microorganisms

MDPI

Article

BZINB Model-Based Pathway Analysis and Module Identification Facilitates Integration of Microbiome and Metabolome Data

Bridget M. Lin [1], Hunyong Cho [1], Chuwen Liu [1], Jeff Roach [2], Apoena Aguiar Ribeiro [3], Kimon Divaris [4,5] and Di Wu [1,6,7,*]

1 Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
2 Research Computing, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
3 Division of Diagnostic Sciences, Adams School of Dentistry, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
4 Division of Pediatric and Public Health, Adams School of Dentistry, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
5 Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
6 Division of Oral and Craniofacial Health Sciences, Adams School of Dentistry, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
7 Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
* Correspondence: did@email.unc.edu; Tel.: +1-919-537-3277

Abstract: Integration of multi-omics data is a challenging but necessary step to advance our understanding of the biology underlying human health and disease processes. To date, investigations seeking to integrate multi-omics (e.g., microbiome and metabolome) employ simple correlation-based network analyses; however, these methods are not always well-suited for microbiome analyses because they do not accommodate the excess zeros typically present in these data. In this paper, we introduce a bivariate zero-inflated negative binomial (BZINB) model-based network and module analysis method that addresses this limitation and improves microbiome–metabolome correlation-based model fitting by accommodating excess zeros. We use real and simulated data based on a multi-omics study of childhood oral health (ZOE 2.0; investigating early childhood dental caries, ECC) and find that the accuracy of the BZINB model-based correlation method is superior compared to Spearman's rank and Pearson correlations in terms of approximating the underlying relationships between microbial taxa and metabolites. The new method, BZINB-iMMPath, facilitates the construction of metabolite–species and species–species correlation networks using BZINB and identifies modules of (i.e., correlated) species by combining BZINB and similarity-based clustering. Perturbations in correlation networks and modules can be efficiently tested between groups (i.e., healthy and diseased study participants). Upon application of the new method in the ZOE 2.0 study microbiome–metabolome data, we identify that several biologically-relevant correlations of ECC-associated microbial taxa with carbohydrate metabolites differ between healthy and dental caries-affected participants. In sum, we find that the BZINB model is a useful alternative to Spearman or Pearson correlations for estimating the underlying correlation of zero-inflated bivariate count data and thus is suitable for integrative analyses of multi-omics data such as those encountered in microbiome and metabolome studies.

Keywords: correlation; microbiome; metabolomics; multi-omics; zero-inflation; counts; caries; clustering; pathways; network

Citation: Lin, B.M.; Cho, H.; Liu, C.; Roach, J.; Ribeiro, A.A.; Divaris, K.; Wu, D. BZINB Model-Based Pathway Analysis and Module Identification Facilitates Integration of Microbiome and Metabolome Data. *Microorganisms* **2023**, *11*, 766. https://doi.org/10.3390/microorganisms11030766

Academic Editors: Patrick Veras Quelemes and Gláuber Campos Vale

Received: 31 January 2023
Revised: 4 March 2023
Accepted: 12 March 2023
Published: 16 March 2023

1. Introduction

Microbiome data are essential for advancing our understanding of the biological basis of many human diseases and are becoming increasingly available. While descriptions of taxonomic aspects of the human microbiome are valuable, functional insights are arguably more informative. Accordingly, characterizations of the ways that bacteria interact with the host and the environment via metabolic byproducts and other biochemicals can offer important biological insights into disease pathogenesis and offer targets for prevention and treatment. However, the complexity of these interactions cannot be underestimated. For example, relevant metabolites can be microbial products, whereas host- or environment-derived metabolites may serve as nutrients or environmental stressors for microbial communities. While the availability of microbiome–metabolome and health-disease associated phenotype data is increasing, suitable analysis method development has not kept pace. Leveraging data on microbiome–metabolome interactions could help illuminate important biological pathways at play and identify bacterial species that influence each other via inter-species activities [1,2]. Importantly, these biological networks and microbial correlations may be influenced by the environment and differ between states of health and disease, as in the case of the oral biofilm microbiome–metabolome and dental caries [3,4]. Therefore, defining and measuring networks among microbial taxa, pathways in which taxa and metabolites are involved, and clusters of inter-correlated taxa are critical for understanding the function of microbial communities in health and diseases. Curated pathway datasets such as KEGG can provide known metabolic pathways involving metabolite networks, but they are not context-specific. The newly available Whole Genome Sequencing shotgun (WGS) DNAseq for metagenomics and RNAseq for metatranscriptomics (providing information at the at the taxon or gene level), or the earlier 16S sequencing for bacterial taxonomic abundance, paired with metabolome data from the same biofilm samples can provide unique new opportunities for context-specific integrative microbial pathway analyses.

Although joint network analyses of microbiome and metabolome data are critical for understanding host–microbiome interactions, the existing computational methods have not been designed for the specific characteristics of microbiome data. Until recently, Pearson or Spearman correlation-based pathway analyses [5] have been popular and robust for gene–gene network analysis for gene expression data; however, these approaches do not consider the excess zeros in microbiome data. Kendall's Tau and Mutual Information (MI) have been suggested as possible replacements for Pearson or Spearman correlations for non-normal distributions, such as in single-cell RNAseq data [6–8]; however, MI is sensitive to threshold grids in data with excess zeros, whereas Kendall's Tau loses information on the continuous scale. More recently, copula-based pathway analysis [9] has been developed to model interactions between genes in single-cell RNAseq data while accommodating their non-normal distribution. Moreover, most existing approaches do not allow for testing pathway changes among sample groups. Therefore, it is challenging to infer, for example, disease-specific microbiome–metabolome pathways and the essential hubs of microbial taxa and metabolites.

We propose a de novo pathway discovery analysis that is independent of prior pathway knowledge and learns from the observed microbiome and metabolome data generated from matched samples, or at least from the same body sites or subjects, as long as a biological interaction hypothesis is valid. Our proposed method, BZINB-based integration of microbiome and metabolome for pathway analysis (BZINB-iMMPath), uses the newly developed bivariate zero-inflated negative binomial (BZINB) model to directly model the joint distribution of a pair of count vectors, where one vector represents microbial species and the other vector represents metabolites, to estimate model-based correlations. The advantage of our method, which uses BZINB, is that we can rigorously handle the excess zeros in the distribution of microbiome counts [10].

Similar to single-cell RNAseq data, microbiome data typically exhibit large numbers of zeros ("excess zeros") for several possible reasons, including structural zeros (e.g., due to the fact that some species may not be present in some samples, also referred to as biological zeros), or sampling zeros (e.g., due to technical artifacts, frequently referred to as "dropout events"). Specifically, two advantages of using BZINB include the realistic assumption of dropouts [11] in the zero-inflated negative binomial (ZINB) distribution that allows for the flexible modeling of both biological zeros (in the negative binomial component) and structural zeros (in logistic regression) to improve model fitting, and the feasibility of estimating correlations in the bivariate negative binomial (BNB) component conditional of the zero inflation component to reflect the underlying correlations.

We additionally propose, as another component of BZINB-iMMPath, the use of BZINB correlation measurements to represent the similarities [12] between species in species-wise clustering analysis to identify species modules (i.e., clusters) wherein species are highly correlated. Because the BZINB model accounts for zero inflation in a pair of species, or in individual species and metabolites when investigating microbiome–metabolome correlations, most species and metabolites can be retained in the analysis rather than excluded because of zero inflation, a feature that may be of biological importance.

To compare the accuracy of BZINB-based correlation with other popular correlation measures, we simulated pairs of correlated microbiome species and metabolite count vectors using the bivariate lognormal distribution and the BZINB distribution. We carried out simulations and applications using matched microbiome–metabolome data from a community-based study of childhood oral health/disease (ZOE 2.0 study, investigating early childhood caries or ECC) that sampled 3–5-year-old children's supragingival dental biofilm. We also evaluated the accuracy of module identification using BZINB as a measure of similarity for cut-based clustering by crafting co-varying clusters of count vectors to represent species in semi-parametric simulations. We show that, in real data applications, the new method can identify the crafted clusters with high accuracy. Moreover, the integrated pathway analysis identified biologically significant and disease-specific microbial–metabolite pathways and meaningful inter-species interactions.

The BZINB framework introduces the correlation between the two variables by adding one variable, so it only allows non-negative correlations between species and metabolites or between species, which can limit the utility of our method. However, in most omics contexts, positively correlated features are arguably of greatest interest. For example, in gene expression data, the vast majority of genes do have positive or near-zero correlation [13,14]. Positive correlations among bacterial species are also more common compared to negative correlations (Figure 1a, top). Although we observe more negative correlations between microbial species and metabolites (Figure 1a, bottom), positive correlations are overall larger are more biological and clinically interesting, as that may, for example, reflect the metabolites that provide nutrition to bacteria or metabolites that are produced by bacteria.

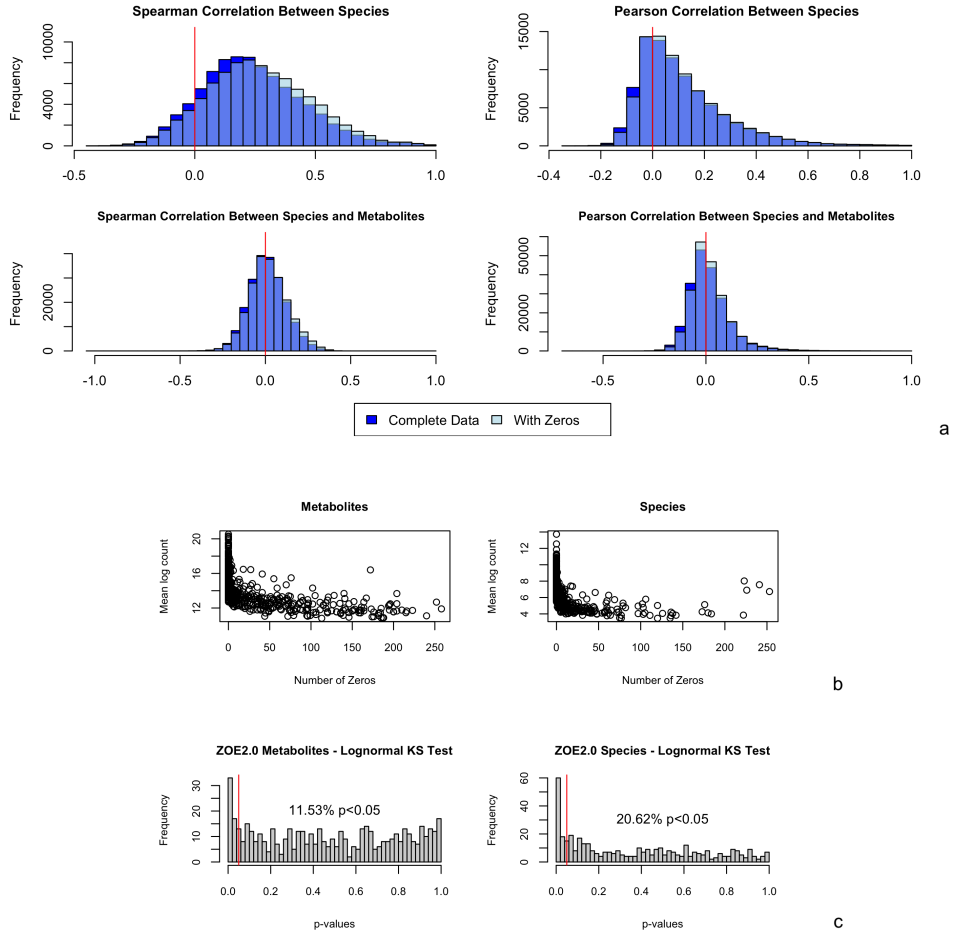


Figure 1. (**a**) Observed Spearman and Pearson correlations between pairs of (Kraken2/Bracken) species (417); and between pairs of (Kraken2/Bracken) species (417) and metabolites (503); in ZOE 2.0 (n = 289). Correlations among complete data exclude subjects with one or more zeros in the pair; correlations among data with zeros include all subjects. The red vertical line in each plot represents a correlation of 0. (**b**) Number of zeros plotted against mean log nonzero count for each metabolite and number of zeros plotted against mean log nonzero count for each Kraken2/Bracken species. (**c**) *p*-values obtained from lognormal (parameters from models fitted on nonzero counts for each metabolite and species) Kolmogorov–Smirnov test for ZOE 2.0 metabolites and Kraken2/Bracken microbiome species. The red vertical line represents a *p*-value of 0.05 so that *p*-values below it indicate statistical significance in the tests.

2. Materials and Methods

2.1. Description of BZINB Model

2.1.1. ZINB Model

Similar to single-cell data analysis, the probability of dropout per species per sample can be modeled using logistic regression in the framework of a zero-inflated model. The ZINB model has been previously proposed for the analysis of single-cell RNAseq data as a superior and more flexible model fitting compared to Poisson-based methods [15] for individual gene analyses in scRNAseq data, by allowing for both excess zeros and overdispersion.

2.1.2. BNB Model

Cho et al. 2021 [10] began by introducing a bivariate negative binomial (BNB) model based on the Poisson–Gamma mixture model. First, let $R_j \sim Gamma(\alpha_j, \beta)$ for $j = 0, 1, 2$. Consider a pair of random variables (X_1, X_2), where X_1 and X_2 are each Poisson-distributed with means of $R_0 + R_1$ and $\delta(R_0 + R_2)$, respectively, where $\delta \in R^+$. These two mean

variables are related through a common Gamma-distributed component, R_0. Therefore, marginally, X_1 and X_2 each follow the negative binomial distribution such that $X_i \sim NB\left(\alpha_0 + \alpha_i, \frac{1}{\beta_i+1}\right)$ for $i = 1, 2$, where $\beta_1 = \beta, \beta_2 = \delta\beta$. Thus, $\text{mean}(X_i) = \frac{\alpha_0+\alpha_i}{\beta_i}$, $\text{var}(X_i) = \frac{(\alpha_0+\alpha_i)(\beta_i+1)}{\beta_i^2}$, and $\rho_{BNB} = \text{Cor}(X_1, X_2) = \frac{\alpha_0}{\sqrt{(\alpha_0+\alpha_1)(\alpha_0+\alpha_2)}}\sqrt{\frac{\beta_1\beta_2}{(\beta_1+1)(\beta_2+1)}}$. We henceforth denote $(X_1, X_2) \sim BNB(\alpha_0, \alpha_1, \alpha_2, \beta_1, \beta_2)$. Therefore, the parameters in ρ_{BNB} are estimated by fitting all the data to the BNB model.

