

Gut microbiome maturation in early childhood interacts with host genetics to predict type 1 diabetes risk

Received: 24 January 2026

Accepted: 19 August 2026

Published online: 21 September 2026

 Check for updates

Danyue Dong^{1,2}, Aaron M. Walsh^{2,3}, Tommi Vatanen^{1,2,4,5,6,7}, George Weingart³, Mondher Khdhiri^{2,3}, Meir J. Stampfer^{1,8,9}, Kendra Vehik¹⁰, Eric A. Franzosa^{1,2,3}✉, Curtis Huttenhower^{1,2,3}✉, Dong D. Wang^{1,2,5}✉ & The TEDDY Study Group*

Prospective evidence linking early-life microbiome development and host genetics with type 1 diabetes (T1D) risk is limited. Here, we describe how gut microbiome maturation and host genetics influence T1D risk in the TEDDY study. We analysed 12,151 longitudinal metagenomes and host genetic data from 887 children at high genetic risk for T1D followed for up to 6 years. We identify three microbiome maturational patterns: Early Matured, Late Matured and Early Plateaued, driven primarily by non-linear changes in species from the *Bifidobacterium* and *Ruminococcus* genera. The Early Matured pattern is enriched in galactose metabolism and exhibits higher production of aromatic amino acids and B-group vitamins at early follow-ups, whereas the Early Plateaued pattern has increased microbial production of branched-chain amino acids. Notably, the Early Plateaued pattern is associated with a threefold elevated risk of T1D, whereas other patterns are not associated with T1D risk. Furthermore, we find that host genetic variants related to antimicrobial and antiviral immune responses modify the association between the Late Matured pattern and T1D risk. These findings highlight the role of early microbial exposures and host genetics in T1D susceptibility.

Type 1 diabetes (T1D) is a disorder characterized by the autoimmune-mediated destruction of pancreatic β -cells, resulting in insulin deficiency and subsequent hyperglycaemia¹. Globally, T1D affects approximately 8.5 million individuals, of which, 1.5 million are below the age of 20 years². Despite advancements in the understanding of T1D pathophysiology and its treatment and management, T1D remains incurable, necessitating lifelong insulin therapy for affected individuals^{1,3}. Notably, T1D can be driven both by complex genetic factors⁴ and by a

combination of environmental exposures, particularly in early life⁵, making it crucial to understand the interplay of genetics and development during this vulnerable period.

The gut microbiome is one of the most pervasive exposures throughout early life, playing a crucial role in both shaping and being shaped by immune, dietary and physiological development⁶. Indirect evidence from recent studies suggests that early-life factors, such as breastfeeding and delivery mode, may influence T1D risk later in

¹Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA.

²Broad Institute of MIT and Harvard, Cambridge, MA, USA. ³Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA.

⁴Institute of Biotechnology, Helsinki Institute of Life Science, University of Helsinki, Helsinki, Finland. ⁵Department of Microbiology, Faculty of Agriculture and Forestry, University of Helsinki, Helsinki, Finland. ⁶Research Program for Clinical and Molecular Metabolism, Faculty of Medicine, University of Helsinki, Helsinki, Finland. ⁷Liggins Institute, University of Auckland, Auckland, New Zealand. ⁸Department of Nutrition, Harvard T.H. Chan School of Public Health, Boston, MA, USA. ⁹Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA. ¹⁰Health Informatics Institute, Morsani College of Medicine, University of South Florida, Tampa, FL, USA. *A list of authors and their affiliations appears at the end of the paper.

✉e-mail: franzosa@hsph.harvard.edu; chuttenh@hsph.harvard.edu; dow471@mail.harvard.edu

life, potentially through their impact on microbial colonization^{5,7}. Several studies involving children at high risk for T1D have identified cross-sectional associations in which those who developed the disease exhibited disruptions in the gut microbiome^{7–10}. However, these were of generally modest effect and predictive power, impeding their utility either as a risk biomarker or as a means of better understanding related mechanisms in T1D.

In contrast with the microbial factors identified so far, genetic factors, including *HLA-DRB1*, *HLA-DQA1* and *HLA-DQB1* alleles, are among the strongest predictors of T1D risk, with other genes such as *INS*, *PTPN22*, *CTLA4* and *IL2RA* contributing to susceptibility^{4,11}. Overall, about 50% of T1D risk is thought to be heritable¹². Conversely, although genome-wide association studies in adults have identified limited associations between host genetics and the microbiome, such as the *LCT* gene's association with *Bifidobacterium* species¹³ and the *FUT2*/*FUT1* locus's link to *Ruminococcus torques* abundance¹⁴, these findings suggest a modest genetic influence on the gut microbiome in later life. However, this does not preclude the possibility that host genetics may exert a more pronounced influence on the gut microbiome during a critical period of early childhood development. Yet, this hypothesis has remained untested in human cohorts.

To elucidate the interrelationship among host genetics, early-life microbiome development and the risk of T1D, we conducted a prospective analysis of 12,151 metagenomes (1,238 not previously published) and host genetic data from 887 children in Finland, Germany, Sweden and the USA during a follow-up of up to 6 years within the TEDDY (The Environmental Determinants of Diabetes in the Young) study¹⁵. We discovered a natural grouping of individuals' microbiome development trajectories into three distinct patterns, which we termed 'Early Matured', 'Late Matured' and 'Early Plateaued'. Of note, our analyses revealed a significant threefold higher T1D risk in children with the Early Plateaued pattern, which is only detectable when considering the microbiome's developmental trajectory over time rather than cross-sectionally. Additionally, we discovered that components of the host genetic background modified the association between the Late Matured pattern and T1D risk. Further, we assessed the direct impact of host genetics on gut microbiome development, finding significant, albeit limited, associations of the overall genetic structure and specific genetic variants with the development of certain microbial species. Our study, employing a prospective design, represents the first research effort to demonstrate that the development of the early-life gut microbiome, interacting with the host genetic background, strongly predicts the incidence of T1D in a human cohort.

Results

Integration of longitudinal gut microbiome profiles and host genetics to study T1D aetiology in a prospective cohort

To investigate the link between the gut microbiome, host genetics and the development of T1D, we assembled genetic data and longitudinally collected information on microbial profiles, serum islet autoantibodies (IAs) and T1D diagnosis in 887 children from the TEDDY cohort. Participants were enrolled from six clinical centres in Finland, Sweden, Germany and three US states: Washington, Georgia/Florida and Colorado (Methods and Fig. 1a). Stool samples were collected approximately monthly through 48 months of age and quarterly thereafter, per the TEDDY protocol. Because TEDDY is a dynamic prospective cohort study with right-censored follow-up at the analytic cutoff, per-participant follow-up duration is variable. A subset of these stool samples was selected for shotgun metagenomic sequencing, yielding 12,151 metagenomes from 887 participants (mean 13.7 ± 8.7 sequenced samples per participant over follow-up) (Fig. 1b and Supplementary Fig. 1). Using bioBakery 2.0 workflows (to match previously published data), we generated taxonomic and functional profiles from these metagenomes, extending previously published shotgun metagenomics^{8,10} by 1,238 samples. After quality control, these profiles comprised 92 species,

1,318 level-4 Enzyme Commission (EC) categories and 348 biochemical pathways (Methods and Extended Data Fig. 1). Participants were genotyped using Illumina ImmunoChip single-nucleotide polymorphism (SNP) array¹⁶, covering around 200,000 polymorphisms relevant to autoimmune and inflammatory diseases (Fig. 1c). The study's primary end point, as predefined by the main TEDDY cohort¹⁷, was a composite outcome of IA seroconversion, defined as a positive IA test in at least two consecutive blood samples or a clinical T1D diagnosis. Additionally, the TEDDY cohort collected extensive data on potential confounders and major T1D risk factors¹⁵, including breastfeeding status, birth mode, probiotic use, antibiotic use and dietary intake (Methods, Fig. 1a and Supplementary Table 1).

Host genetics minimally influence the overall configuration of the early-life gut microbiome

Using these data, we first investigated the influence of host genetic and phenotypic factors on overall microbiome structure. We used principal-component analysis (PCA) to characterize genetic population structure and principal-coordinates analysis (PCoA) based on species-level Bray–Curtis dissimilarity to summarize gut microbiome structure. As expected, we observed a strong correlation between the first genetic principal component (PC1) and participants' geographical location ($\chi^2 = 589.16$, d.f. = 3, $P < 2.2 \times 10^{-16}$; 72.8% of the variance in PC1 explained by country), which particularly aligns with the well-documented genetic bottleneck in the Finnish population¹⁸ (Fig. 1d). No strong correlations were found between geographical location and the remaining top principal components (PCs; PC2 to PC10), so these were used as summary statistics for genetic predisposition to autoimmune and inflammatory diseases in subsequent analyses. As anticipated for infants in this age range¹⁹, the majority of variation in the gut microbial community was driven by a trade-off between the *Actinobacteria* and *Firmicutes* phyla (Extended Data Fig. 2a). Similarly, we used principal coordinate (PCo) scores to summarize the variation in gut microbial communities.

Based on previous studies on host genetics and the gut microbiome²⁰, we hypothesized that host genetics would have a smaller impact on microbial community structure compared with the impact exerted by host phenotypic factors. Consistent with this hypothesis, genetic PC1 showed only a weak association with microbiome PCo1 at baseline (Spearman's $\rho = 0.12$, $P = 7.6 \times 10^{-4}$; Fig. 1e and Supplementary Fig. 2), and no significant associations were observed at later time points. Similarly, Mantel and Procrustes analyses comparing host genetic and microbiome distance matrices demonstrated weak and non-significant concordance across all age groups (Supplementary Table 2). This was further supported by the non-significant association between genetic PC1 and overall microbial community variation (permutational multivariate analysis of variance (PERMANOVA), $R^2 = 0.28\%$, $P = 0.99$; 999 permutations). Although PERMANOVA suggested significant associations between genetic PC4 and PC5 with gut microbiome structure (both $P = 0.001$), the explained variance was minimal ($R^2 \leq 0.1\%$). Among all phenotypic variables, age accounted for the largest percentage of variance ($R^2 = 5.90\%$), followed by breastfeeding ($R^2 = 1.48\%$), introduction of solid food ($R^2 = 0.32\%$), and delivery mode ($R^2 = 0.20\%$) (Fig. 1f, Extended Data Fig. 2b and Supplementary Table 3). We, therefore, included these phenotypic variables as confounders in our subsequent analyses.

Gut microbiome maturational patterns predict subsequent risk of type 1 diabetes

The gut microbiome in infancy is highly dynamic, particularly during the first year, undergoing substantial changes ecologically and through interactions with the developing immune system²¹. Previous studies have demonstrated that this early microbiome is crucial for educating the immune system and promoting immune tolerance⁶. Correspondingly, imbalance or delayed development of the gut microbiome has

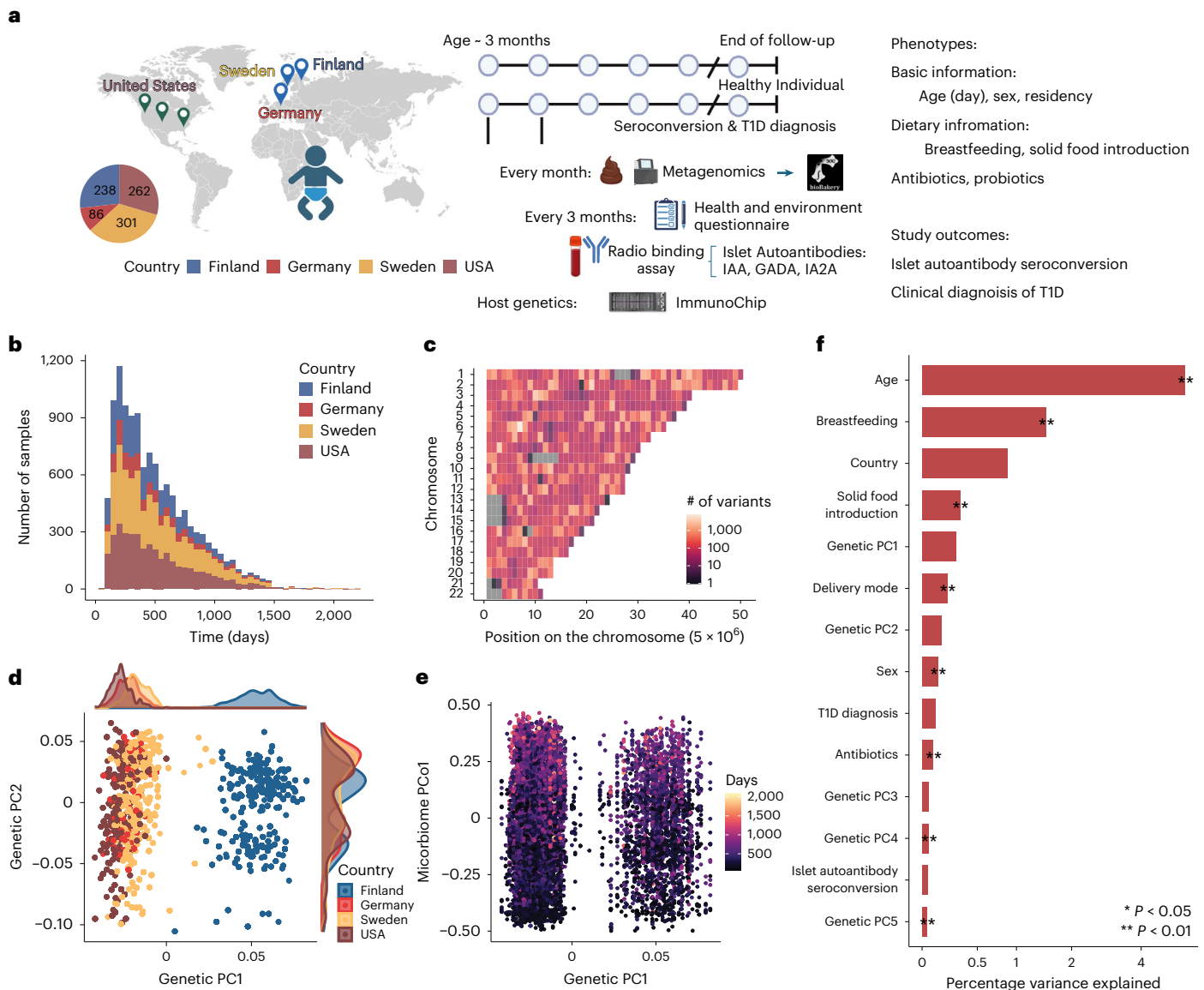


Fig. 1 | A longitudinal study examining the interrelationship among host genetics, early-life development of the gut microbiome, and type 1 diabetes onset. We analysed 12,151 metagenomes, longitudinally collected stool samples spanning early childhood, with dense sampling through the first ~4 years of life and sparser sampling extending up to ~6 years in a subset of participants, paired with host genetic data, and prospectively ascertained IA seroconversion and clinical end point of T1D during the first 6 years of life from 887 children in a longitudinal cohort study^{8,10}. **a**, Participants were recruited from six clinical centres spanning four countries: Finland, Sweden, Germany and the USA (Washington, Georgia and Colorado). The primary end point of this study was the IA seroconversion in at least two consecutive blood samples or the clinical diagnosis of T1D. Additionally, we incorporated longitudinally collected metadata, including breastfeeding status, birth mode, probiotic use, antibiotic use and dietary intake. **b**, The distribution of metagenomes remains consistent across different countries throughout the follow-up period, with dense sampling during the first ~4 years of life and sparser sampling extending up to ~6 years in a subset of participants. Each bar in the stacked bar plot represents the country-specific distribution of metagenomes within a 50-day interval. **c**, Participants underwent genotyping using the Illumina ImmunoChip SNP array¹⁶, which

was designed to detect genetic variations in the human immune system. The plot displays autosomal chromosomes only (chromosomes 1–22), grey indicates bins containing no genotyped variants. **d**, The genetic relationship matrix visualization for the four countries from which our study population was recruited displays a distinct separation of Finnish participants along the PC1 score. This separation indicates a unique genetic architecture within the Finnish population, as expected⁶⁹. **e**, Host genetic architecture, as summarized by PC1, and gut microbial community structure, as summarized by the first principal coordinate (PCo1) based on species-level Bray–Curtis dissimilarity. Points are coloured by age, illustrating the age-related changes in microbiome community structure during early life. **f**, Age predominates over host genetics and other covariables in shaping the early gut microbiome. Variation in gut microbial composition attributable to basic demographic information, dietary factors, antibiotic usage and T1D-associated autoantibodies was assessed using PERMANOVA based on species-level Bray–Curtis dissimilarity. P values were derived from the upper tail of the permutation distribution of the pseudo- F statistic (999 permutations), for which a one- versus two-sided distinction does not apply. Some elements of panel **a** were created in BioRender; Dong, D. <https://BioRender.com/80mnepi> (2026).

been linked to immune-linked conditions such as asthma²². We therefore sought to identify distinct microbiome maturational patterns, hypothesizing that different patterns might subsequently be associated prospectively with varying T1D risk.

