scientific reports

OPEN

Topic modeling for multi-omic integration in the human gut microbiome and implications for Autism

Christine Tataru[1]✉, Marie Peras[3], Erica Rutherford[3], Kaiti Dunlap[4], Xiaochen Yin[3], Brianna S. Chrisman[4], Todd Z. DeSantis[3], Dennis P. Wall[5,6], Shoko Iwai[3] & Maude M. David[1,2]✉

While healthy gut microbiomes are critical to human health, pertinent microbial processes remain largely undefined, partially due to differential bias among profiling techniques. By simultaneously integrating multiple profiling methods, multi-omic analysis can define generalizable microbial processes, and is especially useful in understanding complex conditions such as Autism. Challenges with integrating heterogeneous data produced by multiple profiling methods can be overcome using Latent Dirichlet Allocation (LDA), a promising natural language processing technique that identifies topics in heterogeneous documents. In this study, we apply LDA to multi-omic microbial data (16S rRNA amplicon, shotgun metagenomic, shotgun metatranscriptomic, and untargeted metabolomic profiling) from the stool of 81 children with and without Autism. We identify topics, or microbial processes, that summarize complex phenomena occurring within gut microbial communities. We then subset stool samples by topic distribution, and identify metabolites, specifically neurotransmitter precursors and fatty acid derivatives, that differ significantly between children with and without Autism. We identify clusters of topics, deemed "cross-omic topics", which we hypothesize are representative of generalizable microbial processes observable regardless of profiling method. Interpreting topics, we find each represents a particular diet, and we heuristically label each cross-omic topic as: healthy/general function, age-associated function, transcriptional regulation, and opportunistic pathogenesis.

Autism is a complex neurodevelopmental condition that occurs in 1 out of every 54 children in the United States. It is characterized by a specific set of behaviors, although many autistic people may exhibit only a subset of them[1]. Autism is frequently called "Autism Spectrum Disorder" in the scientific literature, however, surveys and public opinion from people in the community have demonstrated that the preferred term is simply "Autism"[2].

Autistic people have a statistically higher likelihood to experience gastrointestinal (GI) issues[3,4]. These GI issues can include symptoms such as chronic constipation, diarrhea, abdominal pain, and potential signs of GI inflammation such as vomiting and bloody stools[5]. There is also growing evidence to suggest a link between gut microbiome dysbiosis and Autism. Fecal microbiota transplants in children with Autism have demonstrated some improvements in GI symptoms and a decrease in Autism-associated behaviors[6,7]. Probiotic supplementation has also been observed to improve GI symptoms as well as multisensory processing and adaptive functioning in autistic preschoolers[8]. However, the specifics of this relationship are still little understood. Many studies using 16S sequencing to profile microbial communities have reported differences in the abundance of certain species and genera - often these reports are not reproducible in independent studies[9–11].

Molecular-level processes have also been implicated in Autism. Neurotransmitter biosynthesis and breakdown, specifically tryptophan and glutamine metabolism, are often observed to be dysregulated in Autism and potentially influenced by shifts in the gut microbiome[12–17] . Additionally, the chance of mitochondrial disease

[1]Department of Microbiology, Oregon State University, SW Campus Way, Corvallis, USA. [2]School of Pharmacy, Oregon State University, SW Campus Way, Corvallis, USA. [3]Second Genome Inc, 1000 Marina Blvd, Suite 500, Brisbane, CA 94005, USA. [4]Department of Bioengineering, Serra Mall, Stanford, USA. [5]Department of Biomedical Data Science, Serra Mall, Stanford, USA. [6]Department of Pediatrics (Systems Medicine), Stanford, 1265 Welch Road, Stanford, USA. ✉email: tataruc@oregonstate.edu; maude.david@oregonstate.edu

within the autistic population is about 5.0%, 500 times higher than that found in the general population (0.01%), and 30% of children with Autism may experience metabolic abnormalities[18]. The metabolisms of the resident gut bacteria is suggested to be involved in both the phenomena of oxidative stress and Autism, respectively[19]. However, our understanding of how gut microbiome environments relate to both is hindered by the heterogeneity of the categorization of Autism, high variation between and within human gut microbial ecosystems, and differences in conclusions based on different profiling techniques that may be used to measure microbiomes.

There are certain techniques in computer science, specifically in natural language processing, that excel at summarizing highly heterogeneous data. The topic modeling strategy used here, Latent Dirichlet Allocation (LDA), was originally used to identify topics within heterogeneous written text documents[20]. Based on the shared distribution of words across documents, LDA defines a pre-determined number of topics, each of which is a probability distribution across words. Documents are simultaneously defined by a probability distribution across topics. In this way, a document from a food blog may be 60% about restaurants, and 40% about travel. This mixed membership model set up allows models to capture complex phenomenon as would be expected in a microbial ecosystem.

In this study, we pursued two objectives. First, we sought to deepen our understanding of related variables within the human gut microbiome as represented by multiple omic profiling technologies. We define an "omic" as one sample by feature table as produced by one profiling method (i.e. 16S rRNA, shotgun metagenomics, shotgun metatranscriptomics, and untargeted metabolomics). To address this objective, we applied Latent Dirichlet Allocation, a mixed membership statistical method, defining microbial features (i.e. amplicon sequencing variants (ASVs), Kegg Orthologs (KOs), or compounds) as words, and samples as documents. This topic analysis defined specific sub-processes or topics that were represented in a robust manner across multiple gut microbiome profiling techniques. Second, we evaluated potential associations between features of the gut microbial ecosystem and Autism. To do this, we first used the identified microbiome topics to cluster samples with similar "microbial landscapes" together, then applied traditional differential abundance analysis techniques.

Our first objective was motivated by the fact that human gut microbiomes are highly heterogeneous within and between individuals on a taxonomic and functional level[21–24]. Clustering approaches are commonly used to define community structure and have led to the adoption of the concept of enterotypes, distinct sub-types of microbiomes that are dominated by either Bacteroides, Prevotella, or Ruminococus genera[25]. However, while this categorization serves as an important dimensionality reduction tool for these complex datasets, it tends to oversimplify the community structure, and hides complexity in the microbial communities that exist in the space between enterotypes[26]. Newer work argues that host-associated microbiomes should be considered to be on a spectrum, and that mixed membership models, and in particular topic models, are powerful and robust tools to learn and define that spectrum[27–32].

Our second objective, to identify associations between Autism and gut microbial features, was motivated by the high heterogeneity between individual microbiomes, and was enabled by topic modeling. Reports of how the gut microbiome relates to Autism are highly variable across the literature, and while many agree that the gut-brain axis plays a role, there is only moderate consensus about the specific bacteria and processes involved[9–11]. Differential abundance analysis between highly heterogeneous samples provides limited power and is more likely to produce spurious positive results. To address this, we use topics identified in multi-omic data to cluster samples, and perform differential abundance analysis on each cluster independently, making each group less heterogeneous in microbial structure.

Results

Topic modeling for multi-omic integration.

We used Latent Dirichlet Allocation (LDA), a form of unsupervised topic modeling, as an integrative approach to reduce the thousands of features across 16S rRNA (16S), shotgun metagenomic (MTG), shotgun metatranscriptomic (MTT), and untargeted metabolomic (MBX) profiling datasets all performed on the same stool samples from children with and without Autism. Please see Table 1 for a demographic description.

LDA was originally used for topic modeling in natural language processing where topics are defined by distributions across a vocabulary and documents are modeled as deriving from a distribution of topics. In this study case, we treated samples as documents and omic features as words, and we used the counts of features across samples to learn topics per omic (Fig. 1A). After model fitting, each topic is a weighted vector of feature count probabilities, and each sample is a weighted vector of topic probabilities. Each sample has 24 total associated vectors: 4 original omic count vectors, 4 16s topic vectors, 5 MTG topic vectors, 4 MTT topic vectors, and 7 MBX topic vectors (Fig. 1B). The number of topics were selected to minimize differences between true and model-simulated data (see Methods). Remarkably, topics learned from different omics correlated significantly with one another in true data, but not in data simulated using null models (see Methods), allowing us to define cross-omic topics, or sub-processes that are observable at multiple levels of measurements (Fig. 1C, Supplementary Fig. 2). Using the topics as representative of broad microbial processes, we split samples into two main groups based on their topic distribution (Fig. 1D). Samples in the same topic-derived clusters shared similar dietary patterns (Supplementary Fig. 1). Clustering samples using topics highlighted two distinct Autism-related metabolic profiles: sample cluster 1, which associated with healthier eating habits, implicated neurotransmitter precursors and sample cluster 2, which associated with less healthy eating habits, implicated fatty acids and their derivatives. Lastly, topic interpretation revealed four main processes reflected in all omics: healthy/general function, age-associated function, transcriptional regulation, and opportunistic pathogenesis. See Table 2 for cluster demographics.

	n_samples	n_ASD	n_TD
81 inclusive samples	81	42	39
Cluster 1	39	21	18
Cluster 2	42	21	21

Table 1. Demographics of 81 samples where all four omics were performed on the same stool sample, and of topic-driven sub-clusters. ASD stands for Autism Spectrum Disorder or simply "Autism", and TD stands for typically developing or "neurotypical".

