 diabetes (T1D) is an autoimmune disease that targets pancreatic islet beta cells and incorporates genetic and environmental factors[1], including complex genetic elements[2], patient exposures[3] and the gut microbiome[4]. Viral infections[5] and broader gut dysbioses[6] have been identified as potential causes or contributing factors; however, human studies have not yet identified microbial compositional or functional triggers that are predictive of islet autoimmunity or T1D. Here we analyse 10,913 metagenomes in stool samples from 783 mostly white, non-Hispanic children. The samples were collected monthly from three months of age until the clinical end point (islet autoimmunity or T1D) in the The Environmental Determinants of Diabetes in the Young (TEDDY) study, to characterize the natural history of the early gut microbiome in connection to islet autoimmunity, T1D diagnosis, and other common early life events such as antibiotic treatments and probiotics. The microbiomes of control children contained more genes that were related to fermentation and the biosynthesis of short-chain fatty acids, but these were not consistently associated with particular taxa across geographically diverse clinical centres, suggesting that microbial factors associated with T1D are taxonomically diffuse but functionally more coherent. When we investigated the broader establishment and development of the infant microbiome, both taxonomic and functional profiles were dynamic and highly individualized, and dominated in the first year of life by one of three largely exclusive *Bifidobacterium* species (*B. bifidum*, *B. breve* or *B. longum*) or by the phylum Proteobacteria. In particular, the strain-specific carriage of genes for the utilization of human milk oligosaccharide within a subset of *B. longum* was present specifically in breast-fed infants. These analyses of TEDDY gut metagenomes provide, to our knowledge, the largest and most detailed longitudinal functional profile of the developing gut microbiome in relation to islet autoimmunity, T1D and other early childhood events. Together with existing evidence from human cohorts[7,8] and a T1D mouse model[9], these data support the protective effects of short-chain fatty acids in early-onset human T1D.**

Recent literature has linked several facets of gut health with the onset of T1D in humans and rodent models[4,6,10]. Altered intestinal microbiota in connection to T1D has been reported in Finnish[7,8,11,12], German[13], Italian[14], Mexican[15], American (Colorado)[16] and Turkish[17] children. Common findings include increased numbers of *Bacteroides* species, and deficiency of bacteria that produce short-chain fatty acids (SCFAs)[7,8] in cases of T1D or islet autoimmunity (IA)[8,11,15,18]. Corroborating these findings, decreased levels of SCFA-producing bacteria were found in adults with type 2 diabetes (T2D)[19]. In addition, increased intestinal permeability[14] and decreased microbial diversity[12] after IA but before T1D diagnosis have been reported. Studies using the nonobese diabetic (NOD) mouse model have determined immune mechanisms that mediate the protective effects of SCFAs[9] and the microbiome-linked sex bias in autoimmunity[20]. NOD mice fed specialized diets resulting in high bacterial release of the SCFAs acetate and butyrate were almost completely protected from T1D[9]. A study in a streptozotocin-induced T1D mouse model demonstrated that bacterial products recognized in pancreatic lymph nodes contribute to pathogenesis[21].

Even in the absence of immune perturbation, the first few weeks, months and years of life represent a unique human microbial environment that has only recently been detailed[22,23]. Infants have a markedly different gut microbial profile from adults, characterized by a distinct taxonomic profile, greater proportion of aerobic energy harvest metabolism, and more extreme dynamic change[24]. These differences gradually fade over the first few years of life, particularly in response to the introduction of solid food, and individual microbial developmental trajectories are influenced by environment, delivery mode, breast (versus formula) feeding, and antibiotics[25–27]. Most studies that address the development of the gut microbiome, both generally and in association with T1D, have used gene analysis of 16S rRNA, which leaves open the question of functional and strain-specific differences that are not easily detected by this technology that might contribute to disease pathogenesis[12].

Bridging this gap is one goal of the The Environmental Determinants of Diabetes in the Young (TEDDY) study, a prospective study that aims to identify environmental causes of T1D[28]. It includes six clinical research centres in the United States (Colorado, Georgia/Florida and Washington) and Europe (Finland, Germany and Sweden), which together have recruited several thousand newborns with a genetic predisposition for T1D or first-degree relative(s) with T1D. This has enabled the TEDDY study to collect a range of biospecimens, including monthly stool samples starting at three months of age, coupled with extensive clinical and personal data such as diet, illnesses, medications and other life experiences. To characterize microbial, environmental, genetic, immunological and additional contributors to the development

[1]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [2]Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA, USA. [3]Health Informatics Institute, Morsani College of Medicine, University of South Florida, Tampa, FL, USA. [4]Department of Clinical Sciences, Lund University/CRC, Skåne University Hospital SUS, Malmo, Sweden. [5]Pacific Northwest Research Institute, Seattle, WA, USA. [6]Barbara Davis Center for Childhood Diabetes, University of Colorado, Aurora, CO, USA. [7]Center for Biotechnology and Genomic Medicine, Medical College of Georgia, Augusta University, Augusta, GA, USA. [8]Department of Pediatrics, Turku University Hospital, Turku, Finland. [9]Institute of Biomedicine, Research Centre for Integrative Physiology and Pharmacology, University of Turku, Turku, Finland. [10]Institute of Diabetes Research, Helmholtz Zentrum München, Munich, Germany. [11]Forschergruppe Diabetes, Technische Universität München, Klinikum Rechts der Isar, Munich, Germany. [12]Forschergruppe Diabetes e.V. at Helmholtz Zentrum München, Munich, Germany. [13]National Institute of Diabetes & Digestive & Kidney Diseases, Bethesda, MD, USA. [14]Alkek Center for Metagenomics and Microbiome Research, Department of Molecular Virology and Microbiology, Baylor College of Medicine, Houston, TX, USA. [15]Institute of Cellular Medicine, Newcastle University, Newcastle upon Tyne, UK. [16]Department of Computer Science, Aalto University, Espoo, Finland. [17]Gastrointestinal Unit, Center for the Study of Inflammatory Bowel Disease, and Center for Computational and Integrative Biology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA. [18]Center for Microbiome Informatics and Therapeutics, MIT, Cambridge, MA, USA. [19]Present address: Janssen Human Microbiome Institute, Janssen Research and Development, Cambridge, MA, USA. [20]These authors jointly supervised this work: Curtis Huttenhower, Ramnik J. Xavier. *e-mail: vatanen@broadinstitute.org; chuttenh@hsph.harvard.edu; xavier@molbio.mgh.harvard.edu

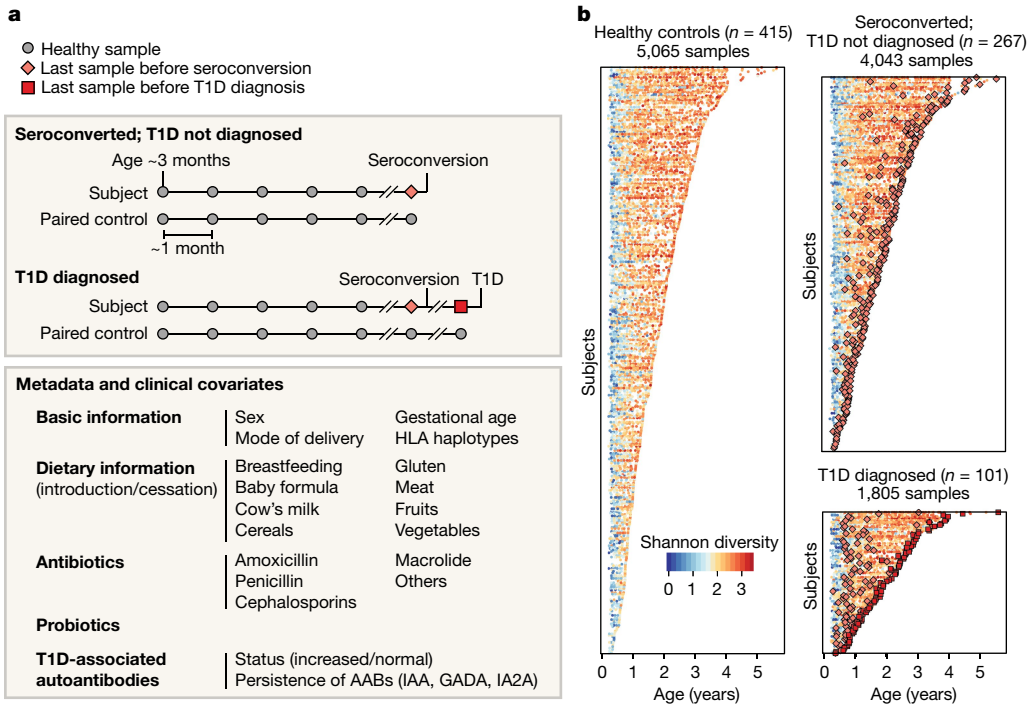



**Fig. 1 | More than 10,000 longitudinal gut metagenomes from the TEDDY T1D cohort.** We analysed 10,913 metagenomes collected longitudinally from 783 children (415 controls, 267 seroconverters, and 101 diagnosed with T1D) approximately monthly over the first five years of life. **a**, Subjects were recruited at six clinical centres (Finland, Sweden, Germany, Washington, Georgia and Colorado). Primary end points were seroconversion (defined as persistent confirmed IA) and T1D diagnosis. Additional metadata analysed for subjects and samples included the status of breastfeeding, birth mode, probiotics, antibiotics, formula feeding, and other dietary covariates. **b**, Overview of stool samples collected and microbiome development as summarized by Shannon's alpha diversity and stratified by end point. Median number of samples per individual $n = 12$ (healthy controls $n = 10$, seroconverters $n = 13$, T1D cases $n = 16$).

of T1D, the TEDDY study group further assembled nested case–control studies for IA ($n = 418$ case–control pairs) and T1D ($n = 114$)[29]. Case–control pairs were matched by clinical centre, sex and family history of T1D, which are all known confounding factors for T1D susceptibility and microbiome composition.

Here, we assessed 783 children followed from three months to up to five years of age from six clinical centres in four countries (Finland, Germany, Sweden and the United States) who either progressed to persistent IA or T1D or were matched as controls (Fig. 1a, b, Extended Data Table 1). Stool samples were collected, on average, monthly starting at three months of age and continuing until the clinical end point (IA or T1D). This study focused solely on analysing metagenomic sequencing data ($n = 10{,}903$ samples, $n = 783$ subjects), while a companion paper by Stewart et al.[30] interrogated corresponding 16S rRNA amplicon sequencing information.

We first investigated the taxonomic composition of early gut metagenomes at the species level. Principal coordinate analysis ordination of Bray–Curtis beta diversities showed a strong longitudinal gradient and marked heterogeneity among the earliest samples (Fig. 2a, Extended Data Fig. 1a–k, Supplementary Note 1). Permutational analysis of variance (ANOVA) of Bray–Curtis beta diversities indicated that inter-subject differences explained 35% of microbial taxonomic variation (permutation test, $P < 0.001$, 1,000 permutations), followed by age at stool sampling at roughly 4% of variance ($P < 0.001$). Using cross-sectional analysis to test for associations between taxonomic beta diversities and other collected metadata, we found that in addition to subject ID and age, geographical location and breastfeeding had strong and systematic effects on the composition of the microbial community (Supplementary Table 1, Extended Data Fig. 2a–d, Supplementary Note 1). To investigate the stability and individuality of the microbial profiles further, we compared intra- and inter-subject Bray–Curtis beta diversities. The gap between individual stability and similarity within or across clinical centres was largest at the beginning of the sampling period, indicating that the children had particularly dissimilar microbiota during these early months (Fig. 2b, Supplementary Note 1). Finally, we tested microbial alpha diversity (Shannon's diversity index) of taxonomic profiles for associations with collected metadata, and found that the cessation of breastfeeding had the largest effect (ANOVA, partial $\eta^2 = 0.053$) in the accrual of alpha diversity in early life (Supplementary Table 2, Extended Data Fig. 3a–e, Supplementary Note 1).

We next investigated the effects of antibiotics on the early life microbiome. Courses of oral antibiotics disrupted microbial stability, with a larger effect in the earliest comparisons (Fig. 2c, Extended Data Fig. 4a–f, Extended Data Table 2, Supplementary Note 2). Previous studies have found *Bifidobacterium* species to be especially vulnerable to antibiotics[31,32], leading us to investigate how antibiotic perturbations influenced these common dominant members of the early gut. Comparing microbial relative abundances before and after antibiotics (assuming that the given species was present in the preceding sample), we saw a decrease in the abundances of the *Bifidobacterium* members *B. bifidum*, *B. pseudocatenulatum*, *B. adolescentis*, *B. dentium* and *B. catenulatum*, whereas *B. longum* and *B. breve* did not systematically decline owing to antibiotics (Fig. 2d), suggesting that certain *Bifidobacterium* species are particularly susceptible to out-competition by other community members after depletion by antibiotics. Given their dominance in the typical developing gut microbiota and finely tuned balance of metabolic interactions with breast milk, this finding underscores the importance of approaching antibiotic prescriptions in early childhood with care, especially during breastfeeding.

