communications biology

Article

https://doi.org/10.1038/s42003-024-07015-6

Longitudinal analyses of infants' microbiome and metabolome reveal microbes and metabolites with seemingly coordinated dynamics

Hao Wu [1,10], Douglas V. Guzior [2,3,10], Christian Martin [3,4], Kerri A. Neugebauer[5], Madison M. Rzepka[3], Julie C. Lumeng[6,7], Robert A. Quinn [3] ✉ & Gustavo de los Campos [1,8,9] ✉

Population studies have shown that the infant's microbiome and metabolome undergo significant changes in early childhood. However, no previous study has investigated how diverse these changes are across subjects and whether the subject-specific dynamics of some microbes correlate with the over-time dynamics of specific metabolites. Using mixed-effects models, and data from the ABC study, we investigated the early childhood dynamics of fecal microbiome and metabolome and identified 83 amplicon sequence variants (ASVs) and 753 metabolites with seemingly coordinated trajectories. Enrichment analysis of these microbes and molecules revealed eight ASV families and 23 metabolite groups involving 1032 ASV-metabolite pairs with their presence-absence changing in a coordinated fashion. Members of the *Lachnospiraceae* (464/1032) and metabolites related to cholestane steroids (309/1032) dominated proportional shifts within the fecal microbiome and metabolome as infants aged.

Background

In the last two decades, microbiome research has revealed that the diversity and composition of the gut microbiome are related to lifestyle and can be predictive of health outcomes[1]. More recently, scientists began to investigate the association of gut microbiome with its corresponding metabolome (i.e., the collection of biomolecules resulting from host or microbial metabolism). Accumulating evidence points to inter-dependencies between these two omics which play diverse roles in physiology and pathology[2–5]. These interactions between microbes and metabolites can be complex. For instance, an over-time increase in the relative abundance of a microbe may lead to an increased abundance of the metabolites it produces and a simultaneous decrease in the abundance of the metabolites it consumes. Identifying metabolite-microbe pairs with correlated dynamics could offer deeper insights into the pathways linking lifestyle (e.g., dietary factors) to chronic and acute diseases and could provide potential diagnostic and therapeutic targets[6].

The repertoire of methods used for the analysis of microbiome-metabolome data is growing. However, most commonly used methods (e.g., correlation analysis[7,8] and multivariate reduced-rank methods[9–13]) are designed for cross-sectional data and cannot capture the longitudinal dynamics within subjects. This is particularly important for microbiome-metabolome studies involving infants because both the metabolome and the microbiome are known to undergo important changes in early childhood from infancy to age three years[14].

Longitudinal analysis[15] is a well-established area of research in statistical learning. In a random regression model, the longitudinal trajectory of an outcome (e.g., the presence-absence or the abundance of a metabolite) is modeled using a population curve (a trajectory shared across individuals)

[1]Department of Epidemiology and Biostatistics, Michigan State University, East Lansing, MI, USA. [2]Department of Microbiology, Genetics, and Immunology, Michigan State University, East Lansing, MI, USA. [3]Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI, USA. [4]Mass Spectrometry and Metabolomics Core, Michigan State University, East Lansing, MI, USA. [5]Department of Plant Soil and Microbiology, Michigan State University, East Lansing, MI, USA. [6]Department of Pediatrics, University of Michigan Medical School, Ann Arbor, MI, USA. [7]Department of Nutritional Sciences, University of Michigan School of Public Health, Ann Arbor, MI, USA. [8]Institute for Quantitative Health Science and Engineering, Michigan State University, East Lansing, MI, USA. [9]Department of Statistics and Probability, Michigan State University, East Lansing, MI, USA. [10]These authors contributed equally: Hao Wu, Douglas V. Guzior. ✉e-mail: quinnrob@msu.edu; gustavoc@msu.edu

and a subject-specific component that captures the individual's deviations from the population trajectory[16]. A few studies have analyzed longitudinal microbiome and metabolome data using mixed-effects models[17–19]; however, these studies focused on population changes (i.e., average changes across subjects) without zooming into inter-individual differences in trajectories. As the metabolome and microbiome composition could vary markedly between individuals[20], we cannot underrate the variability between subjects in the over-time changes in the metabolome and the microbiome and whether these changes show signs of coordination.

Therefore, in this study we used random regression models[21] and longitudinal microbiome (bacterial 16S rRNA amplicon sequence) and metabolome (untargeted liquid-chromatography tandem mass spectrometry, LC-MS/MS) data from the ABC Baby Study[22] to investigate longitudinal changes in these omics in early childhood. We used the inferred subject-specific longitudinal trajectories to identify metabolite-microbe pairs that, within subjects, appear to change in a coordinated fashion.

Results

The ABC Baby Study[22] recruited 284 infant-mother dyads from communities within 1-h driving distance from Ann Arbor, Michigan, between October 2015 and February 2019. Dyads were recruited when infants were between 1 and 10 weeks of age.

Repeated assessments of infants' body weight, length, feeding and eating behavior, and dietary records occurred repeatedly across ages 2 weeks, 2 months, 4 months, 6 months, 9 months, and 12 months of age. Of the enrolled infants, 118 provided stool samples collected using swabs with wooden handles for analysis. These samples were used to generate ASV (amplicon sequence variant) counts using 16S rRNA amplicon sequencing and untargeted LC-MS/MS metabolomics collected at the above time points in the first year of life for infant subjects.

Data pre-processing included removing molecules absent in over 90% of samples, removing subjects with measurements at less than three different time points, and normalization. This resulted in 147 ASVs being preserved across 369 samples and 2422 metabolites across 510 samples from 99 individuals (see Methods for more details). Of the ASVs and metabolites that passed the pre-processing steps, the average detection rates were relatively low (26.8% for metabolites and 28.2% for ASVs), indicating that either the specific metabolite or ASV was not present or, if it was, the abundance was below the minimum detectable level.

Among the infants that provided stool samples, approximately 50% were female (Supplementary Table S1); 61% were non-Hispanic white, 12% were non-Hispanic black, 8% were Hispanic, and 19% were from another ancestry. The distribution of the ancestry and ethnicity of the subjects included in this study (i.e. 99 subjects with at least three different time point measurements) resembled the Southeast Michigan population[23]. Additional demographic, anthropometric, and milk-feeding characteristics of the participants are provided in Supplementary Table S1.

Infant metabolomes and microbiomes show contrasting trajectories in diversity driven by changing feature richness and evenness

We used the Shannon index[24] to investigate the diversity in the microbiome and metabolome through the first year of life. The metabolome had a much higher diversity than the microbiome–this is consistent with previous reports[17]. The longitudinal profile of the Shannon Indices showed that while infant microbiomes become more diverse over time (Fig. 1b, $\hat{\beta} = 0.0811 \pm 0.0087$), the metabolome profiles appear to have a more stable

