Network-based representation learning reveals the impact of time and diet on the gut microbial and metabolomic environment of infants

Adelle Price[1], Sakaiza Rasolofomanana-Rajery[1], Keenan Manpearl[1], Charles E. Robertson[2], Nancy F. Krebs[3], Daniel N. Frank[2], Audrey E. Hendricks*[1], Arjun Krishnan*[1], Minghua Tang*[4]

*These authors contributed equally.

Affiliations:

[1]Department of Biomedical Informatics, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA
[2]Department of Medicine, Division of Infectious Diseases, University of Colorado Anschutz Medical Center, Aurora, CO 80045
[3]Department of Pediatrics, Section of Nutrition, University of Colorado Anschutz Medical
[4]Department of Food Science and Human Nutrition, Colorado State University, Fort Collins, CO 80526

Corresponding Author: Adelle Price

Abstract

While studies have explored differences in gut microbiome development regarding infant liquid diets (breastmilk, formula), surprisingly little is known about the impact of complementary foods on the infant gut microbiome. Indeed, current dietary recommendations for infants and toddlers are not formulated with knowledge of how the developing gut microbiome metabolizes complementary foods. Here, we investigated how different protein-rich foods (i.e., meat vs. dairy) affect fecal metagenomics and metabolomics during early complementary feeding from 5-12 months in formula-fed infants. We used a network node embedding to model the time-dependent, complex interactions between microbiome features, metabolomic compound features, and diet. We then used the embedded space to detect features associated with baseline or endpoint and meat or dairy diet- finding enriched networks of microbiome and metabolomic features, and compared the results to those found using a more traditional differential abundance analysis.

Introduction

Early-in-life colonization of microbes plays a critical role in the establishment and maturation of gut microbiome throughout childhood and into adulthood. Disruption of the optimal microbial succession may result in long-term health impairment[2]. Therefore, manipulation of the gut microbiome during infancy could result in a long-term programming impact on metabolism and disease risks of the host[3]. Research of the gut microbiome goes beyond 16S sequencing and usually includes the genetic features (fecal metagenomics) and metabolites (fecal metabolomics) of the gut microbiota. Gut microbes produce myriad small molecules through the metabolism of dietary components, *de novo* synthesis, and modification of excreted host molecules. Emerging research showed that colonic fermentation of ingested macronutrients results in metabolites with a direct impact on host physiology. As infants transition to complementary foods, gut microbial communities change in composition and functional capacity due to the altered nutritional milieu in the gut as well as concomitant physiological and immunological development by the host.

Among all modifiable factors that could impact infant gut microbiome development, diet is considered one of the most influential. Infant diet includes the liquid diet (breastmilk, formula) and the complementary diet. Early complementary feeding (~5 to 12 months of age) represents the time when solid foods are progressively introduced to infants as they no longer rely solely on breastmilk or formula. Different from toddlerhood (13-24 months), when the individual is fully exposed to table foods, early complementary feeding has less diversity, which makes it challenging and yet critical to provide evidence-based recommendations to meet nutrient needs and promote optimal gut microbial colonization. Unlike the liquid diet phase (breastmilk and infant formula), we know surprisingly little about the effect of complementary foods on gut microbiome development. Additionally, little is known about how and which microbiome and fecal metabolomic features interact with each other, especially at different developmental times and under different nutritional contexts. Therefore, current dietary recommendations for infants and toddlers are not formulated with knowledge of how the developing gut microbiome metabolizes complementary foods. Thus, it is critical to elucidate how complementary foods impact gut microbiome development.

Here, we use a three-step machine learning framework to enable analysis of high-dimensional, multiomic data over two time points investigating how different protein-rich foods (i.e., meat vs. dairy) affect fecal metagenomics and metabolomics during early complementary feeding in formula-fed infants. Specifically, we (1) embed high-dimensional fecal microbiome and metabolomics data along with individual characteristics (e.g., diet and timepoint) into the same space, (2) detect diet or timepoint associations, and (3) identify networks for microbe and metabolomic features that are enriched for different timepoints or diet.

Results

Experimental Overview

We performed a randomized controlled feeding trial in a cohort of 64 exclusively formula-fed 5-months-old infants recruited from the Denver metro area in Colorado, U.S. Infants were randomized to receive either pureed meats (n=32) or dairy foods (n=32) such as infant yogurt, cheese, and whey protein powder. Fruit and vegetable intake was not restricted and parents were encouraged to allow the infant's appetite to dictate their total intake. Infant stool samples used in the analyses here were collected at baseline (5 months) and at the end of the intervention (12 months). See **Methods** for a complete description.

We used untargeted metagenomic sequencing and metabolomics profiling on the longitudinal stool collections to interrogate the gut microbial profile. Sequenced microbiome features are reported with species-level taxonomic classifications as well as Cluster of Orthologous Groups (COG), KEGG Orthology