2.1.3. BZINB Model and BZINB-Based Correlation

For correlation between a pair of genes in scRNaseq data, a bivariate zero-inflated (BZINB) model was proposed by Cho et al. 2021 [10] that has the ZINB marginals, more parameters to flexibly accommodate the complexity of the single-cell biology, and the estimated correlation conditional for the non-dropout events. With similar assumptions of dropouts observed as excess zeros and the overdispersion problem accentuated in microbiome data, here we extend the BZINB framework for microbial data modeling to compute a unique correlation measured between species or between species and metabolites. This new unique correlation analysis approach (i.e., BZINB-iMMPath) is model-based and uses the parameters estimated for the BNB component that is conditional on the probability of there being non-dropouts in the BZINB model, defined as described below.

A pair of Bivariate Zero-Inflated Negative Binomial (BZINB) variables $(Y_1, Y_2) \sim BZINB(\alpha_0, \alpha_1, \alpha_2, \beta_1, \beta_2, \pi_1, \pi_2, \pi_3, \pi_4)$ follows a zero-inflated extension of the Bivariate Negative Binomial (BNB) distribution, where π_1, π_2, π_3, and π_4, respectively, represent the probabilities of observing nonzero Y_1 and Y_2, nonzero Y_1 only, nonzero Y_2 only, and zero Y_1 and Y_2. Note that πs do not represent the probability of observing zero but represent the probability of zero inflation. Without zero inflation, we could still observe zeros. In other words, with higher $\pi_3 + \pi_4$ and $\pi_2 + \pi_4$ values, we expect to observe extra zero values in Y_1 and Y_2, respectively. Therefore, there is an underlying BNB component of the BZINB model, which is partially unobserved. Marginally, $Y_j \sim ZINB\left(\alpha_0 + \alpha_j, \frac{1}{\beta_j+1}, \pi_{4-j} + \pi_4\right)$ for $j = 1, 2$. In other words, without zero inflation, Y_j follows $NB\left(\alpha_0 + \alpha_j, \frac{1}{\beta_j+1}\right)$. Y_j is masked with zeros with probability $\pi_{4-j} + \pi_4$.

Based on our understanding of excess zeros in the microbiome, the BNB components—which can include zeros from the negative binomial distribution—in the BZINB model reflect the underlying correlation between species after accounting for the dropouts (whether structural or technical) in BZINB. It follows that we use the same formula as ρ_{BNB} as in the model-based correlation. Therefore, we have $\rho_{BZINB} = \text{Cor}(Y_1, Y_2) = \frac{\alpha_0}{\sqrt{(\alpha_0+\alpha_1)(\alpha_0+\alpha_2)}}\sqrt{\frac{\beta_1\beta_2}{(\beta_1+1)(\beta_2+1)}}$, which is seemingly the same as the BNB correlation ρ_{BNB}. The difference is that we estimate all the parameters $(\alpha_0, \alpha_1, \alpha_2, \beta_1, \beta_2, \pi_1, \pi_2, \pi_3, \pi_4)$ and use the BNB component parameters $(\alpha_0, \alpha_1, \alpha_2, \beta_1, \beta_2)$ only for ρ_{BZINB}, while the BNB correlation ρ_{BNB} is obtained by estimating $(\alpha_0, \alpha_1, \alpha_2, \beta_1, \beta_2)$ only—the latter does adjust for zero inflations in BNB. This difference reflects the different assumptions of the presence of zero inflation.

There is a naive correlation of the BZINB model; namely, a correlation measure without adjustment for zero inflation in the BZINB model. This correlation involves all BZINB parameters:

$$\tilde{\rho}_{\text{BZINB}}(Y_1, Y_2) = \frac{\sigma_{12}}{\sigma_1\sigma_2},$$

where $\sigma_{12} = \{\alpha_0 + (\alpha_0 + \alpha_1)(\alpha_0 + \alpha_2)\}\beta_1\beta_2\pi_1 - (\alpha_0 + \alpha_1)(\alpha_0 + \alpha_2)\beta_1\beta_2(\pi_1 + \pi_2)(\pi_1 + \pi_3)$ and $\sigma_j^2 = (\alpha_0 + \alpha_j)^2\beta_j^2(\pi_1 + \pi_j + 1)(1 - \pi_1 - \pi_j + 1) + (\alpha_0 + \alpha_j)\beta_j(\beta_j + 1)(\pi_1 + \pi_j + 1)$, $j = 1, 2$. Simulation results (not shown) suggest this "naive BZINB correlation" introduces noise in the estimation and decreases the estimation accuracy of the underlying correlations.

2.2. Existing Correlation Calculation Methods for Network/Pathway Analysis

In correlation-based analyses such as network estimation for multi-omics count data, Pearson's correlations are often used with the assumption of linearity. Previously, weighted correlation network analysis (WGCNA) has been used [5] to identify co-expressed clusters (modules) of highly correlated genes or other features. However, both microbiome and metabolome data contain excessive zeros and therefore there may be excessive ties in the data. In this case, Spearman's rank correlation, even with less stringent assumptions compared to Pearson's correlation, may still not be an appropriate measure.

In this study, we compare ρ_{BZINB} used in BZINB-iMMPath to both ρ_{BNB} and the Spearman and Pearson correlations. The formula for the Spearman correlation between vectors $X_1 = (X_{1,1}, X_{1,2}, \ldots, X_{1,n})$ and $X_2 = (X_{2,1}, X_{2,2}, \ldots, X_{2,n})$ is $\rho_{Spearman} = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$, where $d_i = \text{rank}(X_{1i}) - \text{rank}(X_{2,i})$. In the case of ties, the average of the ranks is used. The formula for the Pearson correlation is $\rho_{Pearson} = \frac{\sum(X_{1,i}-\overline{X_1})(X_{2,i}-\overline{X_2})}{\sqrt{\sum(X_{1,i}-\overline{X_1})^2(X_{2,i}-\overline{X_2})^2}}$.

2.3. Description of Microbiome and Metabolome Data from the ZOE 2.0 Study

The ZOE 2.0 study includes 6404 3–5-year-old children enrolled in public preschools in North Carolina, United States, who underwent clinical dental examinations and biospecimen collection [16]. Of those, a subset of 300 participants' supragingival biofilm samples were analyzed and made available for multi-omics (including metagenomics, metatranscriptomics, and metabolomics) analyses. Accordingly, 300 children have metagenomics data (WGS DNAseq, called DNA in this paper), 297 have metatranscriptomics (RNASeq) data, and 289 have metabolite data. Microbiome data have been made available via https://www.ncbi.nlm.nih.gov/bioproject/671299; (accessed on 4 March 2023) and metabolome data via https://www.ebi.ac.uk/metabolights/MTBLS2215; (accessed on 4 March 2023). As in a previous investigation [17], ten participants with greater than 30% missing metabolite data and one ineligible participant were excluded. Among the 289 with metabolite data, 109 met the clinical criteria for ECC (i.e., cases) and 180 did not (i.e., non-cases) [18,19].

To allow for comparisons of goodness-of-fit and variations in data sparsity (i.e., percentage of zeros), we used microbiome data generated by two different popular procedures for mapping and preprocessing metagenomics. Primarily, microbiome DNA data were classified into species-level profiles using a pipeline based on Kraken2 [20] and Bracken 2.5 [21], referred to as Kraken2/Bracken in this paper. The pipeline was built using a custom database including human, fungal, bacterial, and the expanded Human Oral Microbiome Database (eHOMD) [22] for microbial reference genomes. There were 417 microbial species identified as "core species" after excluding rare and low-prevalence taxa that were kept in the analysis [23]. In a secondary procedure, the same DNA sequence reads were processed using MetaPhlAn2.2 through the HUMAnN 2.0 pipeline [24,25] with the default microbial reference genome in HUMAnN 2.0. Viruses, biosatellites, and unidentified species were filtered out, resulting in 205 species-level taxa remaining available for analysis. The advantage of Kraken2/Bracken for our application is due to the fact that it allowed for the use of a custom and contemporary oral microbiome reference database and thus mapped oral/dental species more accurately than HUMAnN 2.0. On the other hand, HUMAnN 2.0 allowed not only for the identification of species, but also for the generation of gene family and pathway-level data that can be of interest and value in some applications. The real data application of BZINB-iMMPath was done only using Kraken2/Bracken species-level data. Of note, all presented results rely on Kraken2/Bracken data unless HUMAnN 2.0 is explicitly mentioned, such as in goodness-of-fit and percentage of zeros comparisons that are presented in Appendix A.

The focus of the work reported in this paper is metagenomics data at the species level, but our new method can be applied to metatranscriptomics (i.e., RNAseq) as well as

other levels of data, including gene family or genes, because all data types are similarly characterized by excess zeros and overdispersion [26].

To obtain metabolomics data, samples were processed using Metabolon's Ultra Performance Liquid Chromatography-tandem Mass Spectrometry pipeline [27,28]. A total of 503 named metabolites were identified through peak identification, QC, and correction for day-dependent technical variations [16]. Procedures and descriptions of the obtained metabolite data have been previously reported in detail [17,29].

2.4. Simulation Study

2.4.1. Lognormal Based Simulation

We simulated data based on the bivariate lognormal distribution, then replaced some numbers with zero to mimic the excess zeros in the following way, so that each simulated vector can be considered to be drawn from the zero-inflated lognormal distribution. We simulated vectors representing pairs of metabolites and species, with theoretical correlations of 0.05, 0.1, 0.3, and 0.5, representing weak to strong correlations, based on the empirical distribution of correlations between the observed counts of pairs of species and metabolites (Figure 1a). Each vector consisted of 300 elements drawn from a lognormal distribution, representing natural log-transformed counts. After transformation, the numbers were rounded to the nearest integer to represent counts. For simplicity, the marginal variance of the log counts in each vector was set to 1, which was well within the range of the sample variances of the metabolite- and species-wise log counts in ZOE 2.0. Assuming that most missing values in metabolite data are due to low concentration, the counts in each metabolite vector were ranked and assigned a probability based on their rank. These probabilities spanned an interval of 0.3, centered at the pre-determined proportion missing. Let $rank_i$ represent the rank of the ith element in the metabolite vector, and let p_{zero} be the proportion of zeros in the vector. Then, the ith element of the vector is set to zero with a probability of $p_i = (0.5 - (rank_i)/300) * 0.3 + p_{zero}$. Under the assumption that zeros in microbiome species are typically structural zeros, the elements in each vector representing a species were randomly chosen to be set to zero after the counts were simulated. Figure 1b and Appendix A Figure A4 illustrate the number of zero counts against the mean of nonzero counts of each metabolite and species. These data revealed a decreasing trend in mean counts as the number of zeros increased and informed the selection of simulation parameters. Therefore, vector pairs representing metabolites and species were simulated under the scenarios outlined in the first four rows in Table 1.

In addition, the four correlation types were compared in simulated vector pairs that represent the relationships between two microbial species. These vectors were simulated based on the scenarios in the last three rows in Table 2. Zero counts were assigned randomly.

Table 1. Marginal log-scale means (before introducing zeros) and number of zeros for simulation of bivariate lognormal vectors with excess zeros that represent metabolite–species pairs (where X_1 represents a metabolite and X_2 represents a species) and species–species pairs. Levels of zero inflation include balanced (similar number of zeros in each vector), with either low or high zero inflation, and unbalanced (one vector has substantially more zeros than the other). The letter labels (a, b, c, d) of each combination denote different simulation scenarios in that the different zero inflation parameters were used to obtain results for each line of Appendix A Table A3.

Relationship		Zero Inflation	Number of Zeros	Means
Metabolite–Species	a	Balanced, low	30, 60	14, 11
	b	Balanced, high	150, 200	12, 9
	c	$N_{zero,X_1} < N_{zero,X_2}$	30, 200	14, 9
	d	$N_{zero,X_1} > N_{zero,X_2}$	150, 60	12, 11
Species–Species	a	Balanced, low	60, 60	11, 11
	b	Balanced, high	200, 200	9, 9
	c	$N_{zero,X_1} < N_{zero,X_2}$	60, 200	11, 9

Table 2. Zero inflation parameters and resulting expected zeros for simulation of bivariate zero-inflated negative binomial vectors to represent pairs of metabolites and species (where X_1 represents a metabolite and X_2 represents a species) and pairs of species. Similar to the scenarios outlined in Table 1, there are balanced and unbalanced levels of zero inflation. The other BZINB model parameters are outlined in Appendix A Tables A1 and A2. The letter labels of each combination denote the zero inflation parameters used to obtain results for each line of Appendix A Table A3.

	Zero Inflation	Expected Zeros	π_1	π_2	π_3	π_4
a	Balanced, low	30, 60	0.75	0.15	0.05	0.05
b	Balanced, high	210, 240	0.1	0.2	0.1	0.6
c	$N_{zero,X1} < N_{zero,X2}$	60, 225	0.2	0.6	0.05	0.15
d	$N_{zero,X1} > N_{zero,X2}$	225, 60	0.2	0.05	0.6	0.15

2.4.2. BZINB-Based Simulation

To represent typical pairs, also called vector pairs, as in the real data with various amounts of pairwise and non-pairwise zeros, vector pairs, we carried out simulations using several combinations of parameters, as summarized in Appendix A Table A1. For computational efficiency, these vector pairs represent rescaled pairs of count vectors obtained from the real data $\left(X_i / \frac{\mathrm{sd}(X_i)}{30}, i = 1, 2\right)$, without altering the correlations. We considered underlying correlations of $\rho_{BNB} = 0.05,\ 0.1,\ 0.30$, and 0.5 by using different combinations of shape and scale parameters in the BZINB distribution (Appendix A Table A1). For each combination of shape and scale parameters (and accordingly, level of correlation of the nonzero counts), we conducted simulations using four combinations of zero inflation parameters $(\pi_1, \pi_2, \pi_3, \pi_4)$, representing balanced low, two combinations of unbalanced, and balanced high zero inflations (Table 2).