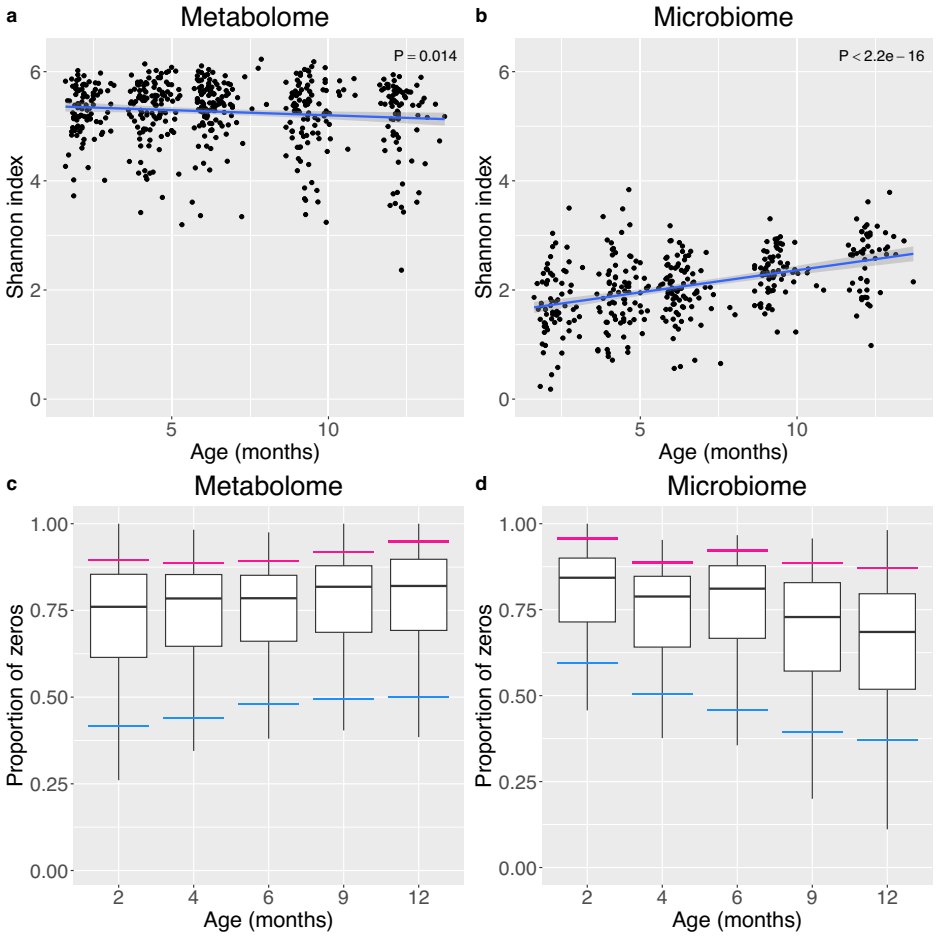

Fig. 1 | Trajectories in alpha-diversity and presence/absence vary in metabolome and microbiome. Temporal changes in Shannon diversity within the infant fecal **a** metabolome and **b** microbiome, with the blue lines and shadows indicating the regression lines and 95% confidence regions. The proportion of zeros present within the infant **c** metabolome and **d** microbiome with blue and pink lines indicating the 10th and 90th percentiles, respectively.

diversity with only a slight (statistically significant) over-time decrease in Shannon Index (Fig. 1a, $\hat{\beta} = -0.0189 \pm 0.0077$).

Regarding the between-sample dissimilarity metrics, there was a small over-time increase in dissimilarity for the metabolome (Supplementary Fig. S1a, $P < 2e{-}16$) and no significant over-time changes in between-sample dissimilarity in the microbiome (Supplementary Fig. S1b, $P = 0.0622$).

The changes in diversity described above were paralleled by a slight reduction in metabolite richness and evenness, and an increase in microbial richness and evenness (Supplementary Fig. S2). Likewise, there was an over-time reduction in the proportion of microbes that were not detected (Fig. 1d) and an over-time increase in the proportion of un-detected metabolites (Fig. 1c).

Modeling longitudinal trajectories using random regressions

In the previous section we documented changes in the diversity of the microbiome and the metabolome using metrics that summarize what is observed across samples, metabolites and microbes. To uncover finer details on the longitudinal trajectories of individual microbes and metabolites, we used random regression models.

A longitudinal trajectory, $f(t;\boldsymbol{\theta})$, describes how the expected value of an outcome (e.g., the probability that a metabolite or microorganism is detected in a sample) changes over time (t). This trajectory is indexed by a set of parameters ($\boldsymbol{\theta}$, e.g., an intercept and a slope in the case of a linear model). In random regression models, these parameters are represented as the sum of a population parameter ($\boldsymbol{\theta}$) and subject-specific deviations ($\boldsymbol{\theta}_i$; here, $i = 1, \ldots, n$ indexes subjects in the sample) that make the longitudinal trajectory, $f\left(t;\boldsymbol{\theta} + \boldsymbol{\theta}_i\right)$ subject-specific. In standard random regression models, these subject specific parameters are assumed to follow IID (independent and identically distributed) Gaussian distributions with mean zero and parameter-specific variances. The random regression framework just described allows inferring population trajectories, $f(t;\boldsymbol{\theta})$ describing average changes, as well as subject-specific dynamics, $f\left(t;\boldsymbol{\theta} + \boldsymbol{\theta}_i\right)$.

Considering the large proportion of zeroes, we fit mixed-effects longitudinal logistic regressions[21] of the form:

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = (\mu + \mu_i) + t_{ij}(\beta_T + \beta_{Ti}) + F_i\beta_F + B_i\beta_B + H_i\beta_H + O_i\beta_O + M_{Bij}\beta_{M_B} + M_{Fij}\beta_{M_F} + M_{Cij}\beta_{M_C} + M_{Oij}\beta_{M_O}$$

to each ASV and metabolite. Above π_{ij} is the probability that subject i had a given ASV or metabolite present at time point j, $\mu + \mu_i$ is a subject-specific intercept, which has a population component (μ) and a subject-specific random effect $\mu_i \overset{iid}{\sim} N\left(0, \sigma^2_\mu\right)$; likewise, $\beta_T + \beta_{Ti}$ is a subject-specific slope that has a population term (β_T) and subject-specific deviations, $\beta_{Ti} \overset{iid}{\sim} N\left(0, \sigma^2_{\beta_T}\right)$. The remaining terms in the right-hand-side of the logit equation include a dummy variable for female (F_i), dummy variables for black (B_i), Hispanic (H_i), and other races (O_i, Asian, multiracial, or other race, non-Hispanic), and dummy variables for breast milk feeding (M_{Bij}), formula milk feeding (M_{Fij}), cow's milk feeding (M_{Cij}), and other milk feeding (M_{Oij}), which were included in the model to account for sex, ancestry, and milk feeding differences (Supplementary Data 1 and 2).

Random regressions had superior performance in predicting metabolite and ASV abundance

To validate the random regression models, we estimated the cross-validation AUC (area under the receiver operating characteristic curve)[25] for each metabolite and ASV separately. To perform this prediction analysis, for each microbe and metabolite, we randomly masked the abundance at some time points, used the remaining data to fit models, and used the fitted model to predict the masked abundance (see Methods for further details). The analysis revealed that 81.5% of the metabolites and 79.5% of the ASVs were better predicted with the mixed-effects logistic model compared to the standard logistic regression model (Supplementary Fig. S3). The standard logistic regression model assumes homogeneity of the trajectories across subjects. Therefore, the superior performance of the random-effects logistic regression models provides evidence of heterogeneity across subjects and shows that this approach was able to learn subject-specific aspects of the longitudinal trajectories. We use the AUC results to filter the metabolites and ASVs; specifically, further analyses were based on 85 ASVs and 1107 metabolites that had AUC $\geq$ 0.65 (Supplementary Fig. S4). For our sample size, an AUC $\geq$ 0.65 is significantly greater than 0.5 (the AUC of a random classifier) at a significance level of 0.05.

Population trajectories describe general prevalence trends in microbiome-metabolome data

We first investigated the average changes in the proportion of zeros for each metabolite and ASVs (i.e., $f(t;\boldsymbol{\theta})$ the estimate populations curves). These trajectories describe whether a specific ASV or metabolite became more prevalent or rarer over time.

