npj | biofilms and microbiomes

Article

Published in partnership with Nanyang Technological University

https://doi.org/10.1038/s41522-025-00868-7

Compiling an early life human gut microbiome atlas and identification of key microbial drivers

Chiara Tarracchini[1], Giulia Longhi[1], Emma Gennaioli[1], Aryanna Muscò[1], Sonia Mirjam Rizzo[1], Alice Viappiani[2], Salvatore Giovanni Vitale[3], Leonardo Mancabelli[4,5], Gabriele Andrea Lugli[1,5], Stefano Angioni[3], Francesca Turroni[1,5], Douwe van Sinderen[6], Christian Milani[1,5] & Marco Ventura[1,5] ✉

During the first year after birth, the infant gut microbiome undergoes a rapid and profound compositional and functional transformation, impelled by an intricate network of intrinsic and extrinsic factors. This process results in increased taxonomic and functional diversification, alongside greater interindividual variability. To better understand this early-life ecosystem, this study assessed the interindividual variability of the infant gut microbiome using a comprehensive infant gut microbiome database of 5288 fecal metagenomic data from healthy, full-term infants across various geographical locations. Our study identified six reference microbial communities, termed Early-Life Community State Types (ELi-CSTs), which not only capture specific compositional profiles and heterogeneity of the infant gut microbiome, but also record the extensive transformation experienced by this developing microbial community during the first year of human life. Indicative Species analysis and Random Forest modeling assisted the precise identification of unique, key taxonomic signatures that are critical to the structure of each ELi-CST, highlighting microbial taxa with pivotal roles in shaping the infant gut microbiota. To complement these findings, we established a bacterial biobank through dedicated cultivation efforts of the infant microbiota, comprising 182 genome-sequenced isolates corresponding to key taxa involved in early life gut microbiota assembly. This biobank provided the basis for co-cultivation experiments combined with transcriptome analyses, thereby enabling in vitro investigations into microbial cross-talk among key modulators, and yielding novel insights into the molecular interactions and cooperative dynamics behind early microbiome development.

Although several studies have controversially suggested that assembly of the gut microbiota commences *in utero*[1,2], it is now well established that formation of this intestinal community initiates in earnest at birth, progressively developing as it adapts and co-evolves with its human host, ultimately maturing into the complex and stable ecosystem characteristic of adulthood[3,4].

Besides well-established perinatal factors, such as mode of delivery (vaginal or cesarean) and gestational age at birth (term or preterm), which influence the initial colonization of the infant gut[5,6], other key elements come into play during the first year of life. These factors include antibiotic usage, feeding practices (breastfeeding or formula feeding), and the introduction of solid foods during the weaning period[7–11]. In particular, the weaning stage marks a significant microbiota developmental occurrence, as the transition from an exclusive milk-based diet to the introduction of solid foods fundamentally reshapes the nutritional landscape of the gut[12–14]. However, this transition is likely influenced by factors such as timing, cultural practices, and local dietary habits.

The interplay of all these factors, occurring within the relatively short timeframe of the first twelve months following birth, is the driving force that causes increased gut microbial diversification and enhanced interindividual divergence. To capture the complexity and dynamics of this early-life state, several publications have reported on the classification of the gut microbiota

[1]Laboratory of Probiogenomics, Dept. Chemistry, Life Sciences and Environmental Sustainability, University of Parma, Parma, Italy. [2]GenProbio srl, Parma, Italy. [3]Division of Gynecology and Obstetrics, Department of Surgical Sciences, University of Cagliari, Cagliari, Italy. [4]Dept. Medicine and Surgery, University of Parma, Parma, Italy. [5]Microbiome Research Hub, University of Parma, Parma, Italy. [6]APC Microbiome Institute and School of Microbiology, Bioscience Institute, National University of Ireland, Cork, Ireland. ✉e-mail: marco.ventura@unipr.it

into distinct Community State Types (CSTs), which represent recurring dominant bacterial communities at specific time points or under varying physiological and environmental conditions[15–17]. However, while these studies have offered valuable descriptive frameworks, the mechanistic basis of early-life community assembly remains only partially resolved. In this context, the current study was aimed not only to identify CSTs from a comprehensive global collection of healthy infant metagenomes, but also to delineate key species characteristic of each CST and investigate the microbe-microbe interactions which shape and stabilize these configurations during early-life development. To this end, we integrated large-scale in silico analyses with an extensive culturomics effort, establishing a dedicated infant gut microbial biobank of 182 novel isolates. This resource enabled experimental validation of ecological interactions through targeted co-cultivation, cross-talk investigation, and metatranscriptomics analysis. By combining metagenomic, ecological, and experimental approaches, our study provides mechanistic insights into the ecological drivers of early-life CSTs, thereby extending beyond previous descriptive meta-analyses and addressing fundamental questions in infant gut microbiome development.

Results

Infant study cohort

With the aim of exploring the gut microbiota during the first year of infant development, we collected 5288 publicly available metagenomic data sets derived from fecal samples of healthy, full-term infants, from diverse countries across the world (Supplementary Data S1). Overall, 39% of the corresponding fecal samples originated from Northern Europe, 14% from Central Europe, and 6% from Africa. The remaining samples had been collected across regions located at the western (22%, predominantly USA) or eastern (19%, Asian countries) parts of the world, reflecting the global nature of the dataset (Fig. 1a). Based on the related metadata and consistent with previously published methodologies[18,19], data sets were stratified according to corresponding infant age into three distinct categories, encompassing the neonatal ($n = 974$; 0-1 month), pre-weaning ($n = 1784$; 1–6 months), and post-weaning ($n = 2530$; 6–12 months) stages (Supplementary Data S1, Fig. 1a). This categorization was intended to address age-related variability, recognizing the main dynamic changes in the establishment of the infant gut microbiota. Indeed, consistent with previous research, the first few weeks of life are characterized by high variability and significant inter-individual differences, largely due to the strong impact of external factors associated with birth and early-life environmental exposures, including maternal skin, vaginal microbiome, and hospital settings[20,21]. During the first six months following birth, the infant diet is almost exclusively based on milk, preferably provided through breastfeeding (according to WHO recommendations[22]), thereby supporting growth and associated metabolic activity of human milk oligosaccharide consumers, particularly key members of the *Bifidobacterium* genus[23]. Following this, the period from six to 12 months is marked by substantial shifts in microbial composition, especially related to the introduction of complementary food around six months (as suggested by WHO guidelines[24,25]), when most infants discontinue breastfeeding (or formula milk feeding) or combine it with solid foods, triggering rapid gut microbiota changes[12,20,26].

Envfit analysis confirmed well-known significant associations between microbial community variation and clinical variables, such as age, geography, delivery mode, and feeding type, with geography and age exerting the strongest influence on the infant gut microbiota structure ($r^2 = 0.23$ and 0.16, respectively), followed by feeding type and delivery mode ($r^2 = 0.05$ and 0.03) (Supplementary Figure S1). Notably, delivery mode had a pronounced impact in the first month of life, with vaginal delivery enriching gut commensals like *Escherichia coli* and *Bifidobacterium longum*, whereas feeding practices became more influential during the suckling period and persisted after solid food introduction, with exclusive breastfeeding enriching *Segatella copri* and *Bifidobacterium bifidum* (Supplementary Fig. S1). In contrast, geography emerged as the dominant driver of microbial composition after weaning, likely reflecting regional dietary habits and sociocultural factors associated with the introduction of solid foods (Supplementary Fig. S1). Age-dependent compositional patterns linked to clinical variables (i.e., feeding type and delivery mode) remained substantially consistent within individual geographic regions, indicating that geography alone does not explain these effects. Full statistical details and taxa associations are available in Supplementary Data S2, S3, and S4.

Global compositional patterns of the gut microbiota from birth to one year

To explore global interindividual variability of the infant gut microbiota throughout the first year of life, we employed hierarchical clustering guided by silhouette analysis and beta-diversity quantification to identify recurrent compositional clusters, also known as Community State Types (CSTs) within our infant cohort[17,27–29]. This approach revealed six CSTs, hereafter referred to as Early-Life CSTs (ELi-CSTs) (Fig. 1b, c, Supplementary Data S5), which were shown to significantly contribute to explaining the observed variability in microbial species composition within the infant gut microbiota (PERMANOVA, $p < 0.001$) (Supplementary Data S5). Among these, ELi-CST1 and ELi-CST4 emerged as the most prevalent, each detected in 23-24% of datasets (Fig. 1c), highlighting the widespread occurrence of *Bacteroidaceae* and *Bifidobacterium* across the infant cohort. Notably, while ELi-CST2 (Ec) exhibits a single-species dominance by *Escherichia coli*, ELi-CST4 and ELi-CST5 are characterized by a diversity and high abundance of *Bifidobacterium* species (Fig. 1c, Supplementary Data S5). These bifidobacterial communities were clearly partitioned into either *B. longum*-dominant (ELi-CST4 Blo) or *B. breve*-dominant (ELi-CST5 Bbr) profiles, aligning with previous observations[30–33]. This latter divergence has been attributed to priority effects, where early colonization provides a competitive advantage, at least in part, driven by fucose utilization, which is a key HMO-derived monosaccharide. Consequently, if *B. breve* is present in the infant gut microbiota shortly after birth, it is more likely to establish dominance within the community by four months of age[34].

In contrast, ELi-CST3 (mix), and ELi-CST6 (mix), and ELi-CST1 (Ba), the latter characterized by a heterogeneous composition with a high abundance of *Bacteroides* members that distinguishes it from the other groups, reflect a diversification into more complex and varied, multi-genera microbial consortia resembling those commonly found in the adult gut microbiota (Fig. 1c, Supplementary Data S5). Although ELi-CST1 and ELi-CST3 appear close in PCoA, PERMANOVA confirmed that they form statistically distinct clusters, albeit with a relatively low R^2 (p-value < 0.05; $R^2 = 0.05811$; Supplementary Data S5), reflecting their partial overlap. Consistently, this statistic is reflected in clear species-level compositional differences (Fig. 1c), supporting their interpretation as distinct yet closely related community states.

Investigation into the influence of developmental stages on these early gut microbiota configurations revealed a significant association between ELi-CST diversification and age. Consistently, as shown in Fig. 2a, ELi-CSTs were shown to exhibit distinct distributions across developmental stages (Pearson's Chi-Squared Test, $p < 0.001$) (Supplementary Data S6), with ELi-CST1 (Ba) and ELi-CST6 (mix) predominantly associated with the post-weaning phase and ELi-CST2 (Ec) characteristic of the neonatal age, thereby emphasizing the highly dynamic nature of the infant gut microbiota during the first year of life.

More specifically, the most pronounced age dependence was observed for ELi-CST1 (Ba) and ELi-CST6 (mix) (Chi-Square Tests, X-squared values = 479.33 and 1,105, respectively; $p < 0.001$) that became increasingly prevalent as infants transitioned into the post-weaning phase (Fig. 2a, Supplementary Data S6). These findings, supported by high standardized residuals in chi-square testing (8.76 and 21.95, respectively; FDR-corrected p-values = 0) (Fig. 2b, Supplementary Data S6), show that ELi-CST1 (Ba) and ELi-CST6 (mix) represent microbial assemblies dominating during the transition toward a more mature assembly after the introduction of complementary feeding, aligning with the presence of adult-like microbial taxa observed in these gut metagenome data sets (Fig. 2b, c, Supplementary Data S5).