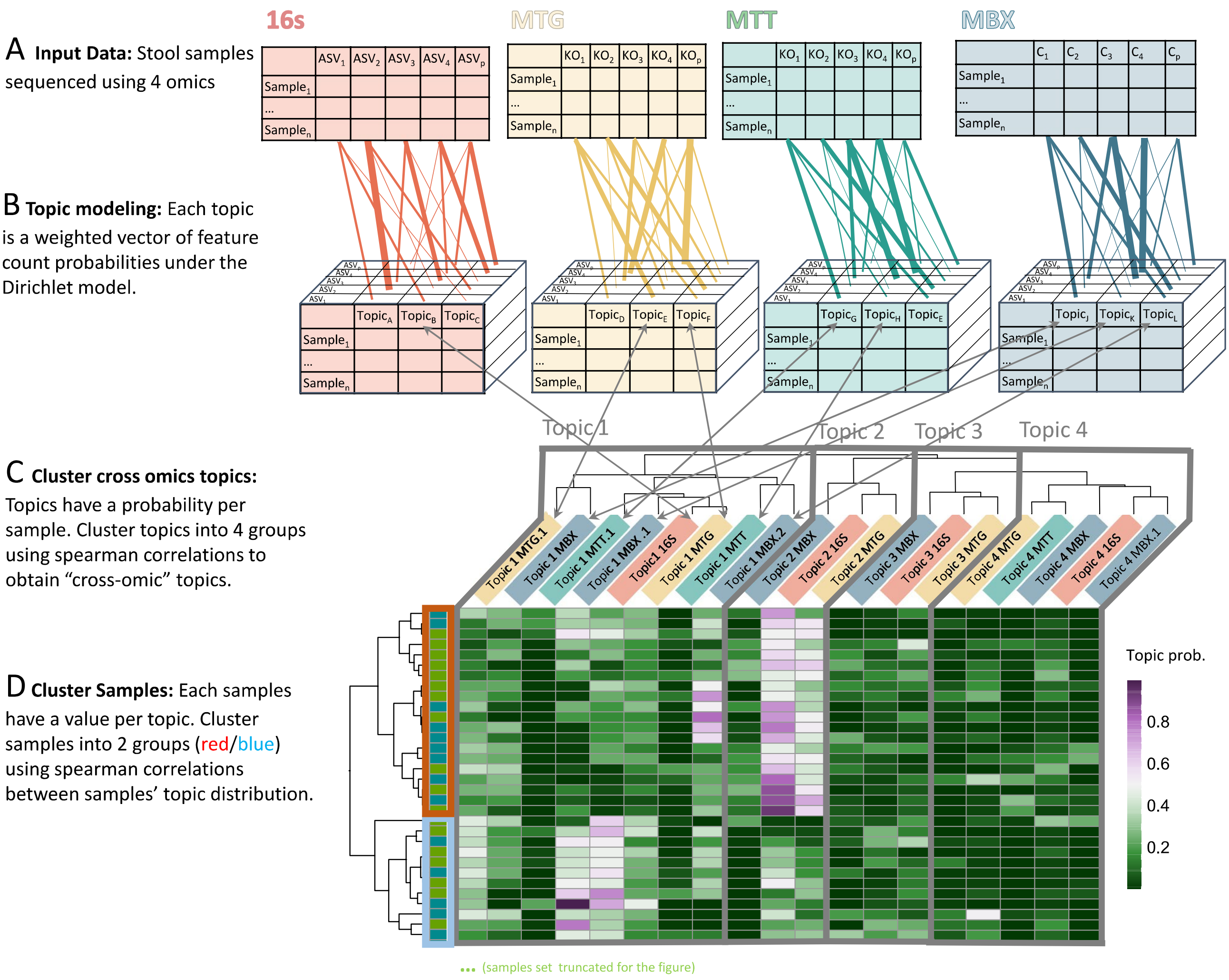
16s
MTG
MTT
MBX
A Input Data: Stool samples sequenced using 4 omics
ASV1 ASV2 ASV3 ASV4 ASVp
Sample1
...
Samplen
KO1 KO2 KO3 KO4 KOp
C1 C2 C3 C4 Cp
B Topic modeling: Each topic is a weighted vector of feature count probabilities under the Dirichlet model.
TopicA TopicB TopicC
TopicD TopicE TopicF
TopicG TopicH TopicE
TopicJ TopicK TopicL
Topic 1
Topic 2
Topic 3
Topic 4
C Cluster cross omics topics: Topics have a probability per sample. Cluster topics into 4 groups using spearman correlations to obtain "cross-omic" topics.
Topic 1 MTG.1
Topic 1 MBX
Topic 1 MTT.1
Topic 1 MBX .1
Topic1 16S
Topic 1 MTG
Topic 1 MTT
Topic 1 MBX.2
Topic 2 MBX
Topic 2 16S
Topic 2 MTG
Topic 3 MBX
Topic 3 16S
Topic 3 MTG
Topic 4 MTG
Topic 4 MTT
Topic 4 MBX
Topic 4 16S
Topic 4 MBX.1
D Cluster Samples: Each samples have a value per topic. Cluster samples into 2 groups (red/blue) using spearman correlations between samples' topic distribution.
Topic prob.
0.8
0.6
0.4
0.2
... (samples set truncated for the figure)

Figure 1. Topic modeling process.Stool samples were measured using 4 techniques, 16s, metagenomics (MTG), metatranscriptomics(MTT), and metabolomics (MBX) (**A**). Latent Dirichlet Allocation (LDA) was applied to each dataset independently to model latent variable "topics" from count data . Each sample is modeled as a Dirichlet distribution of topics, and each topic is modeled as a Dirichlet distribution of features (ASV, KO, or C) (**B**). Topics from each omic were then clustered by their distribution across samples using hierarchical clustering on spearman correlations (**C**). We call the cluster of topics "cross-omic topics", and discuss their interpretations and biological implications. Multiple topics from an omic may be included in the same cross-omic topic, denoted as ".1" or ".2" etc. Samples were clustered using hierarchical clustering on spearman correlations between their topic distributions (**D**). Groups enabled comparisons between samples with similar diets and similar 'microbial feature landscapes'.

Modeling identifies cross-omic topics that correspond to diet and microbial functions. We interpreted topics by observing which features have highest attribution weights for each topic, and also by correlating topic distributions across samples to dietary characteristics (Fig. 2). We find that dietary variables cluster into two main groups (Fig. 1E) with the consumption of fruits, vegetables, fermented vegetables, seafood, meat, home-prepared meals, probiotics, vitamin B, and vitamin D constituting one group, and consumption of sugary

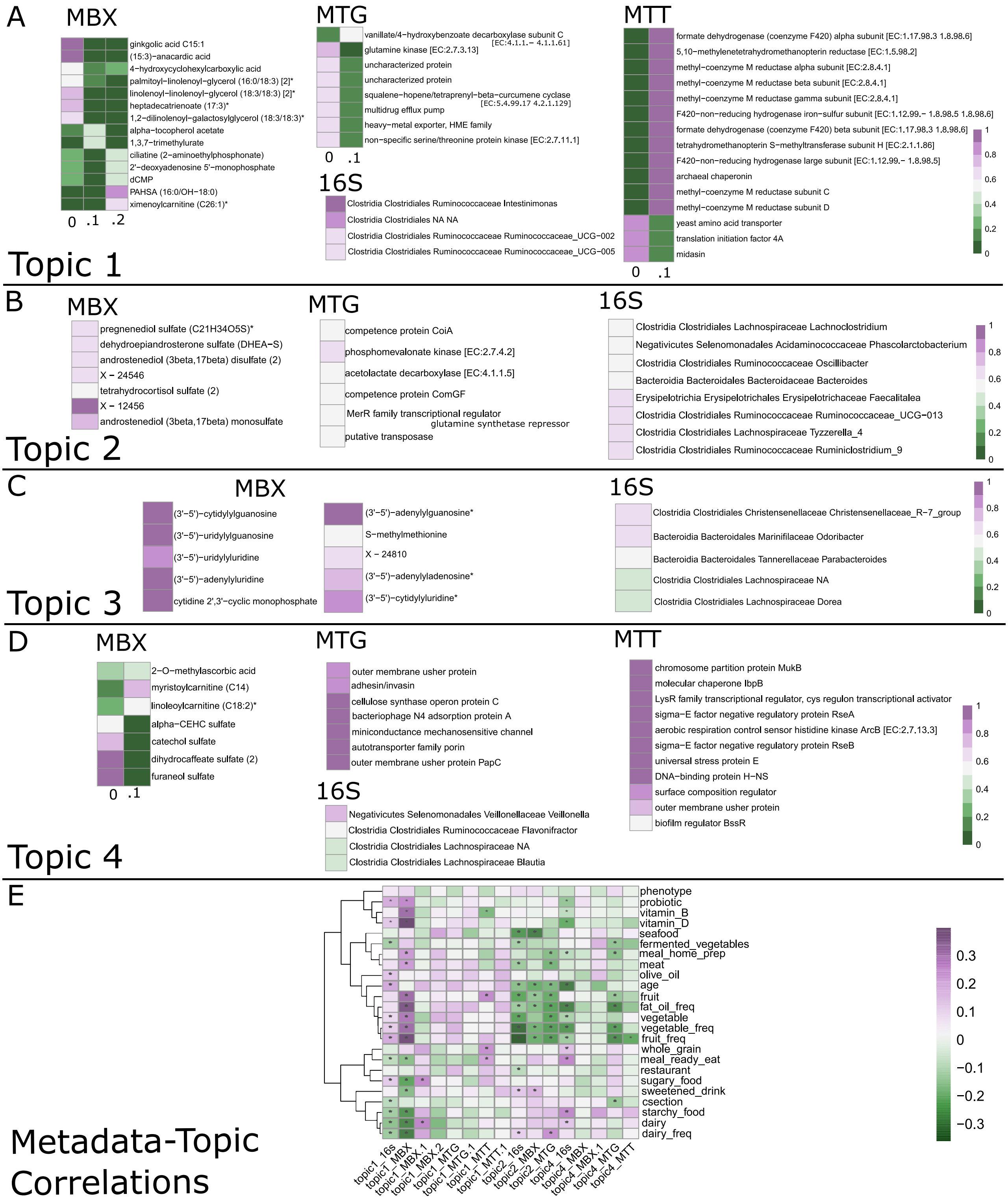

Figure 2. Topic interpretation using feature weights Each topic may be defined by those features with the highest weights. Features detailed here had weights with max value more than 1.5 standard deviations over the median (exceptionally high value) and high values along axes of maximal variation between topics (exceptionally attributable to a single topic). Topic interpretations are for cross-omic topic 1 (**A**), cross-omic topic 2 (**B**), cross-omic topic 3 (**C**) and cross-omic topic 4 (**D**). Each cross-omic topic may contain multiple single omic topics, as denoted by 0, .1 or .2. We consider any features with high weight to be representative of the cross-omic topic. Topic correlations with dietary features are in (**E**). Dietary variables with keyword "freq" represent long-term self-reported variables, while variables without "freq" are about the most recent week before sampling. Cells in A-D are colored by topic value (probability between 0 and 1), while cells in E are colored by spearman ρ value. A star signifies a significant (p<0.05) spearman correlation.

food, sweetened drinks, starchy food, dairy, whole-grains, pre-packaged meals, and restaurant food constituting another. It should be noted that dietary variables are not included into the topic modeling process itself to focus analysis on the true microbial landscapes and erase bias due to self-reported data and pre-conceived notions of healthy eating.

Topic1 is characterized by high consumption of the first, healthier, group of dietary variables (Fig. 1E), as well as with bacterial alpha diversity (Supplementary Fig. 4). There are three MBX topics that cluster into cross-omic Topic1. They are characterized by high values of fats, specifically 18:3 containing glycerols, anacardic and heptadecatrienoic acids, PAHSA, ximenoylcarnitine, as well as alpha-tocopherol acetate, trimethylurate, and AAMU (5-acetyleamino-6-amino-3-methyluracil) (Fig. 2). There are two MTG topics that cluster into cross-omic Topic1. They are characterized by ubiquitous enzymes like glutamine and serine/threonine kinases and heavy metal and drug exporter pumps. There are two MTT topics that cluster into cross-omic Topic1 as well. They are characterized by high values of methyl-coenzyme M subunits used in methane metabolisms, as well as functions like yeast plasma membrane ATPase, yeast amino acid transporter, eukaryotic translation initiation factors, pyruvate decarboxylase, heat shock protein, and MFS sugar transporters. There is one topic from 16s that clusters into cross-omic Topic1; it is high in values for species from Ruminococcacea, as well as other unidentified Clostridiales order members.

Additionally, we found Topic1 values to be correlated with specific behavioral characteristics from the parent-reported behavioral questionnaire (see Methods). In particular, Topic1 values correlated with observed imaginative play, both along and with others, as well as language and speech skills and a lack of self harm behavior (Supplementary Fig. 5). It should be noted that these behaviors and topics may be related to age, although significance of correlations between Topic1 and age vary (Supplementary Fig. 6).