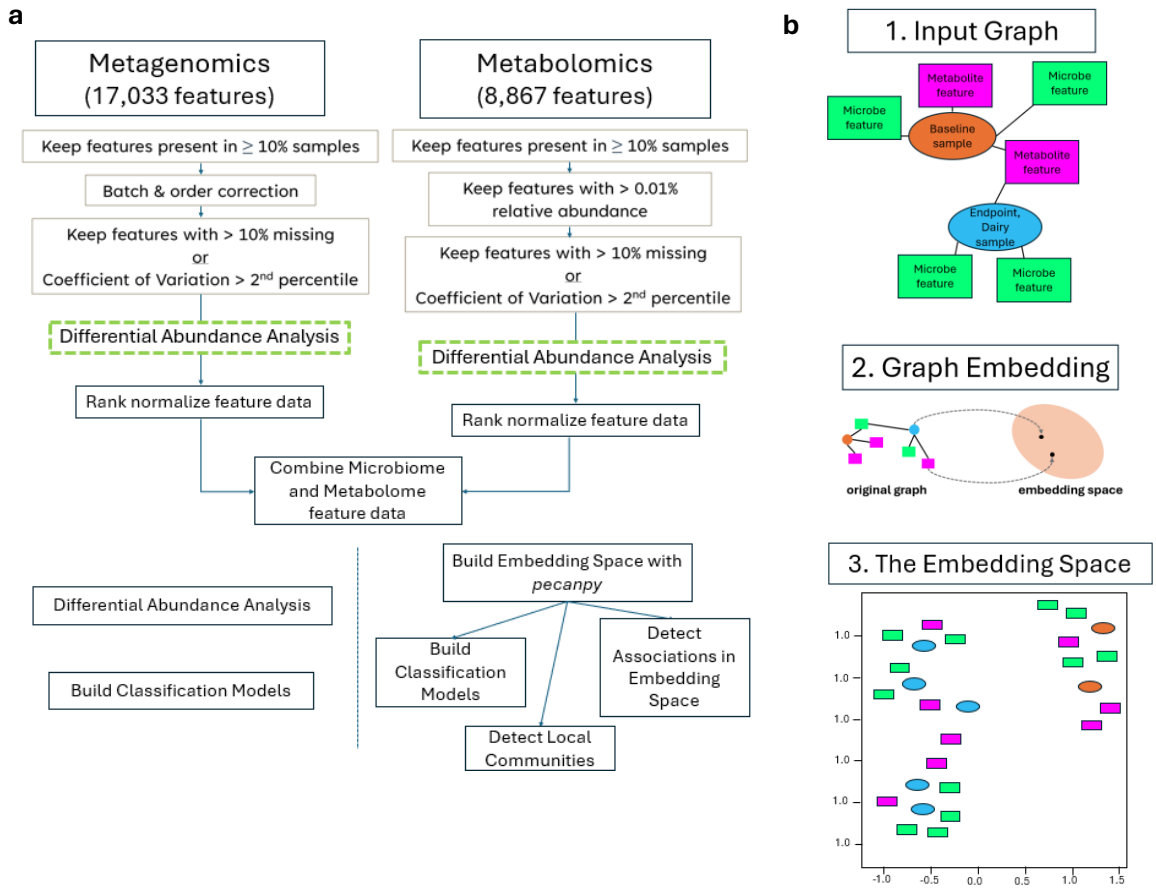

Figure 1. Experimental Overview. a, Data filtering pipelines for microbiome and metabolite data and study design to detect differences in timepoint and diet type. **b**, Schematic representation of graph embedding with *pecanpy*[1]: 1) Input 1-dimensional graph into *pecanpy*, 2) samples and features are mapped into an embedding space using weighted random walks, and 3) the embedding space captures spatial sample-sample, feature-feature, and sample-feature relationships.

(KO), and Protein Families Database (Pfam) functional classifications. Identified metabolites were annotated with Chemical Abstracts Service (CAS), Kyoto Encyclopedia of Genes and Genomes (KEGG), Human Metabolome Project (HMP) database, and LIPID MAPS identifiers. After filtering on missingness and coefficient of variation, 17,033 microbiome and 8,867 metabolome features were used for analysis (**Figure 1a**). See **Methods** for a complete description of the data filtering.

Here, we begin by quantifying differences between diet (meat and dairy) and timepoint (baseline and endpoint) with differential abundance analysis. However, because the landscape of microbes and their metabolites changes across both diet type and time here, differences in diet groups are not well captured with traditional approaches. To circumvent this challenge, we then use a machine learning approach to model the high-dimensional, time-dependent interactions between samples, microbe features, and metabolite features as a network with graph embedding (**Figure 1b**). Using the sample-microbe-metabolite network node embedding, we build and assess diet and time classification models, detect microbe and metabolite feature associations with time and diet, and identify closely linked microbe and metabolite features within the node embedding space.

Differential abundance analyses identifies changes in infant microbiome and metabolome over time but not between diet types

Principal Coordinate Analysis (PCoA) on both the microbiome and metabolomic compound rank normalized features across meat (n=23) and dairy (n=29) diets at 12 months showed no clear separation between diet types (**Figure 3a**). Additionally, tests for differences in variation between the meat and dairy groups across both microbiome and metabolome rank normalized features using omnibus testing with PERMANOVA of Bray-Curtis dissimilarities were not significant (R2=.02; unadjusted pval=0.47 and R2=.01; unadjusted pval=0.96 for microbiome and metabolome features, respectively) (**Figure 3a**). Lastly, we assessed differential abundance between meat and dairy groups using Man-Whitney U tests and found no significant differences in relative abundance of microbiome or metabolite features between the meat and dairy diet groups (**Figure 2c,d**).

Conversely, we observed distinct clustering in PCoA between the baseline (6 months, n=48) and endpoint (12 months, n=48) samples (**Figure 3c**). Additionally, PERMANOVA testing showed timepoint explained a small but significant amount of variation within the microbiome (R2=.06; p=1e-04) and metabolome features (R2=.04; p=1e-04) (**Figure 3c**). Differential abundance analysis between baseline and endpoint groups using paired Wilcoxon-rank sum tests revealed 5,842 significant differences in microbiome relative abundance and 1,949 significant differences in metabolite compound abundance between the baseline and endpoint samples. Microbiome and metabolome features with significant differential abundance across baseline and endpoint samples are reported in **Supplementary Table 1** and **Supplementary Table 2**, respectively. Enrichment analysis of microbiome taxa with significant (indicated by FDR<=0.05) abundance differences between baseline and endpoint samples revealed family-level significant enrichments at the baseline timepoint for *f_Enterobacteriaceae* and *f_Bifidobacteriaceae*. Endpoint was significantly enriched with family-level taxa *f_Lachnospiraceae, f_Oscillospiraceae*, *f_Sutterellaceae* and genus-level taxa *g_Roseburia*, *g_Blautia*, *g_Dorea*, *g_Eisenbergiella*, *and g_Holdemania*. Taxa and other microbiome feature enrichments in baseline and endpoint groups provided in **Supplementary Table 3**.