Accompanying our taxonomic profiling, functional profiling of these metagenomes suggested the development of a consistent microbial functional core during infancy, with a smaller subject-specific variable functional pool (Extended Data Fig. 5a, b, Supplementary Note 3). As in most microbial community studies[33], microbial gene families of uncharacterized function made up a substantial fraction of these

profiles, averaging roughly 50% based on Gene Ontology[34] annotations (Extended Data Fig. 5c) and more than 90% based on more functionally specific MetaCyc pathways (Extended Data Fig. 5d). We observed an increasing longitudinal trend in the proportion of unmapped reads (Extended Data Fig. 5e, Pearson's $r = 0.318$, $P < 2.2 \times 10^{-16}$). However, within the reads that mapped to either microbial pangenomes or known protein sequences (the proportion of which decreased with age), we saw an increase in the proportion of reads with MetaCyc annotation, mainly during the first year (Extended Data Fig. 5f, Pearson $r = 0.391$, $P < 2.2 \times 10^{-16}$). This suggests that although the early life microbiome is relatively well-covered by current microbial reference genomes, less functional and biochemical characterization has been carried out on gene families within these microorganisms, which will thus particularly benefit from future work.

In addition to broadly conserved and subject-specific functions, we identified a range of microbial metabolic enzymes that consistently increased or decreased in abundance over the first year of life, paralleling shifts in community structure and infant diet (Fig. 3, Supplementary Note 3, Supplementary Table 3). For example, the enzyme L-lactate dehydrogenase (1.1.1.27), which is well-characterized in Bifidobacteria for its role in milk fermentation[35], was among the most consistently declining enzymes over this period, notably coinciding with the cessation of breastfeeding in many infants (from 73% breastfed at month 3 to 28% at year 1). Conversely, the enzyme transketolase (2.2.1.1), which has been implicated previously[36] in the metabolism of fibre, was among the most consistently increasing enzymes, which also coincided with increased incorporation of solid food (a component of 53% of infants' diets at month 3 versus 100% at year 1). Hence, these notable changes in community functional potential highlight the unique metabolic environment of the early infant gut, and the subsequent transition to a more adult-like gut microbiome that is adapted to variable, fermentative energy sources.

Combining taxonomic and functional profiles to test for differences between cases and controls, we used linear mixed-effects modelling and identified a relatively small number of individual taxonomic and functional features that were associated with case–control outcome (Supplementary Table 4), most with borderline statistical significance (false discovery rate (FDR) corrected $q$-values indicated below). We confirmed separation between cases and controls by random forest classifiers (Extended Data Fig. 6a, b, Supplementary Note 4). In the IA case–control cohort, healthy controls contained higher levels of *Lactobacillus rhamnosus* ($q = 0.055$), supporting protection against IA by early probiotic supplementation[37] (Extended Data Fig. 6c, d, Supplementary Note 5). IA controls also had more *Bifidobacterium dentium* ($q = 0.054$), whereas IA cases had on average higher abundance of *Streptococcus* group *mitis/oralis/pneumoniae* species ($q = 0.11$). In T1D case–control comparisons, controls had higher levels of *Streptococcus thermophilus* ($q = 0.078$) and *Lactococcus lactis* ($q = 0.094$) species, both common in dairy products, whereas cases contained higher levels of species such as *Bifidobacterium pseudocatenulatum* ($q = 0.078$), *Roseburia hominis* ($q = 0.11$) and *Alistipes shahii* ($q = 0.14$). Even though our modelling approach controlled for regional differences in clinical centres, we found additional but often weak associations with outcome in some clinical centres when tested separately (Supplementary Table 4). Finnish IA cases had more *Streptococcus* group *mitis/oralis/pneumoniae* species ($q = 0.0008$), IA controls from Colorado had more *Streptococcus thermophilus* ($q = 0.0059$), and Swedish IA cases contained more *Bacteroides vulgatus* ($q = 0.090$).

Pathways with the highest statistical significance in case–control comparisons were related to bacterial fermentation (Supplementary Table 4). The superpathway of fermentation (MetaCyc identifier PWY4LZ-257) was increased in controls in the T1D cohort ($q = 0.019$) and Finnish IA cohort ($q = 0.049$). SCFAs such as butyrate, acetate and propionate are common by-products of bacterial fermentation, and butyrate and acetate protected NOD mice against T1D[9]. Consistently, we observed that several bacterial pathways that contribute to the biosynthesis of short-chain fatty acids were increased in healthy controls.

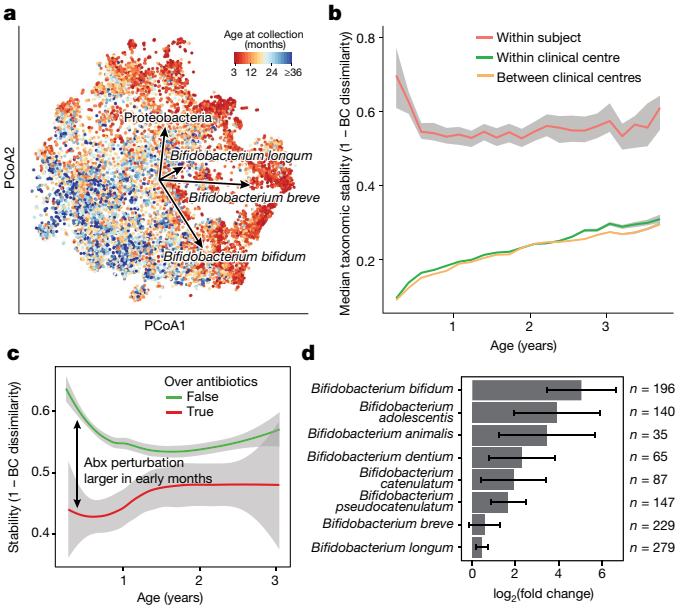



**Fig. 2 | The early gut microbiome is characterized by early heterogeneity of *Bifidobacterium* species and individualized accrual of taxa over time. a**, Principal coordinate analysis (PCoA) ordination of microbial beta diversities ($n = 10{,}913$ samples), measured by Bray–Curtis dissimilarity. Arrows show the weighted averages of key taxonomic groups. **b**, Microbiota stability, measured by Bray–Curtis (BC) dissimilarity ($n = 10{,}750$ samples) in three-month time windows, over two-month increments, stratified into three groups: within subject, within clinical centre, and between clinical centres. Lines show median values per time window. Shaded area denotes the estimated 99% confidence interval. Gut microbial communities were highly individual. **c**, Influence of antibiotic (Abx) courses on microbial stability, measured by Bray–Curtis dissimilarity over consecutive stool samples (<50 days apart) from the same individual during the first three years of life, and stratified by whether antibiotics were given between the two samples ($n = 654$ observations with antibiotics, $n = 6{,}734$ observations without antibiotics). Curves show locally weighted scatterplot smoothing (LOESS) for the data per category. Shaded areas show permutation-based 95% confidence intervals for the fit. **d**, Decreases in the most common *Bifidobacterium* species in connection to oral antibiotic treatments. Fold change was measured between consecutive samples with an antibiotic course between them, given that the species in question was present in the first of the two samples. Sample size per species ($n$) indicates the number of sample pairs in which the species in question was present in the sample before the antibiotic treatment. Bars show bootstrapped mean $\log_2$(fold change) (that is, decrease), and error bars denote s.d. ($n = 1{,}000$ bootstrap samples).

Among pathways involved in butyrate production, the degradation of L-arginine, putrescine and 4-aminobutanoate (ARGDEG-PWY) superpathway was increased in T1D controls cohort-wide ($q = 0.043$), whereas the fermentation of acetyl coenzyme A to butanoate (PWY-5676) was more abundant in the Finnish T1D controls ($q = 0.053$). The degradation of acetylene (P161-PWY), which contributes to acetate production, was increased in T1D controls cohort-wide ($q = 0.14$), and the degradation of L-1,2-propanediol (PWY-7013), which is involved in propionate biosynthesis, was higher in the German T1D controls ($q = 0.019$). These findings support existing evidence for the protective effects of SCFAs in human T1D[7,8] and T2D[19] cohorts and the NOD mouse model[9].

As reflected by the community-level analyses, human milk with its pro- and prebiotic functions is one of the main factors that determine the community composition of the infant gut microbiome. *Bifidobacterium longum* subsp. *infantis* is a particularly versatile degrader of human milk oligosaccharide (HMO) that is often found in stool samples collected during breastfeeding[38]. By following the families representing genes in the *B. longum* subsp. *infantis* HMO gene cluster[39,40] in our data, we found that an additional 30 bacterial species

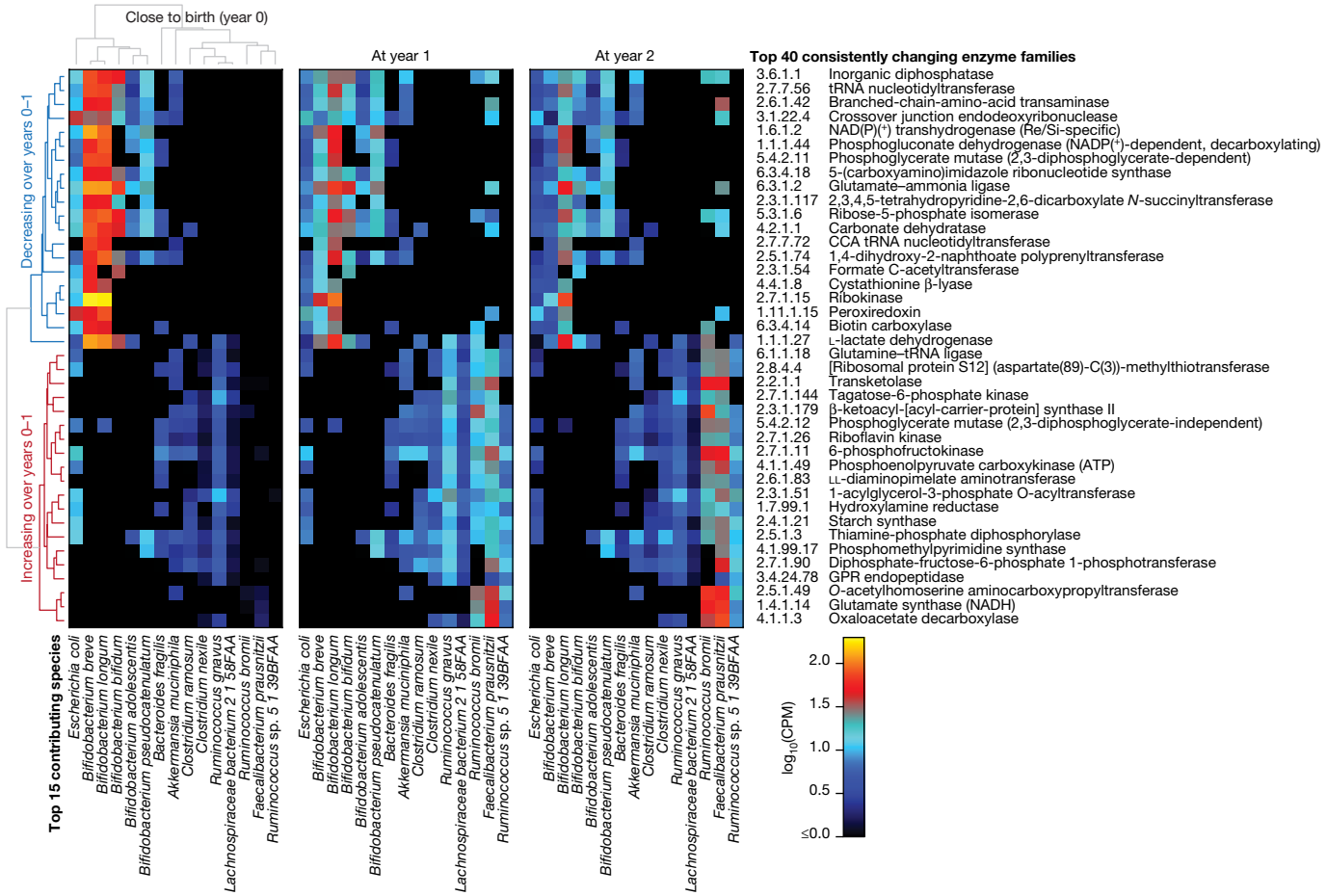



**Fig. 3 | Consistent changes in enzymatic content of the gut microbiome in early life.** We identified enzyme families (level-4 Enzyme Commission (EC) categories) that exhibited the most consistent within-subject changes in total community abundance between the ages of 3 months and 1 year. The top 20 most consistent increases or decreases are presented and stratified according to their top 15 contributing species. Heat map values reflect the mean contribution of each species to each enzyme over samples ($n = 733$ at 3 months; 675 at 1 year; and 382 at 2 years). Values reflect units of copies per million (CPM) normalized to total read depth (including unmapped reads and reads mapped to gene families lacking EC annotation). Rows (enzymes) and columns (species) are clustered according to Spearman correlation at 3 months; subsequent years are ordered according to clustering at 3 months.

carried at least one homologue with more than 50% sequence identity to one or more HMO utilization genes (Supplementary Table 5). As expected, many Bifidobacteria carried several homologues, but surprisingly three *Enterococcus* species (*E. casseliflavus*, *E. faecalis* and *E. faecium*) also carried seven or more homologues (Supplementary Table 5).