To identify different microbiome maturational patterns, we used a trajectory-clustering algorithm²³ based on the Bray–Curtis distance to measure changes in gut microbial composition from baseline over the first 800 days in a subpopulation with four or more metagenomic

samples (Methods and Supplementary Table 4). Here, we define ‘microbiome maturation’ as the longitudinal change in each child’s gut microbial community relative to their own baseline state. Our algorithm identified three major patterns (Fig. 2a and Supplementary Figs. 3–5) that summarize variation across a continuous spectrum of microbiome development. These patterns were largely independent of baseline demographic and clinical characteristics (Supplementary Table 5). The first pattern, named ‘Early Matured’, was characterized by rapid early divergence from baseline within the first 400 days and remained relatively stable thereafter (Extended Data Fig. 3). The second pattern, termed ‘Late Matured’, exhibited slower initial divergence from baseline and lower microbial diversity during the first 400 days, but subsequently showed accelerated divergence and converged towards the Early Matured trajectory. The third pattern, called ‘Early Plateaued’, maintained low diversity throughout the first 800 days and exhibited limited divergence from baseline, reaching an earlier plateau than the other trajectories. We applied the same trajectory-clustering framework to the independent DIABIMMUNE three-country cohort²⁴, and observed three qualitatively similar microbiome maturation trajectories (Supplementary Fig. 6).

Next, we determined the prospective associations between microbiome maturational patterns and the subsequent risk of T1D using conditional logistic models. The maturational patterns were included in the models as exposures, and the composite end point of IA seroconversion or T1D was treated as the outcome. The models were simultaneously adjusted for multiple covariates, including risk sets, family history of T1D, clinical centre, sex, delivery mode, antibiotic use, probiotic use, solid food introduction, breastfeeding cessation and the top five genetic PCs (Methods). Compared with children with the Early Matured pattern, those with the Early Plateaued pattern had more than three times the risk of developing T1D (risk ratio (RR) = 3.21; 95% confidence interval (CI) 2.16–4.76; $P = 6.69 \times 10^{-9}$) (Fig. 2b and Supplementary Table 6). A similarly elevated risk was observed when comparing the Early Plateaued and Late Matured patterns (RR = 3.33; 95% CI 2.32–4.78; $P = 8.01 \times 10^{-11}$), after adjusting for covariates. No significant difference in T1D risk was observed between the Late Matured and Early Matured patterns (RR = 0.96; 95% CI 0.72–1.28; $P = 0.80$). These associations remained consistent in sex-stratified analyses (Supplementary Table 7). On an absolute scale, the corresponding event rates were approximately 0.86, 0.95 and 1.52 cases per 1,000 person-months for the Early Matured, Late Matured and Early Plateaued patterns, respectively (Methods). We also compared the variance explained across nested models, in which, microbiome maturation patterns substantially increased the explained variation in T1D risk ($\Delta R^2 = 0.06$), whereas genetic background contributed only modestly ($\Delta R^2 = 0.007$). Secondary analyses, which examined IA

seroconversion and the clinical end point of T1D separately, yielded similar results (Supplementary Table 6).

Given the observed varying risk of T1D, which potentially indicates different pathologies associated with the maturational patterns, we then quantified the microbial compositional differences across these patterns by performing differential abundance analyses within seven approximately 100-day intervals. Using linear mixed models with MaAsLin 2 (ref. 25) (Methods), we identified microbial species that significantly differed across the three maturational patterns (Extended Data Fig. 4 and Supplementary Table 8). To maximize statistical power, we combined data from the Late Matured and Early Plateaued patterns due to their similar development during the initial 400 days and compared this combined group to the Early Matured pattern. Similarly, during the subsequent 400 days, we compared the Early Plateaued pattern to a combined group of Late Matured and Early Matured patterns (Fig. 2c and Supplementary Table 9). We identified several species that differed significantly across maturational patterns within at least one time interval.

Notably, participants with the Early Matured pattern had fewer *Bifidobacterium* spp. and more *Ruminococcus* during the first 400 days (Fig. 2c). *Bifidobacterium* species are among the first to colonize the infant’s gut and play crucial roles in gut health²⁶; their numbers then decrease over time in parallel with the maturation of the immune system and dietary changes²⁷. *Ruminococcus* species are known for their involvement in adult nutrient metabolism, particularly fermenting dietary fibres and complex carbohydrates²⁸. These findings suggest that the Early Matured pattern represents a more mature microbiome functional capacity to digest essential nutrients and interact with external environmental stimuli at an earlier stage. In the subsequent 400 days, individuals with Early Plateaued pattern showed significantly higher abundance of *Dorea longicatena* and lower abundance of *Clostridium hathewayi* (Fig. 2c). *D. longicatena* has been associated with higher rates of metabolic and inflammatory diseases and impaired gut permeability²⁹. Collectively, these findings indicate that variations in the temporal dynamics of individual species’ abundances contribute to the compositional differences observed across distinct maturational patterns.

Differences in essential amino acid biosynthesis and the metabolism of B vitamins and lactose explain functional variation across microbiome maturational patterns

We next examined whether these maturational patterns showed differences in their community-level microbial functions, including microbial enzymes and biochemical pathways. Using the same modelling approach in MaAsLin 2 described above, we compared the temporal changes in microbial enzymes and biochemical pathways across

Fig. 2 | Gut microbiome maturational patterns predict subsequent risk of type 1 diabetes. **a**, A data-driven approach revealed three maturational patterns of the gut microbiome in early life: Early Matured ($n = 172$), Late Matured ($n = 311$) and Early Plateaued ($n = 111$). The patterns were identified using a trajectory-clustering algorithm that employs the Bray–Curtis distance to quantify changes in the gut microbial composition from baseline across subsequent visits during the initial 800-day period. Trajectories were smoothed and visualized using the LOESS algorithm. Lines represent LOESS-smoothed fitted values, and shaded areas indicate 95% confidence intervals. **b**, Compared with other maturational patterns, the Early Plateaued pattern predicts a significantly higher risk of T1D. RRs were estimated from conditional logistic models with a composite end point of IA seroconversion or diagnosis of T1D as the outcome and the maturational patterns as the exposures, adjusting for risk sets, family history of T1D, clinical centre, sex, delivery mode, antibiotic use, probiotic use, solid food introduction, breastfeeding cessation and the top five genetic PCs. The lines represent adjusted cumulative incidence estimates extrapolated to the full TEDDY cohort over follow-up; shaded bands indicate 95% CIs. Adjusted estimates were derived from the multivariable models incorporating sampling weights to account for

risk-set sampling in the nested case–control design. **c**, Variations in temporal changes in the abundance of individual species account for compositional differences across different maturational patterns. We employed linear mixed models in MaAsLin 2 (ref. 25), incorporating individual IDs as random effects and clinical centre, sex, delivery mode, antibiotic use, solid food introduction and breastfeeding status as fixed effects, to identify microbial species with differential abundance across maturational patterns within each 100-day interval. The bar plots show comparisons between the Early Matured pattern and a combined group of Late Matured and Early Plateaued patterns during the initial 400 days of follow-up, as well as comparisons between the Early Plateaued pattern and a combined group of Late Matured and Early Matured patterns during the subsequent 400 days of follow-up. RA, relative abundance. The false discovery rate (FDR)-adjusted P values (q) were calculated using the Benjamini–Hochberg method with a target rate of 0.25. * $q < 0.25$, ** $q < 0.01$, *** $q < 0.001$. As descriptive illustrations, we further highlight four selected species as insets to illustrate distinct temporal changes across maturational patterns, characterized by either early high or low abundance in the Early Matured pattern and later high or low abundance in the Early Plateaued pattern.



different maturational patterns (Supplementary Tables 10 and 11 and Extended Data Figs. 5 and 6).

Metagenomic potential for microbial biosynthesis of host-essential amino acids was significantly increased in the Early Plateaued pattern. Specifically, aromatic amino acid (AAA) pathways were more abundant in the later stages, whereas branched-chain amino acids (BCAAs) were abundant throughout the 800 days (Fig. 3a and Supplementary Fig. 7). This trend was consistent at the enzyme level (Supplementary Table 12), exemplified by the elevated presence of a multifunctional enzyme necessary for the final step in BCAA biosynthesis (EC 2.6.1.42: branched-chain amino acid transaminase; Fig. 3b). BCAAs and AAAs are crucial for protein synthesis and energy metabolism in early life^{30,31}. Moreover, among the major microbial metabolites derived from AAAs, aromatic lactic acids, produced predominantly by *Bifidobacterium* spp., play a critical role in shaping the immune system during this developmental stage³². Similarly, we found a higher functional capacity for B vitamin metabolism in the Early Plateaued pattern during later follow-ups (Fig. 3a and Supplementary Fig. 7). The trajectory of amino acid and B vitamin metabolism pathways in the Early Plateaued group likely reflects a pattern in which microbiome maturation is impaired, characterized by the persistent dominance of early-colonizing taxa with high biosynthetic potential but limited taxonomic diversity.

In the Early Matured pattern, there was also a higher abundance of biochemical pathways involved in galactose degradation during early follow-ups compared with the other two patterns. This pathway's abundance was significantly higher in the Early Plateaued pattern during later follow-ups (Fig. 3a). As our analysis adjusted for breastfeeding and the introduction of solid food, this finding cannot be fully attributed to dietary intake. The early higher abundance of galactose degradation capacity may indicate that the Early Matured pattern's microbiome is better equipped to digest nutrients in human milk³¹. In contrast, during the late stage, especially with the introduction of solid food, the Early Plateaued pattern remains oriented towards digesting human milk, potentially leading to suboptimal nutrition.

As expected, we observed an increasing trend in microbial species diversity that has contributed to these functions over time. In the Early Matured pattern, the microbial enzymes were encoded by a more diverse set of species, whereas in the Early Plateaued pattern, fewer species contributed to the enzymatic profiles, particularly during later follow-ups (Fig. 3a and Supplementary Fig. 8).

Association between the Late Matured pattern and T1D risk varies by host genetic background in immune response regulation

Host genetics is a dominant risk factor for T1D¹². Although our findings indicate that genetics minimally influenced the early development of the gut microbiome, as has been previously described in adult populations²⁰, it remains possible that genetic background modifies the relationship between microbiome maturation and T1D risk. We therefore evaluated whether the associations between the three maturation patterns and T1D differed by genetic background, summarized by the top five genetic PCs. We found that the association between the Late Matured pattern and T1D risk significantly varied with genetic

PC3 levels ($P_{\text{interaction}} = 0.02$; Supplementary Table 13). The RR for T1D or seroconversion comparing the Late Matured with the Early Matured pattern varied according to PC3 levels, with $RR > 1$ at high PC3 values and $RR < 1$ at low PC3 values. By contrast, the Early Plateaued pattern was associated with consistently elevated risk ($RR > 1$) relative to the Early Matured pattern across the entire PC3 range (Fig. 4a,b).

To better understand the unique biological basis of this gene-microbiome interaction for PC3, we compared SNP loadings across the top five PCs to examine how PC3 differs from the others. We found that in the top five PCs, SNPs with high loadings were predominantly located on chromosomes 3, 6 (including the HLA region) or 8 (Fig. 4c and Supplementary Fig. 9). Notably, regions with high SNP loadings were similar between PC2 and PC3, but only PC3, not PC2, modified the associations between microbiome maturational patterns and T1D risk. Motivated by this discrepancy, we compared the genetic features of PC2 and PC3, including their top contributing SNPs (Supplementary Fig. 10), to better understand the basis of the PC3-specific association. Functional annotation of SNP loadings further revealed several differences between PC2 and PC3. In particular, PC3 showed unique enrichment for Gene Ontology terms related to antimicrobial humoral response and antiviral immune response, which were not observed for PC2 (Fig. 4d, Supplementary Fig. 11 and Supplementary Table 14).

Host genetic background contributes to variation in specific microbes and their developmental trends over time

Last, we investigated the impact of genetics on the initial composition and developmental trajectories of the gut microbiome. Previous studies have shown that genetic influence on the adult microbiome is minimal²⁰. Our analysis confirmed this finding for the overall microbiome configuration in children. However, this does not rule out our hypothesis that host genetic backgrounds influence specific microbial features and their developmental trends over time. To test this hypothesis, we used linear mixed-effect models (Methods). Based on our previous study in TEDDY⁸, which classified different developmental patterns in infants and toddlers, we categorized samples into two age groups: the infant group (up to 18 months) and the toddler group (up to 3 years).

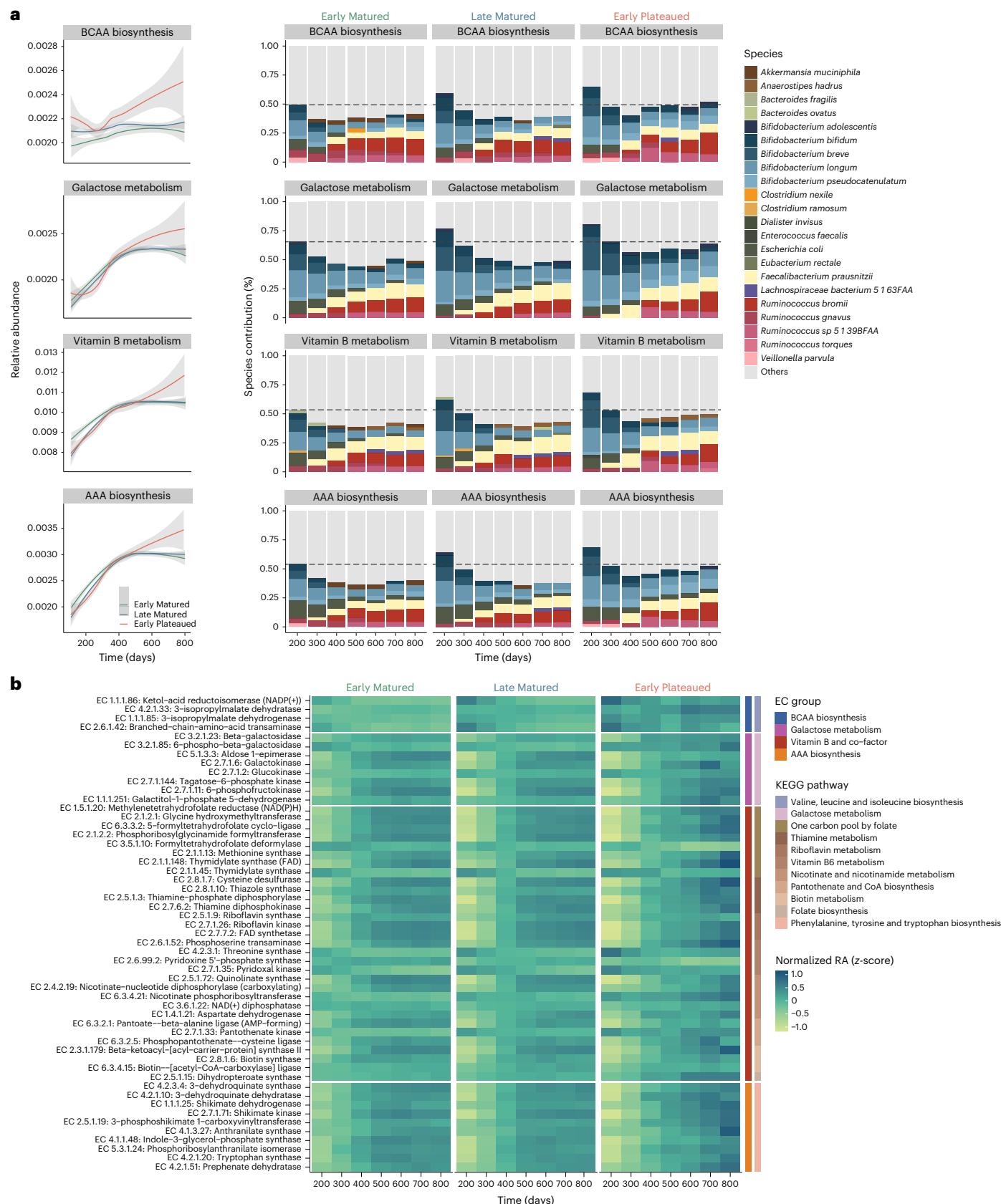
The population genetic structure, again summarized by genetic PC scores, was associated with various microbial species at baseline (quantified by the β coefficients of the PCs) and with differences in their rates of change over time trajectories (quantified by the β coefficients of the interaction terms between a PC and time) (Supplementary Tables 15–17). These associations were particularly prominent in infant participants after the cessation of breastfeeding (Fig. 5a, Supplementary Fig. 12 and Supplementary Table 16). Notable species include the inflammatory polysaccharide producer *Ruminococcus gnavus* and the probiotic bacteria *Lactobacillus casei/paracasei* group (Fig. 5b and Supplementary Table 16). During breastfeeding, the development of *Streptococcus salivarius* significantly varied across genetic PC1 (Fig. 5a and Supplementary Table 15). In toddlers, no species was associated with genetic PC1, which is highly correlated with country of origin. However, the development of plant polysaccharide-degrading species *Dorea longicatena* and butyrate producer *Faecalibacterium prausnitzii* was significantly associated with genetic PC4 in toddlers (Fig. 5b and Supplementary Table 15). For SNP

Fig. 3 | Differential development of microbial processes involved in galactose, vitamin B, and amino acid metabolism contribute to functional variation across microbiome maturational patterns. a, Temporal changes in enzymes encoded by microbial community members, which vary across distinct microbiome maturational patterns, are involved in the biosynthesis of AAAs, BCAAs and vitamin B and galactose metabolism. Trajectories were smoothed and visualized using the LOESS algorithm. Lines represent LOESS-smoothed fitted values, and shaded areas indicate 95% CIs. The right bar plots show the top eight microbial species contributing most to each functional group during the follow-up. The dashed line shows the cumulative contribution of the top species in the

Early Matured group at the first early bin (100–200 days), serving as a reference for comparing differences across clusters. Microbial enzymes are encoded by a more diverse group of microbial species in the Early Matured pattern, whereas a less diverse set of microbial species contributes to the enzymatic profiles, particularly during later follow-up. **b**, Changes in select microbial enzymes across microbiome maturational patterns during follow-up. Enzymes were ordered according to their pathway order. The yellow–green gradient represents relative abundance normalized using z-scores across the follow-up period for each enzyme. KEGG, Kyoto Encyclopedia of Genes and Genomes.

associations, multiple SNPs were linked to species development (Fig. 5c and Supplementary Tables 18 and 19), with fewer associations observed in the infant group. In toddlers, rs13273009, located near *MSRA*, which encodes a methionine sulfoxide reductase involved in oxidative stress

and inflammatory regulation, was linked to *F. prausnitzii*. Similarly, rs584828 near *IGFBP4*, which encodes a regulator of the insulin-like growth factor axis, was linked to the polysaccharide-degrading species *Bacteroides caccae* (Fig. 5d).