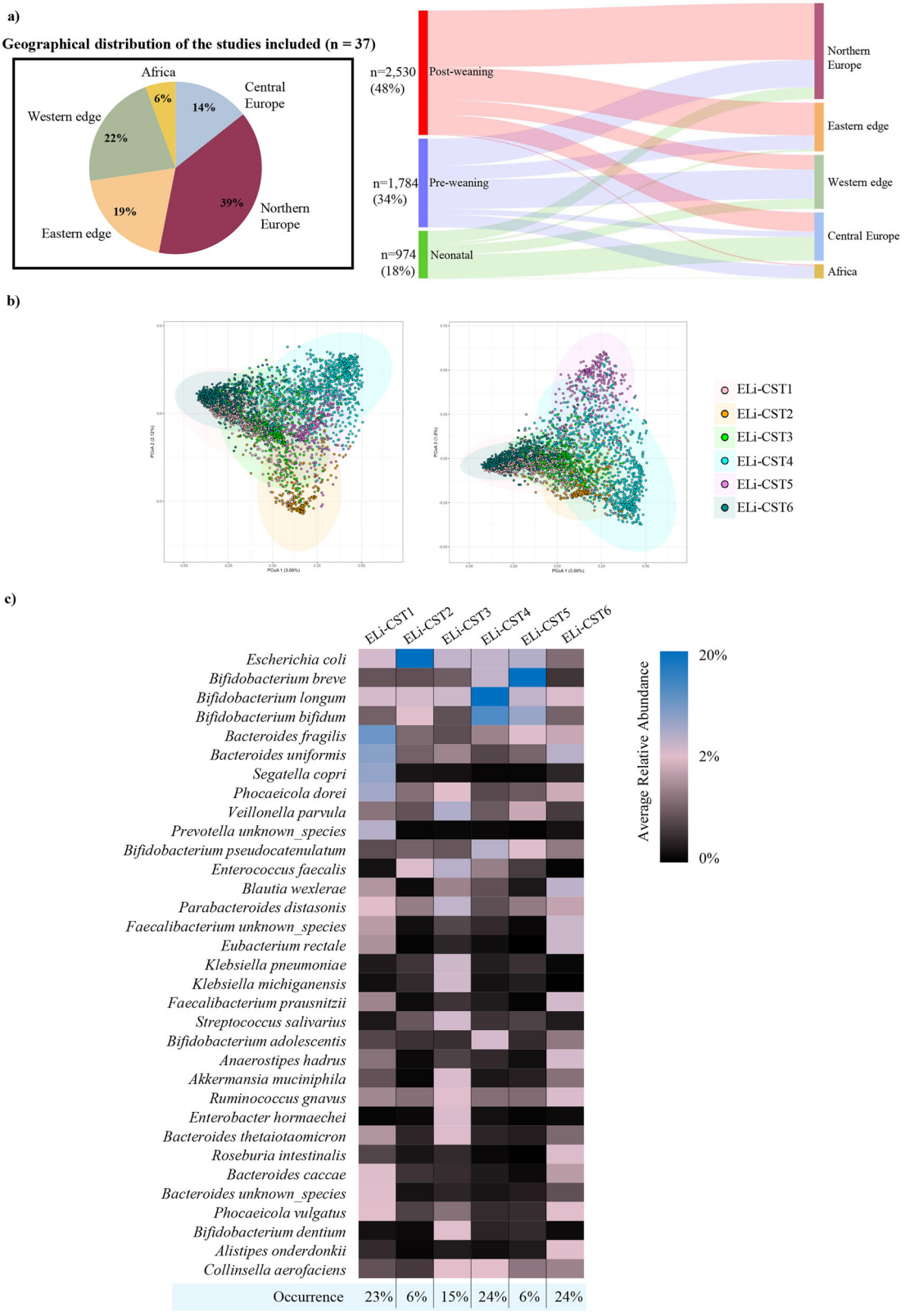
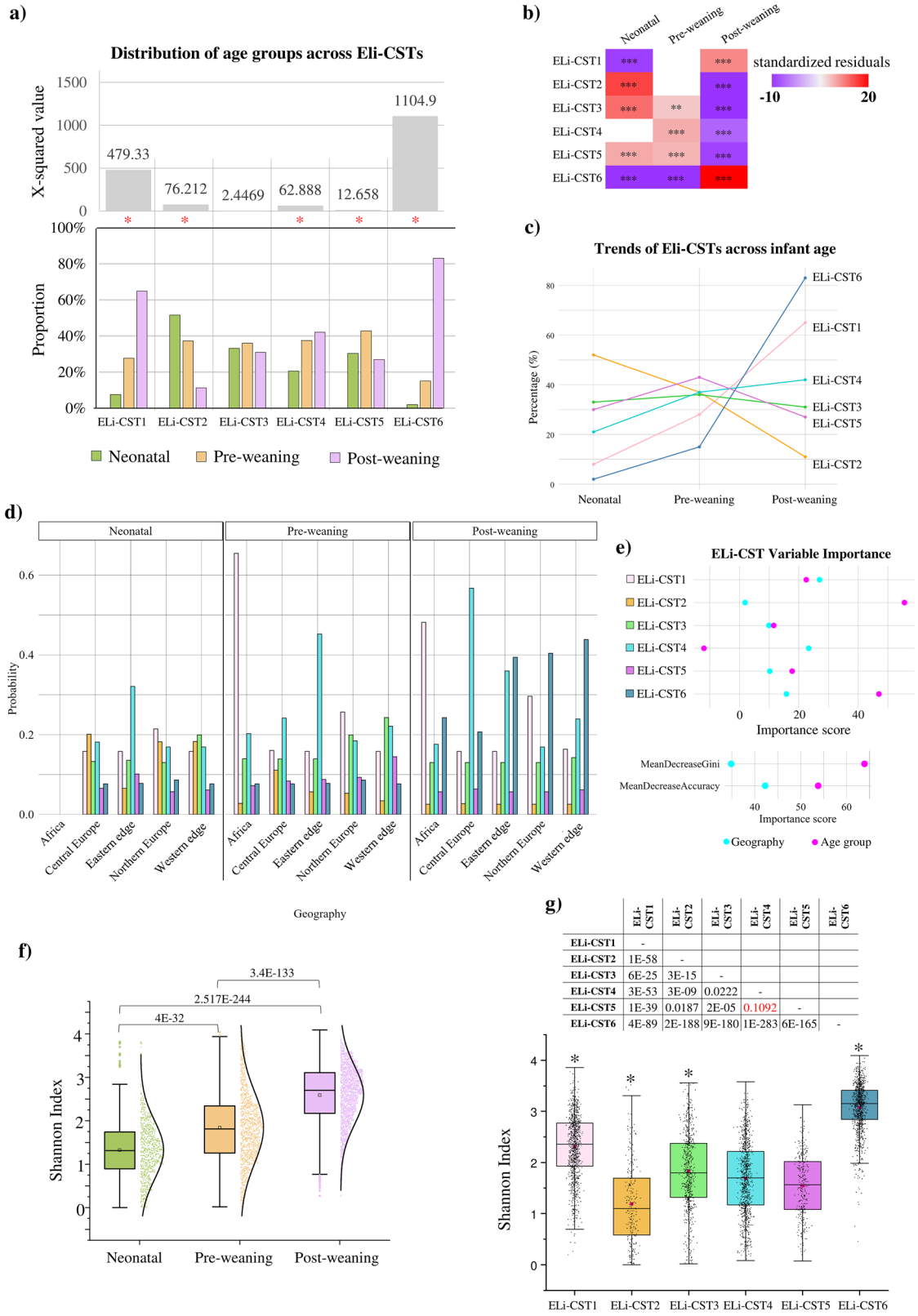


Fig. 1 | Recurrent microbial community state types of the infant gut in the first year of life. In panel (**a**), the pie chart shows the proportion of participants from each geographical region across the full dataset, which spans 37 independent studies. The Sankey diagram illustrates the distribution of study participants across developmental stages (neonatal, pre-weaning, post-weaning) and geographical regions, with the width of the flows proportional to the number of participants in each category. The Eastern edge refers to Asian countries, whereas the Western edge predominantly corresponds to the USA. Panel (**b**) shows the principal coordinate analysis (PCoA) of the six early-life microbial configurations (ELi-CSTs) identified in the assessed infant cohort. The colored ellipses represent 95% confidence intervals around each cluster, calculated based on the covariance matrix and the centroid of each cluster on the specified axes. In panel (**c**), the microbial composition of the six ELi-CSTs is depicted through heat map built using average relative abundance data. Only species with a maximum average relative abundance > 2% across the six ELi-CSTs are included. The percentages at the bottom of the heatmap indicate the occurrence of each ELi-CST within the infant population.



	ELi-CST1	ELi-CST2	ELi-CST3	ELi-CST4	ELi-CST5	ELi-CST6
ELi-CST1	-					
ELi-CST2	1E-58	-				
ELi-CST3	6E-25	3E-15	-			
ELi-CST4	3E-53	3E-09	0.0222	-		
ELi-CST5	1E-39	0.0187	2E-05	0.1092	-	
ELi-CST6	4E-89	2E-188	9E-180	1E-283	6E-165	-

Conversely, ELi-CST2 (Ec) shows a distinct, inverse trend, with significant over-representation among 0–1-month-old infants (standardized residual = 14.52, FDR-corrected p-value = 0) and a decreasing prevalence as infants increase in age, indicating that this gut microbiota configuration is more prominent during the neonatal phase (Fig. 2b, c, Supplementary Data S6). Consistently, in ELi-CST2 (Ec), *E. coli* reached an average relative abundance above 60%, whereas in the other ELi-CSTs its levels remained limited, ranging between 1.15% and 5.88% (Supplementary Data S5). This species, along with the low abundance of various members of the *Bifidobacterium* genus, appears to reflect a community composition characterized by early-life pioneer species acquired during or immediately after birth.

Fig. 2 | Association between gut microbiome structure and infant age. Panel (**a**) illustrates the association between ELi-CSTs and specific infant age as determined by Chi-squared analysis. Statistically significant age-dependence of each Eli-CST is indicated by a red asterisk (Chi-squared test, p-values < 0.05), with the strength of associations represented by the respective Chi-squared values. Panel (**b**) displays a heatmap of standardized residual values from the chi-squared test assessing the association between Eli-CSTs and infant age groups. Only Bonferroni-adjusted chi-squared p-values < 0.05 are shown. *p < 0.05, **p < 0 0.01, ***p < 0 0.001. Panel (**c**) depicts the occurrence trends for each ELi-CST across the three infant ages. In panel (**d**), predicted ELi-CST probabilities and variable importance derived from Random Forest model are shown through bar plots and a dot plot, respectively. Bar plots display the probability of belonging to each ELi-CST based on age group (neonatal, pre-weaning, post-weaning) and geographic origin. In panel (**e**), the dot plot represents the relative importance of each predictor (Geography and Age group) in determining ELi-CST assignment. Variable importance was assessed using both the Mean Decrease in Accuracy and the Gini Index metrics, and visualized through the dot plot, with higher values indicating stronger contributions to classification performance. Numbers shown above the brackets indicate significant pairwise comparisons based on the Kruskal–Wallis test with Dunn's post hoc test (FDR-adjusted p-values < 0.05). Panel (**f**) represents the alpha diversity (species diversity) measured as Shannon Index in the total cohort (n = 5242) categorized in age groups: neonatal (0–1 month of age; n = 967), pre-weaning (1–6 months of age; n = 1745), and post-weaning (6-12 months of age; n = 2530). Panel (**g**) shows the species diversity across the six identified ELi-CSTs, estimate through Shannon index. The boxes are determined by the 25th and 75th percentiles, and the whiskers are determined by the 5th and 95th percentiles. A matrix of FDR-corrected pairwise p-values (Kruskal–Wallis test with Dunn's *post hoc* test) is included, and an asterisk above the corresponding boxplot highlights the cluster significantly different from all others (FDR-corrected p-values < 0.05 in all pairwise comparisons).

With increasing age, 1–6 month-old infants were moderately but significantly over-represented within the bifidobacteria-dominant ELi-CST4 (Blo), which remained prevalent among infants transitioning through the weaning period (Fig. 2c). Thus, although the microbial configuration enriched in *E. coli* is common during the first month after birth, the observed decreasing trend of ELi-CST2 (Ec) suggests a progressive resolution of this early single-species dominance within the first six months of life, implying a physiological transition (Fig. 2c). Consistent with host-microbe coevolution models, *E. coli* acts as a pioneer species when oxygen levels promote growth of facultative anaerobes[35,36]. By progressively consuming oxygen, this process has been demonstrated to facilitate establishment of obligate anaerobes, such as members of the *Bifidobacterium* genus, this being supported by an exclusive milk-based host diet, followed by adult-associated species, including *Faecalibacterium prausnitzii*, *Agathobacter rectalis* (formerly *Eubacterium rectale*), *S. copri*, and *Ruminococcus gnavus*, the latter corresponding to the ELi-CST6 configuration[37].

While the developmental stage appeared to be an important determinant of ELi-CST distribution, we evaluated if geography may further modulate CST assignment, potentially interacting with age. To account for the joint and potentially non-linear effects of these variables, we adopted a predictive modeling approach based on the random forest algorithm, which estimated the probability of ELi-CST membership across all combinations of age groups and geographical origins (Fig. 2d, Supplementary Data S7). This method confirmed that age is the most influential predictor of ELi-CST, as indicated by the highest Mean Decrease in Accuracy and Gini Index scores (Age group: 55.9 and 64.6, respectively; Geography: 35.8 and 33.4, respectively) (Fig. 2d, Supplementary Data S7). Particularly, classification of ELi-CST2 (Ec) and ELi-CST6 (mix) was primarily driven by developmental age, with limited contribution from geographic origin (Fig. 2d, Supplementary Data S7). Nevertheless, for other early-life CSTs, the influence of geography was shown to be substantial. Notably, in ELi-CST4 (Blo), which is dominated by *Bifidobacterium* members, geography exerted a markedly stronger predictive power compared to age, suggesting that Eli-CST4 (Blo) tend to occur across all age groups but with varying prevalence depending on the geographical origin of the fecal sample. Indeed, the predicted probabilities showed a high assignment to ELi-CST4 (Blo) in the most easterly regions (≥32% across all age groups), followed by Central Europe (≥ 18% across all age groups) (Fig. 2d, Supplementary Data S7). This pattern was particularly evident during the neonatal and pre-weaning stages, where the probability of belonging to ELi-CST4 (Blo) for individuals from the most easterly regions was 32% and 45%, respectively (Fig. 2d, Supplementary Data S7). In contrast, pre-weaning infants from Africa showed the highest likelihood of being assigned to ELi-CST1 (Ba) (65%), whereas ELi-CST6 (mix) was more frequently predicted among post-weaning individuals from Northern Europe and most westerly regions (up to 44%) (Fig. 2d, Supplementary Data S7). Though developmental age remains the most important predictive variable for ELi-CST association, these findings highlight a geography-dependent CST prevalence, pointing to environmental parameters playing a determinant role in shaping microbial configurations during early life.