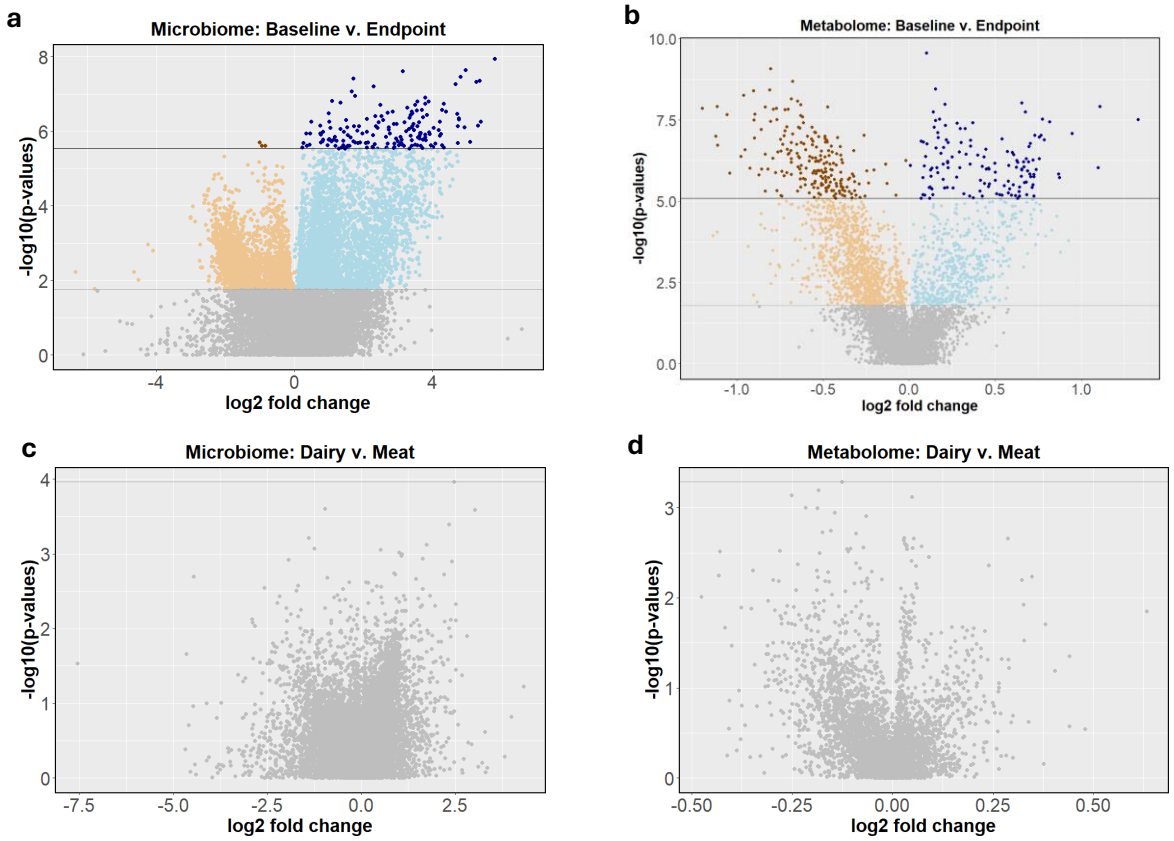

Figure 2. Identifying changes in microbiome and metabolome over time. Differential abundance analysis between **a**, microbiome and **b**, metabolome features at baseline and endpoint. Positive log2 fold changes indicate higher feature abundance at endpoint (dark and light blue) and negative log2 fold changes indicate higher abundance at baseline (dark and light orange). **c**, Differential abundance values of microbiome and **d**, metabolome features in dairy and meat groups. Positive log2 fold changes indicate higher feature abundance in meat groups; negative log2 fold changes indicated higher feature abundance in dairy groups. In all figures, grey lines indicate FDR threshold and black lines indicate Bonferroni correction threshold.

Graph embedding amplifies variation between meat and dairy diet groups

We used a graph embedding approach with *pecanpy*[1] to create a 128-dimension representation of the sample-microbe-metabolite landscape. Characterizing variation between meat and dairy groups within the node embedding space amplified differences between meat and dairy samples and allowed for inference regardless of fold change differences between groups. To assess variation across the sample node embeddings, we employed PCoA across cosine dissimilarity (1 minus cosine similarity) and tested for significant clustering using PERMANOVA of cosine dissimilarity. Notably, PCoA and PERMANOVA tests across diet types using sample network embeddings yielded both distinct and significant clustering among samples (R2=.06, p=0.01), indicating the network embeddings more adequately capture variation between meat and dairy diet samples compared to the rank normalized feature data (**Figure 3a, b**). And as previously seen within rank normalized features in timepoint and time by diet groups, we observed

significant clustering using the node embeddings across the baseline and endpoint samples (R2=.13; p=1e-04) (**Figure 3c, d**).

To further assess the utility of the embedding space and allow for classification of features as diet (meat or dairy) and timepoint (baseline or endpoint) groups we built classification models with sample node embeddings. Across all scenarios, 5 prediction models were built with L1 regularized logistic regression using the *glmnet* package in R. F1 score was used to evaluate model performance. Additionally, we used the same framework to build and assess classification models using the rank normalized metabolome and microbiome data to allow a point of performance comparison for classification models built with node embeddings. Models predicting diet groups (meat or dairy) built with node embedding improved the average F1 score by ~18% compared to models built with the rank normalized data (F1 score=65% and 83% for models built rank normalized data and node embeddings, respectively). Models built with node embeddings predicting time groups, however, had a ~9% decrease in performance compared to those built using rank normalized microbiome and metabolome data (F1 score= 96% and 87% for models