Discussion

The increasing global incidence of T1D cannot be fully explained by genetic susceptibility alone, underscoring the need to explore additional predictive and aetiological factors, such as the gut microbiome⁵. In this study, we conducted a prospective analysis that comprehensively investigates the relationship between host genetics, early-life microbiome development and T1D risk. Utilizing longitudinally collected metagenomic data from the TEDDY study^{15,33}, we identified three distinct microbiome maturational patterns during the first 2 years of life: Early Matured, Late Matured and Early Plateaued. We interpret these as a discretization of an underlying continuum of microbiome development rather than discrete biological categories³⁴. These patterns were associated with varying risks of developing T1D, providing strong evidence that disturbances in early microbiome development are linked to T1D onset. Additionally, we found that the association between the Late Matured pattern and T1D risk varied depending on genetic backgrounds, particularly in pathways related to microbial defence, antiviral immune responses, and broader immune regulation functions. Moreover, we investigated the genetic impact on initial microbiome composition and identified that specific microbial species, particularly during early colonization, were significantly—albeit often subtly—affected by genetic factors. Although our analyses were not designed to establish causality, these findings highlight that host genetic background can modulate the consequences of deviations in early microbiome maturation, even in the context of relatively modest direct genetic effects on microbial composition. This interaction-based perspective provides biological context for how genetics and early microbial development may jointly contribute to T1D susceptibility.

Our study uncovered distinct maturational patterns of the developing gut microbiome during early life, driven largely by distinct temporal trajectories of specific species belonging to the *Bifidobacterium* and *Ruminococcus* genera, as well as *C. hathewayi* and *D. longicatena*. These patterns capture the dynamic nature of the microbiome as it adapts to the host's immune system maturation and the dramatic shifts in nutritional demands during breastfeeding, weaning and the introduction of solid foods^{21,35}. Such dynamic, non-linear maturational patterns may have been overlooked in previous studies^{10,36}, primarily due to insufficient sample sizes, follow-up length and the power of trajectory modelling approaches with dense time series. To some extent, these patterns may reflect early-life risk factors such as breastfeeding, delivery mode and dietary intake, as well as genetic influences. However, our subsequent analysis demonstrated that these patterns were associated with varying risks of developing T1D, independent of these known risk factors. This suggests that the patterns capture additional variability intrinsic to the biological processes encoded by the gut microbiome that are unlikely to be explained by host factors alone. We also note that, although geographic differences in the gut microbiome are well documented between populations with markedly different environmental, dietary or lifestyle exposures, most notably in

cross-continental comparisons between industrialized and traditional populations, such effects tend to be more modest within high-income, industrialized cohorts such as TEDDY, and are further attenuated in early life, when microbiome variation is dominated by age, feeding practices and delivery mode^{19,37}.

The functional differences identified across these patterns are consistent with previous reports on the role of microbial functions in T1D development^{8,9}. For instance, efficient galactose metabolism by *Bifidobacterium* spp. supports the production of short-chain fatty acids (SCFAs), which are crucial for maintaining gut health and metabolic balance^{28,38}. Disruptions in this process can impair SCFA production and hinder immune system maturation, thereby increasing T1D risk^{8,9}. Similarly, we observed potential compensatory mechanisms during early-life microbiome maturation. For example, vitamins synthesized by gut bacteria, such as B vitamins, are vital for metabolic and immune functions³⁹. A robust vitamin-producing microbiome supports metabolic health and may reduce T1D risk⁴⁰. Additionally, essential amino acids processed by the microbiome influence host metabolic pathways and immune responses, with imbalances in amino acid metabolism leading to metabolic and immune dysfunctions that elevate T1D susceptibility^{9,32}.

Over the past decade, several human studies have examined the role of the early-life microbiome in the development of T1D^{7,8,10,36}. In our previous study in the TEDDY study, we found that children with T1D exhibited reduced microbial diversity and increased levels of inflammation-related bacteria, particularly from the *Bacteroides* and *Veillonella* genera⁸. These microbial shifts were accompanied by functional alterations in amino acid and SCFA metabolism. These findings align with earlier, smaller studies using 16S rRNA amplicon sequencing, such as the studies by Kostic et al.⁷ and Davis-Richardson et al.³⁶. Stewart et al.¹⁰, also in the TEDDY study, applied 16S rRNA sequencing and identified higher levels of *Bacteroides* and lower levels of *Prevotella* associated with IA seroconversion. They also observed functional changes, including upregulated metabolism of vitamins and amino acids, as well as immune response pathways in those who experienced seroconversion¹⁰. Furthermore, the important role of *Bifidobacterium* spp. in infant microbiomes is well-established³², and recent findings showed that bifidobacteria-driven maturation of the microbiome cannot be entirely explained by breastfeeding alone⁴¹, suggesting the presence of other influencing factors. These human studies, supported by mechanistic evidence from animal studies^{5,42}, underscore the potential importance of microbiome development in T1D onset.

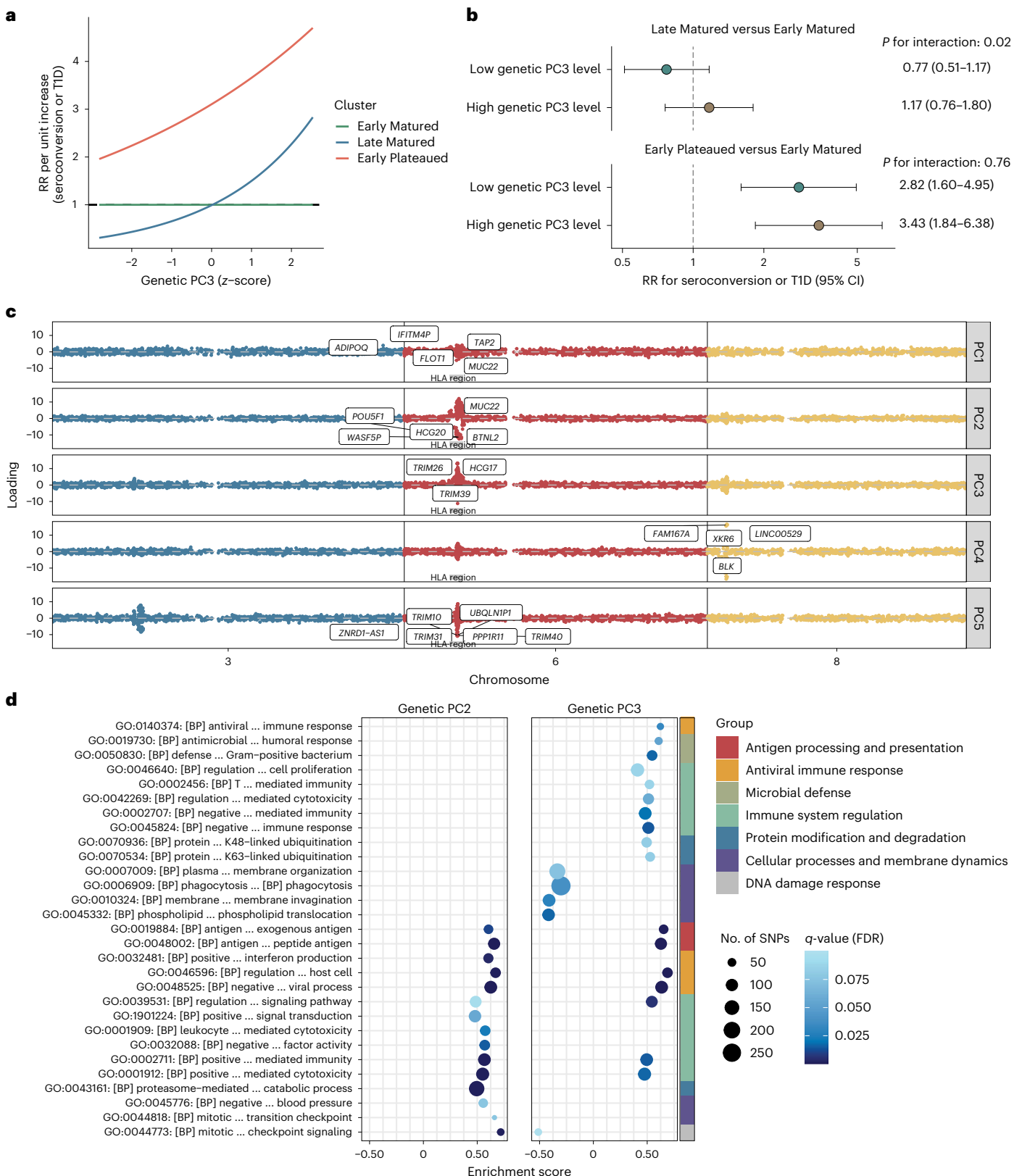
Our study demonstrates that the association between gut microbiome maturation and T1D risk depends on the host genetic background, highlighting the complex interplay between genetic factors and the microbiome in shaping immune and metabolic health during early life. Previous studies have shown that T1D is strongly associated with genetic variations in the HLA region⁴, as well as with a subset of non-HLA loci, including genes involved in immune regulation such as

Fig. 4 | Association between the Late Matured pattern and type 1 diabetes risk varies by individuals' microbial interaction genetics. **a**, Marginal associations between microbiome maturation patterns and T1D risk across PC3. Curves represent model-based marginal RRs for T1D or seroconversion, comparing the Late Matured and Early Plateaued patterns to the Early Matured pattern (as reference group) at equivalent PC3 values. The Late Matured pattern shows PC3-dependent directionality, with RR > 1 at high PC3 levels and RR < 1 at low PC3 levels. In contrast, the Early Plateaued pattern shows consistently elevated risk (RR > 1) across the full range of PC3 values. **b**, RRs and 95% CIs comparing maturation patterns at low versus high PC3 levels. The study population was dichotomized into high and low PC3 groups based on the population median. The numbers of participants in the low genetic PC3 level group were Early Matured ($n = 86$), Late Matured ($n = 163$), and Early Plateaued ($n = 48$); corresponding numbers in the high genetic PC3 level group were Early Matured ($n = 86$), Late Matured ($n = 148$) and Early Plateaued ($n = 63$). Points represent

estimated RRs (the measure of centre), and horizontal bars indicate the corresponding 95% confidence intervals. Comparisons are shown for Late Matured versus Early Matured (top) and Early Plateaued versus Early Matured (bottom), stratified by PC3 level. **c**, The distributions of SNP loadings along the top five genetic PCs on chromosomes 3, 6 (with HLA region indicated by the grey line) and 8. SNPs with high loadings in PC2 and PC3 are concentrated within the major histocompatibility complex. The top absolute loading SNPs in each PC were annotated using the nearest gene within 5 kb. **d**, Comparison in functional annotation between PC2 and PC3 reveals the biological basis of the interaction between PC3 and the Late Matured pattern in relation to T1D risk. We performed functional annotation using gene set enrichment analysis based on the SNP loadings and Gene Ontology (GO) terms related to biological processes. The bubble size represents the number of SNPs, while the colour gradient indicates the statistical significance of each GO term. We corrected for multiple comparisons using the Benjamini–Hochberg procedure, targeting an FDR of 0.1.

PTPN22, *IL2RA* and *IFIH1* (ref. 12). Our study identified genetic factors related to antimicrobial humoral response, antiviral immune response and broader immune regulation as key modulators of the relationship between gut microbiome maturation and T1D risk. Variants affecting antimicrobial and antiviral pathways influence the host's ability to control pathogenic bacteria while supporting beneficial taxa, thereby

shaping microbiome composition and function⁴³. By promoting more regulated immune–microbiome interactions in early life, such genetic variation may help prevent dysbiosis and reduce the release of immunogenic microbial components into the periphery, ultimately lowering T1D risk⁴⁴. Together with previous TEDDY evidence that persistent *Enterovirus B* increases autoimmune risk, these findings suggest that



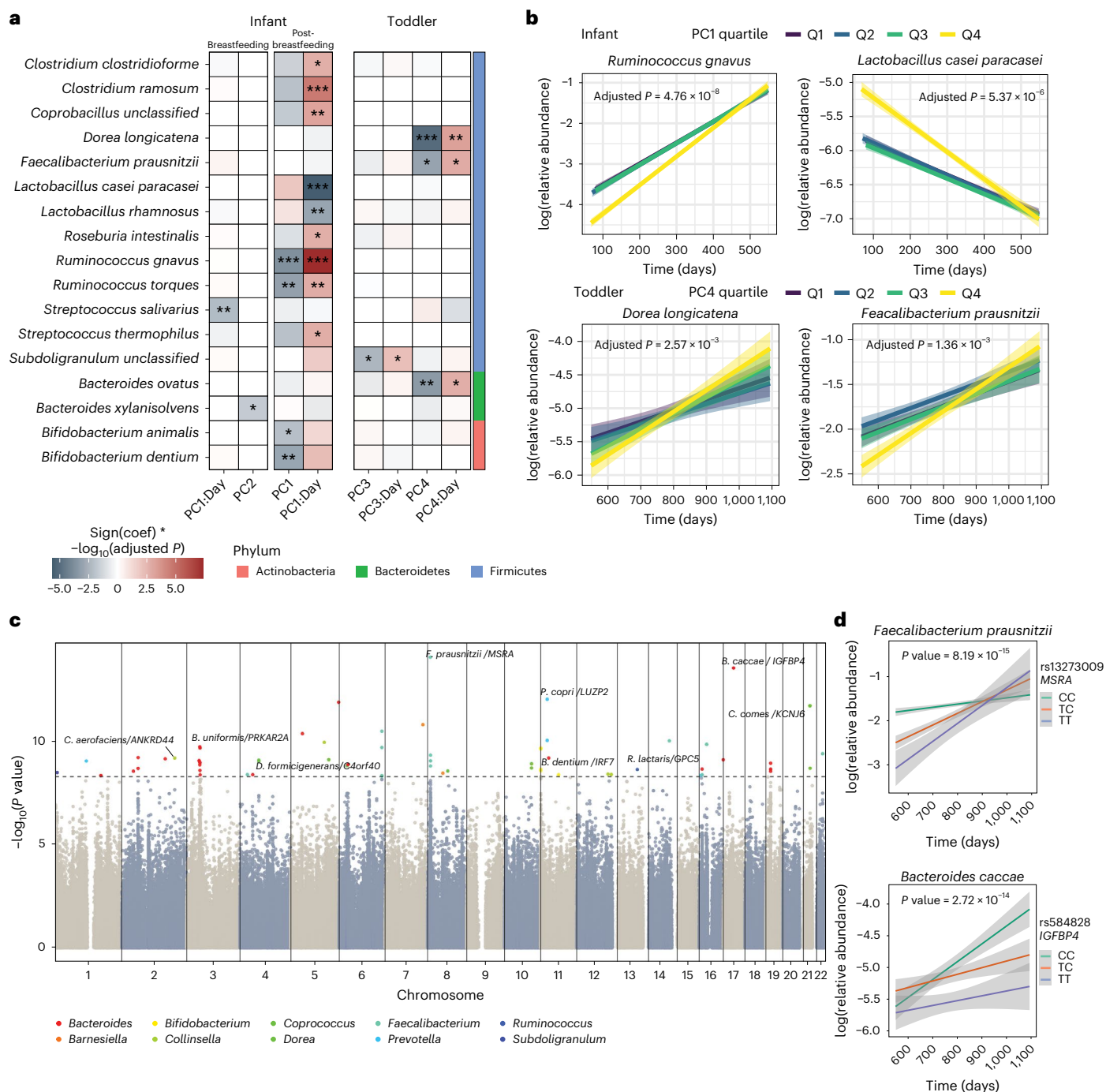


Fig. 5 | Host genetic backgrounds influence specific microbial features and their developmental trajectories during early life. a, The impact of population genetic structure on baseline composition and temporal changes in the gut microbiomes of infants (0–540 days) and toddlers (541–1,095 days). The population genetic structure, summarized by genetic PC scores, was associated with diverse microbial species at baseline (quantified by the β coefficients of the PCs) and their developmental trajectories (quantified by the β coefficients of the interaction terms between a PC and time). These associations were quantified using linear mixed models that included participants' IDs as random effects and clinical centre, genetic PCs, time and interaction terms between a genetic PC and time as fixed effects. Statistical significance was assessed using two-sided Wald tests. These associations were particularly prominent in infants after the cessation of breastfeeding. The colour gradient from blue to red represents both the significance level and direction of these associations, quantified by β coefficients and Bonferroni-adjusted P values from linear mixed models. *Adjusted $P < 0.05$, **Adjusted $P < 0.01$, ***Adjusted $P < 0.001$. **b**, A subset of microbial species showed varying developmental trajectories across different

genetic backgrounds. Different colours distinguish different quartiles of PC1, and straight lines represent fitted linear developmental trajectories. Adjusted P values for PC $\times$ time interaction effects were calculated using the Bonferroni method as described above. **c**, Distribution of microbiome-associated signals across ImmunoChip loci. Each dot represents the $-\log_{10}(P \text{ value})$ of the interaction term between a genetic variant and time, derived from the linear mixed models described above and plotted by chromosomal position. The horizontal dashed line indicates the study-wide significance threshold accounting for multiple comparisons ($P = 5 \times 10^{-8}$). Significant results are coloured by microbial genus. **d**, Developmental trajectories of two gut microbial species, *F. prausnitzii* and *Bacteroides caccae*, showing genotype-specific differences over time. Each panel displays the longitudinal trajectories of log-transformed relative abundance for the indicated species stratified by SNP genotype. Lines represent linear regression fits for visualization of genotype-specific trajectory patterns, and shaded areas represent 95% CIs. The P values correspond to the interaction terms in the linear mixed model. For each SNP, the nearest annotated gene is indicated.

reduced antiviral capacity captured by PC3 may heighten vulnerability to viral persistence in children with delayed microbiome maturation.