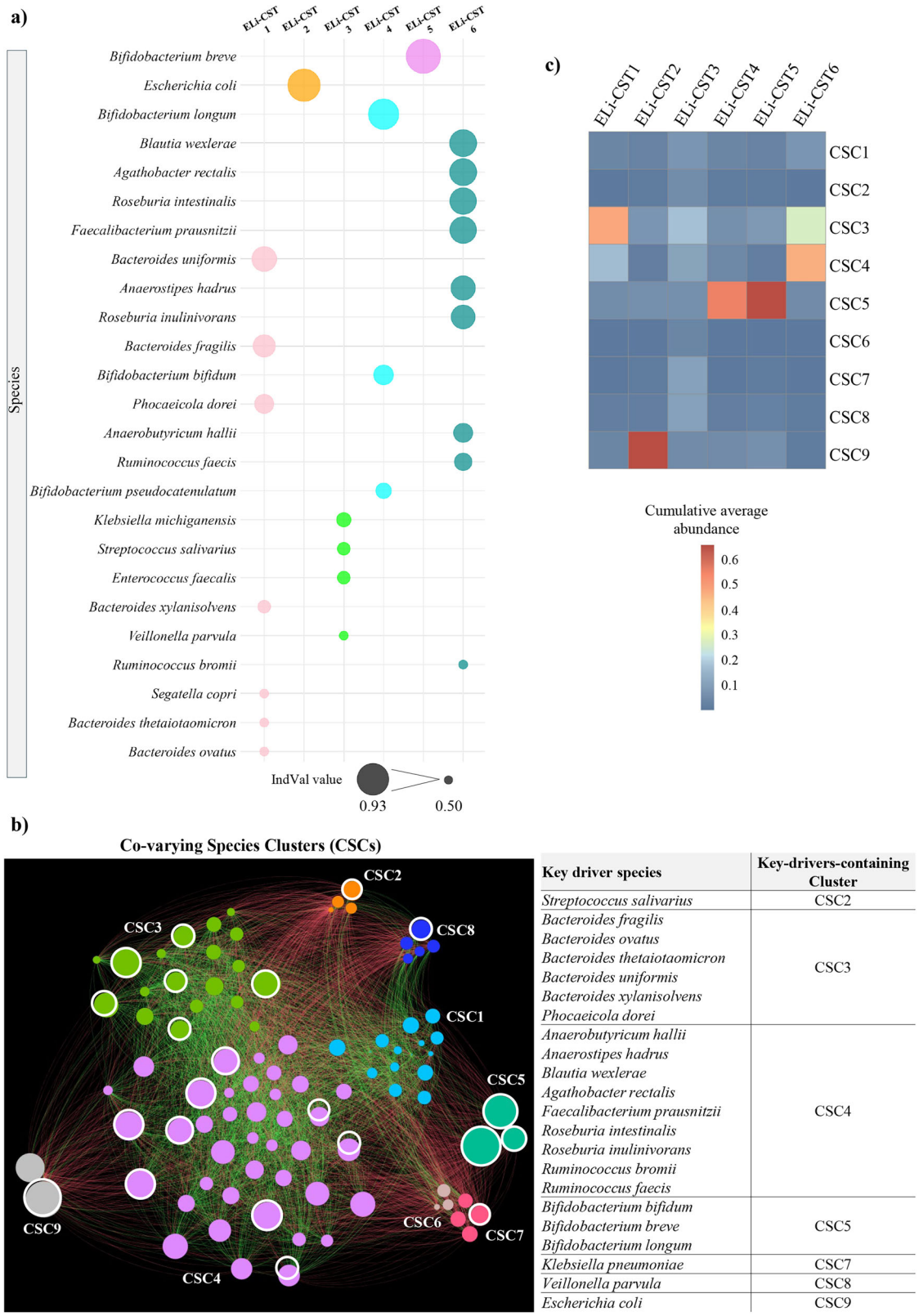
Moreover, similarly to the taxonomic profiles, biodiversity was shown to exhibit a trend associated with developmental age, with a significant increase in alpha diversity (estimated as the Shannon index) as the infant ages (Fig. 2f, Supplementary Data S8), and higher values attributed to the ELi-CSTs dominated by infants in post-weaned age (Fig. 2g, Supplementary Data S8).

Unveiling microbial key drivers of infant gut microbiota development using machine learning

To precisely characterize the interindividual variability of the infant gut microbiota from a taxonomic perspective, the random forest algorithm[38,39] was used to calculate variable importance scores, including age, geography, and study as additional covariates in the model to evaluate their role as potential confounders (Supplementary Data S9). To enhance interpretability in this high-dimension setting, an a priori 90th-percentile threshold was applied to focus on the top 10% of taxa with the highest discriminative power in distinguishing Eli-CSTs. Model performance was assessed with five-fold cross-validation (80:20 training:testing split), showing an overall mean accuracy of 0.86 (Supplementary Data S10). Although ELi-CST3 (mix) showed a lower recall (0.70), consistent with its partial compositional overlap in PCoA (Fig. 1b), per-class accuracy was uniformly high across the six ELi-CSTs (0.82-0.94), including the minority ELi-CST2 (Ec), ELi-CST3 (mix), ELi-CST5 (Bbr), indicating robust discrimination (Supplementary Data S11).

In addition to random forest, we incorporated indicator species analysis[40,41], a statistical method that allows identification of unique microbial signatures representative of each ELi-CST (Supplementary Data S12). Specifically, microbial species with an IndVal greater than 0.5 (permutation test, nperm = 999; p-value < 0.05) were considered crucial contributors[42], further refining the identification of CST-discriminant microbial modulators and highlighting ELi-CST-specific key species (Fig. 3a).

As detailed in Table 1, these combined analyses identified 25 key species that emerged as the most influential in shaping the infant gut microbiota and contributing to the establishment of ELi-CSTs, according to at least one method (importance score by random forests within the 90th percentile; IndVal ≥ 0.5), and were therefore designated as Key Infant Microbial Modulators (KIMMs). Notably, 13 of these taxa were identified as the most significant by both methods (Table 1). However, none of these high-indicator taxa were significantly associated with the ELi-CST3 (mix) (Table 1). Indeed, although this cluster was characterized by four key signatures, i.e., *Veillonella parvula*, *Klebsiella michiganensis*, *Streptococcus salivarius*, and *Enterococcus faecalis*, they achieved IndVal only slightly above the chosen threshold of 0.5 (ranging from 0.50 to 0.52) and importance scores falling outside the 90th percentile in the random forest analysis (Table 1). The absence of statistically robust microbial key modulators for ELi-CST3 (mix) supports the hypothesis that this microbiota configuration



Key driver species	Key-drivers-containing Cluster
Streptococcus salivarius	CSC2
Bacteroides fragilis *Bacteroides ovatus* *Bacteroides thetaiotaomicron* *Bacteroides uniformis* *Bacteroides xylanisolvens* *Phocaeicola dorei*	CSC3
Anaerobutyricum hallii *Anaerostipes hadrus* *Blautia wexlerae* *Agathobacter rectalis* *Faecalibacterium prausnitzii* *Roseburia intestinalis* *Roseburia inulinivorans* *Ruminococcus bromii* *Ruminococcus faecis*	CSC4
Bifidobacterium bifidum *Bifidobacterium breve* *Bifidobacterium longum*	CSC5
Klebsiella pneumoniae	CSC7
Veillonella parvula	CSC8
Escherichia coli	CSC9

Fig. 3 | Key modulators of the infant gut microbiota and clusters of co-varying species (CSCs). Panel (**a**) displays the results of the Indicator Species Analysis for the six ELi-CSTs, identifying microbial signatures significantly associated with each microbial configuration. Balloon sizes are proportional to the IndVal value, and only species with an IndVal > 0.5 and p-value < 0.001 are shown. Panel (**b**) depicts the microbial interaction networks of co-varying species clusters (CSCs) identified with the walktrap algorithm among members of the infant gut microbiota. Each node represents a bacterial species and is colored based on the CSC membership. Node size is proportional to IndVal. Edge colors distinguish between positive associations (green), and negative associations (red). Key indicative species within the CSCs are highlighted with a white node outline and are listed alongside the network. The complete list of species identified in each CSC is provided in Supplementary Data S13. Panel (**c**) illustrates the association between CSCs and ELi-CSTs using a heatmap, which represents the cumulative abundance of the members of each CSC within each ELi-CST.

Table 1 | Indicator values obtained by random forests and IndVal method for the six Eli-CSTs identified in the infant cohort

Species	Random forest method	IndVal method					Average relative abundance
	Importance score	Best match	Specificity	Fidelity	Indicative value	p-value	
Bifidobacterium longum	0.084773682	Eli-CST4	0.67	0.97	0.81	0.001	0.35
Bifidobacterium breve	0.074838215	Eli-CST5	0.87	1	0.93	0.001	0.51
Escherichia coli	0.043080347	Eli-CST2	0.76	1	0.87	0.001	0.64
Bifidobacterium bifidum	0.033058674	Eli-CST4	0.52	0.66	0.58	0.001	0.14
Bacteroides fragilis	0.031443827	Eli-CST1	0.6	0.63	0.62	0.001	0.11
Bacteroides uniformis	0.031229816	Eli-CST1	0.49	0.93	0.67	0.001	0.09
Faecalibacterium prausnitzii	0.029730657	Eli-CST6	0.58	0.85	0.71	0.001	0.03
Blautia wexlerae	0.028580017	Eli-CST6	0.54	0.95	0.72	0.001	0.05
Roseburia intestinalis	0.026918802	Eli-CST6	0.62	0.81	0.71	0.001	0.03
Agathobacter rectalis	0.024065712	Eli-CST6	0.64	0.81	0.72	0.001	0.04
Anaerostipes hadrus	0.020987329	Eli-CST6	0.54	0.81	0.66	0.001	0.03
Bacteroides thetaiotaomicron	0.020284845	Eli-CST1	0.25	0.83	0.45	0.001	0.02
Segatella copri	0.018698281	Eli-CST1	0.89	0.27	0.5	0.001	0.08
Phocaeicola dorei	0.017858311	Eli-CST1	0.51	0.62	0.57	0.001	0.07
Bifidobacterium pseudocatenulatum	0.016173577	Eli-CST4	0.46	0.61	0.53	0.001	0.05
Bacteroides xylanisolvens	0.012491324	Eli-CST1	0.4	0.65	0.51	0.001	0.01
Roseburia inulinivorans	0.011104455	Eli-CST6	0.63	0.66	0.65	0.001	0.02
Bacteroides ovatus	0.008445846	Eli-CST1	0.41	0.61	0.5	0.001	0.01
Ruminococcus faecis	0.00730162	Eli-CST6	0.49	0.62	0.55	0.001	0.01
Veillonella parvula	0.005961195	Eli-CST3	0.55	0.46	0.5	0.001	0.06
Anaerobutyricum hallii	0.005848688	Eli-CST6	0.6	0.55	0.57	0.001	0.01
Enterococcus faecalis	0.00335126	Eli-CST3	0.54	0.48	0.51	0.001	0.05
Klebsiella michiganensis	0.002458112	Eli-CST3	0.75	0.35	0.52	0.001	0.03
Streptococcus salivarius	0.001996813	Eli-CST3	0.54	0.48	0.51	0.001	0.03
Ruminococcus bromii	0.001955333	Eli-CST6	0.54	0.46	0.5	0.001	0.01

The table reports the Key Infant Microbial Modulators (KIMMs) identified based on either or both approaches. Importance scores by random forests within the 90th percentile are delineated by a blue line.

is better defined by the lack of a clear taxonomic signature. This unique pattern may reflect the possibility that ELi-CST3 (mix) does not predominantly occur within a specific age range, as observed above, potentially representing a transitional or mixed microbial state that is less specialized and persists throughout the first year of life, rather than a distinct and age-specific configuration of the developing infant gut microbiota.