Cross-omic Topic2 is characterized by high consumption of the second, less healthy, group of dietary variables (Fig. 2E). In MBX data, we see high values of steroid sulfates, specifically androstenediol mono and di sulfate, tetrahydrocortisol sulfate, pregnenediol sulfate, and DHEA-S (Fig. 2B). In MTG, we see proteins used by bacteria for horizontal gene transfer, namely competence proteins and transposases, as well as a glutamine synthetase repressor, acetolactate decarboxylase, and phosphomevalonate kinase. In 16s, we see high Ruminiclostridium species, Blautia and Tyzzerella and Faecalitalea genera members. Topic2 is not represented in MTT data. Topic2 is also the only cross-omic topic consistently correlated (negatively) with age (Supplementary Fig. 6).

Topic3 is weakly associated with the first, healthy, set of dietary variables (Fig. 2E). In MBX data, we see high values of dinucleoside monophosphates (Fig. 2C). In MTG data, we see higher values of sensor histidine kinase DegS and manganese/zinc/iron transport permease protein. In MTT data, Topic3 is not clearly defined by any set of features, but is dramatically the topic assigned the highest attribution across all samples (the chance of a sample having high Topic3 values is very high) (Supplementary Fig. 7). In 16s data, we see high values for two species, one from the family Christensenellaceae and the other from Odoribacter.

Topic4 is characterized less strongly by metadata variables, but still represents lower intake of vegetable, fruit, fat/oil, meat, and home prepared meals (Fig. 2E). In MBX data, we see high values of sulfates, specifically catechol, dihydrocaffeate, and furaneol sulfates (Fig. 2D). Additionally, myristoylcarnitine, linoleoylcarnitine, 7-ketolithocholate, 3-hydroxyisobutyrate, and 2-O-methylascorbic acid are all high in Topic4 MBX. In MTG, we see a strong contribution from genes associated with pathogenesis, specifically autotransporter family proteins, usher proteins, adhesin/invasin, and bacteriophage proteins. In MTT, we see similarly high values of universal stress protein E, outer membrane usher protein, biofilm regulator BssR, as well as regulatory proteins RseB, molecular chaperone IbpA, and chromosome partition protein MukB. Lastly, 16s data shows particularly high values for a Veillonella species, as well as for species from the Blautia, Faecalibacterium, Intestinibacter, and Anaerostipes genera.

Topics may be observed in independent datasets. To demonstrate that topics may be observed in other independently collected datasets, we utilized data from David et. al and Telleria et. al, which offered 16s and metabolomics data respectively with similar data collection and processing methods. We trained LDA models with the same parameters as used in this paper, calculated feature-feature cosine distance matrices based on the feature-topic distributions (betas) in each model. We compared the respective feature-feature matrices derived from the independent studies to those from this study using a mantel test, which reported a significant relationship between the 16s-based ($p = 0.001$) and metabolomics-based ($p = 0.001$) feature-feature relationship matrices across studies. Then, we permuted the rows of one feature-feature distance matrix to simulate mantel statistics under the null hypothesis, and observed the true mantel z-score was significantly higher than would be expected due to random chance (Supplementary Figure 8). This result implies that topics define consistent feature-feature relationships, in essence grouping features in a similar way, regardless of dataset.

Subsetting population by topics reveals metabolic differentials between Autism and Typically Developing cohorts. We defined two major clusters of samples using hierarchical clustering of topic distributions (Fig. 3A). We found that 92% of families (88 families) clustered into the same groups as each other, while only 8% (8 families) clustered separately (data not shown). Before clustering, autistic and typically developing participants had large differences in multiple dietary variables that significantly affected multi-omic features (Supplementary Fig. 3). After separating samples into subpopulations, we found that metadata differences between autistic and typically developing phenotypes became less stark (Supplementary Fig. 9). The exception to this statement was autistic individuals in Cluster2 still ate less dairy than their typically developing counterparts. This grouping therefore allowed us to account, to some extent, for differences in dietary patterns in our analysis.

We find samples in cluster1 have high cross-omic topic1 values, while samples in cluster2 have high cross-omic topic2 and topic4 values (Fig. 3B). Differential analysis between autistic and typically developing individuals

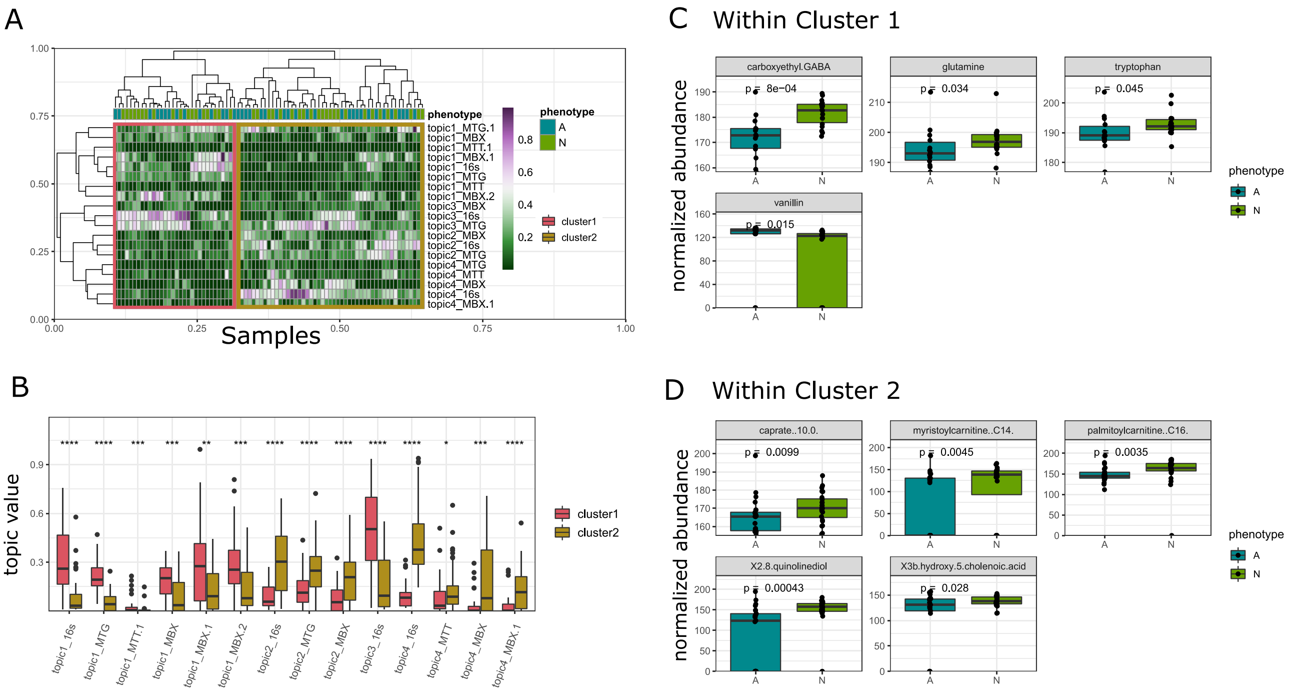

Figure 3. Differential analysis of metabolites between autistic and typically developing children within sample clusters as determined by topics. Hierarchical clustering on samples' topic distribution produces two clusters (**A**). Cluster1 is defined by high topics 1 and 3, while cluster 2 is defined by high topics 2 and 4 (**B**). In cluster 1, CEGABA, glutamine, tryptophan, and vanillin are differentially abundant (**C**). In cluster 2, caprate, myristoylcarnintine, palmitoylcarnitine, quinolinediol, and 3hydroxy5cholenoic acid are differentially abundant (**D**). Features are first filtered, and then *p* values are reported using Wilcox rank sum test. Multiple single-omic topics may belong to the same cross-omic topic, denoted here as ".1" or ".2".

	n_samples	n_ASD	n_TD	n_individuals
16s, used for topic modeling	456	228	228	152
MTG, used for topic modeling	193	98	95	186
MTT, used for topic modeling	178	89	89	178
MBX, used for topic modeling	175	87	88	175
81 all-inclusive samples, used for differential abundance testing	81	42	39	81
Cluster1	39	21	18	39
Cluster2	42	21	21	42

Table 2. Sample demographics for all samples used in topic modeling and differential abundance approaches. 81 samples had all four omics measured on the fecal samples. Remaining samples were collected as part of the same study, but had one or more omics missing.

within these subset populations reveals distinct metabolic signatures. In Cluster1, autistic individuals have significantly lower stool abundances of neurotransmitter precursors, namely caroxyethyl GABA, glutamine, and tryptophan, and higher abundance of vanillin (Fig. 3C). In Cluster2, autistic individuals have significantly lower stool abundances of fatty acids (caprate), acylcarnitines (myristoylcarnitine, palmitoylcarnitine), quinolinediol, and 3-hydroxy-5-cholenoic acid (Fig. 3D). Differential abundance plots for these metabolites across the entire cohort can be found in Supplementary Figure 10.

Additionally, we observed differentials in metagenome, metatranscriptome, and 16s features between autistic and typically developing individuals in each subgroup (Supplementary Fig. 11)).

Discussion

The gut microbiome has been implicated in a plethora of complex and idiopathic conditions, but thus far remains challenging to interpret due to the high variability and many avenues for analysis. We may observe microbiome structure and function using 16s rRNA amplicon sequencing, shotgun metagenomics, shotgun metatranscriptomics, and untargeted metabolomic profiling (among many others), and each of these data modalities reveals a different understanding of the composition and dynamics of the microbial community at large. While none of these methods singularly assures accurate conclusions, this paper simultaneously integrates all four omic data modalities to create a more complete and unbiased representation of microbiome structure and function. Our approach reveals which dietary habits are reflected across the microbiome to what extent, and presents an objective approach to subtyping a population based on microbiome composition and function.

We define 'cross-omic topics' that summarize patterns observed in 16s, metagenome, metatranscriptome, and metabolomic data, and their relationship with each other. We find that topics correlate with dietary variables, and these correlations can be used to group dietary habits in a way that reflects current dietary health standards. We associate features from each omic with those groups, and hypothesize that these feature groups may be used to obtain a more complete and data-driven understanding of overall health of the microbiome, especially in complex diseases associated with microbiome dysbiosis. We then clustered stool from autistic/typically developing participants using topic values, and found two distinct subpopulations. Metabolic differences between autistic and typically developing individuals within each subpopulation implicated neurotransmitter precursosr and medium to long chain fatty acid availability as relevant factors in Autism worthy of further exploration. We argue that this type of topic modeling approach could be used in other complex disease studies to integrate multi-omic data and more accurately elucidate the relevant factors in the human-microbiota relationship.

Value of LDA. Latent Dirichlet Allocation (LDA) has been used with great success to identify topics in natural language processing contexts. Given a set of documents, LDA uses word count distributions to define a set of topics that each document discusses. For example, documents may primarily discuss politics, cooking, or sports. Gibbs sampling is used to learn unlabeled topics by a weighted vector of vocabulary words. Simultaneously, documents are assigned weighted vectors of topic distributions. The resulting topic vectors as defined by vocabulary are unlabeled, however, by interpreting those words with high weight in a given topic, one may assign a label to the topic heuristically. For example, one topic may contain high weights for president, prime minister, counselor, and court room, while another may contain high weight for the words spatula, bruschetta, and baguette. A human interpreter could see these patterns and call the first topic "politics" and the second topic "cooking".

The Dirichlet distribution is commonly used to model microbiome data[33–35], and LDA in particular has proven to be a powerful and robust tool for modeling 16S data in various environments[26,28–32]. LDA is a parametric form of dimensionality reduction; it has the advantage of additional power as compared to other non-parametric forms like PCoA, GloVe, or transformer neural networks[36]. In the case of multi-omic analysis, where data representing multiple omic measurements on the same samples is very limited, parametric methods are particularly useful. LDA allows for mixed membership, so a sample can be represented by multiple sub-communities of bacteria, genes, or metabolites, making it more appropriate for capturing complex phenomena.