Last, by leveraging the unique TEDDY cohort, we conducted what we believe to be the first study examining the genetic contribution to both the initial composition and developmental trajectories of the gut microbiome. Our results emphasize the nuanced role of host genetics in shaping early microbiome development in children. While the overall genetic contribution to the developing microbiome seems moderate, it is still apparently larger than most genetic influences later in life. Our findings indicate that the colonization of specific microbial species, such as *R. gnavus* and *L. casei/paracasei* during infancy, is strongly associated with genetic factors. In toddlers, although the broader genetic influence diminishes, certain species like *F. prausnitzii* and *D. longicatena* continue to show significant genetic associations, suggesting a persistent role of genetics in shaping metabolic functions as the diet becomes more diverse. Notably, we did not recover the *LCT* (rs4988235)–*Bifidobacterium* association that represents one of the strongest adult host–genome–microbiome signals reported to date^{45,46}. This is biologically consistent with the developmental context of our cohort: intestinal *LCT* expression is near-universally high during infancy regardless of rs4988235 genotype, and the adult lactase-mediated mechanism, in which undigested colonic lactose selectively promotes *Bifidobacterium* growth in lactase-non-persistent adults, does not yet operate in this age range. Early-life *Bifidobacterium* abundance is instead predominantly driven by feeding mode, human milk oligosaccharide (HMO) exposure and perinatal factors, which we account for in our models. These results align with our hypothesis that genetic effects on microbiome development during early life are slightly more pronounced than in adulthood^{13,14}, although the effect sizes remain moderate. Of note, we cannot rule out the possibility that these genetic factors, by modulating microbiome development during this critical period, may have long-term implications for health outcomes.

Our study has several strengths. It is based on a prospective risk-set case–control design that accounts for both the onset and timing of T1D, allowing us to establish the temporal relationship between microbiome maturation and disease development. The densely sampled, repeatedly measured microbiome data enabled us to characterize dynamic changes in early-life microbial trajectories. The use of shotgun metagenomic sequencing provided high-resolution information on microbial composition and functional capacity, addressing limitations of earlier 16S-based studies. We also adjusted for a broad range of potential confounders and known T1D risk factors, supporting the robustness of the observed associations. In addition, the inclusion of participants from multiple international centres enhances the generalizability of our findings, several of which we already also observed in additional, external cohort.

Conversely, there are some study limitations. First, as an observational study, it shares inherent limitations with most other population-scale microbiome investigations. Despite our efforts to adjust for major confounders, we cannot completely rule out the possibility of residual confounding. In particular, ancestry-related factors may contribute to some of the observed associations. Although our models adjusted for clinical centre and genetic PCs, residual ancestry-related confounding cannot be entirely excluded. Second, while the ImmunoChip genetic array is valuable for capturing immune-related genetic variants, it lacks the genome-wide coverage needed to explore a broader range of variants. Additionally, taxonomic and functional profiling was performed with bioBakery 2.0 (MetaPhlAn 2 and HUMAnN 2), which has since been updated. The reference databases available at that time were smaller than the most updated ones, so poorly characterized taxa may be underrepresented and our functional annotations less complete. We retained these versions for comparability with previous TEDDY analyses, and because all samples were processed through a single identical pipeline, any

resulting misclassification is non-differential with respect to T1D status and would tend to attenuate rather than create associations. All software versions and analytical parameters used in this study have been documented to facilitate reproducibility. Lastly, our study population is restricted to children with high genetic risk for T1D, which may limit the generalizability of our findings to populations with more typical genetic risk distributions. We thus cannot be certain whether the proportions of children falling into these patterns, or their relationship with T1D, is typical. Future studies in more diverse and representative birth cohorts will be essential to determine the prevalence of these trajectories and to assess whether the observed associations with T1D risk extend beyond genetically high-risk children.

Although our study is not an intervention trial and thus cannot establish a direct causal link, it provides the most comprehensive prospective evidence to date on the involvement of early-life microbiome maturation in the pathogenesis of T1D from a population-based perspective. These findings lay a solid foundation for future mechanistic studies. Additionally, we identified both gene–microbiome interactions related to T1D risk and the direct role of genetics in shaping the developmental microbiome, offering a more nuanced and novel understanding of T1D pathogenesis. Our findings are consistent with conceptual models, such as the hygiene hypothesis, which emphasize the role of early-life microbial ecology in immune development. These results also highlight the potential for microbial biomarkers to serve as targets for future interventional strategies aimed at promoting healthy microbiome development in children, as well as for diagnostic and disease risk prediction applications.

Methods

Ethics statement

The TEDDY study obtained written informed consent for all study participants from a parent or primary caretaker, separately, for genetic screening and participation in the prospective follow-up. The study was approved by local US Institutional Review Boards (IRBs) and European Ethics Committee Boards as follows: Colorado's Colorado Multiple IRB, Georgia's Medical College of Georgia Human Assurance Committee (2004–2010), Georgia Health Sciences University Human Assurance Committee (2011–2012), Georgia Regents University IRB (2013–2015), Augusta University IRB (2015–present), Florida's University of Florida Health Center IRB, Washington state's Washington State IRB (2004–2012) and Western IRB (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 an External Advisory Board formed by the National Institutes of Health (NIH).

Study population and design

The TEDDY study is a prospective cohort study funded by the NIH aimed at identifying environmental causes of T1D. Detailed study design and methods have been previously published^{15,33,47}. Between September 2004 and February 2010, the study screened 424,788 newborns who were younger than 4 months. Among them, 21,589 had HLA genotypes associated with an increased risk of T1D. The parents of 8,676 infants consented to participate in the follow-up study. The TEDDY study involves six clinical research centres: three in the United States (Colorado, Georgia/Florida and Washington) and three in Europe (Finland, Germany and Sweden). Eligible HLA genotypes differed between participants from the general population (four haplogenotypes, DR3/4, DR4/4, DR4/8 and DR3/3, with *HLA-DRB1*04:03* as an exclusion allele) and those with a first-degree relative with T1D (nine haplogenotypes, DR3/4, DR4/4, DR4/8, DR3/3, DR4/4b, DR4/1, DR4/13, DR4/9 and DR3/9, ensuring broad HLA diversity)⁴⁷. Children are followed from 3 months to 15 years of age, with study visits every 3 months until 4 years of age and

every 3 or 6 months thereafter depending on autoantibody status. Sex was recorded at enrolment based on parental report for all participants; gender was not collected.

Our analysis included data from two nested case–control studies within the TEDDY cohort¹⁷. Controls were sampled using a risk-set approach¹⁷. A risk set is defined as the set of individuals who are at risk of the outcome at the specific time point when a case event occurs⁴⁸. In this study, each risk set was operationalized as an event–time-matched stratum containing one case of persistent, confirmed IA or T1D and one control who remained free of IA or T1D at the same follow-up time (± 45 days)¹⁷. This approach ensured a representative sample of person–time accrued during follow-up and allowed for a direct estimation of the incidence rate ratio. Controls were matched by clinical centre, sex and family history of T1D to control for regional, genetic and data-handling differences. The final analytical dataset consisted of 12,151 metagenomes from 887 individuals, including 403 female and 484 male participants. For analyses incorporating genetic data, we included 877 genotyped individuals. To derive microbiome maturational patterns through trajectory analysis, we used a subset of 594 individuals, excluding genetic outliers and those with fewer than four metagenomic samples.

DIABIMMUNE three-country cohort

The DIABIMMUNE three-country cohort included 212 infants recruited from Finland ($n = 71$), Estonia ($n = 71$) and Russia ($n = 70$), countries with substantial differences in T1D and allergy incidence. Infants were selected to have comparable HLA-conferred genetic risk for T1D and were matched by sex. Monthly stool samples were collected during the first 3 years of life, together with clinical assessments and questionnaire-based information on breastfeeding, diet, allergies, infections, family history and medication use²⁴. We included 785 stool samples with repeated measures generated by whole-genome shotgun sequencing from these 212 infants to characterize the gut microbiome. The study was conducted in accordance with the Declaration of Helsinki and approved by the local ethical committees of the participating hospitals, with written informed consent obtained from the parents or study participants before sample collection.

Blood and faecal sample collection and preprocessing

Blood samples were obtained every 3 months until 4 years of age and biannually thereafter for the analysis of IAs (insulin, glutamic acid decarboxylase (GAD) and insulinoma antigen-2 autoantibodies). Whole blood was drawn into a syringe and transferred into a serum separation tube, allowed to clot at room temperature and centrifuged (800g, 4 °C, 20 min) to separate serum from the clot. Stool samples were collected monthly from 3–48 months of age, every 3 months thereafter until 10 years of age, and biannually subsequently following standardized TEDDY sample collection protocols; collection ceased in August 2018. Children who remained antibody-negative after 4 years of age were encouraged to submit samples four times annually despite the transition to a biannual visit schedule. Samples were shipped at ambient temperature or +4 °C with guaranteed delivery within 24 h to the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) repository (US participants) or to the affiliated clinical centre (European participants), which sent monthly bulk shipments of frozen stool to the repository. Full protocols are available in the TEDDY Manual of Operations (<https://repository.niddk.nih.gov/studies/teddy/>).

Serum islet autoantibody measurements

Serum IAs were measured by radio-binding assays as previously described⁴⁹. Samples were collected longitudinally according to the TEDDY follow-up schedule (every 3 months from 3–48 months of age, and every 3 or 6 months thereafter depending on autoantibody status). Glutamate decarboxylase autoantibodies (GADA), insulin autoantibodies (IAA) and islet antigen-2 autoantibodies (IA-2A) were measured in

two TEDDY reference laboratories according to the location of the clinical centre (the Barbara Davis Center for Childhood Diabetes for US centres and the University of Bristol for European centres). From February 2010, harmonized assays were used for GADA and IA-2A in both laboratories, whereas IAA continued to be measured with the original TEDDY assay. All samples testing positive in one reference laboratory, together with 5% of negative samples, were re-tested in the other laboratory and were considered confirmed only if concordant. Positivity was determined against laboratory-specific thresholds; results for GADA and IA-2A are expressed in DK U ml⁻¹ and those for IAA as an index.

Autoimmunity and diagnosis of type 1 diabetes

In this study, persistent autoimmunity was defined as positivity for IAs in two consecutive visits. Positive for IAs was defined as positive for at least one IA in two reference laboratories on at least two consecutive visits. The date of seroconversion (time to first autoantibody) was defined as the collection date of the first sample testing positive. T1D was defined according to the American Diabetes Association criteria for diagnosis⁵⁰. This classification is based on pathogenesis rather than the requirement for insulin therapy.

Shotgun metagenomic sequencing and initial bioinformatics

Samples were metagenomically sequenced as one library each multiplexed through Illumina HiSeq 2000 machines using a 2 × 100-bp paired-end read protocol, and all downstream processing was conducted on paired-end reads. A minimum DNA input of 20 ng was required for whole-genome shotgun library preparation. Sequencing runs were considered passing if >60% of reads passed Illumina quality filtering, with a target sequencing depth of 1 Gb per sample and a minimum acceptable threshold of ~500 Mb when necessary. Samples failing to meet these criteria or yielding insufficient high-quality reads were excluded.

Raw FASTQ files generated by Casava v.1.8.2 (Illumina) were adaptor-trimmed with cutadapt v.1.9dev2, quality-trimmed with Trim Galore v.0.2.8 (Babraham Bioinformatics) and de-replicated and low-complexity-filtered with PRINSEQ v.0.20.5. Host and non-bacterial reads were removed using Bowtie2 v.2.2.3 (--end-to-end --sensitive) by alignment against a reference database derived from the NCBI whole-genome sequencing archive (March 2015 release), including bacterial, viral, human and vector genomes; human reads were filtered using GRCh38 (hg38), and reads whose best alignment was non-bacterial were excluded. Default parameters were used unless otherwise specified. Non-default parameters were: Trim Galore/cutadapt, -q20, --stringency 6, --length 50 and --retain_unpaired with --length 151 and --length 251; PRINSEQ, -derep 12, -lc_method dust, -lc_threshold 5, -trim_ns_left 1, -trim_ns_right 1 and -no_qual_header; Bowtie2, --no-unal, --no-sq and --no-head.

Microbial taxonomical and functional profiling

Taxonomic and functional profiling were performed using the standardized bioBakery 2.0 workflow, consistent with previous TEDDY microbiome studies⁸, to ensure comparability across the TEDDY metagenomic resource. Taxonomic profiling was conducted using MetaPhlAn 2 (v.2.6.0)⁵¹ with default settings, which employs a library of clade-specific markers to quantify bacterial and archaeal communities at the species level.

Functional profiling was carried out using HUMAnN 2 (v.0.9.4)⁵² to characterize community pathway and gene-family abundance. For each metagenome, HUMAnN 2 constructs a sample-specific reference database from the pangenomes⁵³ of species detected by MetaPhlAn 2, maps sample reads against this database to quantify gene abundance in a species-stratified manner, and subjects unmapped reads to a translated search against UniRef90⁵⁴ to capture taxonomically unclassified but functionally distinct gene families. Pathway abundances were

reconstructed from gene families annotated to metabolic reactions using MetaCyc definitions⁵⁵ and enzyme abundance (level-4 EC categories) by summing gene families annotated to each EC number using UniRef90-EC annotations from UniProt⁵⁶. Viral functional content was not analysed, as the underlying ChocoPhlAn and UniRef reference databases contain limited viral content.

Genotype analysis

TEDDY children were genotyped by the Center for Public Health Genomics at the University of Virginia using the Illumina ImmunoChip, a custom array of SNPs from regions robustly associated with autoimmune disease, selected for 12 autoimmune diseases by the ImmunoChip Consortium^{57,58}. Genotype quality control (QC) was performed using KING (v.2.3.0) and PLINK (v.1.90b7). KING --autoQC was used for sample-level, SNP-level QC (95% call rate) and sex QC. In PLINK, SNPs with a minor allele frequency of less than 0.05 were removed, as were SNPs that failed the Hardy–Weinberg test ($P < 0.001$). SNPs in linkage disequilibrium (window 50, step 5, $R^2 = 0.2$) were pruned before calculating heterozygosity, and individuals outside $\mu \pm 3\sigma$ were removed. The proportion of Identity-by-descent was calculated to identify related individuals; for pairs with $PI_HAT > 0.2$, the member with higher missingness was removed. PCA was performed, and the top five PCs were used for downstream analysis. Individuals with PC scores below $Q1 - 1.5 \times$ interquartile range or above $Q3 + 1.5 \times$ interquartile range were considered genetic outliers and removed. BEDTools (v.2.31.1) was used to identify the nearest gene to each SNP using GRCh37.

Covariable measurement

Metadata were collected using validated questionnaires⁴⁷. Information about mothers, pregnancy and birth was collected by questionnaire at the 3-month clinic visit and included mode of birth, 5-min Apgar score, pregnancy complications, maternal diabetes, gestational age, and maternal medication use during pregnancy. Infant diet was assessed through mailed questionnaires completed before the first clinic visit, structured interviews at each clinic visit, and records kept by the mother in the TEDDY Book, covering breastfeeding duration, age at introduction of foods, infant formula type, drinking water source, elimination diets, and dietary supplements. A 24-h dietary recall was obtained at the first 3-month clinic visit to aid caretaker training and to capture the infant's diet at that age. Primary caretakers were trained at the 3-month visit to keep 3-day food records at 3-month intervals during the first year and biannually thereafter, reviewed with trained study personnel at each visit.