To identify strain-level adaptation similar to *B. longum* subsp. *infantis*, we further examined whether any of these genes showed contrasting prevalence between samples collected during breastfeeding and after weaning, given that the carrier species itself was present. In total, 41 gene families were observed more often during breastfeeding (Supplementary Table 5, test of proportions, adjusted $P < 0.001$); most (37 out of 41) were carried by *B. longum* (Fig. 4), and *B. pseudocatenulatum* contained four such gene families (Extended Data Fig. 7, Supplementary Table 5). In samples with *B. longum*, this implicated a clear strain shift after weaning, when fewer *B. longum* strains carried these genes (Fig. 4). In samples with *B. pseudocatenulatum*, four gene families showed a similar but less contrasting pattern (Extended Data Fig. 7). Overall, these observations identify new candidate species that contribute to HMO processing or exploitation, and link strain composition to specific driving molecular functions that potentially explain selective sweeps during microbiome development, in this case specifically related to breastfeeding.

Despite ample sample size, scrutiny of the study design, and thorough statistical analyses, most of the taxonomic and functional signals we detected in case–control comparisons were modest in effect size and statistical significance. This could be due to several reasons—differences between T1D endotypes, temporally diffuse signals, geographical heterogeneity, or lack of stool samples for the first two months of life —and these should be considered in future investigations (Supplementary Note 6). Furthermore, the data used in these investigations was composed of samples from the genetically predisposed and mostly white, non-Hispanic case–control groups designed into the TEDDY study. Results cannot be guaranteed to reflect the whole TEDDY cohort or child populations in the respective countries.

Future targeted approaches to identify subject-specific connections between the gut microbiota and T1D pathogenesis may be beneficial, particularly given the apparent population-level heterogeneity revealed here. For example, laboratory experiments involving dietary factors that have been associated with the onset of T1D[3] may reveal biochemically specific signals that are mediated by the microbiome. Different endotypes of disease, such as differences in the first appearing autoantibody (IAA versus GADA), the number of appearing autoantibodies, the time from seroconversion to T1D diagnosis, genetic host risk alleles and ethnic backgrounds, may be characterized by distinct microbial configurations (Supplementary Note 6). Finally, components of the microbiome that were poorly measured in these data may also have crucial roles: viruses, fungi, microbial transcription or small-molecule biochemistry. By surveying these additional molecular activities by cross-sectional analysis and in more detailed longitudinal populations, this study lays

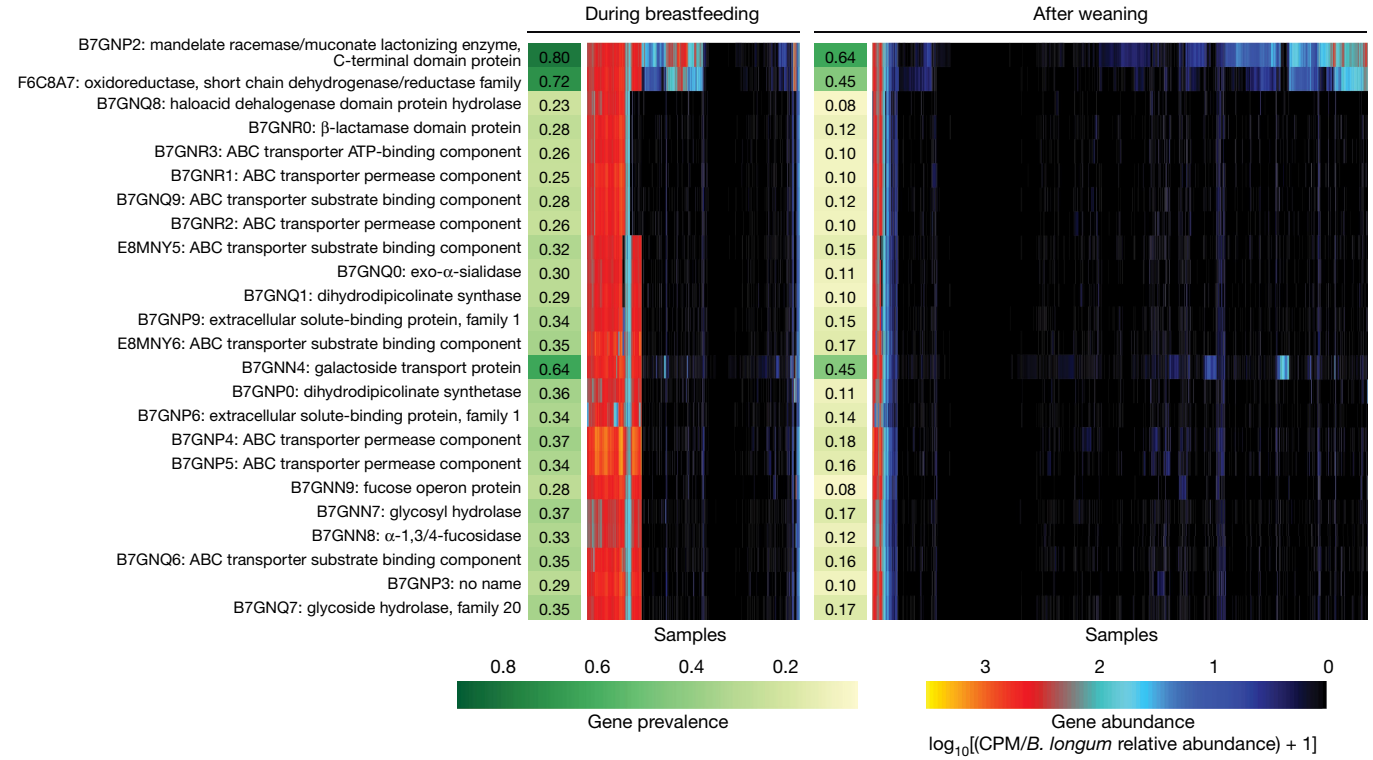



**Fig. 4 | *Bifidobacterium longum* strains are characterized by HMO gene content and stratified by breastfeeding status.** Gene families involved in HMO utilization and showing contrasting presence in *B. longum* genomes during breastfeeding ($n = 1{,}584$ samples) compared to after weaning ($n = 3{,}705$ samples). Abundance heat map columns represent stool samples in which the relative abundance of *B. longum* species was more than 10% ($n = 5{,}289$ samples). Rows and columns were ordered by hierarchical clustering using the complete linkage method. As in Fig. 3, values reflect units of CPM and were further divided by relative abundance of *B. longum* to obtain quantifications that are comparable between samples. UniRef90 identifiers and gene names or families are indicated on the left.

the foundation to identify further gut microbial components that are predictive, protective or potentially causal in T1D risk or pathogenesis.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0620-2.

Received: 16 November 2017; Accepted: 6 September 2018;
Published online 24 October 2018.

1. Katsarou, A. et al. Type 1 diabetes mellitus. *Nat. Rev. Dis. Primers* **3**, 17016 (2017).
2. Pociot, F. & Lernmark, Å. Genetic risk factors for type 1 diabetes. *Lancet* **387**, 2331–2339 (2016).
3. Rewers, M. & Ludvigsson, J. Environmental risk factors for type 1 diabetes. *Lancet* **387**, 2340–2348 (2016).
4. Knip, M. & Siljander, H. The role of the intestinal microbiota in type 1 diabetes mellitus. *Nat. Rev. Endocrinol.* **12**, 154–167 (2016).
5. Hober, D. & Sauter, P. Pathogenesis of type 1 diabetes mellitus: interplay between enterovirus and host. *Nat. Rev. Endocrinol.* **6**, 279–289 (2010).
6. Paun, A., Yau, C. & Danska, J. S. The influence of the microbiome on type 1 diabetes. *J. Immunol.* **198**, 590–595 (2017).
7. de Goffau, M. C. et al. Aberrant gut microbiota composition at the onset of type 1 diabetes in young children. *Diabetologia* **57**, 1569–1577 (2014).
8. de Goffau, M. C. et al. Fecal microbiota composition differs between children with β-cell autoimmunity and those without. *Diabetes* **62**, 1238–1244 (2013).
9. Mariño, E. et al. Gut microbial metabolites limit the frequency of autoimmune T cells and protect against type 1 diabetes. *Nat. Immunol.* **18**, 552–562 (2017).
10. Needell, J. C. & Zipris, D. The role of the intestinal microbiome in type 1 diabetes pathogenesis. *Curr. Diab. Rep.* **16**, 89 (2016).
11. Davis-Richardson, A. G. et al. *Bacteroides dorei* dominates gut microbiome prior to autoimmunity in Finnish children at high risk for type 1 diabetes. *Front. Microbiol.* **5**, 678 (2014).
12. Kostic, A. D. et al. The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. *Cell Host Microbe* **17**, 260–273 (2015).
13. Endesfelder, D. et al. Compromised gut microbiota networks in children with anti-islet cell autoimmunity. *Diabetes* **63**, 2006–2014 (2014).
14. Maffeis, C. et al. Association between intestinal permeability and faecal microbiota composition in Italian children with beta cell autoimmunity at risk for type 1 diabetes. *Diabetes Metab. Res. Rev.* **32**, 700–709 (2016).
15. Mejía-León, M. E., Petrosino, J. F., Ajami, N. J., Domínguez-Bello, M. G. & de la Barca, A. M. Fecal microbiota imbalance in Mexican children with type 1 diabetes. *Sci. Rep.* **4**, 3814 (2014).
16. Alkanani, A. K. et al. Alterations in intestinal microbiota correlate with susceptibility to type 1 diabetes. *Diabetes* **64**, 3510–3520 (2015).
17. Soyucen, E. et al. Differences in the gut microbiota of healthy children and those with type 1 diabetes. *Pediatr. Int.* **56**, 336–343 (2014).
18. Endesfelder, D. et al. Towards a functional hypothesis relating anti-islet cell autoimmunity to the dietary impact on microbial communities and butyrate production. *Microbiome* **4**, 17 (2016).
19. Zhao, L. et al. Gut bacteria selectively promoted by dietary fibers alleviate type 2 diabetes. *Science* **359**, 1151–1156 (2018).
20. Markle, J. G. et al. Sex differences in the gut microbiome drive hormone-dependent regulation of autoimmunity. *Science* **339**, 1084–1088 (2013).
21. Costa, F. R. et al. Gut microbiota translocation to the pancreatic lymph nodes triggers NOD2 activation and contributes to T1D onset. *J. Exp. Med.* **213**, 1223–1239 (2016).
22. Yatsunenko, T. et al. Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (2012).
23. Palmer, C., Bik, E. M., DiGiulio, D. B., Relman, D. A. & Brown, P. O. Development of the human infant intestinal microbiota. *PLoS Biol.* **5**, e177 (2007).
24. Koenig, J. E. et al. Succession of microbial consortia in the developing infant gut microbiome. *Proc. Natl Acad. Sci. USA* **108**, 4578–4585 (2011).
25. Dominguez-Bello, M. G. et al. Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc. Natl Acad. Sci. USA* **107**, 11971–11975 (2010).
26. Bäckhed, F. et al. Dynamics and stabilization of the human gut microbiome during the first year of life. *Cell Host Microbe* **17**, 690–703 (2015).
27. Yassour, M. et al. Natural history of the infant gut microbiome and impact of antibiotic treatment on bacterial strain diversity and stability. *Sci. Transl. Med.* **8**, 343ra81 (2016).
28. Hagopian, W. A. et al. The Environmental Determinants of Diabetes in the Young (TEDDY): genetic criteria and international diabetes risk screening of 421 000 infants. *Pediatr. Diabetes* **12**, 733–743 (2011).
29. Lee, H. S. et al. Biomarker discovery study design for type 1 diabetes in The Environmental Determinants of Diabetes in the Young (TEDDY) study. *Diabetes Metab. Res. Rev.* **30**, 424–434 (2014).
30. Stewart, C. J. et al. Temporal development of the gut microbiome in early childhood from the TEDDY study. *Nature* https://doi.org/10.1038/s41586-018-0617-x (2018).