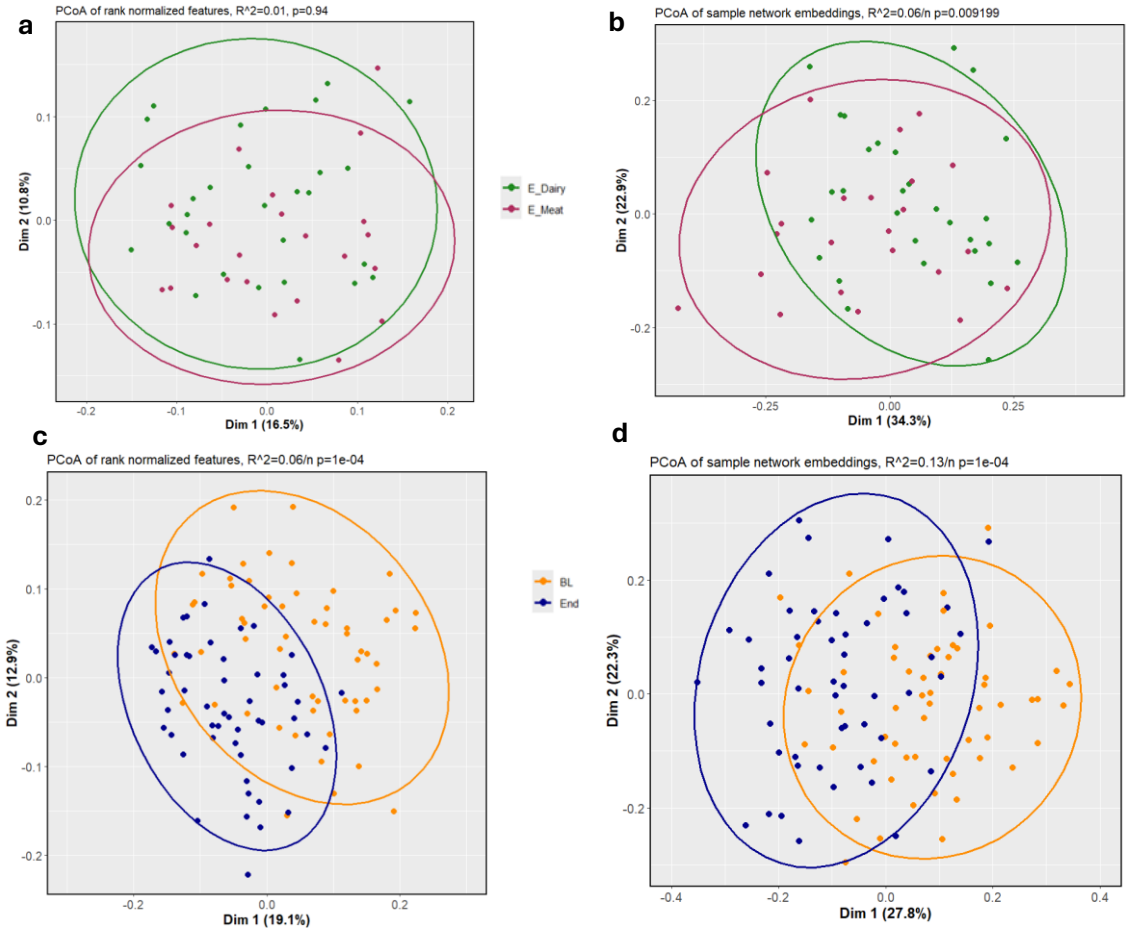

Figure 3. Variation between timepoint and diet type groups. PCoA of **a**, rank normalized microbiome and metabolome feature data and **b**, sample graph embeddings across dairy (green) and meat (purple) groups. **c**, PCoA of rank normalized microbiome and metabolome feature data and **d**, sample graph embeddings across baseline (orange) and endpoint (blue) groups.

built rank normalized data and node embeddings, respectively). This indicates the rank normalized feature data is sufficient to distinguish across sample timepoints; however, the node embedding for the sample time and type diet network enables substantially more reliable predictions of diet type.

Graph embedding identifies microbiome and metabolome features associated with meat diet

As previously noted, the graph embedding framework used here created a 128-dimension representation of the sample-microbiome-metabolome interactions across both changes in time as well as diet (**Figure 1b**). Within the graph embedding space, the positional similarity of any two features, two samples, or a feature-sample pair indicates a functional relationship (e.g. if a given microbiome feature is significantly closer to baseline samples compared to endpoint in the embedding space the microbiome feature may play a more significant role in the gut of infants at 5 months compared to 12 months). We used cosine similarity to capture the positional similarity of all features and samples within the embedding space, then tested the distribution of similarity between a given microbiome or metabolite feature and all meat or dairy samples using a Mann-Whitney U-test, and all baseline or endpoint samples using a paired Wilcoxon rank sum test.

Within the embedding space, we found that 238 microbiome features (including 10 microbiome taxa) and 59 metabolite features were significantly associated with the meat diet and only 1 microbiome feature was significantly associated with the dairy diet (**Supplementary Table 4**). Enrichment analysis of non-taxa microbiome features revealed significant (FDR<0.05) COG feature enrichments of DNA replication, recombination, and repair in the meat diet. Meat diet was also significantly enriched in high level KO database features related to metabolism, genetic information processing, and translation, and lower-level KO features of enzymes with EC numbers, ribosomal proteins, chromosome and associated proteins, and ABC transporters. (**Supplementary Table 5**).

Additionally, we found that 157 microbiome features and 43 metabolites were significantly associated (FDR<0.05) with baseline samples in the embedding space. Conversely, 4,292 microbiome features and 2,007 metabolites were significantly associated with endpoint samples (**Supplementary Table 6**). Within the microbiome features, 173 microbiome taxa were associated with endpoint and only 1 was associated with baseline. Enrichment analysis of baseline associated microbiome features showed significant enrichments in COG features involved in energy production and conversion as well as inorganic ion transport and metabolism (**Supplementary Table 7**).

Endpoint was significantly enriched in family-level taxa *f_Lachnospiraceae*, *f_Oscillospiraceae* and marginally enriched with genus-level *g_Roseburia* (FDR=0.062). Endpoint was also significantly enriched with lower-level KO translation proteins, high level KO cell growth and mitochondrial biogenesis proteins; with marginal enrichment in Pfam features related to ribosomal proteins (FDR=0.061) and rRNA binding proteins (FDR=0.061) (**Supplementary Table 7**).