Among the metabolites that had a significant over-time change (FDR-adjusted p-value < 0.05) in the probability of being detected, most of them showed an increase in the proportion of zeros over time, indicating that these molecules were detected in a smaller proportion of the samples as time progressed (Table 1). On the other hand, among the ASVs that had a significant change in the proportion of zeroes, most had a significant increase in the probability of being detected. While, on average, the probability of zeros increased in the metabolome and decreased in the microbiome, within each of these omics we found metabolites and ASVs that had (statistically significant) longitudinal changes in the opposite direction. For instance, there were 78 metabolites for which we detected a significant over-time decrease in the probability of being not detected (Table 1 and Fig. 2a, blue lines). Standard errors for the predicted curves of each metabolite and microbe for various time points are summarized in Supplementary Table S2.

Among the metabolites with a relatively high initial proportion of zeros (e.g., larger than 0.8), most showed an over-time decline in the proportion of zeroes–these were metabolites that were detected in a small proportion of the samples at the first time point and became more common as children aged. Conversely, metabolites with a low proportion of zeros at the first time point were more likely to exhibit an over-time increase in the proportion of zeros (i.e., become rarer). However, even within this group, many molecules appear to have become detectable as children grow.

Longitudinal trajectories are highly heterogeneous across subjects

Using the fitted random-effects models we predicted the expected changes in the probability of zero in the first year of life for each subject-feature (metabolite or ASV) combination. The results showed that for every metabolite and microbe there was high heterogeneity between individuals (Fig. 3). For instance, even for metabolites with a significant increase in the probability of not being detected (yellow triangles in panel a of Fig. 3) we found several individuals with the opposite trajectory (gray dots to the left of zero). To illustrate how heterogeneous these trajectories can be, we show in Supplementary Fig. S5 three ASVs: one with high heterogeneity in trajectories, one with intermediate heterogeneity, and one with low heterogeneity.

Interestingly, for the metabolites and ASVs without a significant change in the proportion of zeroes at the population level (gray triangles, especially those enclosed in a purple dashed box in Fig. 3), we observed bi-modals trajectories. This happened because some subjects had such metabolites or ASVs absent at the early time point and present at a later time point, or vice versa. Meanwhile, more subjects had such metabolites or ASVs constantly present or absent. Overall, the changes in the population became close to zero.

Microbiome maturation correlates with changes in the metabolome

We used the estimated subject-level expected change for the proportion of zeroes (gray dots in Fig. 3) to identify metabolite-ASV pairs that appear to change in a coordinated manner. To do this, for each metabolite-ASV pair, we computed Pearson's correlation (across subjects) between the predicted change in the probability of not being detected and tested if the correlation significantly differed from zero (FDR-adjusted p-value < 0.05). To mitigate the influence of outliers and to make our analysis more robust, we used Bootstrapping[26] and reported an (approximately) unbiased estimate for the correlation coefficients[27]. Here, a positive correlation indicates that the metabolite and the ASV pair show concordant trajectories (e.g., both increasing or decreasing in the first-year change of not-being-detected probability, Supplementary Fig. S6a–c) and a negative correlation indicates discordant trajectories (Supplementary Fig. S6d–f). Among the 94,095 metabolite-ASV pairs, 3990 pairs had correlations significantly different than zero (FDR-adjusted p-value < 0.05) suggesting that these pairs may be changing, within subject, in a coordinated manner. These pairs involved 83 ASVs and 753 metabolites. Of these correlations, 1974 (49.5%) were positive and the remaining were negative.

Table 1 | Numbers of molecules with (statistically significant) over-time change in the proportion of zeros

	Increasing	Decreasing	No significant change
Metabolome	362 (32.7%)	78 (7.0%)	667 (60.3%)
Microbiome	5 (5.9%)	41 (48.2%)	39 (45.9%)

Fisher's exact test p-value < 2.2e−16.

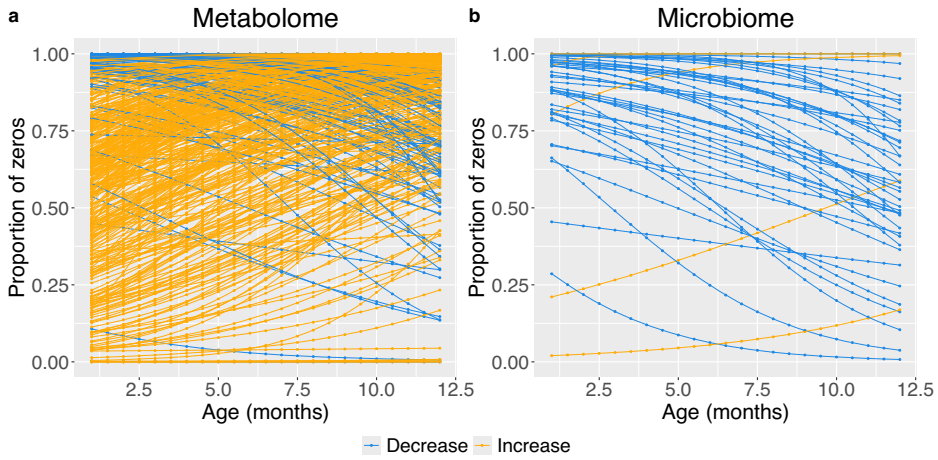

Fig. 2 | Longitudinal changes in the proportion of zeros for metabolome and microbiome reveal population-level omics dynamics. Each line represents the estimated probability that an individual metabolite (**a**) and an individual ASV (**b**) are not detected versus age. Yellow and blue colors are used for metabolites and ASVs that showed statistically significant over-time changes in the probability of not being detected (blue is used for a decline and yellow is used to signal an increase in the probability of not being detected). ASVs and metabolites that did not have a significant over-time change are not shown.

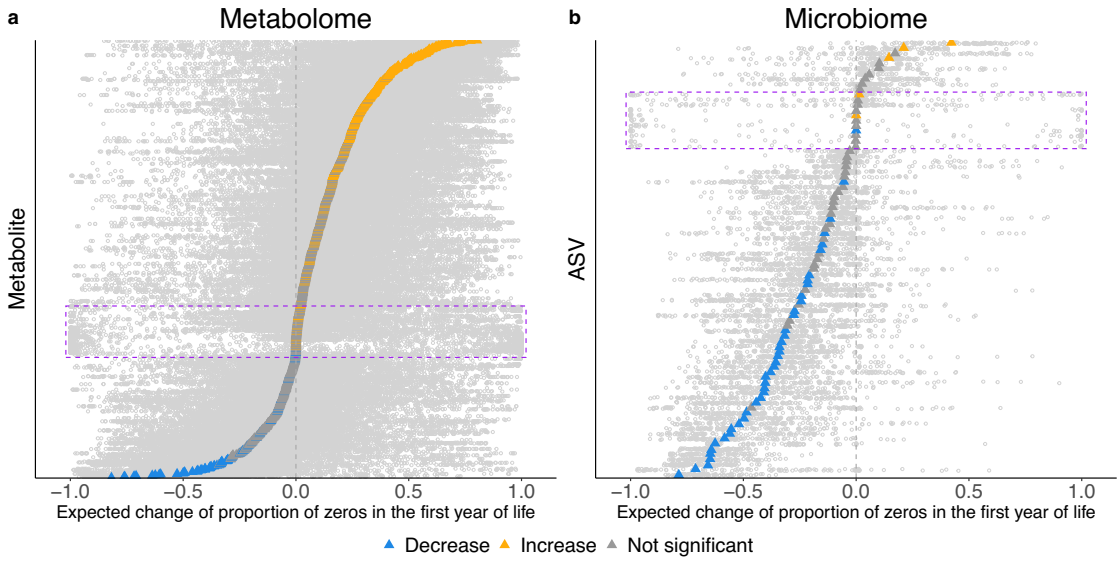

Fig. 3 | Expected change in probability of zero for metabolites and ASVs by feature and individuals. Each gray dot represents the expected change in the probability of not being detected (probability of zero) for a metabolite (**a**) or an ASV (**b**) in a subject. Dots within the same row correspond to the same feature (metabolite or ASV). The colored triangle is the average change in probability of zero for the corresponding feature (the color used for the average expected changes matches the colors of the curves in Fig. 2: blue indicates a significant decline and yellow an increase in the probability of not being detected. Besides, gray represents features with no significant change in prevalence over time). The horizontal dispersion of dots within a row represents how heterogeneous a metabolite or ASV trajectory was across subjects. The dots within the purple rectangle include metabolites and ASVs with overall non-significant changes in the proportion of zeros.