Diversity and ordination-based analyses

The structure of the gut microbiome was examined using the vegan package (v.2.5.6). The Bray–Curtis distances between samples were calculated based on their MetaPhlAn 2 profiles using the `vegan::vegdist()`. To measure both species richness and evenness within a sample, the Shannon diversity index was computed using the `vegan::diversity()` (Extended Data Fig. 3). PCoA was performed based on the Bray–Curtis distances using the `ape::pcoa()` from the `ape` package (v.5.4.1). Additionally, PERMANOVA was performed on Bray–Curtis distances using `vegan::adonis2()`. To respect the non-independence of repeated longitudinal samples, we used a restricted permutation design in which whole individuals were treated as the exchangeable unit: individuals were freely permuted, while no permutation was performed within individuals. The scheme is appropriate for testing between-subject covariates such as genetic PCs and country of origin, as it keeps repeated measures intact and prevents anti-conservative inflation of significance from pseudo-replication. Significance was assessed using 999 permutations. The stability of the gut microbiome over time was evaluated by calculating the Bray–Curtis distance from the initial time point for each individual.

Association between host genetic variation and microbiome composition

Mantel and Procrustes analyses were performed within predefined age intervals (<1, 1–2, 2–3, 3–4, 4–5 and 5–6 years). To ensure sample independence, only the earliest stool sample from each participant within each age interval was retained. Participants without a microbiome sample in a given interval were excluded from that analysis. Microbiome composition was represented by species-level relative abundance profiles using Bray–Curtis dissimilarities, whereas host genetic variation was represented by Euclidean distances calculated from standardized top five genome-wide genotype-derived PCs (PC1–PC5).

Mantel tests were used to evaluate the correspondence between microbiome and genetic distance matrices using Spearman rank correlations with significance assessed by 999 permutations. For Procrustes analysis, PCoA was performed on the Bray–Curtis distance matrix, and the top five microbiome coordinate axes were compared with PC1–PC5 using the `vegan` package in R. Statistical significance was evaluated using the PROTEST permutation procedure with 999 permutations.

Microbiome maturational pattern identification

The maturational patterns were identified using a trajectory-clustering algorithm implemented in the R package `traj` (v.2.0.1), which employs the Bray–Curtis distance to quantify changes in gut microbial composition from baseline across subsequent visits during the initial 800-day period. Baseline was defined as the earliest available stool sample, which reflects a relatively consistent microbial state before major environmental exposures and dietary diversification. `Traj` implements a three-step procedure to identify clusters of individual longitudinal trajectories. This procedure involves (1) computing a number of ‘measures of change’ capturing various features of the trajectories; (2) using a PCA-based dimension reduction algorithm to select a subset of measures; and (3) using the *k*-means clustering algorithm to identify clusters of trajectories.

The optimal number of maturational patterns was determined by balancing the results from various diagnostic metrics, including the Calinski–Harabasz Index⁵⁹, the gap statistic⁶⁰ and the elbow method (Supplementary Fig. 3). The elbow plot showed a marked reduction in slope around $k = 3$, beyond which additional clusters yielded diminishing returns. Although the Calinski–Harabasz Index favoured $k = 2$, the gap statistic indicated that this solution was under-separated. Considering these diagnostics together, as well as statistical efficiency for downstream association analyses, we selected $k = 3$ as the most parsimonious and robust solution.

Estimation of relative risk of type 1 diabetes

Due to the risk-set sampling design of this study, we were able to estimate RRs^{61,62} for a composite end point encompassing either the development of IAs (seroconversion) or a clinical diagnosis of T1D, as well as for each of these event types individually, by comparing a microbiome maturation pattern to Early Matured pattern. Utilizing the equivalence between the partial log-likelihood of the Cox proportional hazards model and the log-likelihood of conditional logistic regression⁶³, RR estimation was conducted using the R package `survival` (v.3.5-5), with the model stratified by risk sets. The primary exposures of interest were the maturational patterns of the gut microbiome. The models were adjusted for potential confounding factors known to influence the risk of seroconversion and T1D. These factors included the matching variables in the case–control design (clinical centre, sex and family history of T1D) as well as additional covariates, including genetic PCs, interactions between clusters and PCs, delivery mode, breastfeeding, antibiotic use, probiotic use and timing of the introduction of solid foods. We applied methods from Langholz et al.⁶⁴ and Samuelsen et al.⁶⁵ to calculate adjusted cumulative incidence curves extrapolated to the full TEDDY cohort from multivariable models incorporating sampling weights to account for risk-set sampling in the nested case–control design.

Partitioning of variance in T1D risk

To assess the relative contribution of different components to T1D risk, we quantified the variance explained using a likelihood-based pseudo- R^2 , which reflects the proportion of variation in the model log-likelihood explained by the included predictors. We compared pseudo- R^2 values across a set of nested models designed to partition the contribution of each component. The baseline model (M0) included demographic and clinical covariates (for example, sex, clinical centre and delivery mode). M1 added the microbiome maturation pattern; M2 additionally incorporated host genetic PCs (PC1–PC5); and M3 further included interaction terms between the genetic PCs and the maturation pattern. Differences in pseudo- R^2 between models were used to estimate the incremental variance explained by each component.

Taxonomic and functional feature association analyses

For a microbial species or a MetaCyc pathway to qualify for downstream analyses, it needed to be present in at least 10% of samples with a minimum detectable relative abundance of 0.0002. The criterion for EC numbers was that at least 10% of samples with that feature were detectable at a minimum relative abundance of 0.00002. Additionally, we removed functional features with high correlations to others by selecting the most abundant feature from each cluster as its representative.

We employed linear mixed models in MaAsLin 2 (ref. 25), using gut microbiome features as dependent variables and microbiome maturational patterns as the independent variable. Individual IDs were incorporated as random effects, whereas clinical centre, sex, delivery mode, antibiotic use, introduction of solid foods and breast-feeding status were included as fixed effects. This approach allowed us to identify microbial species with differential abundance across maturational patterns within each 100-day interval. The false discovery rate (FDR)-adjusted P values (q) were calculated using the Benjamini–Hochberg method.

Gene set enrichment analysis

To characterize further the biological processes associated with genetic PCs, we performed a SNP-level pre-ranked Gene Ontology (GO) term⁶⁶ enrichment analysis in which SNPs were ranked by their loading values for each PC. Significance was assessed by permuting SNP labels within SNP sets, with FDR control. This directly tests whether the observed concentration of high-loading SNPs within a given ontology is greater than expected by chance, given the fixed PC loadings. We adopted the definition of ‘informative GO terms’ from previous studies⁶⁷ to curate a list of GO terms as input for the enrichment analysis. A GO term was defined as informative if it contained more than 20 genes, and all its child terms contained fewer than 20 genes. The enrichment analysis used the ranking of SNP loading factors for each genetic PC. Enrichment scores were calculated using the fgsea R package with 1,000 permutations against curated GO terms, and terms with $q < 0.10$ were considered significantly enriched. Visualization was generated using fgsea (v.1.22.0) and ggplot2 (v.3.3.6).

Genetic and microbiome association analyses over time

We used the following linear mixed-effects model with random slopes to identify microbiome features that were associated with genetic features:

$$\text{microbiome} \sim \text{genetics} + \text{age} + \text{genetics:age} + \text{clinical centre} + (1 + \text{age} | \text{individual})$$

In this model, age refers to the child’s age in days at the time of stool collection. We include a random slope on age so that each individual can have their own microbiome developmental trajectory, which accommodates the substantial inter-individual heterogeneity characteristic of early-life microbiome maturation.

The genetic main effect estimates whether genetic variation is associated with baseline microbiome composition, independently of

time. The genetics:age interaction term tests whether genetic background is associated with differences in the rate of microbiome change over time, rather than at a single time point. Thus, the interaction evaluates whether the slope of microbial development varies with genotype, after adjusting for clinical centre to account for geographic differences in environment and sampling.

Samples were subset into two age groups: infant samples collected before 18 months and toddler samples collected after. Models were run separately for these age groups because gut microbiome development up to three years may be non-linear. In each age group, species were included for modelling if they were present at >1% relative abundance in at least one sample among 25% or more participants. Similarly, pathways were included for modelling if they were present at >1 copy per million reads in at least one sample from 25% of individuals. Microbiome feature abundances were transformed using the formula:

$$\log(\text{abundance} + \varepsilon)$$

where ε is half of the minimum non-zero abundance of the microbiome feature of interest. For SNPs, minor homozygous alleles were included if they were present in >30 individuals per age group. Models were implemented using the lmerTest::lmer() function from lmerTest v.3.1-3, and an analysis of variance test was used to determine the significance of associations in the models. For PC-based models, the P values were adjusted using the Bonferroni correction methods. For SNP models, a P value of 5×10^{-8} was used as the threshold of significance.

Statistics and reproducibility

This study is an observational analysis of two nested case–control studies within the prospective TEDDY birth cohort; no experimental replication was performed, and all analyses were conducted on the full set of available samples meeting the quality criteria described above. No statistical method was used to predetermine sample size; sample size was determined by the number of eligible case–control sets available within the existing TEDDY nested case–control design.

Data were excluded as follows: metagenomic samples whose sequencing run failed quality criteria (<60% of reads passing Illumina quality filtering) or that yielded insufficient high-quality reads; individuals failing genotype QC (call rate <95%, heterozygosity outside $\mu \pm 3\sigma$ or genetic outlier status by PCs) and, for related pairs (PI_HAT > 0.2), the member with higher missingness; individuals with fewer than four metagenomic samples for trajectory analyses; individuals without a stool sample in a given age interval for the corresponding interval-specific analysis; and microbial features not meeting the prevalence and abundance thresholds described above.

The experiments were not randomized; exposure status was observational and controls were selected by risk-set sampling matched on clinical centre, sex and family history of type 1 diabetes rather than by randomization. The Investigators were not blinded to allocation during experiments and outcome assessment.

All analyses were performed in R v.4.2.3. Microbiome maturational patterns were derived by trajectory-based clustering of longitudinal taxonomic profiles; community-level comparisons used distance-based ordination and PERMANOVA; feature-level associations between microbiome features and maturational patterns, host genetic variation and covariates were tested using linear mixed models implemented in MaAsLin 2, with individual identifiers as random effects; and relative risks of type 1 diabetes were estimated within the matched case–control sets. Full model specifications, including covariates and random-effects structures, are given in the corresponding Methods subsections. All tests were two-sided unless otherwise stated. Multiple testing was controlled by the Benjamini–Hochberg FDR, except for PC-based linear mixed models (Bonferroni) and SNP-level models ($P < 5 \times 10^{-9}$). Permutation-based tests used 999 permutations unless otherwise specified.

Reproducibility was addressed in three ways. All 12,151 metagenomes from 887 children were processed through a single identical

bioinformatic pipeline, with software versions, reference databases and analytical parameters reported above. The trajectory-clustering framework was applied independently to the DIABIMMUNE three-country cohort, in which qualitatively similar maturational patterns were recovered. The primary association between maturational pattern and T1D risk was examined in sensitivity analyses, including sex-stratified models and models with additional covariate adjustment.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

TEDDY microbiome whole-genome sequencing and SNP data that support the findings of this study are available in the NCBI database of Genotypes and Phenotypes (dbGaP) under primary accession code [phs001442.v4.p3](https://www.ncbi.nlm.nih.gov/bioproject/500000), in accordance with the dbGaP controlled-access authorization process. Data from TEDDY at <https://doi.org/10.58020/y3jk-x087> (ref. 68) reported here are available upon request through the NIDDK Central Repository website, Resources for Research (<https://repository.niddk.nih.gov/>). Processed taxonomic profiles and meta-data from the DIABIMMUNE three-country cohort are available through the DIABIMMUNE website (<https://diabimmune.broadinstitute.org/diabimmune>). Source data are provided with this paper.

Code availability

This study relies mainly on open-source bioinformatic tools. The analysis-specific programs are available on GitHub at https://github.com/biobakery/teddy_genetics.

References

- bioinformatic pipeline, with software versions, reference databases and analytical parameters reported above. The trajectory-clustering framework was applied independently to the DIABIMMUNE three-country cohort, in which qualitatively similar maturational patterns were recovered. The primary association between maturational pattern and T1D risk was examined in sensitivity analyses, including sex-stratified models and models with additional covariate adjustment.
- ## Reporting summary
- Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.
- ## Data availability
- TEDDY microbiome whole-genome sequencing and SNP data that support the findings of this study are available in the NCBI database of Genotypes and Phenotypes (dbGaP) under primary accession code [phs001442.v4.p3](#), in accordance with the dbGaP controlled-access authorization process. Data from TEDDY at <https://doi.org/10.58020/y3jk-x087> (ref. 68) reported here are available upon request through the NIDDK Central Repository website, Resources for Research (<https://repository.niddk.nih.gov/>). Processed taxonomic profiles and meta-data from the DIABIMMUNE three-country cohort are available through the DIABIMMUNE website (<https://diabimmune.broadinstitute.org/diabimmune>). Source data are provided with this paper.
- ## Code availability
- This study relies mainly on open-source bioinformatic tools. The analysis-specific programs are available on GitHub at https://github.com/biobakery/teddy_genetics.
- ## References
- Quattrin, T., Mastrandrea, L. D. & Walker, L. S. K. Type 1 diabetes. *Lancet* **401**, 2149–2162 (2023).
 - Gregory, G. A. et al. Global incidence, prevalence, and mortality of type 1 diabetes in 2021 with projection to 2040: a modelling study. *Lancet Diabetes Endocrinol.* **10**, 741–760 (2022).
 - Herold, K. C. et al. The immunology of type 1 diabetes. *Nat. Rev. Immunol.* **24**, 435–451 (2024).
 - Pociot, F. & Lernmark, A. Genetic risk factors for type 1 diabetes. *Lancet* **387**, 2331–2339 (2016).
 - Knip, M. & Siljander, H. The role of the intestinal microbiota in type 1 diabetes mellitus. *Nat. Rev. Endocrinol.* **12**, 154–167 (2016).
 - Round, J. L. & Mazmanian, S. K. The gut microbiota shapes intestinal immune responses during health and disease. *Nat. Rev. Immunol.* **9**, 313–323 (2009).
 - 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).
 - Vatanen, T. et al. The human gut microbiome in early-onset type 1 diabetes from the TEDDY study. *Nature* **562**, 589–594 (2018).
 - Yuan, X. et al. Functional and metabolic alterations of gut microbiota in children with new-onset type 1 diabetes. *Nat. Commun.* **13**, 6356 (2022).
 - Stewart, C. J. et al. Temporal development of the gut microbiome in early childhood from the TEDDY study. *Nature* **562**, 583–588 (2018).
 - Sharma, A. et al. Identification of non-HLA genes associated with development of islet autoimmunity and type 1 diabetes in the prospective TEDDY cohort. *J. Autoimmun.* **89**, 90–100 (2018).
 - Steck, A. K. & Rewers, M. J. Genetics of type 1 diabetes. *Clin. Chem.* **57**, 176–185 (2011).
 - Kurilshikov, A. et al. Large-scale association analyses identify host factors influencing human gut microbiome composition. *Nat. Genet.* **53**, 156–165 (2021).
 - Zhernakova, D. V. et al. Individual variations in cardiovascular-disease-related protein levels are driven by genetics and gut microbiome. *Nat. Genet.* **50**, 1524–1532 (2018).
 - TEDDY Study Group. The environmental determinants of diabetes in the young (TEDDY) study. *Ann. NY Acad. Sci.* **1150**, 1–13 (2008).
 - Hadley, D. et al. HLA-DPB1*04:01 protects genetically susceptible children from celiac disease autoimmunity in the TEDDY study. *Am. J. Gastroenterol.* **110**, 915–920 (2015).
 - 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).
 - Kurki, M. I. et al. FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature* **613**, 508–518 (2023).
 - 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).
 - Rothschild, D. et al. Environment dominates over host genetics in shaping human gut microbiota. *Nature* **555**, 210–215 (2018).
 - Tanaka, M. & Nakayama, J. Development of the gut microbiota in infancy and its impact on health in later life. *Allergol. Int.* **66**, 515–522 (2017).
 - Stokholm, J. et al. Maturation of the gut microbiome and risk of asthma in childhood. *Nat. Commun.* **9**, 141 (2018).
 - Leffondré, K. et al. Statistical measures were proposed for identifying longitudinal patterns of change in quantitative health indicators. *J. Clin. Epidemiol.* **57**, 1049–1062 (2004).
 - Vatanen, T. et al. Variation in microbiome LPS immunogenicity contributes to autoimmunity in humans. *Cell* **165**, 842–853 (2016).
 - Mallick, H. et al. Multivariable association discovery in population-scale meta-omics studies. *PLoS Comput Biol* **17**, e1009442 (2021).
 - Saturio, S. et al. Role of *Bifidobacteria* on infant health. *Microorganisms* **9**, 2415 (2021).
 - Sommer, F. & Backhed, F. The gut microbiota—masters of host development and physiology. *Nat. Rev. Microbiol.* **11**, 227–238 (2013).
 - Flint, H. J., Scott, K. P., Duncan, S. H., Louis, P. & Forano, E. Microbial degradation of complex carbohydrates in the gut. *Gut Microbes* **3**, 289–306 (2012).
 - Schirmer, M. et al. Linking the human gut microbiome to inflammatory cytokine production capacity. *Cell* **167**, 1125–1136. e1128 (2016).
 - Hendrix, G. et al. Catabolism of fats and branched-chain amino acids in children with type 1 diabetes: association with glycaemic control and total daily insulin dose. *Endocrinol. Diabetes Metab.* **6**, e448 (2023).
 - Ferretti, P. et al. Assembly of the infant gut microbiome and resistome are linked to bacterial strains in mother’s milk. *Nat. Commun.* **16**, 11536 (2025).
 - Laursen, M. F. et al. *Bifidobacterium* species associated with breastfeeding produce aromatic lactic acids in the infant gut. *Nat. Microbiol.* **6**, 1367–1382 (2021).
 - Rewers, M. et al. The environmental determinants of diabetes in the young (TEDDY) study: 2018 update. *Curr. Diab. Rep.* **18**, 136 (2018).
 - Jeffery, I. B., Claesson, M. J., O’Toole, P. W. & Shanahan, F. Categorization of the gut microbiota: enterotypes or gradients? *Nat. Rev. Microbiol.* **10**, 591–592 (2012).
 - Mueller, N. T., Bakacs, E., Combellick, J., Grigoryan, Z. & Dominguez-Bello, M. G. The infant microbiome development: mom matters. *Trends Mol. Med.* **21**, 109–117 (2015).
 - 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).