Value of cross-omic microbiome topics. Cross-omic topics represent gut microbiome landscapes that can be viewed through the lens of individual omics, but that are present regardless of the lens. We are not the first to consider microbial landscapes as a basis for comparing samples.

Perhaps the most commonly cited example of considering microbial landscapes is that of enterotypes. In a 2011 paper, Arumugam et. al introduced the idea that the human gut can come in three varieties, those dominated by Bacteroides, Prevotella, and Ruminococcus genera respectively. Later work by Knights et. al suggested that the enterotype distribution is actually continuous rather than discrete[26]. Most recently, Symul et. al consider how we might use approaches like topic modeling to identify sub-communities of samples within this continuous microbiome space in the vaginal ecosystem[27]. To our knowledge, the approach of topic modeling to define microbial sub-communities has been reported only for 16S data, with other omic data as supporting evidence. Here, we illustrate that the same sub-community or sub-process structures can be observed through the lens of multiple levels of biology (e.g. community membership, genomic function, and metabolic result), and that the sub-processes (topics) identified by each omic do correspond significantly with one another to define sub-processes across omics.

By comparing only samples with similar community structure, and only then identifying the specific factors by which they differ, our results become more meaningful. For example, in macroecology, it is standard practice to compare population sizes between species in similar environments only, as comparisons between vastly different habitats may be correct, but not meaningful. In microbial ecology, we often perform differential abundance analysis to identify features that differ between host phenotypes without consideration for the broader microbial landscape. We argue that it is not enough that samples come from the same tissue (e.g. gut), we must also consider the ecology between individuals as a relevant factor. One approach to accomplish this grouping of similar samples is to cluster samples based on similarities of diet and/or host genotype, however, the effect these factors have on the microbial environment is not fully understood, and does not account for all differences observed in broad microbial landscapes. Comparisons made on the basis of dietary similarities are also usually subjective to self-report bias. Instead, methodologies that act on the microbial landscapes themselves promise to be more consistent across studies and inherently mitigate bias derived from an incomplete understanding of microbiome-diet relationships.

Additionally, topics may represent underlying metabolic processes that are independently informative of microbial ecology principles. In the "Interpreting Topics by Feature Weights" section, we observe that biologically related features (ex. sets of dinucleotides or genes utilized by bacteria in stressful environments) are all attributed to the same topic. This consistency indicates that biological relationships may be captured in this methodology, and suggests that topic modeling may be used to broaden our understanding about gene, metabolite, and taxonomic species relationships. We found that topics learned in each omic dataset independently significantly correlated with other topics from other omics, which was not true of null simulated data. We hypothesize that this implies some universality of the processes represented by cross-omic topics, however, further research is needed to validate these topic distributions on independent datasets.

One benefit of identifying sub-processes based on multiple omics is mitigation of bias. Each omic technique provides useful information about one level of biological processing, but each is inherently incomplete and subject to unique sources of bias during processing. 16S data acts as a census of the bacterial species represented in the community, and their relative abundances. Because it involves an amplification step, it represents a deeper sampling of the microbial diversity present, but is also subject to amplification bias[37]. It also has very limited capacity to represent function or metabolism within the bacterial community[37]. Shotgun metagenomic (MTG) data provides a shotgun representation of the community gene pool. Because it lacks an amplification step, it represents only the most common genes. It also represents only genetic potential, as opposed to those genes undergoing expression or translation[38]. Shotgun metatranscriptomic (MTT) data provides a shotgun representation of the community translation pool. Like MTG, it represents only the most common genes, and can be additionally highly variable between timepoints[39]. Metabolomic (MBX) data provides a measurement of the metabolites and relative concentrations present in an environment. Metabolites may originate from bacteria or host, and data is additionally limited to those most common metabolites[40]. Individually, all are subject to unavoidable bias, however, integrative multi-omic approaches such as topic modeling have the potential to mitigate these biases - phenomena that are represented in all data modalities are more likely to be ubiquitous, universal, and consistent phenomena.

Dietary habits as related to cross-omic microbiome topics. It is commonly accepted that diet influences microbiome composition and function and vice versa, and it is not known explicitly what dietary habits promote the healthiest microbiomes. We found that correlations between dietary variables and cross-omic topics, which represent the relationship strength between specific dietary practices and microbiome composition and function, spontaneously group dietary choices into two groups. Group1 consists of fruit, vegetable, fermented vegetables, meat, seafood, home prepared meals, and vitamins, while group2 consists of sugary and starchy foods, sweetened drinks, dairy, and ready to eat meals. Interestingly, self-assembled groups clearly fall in line with current pre-determined dietary health standards[30].

In this case, it is impossible to distinguish between effects on the microbiome driven by lack of an important "healthy" food group versus presence of an "unhealthy" food group. Further studies may be designed to include participants actively eating all the necessary food groups in addition to some unhealthy options, as well as participants lacking some healthy food group choices to understand whether presence or absence of healthy eating choices is more influential on the gut microbiome.

Cross-omic microbiome topics as defined by microbial functions. *Topic1: healthy/general phenomena.* We hypothesize that Topic1 may represent base metabolism functionalities across many forms of life, including beta-oxidation (metabolism of fats), cytochrome functionality, methane metabolism, and ion transport. This is supported by a strong correlation between Topic1 values and 16s-based alpha diversity metrics, which are strongly linked to increased resilience and general healthy status[41–45]. Topic1 correlates most strongly with dietary variables such as fruit, vegetable, fat/oil, home prepared meal, meat, seafood, and vitamin consumption frequency, which are largely in line with the Healthy Eating Index "adequacy components" (eating more corresponds to a healthier diet)[46].

Within the MBX topics, there are 3 that belong to cross-omic Topic1. The first is represented by high values of 18 carbon, 3 times unsaturated lipids such as anacardic acid, heptadecatrienoate, and 18:3 glycerols. The second is represented by higher values of metabolites related to CYP1A2 metabolism of caffeine such as 1,3,7 trimethylurate, AAMU (CYP1A2 caffeine metabolism products), and alpha tocopherol acetate (CYP1A2 inhibitor)[47–49]. We hypothesize that this topic may represent the products of CYP1A2 activity of the host, which may be modified with caffeine consumption (in the form of coffee, tea, or chocolate), as well as by cruciferous and apiaceous vegetable consumption[50,51]. The last is represented by high values of PAHSA and ximenoylcarnitine, long chain and very long chain fatty acids that relate to mitochondrial function. Specifically, PAHSAs are known for their anti-inflammatory and insulin controlling action, while ximenoylcarnitine is less studied, but as an acylcarnitine likely has to do with beta-oxidation activity in the mitochondria[52,53]. There are two MTG topics that cluster into cross-omic Topic1. They are characterized by ubiquitous functions like glutamine and serine/threonine kinases and by environmental adaptation genes like heavy metal and drug exporter pumps[54,55]. There are two MTT topics that cluster into cross-omic Topic1 as well. They are characterized by high values of methyl-coenzyme M subunits used in methane metabolisms, as well as eukaryote-specific genes such as yeast plasma membrane ATPase, and yeast amino acid transporter. This same topic is also high in ubiquitous functions like eukaryotic translation initiation factors, pyruvate decarboxylase, heat shock protein, and MFS sugar transporters[56]. While the methane metabolism transcripts most likely represent archaea activity in the gut[57], the second MTT Topic1 more likely relates to eukaryotic metabolisms, specifically yeast. Contributing to this conclusion are the eukaryotic specific transcription factor and yeast specific genes that define this topic, as well as the fact that all transcripts mapping to the human genome were removed (see Methods), eliminating the main source of eukaryotic gene transcripts. Lastly, we see some Ruminococcaceae species along with unidentified Clostridia members highly represented, that, when combined with the high vegetable associations and hypothesized methane metabolisms, imply prevalent fermentation processes[58].

Topic2: age-associated phenomena. In contrast, Topic2, which is also significantly inversely correlated with age, correlates with low quantities of the above food groups, and high quantities of starchy, sugary, pre-packaged and restaurant foods, which are largely in line with the Healthy Eating Index "moderation" components (eating less corresponds to a healthier diet)[46]. In MBX Topic2, we see high values of steroid sulfates, specifically androstenediol mono and di sulfate, tetrahydrocortisol sulfate, and pregnenediol sulfate, DHEA sulfate, and alpha -CEHC sulfate. These steroid compounds are responsible for physiological changes seen during puberty, but are inactive in their sulfonated forms[59]. Thus, we hypothesize that Topic2 may capture the effect of age with an inverse relationship. While steroid sulfates are not considered actively hormonally, some do act as neurosteroids, which can modulate GABA and NMDA receptors among other targets[60]. Steroid sulfatase is a potential modifier of cognition in attention deficit hyperactivity disorder[61]. It is suggested that STS dysfunction (too many sulfates on hormones) predisposes an inattentive subtype of ADHD[62]. We did not find Topic2 to be associated with Autism severity, and it was not able to distinguish between autistic and TD individuals (data not shown).

In Topic2 MTG, we observed high values of a key enzyme in steroid hormone synthesis, phosphomevalonate kinase, which is consistent with the steroid metabolites observed in the MBX data[63]. We also observe high values for genes used in bacterial horizontal gene transfer, namely competence proteins and transposases. Upregulation of these gene sets has been observed from bacterial in stressful environments, and diets heavy in starch and sugars can affect biofilm formation as well as competence[64,65,65,66,66,67]. Topic2 also represents high values Ruminiclostridium species, Blautia and Tyzzerella and Faecalitalea genera members, many of which are associated with chronic inflammation. Ruminiclostridium is considered a potential pathogen associated with obesity and inflammation, and has been found to be increased in younger individuals in macaques, dairy cows, and humans[68–70]. Tyzzerella was found to be profoundly overrepresented in Crohn's disease patients, increased in patients with high cardiovascular disease risk profiles, and correlated with lower healthy eating scores[71–73]. It should be noted that while Topic2 was strongly inversely related to age, participants were no younger than 2 years old and were not being breast-fed, and so Topic2 is not representative of a newborn gut microbiome. The precise relationship between the observed competence proteins, pro-inflammatory genera, and age remains to be elucidated.

Topic3: transcriptional regulation phenomena. In Topic3, we see high values of dinucleoside monophosphates and genes for RNA polymerase subunits, elongation factor G, and HSP20. Dinucleoside monophosphates have been recently suggested to function as RNA caps in bacteria, to either initiate transcription or protect against RNA degradation[71,74]. Alternatively, they may also be an artifact of metabolomics pipeline processing. In 16s data, we see high values for two species, one from the family Christensenellaceae and the other from Odoribacter. MTG and MTT data do not present strong feature attributions for Topic3, however, we do see that the overall topic weight across samples attributed to Topic3 is incredibly high (Supplementary Fig. 7), supporting the interpretation of Topic3 as transcription related factors.