Building on the concept that microbial communities are not randomly assembled but respond to ecological forces and specific functions, we investigated the network of interactions within the infant gut microbial community, with a specific focus on the role of the KIMMs described above in structuring and stabilizing these microbial communities. Notably, we identified nine co-varying species clusters (CSCs), revealing structured microbial associations that reflect potential ecological interactions within the developing gut microbiome (Fig. 3b, Supplementary Data S13). Among these, seven CSCs encompassed KIMMs that strongly correlated with other background species, suggesting the establishment of functionally cohesive microbial consortia centered around crucial modulators (Fig. 3b). These clusters were characterized by strong intra-cluster positive associations and negative relationships with members of other CSCs, supporting the notion that they represent distinct ecological units shaped by functional interdependencies or shared adaptation to host-driven selective pressures (Fig. 3b).

Interestingly, four CSCs, which included KIMMs, showed close association with specific ELi-CSTs (Fig. 3c), reflecting an alignment between co-occurrence network modules and sample community states. This pattern is compatible with the presence of potent microbial interactions that may contribute to ELi-CSTs stability and may partly explain why certain configurations appear more persistent across temporal perturbations or dietary changes[19,43]. For instance, *B. longum*, *B. bifidum*, and *B. breve*, included in CSC5 and associated with ELi-CST4 (Blo) and ELi-CST5 (Bbr) (Fig. 3b, c), are known to be tightly intertwined through their distinct strategies for accessing Human Milk Oligosaccharides (HMOs)[44,45]. These cooperative metabolic interactions likely promote mutualistic networks, reinforcing the resilience and stability of bifidobacterial-dominated communities and eventually contributing to the persistence of ELi-CST4 (Blo) and ELi-CST5 (Bbr) across infant ages[30,46,47].

Conversely, other CSCs were found across multiple ELi-CSTs, suggesting that they represent flexible microbial units that facilitate transitions between community states. Notably, CSC2, CSC7, CSC8, which were not specifically associated with any particular ELi-CST (Fig. 3c), were primarily composed of weakly key species of ELi-CST3 (mix). Given that ELi-CST3 was described above as a non-age-specific and potentially transitional microbial stage, the dispersed network structure of these CSCs further supports the notion that they contribute to microbiota plasticity, potentially mediating shifts between community states as the gut ecosystem adapts to developmental and environmental changes, as is characteristic of the fast-evolving infant gut microbiota[37].

Overall, these findings underscore the pivotal role of identified key microbial players in shaping and stabilizing the structure of the infant gut microbiota. By engaging in close interactions within clusters of co-varying species, these taxa act as central interactive components of microbial community organization, that enhance ecological stability.

Culturing the infant gut microbiota and establishing a microbial biobank to investigate species interactions

To explore whole-genome diversity of the infant gut microbiota, we performed a large-scale bacterial cultivation effort directed toward isolating the identified KIMMs, which play a crucial role in the establishment of

Table 2 | Cultivation media and corresponding supplements used for the cultivation of infant gut-resident bacteria

Media	Supplement
M17 broth	
modified-M17	D-Lactose
Rogosa	
LPSM (*L. plantarum* Selective Medium)	
GAM (Gifu Anaerobic Medium)	
modified-GIFU	
modified-PYG	Volatile fatty acids (acetic acid, propionic acid, isobutyric acid, n-valeric acid, iso-valeric acid)
Schaedler Anaerobe broth	
modified-Schaedler Anaerobe	Xylan, inulin, arabinogalactan
Anaerobe Basal broth	
Wilkins-Chalgren broth	
BHI	
YBHI	Yeast extract, cellobiose, maltose
YCFA	
modified-YCFA	
IGSM (Infant Gut Super Medium)	Filtered rumen
Columbia	Defibrinated sheep blood
TH (Todd-Hewitt)	
effluent-MacFarlane-sugar (EMS) medium	Vitamins solution (B7, B9, B5, B1, B2, B3, B6, B12, lipoic acid, p-aminobenzoic acid), hemin solution
Veillonella medium	
Prevotella medium	Xylan, inulin, pectin
CHOPPED MEAT MEDIUM	Yeast extract, D-glucose, cellobiose, maltose, starch, casitone, dipotassium phosphate
EG	Defibrinated sheep blood
MRC (Modified Reinforced Clostridia)	
MRS broth	
modified-MRS	Clindamycin and Ciprofloxacin

ELi-CSTs. For this purpose, fresh infant fecal samples were inoculated on 26 different media, with selected media supplemented with up to 32 growth factors (alone or in combination) to promote growth of fastidious microorganisms (Table 2).

As a result, we obtained 182 bacterial isolates which, following whole-genome sequencing, led to the definition of a representative biobank encompassing 158 genome sequences with more than 90% completeness, corresponding to 23 different genera and 55 species (Supplementary Data S14). To confirm taxonomic identities, a preliminary whole-genome similarity analysis was accomplished by calculating pairwise ANI values against all conspecific genome sequences available in the NCBI database (last access Jun 2025). In this context, our isolate *Bacteroides* I13-3 did not meet the threshold for similarity to any previously identified species (ANI cutoff < 94%) (Supplementary Data S15), indicating that it may represent a novel species within the *Bacteroides* genus.

This established microbial biobank was used to construct genome-scale metabolic network models to further investigate species interactions within ELi-CSTs, with a primary focus on the *Bifidobacterium*-dominated ELi-CST4 (Blo) and *Bacteroides*-rich ELi-CST1 (Ba), which were the most common microbial community configurations in pre-weaning and post-weaning microbiota, respectively. To achieve this, we included all available genome-sequenced strains from our biobank corresponding to species that represent components of ELi-CST1 (Ba) and ELi-CST4 (Blo) with a relative abundance greater than 2%. We then applied the *RevEcoR* R package[48–50] to assess metabolic interplay among these key early colonizers of the infant gut by computing pairwise indices of metabolic competition and complementarity (Supplementary Data S16).

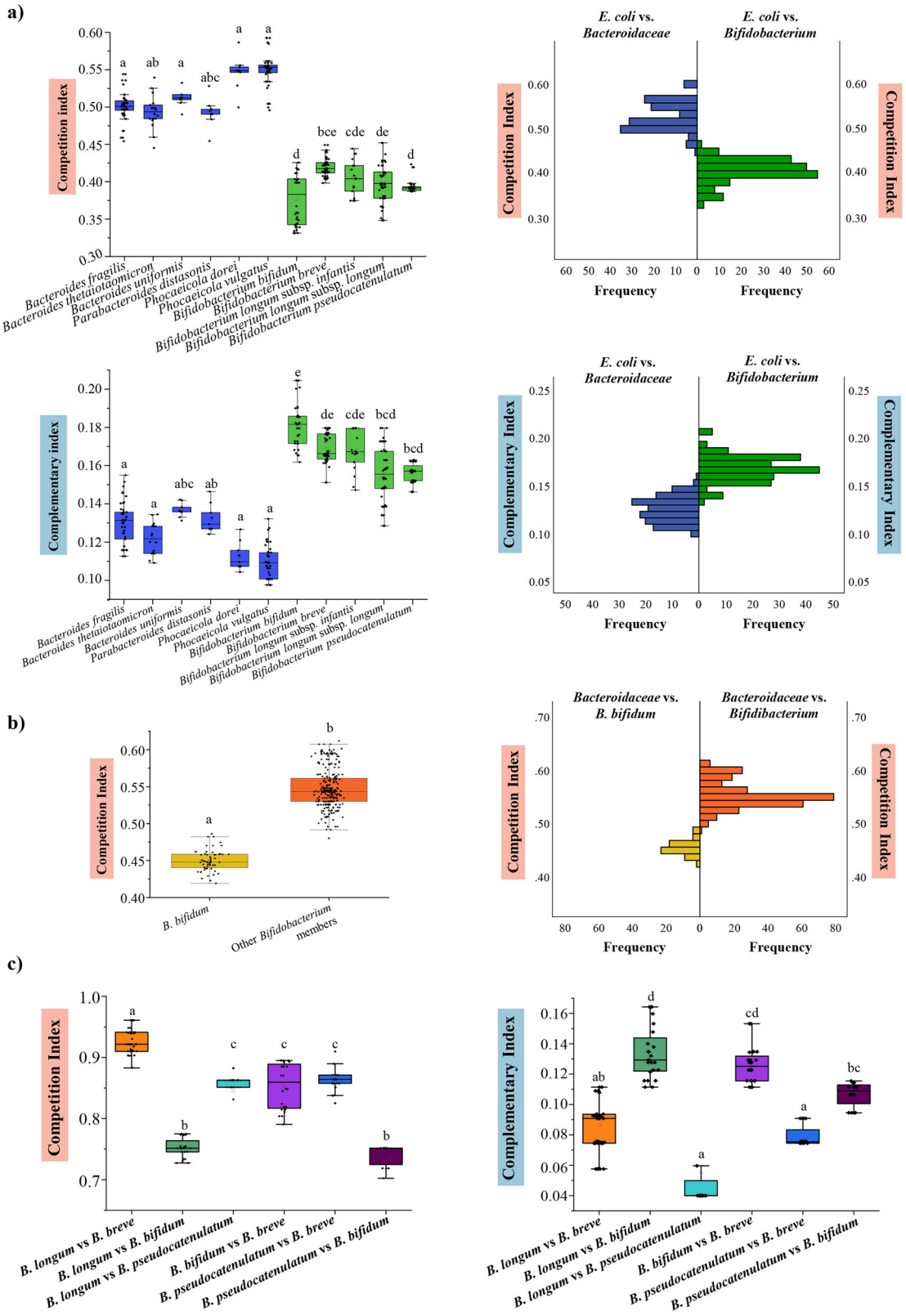
This analysis predicted that *E. coli* engaged in lower competition and a higher complementarity with bifidobacteria of ELi-CST4, including *B. longum*, *B. bifidum*, *B. breve*, and *B. pseudocatenulatum* (average competition index = 0.40, average complementarity index = 0.16), compared to its interaction with *Bacteroidaceae* of ELi-CST1 (average competition index = 0.52, average complementarity index = 0.12) (Kruskal-Wallis with Dunn *post-hoc* test, Bonferroni-corrected $p < 0.05$) (Fig. 4a, Supplementary Data S16). These predicted interactions suggest a degree of metabolic affinity between *E. coli* and bifidobacteria, aligning with their co-occurrence in the early-life infant gut. Conversely, *Bacteroidaceae* of ELi-CST1 (Ba), such as members of the *Bacteroides* and *Phocaeicola* genera, were shown to display a relatively high average competition index with bifidobacteria (0.53), reflecting their mutual exclusion in our CST-based analysis. Nevertheless, competition was reduced when the interacting partner was *B. bifidum* (0.45) (Mann–Whitney *U* test, $p < 0.001$) (Fig. 4b). In this regard, the well-known extracellular hydrolysis of HMOs by *B. bifidum* may generate metabolic intermediates that can be shared with other species, potentially mitigating direct competition in the infant gut[44,51].

Although competition and complementarity indexes are inherently influenced by genomic similarities, often resulting in higher predicted antagonism among closely related bacteria, *B. longum* subsp. *longum* was shown to exhibit a particularly high degree of competition with *B. breve* (average competition index = 0.91) compared to other pairs of infant gut-associated bifidobacterial species (Kruskal-Wallis with Dunn *post-hoc* test, Bonferroni-corrected *p*-values < 0.01) (Fig. 4c, Supplementary Data S18). Consistently, this dynamic likely contributed to the emergence of distinct *Bifidobacterium*-rich ELi-CSTs, each dominated by either *B. longum* (ELi-CST4) or *B. breve* (ELi-CST5). Similarly, the interactions of *B. bifidum* also appeared to be reflected in ELi-CST composition. Indeed, this species emerged as the most permissive and least competitive partner, displaying the lowest competitiveness with other infant gut-associated bifidobacteria (average competition index 0.79), while exhibiting a certain degree of metabolic complementarity, particularly with certain strains of *B. longum* subsp. *longum* and *B. breve* (Fig. 4c, Supplementary Data S18). These predicted interactions are consistent with our CST-based stratification of the infant gut microbiota, in which *B. bifidum* was shown to co-occur with *B. longum* in ELi-CST4 (Blo) and with *B. breve* in ELi-CST5 (Bbr), albeit at a lower relative abundance than its bifidobacterial counterparts. Overall, these findings suggest that commonly co-occurring taxa tend to engage in stronger ecological interactions, either through competition or metabolic complementarity.