We also simulated vector pairs under the BZINB distribution to represent typical pairs of microbial species. These vectors had the same zero inflation parameters as the microbiome–metabolome simulated vector pairs (Table 2) but different means and slightly different correlations. The corresponding shape and scale parameters are presented in Appendix A Table A2.

2.5. Spectral Clustering for Module Identification

2.5.1. Approach for BZINB Application in Spectral Clustering

Spectral clustering is a flexible method for partitioning networks using the eigenvectors of nodes' similarity matrices [12] and has been used in many applications, including bioinformatics. Although similarity is typically quantified through the Gaussian kernel, other measures such as cosine similarity [30] have been used to better represent certain data types. In correlation networks, the positive correlation between a pair of nodes (or, in our data, species or metabolites) is scale-invariant and is often used as a measure of similarity when the co-varying dynamics of the nodes is of interest. Therefore, one can reasonably use the estimated correlations in constructing affinity matrices in applications such as spectral clustering to discover novel pathways that differ between study groups or that are potentially associated with health or disease states. In this paper, we compare the Spearman, BNB, and BZINB correlations in spectral clustering for microbiome count data.

For vectors x_i and x_j, the affinity a_{ij} is a measure of similarity such that a_{ij} is bounded by 0 and 1, a_{ij} is closer to 1 as x_i and x_j are more similar, and $a_i j = 0$ when $i = j$. To obtain each affinity matrix from a correlation matrix, we set the diagonal entries to zero. Since the BZINB model-based correlation can only be positive, we force any negative values obtained from Spearman correlations to be zero. This allows us to only predict clusters with and based on positive inter-dependencies. Next, we cluster the nodes using SpectraLib_A [31]. While the affinity matrices are all symmetric, this method can account for directed networks, for example, to incorporate known interactions between species or metabolites, by using asymmetric affinity matrices.

2.5.2. Evaluation of Cut-Based Spectral Clustering Using Crafted Semi-Parametric Simulation

We simulated correlated clusters to compare the accuracy of the three types of affinity matrices as follows. We permuted the first 400 species in the caries-free (i.e., healthy group) ZOE 2.0 participants and split them into 10 clusters of 40 species each. For each cluster k, we generated a random vector $R_k \sim Pois(17,968)$ (since 17,968 was the mean count of the 400 species). For the nonzero counts of each species j in cluster k, we computed a weighted sum, $Z_j = 0.9 * Y_j + 0.1 * R_k$, of each original species' counts (Y_j) and the random vector to introduce additional correlation within each cluster. We then estimated the Spearman, BNB, and BZINB correlations between all 400 species to construct three types of affinity matrices. Then, we clustered the species for each affinity matrix using SpectraLib_A with 10 clusters. In cases where biological knowledge exists regarding the direction of effects in relationships between different omics layers, the affinity matrix can be altered to reflect it.

To evaluate the accuracy of each correlation type in spectral clustering, we contrasted predicted and assigned clusters to optimize the prediction accuracy as follows:

1. If the most common predicted cluster for an assigned cluster is the same as the most common assigned cluster for that predicted cluster, those clusters are matched.
2. Then, the overall proportion of accurate predicted cluster assignments is calculated for each possible combination of the remaining clusters.
3. The remaining clusters are matched with the combination that maximizes the proportion of accurate predicted cluster assignments.

2.6. Network Visualization

To create visual representations of networks, we represented each metabolite and species as a node and each correlation as an edge. For easier interpretation of the network diagrams, we included only a subset of metabolites and species. Heimisdottir et al. 2021 [17] identified 16 metabolites, and Cho et al. 2022 [23] identified 15 species in ZOE 2.0 that were significantly associated with the childhood dental disease outcome of interest (i.e., ECC). In this work, we focused on the patterns of co-occurrence between these species and metabolites and examined whether they differ between health and disease states. In network visualizations, we included only the strongest correlations that were of interest. We visually assessed histograms of all correlations for each correlation type and disease group to determine optimal correlation cutoff points. We applied Cytoscape's Organic layout and removed node overlaps. To accomplish this, we first obtained the BZINB-based and Spearman correlations between each pair of 16 metabolites and 15 species of interest, as well as between each pair of the 15 species in ZOE 2.0 in each of the two heath/disease (non-ECC/ECC) participant groups. Next, we sought to determine optimal cutoff correlation values to prevent the network visualization from being too large, even for 16 metabolites and 15 species. Therefore, we created network visualizations only for the most correlated species and the most correlated species–metabolites for the ECC and the non-ECC groups. To maintain comparability of the network diagrams, we used the same percentage of strongest correlations for each. After comparing several cut-off values, we determined that using the top 30% of metabolite–species correlations resulted in approximately 100 edges when the two disease groups were plotted on the same diagram, so that the edges and nodes were mostly visible when the network was large enough to illustrate high-degree nodes.

Network visualizations were generated with Cytoscape 3.9.1 [32]. Metabolite superpathways were highlighted by node color, and edge stroke color was used to denote health/disease (non-ECC/ECC) when correlations from both participant groups were plotted together.

3. Results

3.1. BZINB Model Is a Good Fit for the ZOE 2.0 Microbiome and Metabolome Data

First, we sought to identify suitable distributions to model the paired metabolome and species-level microbiome count data. We assumed that proper normalization in microbiome and metabolome data had been carried out. Zeros present in the original counts remain as zeros after normalization (RPK, RPKM, or CPM).

Specifically, we evaluated model fits for three distributions with multiple randomly selected pairs of species and metabolites from ZOE 2.0. Count data naturally correspond to a Poisson distribution, while the negative binomial distribution is an extension of Poisson that allows for overdispersion. Non-zero data can be transformed to lognormal to improve fit, particularly due to the long right-tailed distribution. It is important to consider that many species and metabolites exhibit large proportions of zeros. Therefore, candidate distributions included: (1) zero-inflated Poisson, (2) zero-inflated negative binomial, and (3) zero-inflated lognormal. For each vector, model parameters were estimated using the nonzero counts from the real data. Numbers of zeros were simulated following a binomial distribution with probability p equal to the proportion of zeros in the real data vector, and the remaining counts were simulated based on the estimated model parameters.

The simulated vectors from the zero-inflated Poisson distribution did not capture the overdispersion in most of the real data vectors (Appendix A Figure A2). The zero-inflated negative binomial distribution was found to adequately capture the data distribution of metabolites and microbiomes (Appendix A Figure A1). Because the negative binomial distribution takes on discrete values, we did not evaluate goodness-of-fit using the Kolmogorov–Smirnov test in this case.

Further, using the Kolmogorov–Smirnov test, we assessed the goodness-of-fit of the lognormal distribution for metabolite and species data in ZOE 2.0 (Figure 1c). Because the Kolmogorov–Smirnov test is only applicable to continuous distributions, only the nonzero counts were included. Regarding metabolites, 11.5% had p values less than 0.05, suggesting that the zero-inflated lognormal distribution was a good fit for most metabolite data. On the other hand, the zero-inflated lognormal distribution was not a good fit for over 20% of the Kraken2/Bracken species, while it was a good fit for almost all HUMAnN 2.0-derived species in ZOE 2.0 (Appendix A Figure A3). Additionally, based on a visual comparison of Kraken2/Bracken real data and simulated zero-inflated lognormal count vectors (Appendix A Figure A1), the zero-inflated lognormal distribution appeared to represent species data well.

3.2. Estimation Accuracy of Underlying Correlation in Simulated Correlated Pairs of Count Data Vectors

We evaluated the estimation accuracy of underlying correlations across our measures of correlation for each simulated pair of vectors. The four methods are: (1) correlation based on the BZINB model (fitted with, at most, 1000 E–M iterations); (2) correlation based on the BNB model (fitted with, at most, 1000 E–M iterations); and (3) Pearson and (4) Spearman correlations for the vectors after elements were set to zero. For each of these simulations, the mean and median correlation approximations were based on 1000 replicates.

In nearly all cases, BZINB and BNB-based correlations were closer to the true and theoretical correlation compared to the Spearman correlation (Figure 2, Appendix A Table A3). As the number of zeros in either vector increased, the Spearman and model-based correlations tended to be lower than the true value. Similarly, as the theoretical correlation increased, the Spearman and model-based correlations also tended to be lower than the true value. These patterns were more noticeable with the Spearman correlation compared to the model-based correlations. BZINB-based correlations were more accurate than Spearman and BNB-based correlations in cases of high simulated underlying correlation or with more zeros, which was more noticeable when the simulated correlation was approximately 0.3 or higher.

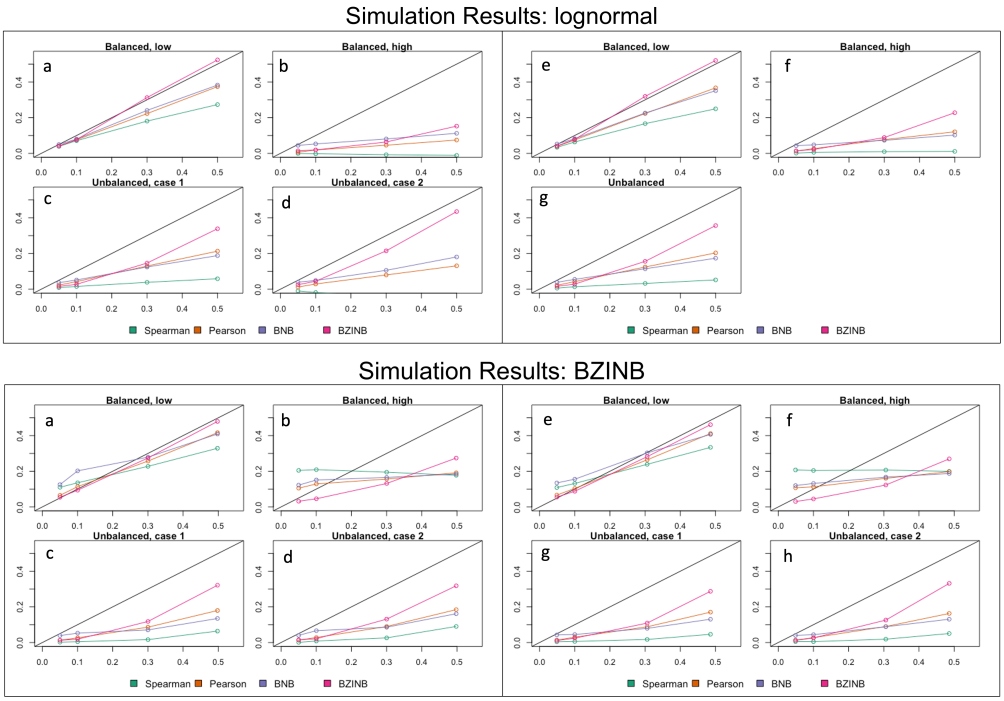


Figure 2. (**Upper panel**) Mean approximated correlation for simulation of lognormal vectors representing pairs of metabolites and species corresponding to the (**a**) balanced, low, (**b**) balanced, high, (**c**) unbalanced, case 1, (**d**) unbalanced, case 2 expected numbers of zeros (parameters in Table 1); mean approximated correlation for simulation of lognormal vectors representing pairs of species corresponding to the (**e**) balanced, low, (**f**) balanced, high, (**g**) unbalanced expected numbers of zeros (parameters in Table 2). Each figure compares Spearman, Pearson, BNB-based, and BZINB-based correlations for five values of underlying correlation from the distributions from which the simulated vectors are drawn. In each plot, the *x*-axis represents the underlying correlation of the bivariate lognormal distribution from which the simulated vector pairs are drawn; the *y*-axis represents the mean estimated correlation for 1000 simulated replicates. (**Lower panel**) Mean approximated correlation for simulation of BZINB vectors representing pairs of metabolites and species corresponding to the (**a**) balanced, low, (**b**) balanced, high, (**c**) unbalanced, case 1, (**d**) unbalanced, case 2 expected numbers of zeros (parameters in Table 2 and Appendix A Table A1); mean approximated correlation for simulation of BZINB vectors representing pairs of species corresponding to the (**e**) balanced, low, (**f**) balanced, high, (**g**) unbalanced, case 1, (**h**) unbalanced, case 2 expected numbers of zeros (parameters in Table 2 and Appendix A Table A2). Each figure compares Spearman, Pearson, BNB-based, and BZINB-based correlations for five values of underlying correlation from the distributions from which the simulated vectors are drawn. In each plot, the *x*-axis represents the underlying correlation of the BZINB distribution from which the simulated vector pairs are drawn; the *y*-axis represents the mean estimated correlation for 1000 simulated replicates.

3.3. Accuracy Evaluation of Identified Species Modules Using Semi-Parametric Simulation

We sought to evaluate the accuracy of species module identification using BZINB-based correlations compared to other correlations for spectral clustering. The ground truth was simulated using semi-parametric simulations as described in the Methods section.

In the crafted semi-parametric simulated dataset representing counts for species belonging to 10 clusters (Figure 3a,b), we constructed affinity (distance) matrices using correlations from three methods (BZINB, BNB, and Spearman correlations) in spectral clustering of species. To evaluate which method produces the most accurate and robust predicted 10 clusters when different distance matrices were used, we compared: (1) proportions of correctly predicted clusters, (2) the Adjusted Rand Index (ARI), and (3) the distance between the correlation matrices of the count matrices before and after adding cluster signals. For all resulting predicted clusters, there were instances where two or more

separately assigned clusters were predicted to be essentially the same cluster (Figure 3c–e). This is likely due to the underlying similarities between species of different clusters in the original count data.