31. Zhernakova, A. et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* **352**, 565–569 (2016).
32. Korpela, K. et al. Intestinal microbiome is related to lifetime antibiotic use in Finnish pre-school children. *Nat. Commun.* **7**, 10410 (2016).
33. Joice, R., Yasuda, K., Shafquat, A., Morgan, X. C. & Huttenhower, C. Determining microbial products and identifying molecular targets in the human microbiome. *Cell Metab.* **20**, 731–741 (2014).
34. Ashburner, M. et al. Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
35. O'Callaghan, A. & van Sinderen, D. Bifidobacteria and their role as members of the human gut microbiota. *Front. Microbiol.* **7**, 925 (2016).
36. Thurston, B., Dawson, K. A. & Strobel, H. J. Pentose utilization by the ruminal bacterium *Ruminococcus albus*. *Appl. Environ. Microbiol.* **60**, 1087–1092 (1994).
37. Uusitalo, U. et al. Association of early exposure of probiotics and islet autoimmunity in the TEDDY Study. *JAMA Pediatr.* **170**, 20–28 (2016).
38. Underwood, M. A., German, J. B., Lebrilla, C. B. & Mills, D. A. *Bifidobacterium longum* subspecies *infantis*: champion colonizer of the infant gut. *Pediatr. Res.* **77**, 229–235 (2015).
39. Sela, D. A. et al. The genome sequence of *Bifidobacterium longum* subsp. *infantis* reveals adaptations for milk utilization within the infant microbiome. *Proc. Natl Acad. Sci. USA* **105**, 18964–18969 (2008).
40. Vatanen, T. et al. Variation in microbiome LPS immunogenicity contributes to autoimmunity in humans. *Cell* **165**, 842–853 (2016).


**Acknowledgements** This research was performed on behalf of the TEDDY Study Group, which is funded by U01 DK63829, U01 DK63861, U01 DK63821, U01 DK63865, U01 DK63863, U01 DK63836, U01 DK63790, UC4 DK63829, UC4 DK63861, UC4 DK63821, UC4 DK63865, UC4 DK63863, UC4 DK63836, UC4 DK95300, UC4 DK100238, UC4 DK106955, UC4 DK112243, UC4 DK117483, and contract no. HHSN267200700014C from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Institute of Allergy and Infectious Diseases (NIAID), National Institute of Child Health and Human Development (NICHD), National Institute of Environmental Health Sciences (NIEHS), Centers for Disease Control and Prevention (CDC), and JDRF. This work supported in part by the NIH/NCATS Clinical and Translational Science Awards to the University of Florida (UL1 TR000064) and the University of Colorado (UL1 TR001082). C.H. was supported by funding from JDRF (3-SRA-2016-141-Q-R) and NIDDK (U54DE023798, R24DK110499). H.V. and R.J.X. were supported by funding from JDRF (2-SRA-2016-247-S-B, 2-SRA-2018-548-S-B).

**Reviewer information** *Nature* thanks K. Aagaard, C. Lozupone and L. Wen for their contribution to the peer review of this work.

**Author contributions** T.V., E.A.F. and R.S. analysed the metagenomic sequencing data. C.J.S., N.J.A. and J.F.P. generated the metagenomic sequencing data. S.T., T.D.A. and H.V. designed and conducted bacterial growth assays. K.V., Å.L., W.A.H., M.J.R., J.-X.S., J.T., A.-G.Z., B.A. and J.P.K. contributed to the study concept, design and sample acquisition. H.L., H.V., C.H. and R.J.X. served as principal investigators. T.V., E.A.F., H.V., C.H. and R.J.X drafted the manuscript. All authors discussed the results, contributed to critical revisions and approved the final manuscript. Members of the TEDDY Study Group are listed in the Supplementary Information.

**Competing interests** The authors declare no competing interests.

**Additional information**
**Extended data** is available for this paper at https://doi.org/10.1038/s41586-018-0620-2.
**Supplementary information** is available for this paper at https://doi.org/10.1038/s41586-018-0620-2.
**Reprints and permissions information** is available at http://www.nature.com/reprints.
**Correspondence and requests for materials** should be addressed to T.V. or C.H. or R.J.X.
**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## METHODS

**Cohort and study design.** TEDDY is a prospective cohort study funded by the National Institutes of Health with the primary goal to identify environmental causes of T1D. It includes six clinical research centres—three in the United States (Colorado, Georgia/Florida, Washington) and three in Europe (Finland, Germany and Sweden). Detailed study design and methods have been previously published[28,41,42]. Written informed consents were obtained for all study participants from a parent or primary caretaker, separately, for genetic screening and participation in a prospective follow-up. The TEDDY study was approved by local US Institutional Review Boards and European Ethics Committee Boards in Colorado's Colorado Multiple Institutional Review Board, Georgia's Medical College of Georgia Human Assurance Committee (2004–2010), Georgia Health Sciences University Human Assurance Committee (2011–2012), Georgia Regents University Institutional Review Board (2013–2015), Augusta University Institutional Review Board (2015–present), Florida's University of Florida Health Center Institutional Review Board, Washington state's Washington State Institutional Review Board (2004–2012) and Western Institutional Review Board (2013–present), Finland's Ethics Committee of the Hospital District of Southwest Finland, Germany's Bayerischen Landesärztekammer (Bavarian Medical Association) Ethics Committee, Sweden's Regional Ethics Board in Lund, Section 2 (2004–2012) and Lund University Committee for Continuing Ethical Review (2013–present). The study is monitored by External Advisory Board formed by the National Institutes of Health.

This analysis used stool samples and clinical metadata from two nested case–control studies (persistent, confirmed IA or T1D) using risk set sampling[29]. The data used here were collected as of 31 May 2012, as a 1:1 match in which one control per case of persistent confirmed IA or T1D was selected from the full TEDDY cohort. A control was a participant who had not developed persistent, confirmed IA or T1D by the time the case to which it was matched had developed IA or T1D, within $\pm 45$ days of the event time. Matching factors were clinical centre, sex and family history of T1D to control for differences in geographical area, genetic background and in sample or data handling between clinical centres. In all case–control comparisons, we removed all case–control pairs in which the control later progressed to case status (that is, progressed to IA or T1D). In additional, 17 subjects with missing information about breastfeeding together with their matched pairs were excluded from the case–control comparisons to avoid confounding effects from unknown breastfeeding status.

The development of persistent, confirmed IA was assessed every three months. Persistent autoimmunity was defined by the presence of confirmed islet autoantibody on two or more consecutive visits. The date of persistent autoimmunity was defined as the draw date of the first sample of the two consecutive samples that deemed the child persistently positive for a specific autoantibody (or any autoantibody). T1D was defined according to American Diabetes Association criteria for diagnosis[43].

Stool samples were collected monthly starting at three months of age and continuing up until 48 months of age, then every three months until 10 years of age and then biannually thereafter, into the three plastic stool containers provided by the clinical centre. Children who were antibody negative after 4 years of age were encouraged to submit four times a year even though after 4 years their visits schedule switched to biannual. Parents sent the stool containers at either ambient or $+4$ °C temperature with guaranteed delivery within 24 h in the appropriate shipping box to the NIDDK repository if living in the United States or their affiliated clinical centre if living in Europe. The European clinical centres stored the stool samples and sent monthly bulk shipments of frozen stool to the NIDDK repository. The TEDDY Manual of Operations, including the stool sample collection protocol, can be accessed online at https://repository.niddk.nih.gov/static/studies/teddy/teddy_moop.pdf.

A priori power calculations using discrete Cox's proportional hazards regression[44] for the matched IA case–control study estimated 80% power, $\alpha = 0.01$, two-sided test to detect an odds ratio $> 3$ for an exposure with 5% prevalence, to an odds ratio $> 1.8$ for an exposure with 20% prevalence. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

**Metagenomic sequencing and initial bioinformatics.** Samples were metagenomically sequenced as one library each multiplexed through Illumina HiSeq machines using the $2 \times 100$-bp paired-end read protocol. Samples with limited DNA quantity and/or too few high-quality reads were filtered out, resulting in a discrepancy of sample frequencies between the metagenomic data and the 16S rRNA amplicon sequencing data analysed in the companion paper[30]. Casava v1.8.2 (Illumina) output initial FASTQ files from the resulting data were processed using cutadapt v1.9dev2 for adaptor removal, Trim Galore v0.2.8 (Babraham Bioinformatics) for removing low-quality bases and PRINSEQ v0.20.3[45] for sample demultiplexing. Bowtie2 v2.2.3 was used to map reads to the human genome for decontamination before subsequent analysis.

**Taxonomic and functional profiling by MetaPhlAn2 and HUMAnN2.** Taxonomic profiling of the metagenomic samples was performed using MetaPhlAn2[46] v2.6.0, which uses a library of clade-specific markers to provide pan-microbial (bacterial, archaeal, viral and eukaryotic) quantification at the species level. MetaPhlAn2 was run using default settings.

Functional profiling was performed with HUMAnN2[47] v0.9.4. For an input metagenome, HUMAnN2 constructs a sample-specific reference database by concatenating and indexing the pangenomes of species detected in the sample by MetaPhlAn2 (pangenomes are pre-clustered, pre-annotated catalogues of open reading frames found across isolate genomes from a given species[48]). HUMAnN2 then maps sample reads against this database to quantify gene presence and abundance in a species-stratified manner, with unmapped reads further used in a translated search against UniRef90[49] to include taxonomically unclassified but functionally distinct gene family abundances. Finally, for community-total, species-stratified, and unclassified gene family abundance, HUMAnN2 reconstructs metabolic pathway abundance based on the subset of gene families annotated to metabolic reactions (based on reaction and pathway definitions from MetaCyc[50]). Enzyme (level-4 Enzyme Commission (EC) categories) abundances were further computed by summing the abundances of individual gene families annotated to each EC number based on UniRef90-EC annotations from UniProt[51].

**Phenotype and covariate analysis.** This study includes extensive collection of clinical covariates that cover several aspects of common and rare life events in early childhood from infancy up to five years of age. In these analyses, we used information that is, according to the literature, of high relevance in terms of gut microbiome development. Information about mothers, pregnancy and birth was collected during the three-month clinic visit by questionnaire and included the mode of birth (vaginal birth versus caesarean section), gestational age, infant's 5-min Apgar score, information about maternal diabetes (T1D, T2D or gestational diabetes) and maternal insulin and medication use (antibiotics, angiotensin-converting enzyme inhibitors, metformin, glyburide, antihypertensives) during pregnancy. Dietary information used in these analyses includes the start (and end) date for the following dietary compounds: breastfeeding, baby formula, cow's milk, gluten, cereals, meat, vegetables and fruits. The start of solid food (anything other than breast milk or cow's milk) was also analysed separately. The T1D-associated autoantibodies IAA, GADA and IA2A were analysed from serum samples collected at each clinic visit. In addition to IA, defined as persistent, confirmed autoantibody seropositivity, we analysed the data in terms of the persistency and cumulative frequency of autoantibodies (single or multiple autoantibodies). In TEDDY, all prescribed antibiotic courses are recorded. We further stratified these data by the type of antibiotic in five categories: amoxicillin, penicillin, cephalosporins, macrolide and other antibiotics. Information about probiotics covered the dates for starting and stopping probiotic supplementation, but not the specific types of probiotics used. In addition, sex, information about whether first degree relatives in family had T1D, and HLA haplotypes of the subjects were used in these analyses. Subjects screened from the general population were identified with high-risk alleles (89%) including: DRB1*04-DQA1*03-DQB1*03:02/DRB1*03-DQA1*05-DQB1*02:01 (DR3/4), DRB1*04-DQA1*03-DQB1*03:02/DRB1*04-DQA1*03-DQB1*03:02 (DR4/4), DRB1*04-DQA1*03-DQB1*03:02/DRB1*08-DQA1*04-DQB1*04:02 (DR4/8) and DRB1*03-DQA1*05-DQB1*02:01/DRB1*03-DQA1*05-DQB1*02:01 (DR3/3), plus six genotypes specific to first-degree relatives[28].