Topic4: opportunistic pathogenesis phenomena. Topic4 is characterized less strongly by metadata variables, but still represents lower intake of vegetable, fruit, fat/oil, meat, and home prepared meals. In MBX data, we see high values of sulfates, specifically catechol, dihydrocaffeate, and furaneol sulfates. Catechol is used as a pesticide and as a precursor to many chemical products, can be manufactured by multiple bacterial species including those found in the human gut[75], and has been found to have both anti-bacterial and anti-fungal action[76]. Dihydrocaffeic acid is produced during colonic fermentation of wheat[77]. Interestingly, although children ages 2–7 are unlikely to be consuming coffee, all three compounds are known to increase upon coffee consumption[78,79]. Additionally, metabolites that reflect beta-oxidation processes in the host are increased in MBX data, specifically, myristoylcarnitine, linoleoylcarnitine, 7-ketolithocholate, 3-hydroxyisobutyrate, and 2-O-methylascorbic acid. Acylcarnitines along with 3-hydroxyisobutyrate are intermediates of beta oxidation[80,81], while 7-ketolithocholate and other bile acids contribute to fat absorption and may influence the availability of fatty acids for catabolism[82]. 2-O-methylascorbic acid is understudied in the literature, however, 2-O-ethyascorbic acid, or Vitamin C, is a cofactor for carnitine biosynthesis which is necessary for beta-oxidation[83]. Interestingly, bacteria may also use carnitine as an osmoprotectant to increase thermotolerance, cryotolerance and barotolerance, and may use 3-HB as a substrate for the synthesis of polyhydroxybutyrate, which is a reserve material[81,84]. High concentrations of these metabolites contribute to the conclusion that Topic4 may represent a stressful or challenging environment for many microorganisms, and may increase the opportunity for opportunistic pathogen growth. In MTG, we see a strong contribution from genes associated with pathogenesis and metabolism in stressful environments. Autotransporter family proteins are often associated with virulence functions such as adhesion, aggregation, invasion, biofilm formation and toxicity, usher proteins facilitate pillus formation, and adhesin/invasin can induce bacterial aggregation and biofilm formation[85–87]. Miniconductance mechanosensitive channel confers protection against mild hypoosmotic shock[88] and bacteriophage proteins signal a stressful environment for commensals. In MTT, we see similarly high values of universal stress protein E, outer membrane usher protein, biofilm regulator BssR, as well as regulatory proteins RseB, molecular chaperone IbpA, and chromosome partition protein MukB. Many of these genes are upregulated during microbial adaptation to stressful environments, and by opportunistic pathogens in particular[89–91]. Lastly, 16s data shows particularly high values for a Veillonella species, as well as for species from the Blautia, Faecalibacterium, Intestinibacter, and Anaerostipes genera. Veillonella species are the strongest contributors to Topic4; they are well known for their behavior as opportunistic pathogens in the human gut and dental plaque and exhibit strong adhesion and biofilm formation capacities mediated by autotransporters[92].

Subsetting population by topics reveals metabolic differentials between ASD and TD cohorts-neurotransmitters and fatty acids. In cluster1 (high topics 1 and 3), which corresponds to largely healthy eating habits including high quantities of vegetables, fruits, and home prepared meals, we found autistic individuals to have lower stool abundances of neurotransmitter precursors carboxyethyl GABA, glutamine, and tryptophan. Autism is characterized by complex neurobehavioral and neurodevelopmental criteria including social interaction, restricted and repetitive behavior, and altered sensory processing[1,93]. Many have reported dysregulation in the glutamate-glutamine cycle in both plasma and brain tissues resulting in altered concentrations of glutamine and GABA in people with Autism[12–15]. Likewise, multiple components of the tryptophan metabolism pathway are neuroactive, including serotonin, kynurenine, and quinolinic acid[94,95]. Dysregulation in tryptophan metabolism is hypothesized as a major contributing agent to the gut-brain axis, and is associated with Autism[96]. Tryptophan may also become NAD+ through the kynurenine pathway, where dysregulation may imply issues with metabolic regulation[97]. While CEGABA is less studied, evidence suggests that it is active in the central nervous system and may strengthen cortical inhibition and act directly on the lower brain stem[96,98]. It also demonstrated anti-convulsant activity in guinea pigs[99].

In cluster2 (high topics 2 and 4), which corresponds to largely unhealthy eating habits including high quantities of sugary and processed food and drinks and low quantities of fruits and vegetables, we found autistic individuals to have lower stool abundances of fatty acids, acylcarnitines, quinolinediol, and 3-hydroxy-5-cholenoic acid. Mitochondrial dysfunction, as well as fatty acid and acylcarnitine concentrations in serum, has been implicated in subgroups of autistic individuals[100–102]. Concentrations of bile acids like 3-hydroxy-5-cholenoic acid have also been implicated in Autism, perhaps in part because bile acid concentrations directly influence the absorption and therefore availability of fatty acids[103,104]. We were unable to find reference to quinolinediol in ASD literature, however, quinolinic acid, a related compound, was found to be decreased in cerebrospinal fluid of 12 children with autism[105]. Quinolinic acid also serves as a precursor to NAD+, and this pathway may be used as an alternative to synthesize NAD+ in the context of oxidative stress as would be seen during mitochondrial dysfunction[106]. Mitochondrial dysfunction and oxidative stress has been reported in only 5% of autistic individuals, however, this is much higher than the expected 0.01% in the rest of the population[102,106].

Both neurotransmitter imbalances and metabolic dysregulation have been implicated in Autism literature, depending on the study. This study implicates both factors depending on the subtype of the gut microbiome, which itself relates to both diet, microbial gene content and expression, and metabolic environment. This subtyping, along with correlations between some topics and specific behavioral metrics, lead us to suggest that the gut microbiome may be effective at identifying subgroups of Autism, and may give some indication as to the metabolic differences seen between those subgroups of autistic and typically developing children.

Methods

Cohort and Metadata.

Stool samples from 196 children between ages 2–7 from all over the United States were collected via a crowdsourced initiative. Children were all from families with two siblings, one previously diagnosed with Autism by a health care provider and one typically developing. Dietary, lifestyle, demographic, and host health information were reported by parents of guardians via a questionnaire for each child (see[107]). Parents and guardians also filled out a Mobile Autism Risk Assessment (MARA) to document autism-associated behaviors in their children with the Autism diagnosis[108]. Samples were then subjected to multi-omic sequencing. In total, 81 of the stool samples received all four omic measurements, and these data were used in defining cross-omic topics and in differential abundance testing. However, some individuals provided multiple stool samples over time, and in topic modeling on each omic independently, all available data was used: 16S (456 stool samples across 152 individuals), MTG (193 stool samples across 186 individuals), MTT (178 stool samples across 178 invividuals), and MBX (175 stool samples across 175 individuals (Table 1).

Stool collection and storage stool.

Samples were collected by the parents using a preservative buffer (Norgen Biotek, ON, Canada). We collected two samples per child: one sample was preserved at room temperature in a preservative buffer, and the second one was collected in a tube without a preservative buffer but immediately frozen at home at −20. This frozen sample was shipped back overnight with two ice packs provided to the participants, while the samples in preservative were shipped at room temperature. Once received, stool samples were stored at −80 °C until processing.

Profiling techniques.

The following profiling and data processing techniques are originally described in West et al.[109]. All methods were performed in accordance with the relevant guidelines and regulations.

16S sequencing. DNA was extracted using the Qiagen MagAttract PowerMicrobiome DNA/RNA Kit according to manufacturer's guidelines. DNA was quantified using the Qubit® Quant-iT dsDNA High Sensitivity Kit (Invitrogen, Life Technologies, Grand Island, NY). The V4 regions of the 16S rRNA gene were amplified using primers as described in[110], and PCR products were quantified by fluorometric method (Qubit or PicoGreen from Invitrogen, Life Technologies, Grand Island, NY). Equimolar amounts of amplicons were mixed and sequenced using Illumina MiSeq for 250 cycles at Second Genome. All samples from the same families were sequenced in the same batch. Samples with fewer than 20,000 reads were re-sequenced up to three times.

Shotgun metagenomic (MTG) and metatranscriptomic (MTT) sequencing. The MTG library was constructed with the same DNA extracts as 16S-V4. For MTT, DNA extraction and digestion was performed using Qiagen MagAttract PowerMicrobiome RNA Kit and Invitrogen TURBO DNA free kit according to manufacturer's guidelines. rRNA depletion was also performed using Ribo-Zero Gold rRNA Removal Kit (Epidemiology). All samples were prepared for sequencing with the Illumina NexteraXT kit and quantified with Quant-iT dsDNA High Sensitivity assays. Libraries were pooled and run with 150 bp paired-end sequencing protocols on the Illumina NextSeq 550 platform.

Metabolomics. Untargeted metabolomics was performed on the preservative-free stool samples by Metabolon, Inc. Samples were extracted with methanol to precipitate protein and dissociate small molecules bound to protein or trapped in the precipitated protein matrix, followed by centrifugation to recover chemically diverse metabolites. The resulting extract was divided into five fractions: two for analysis by two separate reverse phase (RP)/UPLC-MS/MS methods using positive ion mode electrospray ionization (ESI), one for analysis by RP/UPLC-MS/MS using negative ion mode ESI, one for analysis by HILIC/UPLC-MS/MS using negative ion mode ESI, and one reserved for backup. Ultrahigh Performance Liquid Chromatography-Tandem Mass Spectroscopy (UPLC-MS/MS) was performed. All methods utilized a Waters ACQUITY ultra-performance liquid chromatography (UPLC) and a Thermo Scientific Q-Exactive high resolution/accurate mass spectrometer interfaced with a heated electrospray ionization (HESI-II) source and Orbitrap mass analyzer operated at 35,000 mass resolution. The sample extract was dried then reconstituted in solvents compatible with each of the four methods. Each reconstitution solvent contained a series of standards at fixed concentrations to ensure injection and chromatographic consistency. One aliquot was analyzed using acidic positive ion conditions, chromatographically optimized for more hydrophilic compounds. In this method, the extract was gradient-eluted from a C18 column (Waters UPLC BEH C18-2.1 × 100 mm, 1.7 µm) using water and methanol, containing 0.05% perfluoropentanoic acid (PFPA) and 0.1% formic acid (FA). A second aliquot was also analyzed using acidic positive ion conditions, but was chromatographically optimized for more hydrophobic compounds. In this method, the extract ias gradient eluted from the aforementioned C18 column using methanol, acetonitrile, water, 0.05% PFPA and 0.01% FA, and was operated at an overall higher organic content. A third aliquot was analyzed using basic negative ion optimized conditions using a separate dedicated C18 column. The basic extracts were gradient-eluted from the column using methanol and water, however with 6.5 mM Ammonium Bicarbonate at pH 8. The fourth aliquot was analyzed via negative ionization following elution from a HILIC column (Waters UPLC BEH Amide 2.1 × 150 mm, 1.7 µm) using a gradient consisting of water and acetonitrile with 10 mM Ammonium Formate, pH 10.8. The MS analysis alternated between MS and data-dependent MSn scans using dynamic exclusion. The scan range varied slightly between methods, but covered approximately 70–1000 m/z.