Eight microbial families were enriched for coordinated longitudinal dynamics with several metabolite groups

To gain insight into the correlation analysis reported above, we conducted enrichment analyses to identify groups of metabolites and ASV families that are enriched for significant correlation with elements from the other omic, i.e., ASV families enriched for significant correlations with metabolites and metabolites groups enriched for significant correlation with the longitudinal dynamics of microbes. The enrichment analysis identified eight microbial families (35 ASVs) and twenty-three metabolite classes (169 metabolites) which were significantly enriched (FDR-adjusted p-value < 0.05) for coordinated longitudinal changes. The correlation of the longitudinal dynamics of the top-5 microbial families and top-5 metabolite groups from the enrichment analysis are shown in Fig. 4a, and the full set of results are provided in Supplementary Data 3 and 4. For instance, one of the ASVs within the *Clostridiaceae* family often changed in a coordinated fashion with metabolites classified as cholestane steroids and open-chain polyketides (Fig. 4a, red strings between groups E-F and E-I). Significant associations were dominated by two members of the family *Lachnospiraceae*, a member of the genus *Rikenellaceae*, and a member of the genus *Bifidobacteriaceae*, and the bulk of associations were with compounds related to cholestane steroids, a class of C27 bile acids (Fig. 4b, Supplementary Data 4). The most significant co-correlation was between Cluster6939, an unknown cholestane steroid, and the ASV related to *Lachnospiraceae* (FDR-adjusted p-value = 3.8e−22). Other microbes of interest in the correlation analysis included *Bifidobacterium bifidum*, which is also associated with changing cholestane steroids. This bacterial group is well known for its dynamics in early human life, particularly as one changes from breastfeeding to solid food[28].

We further investigated the 35 ASVs and 169 metabolites from the enriched groups. For the microbiome, as expected from the general trends previously reported in this study, most ASVs appear to have an increase in the probability of being detected over time (Fig. 5b, Supplementary Data 5). These increases are paralleled by the acquisition of members of the *Lachnospiraceae*, common bile acid metabolizing bacteria with several implications in host health[29]. The top ASV significantly increasing in population prevalence based on sample trajectories is a member of the *Veillonellaceae* group, with 5.2% of 2-month samples increasing to 62.1% of 12-month samples containing the ASV. Of the metabolites that were significantly increasing or decreasing in population prevalence, most were structurally related to cholestane steroids as determined by MS^2 spectra (Fig. 5a, Supplementary Data 6). One of the top metabolites that significantly decreased in probability of being detected, Cluster 5583 (646.4167 m/z), did not match any annotated compounds however, molecular networking revealed spectral similarity of this unknown to an annotated cholestane steroid (Cluster 6185), which also decreased in sample prevalence as infants matured. MS^2 spectral alignment supported the relatedness of these metabolites, but Cluster 5583 contained a set of peaks exhibiting a 194.0417 ± 0.0024 Da increase from those matching Cluster 6185 (Supplementary Fig. S7). This shift matches what one would expect following glucuronidation, a known method of molecular detoxification for bile acids[30].

Discussion

Both the microbiome and the metabolome undergo important changes in early childhood. In this study, after performing standard population-level diversity analysis, we proceeded to study the subject-specific dynamics of the metabolome and the microbiome and used this to identify microbe-metabolite pairs with seemingly coordinated trajectories.

Our diversity analyses revealed that as infants matured through the first year, there was a slight increase in the proportion of metabolites whose abundance became lower than the minimum detectable value. This contrasted with a significant over-time increase in the richness of the microbiome. One might expect that as a microbiome becomes more diverse, a phenomenon described in many studies of developing infants[5,31–33], including ours, its chemical diversity will follow; however, our data does not support that hypothesis. As microbiome diversity increases, which is typically due to a more diverse diet, so too does its metabolic potential, especially as microbial members within the community fill various metabolic niches. However, this may not result in a more diverse metabolome if the colonization of diverse microbial populations increases the availability of novel metabolic pathways, resulting in the more complete breakdown of molecules into fewer endpoint metabolites. This would lead to the fecal metabolome becoming relatively less diverse, compared to the changes in diversity in the microbiome, and may result in an over-time increase in metabolite zero-inflation if the resident microbiota metabolizes a greater proportion of molecules present. Another possible explanation for the slight over-time decreases in metabolite diversity observed in our study may involve host absorption of metabolites, or other aspects of microbe-microbe interactions that could contribute to these observations. Clearly, how to study a metabolome that may include both host and microbe contributions and disentangling the contributions of the two is a subject that deserves further investigation.

Previous studies have focused on the over-time dynamics of the microbiome and metabolome at the population level (i.e., average changes across subjects)[17–19]. However, to the best of our knowledge, no previous study has investigated how heterogeneous such dynamics are across subjects and whether there are microbes and metabolites that exhibit, within-subject, correlated trajectories. The random regression analysis conducted enabled us to make inferences for individual metabolites and microbes both at the population and subject-specific levels.

Using random-effects logistic regression models, we found that for every metabolite or microbe, there was high heterogeneity between individuals' longitudinal trajectories. A correlation analysis of the subject-specific dynamics identified 3990 metabolite-ASV pairs with significantly correlated trajectories. A taxonomic analysis of the microbes and metabolites involved in those pairs found eight microbial families and twenty-three metabolite classes that change in seemingly coordinated manners over time. These associations revealed strong associations between members of the *Lachnospiraceae* and cholestane steroids. Other families with significant microbe-metabolite pairs included *Prevotellaceae* and *Bifidoacteriaceae*, the latter known to be a dynamic group changing during the first year of life due to its metabolism of breast milk sugars[28]. The significant role that members of *Lachnospiraceae* play in the development and maturation of both the infant microbiome and metabolome is not well understood, but this analysis indicates they have strong associations with cholestane steroids, likely representing bile acids. The *Lachnospiraceae* are well known for their metabolism of bile acids and many express the enzyme bile salt hydrolase, which is the gateway step for the production of secondary bile acids[28].

The approach used in this study to characterize the subject-specific dynamics of features from two omics could be further applied to other types of omics data, such as transcriptomic or proteomic datasets. Depending on the outcomes of interest, the longitudinal analyses may use models for presence/absence (as done in our study) or regression models for quantitative outcomes (including zero-inflated ones if needed). Longitudinal changing patterns could be obtained from each data set separately. Then, a comprehensive picture of how different molecules coordinate in the human body can be obtained by correlating within-subject changes in pairs of omics features as we did for metabolite-microbe pairs.