Principal coordinate analysis (PCoA) ordination was generated using *t*-distributed stochastic neighbour embedding (*t*-SNE) as implemented in Rtsne package in R with Bray–Curtis dissimilarity as the distance measure and perplexity (a free parameter) equal to 50. Statistical significance of the trends between early clusters and metadata were tested using mixed-effect logistic regression and samples collected during the first year of life as follows. The target variable used was a binary indicator of whether the relative abundance of the taxon of interest (three different *Bifidobacterium* species or phylum Proteobacteria) was greater than 0.5 (definition of the cluster). The age of sample collection, mode of delivery, clinical centre, breastfeeding status (ongoing/stopped), solid food status (binary variable indicating whether solid food was introduced in the diet) and antibiotics status (binary variable indicating whether the subject received antibiotics during the last 30 days) were used as fixed effects, and the subject ID was used as a random effect.

Associations between microbial feature abundances and clinical outcome were determined using MaAsLin[52]. In brief, this multivariate linear modelling system for microbial data selects from among a set of (potentially high-dimensional) covariates to associate with microbial taxon or pathway abundances. Mixed-effects linear models using a variance-stabilizing arcsin square root transform on relative abundances are then used to determine the significance of putative associations from among this reduced set. In the models, subject ID was used as a random effect, and the age of sample collection, mode of delivery, clinical centre (for cohort-wide comparisons), breastfeeding status (ongoing or stopped), solid food status (binary variable indicating whether solid food was introduced in the diet), number of sequencing reads and case–control outcome were used as fixed effects. Nominal

*P* values were adjusted using the Benjamini–Hochberg FDR method. Here, microbial features with corrected $q < 0.25$ were reported. For metabolic pathways, pseudocount $2^6$ was added to CPM values to stabilize the variation in lowly abundant and/or prevalent but highly variable categories, and data were $\log_2$-transformed.

As previously described[40], to associate microbial diversity with covariates while accounting for nonlinear, age-dependent effects, we first fitted a sigmoid function (nls function in R) to account for the longitudinal trend. Residuals of this model were then used as inputs for a mixed-effect model (glmmPQL function in the MASS R package), with subject IDs as random effects to account for repeated measurements in the data. Other factors were included in the model as fixed effects, and their significance levels were evaluated using *P* values reported by the model (Supplementary Table 2).

The association between T1D case–control outcome and microbial alpha diversity in individual clinical centres was tested using a linear mixed-effects model (glmmPQL function in MASS R package) on samples 730 days or less before T1D diagnosis. In the model, the age at stool sample collection and T1D case–control outcome were used as fixed effects, and subject ID was used as a random effect.

**Microbial variance explained by clinical and other covariates.** Variance analysis was conducted using the adonis function in the vegan R package given a Bray–Curtis dissimilarity matrix of the taxonomic profiles and all TEDDY clinical metadata listed above. In brief, adonis conducts multivariate ANOVA using the dissimilarity matrix (that is, partitions the sums of squares) given the metadata as covariates. Statistical significance of the fit was assessed using permutation tests.

**HMO gene homology.** The HMO gene cluster homologues between *B. longum* subsp. *infantis* and multiple taxa were analysed as follows. UniRef90 gene families corresponding to the protein sequences in the *B. longum* subsp. *infantis* HMO gene cluster[39] (protein sequences Blon_2331-Blon_2361 in NCBI protein sequence database) were identified by translated BLAST search against ChocoPhlAn pangenome collection[48] used by HUMAnN2. Identified hits were further filtered by requiring $\geq$50% alignment identity and $\geq$80% mutual coverage. Combining this information with HUMAnN2 species-stratified UniRef90 gene family quantification enabled calling these genes present given that they had sufficient read coverage, here defined as $\log_{10}$(counts per million) $> 0.1$ in at least 50 samples collected during breastfeeding. Differential gene prevalence during breastfeeding was tested using the samples in which the carrier species had >1% relative abundance. Testing was conducted using the test of equal or given proportions (prop.test function in R) and by comparing the prevalence (proportion of the samples for which the species in question harboured the gene according to the metagenomic data) of the gene in samples collected during breastfeeding with the samples collected after weaning. *P* values were adjusted for multiple testing by Benjamini–Hochberg method (p.adjust function in R). All homologues together with their BLAST search metrics, prevalence in the metagenomic data and corresponding *B. infantis* HMO gene are reported in Supplementary Table 5.

**Bacterial growth assays.** *Bifidobacterium bifidum* strain RJX-1201, *Bifidobacterium breve* RJX-1202 and *Bifidobacterium longum* RJX-1203 were streaked on brain heart infusion agar (BD) supplemented with 1% vitamin K/hemin solution (BD; sBHI), and incubated for 48 h in a vinyl anaerobic chamber (Coy Laboratory Products) containing 5% $CO_2$, 5% $H_2$ and 90% $N_2$ and maintained at 37 °C. Cells were transferred to sBHI liquid medium (BHI broth, BD, supplemented as above) and grown for 24 h in anaerobic conditions. Cultures were washed twice with PBS and optical density at 600 nm ($OD_{600}$) was measured using a BioTek PowerWave 340 plate reader. $OD_{600}$ was normalized to 0.2 for all strains and 5 μl bacteria inoculum was added to a final volume of 200 μl containing 10% sBHI and 125 mM carbon source (glucose, fucose, galactose or lactose) in a 96-well plate. $OD_{600}$ was measured in the plate reader every hour for 48 h with 5 s of medium shaking before each measurement. All of the measurements were normalized to a medium-only blank. Experiment was repeated three times ($n = 3$) in triplicate and one representative experiment is shown. Error bars are s.d. of three technical replicates.

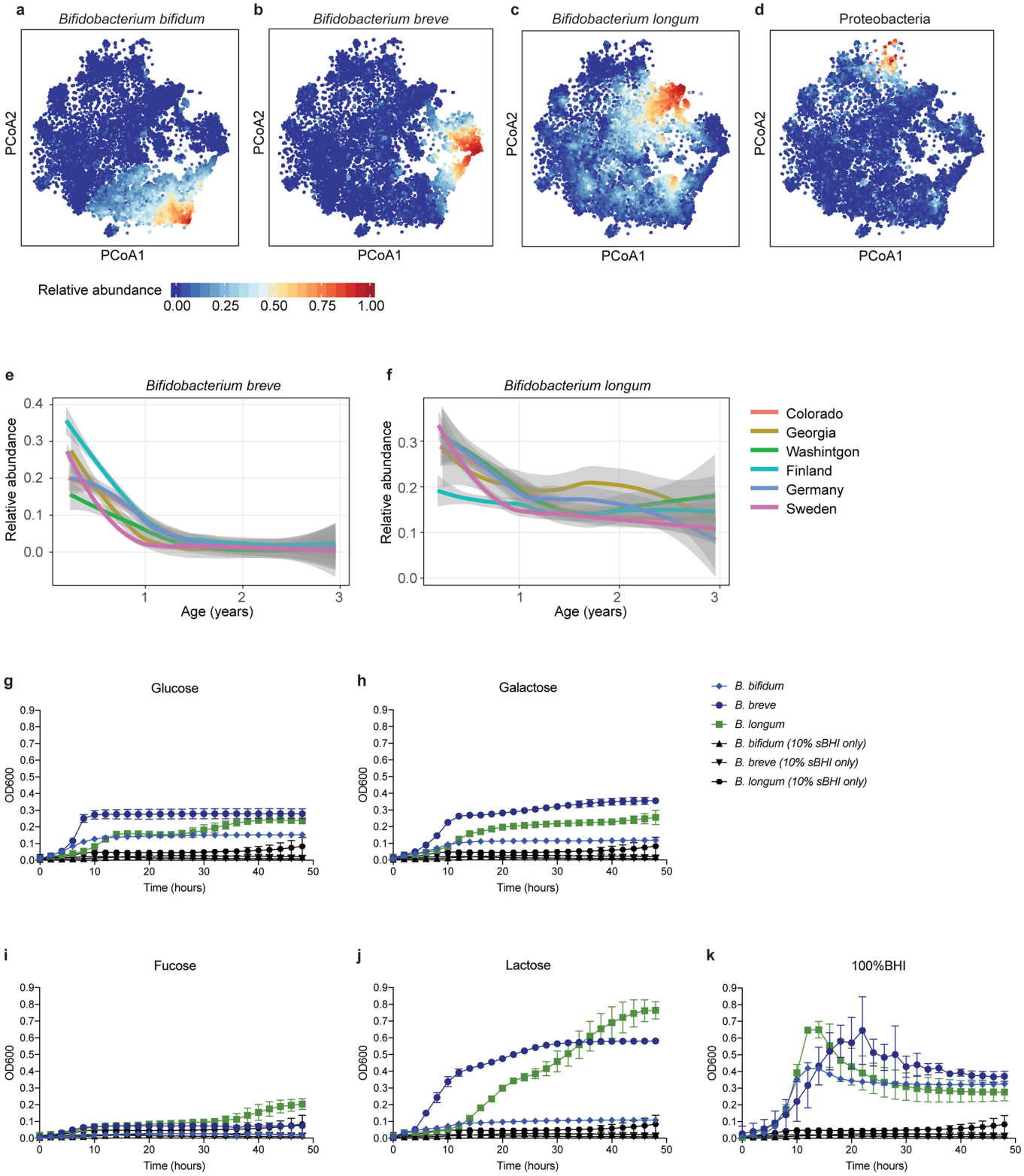
**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

**Code availability.** Code for Random Forest case–control comparisons and cohort wide MaAsLin association analyses in Supplementary Table 4 has been made publicly available at https://github.com/tvatanen/broad_teddy_microbiome_analyses. Other analysis software including quality control, taxonomic, and functional profilers is publicly available and referenced as appropriate.

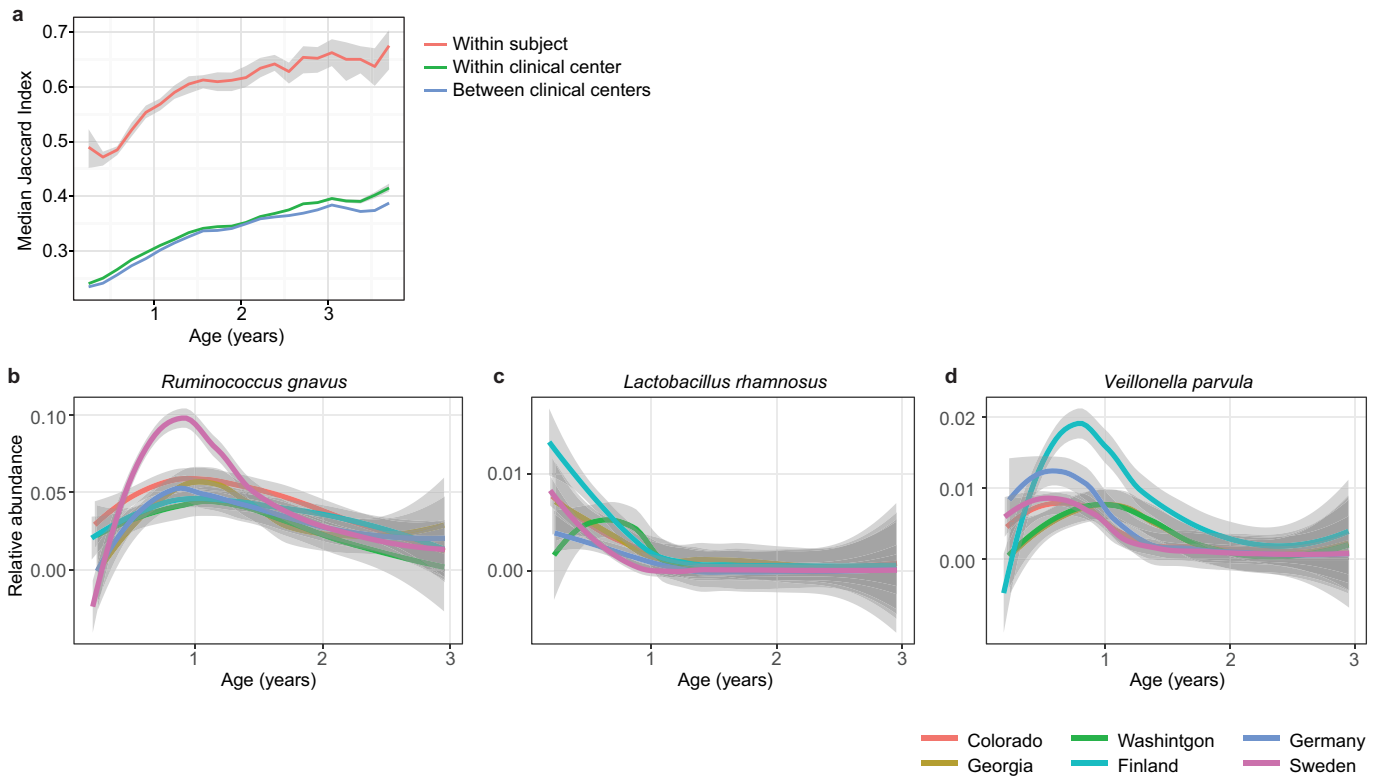
## Data availability

TEDDY microbiome 16S and whole-genome sequencing data that support the findings of this study are available in the NCBI database of Genotypes and Phenotypes (dbGaP) with the primary accession code phs001443.v1.p1, in accordance with the dbGaP controlled-access authorization process. Clinical metadata analysed during the current study are available in the NIDDK Central Repository at https://www.niddkrepository.org/studies/teddy.