37. Yatsunenko, T. et al. Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (2012).
38. De Vuyst, L., Moens, F., Selak, M., Riviere, A. & Leroy, F. Summer meeting 2013: growth and physiology of bifidobacteria. *J. Appl. Microbiol.* **116**, 477–491 (2014).
39. LeBlanc, J. G. et al. Bacteria as vitamin suppliers to their host: a gut microbiota perspective. *Curr. Opin. Biotechnol.* **24**, 160–168 (2013).
40. Levi Mortera, S. et al. Functional and taxonomic traits of the gut microbiota in type 1 diabetes children at the onset: a metaproteomic study. *Int. J. Mol. Sci.* **23**, 15982 (2022).
41. Ennis, D., Shmorak, S., Jantscher-Krenn, E. & Yassour, M. Longitudinal quantification of *Bifidobacterium longum* subsp. *infantis* reveals late colonization in the infant gut independent of maternal milk HMO composition. *Nat. Commun.* **15**, 894 (2024).
42. Brown, E. M., Kenny, D. J. & Xavier, R. J. Gut microbiota regulation of T cells during inflammation and autoimmunity. *Annu. Rev. Immunol.* **37**, 599–624 (2019).
43. Cury, J. et al. Conservation of antiviral systems across domains of life reveals immune genes in humans. *Cell Host Microbe* **32**, 1594–1607 e1595 (2024).
44. Davies, C. S. et al. Immunogenetic variation shapes the gut microbiome in a natural vertebrate population. *Microbiome* **10**, 41 (2022).
45. Goodrich, J. K. et al. Genetic determinants of the gut microbiome in UK twins. *Cell Host Microbe* **19**, 731–743 (2016).
46. Blekhman, R. et al. Host genetic variation impacts microbiome composition across human body sites. *Genome Biol.* **16**, 191 (2015).
47. TEDDY Study Group. The environmental determinants of diabetes in the young (TEDDY) study: study design. *Pediatr. Diabetes* **8**, 286–298 (2007).
48. Rothman, K.J., Greenland, S. & Lash, T.L. Case-control studies. In *Encyclopedia of Quantitative Risk Analysis and Assessment* (eds Melnick, E. L. & Everitt, B. S.) (Wiley, 2008).
49. Bonifacio, E. et al. Harmonization of glutamic acid decarboxylase and islet antigen-2 autoantibody assays for national institute of diabetes and digestive and kidney diseases consortia. *J. Clin. Endocrinol. Metab.* **95**, 3360–3367 (2010).
50. American Diabetes Association. Classification and diagnosis of diabetes. *Diabetes Care* **38**, S8–S16 (2015).
51. Truong, D. T. et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).
52. Abubucker, S. et al. Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput. Biol.* **8**, e1002358 (2012).
53. Huang, K. et al. MetaRef: a pan-genomic database for comparative and community microbial genomics. *Nucleic Acids Res.* **42**, D617–D624 (2014).
54. Suzeck, B. E. et al. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2015).
55. 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).
56. UniProt, C. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **51**, D523–D531 (2023).
57. Torn, C. et al. Role of type 1 diabetes-associated SNPs on risk of autoantibody positivity in the TEDDY study. *Diabetes* **64**, 1818–1829 (2015).
58. Cortes, A. & Brown, M. A. Promise and pitfalls of the immunochip. *Arthritis Res. Ther.* **13**, 101 (2011).
59. Caliński, T. & Harabasz, J. A dendrite method for cluster analysis. *Commun. Stat. Theory Methods* **3**, 1–27 (1974).
60. Tibshirani, R., Walther, G. & Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **63**, 411–423 (2001).
61. Robins, J. M., Prentice, R. L. & Blevins, D. Designs for synthetic case-control studies in open cohorts. *Biometrics* **45**, 1103–1116 (1989).
62. Breslow, N. E. & Day, N. E. Statistical methods in cancer research. Volume I - The analysis of case-control studies. *IARC Sci. Publ.* **1980**, 5–338 (1980).
63. Breslow, N. E., Day, N. E., Halvorsen, K. T., Prentice, R. L. & Sabai, C. Estimation of multiple relative risk functions in matched case-control studies. *Am. J. Epidemiol.* **108**, 299–307 (1978).
64. Langholz, B. & Borgan, O. Estimation of absolute risk from nested case-control data. *Biometrics* **53**, 767–774 (1997).
65. Samuelsen, S. O. A pseudolikelihood approach to analysis of nested case-control studies. *Biometrika* **84**, 379–394 (1997).
66. Gene Ontology, C. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.* **49**, D325–D334 (2021).
67. Zhou, X., Kao, M. C. & Wong, W. H. Transitive functional annotation by shortest-path analysis of gene expression data. *Proc. Natl Acad. Sci. USA* **99**, 12783–12788 (2002).
68. Krischer, J. The Environmental Determinants of Diabetes in the Young (TEDDY) (version 35) [Dataset]. *NIDDK Central Repository* <https://doi.org/10.58020/y3jk-x087> (2026).
69. Kerminen, S. et al. Fine-scale genetic structure in Finland. *G3* **7**, 3459–3468 (2017).

Acknowledgements

We especially acknowledge the TEDDY families for their continued participation in this wonderful study. The computations in this paper were run in part on the FASRC Cannon cluster supported by the FAS Division of Science Research Computing Group at Harvard University.

Author contributions

D.D., E.A.F., C.H. and D.D.W. conceptualized the study. D.D. and A.M.W. performed the data analysis. D.D., E.A.F., C.H. and D.D.W. drafted the paper. T.V. and K.V. managed the data transfer. G.W. processed the data. E.A.F., C.H. and D.D.W. supervised the study. C.H. collected the samples and data and obtained funding. D.D., A.M.W., T.V., G.W., M.K., M.J.S., K.V., E.A.F., C.H. and D.D.W. discussed the results, critically reviewed the text and approved the final paper. Members of the TEDDY Study Group contributed to study design, participant recruitment and follow-up, sample and data collection, laboratory analysis and data curation, and reviewed and approved the final paper.

Funding

The TEDDY study 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, U01 DK124166, U01 DK128847 and contract no. HHSN267200700014C from the NIDDK, National Institute of Allergy and Infectious Diseases (NIAID), Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institute of Environmental Health Sciences, Centers for Disease Control and Prevention and Breakthrough T1D (formerly JDRF). This work is supported in part by the NIH/National Center for Advancing Translational Sciences (NCATS) Clinical and Translational Science Awards to the University of Florida (UL1 TR000064) and the University of Colorado (UL1 TR002535). This manuscript is the result of funding from the NIH and is subject to the NIH Public Access Policy. Through acceptance of this federal funding, NIH has been given a right to make this manuscript publicly available in PubMed Central upon the Official Date of Publication, as defined by NIH. D.D.W. was supported by funding from NIH (RO0DK119412,

R01NR019992, R01AG077489, RF1AG083764, U54AG089325 and R01DK149250). C.H. was supported by funding from NIAID U19AI110820 (PIs V. M. Bruno and D. A. Rasko). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Competing interests

C.H. is a member of the scientific advisory board for Zoe Nutrition, Empress Therapeutics and Seres Therapeutics. All other authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s42255-026-01614-9>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s42255-026-01614-9>.

Correspondence and requests for materials should be addressed to Eric A. Franzosa, Curtis Huttenhower or Dong D. Wang.

Peer review information *Nature Metabolism* thanks the anonymous reviewers for their contribution to the peer review of this work.

Primary Handling Editor: Jean Nakhle, in collaboration with the *Nature Metabolism* team. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026

The TEDDY Study Group

Colorado Clinical Center

Marian Rewers¹¹, Kimberly Bautista¹¹, Judith Baxter¹¹, Daniel Felipe-Morales¹¹, Brigitte I. Frohnert¹¹, Marisa Stahl¹¹, Patricia Gesualdo¹¹, Michelle Hoffman¹¹, Randi Johnson¹¹, Rachel Karban¹¹, Edwin Liu¹¹, Jill Norris¹¹, Holly O'Donnell¹¹, Andrea Steck¹¹ & Kathleen Waugh¹¹

Finland Clinical Center

Jorma Toppari^{12,13}, Olli G. Simell¹², Annika Adamsson¹³, Suvi Ahonen^{14,15,16}, Mari Åkerlund^{14,15,16}, Sirpa Anttila^{17,18}, Leena Hakola^{14,15}, Sanni Heikura^{17,18}, Tiia Honkanen^{17,18}, Heikki Hyöty^{14,15}, Jorma Ilonen¹², Saori Itoshima^{12,13}, Sanna Jokipuu¹³, Taru Karjalainen^{17,18}, Leena Karlsson¹³, Pieta Kempainen^{17,18}, Jukka Kero^{12,13}, Marika Korpela^{17,18}, Jaakko J. Koskeniemi^{12,13}, Miia Kähönen^{17,18}, Mikael Knip^{14,15}, Minna-Liisa Koivikko^{17,18}, Katja Kokkonen^{14,15}, Merja Koskinen^{14,15}, Mirva Koreasalo^{14,15,16}, Kalle Kurppa^{14,15}, Salla Kuusela^{17,18}, Jarita Kytölä^{14,15}, Jutta Laiho¹⁴, Tiina Latva-aho^{17,18}, Siiri Leisku^{14,15}, Laura Leppänen¹³, Katri Lindfors¹⁴, Maria Lönnrot^{14,15}, Elina Mäntymäki¹³, Markus Mattila^{14,15}, Maija E. Miettinen¹⁶, Teija Mykkänen^{17,18}, Tiina Niininen^{14,15}, Sari Niinistö¹⁶, Noora Nurminen^{14,15}, Sami Oikarinen^{14,15}, Hanna-Leena Oinas^{14,15}, Paula Ollikainen^{17,18}, Zhian Othmani¹², Sirpa Pohjola^{17,18}, Jenna Rautanen¹⁶, Mia Rein^{17,18}, Minna Romo¹³, Juulia Rönkä^{17,18}, Nelli Rönkä^{17,18}, Noora Ruotsalainen^{17,18}, Satu Simell¹², Päivi Tossavainen^{17,18}, Mari Vähä-Mäkilä¹², Eeva Varjonen¹³, Riitta Veijola^{17,18}, Irene Viinikangas^{17,18} & Suvi M. Virtanen^{14,15,16}

Georgia/Florida Clinical Center

Richard McIndoe¹⁹, Desmond Schatz²⁰, Diane Hopkins¹⁹, Michael Haller²⁰, Melissa Gardiner¹⁹, Ashok Sharma¹⁹, Laura Jacobsen²⁰, Percy Gordon¹⁹ & Jennifer Hosford²⁰

Germany Clinical Center

Anette G. Ziegler²¹, Ezio Bonifacio²², Cigdem Sanverdi²¹, Anja Heublein²¹, Sandra Hummel²¹, Annette Knopff²¹, Melanie Köger²¹, Sibylle Koletzko²³, Claudia Ramminger²¹, Roswith Roth²¹, Jennifer Schmidt²¹, Marlon Scholz²¹, Joanna Stock²¹, Katharina Warncke²¹, Lorena Müller²¹ & Christiane Winkler²¹

Sweden Clinical Center

Åke Lernmark²⁴, Daniel Agardh²⁴, Carin Andrén Aronsson²⁴, Rasmus Bennet²⁴, Corrado Cilio²⁴, Susanne Dahlberg²⁴, Malin Goldman Tsubarah²⁴, Emelie Ericson-Hallström²⁴, Lina Fransson²⁴, Emina Halilovic²⁴, Susanne Hyberg²⁴, Berglind Jonsdottir²⁴, Naghmeh Karimi²⁴, Helena Elding Larsson²⁴, Markus Lundgren²⁴, Jessica Melin²⁴,

Kobra Rahmati²⁴, Anita Ramelius²⁴, Falastin Salami²⁴, Anette Sjöberg²⁴, Evelyn Tekum Amboh²⁴, Carina Törn²⁴, Terese Wiktorsson²⁴, Eva Andersson²⁴, Marie Andersson Turpeinen²⁴, Rawya Antar²⁴, Maria Ask²⁴, Jenny Bremer²⁴, Sylvia Bianconi Svensson²⁴, Ulla-Marie Carlsson²⁴, Magdalena Delikat Kulinski²⁴, Annika Fors²⁴, Ulla Fält²⁴, Thomas Gard²⁴, Joanna Gerardsson²⁴, Monika Hansen²⁴, Anna Hansson²⁴, Carina Hansson²⁴, Gertie Hansson²⁴, Elin M. Hård af Segerstad²⁴, Hanna Jisser²⁴, Fredrik Johansen²⁴, Linda Jonsson²⁴, Silvija Jovic²⁴, Sigrid Lenrick Forss²⁴, Barbro Lernmark²⁴, Marielle Lindström²⁴, Maria Markan²⁴, Theodosia Massadakis²⁴, Marlena Maziarz²⁴, Zeliha Mestan²⁴, Maria Månsson Martinez²⁴, Caroline Nilsson²⁴, Emma Nilsson²⁴, Yohanna Nordh²⁴, Karin Ottosson²⁴, Sara Rang²⁴, Anna Rosenquist²⁴, Monika Sedig Järvirova²⁴, Sara Sibthorpe²⁴, Birgitta Sjöberg²⁴, Ulrika Swartling²⁴, Erika Trulsson²⁴, Ulrika Ulvenhag²⁴, Anne Wallin²⁴, Ingrid Wigheden²⁴, Åsa Wimar²⁴ & Sofie Åberg²⁴

Washington Clinical Center

William A. Hagopian²⁵, Michael Killian²⁶, Claire Cowen Crouch²⁶, Jennifer Skidmore²⁶, Ben Kim²⁶, Cody McCall²⁶, Arlene Meyer²⁶, Jared Radtke²⁶ & Shreya Roy²⁶

Pennsylvania Satellite Center

Dorothy Becker²⁷, Margaret Franciscus²⁷, MaryEllen Dalmagro-Elias Smith²⁷, Ashi Daftary²⁷, Mary Beth Klein²⁷ & Chrystal Yates²⁷

Data Coordinating Center

Jeffrey P. Krischer²⁸, Rajesh Adusumali²⁸, Sarah Austin-Gonzalez²⁸, Maryouri Avendano²⁸, Sandra Baethke²⁸, Brant Burkhardt²⁸, Martha Butterworth²⁸, Nicholas Cadigan²⁸, Joanna Clasen²⁸, Kevin Counts²⁸, Laura Gandolfo²⁸, Jennifer Garmeson²⁸, Veena Gowda²⁸, Shu Liu²⁸, Xiang Liu²⁸, Kristian Lynch²⁸, Jamie Malloy²⁸, Lazarus Mramba²⁸, Cristina McCarthy²⁸, Hemang M. Parikh²⁸, Cassandra Remedios²⁸, Chris Shaffer²⁸, Susan Smith²⁸, Noah Sulman²⁸, Roy Tamura²⁸, Dena Tewey²⁸, Henri Thuma²⁸, Michael Toth²⁸, Ulla Uusitalo²⁸, Kendra Vehik¹⁰, Ponni Vijayakandipan²⁸, Melissa Wroble²⁸, Jimin Yang²⁸, Kenneth Young²⁸, Michael Abbondandolo²⁸, Lori Ballard²⁸, Rasheedah Brown²⁸, David Cuthbertson²⁸, Stephen Dankyi²⁸, Christopher Eberhard²⁸, Steven Fiske²⁸, David Hadley²⁸, Kathleen Heyman²⁸, Belinda Hsiao²⁸, Christina Karges²⁸, Francisco Perez Laras²⁸, Hye-Seung Lee²⁸, Qian Li²⁸, Colleen Maguire²⁸, Wendy McLeod²⁸, Aubrie Merrell²⁸, Steven Meulemans²⁸, Jose Moreno²⁸, Ryan Quigley²⁸ & Laura Smith²⁸

Autoantibody Reference Laboratories

Liping Yu¹¹, Dongmei Miao¹¹, Kathleen Gillespie²⁹, Kyla Chandler²⁹, Olivia Pearce²⁹, Sarah Stollery²⁹, Elinor Balch²⁹, Hanah Batholomew²⁹ & Zahra Hashmi²⁹

Genetics Laboratory

Stephen S. Rich³⁰, Wei-Min Chen³⁰, Suna Onengut-Gumuscu³⁰, Emily Farber³⁰, Rebecca Roche Pickin³⁰, Jonathan Davis³⁰, Jordan Davis³⁰, Dan Gallo³⁰, Jessica Bonnie³⁰ & Paul Campolieto³⁰

HLA Reference Laboratory

William Hagopian²⁶, Jared Radtke²⁶, Henry Erlich³¹, Steven J. Mack³¹ & Anna Lisa Fear³¹