The analyses presented in this study are, of course, not free of limitations. The subjects involved in this project only included one-third of the ABC Baby Study due to the limited number of enrolled subjects providing stool samples to generate microbiome and metabolome data. However, the distribution of the sex and race/ethnicity of the subjects included in this study resembled the substantial ABC Study population and the Southeast Michigan population[23]. Besides, subject-specific inferences can be influenced by outliers which could happen due to true biological events or technical problems. To mitigate this problem, in the correlation analysis we used a resampling (Bootstrapping) technique which is robust against outliers. Finally, the difficulty in distinguishing the host contribution to the gut metabolome from other sources (e.g., microbes) makes the interpretation of results challenging.

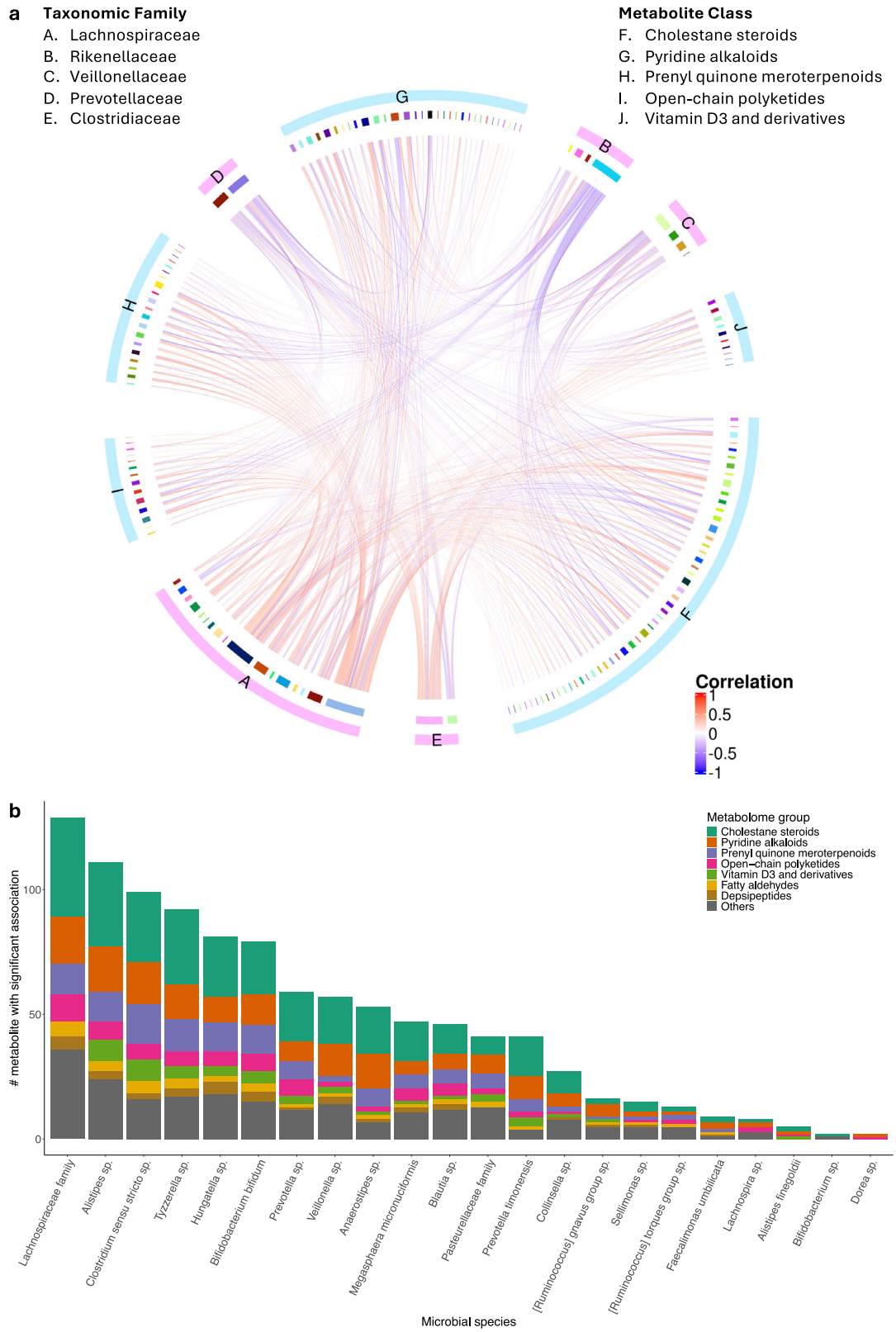

Fig. 4 | Correlation patterns in the longitudinal dynamics of the microbial families and metabolite groups identified in the enrichment analysis. **a** Co-correlation network between metabolites and ASVs where each node in the inner circle represents one metabolite or ASV and tracks in the outer circle indicate the taxonomy groups of inner molecules (groups A-E are ASV families, and groups F-J are metabolite groups). The colors of the strings indicate the correlation of longitudinal trends between two nodes. **b** The number of significant metabolite associations for each ASV, colored by Natural Products Classifier, with the most specific taxonomic assignment displayed.