Data processing. *16S.* DADA2 (Callahan et al. 2016) was used to generate Amplicon Sequence Variants (ASVs). Raw sequence reads were processed applying default settings for filtering, learning errors, dereplication, ASV inference, and chimera removal. Truncation quality was set to 2. Ten nucleotides were then trimmed from each terminus of each read, both forward and reverse. ASVs were mapped to an in-house strain database: StrainSelect (StrainSelect, http://strainselect.secondgenome.com/, version 2019) using USEARCH (Edgar 2018). StrainSelect is a repository of monikers for known isolated microbial strains, their various synonyms and their genomic sequence identifiers. A DNA sequence observed from clinical material was assigned a strain-level annotation only when it uniquely matched one and only one strain. 6150 ASVs were generated from this process.

MTG and MTT. Reads were trimmed for adapter sequences and low-quality ends with Trimmomatic (Bolger, Lohse, and Usadel 2014), then contaminant sequences were removed with Bowtie2 (Langmead and Salzberg 2012). Host sequences were identified and removed with Kraken (Wood and Salzberg 2014). For MTT data, an additional step for rRNA removal was performed using SortMeRNA 2.0 (Kopylova, Noé, and Touzet 2012) prior to the host sequence removal. Filtered DNA sequences from MTG and MTT were mapped against a reference database of proteins within the KEGG (May 2019 release) and hits that spanned more than 20 amino acids with more than 80% similarity were collected. A total of 10,543 KOs were observed in MTG data, and 10,625 KOs were detected in MTT data.

MBX. Raw data were extracted, peak-identified, and quality control (QC) processed using Metabolon's hardware and software. Compounds were identified by comparison to library entries of purified standards or recurrent unknown entities. Peaks were quantified as area-under-the-curve detector ion counts. For studies spanning multiple days, a data adjustment step was performed to correct block variation resulting from instrument interday tuning differences, while preserving intraday variance. 1267 metabolites were identified, and 1025 annotated metabolites were used in the downstream analysis.

Data analysis. *Normalization.* To select a normalization method, we sought to minimize the within sibling to between sibling distance ratio using manhattan, euclidean, canberra, clark, bray, kulczynski, jaccard, and altGower distances from the vegan package. We found that DeSeq2 normalization minimized the ratio for all of the distance metrics in 16S data, and the highly related Relative Log Expression (RLE) normalization minimized the ratio for MBX data especially. Differences in the sibling ratio for MTG and MTT were slight, and so RLE was selected for these data as well for the sake of consistency (Supplementary Fig. 12). The R packages DeSeq2 and edgeR were used to normalize data respectively.

The high counts of the MTG, MTT, and MBX datasets made the Gibbs samples strategy for topic model training too computationally expensive. Additionally, in order to mitigate issues associated with heteroskedasticity, we log transformed the counts of each dataset after normalization and before model training.

To validate that the count-based approach of LDA would be appropriate to apply to the continuous MBX dataset, we calculated a distribution of all feature values across the population per omic dataset. We found that after the above transformations, features values possessed similar distributions across omic datasets, with 16s/MBX and MTG/MTT matching the most in general shape 13.

Topic modeling. We performed Latent Dirichlet Allocation (LDA) as a form of dimensionality reduction and multi-omic integration. LDA is a generative statistical model traditionally used in natural language processing that models documents as deriving from a set of topics, and topics as deriving from a set of words. In our case, we treat samples as documents and microbial features as words. As such, each sample is defined by a probability vector over K possible topics, and each topic is defined by a probability vector over V possible microbial features (e.g. ASVs, KOs, metabolites).

Under LDA, to generate each word in a document (w_{d_n}), first get that document's (d) topic probability vector (θ_d) and select a topic (z_{d_n}) by drawing from a multinomial. Then, get that topic's word probability vector ($\beta_{z_{d_n}}$), and select a word (w_{d_n}) using a multinomial. Each of the topic probability vectors is modeled as a Dirichlet distribution over hyperparameter α, and each of the word probability vectors is modeled as a Dirichlet distribution over a different hyperparameter γ. After model fitting, the Θ and β matrices can be accessed to obtain the topic distribution per document and the word distribution per topic. In this case, we access these matrices to obtain the topic distribution per sample and the microbial feature distribution per topic (Fig. 1D). Models were trained using the topicmodels package in R.

$$w_{d_n}|(\beta_k)_{k=1}^{K}, z_{d_n} \overset{iid}{\sim} Mult\left(1, \beta_{z_{d_n}}\right) \text{ for } d = 1, ...D \; and \; n = 1, ..., N_d$$

$$z_{d_n}|\theta_d \overset{iid}{\sim} Mult(1, \theta_d) \text{ for } d = 1, ...D \text{ and } n = 1, ..., N_d$$

$$\theta_d \overset{iid}{\sim} Dir(\alpha) \text{ for } d = 1, ..., D$$

$$\beta_k \overset{iid}{\sim} Dir(\gamma) \text{ for } k = 1, ..., K$$

Selecting number of topics per model. Models for each omic were trained independently, and the number of topics was selected per model. To evaluate model fit, we simulated sample count data drawn from each model, and compared it to the true distributions observed. We used three metrics, correlation of quantiles of sample distributions, correlation of quantiles of feature distributions, and correlation of pairwise marginal distributions (correlation of feature correlations). For each omic, we used the heuristic "elbow method" to select the minimum number of topics that produced the highest correlation across all of the three metrics. The final number of topics selected were 4 for 16s data, 5 for MTG data, 4 for MTT data, and 7 for MBX data (Supplementary Fig. 14).

In selecting the number of topics, we prioritized the pairwise marginal distribution correlation between features, because 1. This was the most variable across numbers of topics and 2. It was the most intuitively insightful (i.e. a model fits well when it captures feature-feature relationships across samples). The decision to choose the number of topics by comparing simulated and observed data stems from the intuition that data generated from a model should fit the actual data distribution well if the model is to be considered a good fit. Unlike using a simple likelihood calculation, this approach provides the added benefit of defining just how many topics is optimal for this modeling strategy, but also how well the final model actually captures the statistical properties of the original data.

Identifying cross-omic topics. To determine the optimal number of cross-omic topics, we selected the minimum number of topic clusters that minimized the gap statistic amongst clustered topics. First, we clustered topics into n groups based on their spearman correlation across samples. Then, for each cluster, we calculated the average distance between topics in that cluster (within group distance), and divided by the average distance between topics in that cluster and all other topics (between group distance). We used manhattan, euclidean, and inverse correlation as distance metrics, and selected the minimum number of clusters that minimized the above ratio adequately. In the end, we selected 4 cross-omic topic clusters (Supplementary Fig. 2A). Topics were visualized by PCoA on the topic distribution across samples using manhattan distance (Supplementary Fig. 2B).

We additionally found that single-omic topics within a cross-omic topic group significantly correlated with one another far more frequently than would be expected by random chance. To test this, we created 15 null count tables, drawing each sample from a multinomial where probabilities are assigned for each feature using a random sample's actual feature distribution. We then fit an LDA model to each of the null count tables using the same process, parameters, and random seeds as were used in the original models. We then clustered the topics produced from the null models into four clusters, as was done with the original data, and counted the number of significant (spearman correlation test, $p < .05$) correlations existed between topics that shared the same cluster. We found that null topics in the same cluster shared very few significant correlations and null topic distributions across 15 permutations had correlation structures stronger than the true data (Supplementary Fig. 2C). We additionally visualized a correlation matrix between null topics (Supplementary Fig. 2D) and between true topics (Supplementary Fig. 2E).

It is relevant to note that the treatment of MTT topics differed slightly from the rest. The total attributable weight across all samples to topic3 MTT was orders of magnitude higher than the other MTT topics (Supplementary Fig. 7). In addition, no specific features differentiated topic3 MTT from the other MTT topics (data not shown). Because this over attribution of topic3 MTT was not informative of differences between samples nor interpretable, and because its inclusion swayed all downstream clustering analyses considerably, we removed this component.

Interpreting topics by feature weights. We used the feature weight vectors (β)s to interpret each cross-omic topic. We limited features to those with high standard deviations across topics, as we wanted to identify features that were specific to certain topics and not ubiquitous across topics. Specifically, we identified features where the max value was less than 1.5 standard deviations over the median. We also limited features to those with high values along axes of maximal variation between topics, as these are the features driving topic definitions and separation most strongly. Specifically, we performed singular value decomposition on the topic by feature weight matrices independently (including all features), and recorded features with weights along the PCA axes above the 99th percentile for 16s and MBX data, and above the 99.9th percentile for MTG and MTT features. Features that fulfilled both of the above criteria are reported in Fig. 2. These thresholds were selected heuristically based on the maximum number of features that could be clearly visually represented.

Generalization of modeling approach to outside datasets. To demonstrate that topics may be observed in other independently collected datasets, we utilized 16s data from David et al.[111] and metabolomics data from Telleria et al.[112], which utilized similar data collection and processing methods to this study. We normalized original data in the same fashion as presented above. We then trained LDA models with the same parameters as used in this paper, and subsequently calculated feature-feature cosine distance matrices based on the feature-topic distributions (betas) in each model. We compared the feature-feature matrices derived from the independent studies to those from this study using a mantel test. Then, we permuted the rows of one feature-feature distance matrix 10,000 times to simulate mantel z-scores under the null hypothesis, and compared this distribution to the true mantel z-score.

Clustering samples by topic distribution. To cluster samples by their topic distribution, we used the R package pheatmap, which performs hierarchical clustering using inverse correlation as a distance metric.

Differential analysis. Differential analysis was performed to determine the association between any given microbial feature between the Autism and typically developing cohorts after clustering. First, we selected likely candidates using the Boruta package in R. The method performs a top-down search for relevant features by comparing original attributes' importance with importance achievable at random, estimated using their permuted copies, and progressively eliminating irrelevant features to stabilize that test[113].

In more detail, filtering features with Boruta works as such: 1) create an extended dataset by adding columns of randomly permuted features; 2) train a random forest classifier on the extended dataset to predict phenotype (Autism vs. typically developing) and calculate feature importance as the mean decrease in accuracy; 3) if the true feature Z score is higher than the maximum Z score of it's permuted versions over 100 runs, consider this feature "important". To increase consistency and generalizability, we repeated this described test 100 times, and tallied the number of times any given feature was considered "important". Features in the top 90th percentile of tallied counts were passed to a subsequent Wilcoxon rank sum, and were reported if their differential was significant ($p < .05$) after Benjamini & Hocherg correction[114].

Number of input features for each omic and number of features that passed the first step of Boruta filtering are reported here: 1187 features input and 7 features passed filtering (MBX), 10543 features input and 15 features passed filtering (MTG), 10625 features input and 8 features passed filtering (MTT), 5265 features input and 4 passed filtering (16S).

Ethical approval and informed consent. This study was authorized by the Stanford University Institutional Review Board protocol number 30205. Informed consent was obtained from all subjects and/or their legal guardian(s).

Data availability

All analyses and processed data files are available at: https://github.com/MaudeDavidLab/multiomics_topic_modeling. All sequence data and de-identified participant responses can be found in NCBI under project PRJNA895487.