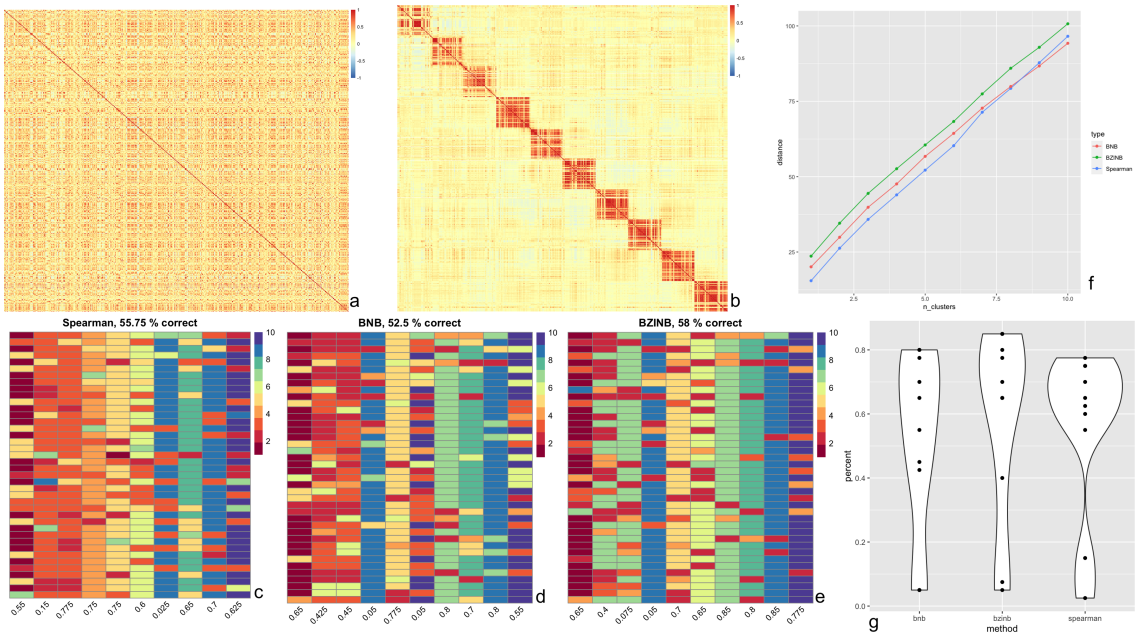


Figure 3. Results of BZINB-based spectral clustering of species. (**a**) Heatmap of BZINB-based correlation between Kraken2/Bracken counts of 400 of the species in ZOE 2.0 in a random order; (**b**) Heatmap of BZINB-based correlation of the Kraken2/Bracken count data (in the same order as in (**a**)) after introducing simulated clusters; (**c**–**e**) Each column of cells represents a true cluster based on simulation (**b**), each cell represents one species, and the color is the indicator of the predicted cluster using the affinity matrix made from (**c**) BNB, (**d**) BZINB, and (**e**) Spearman correlations. There are 10 distinct colors used to represent the 10 clusters; (**f**) Distance (Frobenius norm) between the correlation matrices of nested predicted clusters between data with (as in Figure 3b) and without (as in Figure 3a) increased correlations that represented the clusters: the first set (number of clusters = 1) is the predicted cluster with the greatest distance between correlation matrices. For each increase in the number of clusters, we included an additional cluster in the order of decreasing distances. This was done using the BNB-based, BZINB-based, and Spearman correlation matrices and their corresponding cluster predictions; (**g**) Violin plot of cluster-wise percent accuracy for each of the 10 clusters comparing BNB, BZINB, and Spearman correlation-based affinity matrices.

First, while several approaches exist to quantify clustering accuracy, we considered the proportions of species in each assigned cluster that were predicted to be in the same cluster. We found that, in the data with simulated clusters (simulated as in Section 2.4.2), using the BZINB-based correlation resulted in the highest overall proportion of accurate cluster assignments, while the BNB-based correlation resulted in the lowest accuracy (Figure 3g). Clusters that were generated using BZINB correlations had up to 85% accuracy, and most had at least 65% accuracy. On the other hand, most of the Spearman correlation-based clusters had between 55% to 75% accuracy. There was a moderate percentage (40–55%) of inaccurately predicted BNB correlation-based clusters.

Second, we evaluated the accuracy of the predicted clusters for each correlation type using the ARI. Higher ARI indicates higher consistency between the observed and the simulated cluster membership. In concordance with the proportion of accurate cluster assignments, the affinity matrix based on the BZINB-based correlation resulted in an ARI of 0.43, which was the highest among the three. The ARI for the BNB-based and Spearman correlations were 0.38 and 0.34, respectively. Therefore, BZINB model-based clustering provides the best clustering results.

Third, we compared the three methods according to the distance between correlation matrices. The distance between two correlation matrices (where BZINB correlations were calculated for each pair of species) with partitions representing clusters is one way to compare networks of microbial species or other multi-omics between two health/disease groups. Further, distances between correlation matrices of two health/disease states within each species cluster allows for the determination of clusters that are differentially inter-correlated between these conditions.

Different types of correlation measurements vary in terms of power for detecting between-network differences. Therefore, to compare the correlation types in quantifying the difference between a network with clusters of highly correlated species and a network with clusters of weakly correlated species, we computed distances between the two networks for nested sets of clusters. The first set was the cluster with the greatest distance, and we proceeded by sequentially adding clusters in order of decreasing distances. We used the Frobenius norm of the absolute difference between the correlation (sub-)matrices as the distance measure because it accounts for all matrix entries and is easily understood as an extension of the Euclidean distance between vectors. This was done using the BNB-based, BZINB-based, and Spearman correlation matrices and their corresponding cluster predictions. Distances between two correlation networks were consistently maximized using BZINB correlations, while they were the lowest using Spearman correlations for all but one of the cluster sets (Figure 3f).

3.4. Application in the ZOE 2.0 Study

Interactions among Commensal Species and among ECC-Associated Species

The most abundant species in a microbial community are of natural interest when examining microbial community dynamics in dysbiotic conditions such as those leading to the development of dental caries. They represent a group of commensal species that may be perturbed in the presence of dental disease. Between the top 10 most abundant species in ZOE 2.0, there are stronger correlations in the context of disease (ECC group) compared to the caries-free (non-ECC) group (Figure 4). The Spearman, BNB, and BZINB-based correlations between the 10 most abundant species are very similar because these species have no missing counts. In contrast, when one or more species have higher proportions of zeros, there may be a larger difference between the BNB and BZINB correlations. This is in accordance with simulation results, where all the correlation types were similar under few zeros in both vectors, while the different correlation types were less similar when there were excess zeros in one or more of the vectors.

We also examined interactions between metabolites and species that have been previously shown to be strongly associated with the presence of ECC. Therefore, next, we focused on the set of 15 metabolites and 16 species that have been previously identified to be associated with ECC in differential abundance analyses [17,23]. To understand these ECC-associated interaction networks/pathways, we compared correlations of between-species networks and species–metabolite networks as follows. First, we compared BZINB-based (Figure 5a) and Spearman-based correlations between-species networks (Figure 5b). We found that *Veillonella atypica* is highly correlated with several ECC-associated *Prevotella* species among children affected with ECC using both of these correlations (Figure 5a,b). On the other hand, many of these *Prevotella* species tend to be strongly correlated with *Leptotrichia*, *Lachnospiraceae*, and *Lachnoanaerobaculum* species in children unaffacted by ECC. This points to two possible co-abundance patterns: one where *Prevotella*, *Leptotrichia*, *Lachnospiraceae*, and *Lachnoanaerobaculum* taxa may coexist in biofilms without disease and another pattern of mutual benefit among *V. atypica* and *Prevotella* species when disease is present. In this case, the co-abundance pattern between these two species can be explained by their beneficial interrelation in metabolic activities: carbohydrates and sugar alcohols from the diet are subjected to glycolysis, which creates anaerobic conditions by consuming oxygen and produces pyruvate that can be converted into lactate by *Prevotella*

species. On the other hand, *Veillonella atypica* is an anaerobic bacterium that uses lactate as its sole carbon source, converting into weaker acids, such as acetate and propionate [33].

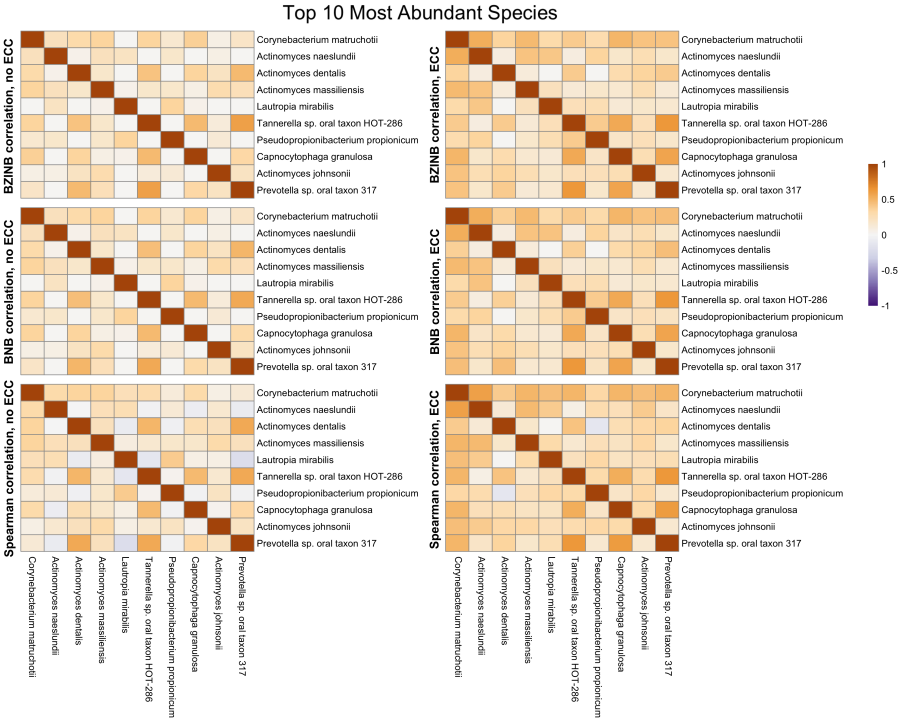


Figure 4. Heatmaps of BZINB-based and Spearman correlations between the top 10 species with the highest overall abundance for each health/dental disease group (non-ECC versus ECC) in the ZOE 2.0 Kraken2/Bracken data.

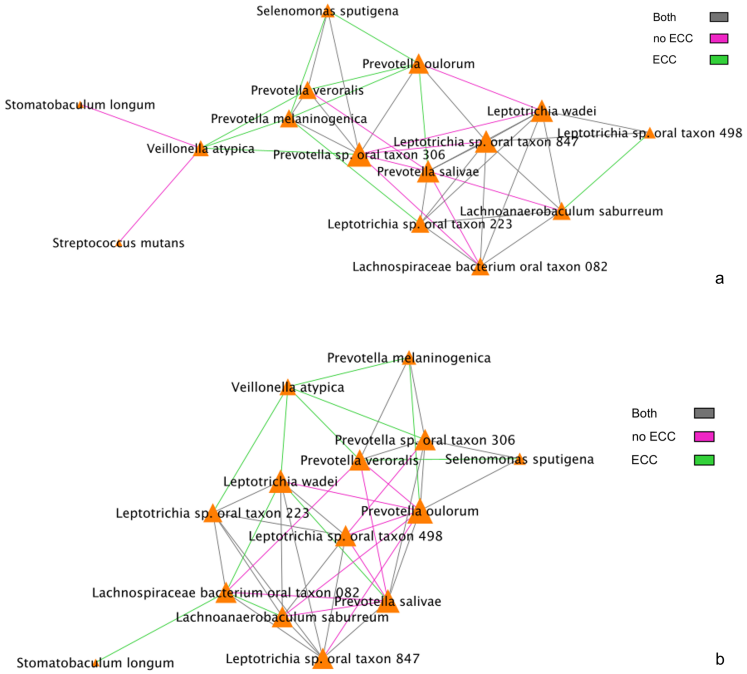


Figure 5. (**a**) BZINB correlations between species. The strongest 30% of correlations are included in the diagrams, and the color of the lines represents whether the correlation was strong in one or both of the health/disease groups. (**b**) Spearman correlations between species. The strongest 30% of correlations are included in the diagrams, and the color of the lines represents whether the correlation was strong in one or both of the health/disease groups.

Among the 15 species of interest, the BZINB correlation network included only one strong correlation involving *Streptococcus mutans* and *Veillonella atypica* in healthy subjects, whereas the Spearman correlation network did not include *Streptococcus mutans* at all. *Streptococcus* and *Veillonella* species are very common in supragingival oral biofilm, and [34] showed a *Streptococcus–Veillonella* link in early dental plaque formation. In fact, *Streptococcus mutans* is well known as a major lactic acid producer from the fermentation of dietary carbohydrates, which benefits *Veillonella* species since it utilizes lactate produced by *Streptococcus mutans* and converts it into weaker acids, such as acetate and propionate, contributing to acid neutralization. Therefore, the identified strong correlation between the two species is expected and reasonable. However, when acid production occurs at a greater rate and frequency than that of acid neutralization, dental caries will develop. In subjects with caries, *Veillonella atypica* was more abundant compared to those without caries (Figure 6). Therefore, the *Streptococcus mutans–Veillonella atypica* dynamic may be somewhat overpowered by *Streptococcus mutans* once disease has been established.