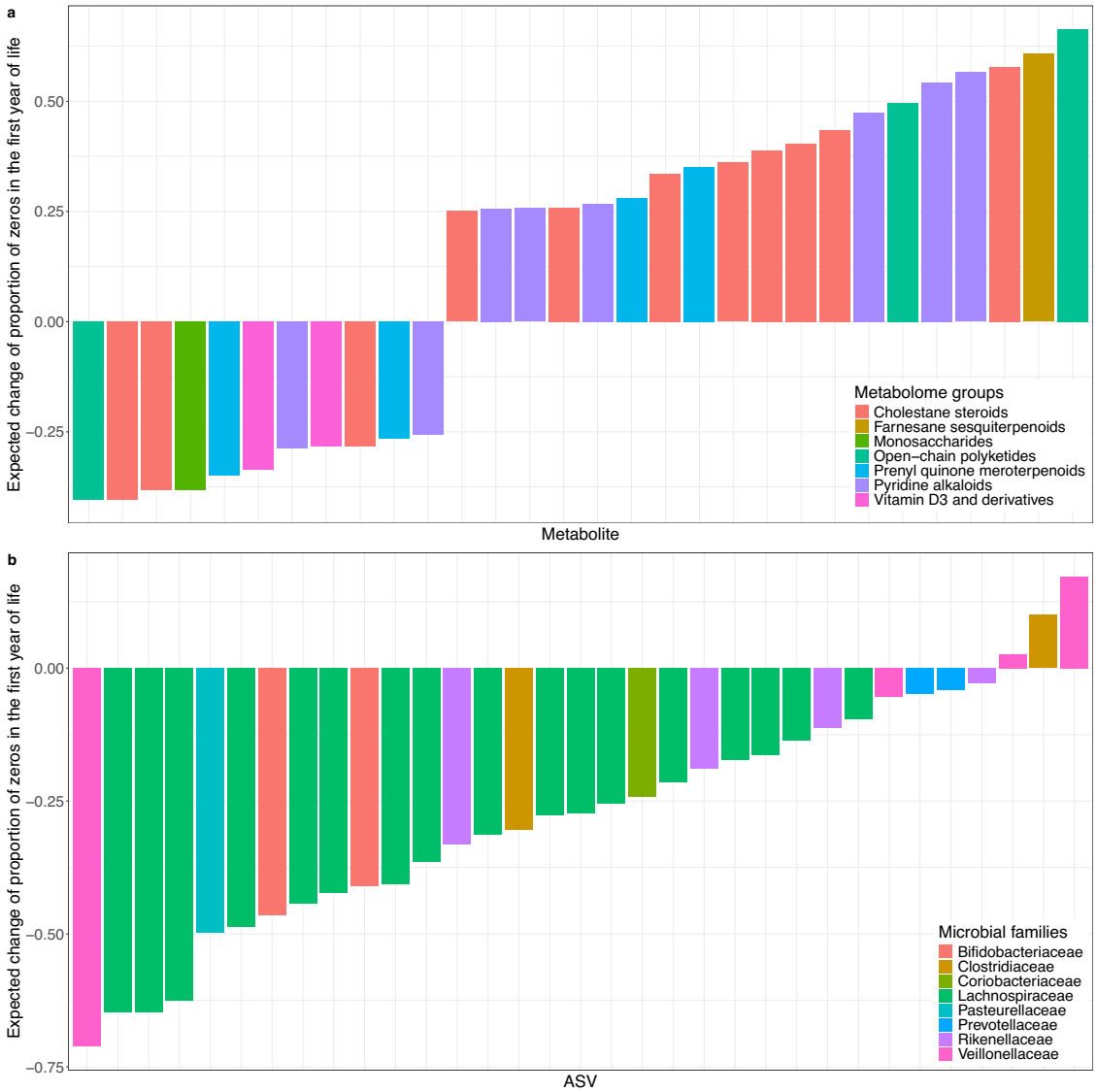

Fig. 5 | Temporal shifts in zero-proportions of metabolite and microbial features identified in the enrichment analysis. The change in probability of not being detected for **a** the top 30 metabolites from significantly enriched groups (FDR-adjusted p-value < 0.05), where metabolites are colored by molecular class as determined by the Natural Products Classifier, and **b** The 35 ASVs from significantly enriched families (FDR-adjusted p-value < 0.05), where ASVs are colored by assigned taxonomic family. Results for all the significantly enriched metabolites and ASVs can be seen in Supplementary Data 5 & 6, respectively.

Conclusion

We proposed a novel approach to investigate the subject-specific longitudinal dynamics of microbiome-metabolome data. Using this approach, and microbiome-metabolome data collected from the ABC Baby Study, we identified several microbes and metabolites that appear to have coordinated dynamics in early childhood. These associations reveal the significant role of *Lachnospiraceae* family and cholestane steroids in the development and maturation of both the infant microbiome and metabolome.

Methods

Sample collection and inclusion criteria

The ABC Baby study obtained written informed consent from mothers for their infants. The study was approved by the University of Michigan institutional review board (HUM00103575).

The ABC Baby Study[22] recruited 284 infant-mother dyads from communities within a 1-h driving distance from Ann Arbor, Michigan, between October 2015 and February 2019. Dyads were recruited when infants were between 1 and 10 weeks of age.

Repeated assessments of infants' body weight, length, feeding and eating behavior, and dietary records occurred repeatedly across ages 2 weeks, 2 months, 4 months, 6 months, 9 months, and 12 months of age.

Of the enrolled infants, 118 provided stool samples collected by the guardians at home using swabs. The guardians were instructed to store the samples in their refrigerator until they were collected by the study team within 72 h and stored frozen until analysis. The swabs used were dual-headed BBL culture swabs with wooden stalks. The amount of feces present on each swab was not controlled but all samples were treated identically after collection. Prior to extraction, the swabs were split for separate microbiome and metabolome analysis of each by placing the swab head in a 96-well deep well plate. These samples were used to generate ASV (amplicon sequence variant) counts using 16S rRNA amplicon sequencing and untargeted LC-MS/MS metabolomics collected at the above time points in the first year of life for infant subjects.

We included in our analyses data from dyads that contributed stool samples for at least three different time points, this led to a total of 99 dyads and 369 stool samples that had associated metabolome and microbiome.

Liquid chromatography-tandem mass spectrometry

Metabolite extractions were performed in 96 deep-well plates (Thermo), where one swab head was inserted into 600 µL cold methanol, the plate was then sealed with a rubber mat, and incubated at 4 °C overnight. Swab heads were then removed. Plates containing the resulting metabolite extracts were centrifuged for 10 min at 4100 × *g* to pellet cell debris followed by storage at −80 °C prior to liquid chromatograph-tandem mass spectrometry analysis (LC-MS/MS). Metabolite extracts were diluted 1:1 (v:v) in 50% methanol (v:v, water) prior to LC-MS/MS analysis. Ultra-high performance liquid chromatography (UPLC) was performed using a Vanquish Autosampler (Thermo) and separation was achieved using an Acquity UPLC BEH C-18 column, 2.1 mm × 100 (Waters). All analyses used a 10 µL injection volume, 0.4 mL min^{-1} flow rate, and 60 °C column temperature. Samples were eluted using a linear solvent gradient of water (A) and acetonitrile (B), each containing 0.1% formic acid, across a 12-min chromatographic run as follows: 0–1 min, 2% B; 1–8 min, 2–100% B; 8–12 min, 100% B; 10–12 min, 2% B. Mass spectrometry was performed using a Q Exactive Hybrid Quadrupole-Orbitrap Mass Spectrometer (Thermo). Data were collected using electrospray ionization in positive mode. MS^1 data were collected using a 35,000 resolution, AGC target of 1e6, maximum injection time of 100 ms, and a scan range set from 100 to 1500 *m/z* during min 1–10. Data-dependent MS^2 spectra were collected for the top 5 most abundant peaks identified in MS^1 survey scans. The resulting raw data files were converted to mzXML format via GNPS Vendor Conversion prior to data mining using MZmine3 (version 3.2.8)[34,35]. Outputs from MZmine were submitted to the Global Natural Products Social Molecular Networking Database (GNPS, University of California at San Diego) for molecular networking and spectral identification[36,37]. Due to the large number of unknown features present, SIRIUS structural prediction and molecular classification were employed for metabolites smaller than 850 Da[38].

DNA isolation and 16S rRNA amplicon sequencing

Genomic DNA from infant fecal swabs in the 96-well plates was isolated using the PowerFecal or PowerFecal Pro DNA extraction kits (Qiagen) according to the manufacturer's protocol. DNA extracts underwent quality control via PCR amplification using GoTaq Green Master Mix (Promega) with primers 27f (5'-AGAGTTTGATCCTGGCTCAG-3') and 1492r (5'-GGTTACCTTGTTACGACTT-3')6. Products were examined via gel electrophoresis. Extracted DNA was then sent to the Michigan State University RTSF Genomics Core for 16S-V4 amplicon sequencing using an Illumina MiSeq (Illumina). Briefly, the V4 hypervariable region of the bacterial 16S rRNA gene was amplified using Illumina-compatible, dual indexed primers 515f and 806r7[39]. Products were batch-normalized using a SequalPrep DNA Normalization plate (Invitrogen) and subsequent products recovered from the plates were pooled. Pools were cleaned and concentrated using a QIAquick Spin column (Qiagen) and AMPure XP magnetic beads (Beckman Coulter). The pool quality was analyzed and quantified using a combination of Qubit dsDNA HS (Invitrogen), 4200 TapeStation HS DNA1000 (Agilent), and Collibri Illumina Library Quantification qPCR (Invitrogen) assays. Pools were loaded onto a single MiSeq v2 Standard flow cell and sequencing was carried out in a 2 × 250bp paired end format using a MiSeq v2 500 cycle reagent cartridge. Custom sequencing and index primers complementary to 515f and 806r oligomers were added. Base calling was performed by Illumina Real Time Analysis (RTA, version 1.18.54, Illumina), and the output of RTA was demultiplexed and converted to FastQ format with Bcl2fastq (version 2.20.0, Illumina).