Received: 12 October 2022; Accepted: 5 July 2023
Published online: 13 July 2023

References

1. Lord, C. *et al.* Autism diagnostic observation schedule: A standardized observation of communicative and social behavior. *J. Autism Dev. Disord.* **19**, 185–212 (1989).
2. Organization for Autism Research. 1,000 people surveyed, survey says.... howpublished (2020). https://researchautism.org/1000-people-surveyed-survey-says/ Accessed 31 Aug 2022.
3. Wasilewska, J. & Klukowski, M. Gastrointestinal symptoms and autism spectrum disorder: Links and risks - a possible new overlap syndrome. *Pediatr. Health Med. Therapeut.* **6**, 153 (2015).
4. Kohane, I. S. *et al.* The Co-Morbidity burden of children and young adults with autism spectrum disorders. *PLoS ONE* **7**, e33224 (2012).
5. Hsiao, E. Y. Gastrointestinal issues in autism spectrum disorder. *Harv. Rev. Psychiatry* **22**, 104 (2014).
6. Kang, D.-W. *et al.* Microbiota transfer therapy alters gut ecosystem and improves gastrointestinal and autism symptoms: an open-label study. *Microbiome* **5**, 1–16 (2017).
7. Kang, D.-W. *et al.* Long-term benefit of microbiota transfer therapy on autism symptoms and gut microbiota. *Sci. Rep.* **9**, 1–9 (2019).
8. Santocchi, E. *et al.* Effects of probiotic supplementation on gastrointestinal, sensory and core symptoms in autism spectrum disorders: A randomized controlled trial. *Front. Psychiatry* **11**, 550593 (2020).

9. Nitschke, A., Deonandan, R. & Konkle, A. T. The link between autism spectrum disorder and gut microbiota: A scoping review. *Autism* **24**, 1328–1344 (2020).
10. Iglesias-Vázquez, L., Van Ginkel Riba, G., Arija, V. & Canals, J. Composition of gut microbiota in children with autism spectrum disorder: A systematic review and meta-analysis. *Nutrients* **12**, 792 (2020).
11. Xu, M., Xu, X., Li, J. & Li, F. Association between gut microbiota and autism spectrum disorder: A systematic review and meta-analysis. *Front. Psychiatry* **10**, 473 (2019).
12. Al-Otaish, H. *et al.* Relationship between absolute and relative ratios of glutamate, glutamine and GABA and severity of autism spectrum disorder. *Metab. Brain Dis.* **33**, 843–854 (2018).
13. Marotta, R. et al. The neurochemistry of autism. Brain Sci. **10** (2020).
14. Cochran, D. M. *et al.* Relationship among glutamine, γ-Aminobutyric acid, and social cognition in autism spectrum disorders. *J. Child Adolesc. Psychopharmacol.* **25**, 314 (2015).
15. Horder, J. *et al.* Reduced subcortical glutamate/glutamine in adults with autism spectrum disorders: A [^{1}H]MRS study. *Transl. Psychiatry* **3**, e279 (2013).
16. Strandwitz, P. *et al.* GABA-modulating bacteria of the human gut microbiota. *Nat. Microbiol.* **4**, 396–403 (2019).
17. Yano, J. M. *et al.* Indigenous bacteria from the gut microbiota regulate host serotonin biosynthesis. *Cell* **161**, 264–276 (2015).
18. Rossignol, D. A. & Frye, R. E. Mitochondrial dysfunction in autism spectrum disorders: A systematic review and meta-analysis. *Mol. Psychiatry* **17**, 290–314 (2012).
19. Hu, T., Dong, Y., He, C., Zhao, M. & He, Q. The gut microbiota and oxidative stress in autism spectrum disorders (ASD). *Oxid. Med. Cell. Longev.* **2020**, 8396708 (2020).
20. Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003).
21. Gurry, T., Nguyen, L. T. T., Yu, X. & Alm, E. J. Functional heterogeneity in the fermentation capabilities of the healthy human gut microbiota. *PLoS ONE* **16**, e0254004 (2021).
22. Wilmanski, T. *et al.* Heterogeneity in statin responses explained by variation in the human gut microbiome. *Med (N Y)* **3**, 388–405 (2022).
23. Laukens, D., Brinkman, B. M., Raes, J., De Vos, M. & Vandenabeele, P. Heterogeneity of the gut microbiome in mice: Guidelines for optimizing experimental design. *FEMS Microbiol. Rev.* **40**, 117–132 (2016).
24. Ho, L. *et al.* Heterogeneity in gut microbiota drive polyphenol metabolism that influences alpha-synuclein misfolding and toxicity. *J. Nutr. Biochem.* **64**, 170–181 (2019).
25. Arumugam, M. *et al.* Enterotypes of the human gut microbiome. *Nature* **473**, 174–180 (2011).
26. Knights, D. *et al.* Rethinking "enterotypes". Cell Host Microbe **16** (2014).
27. Symul, L. *et al.* Sub-communities of the vaginal ecosystem in pregnant and non-pregnant women. *bioRxiv*https://doi.org/10.1101/2021.12.10.471327 *(2022).*
28. Sankaran, K. & Holmes, S. P. Latent variable modeling for the microbiome. *Biostatistics* **20**, 599–614 (2018).
29. Deek, R. A. & Li, H. A Zero-Inflated latent Dirichlet allocation model for microbiome studies. *Front. Genet.* **11**, 602594 (2021).
30. Breuninger, T. A. *et al.* Associations between habitual diet, metabolic disease, and the gut microbiota using latent Dirichlet allocation. *Microbiome* **9**, 1–18 (2021).
31. Sommeria-Klein, G. *et al.* Latent Dirichlet allocation reveals spatial and taxonomic structure in a DNA-based census of soil biodiversity from a tropical forest. *Mol. Ecol. Resour.* **20**, 371–386 (2020).
32. Okui, T. A Bayesian nonparametric topic model for microbiome data using subject attributes. *IPSJ Trans. Bioinformat.* **13**, 1–6 (2020).
33. Holmes, I., Harris, K. & Quince, C. Dirichlet multinomial mixtures: Generative models for microbial metagenomics. *PLoS ONE* **7**, e30126 (2012).
34. Harrison, J. G., Calder, W. J., Shastry, V. & Buerkle, C. A. Dirichlet-multinomial modelling outperforms alternatives for analysis of microbiome and other ecological count data. *Mol. Ecol. Resour.* **20**, 481–497 (2020).
35. Wadsworth, W. D. *et al.* An integrative Bayesian Dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. *BMC Bioinformat.* **18**, 94 (2017).
36. Chin, R. & Lee, B. Y. *Principles and Practice of Clinical Trial Medicine* (Elsevier, 2008).
37. McLaren, M. R., Willis, A. D. & Callahan, B. J. Consistent and correctable bias in metagenomic sequencing experiments. *Elife* **21**, 8279 (2019).
38. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833–844 (2017).
39. Shakya, M., Lo, C. C. & Chain, P. S. G. Advances and challenges in metatranscriptomic analysis. *Front. Genet.* **10**, 904 (2019).
40. Johnson, C. H. & Gonzalez, F. J. Challenges and opportunities of metabolomics. *J. Cell. Physiol.* **227**, 2975–2981 (2012).
41. Xu, Z. & Knight, R. Dietary effects on human gut microbiome diversity. *Br. J. Nutr.* **113**, S1–S5 (2015).
42. Manor, O. *et al.* Health and disease markers correlate with gut microbiome composition across thousands of people. *Nat. Commun.* **11**, 1–12 (2020).
43. Hagerty, S. L., Hutchison, K. E., Lowry, C. A. & Bryan, A. D. An empirically derived method for measuring human gut microbiome alpha diversity: Demonstrated utility in predicting health-related outcomes among a human clinical sample. *PLoS ONE* **15**, e0229204 (2020).
44. Menni, C. *et al.* Gut microbiome diversity and high-fibre intake are related to lower long-term weight gain. *Int. J. Obes.* **41**, 1099–1105 (2017).
45. Fassarella, M. *et al.* Gut microbiome stability and resilience: Elucidating the response to perturbations in order to modulate gut health. *Gut* **70**, 595–605 (2021).
46. Arem, H. *et al.* The healthy eating index 2005 and risk for pancreatic cancer in the NIH-AARP study. *J. Natl. Cancer Inst.* **105**, 1298–1305 (2013).
47. Labedzki, A., Buters, J., Jabrane, W. & Fuhr, U. Differences in caffeine and paraxanthine metabolism between human and murine CYP1A2. *Biochem. Pharmacol.* **63**, 2159–2167 (2002).
48. Nyéki, A., Buclin, T., Biollaz, J. & Decosterd, L. A. NAT2 and CYP1A2 phenotyping with caffeine: Head-to-head comparison of AFMU versus AAMU in the urine metabolite ratios. *Br. J. Clin. Pharmacol.* **55**, 62–67 (2003).
49. Le Marchand, L., Franke, A. A., Custer, L., Wilkens, L. R. & Cooney, R. V. Lifestyle and nutritional correlates of cytochrome CYP1A2 activity: Inverse associations with plasma lutein and alpha-tocopherol. *Pharmacogenetics* **7**, 11–19 (1997).
50. Lampe, J. W. *et al.* Brassica vegetables increase and apiaceous vegetables decrease cytochrome P450 1A2 activity in humans: Changes in caffeine metabolite ratios in response to controlled vegetable diets. *Carcinogenesis* **21**, 1157–1162 (2000).
51. Tantcheva-Poór, I., Zaigler, M., Rietbrock, S. & Fuhr, U. Estimation of cytochrome P-450 CYP1A2 activity in 863 healthy Caucasians using a saliva-based caffeine test. *Pharmacogenetics* **9**, 131–144 (1999).
52. Brejchova, K. *et al.* Understanding FAHFAs: From structure to metabolic regulation. *Prog. Lipid Res.* **79**, 101053 (2020).
53. Schultz Moreira, A. R. *et al.* 9-PAHSA prevents mitochondrial dysfunction and increases the viability of steatotic hepatocytes. *Int. J. Mol. Sci.* **21**, 8279 (2020).
54. Blanco, P. *et al.* Bacterial multidrug efflux pumps: Much more than antibiotic resistance determinants. *Microorganisms* **4**, 14 (2016).