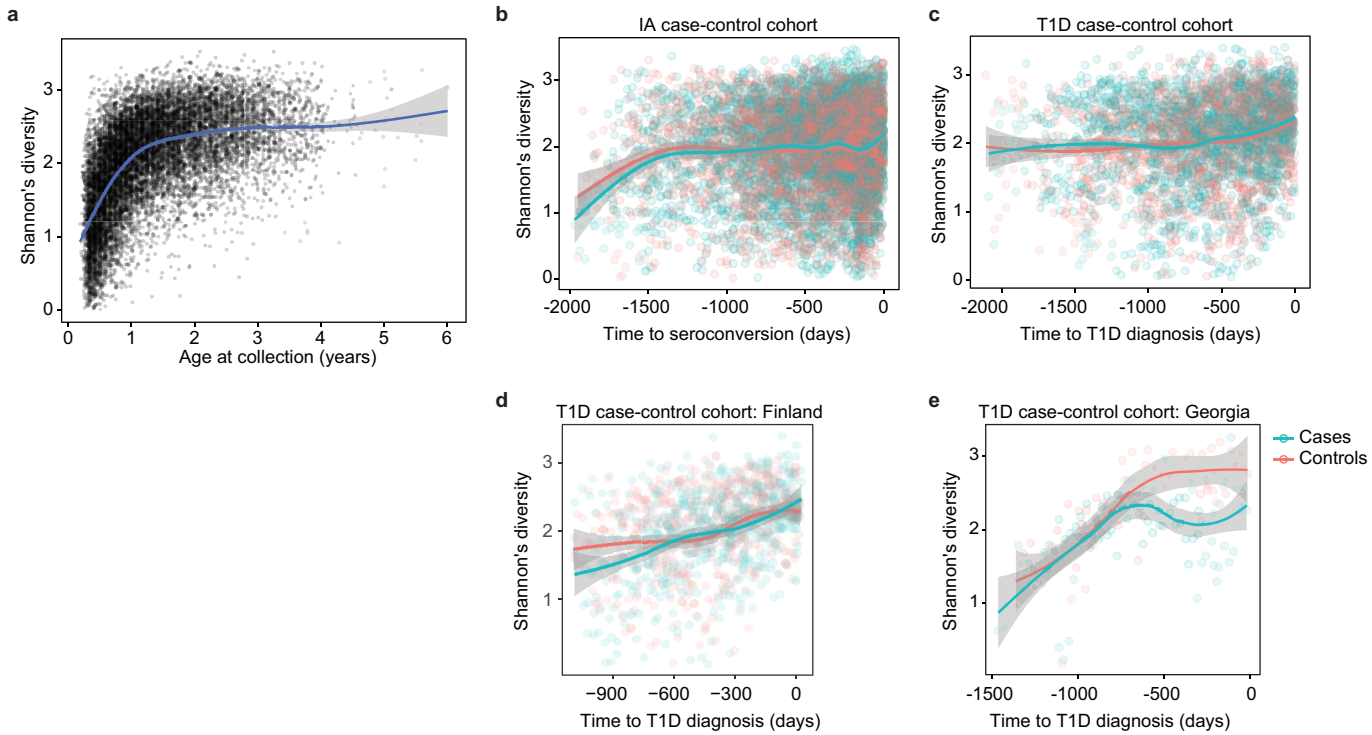

41. TEDDY Study Group. The Environmental Determinants of Diabetes in the Young (TEDDY) study: study design. *Pediatr. Diabetes* **8**, 286–298 (2007).
42. TEDDY Study Group. The Environmental Determinants of Diabetes in the Young (TEDDY) Study. *Ann. NY Acad. Sci.* **1150**, 1–13 (2008).
43. American Diabetes Association. 2. Classification and diagnosis of diabetes. *Diabetes Care* **38**, S8–S16 (2015).
44. Lachin, J. M. Sample size evaluation for a multiply matched case–control study using the score test from a conditional logistic (discrete Cox PH) regression model. *Statist. Med.* **27**, 2509–2534 (2012).
45. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**, 863–864 (2011).
46. Truong, D. T. et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).
47. Abubucker, S. et al. Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLOS Comput. Biol.* **8**, e1002358 (2012).
48. Huang, K. et al. MetaRef: a pan-genomic database for comparative and community microbial genomics. *Nucleic Acids Res.* **42**, D617–D624 (2014).
49. Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B. & Wu, C. H. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2015).
50. Caspi, R. et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **44**, D471–D480 (2016).
51. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169 (2017).
52. Morgan, X. C. et al. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol.* **13**, R79 (2012).



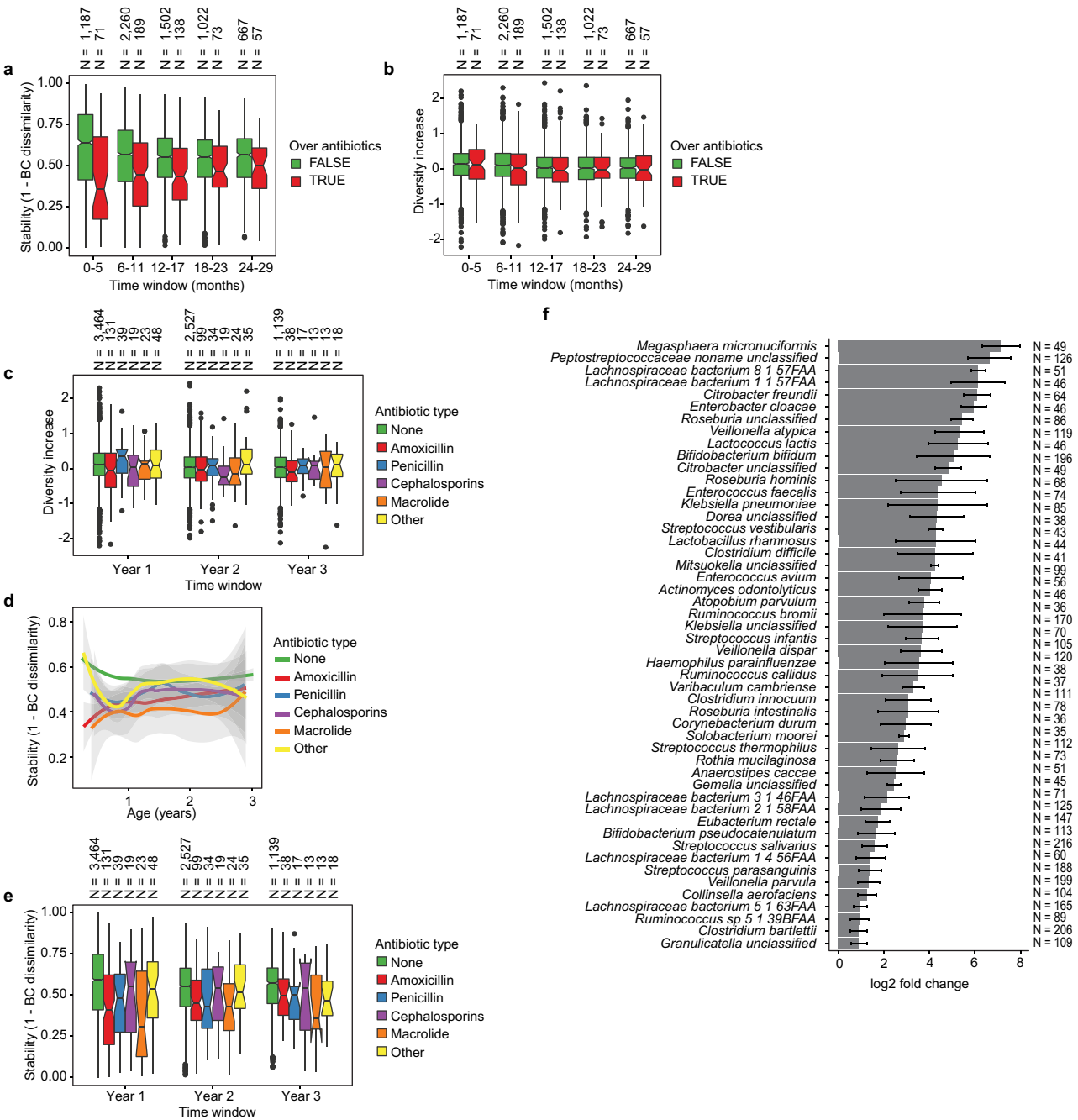

**Extended Data Fig. 1 | Heterogeneity in early taxonomic profiles.** **a**–**d**, Relative abundances of taxonomic groups highlighted by weighted averages in Fig. 2a (arrows) shown separately ($n = 10{,}913$ samples). **e**, **f**, Average longitudinal abundance of *B. breve* (**e**) and *B. longum* (**f**) per clinical centre ($n = 10{,}194$ samples). The curves show LOESS fits for the relative abundances, and shaded area shows 95% confidence interval for each fit, as implemented in geom_smooth function in ggplot2 R package. **g**–**k**, Growth curves of human infant isolates of *B. breve*, *B. bifidum* and *B. longum* grown individually in low-nutrient medium (10% sBHI) supplemented with single carbon sources (glucose (**g**), galactose (**h**), fucose (**i**) and lactose (**j**)) or grown in 100% sBHI (**k**). As a negative control, growth curves of each strain grown in 10% BHI without additional sugar are shown in black for each condition. Data are representative of three independent experiments and are presented as the mean and s.d. of triplicate assessments.