Assessment of in vitro cross-talk through dual co-culture

To experimentally investigate microbial interactions, we aimed to characterize the relationships among key species within the same ELi-CST (co-occurring species) and those between key species from temporally successive CSTs. For this purpose, we selected representative bacterial strains from our in-house biobank, corresponding to dominant KIMMs associated with the pre-weaning ELi-CST4 (Blo) and post-weaning ELi-CST1 (Ba), which represent two of the most prevalent community states observed in infants before and after weaning (Fig. 1c). Strain selection was guided by the Optimal Representative Strain (ORS) selector pipeline[52], which ranked *Bacteroides fragilis* 318 F, *Bacteroides uniformis* 324 F, *Bifidobacterium bifidum* PRL2015 and *Bifidobacterium longum* subsp. *longum* PRL2016 within the 85th-90th percentile of their respective species-specific reference score distributions, thereby supporting their suitability as representative models (Supplementary Figs. S4, S5).

Pairwise co-cultivation assays (bi-associations) were subsequently conducted using infant gut-simulating medium (Infant Gut Super Medium, IGSM), in parallel with monoculture controls, to explore in vitro potential ecological interactions between strains. After 8 h of direct cell-to-cell



contact, quantitative PCR (qPCR) analysis was performed to quantify growth of each strain, calculated as the increase in cell number relative to the initial inoculum (log_{10} fold change). Growth in co-culture was then compared to that observed in the corresponding monocultures to determine the effect of interspecies interactions (Fig. 5a, Supplementary Data S19). Notably, *B. bifidum* PRL2015 and the two *Bacteroides* strains did not appear to benefit from co-cultivation with one another, as their growth was either not significantly different from or significantly lower than that observed in monoculture (log_{10} fold change bi-associations vs monoculture; Mann-Whitney U Test), aligning with their low tendency to co-occur within the same microbial community structures. In contrast, *B. longum* subsp. *longum* PRL2016 showed markedly enhanced growth when paired with *B. bifidum*

Fig. 4 | Prediction of ecological interactions among common members of the infant gut microbiota. In panel (**a**), box-and-whisker plots (left) further detail the distribution of these indices across individual species of *Bacteroidaceae* and *Bifidobacterium*. Groups sharing at least one letter are not significantly different from each other, whereas groups with no letters in common are significantly different (Kruskal-Wallis with Dunn *post-hoc* test, $p < 0.05$). Complete listings of all FDR-adjusted pairwise *p*-values can be found in Supplementary Data S17–S18. Boxes indicate the 25th and 75th percentiles, while whiskers extend to the 5th and 95th percentiles. Pyramid histograms (right) display the distribution of competition (top) and complementarity (bottom) indices for *E. coli* interactions with members of the *Bacteroidaceae* family (left, blue bars) and *Bifidobacterium* genus (right, green bars), capturing statistically significant differences in interaction patterns between the two groups (Mann–Whitney *U* test, $p < 0.05$). In panel (**b**), the box-and-whisker plot (left) and the pyramid histogram (right) illustrate the statistically significant difference (Mann-Whitney U test, *p*-value < 0.05) in competitive interaction scores between *Bacteroidaceae* and either *B. bifidum* (left, blue bars) or other *Bifidobacterium* species (right, green bars). Panel (**c**) illustrates competition (left) and complementarity indices (right) among different *Bifidobacterium* species using box-and-whisker plots. Boxes indicate the 25th and 75th percentiles, while whiskers extend to the 5th and 95th percentiles. Groups sharing at least one letter are not significantly different from each other, whereas groups with no letters in common are significantly different (Kruskal-Wallis test with Dunn's *post hoc* test, Bonferroni-adjusted *p*-value < 0.05). Complete table of all FDR-adjusted pairwise *p*-values in Supplementary Data S17–S18.

and also appeared to promote growth of both *B. bifidum* PRL2015 and *Ba. fragilis* 318 F (log_{10} fold change bi-associations vs monoculture; Mann-Whitney *U* Test, $p < 0.05$) (Fig. 5a, Supplementary Data S19), in line with previous evidence of interspecies compatibility between *B. longum* subsp. *longum* and *Bacteroides* under in vitro conditions[53]. Considering its known persistence in the infant gut[19], this behavior suggests a potential facilitator role for *B. longum* subsp. *longum* in the ecological maturation of the infant gut microbiota. By sustaining early-life colonizers while promoting the integration of post-weaning-associated species, *B. longum* subsp. *longum* may help coordinate the compositional transition of the microbial community during infancy.

To molecularly assess the interaction between bacterial strains, whole-transcriptome sequencing (Supplementary Data S20) was performed on aliquots from the same cultures used in the qPCR assays. Subsequently, a cross-talk index was calculated for each genome in the bi-associations, defined as the proportion of genes differentially expressed relative to monoculture conditions, normalized to the total number of protein-coding genes in the respective genome, as previously described[54,55].

The highest cross-talk index was detected in the co-culture of *B. bifidum* PRL2015 and *B. longum* subsp. *longum* PRL2016, where 23% and 36% of their respective protein-coding genes were transcriptionally modulated in response to the presence of the bifidobacterial partner (Fig. 5b, Supplementary Data S21, S22). Functional classification of these genes into Cluster of Orthologous Groups (COGs) revealed reciprocal upregulation of genes involved in carbohydrate uptake and metabolism, accounting for 6% and 10% of the identified differentially expressed genes, respectively (Fig. 5c, Supplementary Data S21, S22). These were shown to include genes encoding exo-α-sialidase, and β-N-acetylglucosaminidase activities, as well as members of glycosyl hydrolase families 20 and 101, members of which have been shown to be pivotal for host glycan degradation[56–58], suggesting a metabolism-oriented transcriptional shift that enhances fitness of both bifidobacterial partners, thereby presumably contributing to the enhanced growth phenotype observed in co-culture.

Interestingly, *B. longum* PRL2016 was demonstrated to exhibit a broader transcriptional response, also showing significant gene expression changes in the presence of *Ba. fragilis* 318 F and *Ba. uniformis* 324 F, with a cross-talk index of approximately 25% in both co-associations (Fig. 5b, Supplementary Data S22). Specifically, 7% of these transcriptionally modulated genes were upregulated and associated with the uptake and metabolism of carbohydrate sources, including genes encoding GH20 and GH101 family members, while an additional 5% were linked to amino acid transport and metabolism (Fig. 5, Supplementary Data S22). In turn, although the transcriptome of *Ba. fragilis* 318 F appeared to be only marginally influenced by the presence of a cultivation partner (Fig. 5, Supplementary Data S23), co-cultivation with *B. longum* subsp. *longum* PRL2016 enhanced the expression of three distinct Sus-like polysaccharide utilization loci (PULs) (Supplementary Data S23). Notably, these *loci* encode components required for glycan binding and import[59,60], as well as a set of specialized glycosyl hydrolases, such as GH18, GH95, and an α-L-fucosidase, involved in the degradation of host mucin O-glycans and structurally related HMOs[61,62]. A similar transcriptomic response was achieved in the bi-association *B. longum* subsp. *longum* PRL2016 and *Ba. uniformis* 324 F, where the latter upregulated three diverse Sus-like PULs representing genes which are predicted to specify GH76, GH125, GH92, and GH18 family enzyme activities, previously associated with host glycan degradation in the model species *Bacteroides thetaiotaomicron* (Supplementary Data S24)[61].

Consistent with previous findings[53,55] these transcriptional responses support the notion that *B. longum* subsp. *longum* benefits from, and actively engages in, molecular cross-talk with both early- and post-weaning taxa. Notably, its presence appears to provide a nutritional advantage to *Bacteroides* spp., potentially supporting their establishment and persistence in the suckling infant gut as the diet transitions to include more plant-derived saccharides[62]. These findings support the proposed role of *B. longum* subsp. *longum* as a facilitator of ecological maturation within the developing gut microbiota.

Discussion

The first year of life is marked by a highly dynamic gut microbiota, characterized by rapid compositional shifts and considerable interindividual variability[21,63,64]. While several cohort studies and large-scale meta-analyses have catalogued this variability, few have examined the mechanistic basis of how early-life pioneer taxa co-assemble to form the infant gut community.

In this context, our findings offer a detailed ecological perspective supported by an integrated statistical framework, on early-life assembly of the gut microbiota through the identification of six distinct early-life community state types (ELi-CSTs), which serve as a starting scaffold for downstream investigation. These microbiota configurations capture not only temporal transitions, highlighting the rapid transformation of the infant gut during the first year after birth, but also reveal how regional factors contribute to shape microbial trajectories that appear to be age-independent. In particular, ELi-CST4, characterized by a bifidobacteria-dominated profile, emerged as a geography-dependent bifidobacterial community. Its greater prevalence across developmental windows in the Asian populations suggests the existence of microbial trajectories potentially shaped by local vertical transmission dynamics, environmental exposure, or dietary practices during early infancy[65,66].

Building upon these community-level patterns, the identification of 25 Key Infant Microbial Modulators (KIMMs) further reveals how specific microbial taxa appears to act as ecological scaffolds, shaping the structure and stability of early-life gut communities through predicted metabolic complementarity and competitive relationships, reflecting their potential for coexistence. These predicted interaction patterns were experimentally investigated through dual co-culture RNA sequencing, providing molecular insights into the complex microbial cross-talk that shapes the infant gut microbiota during the first year of life. In this context, the molecular interplay between *B. longum* subsp. *longum*, known for its prolonged persistence in the infant gut[19], and *Bacteroides* species associated with post-weaning ELi-CSTs may play a key role in guiding the ecological transition of the infant gut microbiota. Such interactions may support the progressive maturation of the microbial ecosystem during the first year of life, facilitating the integration of taxa associated with post-weaning configurations while maintaining functional continuity across developmental stages[53,67]. Taken together, our CST inference coupled with identification of key species and experimental investigations builds on earlier work while extending it

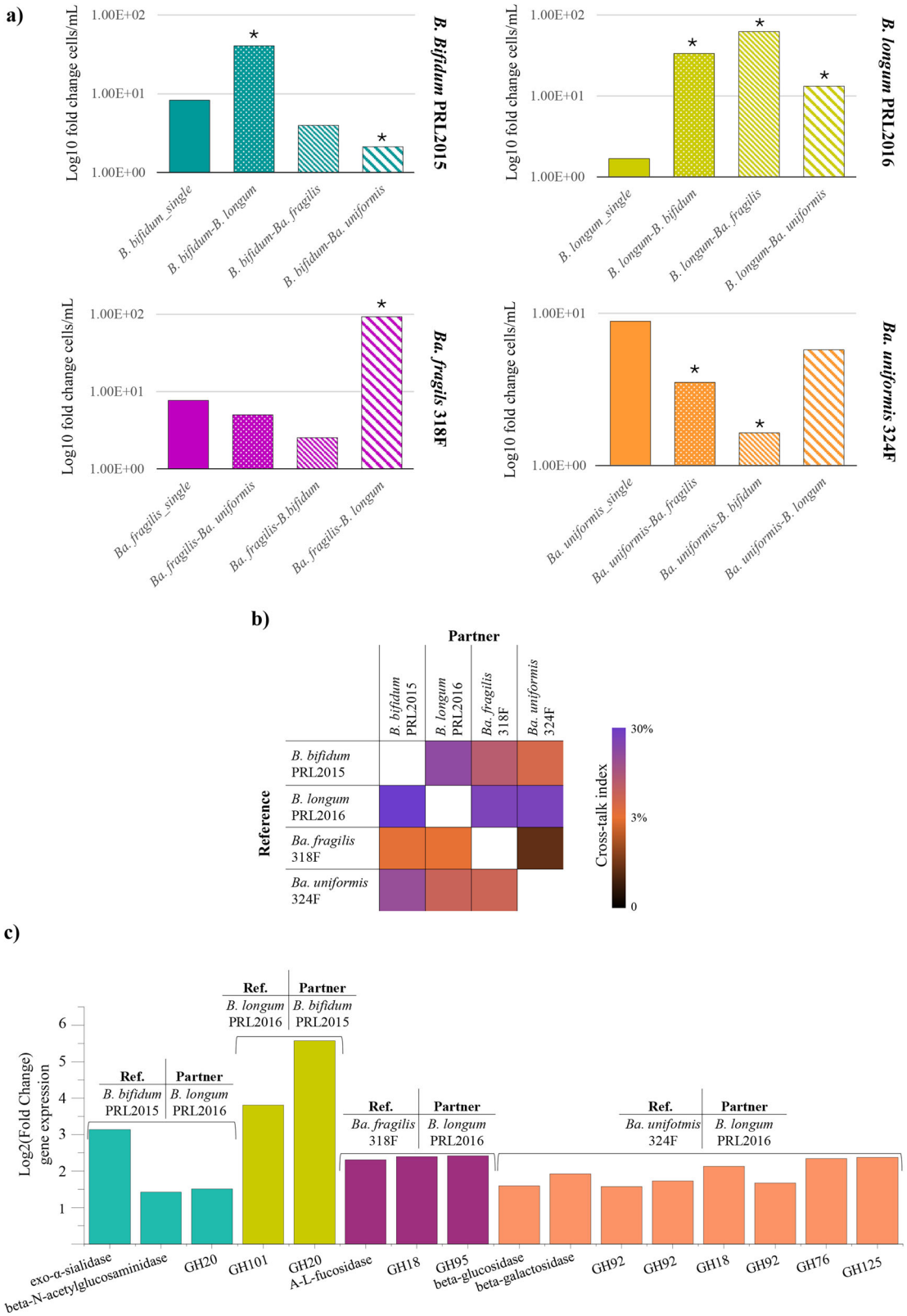


Fig. 5 | Ecological and molecular interaction among early and post-weaning infant gut members. Panel (**a**) displays the bacterial growth, of *B. bifidum* PRL2015, *B. longum* PRL2016. *Ba. fragilis* 318 F, and *Ba. uniformis* 324 F, expressed as $\log_{10}$ fold change in cell/mL relative to the initial *inoculum*, in single cultures and in bi-associations in IGSM medium. For each strain, growth in co-cultures was statistically compared to its corresponding monoculture using the Mann-Whitney *U* test; statistically significant differences are indicated by asterisks. Panel (**b**) shows the cross-talk index, defined as the percentage of differentially expressed genes relative to the total gene repertoire, for each bacterial strain grown in bi-association cultures. Panel (**c**) displays bar plots of genes involved in HMO/host glycan utilization that are significantly upregulated ($\log_2$ fold-change, FDR-adjusted p-values < 0.05) in bi-association conditions compared to monocultures.

toward mechanism, offering a clearer link between community states, candidate microbial drivers, and microbial cross-talk underlying early-life community assembly of the first year of life.