Metagenomics and Microbiome Laboratory

Joseph F. Petrosino³², Richard E. Lloyd³², Nadim J. Ajami³², Tulin Ayvaz³², Tonya Bauch³², Harshavardhan Doddapaneni³³, Richard Gibbs³³, Rachel Hirsch³², Kristi L. Hoffman³², Hanna Houston³², Diane S. Hutchinson³², Sara Javornik Cregeen³², Elizabeth Laval³², Lenka Kusic³², Ginger Metcalf³³, Hannah Moreno³², Donna Muzny³³, Jacqueline L. O'Brien³², Katelyn Payne³², Ashlie Renner³², Lauren Riley³², Matthew C. Ross³², Daniel P. Smith³², Anil Surathu³², Auriole Tamegnon³², Xiangjun Tian³², Michael Tisza³², Nguyen Truong³², Li Wang³², Matthew C. Wong³² & Eduardo Zavala³²

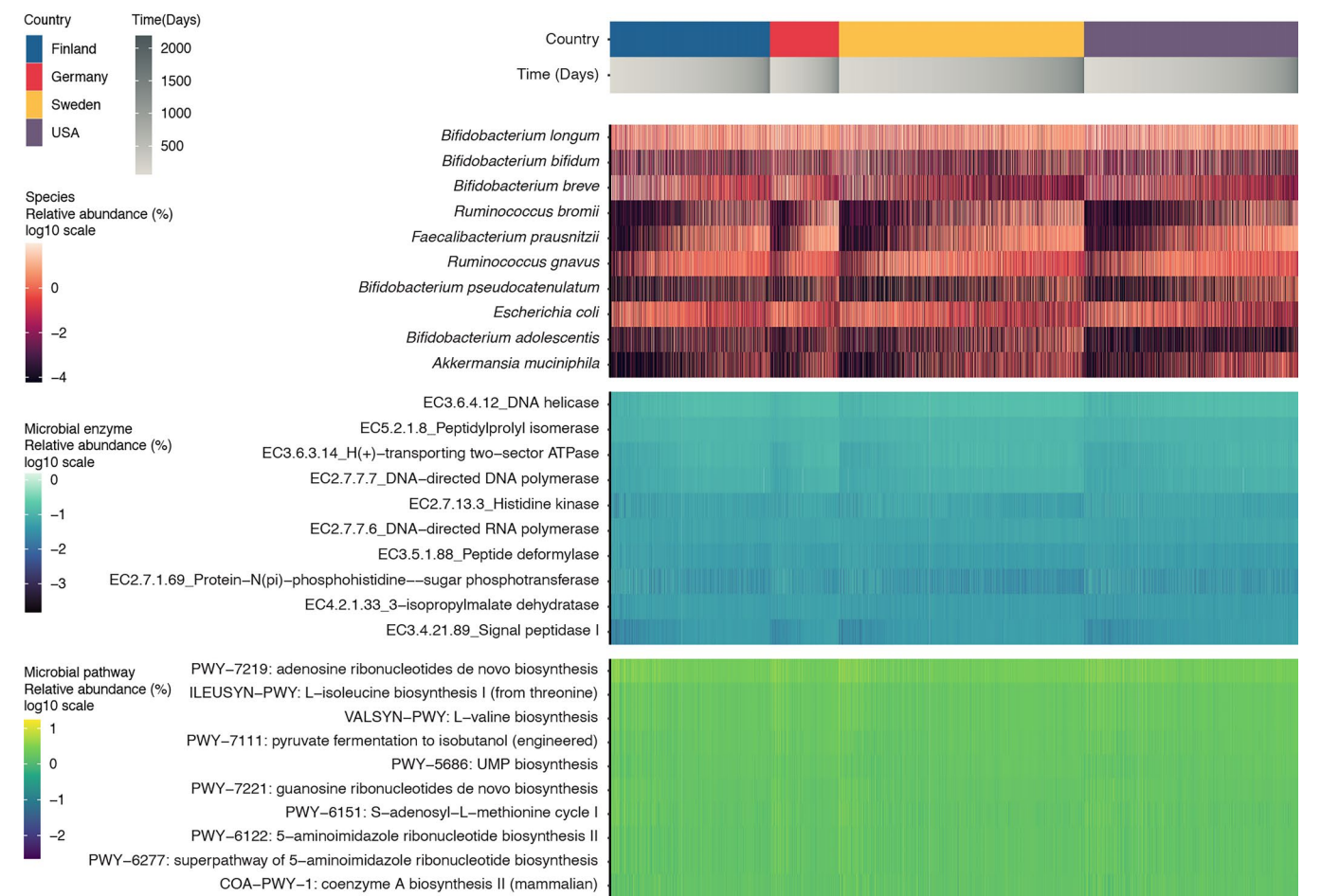
Repository

Chris Deigan³⁴, Ricky Schrock³⁴, Polina Malone³⁴, Sandra Ke³⁴ & Niveen Mulholland³⁴

Project Scientist**Beena Akolkar³⁵**

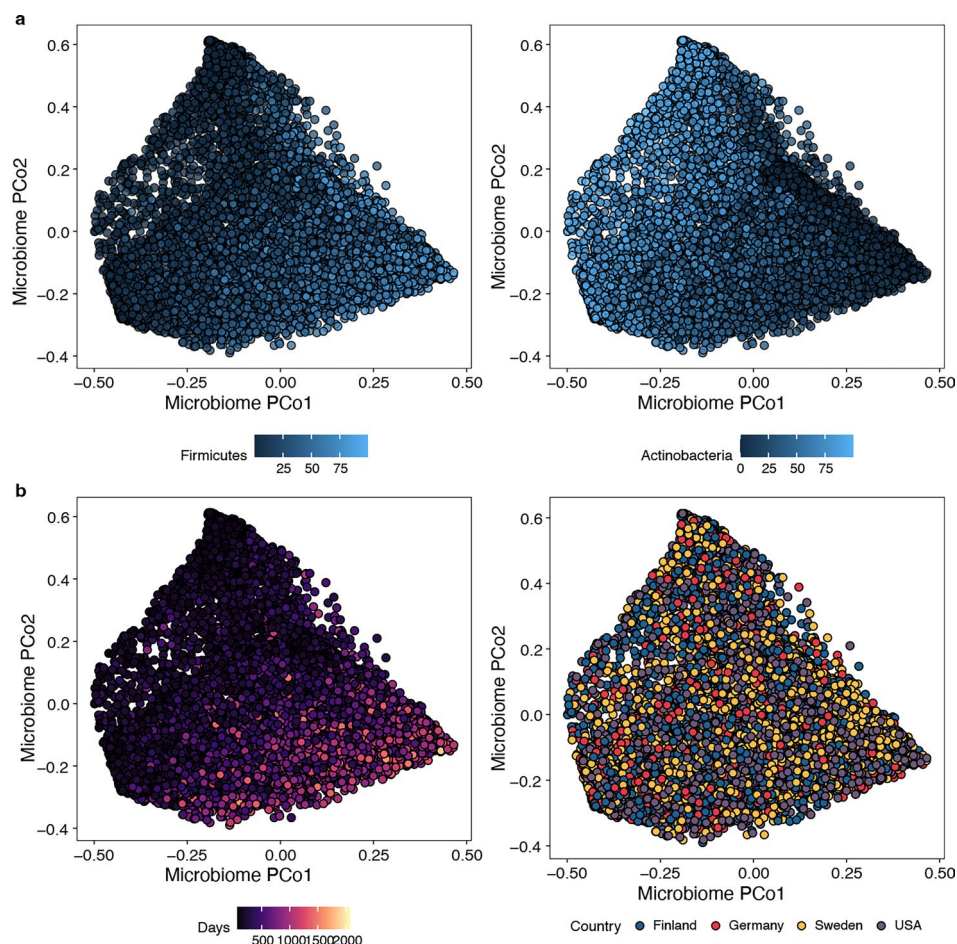
Other contributors**Thomas Bries³⁶, Todd Brusko³⁷, Teresa Buckner³⁸, Suzanne Bennett Johnson³⁹, Eoin McKinney⁴⁰, Tomi Pastinen⁴¹, Steffen Ullitz Thorsen^{42,43} & Eric Triplett³⁷**

¹¹Barbara Davis Center for Childhood Diabetes, University of Colorado Anschutz Medical Campus, Aurora, CO, USA. ¹²University of Turku, Turku, Finland. ¹³Turku University Hospital, Wellbeing Services County of Southwest Finland, Turku, Finland. ¹⁴Tampere University, Tampere, Finland. ¹⁵Tampere University Hospital, Wellbeing Services County of Pirkanmaa, Tampere, Finland. ¹⁶Finnish Institute for Health and Welfare, Helsinki, Finland. ¹⁷University of Oulu, Oulu, Finland. ¹⁸Oulu University Hospital, Wellbeing Services County of North Ostrobothnia, Oulu, Finland. ¹⁹Center for Biotechnology and Genomic Medicine, Augusta University, Augusta, GA, USA. ²⁰Pediatric Endocrinology, University of Florida, Gainesville, FL, USA. ²¹Forschergruppe Diabetes and Institute of Diabetes Research, Helmholtz Zentrum München, and Klinikum rechts der Isar, Technische Universität München, Neuherberg, Germany. ²²Center for Regenerative Therapies, TU Dresden, Dresden, Germany. ²³Department of Gastroenterology, Dr. von Hauner Children's Hospital, Ludwig Maximilians University Munich, Munich, Germany. ²⁴Lund University, Lund, Sweden. ²⁵Indiana University, Indianapolis, IN, USA. ²⁶Pacific Northwest Research Institute, Seattle, WA, USA. ²⁷Children's Hospital of Pittsburgh of UPMC, Pittsburgh, PA, USA. ²⁸University of South Florida, Tampa, FL, USA. ²⁹Bristol Medical School, University of Bristol, Bristol, UK. ³⁰Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA. ³¹Center for Genetics, Children's Hospital Oakland Research Institute, Oakland, CA, USA. ³²Alkek Center for Metagenomics and Microbiome Research, Department of Molecular Virology and Microbiology, Baylor College of Medicine, Houston, TX, USA. ³³Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA. ³⁴NIDDK Biosample Repository at Fisher BioServices, Rockville, MD, USA. ³⁵National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD, USA. ³⁶Columbia University, New York, NY, USA. ³⁷University of Florida, Gainesville, FL, USA. ³⁸University of Northern Colorado, Greeley, CO, USA. ³⁹Florida State University, Tallahassee, FL, USA. ⁴⁰University of Cambridge, Cambridge, UK. ⁴¹The Children's Mercy Hospital, Kansas City, MO, USA. ⁴²Department of Clinical Immunology, University of Copenhagen, Copenhagen, Denmark. ⁴³Department of Pediatrics and Adolescents, Copenhagen University Hospital, Herlev, Denmark.



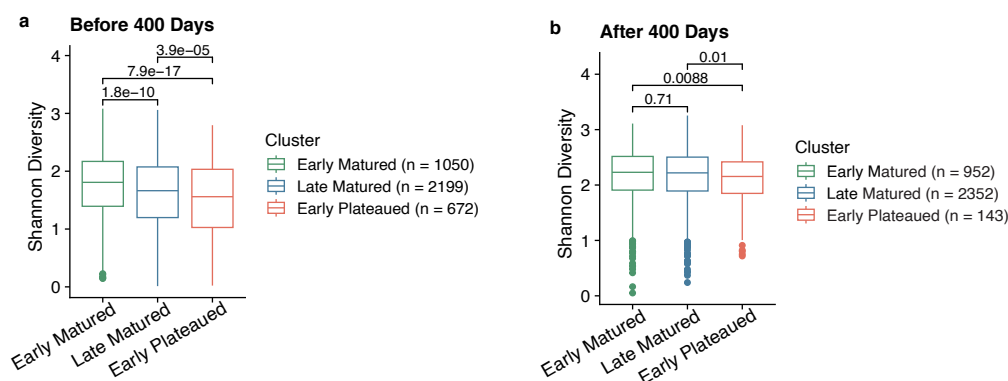
Extended Data Fig. 1 | Data distribution for taxonomic and functional profiles of the gut microbiome. Distributions of the 10 most abundant microbial species, microbial enzymes, and biochemical pathways in the analysed metagenomes are shown at the top, middle, and bottom, respectively. The analysis included 12,151 metagenomic samples. The samples are ordered by country of origin and age.

Polysaccharide degraders and butyrate producers are among the most abundant gut microbes. Enzymes involved in DNA replication, energy production, and protein folding are highly abundant. Nucleotide biosynthesis and amino acid biosynthesis pathways are also highly abundant.



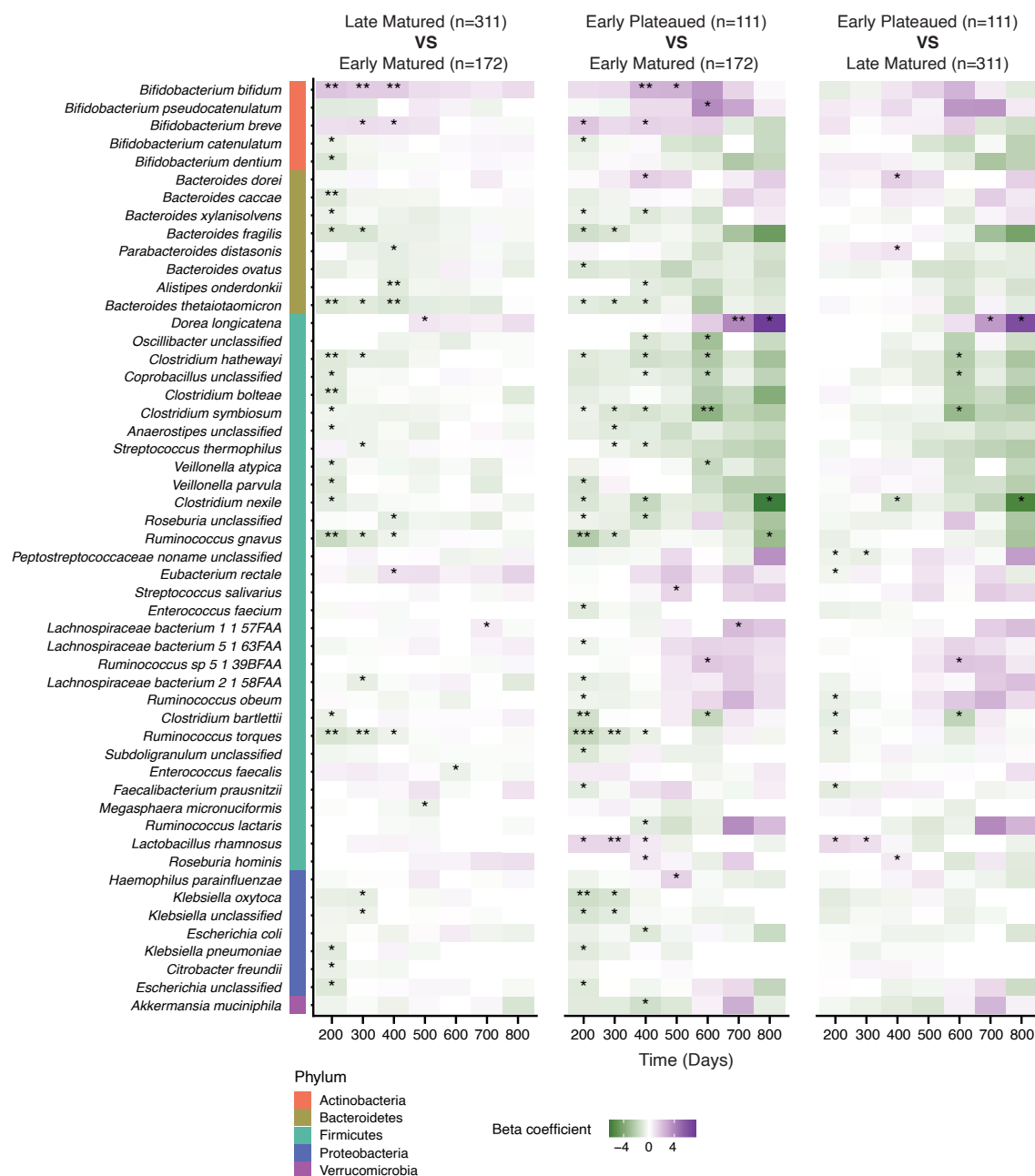
Extended Data Fig. 2 | Principal coordinate analysis of all the microbiome samples based on species-level Bray–Curtis dissimilarity. (a) Principal coordinate analysis of species-level Bray–Curtis dissimilarity coloured in correspondence to the relative abundance of Actinobacteria and Firmicutes phyla. As expected, a majority of variation in the species-level compositional structure of the gut microbiome was driven by a tradeoff between Actinobacteria versus Firmicutes phyla. (b) Age accounted for the largest percentage of variance

($R^2 = 5.90\%$). Samples are coloured by age and country at the time of sampling. R^2 values were calculated using permutational multivariate analysis of variance (PERMANOVA; 999 permutations) and indicate the percentage of variance in the overall microbiome community attributable to age, and geographic location. Each point indicates one metagenomic sample from a genotyped participant who passed genetic quality control ($n = 10,851$).