Graph embedding identifies putative functional microbiome and metabolite networks enriched in time or diet group

Just as the spatial landscape of the node embeddings allows classification of microbiome and metabolome features into time and diet type groups, we can also use the community structure of the latent space to hypothesize clusters of functionally related microbes, microbial features, and metabolite features within the context of this study. To identify possible functional groups, we performed hierarchical clustering on all microbiome and metabolome feature pairwise cosine similarities using the default complete linkage method within the *hclust* function in *R*. Here, we report microbiome-metabolome relationships at or below the heights of 0.15 and 0.2 in the resulting hierarchical dendrogram. For brevity, we limit our discussion of microbe/metabolite clusters to two clusters containing microbe species significantly associated with the meat diet in the embedding space.

At a height of 0.15 (with a maximum height of 0.54) in the *hclust* dendrogram the species *Bifidobacterium dentium*- significantly associated with meat samples in the embedding space- was observed to cluster with the Pfam database PF01751 gene encoding a Toprim domain as well as the KO database K18206 gene (*hypBA2*) encoding β-L-arabinobiosidase. Both the PF01751 and *hypBA2* genes have previously been found in genomic sequences of *Bifidobacterium dentium*[4,5]. When clusters are expanded to a dendrogram height of 0.2, the *Bifidobacterium dentium* cluster includes species *Blautia producta*, 27 metabolites, and 15 additional microbiome features (**Supplementary Table 8**).

A second cluster of interest at a dendrogram height of 0.2 contains the butyrate-producing species *Roseburia inulinivorans* and the KO database K00013 gene (*hisD*) encoding histidinol dehydrogenase. *HisD* is known to be conserved across all *Roseburia* subspecies and plays a role in the histidine biosynthesis pathway[6]. Additionally, the cluster was observed to contain genes coding for ribosome-binding factor A (rbfA), 16S rRNA processing protein RimM, Translation initiation factor 2 (IF-2; GTPase), and smaller subunit ribosomal proteins S19 and S5. RimM, rbfA, and IF-2 play essential roles in the assembly of the 30S ribosomal subunit; the small subunit ribosomal proteins S19 and S5 are components of the 30S ribosomal subunit (**Supplementary Table 8**)[7-11].

Discussion

Here we compared the gut microbiota and metabolites of infants 5 months old and 12 months old across protein enriched diet types of meat and dairy. With traditional differential abundance analyses, we found significant differences between the microbiome and metabolite landscape of 5-month-old infants compared to 12-month-old infants. These findings are consistent with prior studies showing the relative abundance and diversity of microbiome taxa increases greatly in the first year of life[12]. Many metabolite pathways are also known to become enriched with age; particularly pathways in amino acid metabolism[13]. In comparing diet groups of meat and dairy among infants 12 months old, however, differential abundance analyses did not capture any significant differences in the microbiome or metabolome between the two groups- consistent with prior findings in the study released by Minghua et.al[14] in 2023 which found significant height differences between the meat and dairy infant diet groups- but was unable to detect significant differences between gut microbes (ascertained with 16S rRNA gene sequencing of infant fecal samples) in each diet group.

We then applied a graph embedding approach with *pencanpy*[1] to model the sample-microbiome-metabolite interactions within the time and diet type landscape. As the *pecanpy* algorithm maps

samples and features with similar structural properties to nearby locations in the embedding space, we were able to leverage the both microbiome-metabolite feature interactions and the baseline sample landscape to enable more accurate and robust classification of endpoint samples into meat and dairy diet groups.

In addition to more accurate classification, the graph embedding also allowed detection of significant associations between microbe and metabolite features and the meat diet. Notably, among the 10 microbiome taxa significantly associated with the meat diet, the species *Bifidobacterium dentium* (FDR=0.04), *Erysipelotrichaceae* (FDR=0.009), *Pseudomonadales* (FDR=0.03), and the genus *Roseburia* (FDR=0.04) have previously been seen to be positively associated with the consumption of meat or animal protein in adults[15,16].

Lastly, we used the graph embedding to detect spatially correlated microbes and metabolites and created functional microbe-metabolite maps within the time and diet context of this study. We identified putative microbiome-microbiome feature relationships clustered within the embedding space, indicating the embedding space captures functional relationships in addition to enhancing sample classification. In future work, we plan to use the embedding space to develop functional maps of microbe-metabolite interactions.

While graph embedding algorithms are currently used across many disciplines (e.g., understanding protein-protein interactions[17,18], modeling cell differentiation[19], etc), graph embedding has not yet been applied to elucidate gut microbial changes and microbe-metabolite interactions. Here we illustrated the application of graph embedding to the high-dimensional, complex landscape of microbes and metabolites allows for deeper exploration of interactions between microbes and their metabolites-enhancing our understanding diet and gut microbiome development.

Methods

Study Design & Human Subjects

Stool samples from a completed randomized controlled feeding trial[20] were used in this study. In brief, formula-fed infants were recruited from the metro Denver area using mail-in flyers by the Colorado Department of Public Health. After consenting, participants were randomized to consume meat or dairy as the primary source of protein from the complementary foods. Exclusion criteria included low birth weight, cumulative breastfeeding > 1 month, and significant congenital anomalies or known chronic diseases. The primary growth outcome was presently previously[20].

After stratifying by sex and ethnicity, five-month-old, exclusively formula-fed infants were randomized to receive either pureed meats or dairy foods such as infant yogurt, cheese, and whey protein powder. Participants visited the Pediatric Clinical & Translational Research Center (CTRC) at Children's Hospital Colorado (CHCO) at baseline (5 months) and at the end of the intervention (12 months). Compliance and health history were obtained at the monthly home visits. Stool samples were collected at baseline, 9 and 12 months. This study was approved by the Colorado Multiple Institutional Review Board and was registered at ClinicalTrials.gov (NCT02142647).