This study integrates publicly available infant metagenomes across multiple cohorts with the aim of exploring the natural variability present across different populations. This approach inherently introduces heterogeneity in study design. Although we attempted to address known potential confounding factors, residual confounding and batch effects cannot be fully excluded. A further limitation is represented by incomplete or inconsistently recorded metadata for clinically relevant variables (e.g., feeding type, delivery mode), which constrains the scope and precision of multi-cohort analyses.

Looking ahead, prospective longitudinal cohorts with denser time points would better resolve CST transitions and stability of community state. Incorporating multi-omics layers (e.g., metatranscriptomics, metabolomics) will help to connect composition data to function and refine mechanistic interpretation. Finally, complementary experimental work in gut-on-chip systems could test putative interactions suggested by the co-occurrence network analysis.

Methods

Shotgun metagenomic dataset collection

In this study, we utilize 5288 publicly available shotgun metagenomic data sets involving studies that used fecal samples from infants aged between a few days to one year, classified as overall healthy by the original publications. Only datasets annotated as stool samples and associated with subjects identified as "newborn" or "infant", with a valid "infant age" falling within the study range, were considered for inclusion. Metagenomic samples from preterm infants, from infants affected by acute intestinal infections, or from those who had received antibiotic or probiotic treatment at the time of sampling were excluded. Where available, information on feeding type and delivery mode was collected. No further exclusion criteria were applied, as the primary objective was to preserve the natural variability and biodiversity of the global infant gut microbiome.

To minimize platform-driven batch effects, we restricted inclusion to Illumina datasets (manufacturer-standardized chemistries and library kits). All raw reads were reprocessed with a single pipeline, applying uniform adapter/quality trimming, host read removal, and taxonomic profiling with the same tool METAnnotatorX2 and database version; only samples passing common QC thresholds were retained (see below for details on reads processing methods). Samples with fewer reads than 50,000 (after quality filtering and removal of host-associated reads; see below for details on reads processing methods) were excluded from the analysis. To avoid confounding by library size, microbial tables were normalized to relative abundances (counts scaled by total reads per sample).

When multiple samples from the same subject fell within the same age group, only one sample was retained. This allowed us to retain only a single representative sample per individual for each age category, contributing to the integrity of the data. A complete list of samples and metadata is available in the supplementary material (Supplementary Data S1).

Processing and taxonomic classification of metagenomic reads

The raw shotgun sequencing data were quality-filtered using the fastq-mcf program, applying a minimum mean quality score of 20, a window size of 5 bp, and a minimum read length of 100 bp. Next, the reads were mapped to the *Homo sapiens* reference genome (GRCh38.p13 assembly) using BWA software to remove any human-derived metagenomic sequences. The obtained filtered read sequences were subjected to taxonomic profiling, including the assessment of the relative abundance of bacterial species, using the METAnnotatorX2 tool[68]. Any reads demonstrating identical sequence matches to multiple bacterial species were excluded from further analysis to ensure accuracy. Low-abundance and low-prevalence filtering was implemented to exclude taxonomic features with relative abundances below 0.1% and prevalences below 5%, aimed at minimizing background noise in the data. After taxonomic filtering, individuals whose cumulative relative abundance of the remaining taxa did not reach 85% were excluded from further analysis.

Definition of early-life CSTs (ELi-CSTs)

The Bray-Curtis dissimilarity matrix was computed from species abundance data using the "vegdist" function from the *vegan* package (version 2.5-7). Alpha-diversity was calculated using the Shannon index, estimated through the *vegan* package (version 2.5-7), while beta-diversity was visualized through PCoA, based on the Bray-Curtis distance matrix.

To define possible microbial community state types (ELi-CSTs) of the infant gut microbiota, hierarchical clustering was performed with the "hclust" function in R, employing the Bray–Curtis dissimilarity matrix and Complete Linkage method. The optimal number of clusters was determined using the average silhouette width[69], calculated with the "silhouette" function from the *cluster* package (version 2.1.4). Following unsupervised inspection, clusters with small sample size (<1% of the study population) and heterogeneous composition were not retained, resulting in six robust CSTs used for downstream analyses. To further confirm the robustness of the CST classification and to quantify the proportion of variance explained by each comparison, PERMANOVA analysis was performed using the "adonis2" function in the *vegan* R package (version 2.5-7) with Bray-Curtis dissimilarities and 999 permutations (Supplementary Data S4).

To evaluate the association between age groups and ELi-CST distribution, a Chi-squared test was performed in RStudio using proportional count data. Standardized residuals were used for *post hoc* analysis, and adjusted *p*-values were computed to assess statistical significance. To investigate the predictive contribution of age and geography on ELi-CST assignment, a random forest classification model was implemented using the *randomForest* package in R. All categorical predictors were included simultaneously in the model formula, so that CST assignment was predicted from the joint distribution of age group and geography rather than from each variable in isolation (e.g., ELi-CST ~ Age_group + Geography). Random forests inherently account for interdependencies between predictors, as each tree split considers the combined influence of the included variables, thereby controlling for one another in the classification task. The model was trained with 500 trees (ntree = 500). Variable importance was quantified through both the Mean Decrease in Accuracy and the Gini Index metrics, which capture the relative contribution of each predictor while accounting for the presence of the others. The final output included calibrated probabilities of CST assignment for each age-geography combination, obtained by applying Platt scaling (logistic regression on the raw random forest probabilities) to improve probability estimates. These calibrated probabilities were then used for graphical visualization of class distributions and for reporting variable importance scores.

Identification of ELi-CST key indicator taxa

Key species associated with each ELi-CST were identified using the "multipatt" function from the R package *indicspecies*, employing 999 permutations for statistical significance testing. The generalized indicator value index (IndVal) was used to identify microbial species that significantly contribute to the differentiation of ELi-CSTs, combining measures of specificity (how representative a species is for a given group) and fidelity (how consistently a species occurs within that group)[70]. This approach identifies species that are strong indicators of each ELi-CST. IndVal analysis was combined with a random forest model[71]. In this latter, microbial taxa (at species level) were used as predictors to classify samples into ELi-CSTs, with the aim of identifying microbial signatures that best discriminate between CST categories. In random forest analysis, we trained a multiclass random forest classifier with the *ranger* package in R, using species-level abundances as predictors and the six ELi-CSTs as responses. Alongside microbial relative abundances, non-microbial covariates (study, geography, and age group) were included as additional predictors to account for potential confounding effects. The model used 1000 trees (num.trees = 1000), splitrule = "gini", min.node.size = 1, and the number of variables randomly sampled at each split (mtry) was set to the square root of the total number of predictors. To assess robustness, we

ran 5-fold cross-validation (with an 80:20 train:test split within each fold) created with caret::createFolds(class_data$enterotype, k = 5, list = TRUE, returnTrain = FALSE). For each fold, hold-out predictions were compared with true labels using *caret* to obtain confusion matrices, overall (accuracy and Cohen's κ), as well as per-class (precision, recall/sensitivity, F1-score, balanced accuracy) metrics: caret::confusionMatrix(preds_factor, fold_test $classes. Overall and per-class metrics are reported in Supplementary Data S10, S11. From held-out predictions we computed one-vs-rest ROC curves and AUCs, including macro-AUC and multiclass AUC (Supplementary Fig. S2): pROC::roc(response = factor(truth == k, levels = c(FALSE, TRUE)), predictor = probs[, k]); pROC::auc(roc_obj);. pROC::multiclass.roc(truth, as.matrix(probs)).

A final model was then refit on all samples with the same hyperparameters to obtain the full ranked list of predictor importance. For each predictor variable (i.e., assessed taxa), random forest computes an importance value which gives a measure of the intensity of relationship between this variable and the response variable (ELi-CST membership). Feature importance was quantified using permutation importance (mean decrease in accuracy), which measures the predictive contribution of each variable by assessing the decrease in model accuracy when its values are randomly permuted.

Packages/functions used: ranger (ranger, predict), caret (createFolds, confusionMatrix), pROC (roc, auc, multiclass.roc), ggplot2 (plotting).

Definition of clusters of co-varying species and network representation

Spearman correlation coefficients were calculated based on taxonomic abundance data to assess associations between species, generating a correlation matrix that was then processed using the Walktrap algorithm to detect clusters of co-varying species (CSCs)[72,73]. The resulting correlation data were visualized using the Gephi (https://gephi.org/) to build a force-driven network, where nodes represent bacterial species, and edges define their relationships. The node size is related to the IndVal obtained in the indicator species analysis, whereas the edge color shows the type of interaction, i.e., positive (green) or negative (red).

Collection of fresh infant stool samples

Fresh stool samples were collected from infants during scheduled postnatal visits. Eligible infants were born at term after an uncomplicated pregnancy and vaginal delivery. Inclusion criteria required that neither the mothers nor the infants had received antibiotics or probiotics during the perinatal period (intrapartum and postnatal), and that the infants were overall healthy at the time of sampling. Stool samples were collected by the parents using sterile containers and immediately stored at 4 °C. All samples were transferred to the laboratory within 24 h of collection and stored at −80 °C until further processing. Informed written consent was obtained from the parents or legal guardians prior to enrolment in the study, in accordance with the principles of the Declaration of Helsinki. The study protocol was approved by the Ethics Committee (Approval Code no. 25103).

Isolation of infant gut-associated bacteria and DNA extraction

To enhance the in vitro cultivation of infant gut microbes, fresh infant fecal samples were processed under strict anaerobic conditions and serially diluted (10^{-1} to 10^{-9}), then handled with two different protocols. In the traditional bacterial isolation procedure, diluted samples were inoculated on agar plates using 26 different media (Table 2). To promote the cultivation of fastidious microorganisms, a total of 32 growth factors were added, individually or in combination, to some media (Table 2). Plates were incubated for three days at 37 °C in an anaerobic chamber, after which at least ten colonies per medium were picked and transferred to Hungate tubes for traditional anaerobic cultivation at 37 °C until visible growth (~ 24h–72h, depending on the colony). In parallel, an enrichment workflow was performed by inoculating the 10^{-2} dilution of each fresh fecal sample into BACT/ALERT FN PLUS anaerobic blood culture bottles (bioMérieux) supplemented with defibrinated sheep blood and 0.2 µm filtered rumen. After incubation at 37 °C for 7, 14, and 21 days, aliquots from each bottle were plated onto the same 26 culture media as mentioned above (Table 2) and incubated at 37 °C. Selected colonies were then transferred to Hungate tubes and grown anaerobically at 37 °C until visible growth. For both workflows, after growth in Hungate tubes, 1 mL of culture was withdrawn, washed twice with phosphate-buffered saline (PBS), and resuspended in 100 µL of sterile water. Identification was performed using matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS), considering a score ≥1.9 as the threshold for reliable species-level identification. For isolates yielding lower scores, identification was pursued through 16S rRNA gene sequencing following DNA extraction.