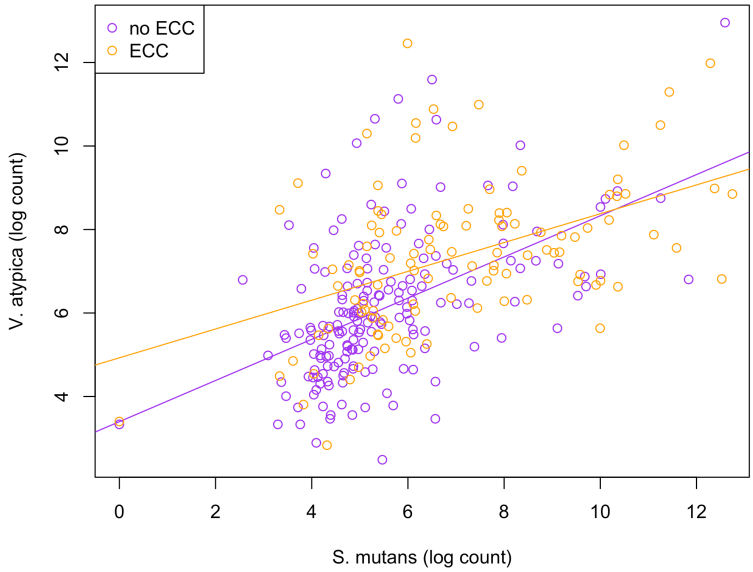


Figure 6. Scatterplot illustrating the comparison of relationships between *S. mutans* and *V. atypica* abundances between healthy (non-ECC) and disease (ECC) groups.

Additionally, we compared BZINB-based (Figure 7) and Spearman correlation-based species–metabolite networks (Figure 8). In the oral biofilm, when diet-associated carbohydrates are present, carbohydrate-degrading species tend to increase in abundance, and the local environment pH may decrease [35]. To observe the differences in species that are highly correlated with carbohydrates of interest in healthy subjects and subjects with ECC, we focused on interpretation of four carbohydrates that were previously shown to be significantly and positive associated with ECC in Heimisdottir et al. 2021 [17]. We used the BZINB-based correlations because some of the species had excess zeros. For each of the five carbohydrates, we compared the strongest 5% of metabolite–species correlations between healthy/disease groups. In caries-affected participants, the amount of three of the carbohydrates (fucose, sedoheptulose-7-phosphate, and N-acetylneuraminate) is strongly correlated with many *Prevotella* species. According to Takahashi et al. 2005 [36], *Prevotella* neutralizes pH but may also favor the presence of other pathogenic species. In healthy subjects, we found the carbohydrates to be correlated with *Streptococcus*, *Fusobacterium*, and *Selenomonas* species, many of which have been described as carbohydrate-degrading or pH-neutralizing in the oral biofilm [36,37] or are a core part of the normal flora. In the BZINB network, 3-(4-hydroxyphenyl)lactate (HPLA) had many strong correlations with various species in participants with ECC but much less among unaffected ones. HPLA is a

metabolite in the tyrosine metabolism pathway that functions similarly to lactate, which has been previously shown to be an important metabolic regulator in multiple pathways (including glucose metabolism) in various parts of the human body [38,39]. The differing strengths of correlations in the two healthy/disease groups could indicate that HPLA is metabolized differently by ECC-associated species in the context of a dental caries-promoting environment and may be a candidate for further investigation in its role in ECC development. Furthermore, HPLA is strongly associated with many *Streptococcus* species in healthy subjects and with many *Prevotella* species among those with ECC, similar to what was found for ECC-associated carbohydrates.

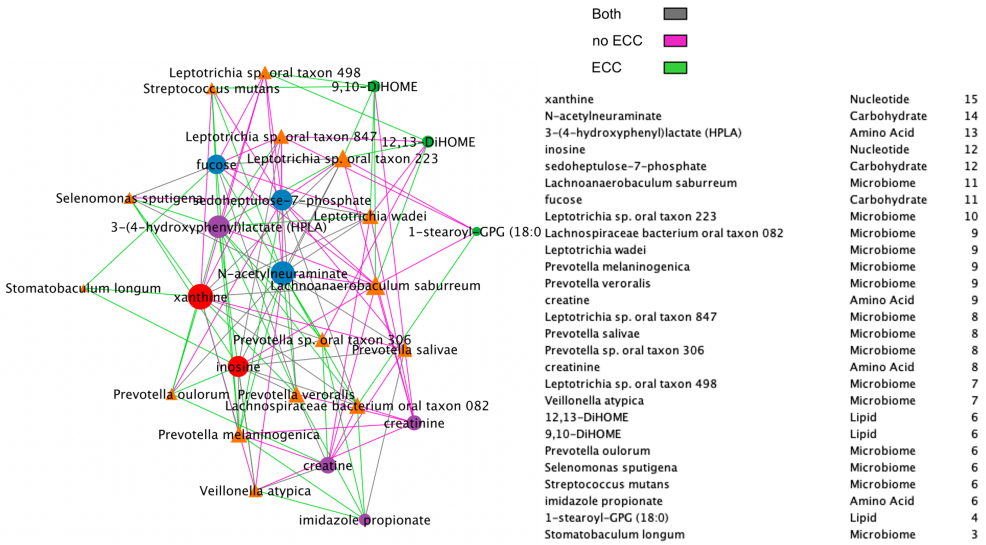


Figure 7. BZINB network between species and metabolites including a node degree table. The strongest 30% of correlations are presented in the diagrams, and line colors represent whether the correlation was strong in one or both of the healthy/disease groups.

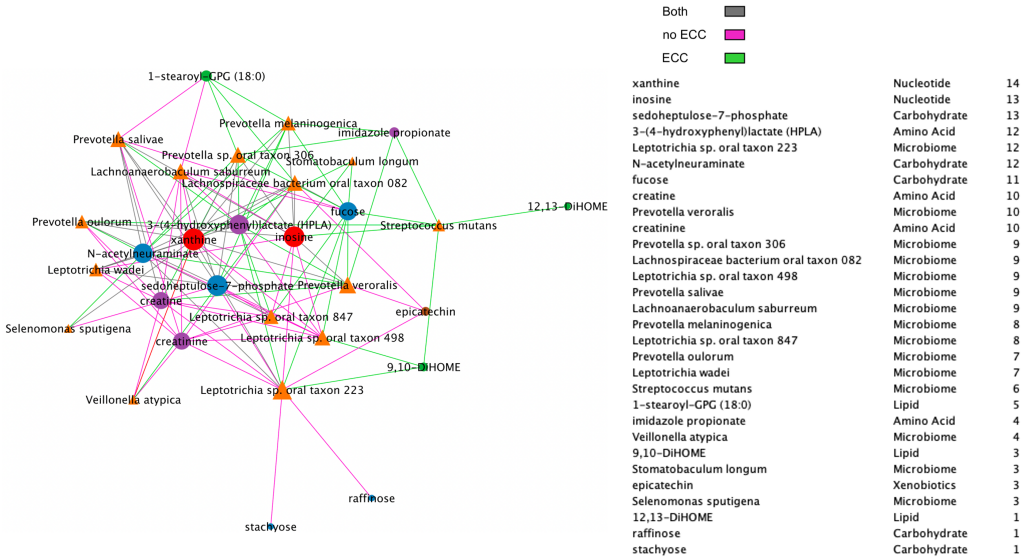


Figure 8. Spearman network between species and metabolites, presenting positive correlations only and including a node degree table. The strongest 30% of positive correlations are presented in the diagram, and line colors represent whether the absolute correlation was strong in one or both of the healthy/disease groups.

Overall, Spearman and Pearson correlations are not suitable for data with excess zeros because Spearman is influenced by ties and Pearson assumes a linear association.

The negative binomial distribution accounts for the presence of zeros, which makes the BNB distribution a better choice for modeling the relationship between a typical pair of species or metabolites. When there are excess zeros in either or both species or metabolites in a pair, the BZINB model can account for the zero inflation while approximating the correlation of the nonzero components.

3.5. Species Modules Identified Using BZINB-Based Correlation and Spectral Clustering

We applied cut-based spectral clustering to the ZOE 2.0 data separately for each healthy/disease group. We compared results between BZINB-based and Spearman correlations when constructing the affinity matrix. To determine the optimal number of clusters, we plotted the eigenvalues of the graph Laplacian for each affinity matrix (Appendix A Figure A5). According to the eigengap method [40], the optimal number of clusters was 2 for each affinity matrix; for more interpretable results, we set the number of clusters to be 6 in each case. To visualize the results of cut-based clustering, we created heatmaps of standardized counts for all species, where the species are grouped and annotated by predicted cluster and the study participants are annotated according to health/disease and batch group. There were visible within-cluster similarities and differences between the clusters for count patterns (Figure 9). Many species that were predicted to be in the second and fifth cluster (shown in blue and orange, respectively, in the top bars of Figure 9) in the healthy group had been classified in the third cluster (shown in green) in the disease group. In other words, some species that were more similar to the first and fifth clusters in the healthy group were instead more similar to the third cluster in subjects with ECC. The different co-varying patterns in these species may be a reflection of differences in the microbial community structure and function in ECC.

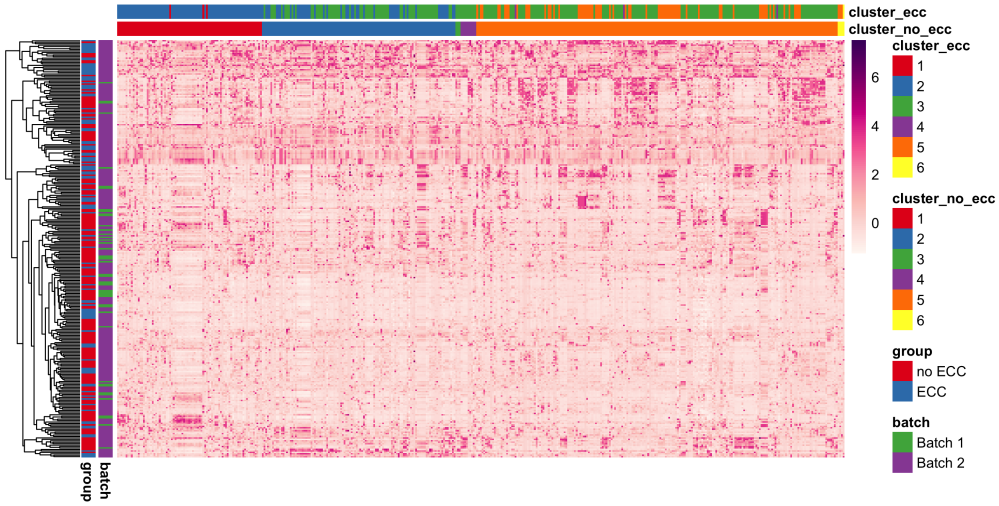


Figure 9. Heatmap of species-wise standardized counts illustrating species module identification results (species are columns and modules are presented with different colors) using BZINB-based species spectral clustering. Each column represents a single species. Columns are ordered by the clusters predicted from the affinity matrix based on the BZINB correlations between species in the healthy (non-ECC) group. The columns are annotated to show and compare the estimated clusters within health (non-ECC) and disease (ECC) groups. Each row represents a participant, and the rows are ordered based on hierarchical clustering. Rows (n = 289) are annotated to denote healthy/disease and batch groups. Standardized counts were calculated separately for each species by first suppressing outliers (defined as greater than 3 × IQR from the first or third quartile), then centering around the median and dividing by the standard deviation.

4. Discussion

In this paper, we introduced a new method, BZINB-iMMPath, entailing a bivariate zero-inflated negative binomial (BZINB) model-based correlation for network analysis

of pairs of vectors of omics count data and module identification. The model makes reasonable assumptions regarding dropouts and excess zeros as structural zeros in the observed microbiome data compared to other types of zeros. Therefore, the microbial correlation distribution is assumed to be that of the latent bivariate negative binomial model. Our approach improves the estimation of correlations compared to the traditional Pearson correlation and the more robust Spearman's rank correlation coefficients. In contrast to Pearson and Spearman correlations, the BZINB model accommodates zeros in a flexible manner (in either or both vectors of each pair) and estimates the correlation under the bivariate negative binomial model. For each pair of omics features, the BZINB model is fitted, and a model-based correlation is computed from the estimated parameters. Using the model, we can calculate the correlations between pairs of omics features in the same layer (i.e., between pairs of microbial species) or between two different layers (i.e., between pairs of metabolites and species). These correlations may then be used in other applications such as networks' visual representations and identification of clusters of omics features. Accordingly, we applied the new method to microbial species and metabolite data obtained in an oral microbiome study of early childhood dental disease. Using visual comparisons and goodness-of-fit tests, we determined that the negative binomial and lognormal distributions were appropriate for modeling most metabolites and species. In addition to accounting for zero inflation, marginally, the negative binomial distribution is a natural choice to model count data. Therefore, our model-based correlation approach has several advantages over conventional measures of correlation when applied to bivariate count data with excess zeros. In addition, correlations estimated from BZINB can be used as the affinity matrix in the cut-based spectral clustering method for species module identification in zero-inflated microbiome data. Modules can be compared between groups of interest (e.g., health versus disease) and help identify species that demonstrate important between-group pattern differences.

To evaluate the performance of BZINB-iMMPath, we used real data-inspired simulations to estimate the accuracy of underlying correlations in microbiome data; real data-based semi-parametric simulations to assess the accuracy of module identification; and finally, we applied it in a sizeable oral microbiome study to identify ECC-associated microbial networks and modules. Specifically, we simulated pairs of count vectors representing typical metabolite and microbial species vectors from ZOE 2.0 to compare the accuracy of Spearman, BNB, and BZINB model-based correlations. We fitted the BZINB model to each metabolite–species and species–species pair to construct visualizations of ECC disease group-specific filtered networks and build affinity matrices for cut-based spectral clustering. Using the simulated vector pairs, the BZINB model-based correlation was on average closer to the underlying correlation when there were more zeros in one or both vectors compared to the Spearman correlation coefficient. Notably, the average BZINB-based correlation was higher than the other correlation types when the underlying correlation was high (>0.3) and when there was zero inflation in at least one of the vectors. Therefore, we recommend using the BZINB-based correlation for the identification of strongly correlated pairs when zero inflation is present. The application in ZOE 2.0 not only highlighted previously known networks involving carbohydrate metabolites but also revealed novel regulation relationships between species and metabolites and ECC-associated species modules.

While our method focuses on identifying pathways through identifying species modules based on correlated abundances and constructing networks, the inferred clusters from BZINB-iMMPath could be used in extension to perform differential network/pathway analysis by testing the difference (represented as distance) of species-wise correlation structures between sample groups in one specific cluster for a more site- or disease-specific perspective.