Extended Data Fig. 3 | The diversity of the gut microbial communities across different microbiome maturational patterns during early and late follow-ups. Samples with an Early Matured pattern show higher microbiome diversity during the first 400 days, whereas those with an Early Plateaued pattern exhibit lower diversity beyond 400 days. We assessed α -diversity using the Shannon Diversity Index. Box plot centre lines represent the medians of the Shannon Diversity Index for different microbiome maturational patterns during the first 400 days of follow-up (a) and beyond 400 days (b). The boxes show the interquartile ranges

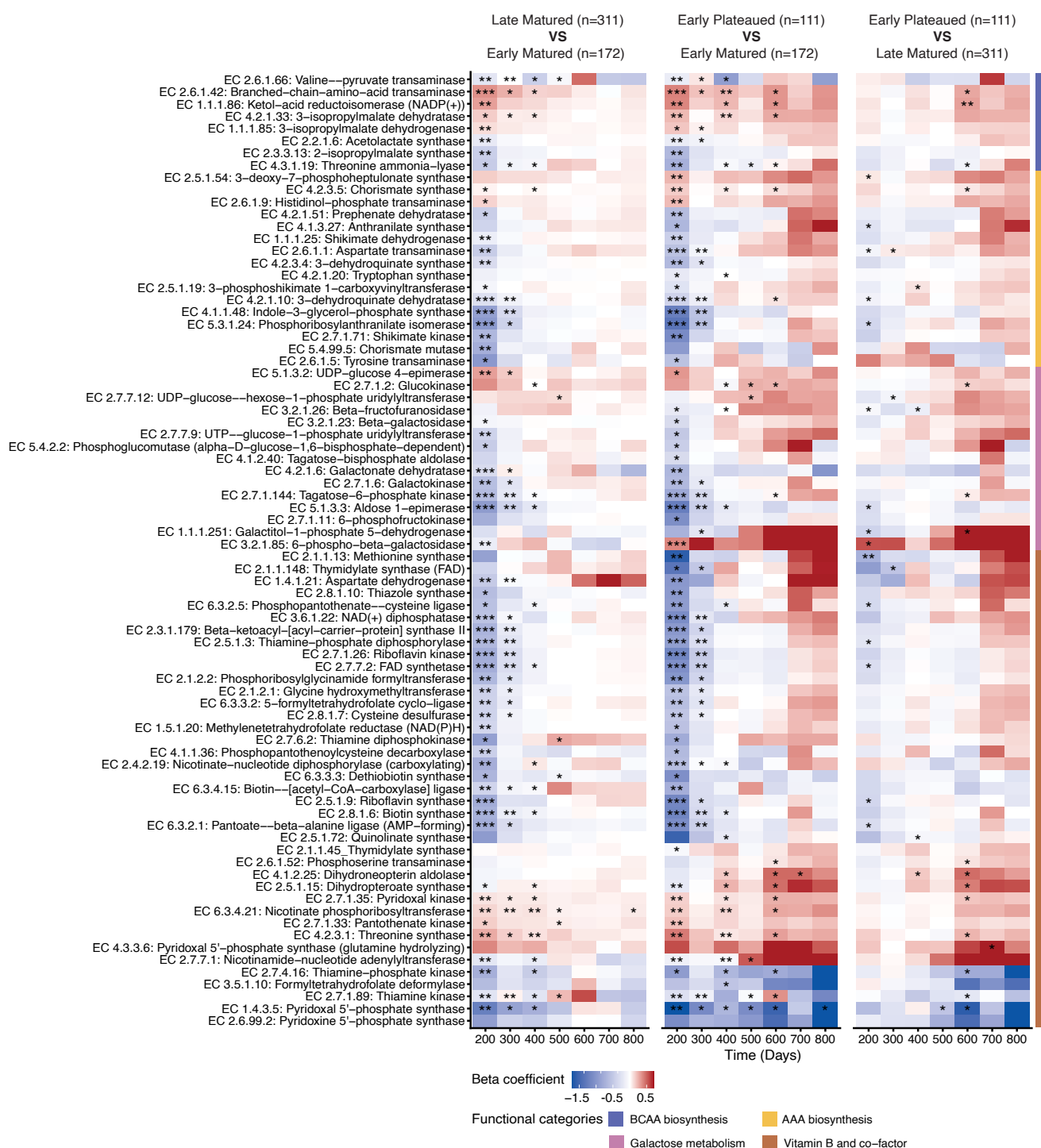
(IQRs; 25th–75th percentiles), and whiskers extend to the most extreme values within 1.5 times the IQR from the lower and upper quartiles, respectively. Points beyond the whiskers represent outliers. *P* values were calculated using two-sided *t*-test, pairwise comparisons included Early Matured vs Late Matured (before 400 days: $P = 1.8 \times 10^{-10}$; after 400 days: $P = 0.71$), Early Matured vs Early Plateaued (before 400 days: $P = 7.9 \times 10^{-17}$; after 400 days: $P = 0.0088$), and Late Matured vs Early Plateaued (before 400 days: $P = 3.9 \times 10^{-5}$; after 400 days: $P = 0.01$). The *n* indicates the number of longitudinal stool samples in each group.



Extended Data Fig. 4 | Differentially abundant microbial species across different maturational patterns during the first 800 days of follow-up.

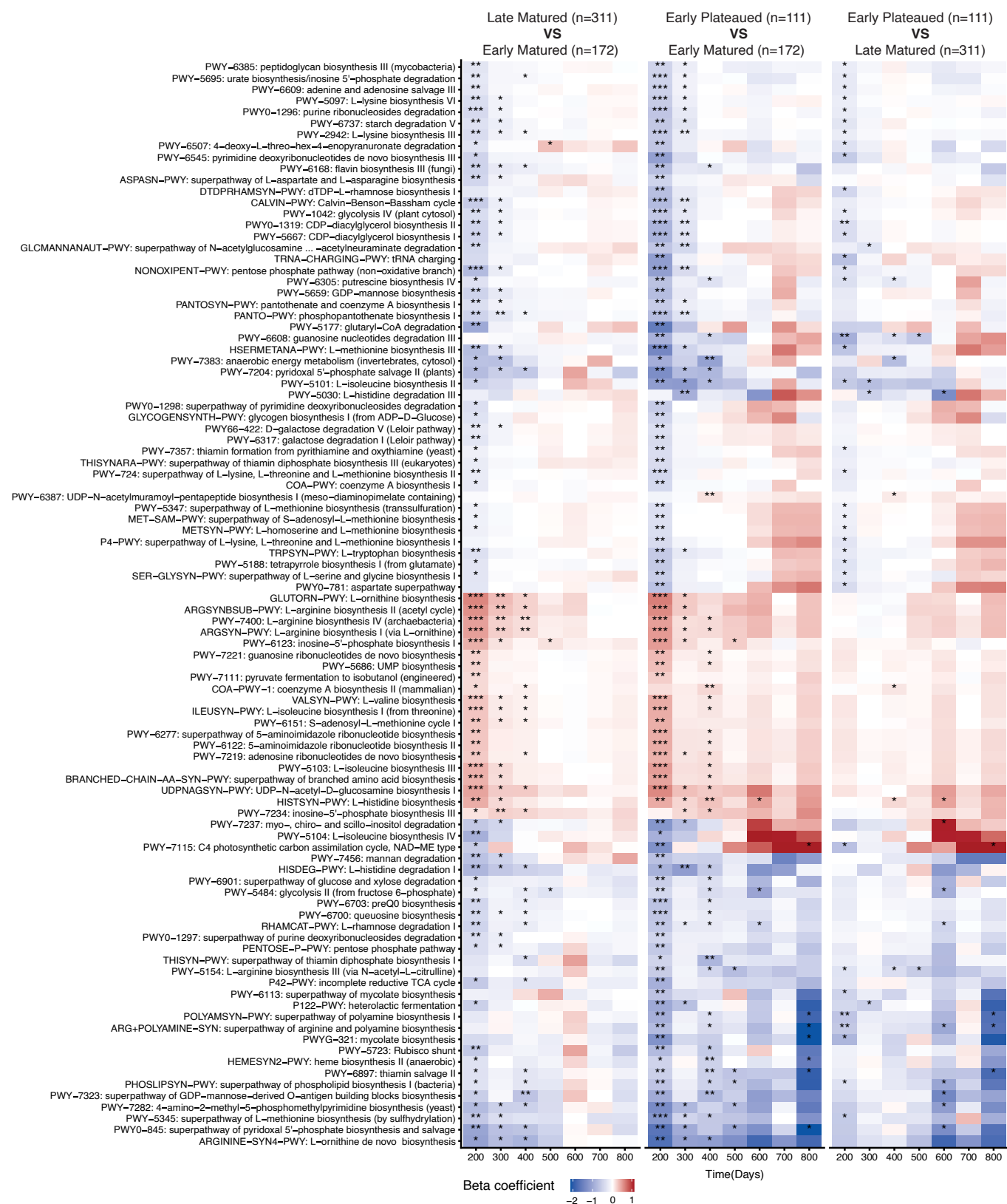
The colours of the heatmap represent the beta coefficients for the comparison between two maturational patterns derived from linear mixed models in MaAsLin 2, with the relative abundance of each microbial species as the dependent variable. The heatmap displays microbial features that were differentially abundant (q -value < 0.25) between maturational patterns at least once during

the follow-up. Each model accounted for the participant's identifier as a random effect and was adjusted for age, clinical centre, sex, delivery mode, antibiotic use, probiotic use, solid food introduction and breastfeeding duration as fixed effects. To correct for multiple comparisons, we applied the Benjamini–Hochberg method to calculate q -values (false discovery rate-adjusted P values). Significance levels are indicated as follows: * $q < 0.25$, ** $q < 0.05$, *** $q < 0.001$.



Extended Data Fig. 5 | Differentially abundant microbial enzymes involved in the biosynthesis of aromatic amino acids, branched-chain amino acids, and the metabolism of vitamin B and galactose maturational patterns during the first 800 days of follow-up. The colours of the heatmap represent the beta coefficients for the comparison between two maturational patterns derived from linear mixed models in MaAsLin2, with the relative abundance of each microbial enzyme as the dependent variable. The heatmap displays microbial features, which were involved in the biosynthesis of aromatic amino acids, branched-

chain amino acids, and the metabolism of vitamin B and galactose, and were differentially abundant (q -value < 0.25) between maturational patterns at least once during the follow-up. Each model accounted for the participant's identifier as a random effect and was adjusted for age, clinical centre, sex, delivery mode, antibiotic use, probiotic use, solid food introduction and breastfeeding duration as fixed effects. To correct for multiple comparisons, we applied the Benjamini-Hochberg method to calculate q -values (false discovery rate-adjusted P values). Significance levels are indicated as follows: * $q < 0.25$, ** $q < 0.05$, *** $q < 0.001$.



Extended Data Fig. 6 | Differentially abundant microbial biochemical pathway across maturational patterns during the first 800 days of follow-up. The colours of the heatmap represent the beta coefficients for the comparison between two maturational patterns derived from linear mixed models in MaAsLin 2, with the relative abundance of each microbial biochemical pathway as the dependent variable. The heatmap displays microbial biochemical pathways that were differentially abundant (q -value < 0.05) between maturational patterns

at least once during the follow-up. Each model accounted for the participant's identifier as a random effect and was adjusted for age, clinical centre, sex, delivery mode, antibiotic use, probiotic use, solid food introduction and breastfeeding duration as fixed effects. To correct for multiple comparisons, we applied the Benjamini-Hochberg method to calculate q -values (false discovery rate-adjusted P values). Significance levels are indicated as follows: * $q < 0.05$, ** $q < 0.01$, *** $q < 0.001$.

Reporting Summary

Nature Portfolio 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 Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- | | |
|-------------------------------------|--|
| n/a | Confirmed |
| ☐ | ☒ The exact sample size (<i>n</i>) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☐ | ☒ A statement on 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
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| ☐ | ☒ 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 statistical parameters 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. <i>F</i> , <i>t</i> , <i>r</i>) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted
<i>Give P values as exact values whenever suitable.</i> |
| ☒ | ☐ 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 <i>d</i> , Pearson's <i>r</i>), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	No software was used for data collection.
Data analysis	Metagenomic sequencing data were processed by 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. Genetic data were processed using KING (v2.3.0) and PLINK (v1.90b7). All the data analysis and plotting were performed using R v4.2.3 (Main package: vegan v2.6-4, Maaslin2 v1.12.0, ggplot2 v3.4.2, lmerTest v3.1-3, traj v2.0.1, survival v3.5-5) The analysis-specific programs are publicly available through https://github.com/biobakery/teddy_genetics

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 and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

TEDDY microbiome whole-genome sequencing and SNP data that support the findings of this study are available in the NCBI database of Genotypes and Phenotypes (dbGaP) under the primary accession code phs001442.v4.p3, in accordance with the dbGaP controlled-access authorization process. Data from The Environmental Determinants of Diabetes in the Young (<https://doi.org/10.58020/y3jk-x087>) reported here are available upon request through the NIDDK Central Repository (NIDDK-CR) website, Resources for Research (R4R), <https://repository.niddk.nih.gov/>. Processed taxonomic profiles and metadata from the DIABIMMUNE three country cohort are available through the DIABIMMUNE website (<https://diabimmune.broadinstitute.org/diabimmune>).

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender

For the TEDDY cohort, sex was recorded at enrolment based on parental report for all participants. Gender was not assessed in this study. The study included 403 (45.4%) female and 484 (54.6%) male participants. Sex was included as a covariate in all analyses to account for potential sex-related influences on the outcomes. In addition, sex-stratified analyses of the primary findings yielded results consistent with the main analyses.

For the DIABIMMUNE three-country cohort, which was used as a secondary replication cohort specifically for the microbiome maturation pattern analysis, sex was recorded as reported in the original study and was not used as a stratification variable.

Reporting on race, ethnicity, or other socially relevant groupings

The study population are 887 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). For the DIABIMMUNE cohort, race and ethnicity were not systematically collected or used in the replication analysis.

Population characteristics

The study included children from the TEDDY and DIABIMMUNE cohorts. TEDDY participants were infants and young children with increased genetic risk for type 1 diabetes based on HLA genotype, including children from the general population and those with a first-degree relative with type 1 diabetes. DIABIMMUNE participants were followed from birth to three years of age and carried HLA alleles associated with increased risk of autoimmunity.

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 TEDDY study. The DIABIMMUNE three-country cohort recruited newborn infants from Finland, Estonia, and Russia with increased HLA-conferred genetic susceptibility. The recruitment procedures for the published datasets were described in the corresponding papers.

Ethics oversight

The 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 an External Advisory Board formed by the National Institutes of Health. The DIABIMMUNE study was conducted according to the guidelines laid down in the Declaration of Helsinki, and all procedures involving human subjects were approved by the local ethical committees of the participating hospitals. The parents and/or study subjects gave their written informed voluntary consent prior to sample collection.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is 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/documents/nr-reporting-summary-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 = 12,151 stool samples) from subjects in TEDDY islet autoimmunity and type 1 diabetes case-control cohorts are being analyzed for maximal power. This includes all participants with islet autoimmunity or type 1 diabetes and their

matched controls in TEDDY study as of May 31, 2012 (N = 887 participants; host genetic data available for 877 participants). For the DIABIMMUNE three-country cohort, 785 stool metagenomic samples from 212 infants and young children were included for the microbiome maturation pattern analysis.

Data exclusions	The exclusion criteria were pre-specified. At the sample level, metagenomes failing quality control were excluded before analysis. At the feature level, microbial features were retained only if they were present in at least 10% of samples, with minimum relative abundance thresholds of 0.0002 for species and pathways and 0.00002 for level-4 EC categories; these filters were specified in advance to avoid spurious associations arising from sparsely detected features. For genetic analyses, sample-level and SNP-level quality control was performed using KING --autoQC and PLINK, and individuals identified as genetic outliers by principal component analysis were removed. For trajectory clustering analyses, individuals with insufficient longitudinal sampling (<4 metagenomic samples) were excluded because reliable developmental patterns could not be inferred.
Replication	Replication of the microbiome maturation pattern was performed in an independent birth cohort (DIABIMMUNE). Other analyses were conducted in the primary TEDDY cohort only.
Randomization	Controls were matched individually to cases as described in detail in the manuscript text. Cases were sampled until diagnosis of T1D and matched control samples were included up until the corresponding day of life.
Blinding	No blinding used, TEDDY is an observational follow-up study

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

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

Clinical data

Policy information about [clinical studies](#)

All manuscripts should comply with the ICMJE [guidelines for publication of clinical research](#) and a completed [CONSORT checklist](#) must be included with all submissions.

Clinical trial registration	NCT00279318
Study protocol	https://repository.niddk.nih.gov/studies/teddy/
Data collection	The TEDDY study involves six clinical centers across four countries: the United States (Colorado, Georgia/Florida, and Washington), Finland, Sweden, and Germany. Participant recruitment took place between September 2004 and February 2010, enrolling 8,676 infants aged 3–4 months who were identified through newborn genetic screening for HLA genotypes associated with increased risk for type 1 diabetes (T1D). Blood samples were collected every three months until participants reached the age of four, transitioning to biannual collections thereafter. Stool samples were collected monthly from three months to 48 months of age, then every three months until 10 years of age, and biannually thereafter. Stool sample collection ceased for all participants in August 2018
Outcomes	Appearance of one or more islet cell autoantibodies: GADA, IAA, or IA-2A confirmed at two consecutive visits. And development of Type 1 Diabetes

Plants

Seed stocks	Not applicable
Novel plant genotypes	Not applicable
Authentication	Not applicable