Bacterial DNA extraction was performed using a standard rapid glass beads protocol. Briefly, 1 mL of culture was centrifuged at 6000 rpm for 5 min, then washed and resuspended in 1 mL of distilled water. The cell suspension was combined with approximately 250 mg of glass beads and homogenized using a bead-beater homogenizer for 2 min. To enhance DNA extraction, aliquots were then placed on ice for 2 min. This process was repeated twice. Finally, samples were centrifuged at 10,000 rpm for 2 min, and resulting the supernatant was collected and quantified through spectrophotometric analysis.

Bacterial genome sequencing and assembly

The genome sequences of 182 infant gut isolated were determined by GenProbio Srl (Parma, Italy) using NextSeq 2000 and MiSeq platforms (Illumina, San Diego, CA). Genome libraries were prepared using Illumina Nextera XT DNA Library Preparation Kit (Illumina, San Diego, CA) and quantified through a fluorometric Qubit quantification system (Life Technologies, USA). The obtained genome libraries were then loaded on a 2200 TapeStation instrument (Agilent Technologies, USA) and normalized to 4 nM. Sequencing was performed using MiSeq Reagent Kit v3 (600-cycles) (Illumina, San Diego, CA), or NextSeq 1000/2000 Reagents Kit (300-cycles) (Illumina, San Diego, CA), with a 1–2% PhiX spike-in. Paired fastq files were parsed with SPAdes assembler v4.0[74]. De novo genomic assemblies were performed using default parameters enabling the flag option "-isolates", coupled with a list of k-mer sizes of 21, 33, 55, 77, 99, and 127. The ORFs of the newly decoded genome were predicted using Prodigal[75] and annotated using the MEGAnnotator2 pipeline[76].

In silico prediction of microbial interaction

To investigate microbial interactions in terms of competition or complementarity, we employed the R package *RevEcoR*[48]. First, the metabolic potential of each microbial genome was inferred using KEGG Orthology and Hidden Markov Model[77], generating KO profiles that were used to identify the set of seed metabolites, i.e., compounds that each organism requires from the environment[78,79]. Pairwise metabolic interaction networks were reconstructed in *RevEcoR* by analyzing the overlap of seed metabolites among microbial species to infer potential competitive relationships (i.e., species requiring the same essential metabolites). Additionally, metabolite exchange was predicted as the fraction of seed metabolites of one species that appeared in the metabolic network, but not in the seed set, of its partner. Statistical significance of these interactions was computed using 1000 permutations ($p < 0.05$).

Batch culture fermentation

The four bacterial strains identified as reference representatives for *Bacteroides fragilis* (318 F), *Bacteroides uniformis* (324 F), *Bifidobacterium longum* subsp. *longum* (PRL2016), and *Bifidobacterium bifidum* (PRL2015) were cultivated both individually (as monocultures) and in pairwise combinations (bi-associations) using an in vitro gut-simulating medium (Infant Gut Super Medium, IGSM), previously described by Alessandri et al.[80], which mimics the conditions of the infant colonic environment. Cells were inoculated at approximately 1×10^7 cells/mL, both in bi-associations and monocultures. The experiments were performed in triplicate and incubated in anaerobe conditions for 8 h. After the incubation time, 2 mL per culture were collected for further analysis, while the remaining volume was

centrifuged (7000 rpm for 15 min) to separate the cell pellet from the supernatant. All aliquots were stored at −80 °C until RNA extraction.

Estimation of bacterial growth by qPCR

The 2 mL aliquot of each condition (monocultures and bi-associations) was subjected to DNA extraction using the QIAmp DNA Stool Mini kit following the manufacturer's instructions (Qiagen, Germany). Quantification of *Bifidobacterium* and *Bacteroides* strains in both monocultures and co-cultures (expressed as cell counts per mL) was carried out via quantitative PCR (qPCR), using DNA extracted from batch-grown cultures. Species-specific primers targeting single-copy genes were used for each strain: *Bacteroides fragilis* (Fw: 5'-GTACACTGCTCGAGATTATG-3', Rv: 5'-GTCGTCCTGAAACACATAG-3')[81], *Bacteroides uniformis* (Fw: 5'-TCTTCCGCATGGTAGAACTATTA-3', Rv: 5'-ACCGTGTCTCAGTTCCAATGTG-3')[82], *Bifidobacterium longum* subsp. *longum* (Fw: 5'-GGCATTCTCGAATCCTGTCT-3', Rv: 5'-ACAACCTTGCCGTAGGTGTC-3'), and *Bifidobacterium bifidum* (Fw: 5'-GCGAACAATGATGGCACCTA-3', Rv: 5'-GTCGAACACCACGACGATGT-3'). For each strain used, standard curves were generated by quantifying chromosomal DNA and performing serial dilutions to obtain concentrations ranging from 10^3 to 10^9 copies of double-stranded DNA per mL. Negative controls (no DNA) were included in each run. The standard curves were generated using the CFX96 software (Bio-Rad).

Bacterial RNA extraction and sequencing

Total RNA from each bacterial culture was isolated as previously described[54]. Briefly, cell pellets were initially suspended in 1 mL of QIAZOL lysis reagent (Qiagen, United Kingdom) and transferred into tubes pre-loaded with 0.8 g of 106 µm glass beads (Sigma). Mechanical cell disruption was carried out using a bead beater, applying alternating cycles of 2 min of vigorous agitation followed by 2 min of rest on ice. After lysis, samples were centrifuged at 12,000 rpm for 15 min, and the aqueous phase, containing the RNA, was carefully recovered. RNA purification was then completed using the RNeasy Mini Kit (Qiagen, Germany) following the manufacturer's guidelines.

RNA integrity was assessed with the TapeStation 2200 system (Agilent Technologies, United States), while concentration and purity were determined spectrophotometrically (Eppendorf, Germany). For transcriptomic sequencing, between 100 ng and 1 µg of total RNA underwent ribosomal RNA depletion using the QIAseq FastSelect kit specific for 5S/16S/23S rRNAs (Qiagen, Germany), according to the provided protocol. The efficiency of rRNA removal was verified using the TapeStation 2200.

Library preparation for whole transcriptome analysis was performed using the TruSeq Standard mRNA kit (Illumina, San Diego, United States), and sequencing was conducted on a NextSeq 500 platform with the high-output v2.5 kit (150 cycles, single-end), following Illumina's recommendations.

Differential gene expression analysis and assessment of cross-talk index

Following sequencing, raw reads were subjected to quality control with METAnnotatorX2[68], filtering out sequences with average quality scores below 20, lengths shorter than 150 bp, and those mapping to bacterial ribosomal loci. The remaining high-quality reads (average ± SD: 2,627,062.1 ± 1,264,713.203 per sample) were subsequently aligned to the genome sequence of each bacterial strain using BWA[83]. Transcript-level quantification was achieved using the htseq-count function in HTSeq, employing the "union" mode[84]. Raw count data were normalized by applying CPM (counts per million) to remove lowly expressed genes (CPM < 1) and TMM (trimmed mean of M-values) normalization for accurate differential expression analysis with the EdgeR package[85]. Expression changes were reported as log2 fold changes (logFC) by comparing each strain in co-culture to its corresponding monoculture. Volcano plots were generated to visualize gene-level changes by integrating logFC and statistical significance (FDR *p*-values). For each individual strain within a co-culture, a cross-talk index was calculated as the proportion of significantly differentially expressed genes relative to its total annotated gene repertoire, as previously described[54,55].

Statistical analysis

PERMANOVA and ANOSIM analyses were performed on Rstudio (version 2024.12.1 + 563, R version 4.3.2) using 999 permutations to assess *p*-values for population differences in PCoA with "adonis2" function in the R package *vegan* (2.5-7). The significance of infant-related variables (predefined infant age groups, feeding types, delivery mode, and geography) in explaining the total variation was assessed using the "envfit" function from the *vegan* package, based on 999 permutations (https://rdrr.io/cran/vegan/man/envfit.html). Since information on feeding type and delivery mode was not consistently available across all studies, their impact on the microbiota was assessed with EnvFit using the subset of samples for which these metadata were available (Neonatal, $n = 868$; pre-weaning, $n = 1261$; post-weaning, $n = 2002$). Mann–Whitney *U* test (using IBM SPSS Statistics software) or Kruskal-Wallis test (using the "kruskal.test" function in R version 4.3.2, Rstudio version 2024.12.1 + 563), followed by Dunn's *post hoc* test, were used to perform non-parametric comparisons. Pyramid histograms were generated using IBM SPSS Statistics (version 25.0) to visualize significant differences. In multiple comparisons, False Discovery Rate (FDR) or Bonferroni (BF) corrections were applied using the "p.adjust" function of R (version 4.3.2) and statistical significance was set at 95% level. The *ggplot2* package was utilized for graphical representation of the data.

Data availability

The raw data from the genomic sequencing of bacteria isolated from infant feces has been deposited in the NCBI database under the accession code PRJNA1279005. All publicly available metagenomic datasets supporting the findings of this study can be obtained through the accession code reported in the Supplementary Materials.

Received: 30 June 2025; Accepted: 14 November 2025;
Published online: 05 December 2025

References

1. Collado, M. C., Rautava, S., Aakko, J., Isolauri, E. & Salminen, S. Human gut colonisation may be initiated in utero by distinct microbial communities in the placenta and amniotic fluid. *Sci. Rep.* **6**, 23129 (2016).
2. Aagaard, K. et al. The placenta harbors a unique microbiome. *Sci. Transl. Med* **6**, 237ra65 (2014).
3. Palmer, C., Bik, E. M., DiGiulio, D. B., Relman, D. A. & Brown, P. O. Development of the human infant intestinal microbiota. *PLoS Biol.* **5**, 1556–1573 (2007).
4. Fahur Bottino, G. et al. Early life microbial succession in the gut follows common patterns in humans across the globe. *Nat. Commun.* **16**, 660 (2025).
5. Linehan, K., Dempsey, E. M., Ryan, C. A., Ross, R. P. & Stanton, C. First encounters of the microbial kind: perinatal factors direct infant gut microbiome establishment. *Microbiome Res. Rep.* **1**, 10 (2022).
6. Collado, M. C., Cernada, M., Baüerl, C., Vento, M. & Pérez-Martínez, G. Microbial ecology and host-microbiota interactions during early life stages. *Gut Microbes* **3**, 352–365 (2012).
7. Biagioli, V., Volpedo, G., Riva, A., Mainardi, P. & Striano, P. From birth to weaning: a window of opportunity for microbiota. *Nutrients* **16**, 272 (2024).
8. Xu, D. & Wan, F. Breastfeeding and infant gut microbiota: influence of bioactive components. *Gut Microbes* **17**, 2446403 (2025).
9. Schwab, C. The development of human gut microbiota fermentation capacity during the first year of life. *Micro. Biotechnol.* **15**, 2865–2874 (2022).
10. Ho, N. T. et al. Meta-analysis of effects of exclusive breastfeeding on infant gut microbiota across populations. *Nat. Commun.* **9**, 4169 (2018).