Dietary Intakes

During the 5 to 12 months dietary intervention, infants consumed a standard, intact milk-protein-based formula *ad lib*. Infant formula (Similac Optigro, Abbott Nutrition, IL) was provided during the intervention to standardize this exposure. The meat-based diet consisted of commercially available pureed meats, and the dairy-based diet consisted of infant yogurt, cheese, and a powdered concentrate of 80% whey protein (specially packaged for this study by Leprino Foods (Denver, CO). Fruit and vegetable intakes were not restricted. Parents were provided tailored feeding guidelines and were encouraged to let the infant's appetite dictate their total intake, as done with previous complementary feeding interventions by our group [21-23]. Three 3-day diet records (de-identified) at baseline, 9 and 12 months were analyzed by the CTRC Nutrition Core (NDSR software, Minneapolis, MN).

Stool Sample Collection

To collect stool samples, sterile diaper liners, gloves and pre-labeled ziplock bags were given to caregivers to place in the infant's diaper. The liner was biodegradable, which effectively collected stool, but allowed passage of urine. Caregivers were given instructions on how to place and remove the liner from the diaper once stool was produced. The soiled liner was then placed by the caregiver in ziplock bags with written collection time and date. The caregiver placed the bagged samples in home freezers (-20° C) and notified the study coordinator immediately. The study coordinator collected the samples within 24 hours and transferred them to University of Colorado Anschutz Medical Campus on ice packs. Stool samples were then collected by research personnel from the liner in a laminar flow hood and stored in sterile vials in -80° C freezers until analysis.

Fecal Untargeted Metabolomics

Fecal homogenates (50ul) were deproteinzied with six times volume of cold acetonitrile:methanol (1:1 ratio), kept on ice with intermittent vortexing for 30 minutes at 4^0C, then centrifuged at 18000xg. 13C6-phenylalanine (3 µl at 250ng/µl) was added as internal standard to fecal homogenates prior to

deproteinization. The supernatants were divided into 4 aliquots and dried down using a stream of nitrogen gas for analysis on a Quadrupole Time-of-Flight Mass Spectrometer (Agilent Technologies 6550 Q-TOF) coupled with an Ultra High Pressure Liquid Chromatograph (1290 Infinity UHPLC Agilent Technologies). Profiling data was acquired under both positive and negative electrospray ionization conditions over a mass range m/z of 100 - 1700 at a resolution of 10,000 (separate runs) in scan mode. Metabolite separation was achieved using two columns of differing polarity, a hydrophilic interaction column (HILIC, ethylene-bridged hybrid 2.1 x 150 mm, 1.7 mm; Waters) and a reversed-phase C18 column (high-strength silica 2.1 x 150 mm, 1.8 mm; Waters) with gradient described previously. [24-26]. Run time was 18 min for HILIC and 27 min for C18 column using a flow rate of 400µl/min. A total of four runs per sample were performed to give maximum coverage of metabolites. Samples were injected in duplicate, wherever necessary, and a pooled quality control (pooled QC) sample, made up of all of the samples from each batch was injected several times during a run. A separate plasma quality control (QC) sample was analyzed with pooled QC to account for analytical and instrumental variability. Dried samples were stored at -20°C till analysis. Samples were reconstituted in running buffer and analyzed within 48 hours of reconstitution. Auto-MS/MS data was also acquired with pooled QC sample to aid in unknown compound identification using fragmentation pattern.

Data alignment, filtering, univariate, multivariate statistical and differential analysis was performed using Mass Profiler Professional (Agilent Inc, USA). Metabolites detected in at least ≥80% of one of two groups were selected for differential expression analyses. Metabolite peak intensities and differential regulation of metabolites between groups were determined as described previously [24,26]. Each sample was normalized to the internal standard and log 2 transformed. Default settings were used except for signal-to-noise ratio threshold (3), mass limit (0.0025 units), and time limit (9 s). Putative identification of each metabolite was done based on accurate mass (m/z) against METLIN database using a detection window of ≤7 ppm. The putatively identified metabolites were annotated as Chemical Abstracts Service (CAS), Kyoto Encyclopedia of Genes and Genomes (KEGG), Human Metabolome Project (HMP) database, and LIPID MAPS identifiers.

Fecal Metagenomics

DNA was extracted from 25-50 mg of each stool sample using the QIAamp PowerFecal DNA kit (QIAGEN, Inc), which employs both mechanical and chemical lysis. DNA concentrations were measured using a Qubit® 2.0 Fluorometer (Invitrogen, Inc.) following the manufacturer's protocol for the dsDNA HS Assay Kit. The NexteraXT DNA sample preparation kit (Illumina, Inc) was used to prepare 1 ng of DNA for multiplexed shotgun metagenomic sequencing, following the manufacturer's protocol. Paired-end 2x150 nucleotide sequencing of pooled libraries was conducted on the Illumina NextSeq 550 platform with the NextSeq High Output Kit v2.5 (Illumina, Inc). Sequences were demultiplexed by library and putative human reads culled by Bowtie2[10] mapping to the hg19 human reference genome. All resulting sequences that did not map to the human genome were deposited in the NCBI Sequence Read Archive under project accession number PRJNA884688. These non-human reads were then processed by Trimmomatic (v0.39[11]) to remove low quality bases (parameters: LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:50) and adapter sequences. This process produced a median of 5.1×10^6 reads per library (IQR: 3.3×10^6 – 9.0×10^6reads/library).

Quality-filtered reads were then annotated using the SqueezeMeta pipeline (v1.4.0[12, 13]), by first co-assembling sequences across all libraries with MEGAHIT (v1.0[14]), removing short contigs (<200bp) with prinseq[15], and then predicting open reading frames (ORFs) in remaining contigs with Prodigal[16]. ORFs were

annotated using DIAMOND[17] to query the NCBI GenBank[18] non-redundant database for taxonomic classification and EggNOG[19] and KEGG[20] databases for functional annotations of Clusters of Orthologous Groups (COG) and KEGG orthology (KO) classifications, respectively. Assignments to the Protein Families Database (Pfam[21]) were made by Hidden Markov Model (HMM) search using HMMER3[22]. Bowtie2[10] was used to map reads to contigs. The final outputs of SqueezeMeta were separate, raw read count files for species-level taxonomic classifications, as well as COG, KO, and Pfam functional classifications.