55. Pereira, S. F. F., Goss, L. & Dworkin, J. Eukaryote-like serine/threonine kinases and phosphatases in bacteria. *Microbiol. Mol. Biol. Rev.* **75**, 192–212 (2011).
56. Henderson, P. & Maiden, M. Homologous sugar transport proteins in *Escherichia coli* and their relatives in both prokaryotes and eukaryotes. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* **326**, 391–410. https://doi.org/10.1098/rstb.1990.0020 (1990).
57. Gaci, N., Borrel, G., Tottey, W., O'Toole, P. W. & Brugère, J.-F. Archaea and the human gut: New beginning of an old story. *World J. Gastroenterol.* **20**, 16062 (2014).
58. Ze, X., Duncan, S. H., Louis, P. & Flint, H. J. Ruminococcus bromii is a keystone species for the degradation of resistant starch in the human colon. *ISME J.* **6**, 1535–1543 (2012).
59. Mueller, J. W., Gilligan, L. C., Idkowiak, J., Arlt, W. & Foster, P. A. The regulation of steroid action by sulfation and desulfation. *Endocr. Rev.* **36**, 526–563 (2015).
60. Gibbs, T. T., Russek, S. J. & Farb, D. H. Sulfated steroids as endogenous neuromodulators. *Pharmacol. Biochem. Behav.* **84**, 555–567 (2006).
61. Stergiakouli, E. *et al.* Steroid sulfatase is a potential modifier of cognition in attention deficit hyperactivity disorder. *Genes Brain Behav.* **10**, 334 (2011).
62. Brookes, K. J. *et al.* Association of the steroid sulfatase (STS) gene with attention deficit hyperactivity disorder. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **147B**, 1531–1535 (2008).
63. Cerqueira, N. M. *et al.* Cholesterol biosynthesis: A mechanistic overview. *Biochemistry* **55**, 5483–5506 (2016).
64. Solomon, J. M. & Grossman, A. D. Who's competent and when: Regulation of natural genetic competence in bacteria. *Trends Genet.* **12**, 150–155 (1996).
65. Klein, M. I. *et al.* Structural and molecular basis of the role of starch and sucrose in streptococcus mutans biofilm development. *Appl. Environ. Microbiol.* **75**, 837 (2009).
66. Ogura, M., Liu, L., Lacelle, M., Nakano, M. M. & Zuber, P. Mutational analysis of ComS: Evidence for the interaction of ComS and MecA in the regulation of competence development in bacillus subtilis. *Mol. Microbiol.* **32**, 799–812 (1999).
67. Cordero, M. *et al.* The induction of natural competence adapts staphylococcal metabolism to infection. *Nat. Commun.* **13**, 1–17 (2022).
68. Zhang, Y. *et al.* Dietary type 2 resistant starch improves systemic inflammation and intestinal permeability by modulating microbiota and metabolites in aged mice on high-fat diet. *Aging* **12**, 9173 (2020).
69. Duan, J. *et al.* Age-related changes in microbial composition and function in cynomolgus macaques. *Aging* **11**, 12080–12096 (2019).
70. Zhang, G. *et al.* The association between inflammaging and age-related changes in the ruminal and fecal microbiota among lactating holstein cows. *Front. Microbiol.* **10**, 1803 (2019).
71. Olaisen, M. *et al.* Bacterial mucosa-associated microbiome in inflamed and proximal noninflamed ileum of patients with crohn's disease. *Inflamm. Bowel Dis.* **27**, 12 (2021).
72. Kelly, T. N. *et al.* Gut microbiome associates with lifetime cardiovascular disease risk profile among bogalusa heart study participants. *Circ. Res.* **119**, 956 (2016).
73. Liu, Y. *et al.* Dietary quality and the colonic mucosa-associated gut microbiome in humans. *Am. J. Clin. Nutr.* **110**, 701–712 (2019).
74. Hudeček, O. *et al.* Dinucleoside polyphosphates act as 5'-RNA caps in bacteria. *Nat. Commun.* **11**, 1–11 (2020).
75. Balderas-Hernández, V. E. *et al.* Catechol biosynthesis from glucose in *Escherichia coli* anthranilate-overproducer strains by heterologous expression of anthranilate 1,2-dioxygenase from Pseudomonas aeruginosa PAO1. *Microb. Cell Fact.* **13**, 136 (2014).
76. Kocaçalişkan, I., Talan, I. & Terzi, I. Antimicrobial activity of catechol and pyrogallol as allelochemicals. *Z. Naturforsch. C J. Biosci.* **61**, 639–642 (2006).
77. Koistinen, V. M. *et al.* Metabolite pattern derived from lactiplantibacillus plantarum-fermented rye foods and in vitro gut fermentation synergistically inhibits bacterial growth. *Mol. Nutr. Food Res.* **66**, e2101096 (2022).
78. Goldstein, D. S. *et al.* Multiple catechols in human plasma after drinking caffeinated or decaffeinated coffee. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* **1185**, 122988 (2021).
79. Tressl, R., Bahri, D., Köppler, H. & Jensen, A. Diphenols and caramel compounds in roasted coffees of different varieties. II. (author's transl). *Z. Lebensm. Unters. Forsch.* **167**, 111–114 (1978).
80. Rutkowsky, J. M. *et al.* Acylcarnitines activate proinflammatory signaling pathways. *Am. J. Physiol. Endocrinol. Metab.* **306**, E1378–E1387 (2014).
81. Nilsen, M. S. *et al.* 3-hydroxyisobutyrate, a strong marker of insulin resistance in type 2 diabetes and obesity that modulates white and brown adipocyte metabolism. *Diabetes* **69**, 1903–1916 (2020).
82. Ahmad, T. R. & Haeusler, R. A. Bile acids in glucose metabolism and insulin signalling - mechanisms and research needs. *Nat. Rev. Endocrinol.* **15**, 701–712 (2019).
83. Rebouche, C. J. Ascorbic acid and carnitine biosynthesis. *Am. J. Clin. Nutr.* **54**, 1147S–1152S (1991).
84. Meadows, J. A. & Wargo, M. J. Carnitine in bacterial physiology and metabolism. *Microbiology* **161**, 1161 (2015).
85. Wells, T. J., Tree, J. J., Ulett, G. C. & Schembri, M. A. Autotransporter proteins: Novel targets at the bacterial cell surface. *FEMS Microbiol. Lett.* **274**, 163–172 (2007).
86. Remaut, H. & Ben-Tal, N. Usher proteins: Lifting the lid on pilus assembly. *eLife* **3**, e04997. https://doi.org/10.7554/eLife.04997 (2014).
87. Sherlock, O., Vejborg, R. M. & Klemm, P. The TibA adhesin/invasin from enterotoxigenic *Escherichia coli* is self recognizing and induces bacterial aggregation and biofilm formation. *Infect. Immun.* **73**, 1954–1963 (2005).
88. Schumann, U. *et al.* YbdG in *Escherichia coli* is a threshold-setting mechanosensitive channel with MscM activity. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 12664–12669 (2010).
89. Siegele, D. A. Universal stress proteins in *Escherichia coli. J. Bacteriol.* **187**, 6253–6254 (2005).
90. Domka, J., Lee, J. & Wood, T. K. YliH (BssR) and YceP (BssS) regulate *Escherichia coli* K-12 biofilm formation by influencing cell signaling. *Appl. Environ. Microbiol.* **72**, 2449–2459 (2006).
91. Miwa, T., Chadani, Y. & Taguchi, H. *Escherichia coli* small heat shock protein IBPA is an aggregation-sensor that self-regulates its own expression at posttranscriptional levels. *Mol. Microbiol.* **115**, 142–156 (2021).
92. Béchon, N. *et al.* Autotransporters drive biofilm formation and autoaggregation in the diderm firmicute veillonella parvula. *J. Bacteriol.* **202**, 10 (2020).
93. Lord, C. *et al.* The autism diagnostic observation schedule-generic: A standard measure of social and communication deficits associated with the spectrum of autism. *J. Autism Dev. Disord.* **30**, 205–223 (2000).
94. Oxenkrug, G. Serotonin-kynurenine hypothesis of depression: Historical overview and recent developments. *Curr. Drug Targets* **14**, 514–521 (2013).
95. Lapin, I. P. Antagonism of kynurenic acid to anxiogens in mice. *Life Sci.* **63**, L231–L236 (1998).
96. Adams, J. B. *et al.* Nutritional and metabolic status of children with autism versus neurotypical children, and the association with autism severity. *Nutr. Metab.* **8**, 1–32 (2011).
97. Castro-Portuguez, R. & Sutphin, G. L. Kynurenine pathway, NAD+ synthesis, and mitochondrial function: Targeting tryptophan metabolism to promote longevity and healthspan. *Exp. Gerontol.* **132**, 110841. https://doi.org/10.1016/j.exger.2020.110841 (2020).

98. Bo, P. *et al.* Experimental study on central effects of carboxyethyl-gamma-aminobutyric acid (CEGABA). *Farmaco Sci.* **43**, 363–372 (1988).
99. Savoldi, F., Ceroni, M., Fussi, F. & Curti, M. Pharmacological effects of cegaba, a new aminoacid occurring in mammalian brain. *Farmaco Sci.* **42**, 77–79 (1987).
100. Barone, R. *et al.* A subset of patients with autism spectrum disorders show a distinctive metabolic profile by dried blood spot analyses. *Front. Psychiatry* **9**, 636 (2018).
101. Jay Gargus, J. & Imtiaz, F. Mitochondrial energy-deficient endophenotype in autism. *Am. J. Biochem. Biotechnol.* **4**, 198–207 (2008).
102. Rossignol, D. A. & Frye, R. E. Mitochondrial dysfunction in autism spectrum disorders: A systematic review and meta-analysis. *Mol. Psychiatry* **17**, 290 (2012).
103. Golubeva, A. V. *et al.* Microbiota-related changes in bile acid & tryptophan metabolism are associated with gastrointestinal dysfunction in a mouse model of autism. *EBioMedicine* **24**, 166–178. https://doi.org/10.1016/j.ebiom.2017.09.020 (2017).
104. Wu, W. L. Association among gut microbes, intestinal physiology, and autism. *EBioMedicine* **25**, 11–12 (2017).
105. Zimmerman, A. W. *et al.* Cerebrospinal fluid and serum markers of inflammation in autism (2005).
106. Sahm, F. *et al.* The endogenous tryptophan metabolite and NAD+ precursor quinolinic acid confers resistance of gliomas to oxidative stress. *Cancer Res.* **73**, 3225 (2013).
107. Tataru, C. *et al.* Longitudinal study of stool-associated microbial taxa in sibling pairs with and without autism spectrum disorder. *ISME Commun.* **1**, 1–12 (2021).
108. Duda, M., Daniels, J. & Wall, D. P. Mobile autism risk assessment. in *PsycTESTS Dataset* (2017).
109. West, K. *et al.* Multi-angle meta-analysis of the gut microbiome in autism spectrum disorder: A step toward understanding patient subgroups. *Sci. Rep.* **12**, 17034 (2022).
110. Caporaso, J. G. *et al.* Ultra-high-throughput microbial community analysis on the illumina HiSeq and MiSeq platforms. *ISME J.* **6**, 1621–1624 (2012).
111. David, M. M. *et al.* Children with autism and their typically developing siblings differ in amplicon sequence variants and predicted functions of stool-associated microbes. *Msystems* **6**, e00193 (2021).
112. Telleria, O. *et al.* A comprehensive metabolomics analysis of fecal samples from advanced adenoma and colorectal cancer patients. *Metabolites* **12**, 550 (2022).
113. Kursa, M. B. & Rudnicki, W. R. Feature selection with the Boruta package. *J. Stat. Softw.* **36**, 1–13 (2010).
114. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodol.)* **57**, 289–300 (1995).

Author contributions

C.T. conceived of and performed the analysis and wrote the manuscript. E.R. curated the metadata. K.D. corresponded with participants. X.Y. processed the raw sequencing data. T.Z.D., D.P.W., S.I., and M.M.D. conceived the project. All authors reviewed the manuscript.

Funding

This work was supported by the NIH (NIDA R44DA04395402) and the Larry W. Martin & Joyce B. O'Neill Endowed Fellowship, as well as NSF (Award Number:2025457). The funding bodies did not play any roles in the design of the study, collection, analysis, and interpretation of data, and in writing the manuscript.

Competing interests

MP, ER, XY, TZD, SI, and MMD have a financial interest in Second Genome Inc, an independent therapeutics company with products in development to treat Inflammatory Bowel Diseases and Cancer, and could potentially benefit from the outcomes of this research. MMD is co-owner of NeuroBiome, LLC., a company specialized in developing biosensors. DPW is cofounder of Cognoa, a company focused on digital methods for healthy child development. CT and BC report no conflict of interest.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-38228-0.

Correspondence and requests for materials should be addressed to C.T. or M.M.D.

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

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