The most noticeable limitation of the new approach is that the BZINB model allows for only positive model-based correlations. Of course, there are cases where negative correlations are of interest; for example, in the context of species competition, other cor-

relation measures could be used. If negative correlations are also of interest in network visualizations and strong (<-0.3) [41] negative Spearman correlations were observed, the negative Spearman correlations could be directly used in place of near-zero BZINB model-based correlations, or the weighted Pearson correlation could be used by weighting the observed abundance counts by the model-based dropout probability as suggested in Cho et al. 2021 [10]. However, incorporating negative correlations can introduce another layer of complexity to network analysis applications for multi-omics and cluster identification. For example, negative correlations may be considered with different importance compared to positive correlations. Further, negative correlations within one layer of omics (such as the microbiome), which could represent competition, may be more of interest compared to negative correlations between layers (e.g., microbiome and metabolome), which could be more complex in terms of direction of influence. This leaves room for future method development, for example, wherein other bivariate (or multivariate) models can be evaluated in terms of goodness-of-fit for certain types of omics data that could accommodate negative correlations; however, there are advantages and disadvantages. For example, the results of goodness-of-fit tests (Figure 1c) suggest that the lognormal distribution is appropriate for modeling the nonzeros for most of the species and metabolites in our dataset so that zero-inflated log-normal (ZILN) based multivariate models can be another option for our purpose. The so-called ZILN distribution is actually a truncated lognormal that has a point mass at zero and a log-normal distribution for positive values. It has been previously used for microbial networks and considers both positive and negative correlations as in Prost et al. 2021 [42]. However, correlation of the multi-LN component may not fully address the different mechanisms (e.g., biological zeros, technical zeros) that generate zeros. In reality, some structural zeros representing non-existing species in the sample can be important, as considered in correlation of the BNB component in BZINB. Meanwhile, identifying positive correlations between bacteria and metabolites is a logical priority because of biological interest regarding (1) which bacteria generate or up-regulate which metabolites, and (2) which biochemicals are associated with bacterial abundance (e.g., possibly growth). Meanwhile, negative correlation (like inhibition or competition) is harder to interpret as detailed above, and in our BZINB model, positive correlations are presented as such and negative correlations are estimated as near-zero.

In our application to the ZOE 2.0 study microbiome data, we determined that: (1) there were relatively fewer zero counts when taxa were identified through the oral health-specific Kraken2/Bracken pipeline, compared to the data from the still widely used HUMAnN 2.0 pipeline; (2) zero inflation does not appear to be a significant issue for many of the named metabolites; and (3) in the absence of excess zeros, other measures of correlation appear to be just as adequate as the BZINB-based correlation. Because HUMAnN 2.0 generated data are very sparse, our method is even more powerful in those data, as well as in similarly sparse gene-level metagenomics or metatranscriptomics data.

In sum, in this paper, we demonstrated that the new method based on the BZINB model is a useful alternative to Spearman or Pearson correlations in estimating underlying correlations for bivariate count data that are zero-inflated in one or both dimensions. Because the model accommodates both technical and true zeros, it is suitable for multi-omics data types, including the microbiome and metabolome. To identify differences between healthy/disease groups, we prioritized and illustrated the strongest correlations within each group, allowing for the visualization of important dynamic relationships and their between-group comparison. Finally, these correlations can also be used in identifying modules, i.e., clusters of correlated metabolites and microbial species, which could be of biological interest both in terms of disease pathogenesis and intervention targeting.

Author Contributions: Conceptualization, B.M.L. and D.W.; methodology, B.M.L., H.C. and D.W.; software, B.M.L., H.C. and C.L.; validation, B.M.L. and A.A.R.; formal analysis, B.M.L.; resources, K.D., J.R. and C.L.; data curation, K.D.; writing—original draft preparation, B.M.L. and D.W.; writing—review and editing, B.M.L., H.C., K.D., A.A.R., J.R. and D.W.; visualization, B.M.L. and D.W.; supervision, D.W. All authors read and agreed to the published version of the manuscript.

Funding: This work was funded by grants from the National Institutes of Health, National Institute of Dental and Craniofacial Research, R03-DE028983 and U01-DE025046.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board (or Ethics Committee) of the University of North Carolina-Chapel Hill (#14-1992, latest approval on 21 February 2022).

Informed Consent Statement: Written informed consent was obtained from legal guardians of all children who participated in the ZOE 2.0 study.

Data Availability Statement: ZOE 2.0 microbiome data are publicly available in the dbGaP repository at https://www.ncbi.nlm.nih.gov/gap; (accessed on 4 March 2023) under the umbrella study name Trans-Omics for Precision Dentistry and Early Childhood Caries or TOPDECC (accession: phs002232.v1.p1) via the Sequence Read Archive (SRA) Bioproject PRJNA671299 at https://www.ncbi.nlm.nih.gov/bioproject/671299; (accessed on 4 March 2023). Metabolomics raw spectral data have been made publicly available via the MetaboLights repository project MTBLS2215 at https://www.ebi.ac.uk/metabolights/MTBLS2215; (accessed on 4 March 2023). The code used for analysis is available at https://github.com/blin24/BZINB-iMMPath; (accessed on 4 March 2023).

Acknowledgments: The authors would like to thank ZOE 2.0 study participants for their contributions.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

ZINB	Zero-inflated negative binomial
BNB	Bivariate negative binomial
BZINB	Bivariate zero-inflated negative binomial
MI	Mutual Information
ECC	Early childhood caries

Appendix A

Table A1. Shape and scale parameters used to obtain various values of correlation for BZINB simulation for vector pairs that represent pairs of metabolites and species. The number labels of each line denote the parameters used to obtain results for each line of Appendix A Table A3.

	α_0	α_1	α_2	β_1	β_2	ρ_{BZINB}
1	0.2	0.1	0.3	20	40	0.498
2	0.3	0.5	0.8	12	21	0.300
3	0.15	1.1	1.5	20	30	0.100
4	0.05	0.85	1	30	50	0.050

Table A2. Shape and scale parameters used to obtain various values of correlation for BZINB simulation for vector pairs that represent pairs of species. The number labels of each line denote the parameters used to obtain results for each line of Appendix A Table A3.

	α_0	α_1	α_2	β_1	β_2	ρ_{BZINB}
1	0.3	0.3	0.3	30	40	0.486
2	0.35	0.7	0.8	20	30	0.306
3	0.1	0.75	1	30	30	0.100
4	0.05	0.9	1	50	50	0.049

Table A3. The estimated correlations for each set of simulations: median (standard deviation). The number label of each line corresponds to the shape and scale parameters in Appendix A Table A1 or Appendix A Table A2, and the letter label of each line corresponds to the zero inflation parameters in Table 1 or Table 2.

		Theoretical	Spearman	Pearson	BNB	BZINB
lognormal (metabolome–microbiome)	1a	0.5	0.273 (0.056)	0.374 (0.115)	0.383 (0.047)	0.524 (0.058)
	1b		−0.011 (0.06)	0.075 (0.122)	0.112 (0.054)	0.153 (0.21)
	1c		0.058 (0.058)	0.213 (0.118)	0.188 (0.049)	0.339 (0.18)
	1d		−0.064 (0.06)	0.131 (0.123)	0.181 (0.07)	0.435 (0.205)
	2a	0.3	0.181 (0.058)	0.223 (0.096)	0.241 (0.058)	0.313 (0.08)
	2b		−0.007 (0.058)	0.046 (0.099)	0.081 (0.048)	0.064 (0.127)
	2c		0.038 (0.06)	0.13 (0.101)	0.124 (0.06)	0.146 (0.144)
	2d		−0.043 (0.062)	0.08 (0.093)	0.106 (0.064)	0.215 (0.171)
	3a	0.1	0.071 (0.058)	0.076 (0.071)	0.08 (0.061)	0.077 (0.068)
	3b		−0.002 (0.058)	0.019 (0.071)	0.053 (0.041)	0.019 (0.05)
	3c		0.015 (0.059)	0.042 (0.071)	0.052 (0.053)	0.027 (0.053)
	3d		−0.019 (0.059)	0.029 (0.07)	0.049 (0.046)	0.045 (0.07)
	4a	0.05	0.039 (0.059)	0.042 (0.065)	0.05 (0.052)	0.042 (0.048)
	4b		0 (0.058)	0.007 (0.063)	0.045 (0.036)	0.014 (0.042)
	4c		0.009 (0.056)	0.023 (0.061)	0.036 (0.042)	0.016 (0.034)
	4d		−0.01 (0.059)	0.012 (0.065)	0.04 (0.042)	0.025 (0.045)
lognormal (within microbiome)	1a	0.5	0.25 (0.058)	0.367 (0.11)	0.351 (0.047)	0.52 (0.065)
	1b		0.011 (0.061)	0.121 (0.13)	0.102 (0.046)	0.228 (0.242)
	1c		0.052 (0.06)	0.203 (0.123)	0.173 (0.051)	0.356 (0.191)
	2a	0.3	0.167 (0.06)	0.223 (0.102)	0.226 (0.056)	0.319 (0.083)
	2b		0.01 (0.057)	0.077 (0.107)	0.073 (0.041)	0.089 (0.151)
	2c		0.032 (0.059)	0.124 (0.099)	0.114 (0.054)	0.156 (0.154)
	3a	0.1	0.064 (0.059)	0.075 (0.072)	0.082 (0.061)	0.078 (0.073)
	3b		0.006 (0.06)	0.027 (0.077)	0.048 (0.037)	0.022 (0.055)
	3c		0.014 (0.057)	0.042 (0.073)	0.055 (0.046)	0.028 (0.055)
	4a	0.05	0.034 (0.059)	0.04 (0.064)	0.052 (0.051)	0.041 (0.05)
	4b		0.002 (0.061)	0.013 (0.063)	0.043 (0.035)	0.014 (0.041)
	4c		0.006 (0.059)	0.022 (0.067)	0.039 (0.04)	0.016 (0.034)
BZINB (metabolome–microbiome)	1a	0.4978	0.329 (0.056)	0.416 (0.113)	0.409 (0.047)	0.48 (0.07)
	1b		0.177 (0.069)	0.192 (0.146)	0.184 (0.067)	0.274 (0.224)
	1c		0.064 (0.063)	0.18 (0.128)	0.135 (0.056)	0.321 (0.2)
	1d		0.09 (0.058)	0.185 (0.118)	0.162 (0.056)	0.318 (0.203)
	2a	0.300	0.228 (0.061)	0.257 (0.083)	0.279 (0.055)	0.275 (0.074)
	2b		0.195 (0.065)	0.156 (0.103)	0.166 (0.056)	0.132 (0.164)
	2c		0.016 (0.057)	0.086 (0.082)	0.071 (0.047)	0.118 (0.151)
	2d		0.026 (0.058)	0.09 (0.085)	0.086 (0.053)	0.131 (0.164)
	3a	0.100	0.135 (0.059)	0.114 (0.067)	0.202 (0.068)	0.094 (0.061)
	3b		0.209 (0.063)	0.129 (0.076)	0.151 (0.047)	0.046 (0.082)
	3c		0.005 (0.057)	0.026 (0.066)	0.053 (0.04)	0.019 (0.046)
	3d		0.008 (0.058)	0.028 (0.065)	0.066 (0.047)	0.021 (0.052)
	4a	0.050	0.11 (0.06)	0.065 (0.065)	0.125 (0.067)	0.054 (0.047)
	4b		0.205 (0.064)	0.105 (0.076)	0.122 (0.049)	0.032 (0.062)
	4c		0.002 (0.057)	0.013 (0.063)	0.039 (0.034)	0.013 (0.027)
	4d		0.001 (0.058)	0.014 (0.063)	0.04 (0.036)	0.015 (0.038)

Table A3. *Cont.*

		Theoretical	Spearman	Pearson	BNB	BZINB
BZINB (within microbiome)	1a	0.486	0.334 (0.057)	0.412 (0.094)	0.407 (0.044)	0.461 (0.061)
	1b		0.2 (0.066)	0.198 (0.13)	0.187 (0.059)	0.27 (0.22)
	1c		0.046 (0.058)	0.17 (0.108)	0.131 (0.047)	0.287 (0.191)
	1d		0.051 (0.06)	0.163 (0.105)	0.131 (0.047)	0.332 (0.194)
	2a	0.306	0.239 (0.058)	0.264 (0.081)	0.303 (0.05)	0.282 (0.071)
	2b		0.207 (0.061)	0.161 (0.093)	0.167 (0.051)	0.123 (0.157)
	2c		0.017 (0.059)	0.088 (0.083)	0.079 (0.047)	0.109 (0.139)
	2d		0.019 (0.061)	0.09 (0.082)	0.089 (0.047)	0.126 (0.151)
	3a	0.100	0.132 (0.059)	0.104 (0.069)	0.156 (0.061)	0.088 (0.058)
	3b		0.204 (0.064)	0.113 (0.077)	0.132 (0.05)	0.045 (0.081)
	3c		0.006 (0.058)	0.03 (0.068)	0.045 (0.038)	0.025 (0.056)
	3d		0.005 (0.058)	0.028 (0.069)	0.044 (0.039)	0.026 (0.058)
	4a	0.049	0.109 (0.06)	0.066 (0.066)	0.135 (0.066)	0.055 (0.047)
	4b		0.208 (0.062)	0.107 (0.07)	0.12 (0.046)	0.03 (0.058)
	4c		0.004 (0.056)	0.014 (0.06)	0.043 (0.034)	0.011 (0.017)
	4d		0.005 (0.056)	0.014 (0.06)	0.04 (0.035)	0.012 (0.022)