11. Arboleya, S., Saturio, S. & Gueimonde, M. Impact of intrapartum antibiotics on the developing microbiota: a review. *Microbiome Res. Rep.* **1**, null–null (2022).
12. Fallani, M. et al. Determinants of the human infant intestinal microbiota after the introduction of first complementary foods in infant samples from five European centres. *Microbiology* **157**, 1385–1392 (2011).
13. Vallès, Y. et al. Microbial succession in the gut: directional trends of taxonomic and functional change in a birth cohort of Spanish infants. *PLoS Genet* **10**, e1004406 (2014).
14. Horwell, E., Bearn, P. & Cutting, S. M. A microbial symphony: a literature review of the factors that orchestrate the colonization dynamics of the human colonic microbiome during infancy and implications for future health. *Microbiome Res. Rep.* **4**, 1 (2024).
15. Shao, Y. et al. Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. *Nature* **574**, 117–121 (2019).
16. Xiao, L., Wang, J., Zheng, J., Li, X. & Zhao, F. Deterministic transition of enterotypes shapes the infant gut microbiome at an early age. *Genome Biol.* **22**, 243 (2021).
17. Lugli, G. A. et al. Comprehensive insights from composition to functional microbe-based biodiversity of the infant human gut microbiota. *NPJ Biofilms Microbiomes* **9**, 25 (2023).
18. Mancabelli, L. et al. Multi-population cohort meta-analysis of human intestinal microbiota in early life reveals the existence of infant community state types (ICSTs). *Comput Struct. Biotechnol. J.* **18**, 2480–2493 (2020).
19. Tarracchini, C. et al. Genetic strategies for sex-biased persistence of gut microbes across human life. *Nat. Commun.* **14**, 4220 (2023).
20. Arrieta, M. C., Stiemsma, L. T., Amenyogbe, N., Brown, E. & Finlay, B. The intestinal microbiome in early life: health and disease. *Front Immunol.* **5**, 427 (2014).
21. 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).
22. Infant and young child feeding n.d. https://www.who.int/news-room/fact-sheets/detail/infant-and-young-child-feeding (accessed June 10, 2025).
23. Bifidobacteria: insights into the biology of a key microbial group of early life gut microbiota - Search n.d. https://www.bing.com/search?q=Bifidobacteria%3A+insights+into+the+biology+of+a+key+microbial+group+of+early+life+gut+microbiota&cvid=b6d628128deb4e00a9ea38a06b32a4b0&aqs=edge..69i57j69i58.294j0j9&FORM=ANAB01&PC=U531 (accessed February 3, 2023).
24. Koletzko, B. et al. National recommendations for infant and young child feeding in the World Health Organization European Region. *J. Pediatr. Gastroenterol. Nutr.* **71**, 672–678 (2020).
25. Pérez-Escamilla, R., Buccini, G. S., Segura-Pérez, S. & Piwoz, E. Perspective: should exclusive breastfeeding still be recommended for 6 months?. *Adv. Nutr.* **10**, 931–943 (2019).
26. Koenig, J. E. et al. Succession of microbial consortia in the developing infant gut microbiome. *Proc. Natl. Acad. Sci. USA* **108**, 4578–4585 (2011).
27. Ravel, J. et al. Vaginal microbiome of reproductive-age women. *Proc. Natl. Acad. Sci. USA* **108**, 4680–4687 (2011).
28. Arumugam et al. Enterotypes of the human gut microbiome. *Nature* **473**, 174–180 (2011).
29. Costea, P. I. et al. Enterotypes in the landscape of gut microbial community composition. *Nat. Microbiol.* **3**, 8–16 (2017).
30. Tannock, G. W. et al. Comparison of the compositions of the stool microbiotas of infants fed goat milk formula, cow milk-based formula, or breast milk. *Appl Environ. Microbiol* **79**, 3040–3048 (2013).
31. Roger, L. C., Costabile, A., Holland, D. T., Hoyles, L. & McCartney, A. L. Examination of faecal Bifidobacterium populations in breast- and formula-fed infants during the first 18 months of life. *Microbiology* **156**, 3329–3341 (2010).
32. Avershina, E. et al. Bifidobacterial succession and correlation networks in a large unselected cohort of mothers and their children. *Appl Environ. Microbiol.* **79**, 497–507 (2013).
33. Turroni, F. et al. Diversity of bifidobacteria within the infant gut microbiota. *PLoS One* **7**, e36957 (2012).
34. Ojima, M. N. et al. Priority effects shape the structure of infant-type Bifidobacterium communities on human milk oligosaccharides. *ISME J.* **16**, 2265–2279 (2022).
35. André, A. C., Debande, L. & Marteyn, B. S. The selective advantage of facultative anaerobes relies on their unique ability to cope with changing oxygen levels during infection. *Cell Microbiol.* **23**, e13338 (2021).
36. Friedman, E. S. et al. Microbes vs. chemistry in the origin of the anaerobic gut lumen. *Proc. Natl. Acad. Sci. USA* **115**, 4170–4175 (2018).
37. Coker, M. O. et al. Infant feeding alters the longitudinal impact of birth mode on the development of the gut microbiota in the first year of life. *Front Microbiol* **12**, 642197 (2021).
38. Knights, D., Costello, E. K. & Knight, R. Supervised classification of human microbiota. *FEMS Microbiol Rev.* **35**, 343–359 (2011).
39. Zhang, L., Wang, Y., Chen, J. & Chen, J. RFtest: a robust and flexible community-level test for microbiome data powerfully detects phylogenetically clustered signals. *Front Genet* **12**, 749573 (2022).
40. De Cáceres, M. & Legendre, P. Associations between species and groups of sites: Indices and statistical inference. *Ecology* **90**, 3566–3574 (2009).
41. De Cáceres, M., Legendre, P. & Moretti, M. Improving indicator species analysis by combining groups of sites. *Oikos* **119**, 1674–1684 (2010).
42. Kubosova, K., Brabec, K., Jarkovsky, J. & Syrovatka, V. Selection of indicative taxa for river habitats: a case study on benthic macroinvertebrates using indicator species analysis and the random forest methods. *Hydrobiologia* **651**, 101–114 (2010).
43. Dizzell, S. et al. Investigating colonization patterns of the infant gut microbiome during the introduction of solid food and weaning from breastmilk: a cohort study protocol. *PLoS One* **16**, e0248924 (2021).
44. Gotoh, A. et al. Sharing of human milk oligosaccharides degradants within bifidobacterial communities in faecal cultures supplemented with Bifidobacterium bifidum. *Sci. Rep.* **8**, 13958 (2018).
45. Sakanaka, M. et al. Evolutionary adaptation in fucosyllactose uptake systems supports bifidobacteria-infant symbiosis. *Sci. Adv.* **5**, eaaw7696 (2019).
46. Shao, Y. et al. Primary succession of Bifidobacteria drives pathogen resistance in neonatal microbiota assembly. *Nat. Microbiol* **9**, 2570–2582 (2024).
47. Xiao, M. et al. Cross-feeding of bifidobacteria promotes intestinal homeostasis: a lifelong perspective on the host health. *NPJ Biofilms Microbiomes* **10**, 47 (2024).
48. Cao, Y., Wang, Y., Zheng, X., Li, F. & Bo, X. RevEcoR: an R package for the reverse ecology analysis of microbiomes. *BMC Bioinforma.* **17**, 294 (2016).
49. Gao, X. et al. Predicting personalized diets based on microbial characteristics between patients with superficial gastritis and atrophic gastritis. *Nutrients* **15**, 4738 (2023).
50. Sarkar, I., Sen, G., Bhattacharyya, S., Gtari, M. & Sen, A. Inter-cluster competition and resource partitioning may govern the ecology of Frankia. *Arch. Microbiol* **204**, 326 (2022).
51. Díaz, R. & Garrido, D. Screening competition and cross-feeding interactions during utilization of human milk oligosaccharides by gut microbes. *Microbiome Res. Rep.* **3**, 12 (2024).
52. Tarracchini, C. et al. Optimal representative strain selector—a comprehensive pipeline for selecting next-generation reference strains of bacterial species. *NAR Genom. Bioinform* **6**, lqae173 (2024).
53. Rios-Covian, D. et al. Interactions between Bifidobacterium and Bacteroides species in cofermentations are affected by carbon

sources, including exopolysaccharides produced by bifidobacteria. *Appl. Environ. Microbiol.* **79**, 7518–7524 (2013).
54. Turroni, F. et al. Deciphering bifidobacterial-mediated metabolic interactions and their impact on gut microbiota by a multi-omics approach. *ISME J.* **10**, 1656–1668 (2016).
55. Rizzo, S. M. et al. Molecular cross-talk among human intestinal bifidobacteria as explored by a human gut model. *Front Microbiol.* **15**, 1435960 (2024).
56. Garrido, D. et al. Endo-β-N-acetylglucosaminidases from infant gut-associated bifidobacteria release complex N-glycans from human milk glycoproteins. *Mol. Cell. Proteom.* **11**, 775–785 (2012).
57. Bell, A. & Juge, N. Mucosal glycan degradation of the host by the gut microbiota. *Glycobiology* **31**, 691–696 (2021).
58. Kitaoka, M. Bifidobacterial enzymes involved in the metabolism of human milk oligosaccharides. *Adv. Nutr.* **3**, 422S–429SS (2012).
59. Glenwright, A. J. et al. Structural basis for nutrient acquisition by dominant members of the human gut microbiota. *Nature* **541**, 407–411 (2017).
60. Martens, E. C., Koropatkin, N. M., Smith, T. J. & Gordon, J. I. Complex glycan catabolism by the human gut microbiota: the bacteroidetes sus-like paradigm. *J. Biol. Chem.* **284**, 24673–24677 (2009).
61. Brown, H. A. & Koropatkin, N. M. Host glycan utilization within the Bacteroidetes Sus-like paradigm. *Glycobiology* **31**, 697–706 (2021).
62. Marcobal, A. et al. Bacteroides in the infant gut consume milk oligosaccharides via mucus-utilization pathways. *Cell Host Microbe* **10**, 507–514 (2011).
63. Wernroth, M. L. et al. Development of gut microbiota during the first 2 years of life. *Sci. Rep.* **12**, 9080 (2022).
64. Stewart, C. J. et al. Temporal development of the gut microbiome in early childhood from the TEDDY study. *Nature* **562**, 583–588 (2018).
65. Nandagire, W. H. et al. Exploring cultural beliefs and practices associated with weaning of children aged 0-12 months by mothers attending services at Maternal Child Health Clinic Kalisizo Hospital, Uganda. *Pan Afr. Med J.* **34**, 47 (2019).
66. Marvin-Dowle, K., Soltani, H. & Spencer, R. Infant feeding in diverse families; the impact of ethnicity and migration on feeding practices. *Midwifery* **103**, 103124 (2021).
67. Martens, E. C., Chiang, H. C. & Gordon, J. I. Mucosal glycan foraging enhances fitness and transmission of a saccharolytic human gut bacterial symbiont. *Cell Host Microbe* **4**, 447–457 (2008).
68. Milani, C. et al. METAnnotatorX2: a comprehensive tool for deep and shallow metagenomic data set analyses. *MSystems* **6**, 101128msystems0058321 (2021).
69. Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput Appl Math.* **20**, 53–65 (1987).
70. Dufreˆne, M., Dufreˆne, D. & Legendre, P. Species assemblages and indicator species: the need for a flexible asymmetrical approach. *Ecol. Monogr.* **67**, 345–366 (1997).
71. Breiman, L. Random forests. *Mach. Learn* **45**, 5–32 (2001).
72. Pons, P. & Latapy, M. Computing communities in large networks using random walks. *Lect. Notes Computer Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinforma.)* **3733**, 284–293 (2005).
73. Yang, Z., Algesheimer, R. & Tessone, C. J. A comparative analysis of community detection algorithms on artificial networks. *Sci. Rep.* **6**, 1–18 (2016).
74. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
75. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinforma.* **11**, 119 (2010).
76. Lugli, G. A. et al. MEGAnnotator2: a pipeline for the assembly and annotation of microbial genomes. *Microbiome Res. Rep.* **2**, 15 (2023).
77. Aramaki, T. et al. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).
78. Handorf, T., Ebenhöh, O. & Heinrich, R. Expanding metabolic networks: scopes of compounds, robustness, and evolution. *J. Mol. Evol.* **61**, 498–512 (2005).
79. Borenstein, E., Kupiec, M., Feldman, M. W. & Ruppin, E. Large-scale reconstruction and phylogenetic analysis of metabolic environments. *Proc. Natl. Acad. Sci. USA* **105**, 14482–14487 (2008).
80. Alessandri, G. et al. Exploring species-level infant gut bacterial biodiversity by meta-analysis and formulation of an optimized cultivation medium. *Npj Biofilms Microbiomes* **8**, 1–12 (2022).
81. Buzun, E. et al. A bacterial sialidase mediates early-life colonization by a pioneering gut commensal. *Cell Host Microbe* **32**, 181–190.e9 (2024).
82. Tong, J., Liu, C., Summanen, P., Xu, H. & Finegold, S. M. Application of quantitative real-time PCR for rapid identification of Bacteroides fragilis group and related organisms in human wound samples. *Anaerobe* **17**, 64–68 (2011).
83. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
84. Anders, S., Pyl, P. T. & Huber, W. HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
85. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

Acknowledgements

We thank GenProbio Srl for the financial support of the Laboratory of Probiogenomics. Part of this research is conducted using the High-Performance Computing (HPC) facility of the University of Parma. M.V. and postdoctoral research fellowship of C.T. were funded by the European Union, NextGeneration EU, PNRR-M4C2- I1.1, PRIN 2022 - Project Code 20229LEB99 - CUP Code D53D23014150006, Project title: Disentangling the molecular interplay between the gut microbiota and the host in the first stages of life (I-MAP). F.T. is funded by the Italian Ministry of Health through the Bando 414 Ricerca Finalizzata (Grant no. GR-2018-12365988). D.v.S. is member of the APC Microbiome Ireland research center, which is funded by Science Foundation Ireland (SFI; now called Research Ireland) under Grant Numbers 12/RC/2273 and 16/SP/3827.

Author contributions

C.T., G.L.: formal analysis, writing—original draft, writing—review and editing. L.M., G.A.L.: methodology, software, formal analysis, writing—review and editing. G.L., E.G., A.M., S.M.R., A.V., S.G.V.: investigation (fecal sample collection, bacterial isolation, co-culture experiments, RNA-seq, DNA sequencing), Data curation. S.A., F.T., D.v.S., C.M., M.V.: conceptualization, supervision, project administration, writing—review and editing. All authors: critical revision and final approval of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41522-025-00868-7.

Correspondence and requests for materials should be addressed to Marco Ventura.