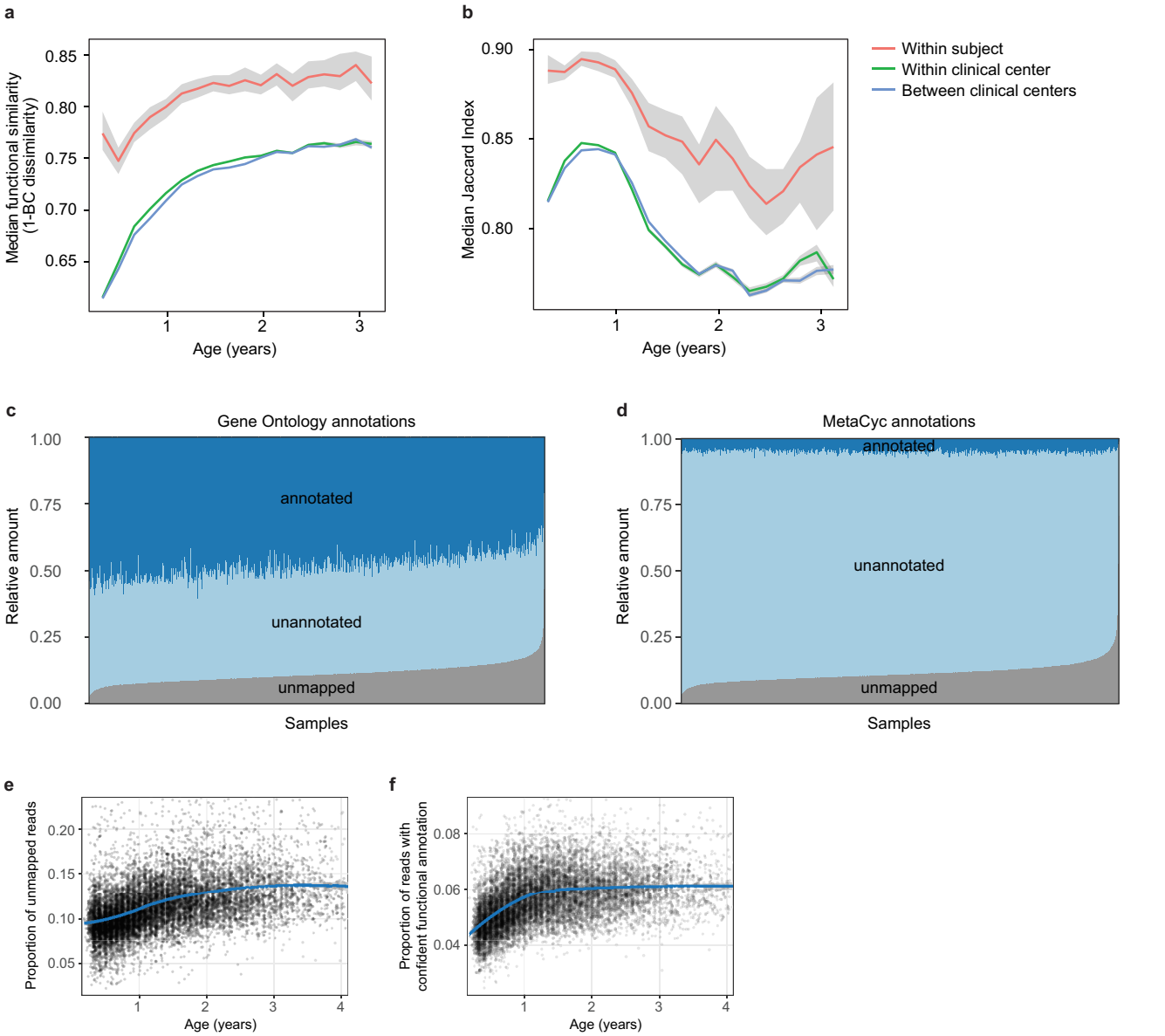

**Extended Data Fig. 2 | Stability and regional differences of taxonomic profiles. a**, Stability of the microbiota, measured by the Jaccard index ($n = 10{,}750$ samples) in three-month time windows, over two-month increments, stratified into three groups: within subject, within clinical centre, and across clinical centres. Lines show the median per time window. Shaded areas show the 99% confidence interval estimated using binomial distribution. Compare to Fig. 2b, which shows the same analysis measured by Bray–Curtis dissimilarity. **b**–**d**, Average longitudinal abundance of *Ruminococcus gnavus* (**b**), *Lactobacillus rhamnosus* (**c**) and *Veillonella parvula* (**d**) per clinical centre ($n = 10{,}194$ samples). The curves show LOESS fit for the relative abundances, as above.



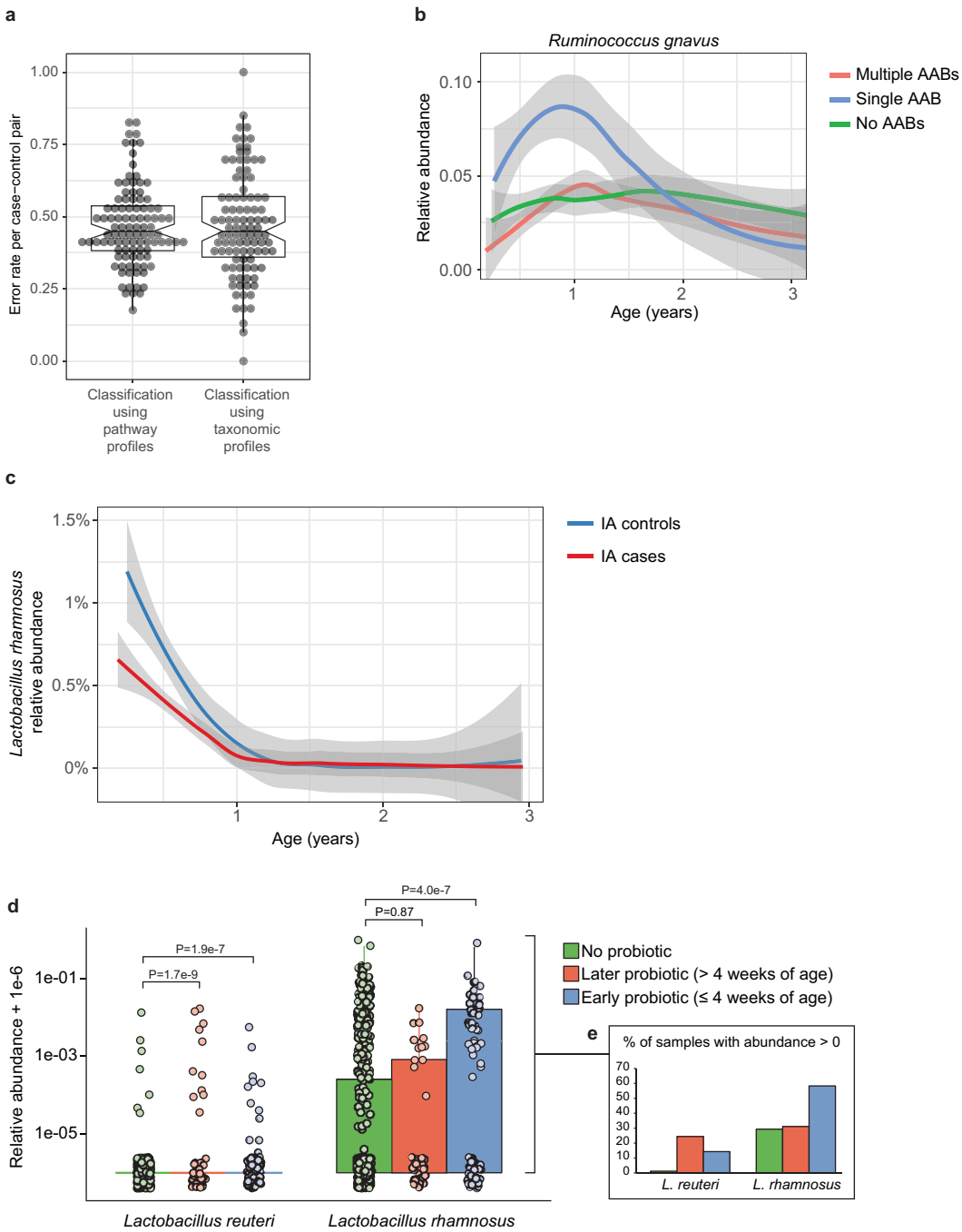

**Extended Data Fig. 3 | Accrual of microbial alpha diversity. a**, Shannon's diversity of the taxonomic profiles of the gut microbial communities ($n = 10{,}913$ samples) with respect to the age at the sample collection. The curve shows the generalized additive model (GAM) fit for the data, and the shaded area shows the 95% confidence interval for each fit, as implemented in geom_smooth function in ggplot2 R package. **b**, Shannon's diversity for the samples in the IA case–control cohort ($n = 7{,}051$) with respect to the time to the appearance of first autoantibody (seroconversion). The curves show LOESS fits for cases and controls separately, and the shaded area shows 95% confidence intervals for each fit. **c**, Shannon's diversity for the samples in the T1D case–control cohort ($n = 3{,}309$) with respect to the time to T1D diagnosis. The curves and shaded areas are as in **b**. **d**, As in **c**, but only for data ($n = 983$ samples) for subjects in Finland. No difference between cases and controls. **e**, As in **c**, but only for data ($n = 142$ samples, $n = 6$ subjects) for subjects in Georgia, USA. Cases show a drop in alpha diversity before the diagnosis of T1D (linear mixed-effects model, $P = 0.0033$).



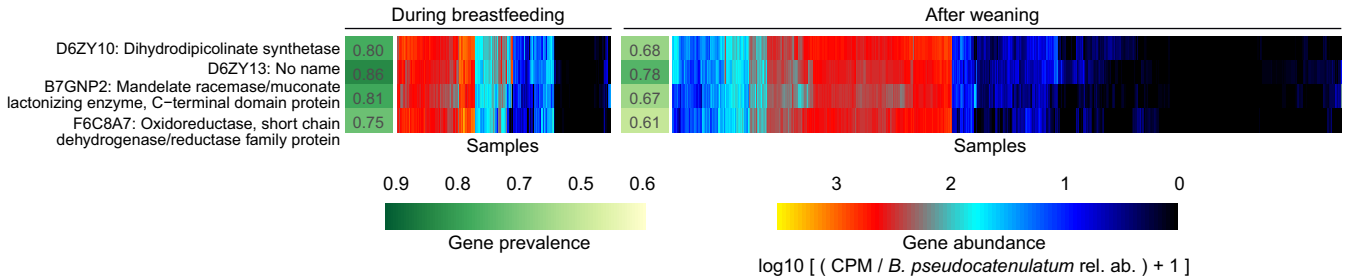

**Extended Data Fig. 4 | Effects of antibiotics. a**, Influence of antibiotic courses on microbial stability, stratified into six-month time windows (*x* axis). Stability was measured by Bray–Curtis dissimilarity over consecutive stool samples (<50 days apart) from the same individual between 3 and 29 months of age, and stratified by whether antibiotics were given between the two samples. For each notched box plot, the box denote the interquartile range (IQR), the horizontal line denotes the median, and the notch denotes the approximation for the 95% confidence interval (notch width = $1.58 \times \mathrm{IQR}/n^{0.5}$, in which $n$ is the number of samples per box plot). Compare to Fig. 2c. **b**, Influence of antibiotic courses on microbial diversity. Notched box plots denote the increase (difference) in diversity between two consecutive stool samples (<50 days apart) stratified by antibiotic administration between the samples. Data show no difference between the groups (antibiotics versus no antibiotics). **c**, Influence of antibiotics courses on microbial diversity by antibiotic type; data from **b** stratified into one-year time windows (*x* axis) and antibiotic types. No significant differences were detected between the antibiotic types. **d**, **e**, Influence of antibiotic courses on microbial stability by antibiotic type; data from Fig. 2c and Extended Data Fig. 3a stratified by antibiotic type. **d**, LOESS fit for the relative abundances (shaded area shows 95% confidence interval for each fit, as implemented in geom_smooth function in ggplot2 R package). **e**, Notched box plots (as in **a** and **b**) for the data per antibiotic type. No significant differences were detected between the antibiotic types. No antibiotics, $n = 7{,}130$; amoxicillin, $n = 268$; penicillin, $n = 90$; cephalosporin, $n = 51$; macrolide, $n = 60$; other, $n = 101$. **f**, Decreases in relative abundance of bacteria over antibiotic courses. Bacteria for which the bootstrapped 95% confidence interval of the fold change does not overlap zero are shown. Fold change was measured between consecutive samples with an antibiotic course between them, given that the species in question was present in the first of the two samples. Sample size per species ($n$) indicate the number of sample pairs in which the species in question was present in the sample preceding the antibiotic treatment. Bars denote bootstrapped mean $\log_2$(fold change) (that is, decrease), and error bars denote s.d. ($n = 1{,}000$ bootstrap samples).




**Extended Data Fig. 5 | Dynamics of species-specific microbial functional potential during early gut development. a**, **b**, Stability of microbial pathways ($n = 10{,}580$ samples) measured by Bray–Curtis dissimilarity (**a**) and the Jaccard index (**b**) and stratified into three groups: within subject, within clinical centre, and across clinical centres. Although the baseline level of functional similarity is significantly greater than that of taxa (see Fig. 2b), functional states and development trajectories also both retain a level of personalization. The stability of the functional profiles was evaluated in three-month time windows, over two-month increments. Lines show the median per time window, and shaded area denotes the 99% confidence interval estimated using binomial distribution. **c**, **d**, Proportion of metagenomic gene abundance with functional annotation through Gene Ontology (**c**) and MetaCyc (**d**) databases. The metagenomic reads were divided into the following categories: reads that could be mapped to genes with functional assignment in the database in question (annotated), and reads with no annotation but alignment to species pangenomes or UniProt proteins (unannotated). The proportion of the unknown genes (unmapped) was estimated using the number of reads with unknown origin. **e**, The proportion of unmapped reads, reflecting the relative abundances of reads not mappable to any microbial pangenomes in the available reference set or to UniProt. An increasing trend of unmapped reads with respect to the age at sample collection continued through approximately two years of age. **f**, The proportion of reads with confident functional annotation in MetaCyc within the genes that mapped to species pangenomes or UniProt proteins. The data again showed an increasing longitudinal trend, implicating a deficit of functional and biochemical annotations within microorganisms that are abundant during the first year of life.