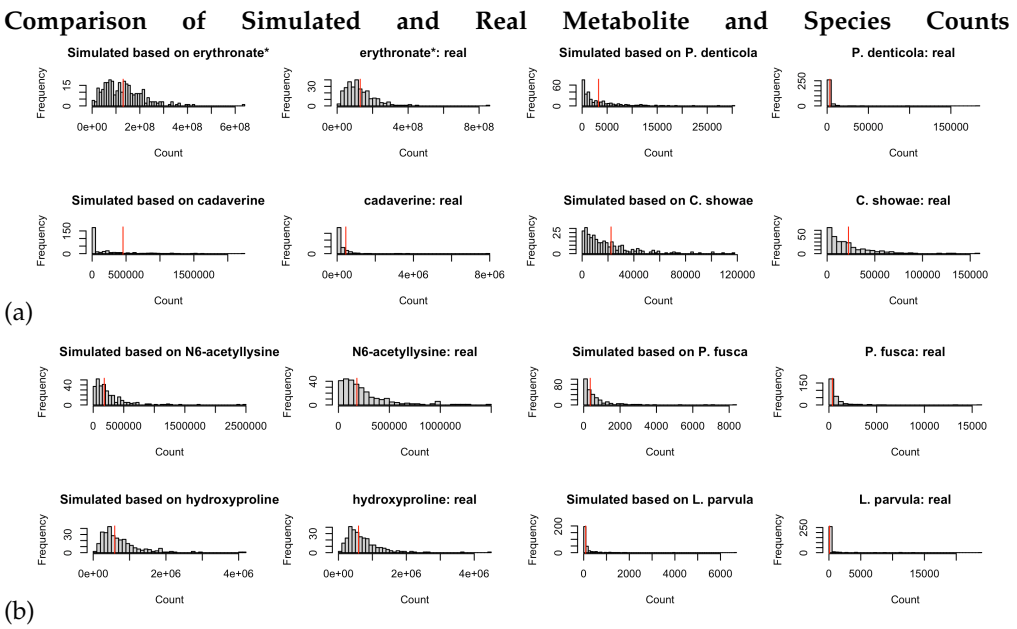


Figure A1. Evaluation of goodness-of-fit for the ZINB/NB and lognormal models by comparing the histogram of frequency between the simulated data and real data processed in Kraken2/Bracken. The 2nd and 4th columns each contain one species or metabolite from ZOE 2.0. The 1st and 3rd columns contain ZINB simulated data using the parameters estimated from the corresponding species. The 1st and 2nd columns illustrate metabolites, and the 3rd and 4th columns present microbial species. The star at the end of some of the metabolite names indicates that Metabolon, where the metabolome data were generated, is confident in the metabolite's identity but it has not been confirmed based on a standard. (**a**) denotes ZINB/NB-based simulation, and (**b**) denotes ZILN/lognormal based simulation. (**a**) For two randomly selected metabolites and two randomly selected species (Kraken2/Bracken), comparison of simulated counts drawn from the ZINB distribution (with parameters obtained from models fitted on the real data) and the real data. If the real data have less than 50 out of 289 zeros, the simulated counts are drawn from the negative binomial distribution with no zero inflation. Red vertical lines represent model-based means for each metabolite and species. (**b**) For two randomly

selected metabolites and two randomly selected species (Kraken2/Bracken), comparison of simulated counts drawn from the (ZI-)lognormal distribution (with parameters obtained from models fitted on the real data) and the real data. If the real data have no zeros, the simulated counts are drawn from the lognormal distribution with no zero inflation. Red vertical lines represent the log-scale means of the counts for each metabolite and species.

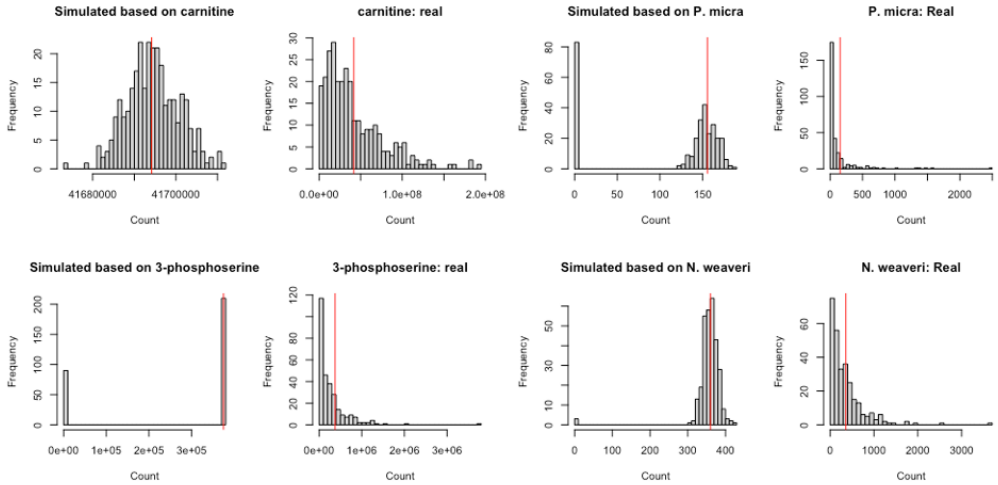


Figure A2. Comparison of simulated counts drawn from (ZI-)Poisson distribution (with parameters from model fitted on the real data) and real data of 4 randomly-selected metabolites and species. The red vertical line on each plot represents the Poisson model-based mean of the nonzero component for each metabolite and species.

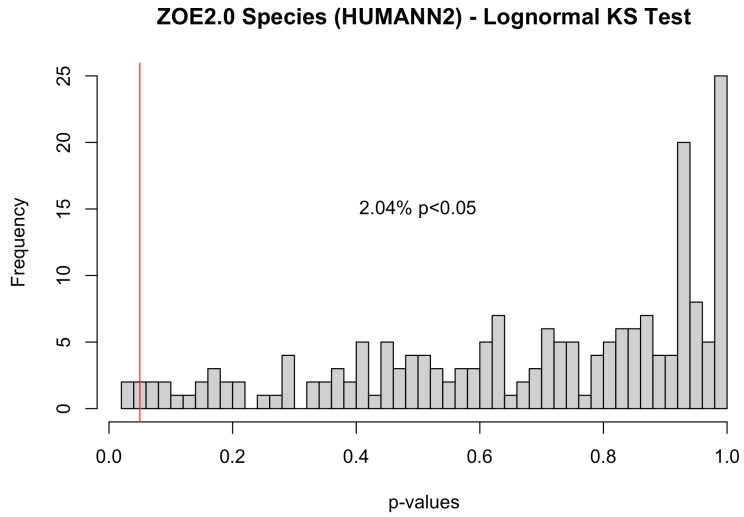


Figure A3. *p*-values obtained from lognormal (parameters from models fitted on nonzero counts for each metabolite and species) Kolmogorov–Smirnov test for ZOE 2.0 metabolites and HUMAnN 2.0 microbiome species. The red vertical line represents a *p*-value of 0.05 so that *p*-values below it indicate statistical significance in the tests.

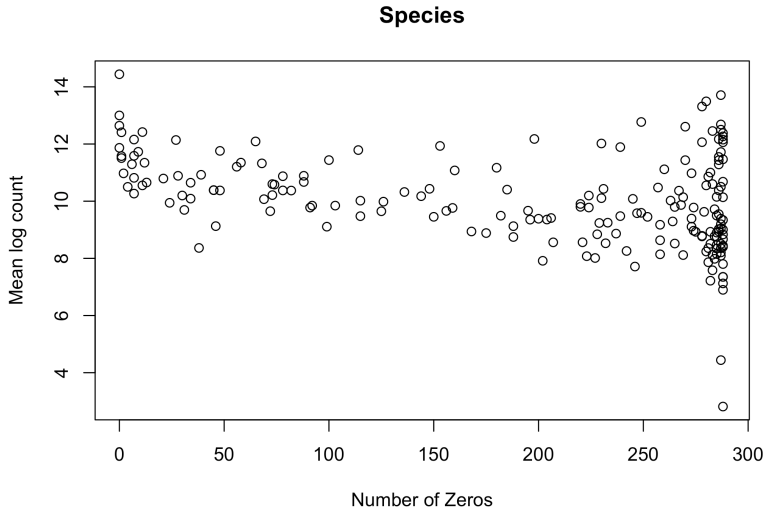


Figure A4. Species-wise (HUMAnN 2.0) numbers of zeros plotted against mean log nonzero counts.

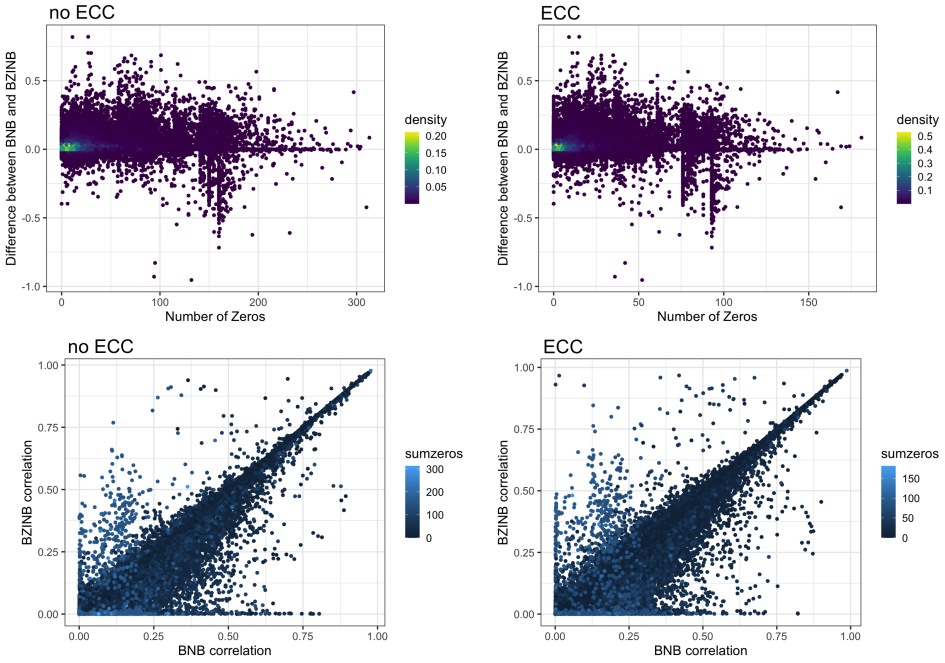


Figure A5. Comparison of BNB and BZINB correlations between all pairs of microbial species in ZOE 2.0 with respect to the total number of zeros in each pair.

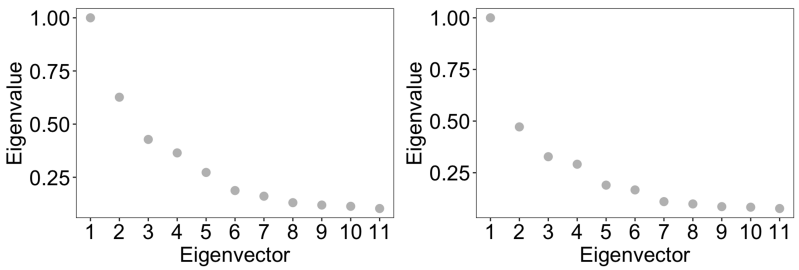


Figure A6. Eigenvalues of the Laplacian graph based on each affinity matrix for healthy and disease (ECC) groups in ZOE 2.0 Kraken2/Bracken microbiome data, used to determine an appropriate number of clusters.

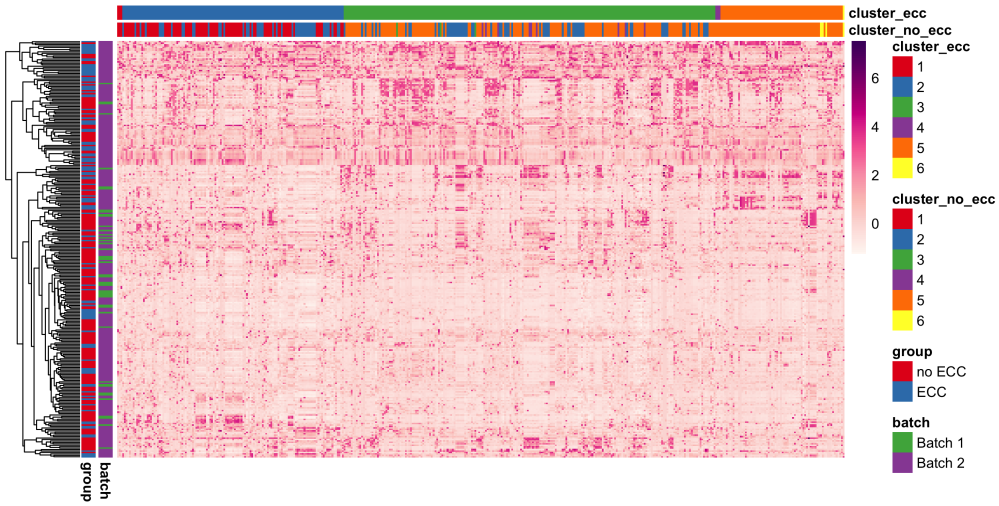
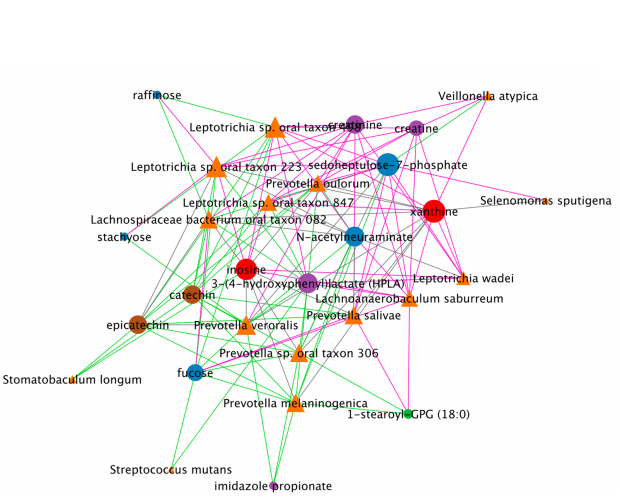


Figure A7. Heatmap of species-wise standardized counts illustrating species modules identified by spectral clustering (shown as the two top bars). Columns represent individual species and are ordered by the clusters predicted from the affinity matrix based on the BZINB correlations between species in the disease (ECC) group. Columns are annotated to show and compare the predicted clusters between healthy (no ECC) and disease (ECC) groups. Each row represents a participant (n = 289). The rows are ordered based on hierarchical clustering and are annotated to illustrate healthy and disease groups and the sequencing batch. Standardized counts were calculated separately for each species by first suppressing outliers (defined by greater than 3 × IQR from the first or third quartile), then centering around the median and dividing by the standard deviation.



sedoheptulose−7−phosphate	Carbohydrate	13
xanthine	Nucleotide	13
inosine	Nucleotide	12
Leptotrichia sp. oral taxon 223	Microbiome	12
Leptotrichia sp. oral taxon 498	Microbiome	12
3−(4−hydroxyphenyl)lactate (HPLA)	Amino Acid	11
Prevotella veroralis	Microbiome	11
N−acetylneuraminate	Carbohydrate	11
catechin	Xenobiotics	10
epicatechin	Xenobiotics	10
Prevotella melaninogenica	Microbiome	10
Prevotella sp. oral taxon 306	Microbiome	10
Lachnospiraceae bacterium oral taxon 082	Microbiome	10
creatinine	Amino Acid	10
Prevotella salivae	Microbiome	10
fucose	Carbohydrate	9
Prevotella oulorum	Microbiome	9
Leptotrichia sp. oral taxon 847	Microbiome	9
Lachnoanaerobaculum saburreum	Microbiome	9
creatine	Amino Acid	8
Leptotrichia wadei	Microbiome	7
1−stearoyl−GPG (18:0)	Lipid	4
Veillonella atypica	Microbiome	4
imidazole propionate	Amino Acid	3
Stomatobaculum longum	Microbiome	3
raffinose	Carbohydrate	3
stachyose	Carbohydrate	3
Streptococcus mutans	Microbiome	2
Selenomonas sputigena	Microbiome	2

Figure A8. Spearman microbiome–metabolome correlation network including a node degree table. The strongest 30% absolute correlations are illustrated. Line colors represent correlations' strength in health, disease (ECC), or both.