Sample fastq files were uploaded to Qiita, a web-based QIIME2 platform, for initial analysis and taxonomy assignment[40,41]. Sequence reads were trimmed to a read length of 200 bases and quality filtered using Deblur (version 2021.9). Taxonomy was assigned using the q2-feature-classifier against the 99% SILVA 16S rRNA database (release 138)[42,43].

Data pre-processing

Data processing and statistical analysis were performed in R 4.1.1[44]. Metabolome samples were filtered to only include samples collected using swabs with wooden handles due to overwhelming signals from plastic-handled swabs severely impacting subsequent analysis (data not shown). Additionally, samples collected at two weeks following birth were not included in this analysis due to a lack of available metadata. Microbiome data were initially rarefied to 5000 counts, without replacement, and samples unable to meet that threshold were excluded from further analysis. Resulting metabolome and microbiome data are featured with zero-inflation. For example, around 99% of the ASVs were not detected in more than 10% of subjects. Although our model could handle zero-inflated data, the molecules with excessive zeros provide little valuable information in the study. Therefore, we removed molecules with proportions of non-zeros less than 10% among samples. Then, to make different omics data more "integrable", we normalized the data by sum. Besides, to ensure enough sample sizes to compute the expected changes for each subject, we removed the subjects with less than three samples (subjects with samples collected at less than three different time points).

Mixed-effects longitudinal logistic regressions

We fitted mixed-effects longitudinal logistic regressions models[21] of the form:

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = (\mu+\mu_i) + t_{ij}(\beta_T+\beta_{Ti}) + F_i\beta_F + B_i\beta_B + H_i\beta_H + O_i\beta_O + M_{Bij}\beta_{M_B} + M_{Fij}\beta_{M_F} + M_{Cij}\beta_{M_C} + M_{Oij}\beta_{M_O}$$

to each ASV and metabolite. Above π_{ij} is the probability that subject *i* had a given ASV or metabolite present at time point *j*, $\mu+\mu_i$ is a subject-specific intercept, which has a population component (μ) and a subject-specific random effect $\mu_i \overset{iid}{\sim} N\left(0, \sigma_\mu^2\right)$; likewise, $\beta_T+\beta_{Ti}$ is a subject-specific slope that has a population term (β_T) and subject-specific deviations, $\beta_{Ti} \overset{iid}{\sim} N\left(0, \sigma_{\beta_T}^2\right)$. The remaining terms in the right-hand-side of the logit equation include a dummy variable for female (F_i), dummy variables for black (B_i), Hispanic (H_i), and other races (O_i, Asian, multiracial, or other race, non-Hispanic), and dummy variables for breast milk feeding (M_{Bij}), formula milk feeding (M_{Fij}), cow's milk feeding (M_{Cij}), and other milk feeding (M_{Oij}).

The models above-described were fitted using the 'glmer' from package 'lme4' (version 1.1-34)[45].

Evaluation of model prediction with testing data

We further examined the models' prediction performance in cross-validations and compared it with standard logistic regression models (without subject-specific random effects on the intercepts, μ_i, and slopes, β_{Ti}). To estimate prediction accuracy, for each metabolite or ASV, we randomly selected one sample from each subject to form the testing set. The remaining samples were used as the training set to fit the models (mixed-effects and standard logistic regression models) and predict the left-out testing samples. We then computed the area under the ROC curve (AUC, using the function 'auc' from package 'pROC' (version 1.18.0)[46]) between the predicted probabilities and the observed presence/absence outcome in the testing data. We repeated the above validation process 20 times and computed the average of the AUC for each metabolite and ASV. Using these validation results, we filtered out metabolites and ASVs that had an AUC that was not significantly greater than 0.5 ($P < 0.05$, for our sample size, that corresponds to an AUC ≥ 0.65, Supplementary Fig. S3). Therefore, all the longitudinal trajectories that we report, as well as the cross-subject correlation and enrichment analysis that followed are based on ASVs and metabolites with testing AUC ≥ 0.65.

Prediction of longitudinal trajectories

We used the fitted mixed-effects longitudinal logistic regression models to predict the expected change in the probability of detection for each metabolite and ASV. At the population level, the expected changes in the proportion of zeros between two different time points ($t_2 > t_1$) are (here we predicted for sex and ancestry indices - male, white, no milk feeding):

$$\Delta P = \left[1 - \frac{1}{1 + e^{-(\mu + t_2\beta_T)}}\right] - \left[1 - \frac{1}{1 + e^{-(\mu + t_1\beta_T)}}\right].$$

T-tests were used to determine if the expected change was significantly different from 0 using FDR-adjusted p-value and a threshold of 0.05 FDR.

Meanwhile, the expected changes in the proportion of zeros between two different time points at the individual level are:

$$\Delta P_i = \left\{1 - \frac{1}{1 + e^{-[(\mu+\mu_i)+t_2(\beta_T+\beta_{Ti})+\tau_i]}}\right\} - \left\{1 - \frac{1}{1 + e^{-[(\mu+\mu_i)+t_1(\beta_T+\beta_{Ti})+\tau_i]}}\right\},$$

$$\boldsymbol{\tau_i} = F_i\beta_F + B_i\beta_B + H_i\beta_H + O_i\beta_O + M_{Bi}\beta_{M_B} + M_{Fi}\beta_{M_F} + M_{Ci}\beta_{M_C} + M_{Oi}\beta_{M_O}.$$

where $i = 1, \ldots, n$ is an index for the subject. The curves and predicted changes we report is for 12 months during the first year of life (since our first samples were collected at month 2, we used $t_1 = 2$ and $t_2 = 14$.) The milk feeding practice was imputed by the closest time points.

Bootstrap analysis of Pearson correlation

To identify metabolite-ASV pairs that were changing in a seemingly coordinated fashion, we correlated the vectors containing the $\Delta P_i's$ for all subjects for a metabolite with that of a vector containing the $\Delta P_i's$ for an ASV. We did this for all metabolite-ASV pairs. For the sample size that we have, Pearson's correlation can be highly affected by outliers; therefore, to smooth out the influence of outliers we reported an (approximately unbiased) Bootstrap estimate for the correlation coefficients[27]

$$\hat{r} = \frac{1}{B}\sum_{b=1}^{B} r_b\left[1 + \frac{1 - r_b^2}{2(n-3)}\right]$$

where r_b is the traditional Pearson correlation coefficient for Bootstrap sample b. A standard error for these estimates was computed using the 100 Bootstrap samples and t-tests were used to determine if the correlation was significantly different from 0 using FDR-adjusted p-value and a threshold of 0.05 FDR.

Enrichment analysis

With the taxonomy grouping information for each metabolite and ASV, as well as the above results indicating whether each of these molecules had a significant longitudinal change, we used hypergeometric tests to identify groups of metabolites or ASVs that change significantly over time. Multiple testing corrections were performed via FDR with a significant threshold of 0.05. These taxonomic groups with expected relative abundance change coordinatively over time were used to form a network.

Data availability

Raw mass spectrometry data is available on MassIVE under ID MSV000092782 at https://doi.org/10.25345/C5DJ58S9M. LC-MS/MS spectral annotation and molecular networking by GNPS are available at https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=7454748a6baa406b909540b1c90a4e7e. 16S rRNA gene amplicon data were deposited in the EMBL-EBI European Nucleotide Archive and can be found under project PRJEB72674, study accession ERP157451. Data are also available on Qiita under study ID 14092 with subsequent analysis and taxonomy assignment available under analysis ID 48437.

Code availability

All the scripts used to implement the analyses presented in this manuscript are available at https://github.com/Harold-Wu/MiMe.