Filtering and preparation of the metabolome and metagenome data

Beginning with 37,194 microbiome features (22,656 COG; 5,733 KO; 7,109 Pfam; 1,696 Bacterial Taxa) across 114 observations, we removed observations with <250k reads or duplicated samples. We then removed features with more than 90% missing across all observations and relative abundance <0.01%. Next, to remove features with low variability across observations, we removed features with low variation or missingness (coefficient of variation (CV)< 2nd percentile of CV across that feature type or missingness<10%). We retained 17,033 microbiome features for analysis.

Beginning with 9,046 metabolite features (1,556 negative aqueous; 2,409 negative lipid; 1,784 positive aqueous; 3,297 positive lipid) across 109 observations, we removed compounds with more than 90% missing across all observations. To correct for batch and drift effects, we used a linear regression for each compound as the outcome and processing batch, order, and batch by order interaction terms as predictors. Outliers were removed and the updated fitted model was used to estimate residuals for all observations including outliers. The residuals were then shifted by the difference between the original minimum value and the minimum value of the residuals to preserve the minimum value from the starting dataset resulting in the final batch and order adjusted values for each compound. And lastly, to focus on variable compounds we removed compounds with low variation or missingness (HS CV < 0.1 or missingness <10%). We retained 8,867 metabolites for analysis.

To transform the data into a weighted, undirected network as input for *pecanypy*[1], we rank normalized the percent relative abundance (microbiome) or compound abundance (metabolites) values across observations. We merged the microbiome and metabolite data by observational unit (infant sample), retaining 109 observations, then reshaped the data into long format with each row containing a sample-feature pair and the corresponding rank-normalized feature value. Note that zeros observed in percent relative abundance or compound abundance values were kept as zeros in the ranking step to preserve the original data structure in the network (Pipeline figure in **Figure 1a**).

Differential expression analysis

Statistical analysis of differential expression for both the log2 transformed microbiome and metabolomic compound relative abundance was performed by a paired Wilcoxon signed-rank test or non-parametric Mann-Whitney U-test for the paired timepoint data and the diet type data, respectively.

Nested cross-validation evaluation scheme

To reduce overfitting to our limited number of samples, we elected to implement a variation of nested cross-validation. We designed 5 training/testing splits that were stratified for both time point and diet type samples. When evaluating an embedding space we trained a diet classifier and a time point classifier for each of the 5 splits by performing 10-fold CV within the training data to select the best logistic regression hyperparameters: penalty type (l1, l2 or elastic net), strength and ratio (if penalty was

elastic net). Since elastic net models did not always converge, we opted to allow all three penalty types. The validation performance was reported as the average of all 10 folds and was used to optimize embedding spaces. All models training and cross validation was done using *scikit-learn.*

Building the network embedding

128 dimension network embeddings were generated using the *pecanpy*[1] implementation of the *Node2Vec+* algorithm[27] To optimize the embedding space we used *wandb* (wandb.ai) hyperparameter sweeps to tune the *Node2Vec+* return parameter (p), in-out parameter (q), and gamma value (g). For all other parameters, the default values were used. An initial random sweep with 100 parameter combinations was done using a range of 0.01-100 for p and q and a value of either 0,1 or 2 for g. The embedding spaces were scored using the average validation F1 for all 5 time point classifiers. Based on the results of the first sweep we conducted a second sweep with 200 models using a range of 0.1-25 for p, 0.1-10 for q and either 0,1, or 2 for g. The same score procedure was used, but this time to get a final embedding score we averaged the embedding score of each classifier type. Since no clear trend in best performing parameters emerged, we elected to do a final sweep informed by the top 20 best performing embedding spaces. To do this we held either p or q constant and varied the other two parameters. We also introduced slight variation into the constant parameter by adding in values +/- 0.1 and 0.5 from the original value. We then assessed the variance of all our embedding spaces generated across all sweeps. To do this we used a combination of median score and IQR for the five models for each classifier type. This resulted in six embedding spaces with a minimum median F1 validation score of 0.85 and a maximum IQR of 0.25.

Identification of features associated with diet and time using node embedding

The class of each microbiome and metabolome feature was predicted using each of the five prediction models built on the node embedding developed with all eight microbiome and metabolomic compound feature categories. The five models predicted on node embeddings of a given feature; average prediction probability across all five models was computed for each class.

Enrichment Analysis

Conducted using one sided Fisher exact tests. Raw p-values were FDR adjusted.

Acknowledgements

We thank Diana Ir for technical assistance in shotgun metagenomic sequence analysis.

Funding

NIH (NIDDK) 1K01DK111665-01, 1R01DK126710, the Beef Checkoff through the National Cattlemen's Beef Association, and the National Pork Board.

References

1 Liu, R. & Krishnan, A. PecanPy: a fast, efficient and parallelized Python implementation of node2vec. *Bioinformatics* **37**, 3377-3379, doi:10.1093/bioinformatics/btab202 (2021).

2 Robertson, R. C., Manges, A. R., Finlay, B. B. & Prendergast, A. J. The Human Microbiome and Child Growth - First 1000 Days and Beyond. *Trends Microbiol* **27**, 131-147, doi:10.1016/j.tim.2018.09.008 (2019).

3 Edwards, C. A. Determinants and Duration of Impact of Early Gut Bacterial Colonization. *Ann Nutr Metab* **70**, 246-250, doi:10.1159/000466711 (2017).

4 Apweiler, R. *et al.* UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* **32**, D115-119, doi:10.1093/nar/gkh131 (2004).