**Extended Data Fig. 6 | Differences between cases and controls. a**, The gut microbiome functional (left) and taxonomic (right) profiles were classified between cases and controls using leave-one-out cross-validation ($n = 3{,}366$ samples), in which one case–control pair was held-out in turn. Data show error rates for classifying these held-out samples per fold (a data point per fold, $n = 100$ folds). This suggests weak but better-than-random classification between cases and controls. Notched box plots are as in Extended Data Fig. 4. **b**, Average longitudinal abundance of *Ruminococcus gnavus* in Finland ($n = 2{,}630$ samples) stratified by the number of observed persistent autoantibodies (AABs); no autoantibodies (that is, healthy control), a single autoantibody, or multiple (two or more) autoantibodies. **c**, Average longitudinal abundance of *Lactobacillus rhamnosus* in IA cases and controls ($n = 7{,}017$ samples). *L. rhamnosus* is more abundant in controls ($q = 0.055$). The curves in **b** and **c** show LOESS fit per group, and shaded areas show 95% confidence interval for each fit, as implemented in geom_smooth function in ggplot2 R package. **d**, Abundance (left) and prevalence (right) of *Lactobacillus reuteri* and *L. rhamnosus* in the first stool sample of each individual (collected at approximately three months of age) in association with early probiotic supplementation. 'No probiotic' indicates no probiotics given before the first stool sample ($n = 583$); 'later probiotic' refers to probiotics given later than the first four weeks but before the first stool sample ($n = 45$); 'early probiotic' refers to probiotics given during the first four weeks of life ($n = 84$). Numbers ($n$) per clinical centre are given in Extended Data Table 2. *L. reuteri* and *L. rhamnosus* were more abundant and prevalent in groups with probiotics supplementation. Visual jitter was added to make data equal to zero distinguishable, and boxes denote the IQR, when applicable. The shown $P$ values were obtained by applying Fisher's exact test (two-sided) to presence or absence count data (counting samples in which the species were present).




**Extended Data Fig. 7 | Contrasting HMO utilization genes in *B. pseudocatenulatum*.** The gene families involved in HMO utilization and that show contrasting presence in *B. pseudocatenulatum* genomes during breastfeeding ($n = 321$ samples) compared to after weaning ($n = 1{,}004$ samples). Columns represent stool samples in which the relative abundance of *B. pseudocatenulatum* species was greater than 10% ($n = 1{,}325$ samples). Rows and columns were ordered by hierarchical clustering using complete linkage method. Compare to Fig. 4, which shows similar data for *B. longum*. UniRef90 identifiers and gene names or families are indicated on the left.

**Extended Data Table 1 | Summary of TEDDY microbiome cohort**

| | US, Colorado | US, Georgia | US, Washington | Finland | Germany | Sweden |
|---|---|---|---|---|---|---|
| T1D cases (samples) | 14 (274) | 3 (89) | 8 (111) | 34 (553) | 13 (246) | 29 (532) |
| IA cases (samples) | 39 (689) | 17 (252) | 25 (368) | 70 (900) | 21 (292) | 95 (1,542) |
| Healthy controls (samples) | 61 (906) | 22 (250) | 36 (399) | 119 (1,273) | 40 (512) | 137 (1,725) |
| **Sex** | | | | | | |
| Male / Female | 61 / 53 | 19 / 23 | 51 / 18 | 117 / 106 | 30 / 44 | 152 / 109 |
| **Ethnic background** | | | | | | |
| White, non-hispanic | 86 (75.4%) | 41 (97.6%) | 56 (81.2%) | N/A | N/A | N/A |
| **Mode of birth** | | | | | | |
| Caesarean section | 41 (36.0%) | 22 (52.4%) | 25 (36.2%) | 42 (18.8%) | 23 (31.1%) | 46 (17.6%) |
| **Probiotic supplementation** | | | | | | |
| Probiotics during first 4 weeks | 0 | 2 (4.8%) | 0 | 67 (30.0%) | 7 (9.5%) | 14 (5.4%) |
| Probiotics during follow-up | 22 (19.3%) | 13 (31.0%) | 9 (13.0%) | 162 (72.6%) | 33 (44.6%) | 58 (22.2%) |
| **Breastfeeding** | | | | | | |
| Median duration (days) | 268 | 301 | 335 | 289 | 278 | 228 |
| duration, 25 percentile | 56 | 145 | 171 | 152 | 140 | 98 |
| duration, 75 percentile | 396 | 365 | 440 | 385 | 367 | 304 |
| Number of subjects never breastfed | 3 | 3 | 1 | 0 | 0 | 0 |
| **Maternal characteristics** | | | | | | |
| Maternal T1D | 7 (6.1%) | 0 | 3 (4.3%) | 14 (6.3%) | 18 (24.3%) | 7 (2.7%) |
| Maternal T2D | 2 (1.8%) | 0 | 0 | 0 | 0 | 0 |
| Gestational diabetes | 5 (4.4%) | 5 (11.9%) | 5 (7.2%) | 32 (14.3%) | 3 (4.1%) | 6 (2.3%) |
| Antibiotics during pregnancy | 21 (18.4%) | 10 (23.8%) | 5 (7.2%) | 40 (17.9%) | 13 (17.6%) | 29 (11.1%) |
| Metformin during pregnancy | 1 (0.9%) | 0 | 0 | 1 (0.4%) | 0 | 0 |
| Glyburide during pregnancy | 2 (1.8%) | 2 (4.8%) | 2 (2.9%) | 0 | 0 | 0 |
| Antihypertensives during pregnancy | 4 (3.5%) | 3 (7.1%) | 4 (5.8%) | 5 (2.2%) | 3 (4.1%) | 0 |
| Insulin during pregnancy | 9 (7.9%) | 0 | 3 (4.3%) | 23 (10.3%) | 19 (25.7%) | 8 (3.1%) |

Data on subjects' ethnic background were not systematically collected in European clinical centres but these study populations were predominantly white, non-Hispanic. Reported antihypertensive drugs were atenolol ($n=2$), bisoprolol ($n=1$), labetalol ($n=6$), methyldopa ($n=1$), methyldopa plus methyldopate ($n=3$), metoprolol ($n=4$) and nifedipine ($n=5$). No use of angiotensin-converting enzyme (ACE) inhibitors was reported. Numbers indicate the number of subjects ($n$) if not specified otherwise.

**Extended Data Table 2 | Antibiotics and probiotics**

| | US, Colorado | US, Georgia | US, Washington | Finland | Germany | Sweden |
|---|---|---|---|---|---|---|
| Subjects with abx prescriptions | 93 (81.6%) | 37 (88.1%) | 54 (78.3%) | 206 (92.4%) | 56 (75.7%) | 192 (73.6%) |
| Median number of abx per subject (25th and 75th percentile) | 2 (1-6) | 5 (2-9) | 2 (1-4) | 6 (3-11) | 2 (0-5) | 2 (0-4) |
| **Number of abx by type (prescriptions per subject)** | | | | | | |
| Amoxicillin | 242 (2.12) | 147 (3.50) | 104 (1.51) | 769 (3.45) | 45 (0.61) | 134 (0.51) |
| Cephalosporins | 87 (0.76) | 65 (1.55) | 31 (0.45) | 127 (0.57) | 51 (0.69) | 23 (0.09) |
| Macrolide | 54 (0.47) | 35 (0.83) | 47 (0.68) | 203 (0.91) | 33 (0.45) | 23 (0.09) |
| Penicillin | 6 (0.05) | 2 (0.05) | 3 (0.04) | 17 (0.08) | 13 (0.18) | 412 (1.58) |
| Other | 76 (0.67) | 80 (1.90) | 33 (0.48) | 521 (2.34) | 77 (1.04) | 154 (0.59) |
| Total | 465 (4.08) | 329 (7.83) | 218 (3.16) | 1,637 (7.34) | 219 (2.96) | 746 (2.86) |
| **Probiotic use in early life** | | | | | | |
| Early probiotic | 0 (0.0%) | 1 (2.9%) | 0 (0.0%) | 63 (30.7%) | 7 (10.0%) | 13 (5.6%) |
| Later probiotic | 1 (0.9%) | 1 (2.9%) | 2 (3.3%) | 16 (7.8%) | 8 (11.4%) | 17 (7.3%) |
| No probiotic | 109 (99.1%) | 32 (94.1%) | 59 (96.7%) | 126 (61.5%) | 55 (78.6%) | 202 (87.1%) |

Top, 3,678 antibiotic prescriptions in the TEDDY microbiome study population by clinical centre. Bottom, early probiotic supplementation in TEDDY clinical centres. Probiotic use was stratified into three categories: probiotics during the first 4 weeks of life (early probiotic); probiotics before the first stool sample (roughly at three months) but not the first 4 weeks (later probiotic); and no probiotics before the first stool sample (no probiotic). Data for probiotics are presented as *n* (percentage).

Corresponding author(s): Tommi Vatanen, Curtis Huttenhower, Ramnik J. Xavier

# Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

## Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

| n/a | Confirmed | |
|---|---|---|
| ☐ | ☒ | The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☐ | ☒ | An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☐ | ☒ | The statistical test(s) used AND whether they are one- or two-sided<br>*Only common tests should be described solely by name; describe more complex techniques in the Methods section.* |
| ☐ | ☒ | A description of all covariates tested |
| ☐ | ☒ | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ | A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☐ | ☒ | For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted<br>*Give P values as exact values whenever suitable.* |
| ☒ | ☐ | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☐ | ☒ | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☐ | ☒ | Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated |
| ☐ | ☒ | Clearly defined error bars<br>*State explicitly what error bars represent (e.g. SD, SE, CI)* |

*Our web collection on statistics for biologists may be useful.*

## Software and code

Policy information about availability of computer code

| | |
|---|---|
| Data collection | No software was used for data collection. |
| Data analysis | Bowtie2 v2.2.3, Casava v1.8.2 (Illumina), cutadapt v1.9dev2, Trim Galore v0.2.8 (Babraham Bioinformatics), PRINSEQ v0.20.3, MetaPhlAn2 v2.6.0, HUMAnN2 v0.9.4, R v3.1.1, Python v2.7.1. Additional details are given in Methods. |

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

## Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

TEDDY Microbiome 16S and WGS data that support the findings of this study are available in NCBI's database of Genotypes and Phenotypes (dbGaP) with the

primary accession code phs001443.v1, in accordance with the dbGaP controlled-access authorization process. Clinical metadata analyzed during the current study are available in the NIDDK Central Repository at https://www.niddkrepository.org/studies/teddy.

# Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

# Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

| | |
|---|---|
| Sample size | All stool samples with metagenomic data (N = 10,903 stool samples) from subjects in TEDDY islet autoimmunity and type 1 diabetes case-control cohorts are being analyzed for maximal power. This includes all subjects with islet autoimmunity or type 1 diabetes and their matched controls in TEDDY study as of May 31, 2012 (N =783 subjects). |
| Data exclusions | In T1D and IA case-control comparisons, all case-control pairs where the control later progressed to case status were removed (i.e. they progressed to IA or T1D). |
| Replication | For bacterial growth assays, the experiment was reproduced 3x with technical replicates in triplicate. |
| Randomization | Randomization was not used. |
| Blinding | No blinding was used, TEDDY is an observational follow-up study. |

# Reporting for specific materials, systems and methods

## Materials & experimental systems

| n/a | Involved in the study |
|---|---|
| ☐ | ☒ Unique biological materials |
| ☒ | ☐ Antibodies |
| ☒ | ☐ Eukaryotic cell lines |
| ☒ | ☐ Palaeontology |
| ☒ | ☐ Animals and other organisms |
| ☐ | ☒ Human research participants |

## Methods

| n/a | Involved in the study |
|---|---|
| ☒ | ☐ ChIP-seq |
| ☒ | ☐ Flow cytometry |
| ☒ | ☐ MRI-based neuroimaging |

## Unique biological materials

Policy information about availability of materials

| | |
|---|---|
| Obtaining unique materials | The bacterial strains isolated for and used in this study will be provided to anyone who requests them. |

## Human research participants

Policy information about studies involving human research participants

| | |
|---|---|
| Population characteristics | This study includes metagenomic profiles of stool samples from children collected monthly starting at three months of age to up to five years of age. The study population are 783 mostly white, non-hispanic children from six different clinical centers in the U.S. (Colorado, Georgia/Florida, and Washington) and Europe (Finland, Germany, and Sweden). The whole study population had a genetic predisposition for T1D or first-degree relative(s) with T1D. |
| Recruitment | Families with a newborn in participating clinical centers with HLA-conferred genetic predisposition for T1D or first-degree relative(s) with T1D were invited to join the study. |