Received: 25 March 2024; Accepted: 4 October 2024;
Published online: 14 November 2024

References

1. Cho, I. & Blaser, M. J. The human microbiome: at the interface of health and disease. *Nat. Rev. Genet.* **13**, 260–270 (2012).
2. Zhernakova, A. et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science* **352**, 565–569 (2016).
3. Vatanen, T. et al. Variation in microbiome LPS immunogenicity contributes to autoimmunity in humans. *Cell* **165**, 842–853 (2016).
4. Fan, Y. & Pedersen, O. Gut microbiota in human metabolic health and disease. *Nat. Rev. Microbiol.* **19**, 55–71 (2021).
5. Stewart, C. J. et al. Temporal development of the gut microbiome in early childhood from the TEDDY study. *Nature* **562**, 583–588 (2018).
6. Vernocchi, P., Chierico, F. D. & Putignani, L. Gut microbiota profiling: metabolomics based approach to unravel compounds affecting human health. *Front. Microbiol.* **7**, 1144 (2016).
7. Morton, J. T. et al. Reply to: Examining microbe–metabolite correlations by linear methods. *Nat. Methods* **18**, 40–41 (2021).
8. Quinn, T. P. & Erb, I. Examining microbe–metabolite correlations by linear methods. *Nat. Methods* **18**, 37–39 (2021).
9. Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **2**, 559–572 (1901).
10. Gerlach, R. W., Kowalski, B. R. & Wold, H. O. A. Partial least-squares path modelling with latent variables. *Anal. Chim. Acta* **112**, 417–421 (1979).
11. Hotelling, H. Relations between two sets of variates. In *Breakthroughs in Statistics, Methodology and Distribution* 162–190 (Springer, 1992).
12. Dolédec, S. & Chessel, D. Co-inertia analysis: an alternative method for studying species–environment relationships. *Freshw. Biol.* **31**, 277–294 (1994).
13. Schönemann, P. H. A generalized solution of the orthogonal procrustes problem. *Psychometrika* **31**, 1–10 (1966).
14. Tang, Z.-Z. et al. Multi-omic analysis of the microbiome and metabolome in healthy subjects reveals microbiome-dependent relationships between diet and metabolites. *Front. Genet.* **10**, 454 (2019).
15. Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. *Applied longitudinal analysis* (John Wiley & Sons, 2012).
16. Henderson, C. R. Jr Analysis of covariance in the mixed model: higher-level, nonhomogeneous, and random regressions. *Biometrics*. **38**, 623–640 (1982).
17. Raghuvanshi, R. et al. High-resolution longitudinal dynamics of the cystic fibrosis sputum microbiome and metabolome through antibiotic therapy. *Msystems* **5**, 10–1128 (2020).
18. Lamichhane, S. et al. Dysregulation of secondary bile acid metabolism precedes islet autoimmunity and type 1 diabetes. *Cell Rep. Med.* **3**, 100762 (2022).
19. Aatsinki, A.-K. et al. Dynamics of gut metabolome and microbiome maturation during early life. MedRxiv Prepr. https://www.medrxiv.org/content/10.1101/2023.05.29.23290441v1 (2023).
20. Johnson, K. V.-A. Gut microbiome composition and diversity are related to human personality traits. *Hum. Microbiome J.* **15**, 100069 (2020).
21. Stiratelli, R., Laird, N., Ware, J. H. Random-effects models for serial observations with binary response. *Biometrics*. **40**, 961–971 (1984).
22. Reynolds, L. A. F. et al. Capacity for Regulation of Energy Intake in Infancy. *JAMA Pediatr.* **177**, 590–598 (2023).

23. Bureau USC. SEMCOG's Census 2020. Available from: https://www.semcog.org/census-2020 (2020).
24. Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).
25. Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36 (1982).
26. Efron, B. Bootstrap methods: another look at the jackknife. In *Breakthroughs in Statistics Methodology and Distribution* 569–593 (Springer, 1992).
27. Olkin, I., Pratt, J. W. Unbiased estimation of certain correlation coefficients. *Ann. Math Stat.* **29**, 201–211 (1958).
28. Laursen et al. Bifidobacterium species associated with breastfeeding produce aromatic lactic acids in the infant gut. *Nat. Microbiol.* **6**, 1367–1382 (2021).
29. Guzior, D. V. & Quinn, R. A. Microbial transformations of human bile acids. *Microbiome* **9**, 1–13 (2021).
30. Trottier, J. et al. Profiling serum bile acid glucuronides in humans: gender divergences, genetic determinants, and response to fenofibrate. *Clin. Pharm. Ther.* **94**, 533–543 (2013).
31. Mueller, N. T. et al. The infant microbiome development: mom matters. *Trends Mol. Med.* **21**, 109–117 (2015).
32. Moore, R. E. & Townsend, S. D. Temporal development of the infant gut microbiome. *Open Biol.* **9**, 190128 (2019).
33. Best, N. V. et al. Bile acids drive the newborn's gut microbiota maturation. *Nat. Commun.* **11**, 3692 (2020).
34. Pluskal, T. et al. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinforma.* **11**, 1–11 (2010).
35. Myers, O. D. et al. One step forward for reducing false positive and false negative compound identifications from mass spectrometry metabolomics data: new algorithms for constructing extracted ion chromatograms and detecting chromatographic peaks. *Anal. Chem.* **89**, 8696–8703 (2017).
36. Wang, M. et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **34**, 828–837 (2016).
37. Nothias, L.-F. et al. Feature-based molecular networking in the GNPS analysis environment. *Nat. Methods* **17**, 905–908 (2020).
38. Dührkop, K. et al. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat. Methods* **16**, 299–302 (2019).
39. Kozich, J. J. et al. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl. Environ. Microbiol.* **79**, 5112–5120 (2013).
40. Gonzalez, A. et al. Qiita: rapid, web-enabled microbiome meta-analysis. *Nat. Methods* **15**, 796–798 (2018).
41. Bolyen, E. et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* **37**, 852–857 (2019).
42. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2012).
43. Yilmaz, P. et al. The SILVA and "all-species living tree project (LTP)" taxonomic frameworks. *Nucleic Acids Res.* **42**, D643–D648 (2014).
44. R Core Team. *R: A language and environment for statistical computing* (R Core Team, 2013).
45. Bates, D. et al. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–48 (2015).
46. Robin, X. et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinforma.* **12**, 1–8 (2011).

Acknowledgements

We thank the participants of the ABC study as well as all the scientists and staff involved in participant recruitment and data collection. This work was supported by the National Institutes of Health under grant R01HD084163 and by Michigan State University.

Author contributions

H.W., D.V.G., J.C.L., R.A.Q., and G. de los C. conceived the study and participated in its design. J.C.L. organized sample collection and acquired funding. D.V.G., C.M., K.A.N., M.M.R., and R.A.Q. processed samples. D.V.G. and R.A.Q. generated data. D.V.G., K.A.N., and M.M.R. performed quality control. H.W. and G. de los C. designed the statistical models. H.W. and D.V.G. performed statistical analysis and visualized results. H.W. and D.V.G. wrote the manuscript. G. de los C. and R.A.Q. revised the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Ethics approval and consent to participate

The study was approved by the University of Michigan Institutional Review Board (HUM00103575). Mothers gave written informed consent for themselves and their infants. All ethical regulations relevant to human research participants were followed.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-024-07015-6.

Correspondence and requests for materials should be addressed to Robert A. Quinn or Gustavo de los Campos.

Peer review information *Communications Biology* thanks Partho Sen, and the other, anonymous, reviewers for their contribution to the peer review of this work. Primary Handling Editors: Sabina Leanti La Rosa and Tobias Goris. [A peer review file is available.]