5 Fujita, K. *et al.* Molecular cloning and characterization of a beta-L-Arabinobiosidase in Bifidobacterium longum that belongs to a novel glycoside hydrolase family. *J Biol Chem* **286**, 5143-5150, doi:10.1074/jbc.M110.190512 (2011).

6 Hillman, E. T. *et al.* Comparative genomics of the genus Roseburia reveals divergent biosynthetic pathways that may influence colonic competition among species. *Microb Genom* **6**, doi:10.1099/mgen.0.000399 (2020).

7 Thurlow, B. *et al.* Binding properties of YjeQ (RsgA), RbfA, RimM and Era to assembly intermediates of the 30S subunit. *Nucleic Acids Res* **44**, 9918-9932, doi:10.1093/nar/gkw613 (2016).

8 Suzuki, S. *et al.* Structural characterization of the ribosome maturation protein, RimM. *J Bacteriol* **189**, 6397-6406, doi:10.1128/JB.00024-07 (2007).

9 Bylund, G. O., Wipemo, L. C., Lundberg, L. A. & Wikstrom, P. M. RimM and RbfA are essential for efficient processing of 16S rRNA in Escherichia coli. *J Bacteriol* **180**, 73-82, doi:10.1128/JB.180.1.73-82.1998 (1998).

10 Nord, S., Bhatt, M. J., Tukenmez, H., Farabaugh, P. J. & Wikstrom, P. M. Mutations of ribosomal protein S5 suppress a defect in late-30S ribosomal subunit biogenesis caused by lack of the RbfA biogenesis factor. *RNA* **21**, 1454-1468, doi:10.1261/rna.051383.115 (2015).

11 Bunner, A. E., Beck, A. H. & Williamson, J. R. Kinetic cooperativity in Escherichia coli 30S ribosomal subunit reconstitution reveals additional complexity in the assembly

landscape. *Proc Natl Acad Sci U S A* **107**, 5417-5422, doi:10.1073/pnas.0912007107 (2010).

12 Yao, Y. *et al*. The Role of Microbiota in Infant Health: From Early Life to Adulthood. *Front Immunol* **12**, 708472, doi:10.3389/fimmu.2021.708472 (2021).

13 Ouyang, R. *et al*. Maturation of the gut metabolome during the first year of life in humans. *Gut Microbes* **15**, 2231596, doi:10.1080/19490976.2023.2231596 (2023).

14 Tang, M. *et al*. Different gut microbiota in U.S. formula-fed infants consuming a meat vs. dairy-based complementary foods: A randomized controlled trial. *Front Nutr* **9**, 1063518, doi:10.3389/fnut.2022.1063518 (2022).

15 Bolte, L. A. *et al*. Long-term dietary patterns are associated with pro-inflammatory and anti-inflammatory features of the gut microbiome. *Gut* **70**, 1287-1298, doi:10.1136/gutjnl-2020-322670 (2021).

16 Wang, Y. *et al*. Meat Consumption and Gut Microbiota: a Scoping Review of Literature and Systematic Review of Randomized Controlled Trials in Adults. *Adv Nutr* **14**, 215-237, doi:10.1016/j.advnut.2022.10.005 (2023).

17 Yang, F., Fan, K., Song, D. & Lin, H. Graph-based prediction of Protein-protein interactions with attributed signed graph embedding. *BMC Bioinformatics* **21**, 323, doi:10.1186/s12859-020-03646-8 (2020).

18 Wang, N., Zeng, M., Li, Y., Wu, F. X. & Li, M. Essential Protein Prediction Based on node2vec and XGBoost. *J Comput Biol* **28**, 687-700, doi:10.1089/cmb.2020.0543 (2021).

19 Bica, I., Andres-Terre, H., Cvejic, A. & Lio, P. Unsupervised generative and graph representation learning for modelling cell differentiation. *Sci Rep* **10**, 9790, doi:10.1038/s41598-020-66166-8 (2020).

20 Tang, M., Hendricks, A. E. & Krebs, N. F. A meat- or dairy-based complementary diet leads to distinct growth patterns in formula-fed infants: a randomized controlled trial. *Am J Clin Nutr* **107**, 734-742, doi:10.1093/ajcn/nqy038 (2018).

21 Tang, M., Sheng, X. Y., Krebs, N. F. & Hambidge, K. M. Meat as complementary food for older breastfed infants and toddlers: a randomized, controlled trial in rural China. *Food Nutr Bull* **35**, S188-192 (2014).

22 Krebs, N. F. *et al*. Comparison of complementary feeding strategies to meet zinc requirements of older breastfed infants. *Am J Clin Nutr* **96**, 30-35, doi:10.3945/ajcn.112.036046 (2012).

23 Krebs, N. F. *et al*. Meat as a first complementary food for breastfed infants: feasibility and impact on zinc intake and status. *J Pediatr Gastroenterol Nutr* **42**, 207-214, doi:10.1097/01.mpg.0000189346.25172.fd (2006).

24 Dutta, T. *et al*. Concordance of Changes in Metabolic Pathways Based on Plasma Metabolomics and Skeletal Muscle Transcriptomics in Type 1 Diabetes. *Diabetes* **61**, 1004-1016, doi:10.2337/db11-0874 (2012).

25 Trushina, E., Dutta, T., Persson, X. M., Mielke, M. M. & Petersen, R. C. Identification of altered metabolic pathways in plasma and CSF in mild cognitive impairment and Alzheimer's disease using metabolomics. *PLoS One* **8**, e63644, doi:10.1371/journal.pone.0063644 (2013).

26 Dutta, T. *et al*. Impact of long-term poor and good glycemic control on metabolomics alterations in Type 1 diabetic people. *The Journal of clinical endocrinology and metabolism*, jc20152640, doi:10.1210/jc.2015-2640 (2016).

27 Liu, R., Hirn, M. & Krishnan, A. Accurately modeling biased random walks on weighted networks using node2vec. *Bioinformatics* **39**, doi:10.1093/bioinformatics/btad047 (2023).