References

1. Bauer, M.A.; Kainz, K.; Carmona-Gutierrez, D.; Madeo, F. Microbial Wars: Competition in Ecological Niches and within the Microbiome. *Microb. Cell* **2018**, *5*, 215–219. [CrossRef]
2. Tong, H.; Chen, W.; Merritt, J.; Qi, F.; Shi, W.; Dong, X. Streptococcus Oligofermentans Inhibits Streptococcus Mutans through Conversion of Lactic Acid into Inhibitory H2O2: A Possible Counteroffensive Strategy for Interspecies Competition. *Mol. Microbiol.* **2007**, *63*, 872–880. [CrossRef] [PubMed]
3. Nyvad, B.; Crielaard, W.; Mira, A.; Takahashi, N.; Beighton, D. Dental Caries from a Molecular Microbiological Perspective. *Caries Res.* **2012**, *47*, 89–102. [CrossRef]

4. Mira, A.; Simon-Soro, A.; Curtis, M.A. Role of Microbial Communities in the Pathogenesis of Periodontal Diseases and Caries. *J. Clin. Periodontol.* **2017**, *44*, S23–S38. [CrossRef] [PubMed]
5. Langfelder, P.; Horvath, S. WGCNA: An R Package for Weighted Correlation Network Analysis. *BMC Bioinform.* **2008**, *9*, 559. [CrossRef]
6. Wu, N.; Yin, F.; Ou-Yang, L.; Zhu, Z.; Xie, W. Joint Learning of Multiple Gene Networks from Single-Cell Gene Expression Data. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 2583–2595. [CrossRef]
7. Zhang, Z.; Zhang, X. Inference of High-Resolution Trajectories in Single-Cell RNA-Seq Data by Using RNA Velocity. *Cell Rep. Methods* **2021**, *1*, 100095. [CrossRef]
8. Gan, Y.; Liang, S.; Wei, Q.; Zou, G. Identification of Differential Gene Groups From Single-Cell Transcriptomes Using Network Entropy. *Front. Cell Dev. Biol.* **2020**, *8*, 588041. [CrossRef]
9. Ray, S.; Lall, S.; Bandyopadhyay, S. CODC: A Copula-Based Model to Identify Differential Coexpression. *NPJ Syst. Biol. Appl.* **2020**, *6*, 20. [CrossRef] [PubMed]
10. Cho, H.; Liu, C.; Preisser, J.S.; Wu, D. A bivariate zero-inflated negative binomial model and its applications to biomedical settings. *bioRxiv* **2020**. [CrossRef]
11. Qiu, P. Embracing the Dropouts in Single-Cell RNA-Seq Analysis. *Nat. Commun.* **2020**, *11*, 1169. [CrossRef]
12. Shi, J.; Malik, J. Normalized Cuts and Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 888–905. [CrossRef]
13. Wu, D.; Smyth, G.K. Camera: A Competitive Gene Set Test Accounting for Inter-Gene Correlation. *Nucleic Acids Res.* **2012**, *40*, e133. [CrossRef] [PubMed]
14. Van Buren, E.; Hu, M.; Cheng, L.; Wrobel, J.; Wilhelmsen, K.; Su, L.; Li, Y.; Wu, D. TWO-SIGMA-G: A New Competitive Gene Set Testing Framework for scRNA-Seq Data Accounting for Inter-Gene and Cell–Cell Correlation. *Briefings Bioinform.* **2022**, *23*, bbac084. [CrossRef] [PubMed]
15. Van Buren, E.; Hu, M.; Weng, C.; Jin, F.; Li, Y.; Wu, D.; Li, Y. TWO-SIGMA: A Novel Two-component Single Cell Model-based Association Method for Single-cell RNA-seq Data. *Genet. Epidemiol.* **2020**, *45*, 142–153. [CrossRef] [PubMed]
16. Divaris, K.; Slade, G.D.; Ferreira Zandona, A.G.; Preisser, J.S.; Ginnis, J.; Simancas-Pallares, M.A.; Agler, C.S.; Shrestha, P.; Karhade, D.S.; Ribeiro, A.d.A.; et al. Cohort Profile: ZOE 2.0—A Community-Based Genetic Epidemiologic Study of Early Childhood Oral Healt. *Int. J. Environ. Res. Public Health* **2020**, *17*, 8056. [CrossRef]
17. Heimisdottir, L.H.; Lin, B.M.; Cho, H.; Orlenko, A.; Ribeiro, A.A.; Simon-Soro, A.; Roach, J.; Shungin, D.; Ginnis, J.; Simancas-Pallares, M.A.; et al. Metabolomics Insights in Early Childhood Caries. *J. Dent. Res.* **2021**, *100*, 615–622. [CrossRef]
18. Ginnis, J.; Ferreira Zandoná, A.G.; Slade, G.D.; Cantrell, J.; Antonio, M.E.; Pahel, B.T.; Meyer, B.D.; Shrestha, P.; Simancas-Pallares, M.A.; Joshi, A.R.; et al. Measurement of Early Childhood Oral Health for Research Purposes: Dental Caries Experience and Developmental Defects of the Enamel in the Primary Dentition. *Methods Mol. Biol.* **2019**, *1922*, 511–523. [CrossRef]
19. Divaris, K.; Shungin, D.; Rodríguez-Cortés, A.; Basta, P.V.; Roach, J.; Cho, H.; Wu, D.; Ferreira Zandoná, A.G.; Ginnis, J.; Ramamoorthy, S.; et al. The Supragingival Biofilm in Early Childhood Caries: Clinical and Laboratory Protocols and Bioinformatics Pipelines Supporting Metagenomics, Metatranscriptomics, and Metabolomics Studies of the Oral Microbiome. *Methods Mol. Biol.* **2019**, *1922*, 525–548. [CrossRef]
20. Wood, D.E.; Lu, J.; Langmead, B. Improved Metagenomic analysis with Kraken 2. *Genome Biol.* **2019**, *20*, 257. [CrossRef]
21. Lu, J.; Breitwieser, F.P.; Thielen, P.; Salzberg, S.L. Bracken: Estimating Species Abundance in Metagenomics Data. *PeerJ Comput. Sci.* **2017**, *3*, e104. [CrossRef]
22. Dewhirst, F.E.; Chen, T.; Izard, J.; Paster, B.J.; Tann, A.C.R.; Yu, W.-H.; Lakshmanan, A.; Wade, W.G. The Human Oral Microbiome. *J. Bacteriol.* **2010**, *192*, 5002–5017. [CrossRef]
23. Cho, H.; Ren, Z.; Divaris, K.; Roach, J.; Lin, B.; Lin, C.; Azcarate-Peril, A.; Simancas-Pallares, M.; Shrestha, P.; Orlenko, A.; et al. Pathobiont-Mediated Spatial Structuring Enhances Biofilm Virulence in Childhood Oral Disease. *bioRxiv* **2022**. [CrossRef]
24. Franzosa, E.A.; McIver, L.J.; Rahnavard, G.; Thompson, L.R.; Schirmer, M.; Weingart, G.; Lipson, K.S.; Kn, R.; Caporaso, J.G.; Segata, N.; et al. Species-Level Functional Profiling of Metagenomes and Metatranscriptomes. *Nat. Methods* **2018**, *15*, 962–968. [CrossRef] [PubMed]
25. Franzosa, E.A.; Sirota-Madi, A.; Avila-Pacheco, J.; Fornelos, N.; Haiser, H.J.; Reinker, S.; Vatanen, T.; Hall, A.B.; Mallick, H.; McIver, L.J.; et al. Gut Microbiome Structure and Metabolic Activity in Inflammatory Bowel Disease. *Nat. Microbiol.* **2018**, *4*, 293–305. [CrossRef]
26. Cho, H.; Qu, Y.; Liu, C.; Tang, B.; Lyu, R.; Lin, B.M.; Roach, J.; Azcarate-Peril, M.A.; de Aguiar Ribeiro, A.; Love, M.I.; et al. Comprehensive Evaluation of Methods for Differential Expression Analysis of Metatranscriptomics Data. *bioRxiv* **2021**. [CrossRef]
27. Evans, A.M.; DeHaven, C.D.; Barrett, T.; Mitchell, M.; Milgram, E. Integrated, Nontargeted Ultrahigh Performance Liquid Chromatography/Electrospray Ionization Tandem Mass Spectrometry Platform for the Identification and Relative Quantification of the Small-Molecule Complement of Biological Systems. *Anal. Chem.* **2009**, *81*, 6656–6667. [CrossRef]
28. Evans, A.M.; Bridg, B.R.; Liu, Q.; Mitchell, M.W.; Robinson, R.J.; Dai, H.; Stewart, S.J.; DeHaven, C.D.; Miller, L.A.D. High Resolution Mass Spectrometry Improves Data Quantity and Quality as Compared to Unit Mass Resolution Mass Spectrometry in High-Throughput Profiling Metabolomics. *J. Postgenomics Drug Biomark. Dev.* **2014**, *4*. [CrossRef]
29. Xie, J.; Cho, H.; Lin, B.M.; Pillai, M.; Heimisdottir, L.H.; Bandyopadhyay, D.; Zou, F.; Roach, J.; Divaris, K.; Wu, D. Improved Metabolite Prediction Using Microbiome Data-Based Elastic Net Models. *Front. Cell. Infect. Microbiol.* **2021**, *11*, 734416. [CrossRef] [PubMed]

30. Berahmand, K.; Nasiri, E.; Pir Mohammadiani, R.; Li, Y. Spectral Clustering on Protein-Protein Interaction Networks via Constructing Affinity Matrix Using Attributed Graph Embedding. *Comput. Biol. Med.* **2021**, *138*, 104933. [CrossRef]
31. Meilă, M.; Pentney, W. Clustering by weighted cuts in directed graphs. In Proceedings of the 2007 SIAM International Conference on Data Mining, Minneapolis, MN, USA, 26–28 April 2007.
32. Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N.S.; Wang, J.T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **2003**, *13*, 2498–2504. [CrossRef] [PubMed]
33. Takahashi, N. Oral Microbiome Metabolism: From "Who Are They?" to "What Are They Doing?". *J. Dent. Res.* **2015**, *94*, 1628–1637. [CrossRef] [PubMed]
34. Mashima, I.; Nakazawa, F. Interaction between *Streptococcus* Spp. and *Veillonella tobetsuensis* in the Early Stages of Oral Biofilm Formation. *J. Bacteriol.* **2015**, *197*, 2104–2111. [CrossRef] [PubMed]
35. Takahashi, N.; Washio, J.; Mayanagi, G. Metabolomic Approach to Oral Microbiota. *Interface Oral Health Sci.* **2011**, *2012*, 334–340. [CrossRef]
36. Takahashi, N. Microbial Ecosystem in the Oral Cavity: Metabolic Diversity in an Ecological Niche and Its Relationship with Oral Diseases. *Int. Congr. Ser.* **2005**, *1284*, 103–112. [CrossRef]
37. Takahashi, N.; Washio, J.; Mayanagi, G. Metabolomic Approach to Oral Biofilm Characterization—A Future Direction of Biofilm Research. *J. Oral Biosci.* **2012**, *54*, 138–143. [CrossRef]
38. Sola-Penna, M. Metabolic Regulation by Lactate. *IUBMB Life* **2008**, *60*, 605–608. [CrossRef]
39. Larrabee, M.G. Lactate Metabolism and Its Effects on Glucose Metabolism in an Excised Neural Tissue. *J. Neurochem.* **2002**, *64*, 1734–1741. [CrossRef]
40. John, C.R.; Watson, D.; Barnes, M.R.; Pitzalis, C.; Lewis, M.J. Spectrum: Fast Density-Aware Spectral Clustering for Single and Multi-Omic Data. *Bioinformatics* **2020**, *36*, 1159–1166. [CrossRef]
41. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*; Erlbaum: Mahwah, NJ, USA, 1988.
42. Prost, V.; Gazut, S.; Brüls, T. A Zero Inflated Log-Normal Model for Inference of Sparse Microbial Association Networks. *PLoS Comput. Biol.* **2021**, *17*, e1009089. [CrossRef]