# Defining and Evaluating Microbial Contributions to Metabolite Variation in Microbiome-Metabolome Association Studies

Cecilia Noecker[1], Hsuan-Chao Chiu[1,2], Colin P. McNally[1], and Elhanan Borenstein[1,3,4,*]

[1] Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA
[2] Current affiliation: Office of Chief Technology Officer, MediaTek Inc., Hsinchu City, Taiwan 30078
[3] Department of Computer Science and Engineering, University of Washington, Seattle, WA 98195, USA
[4] Santa Fe Institute, Santa Fe, NM 87501, USA
* Corresponding Author (elbo@uw.edu)

# Abstract

Correlation-based analysis of paired microbiome-metabolome datasets is becoming a widespread research approach, aiming to comprehensively identify microbial drivers of metabolic variation. To date, however, the limitations of this approach have not been comprehensively evaluated. To address this challenge, we introduce a mathematical framework to quantify the contribution of each taxon to metabolite variation based on uptake and secretion fluxes. We additionally use a multi-species metabolic model to simulate simplified gut communities, generating idealized microbiome-metabolome datasets. We then compare observed taxon-metabolite correlations in these datasets to calculated ground-truth taxonomic contribution values. We find that in simulations of both a model 10-species community and of complex human gut microbiota, correlation-based analysis poorly identifies key contributors, with extremely low predictive value despite the idealized setting. We further demonstrate that the predictive value of correlation analysis is strongly influenced by both metabolite and taxon properties, as well as exogenous environmental variation. We finally discuss the practical implications of our findings for interpreting microbiome-metabolome studies.

# Importance

Identifying the key microbial taxa responsible for metabolic differences between microbiomes is an important step towards understanding and manipulating microbiome metabolism. To achieve this goal, researchers commonly conduct microbiome-metabolome association studies, comprehensively measuring both the composition of species and the concentration of metabolites across a set of microbial community samples, and then testing for correlations between microbes and metabolites. Here, we evaluated the utility of this general approach by first developing a rigorous mathematical definition of the contribution of each microbial taxon to metabolite variation, and then examining these contributions in simulated datasets of microbial community metabolism. We found that standard correlation-based analysis of our simulated microbiome-metabolome datasets identifies true contributions with very low predictive value, and that its performance depends strongly on specific properties of both metabolites and microbes, as well as on the surrounding environment. Combined, our findings can guide future interpretation and validation of microbiome-metabolome studies.

# Introduction

Microbial communities have a tremendous impact on their surroundings, ranging from the degradation of environmental toxins (1) to the production of climate change-relevant metabolites (2). Host-associated communities, in particular, have a substantial impact on their hosts, and often produce a diverse set of metabolites that interact with numerous host pathways. In humans, such microbiome-derived metabolites have been identified as contributing factors to a wide array of diseases including heart disease (3), autism (4), non-alcoholic fatty liver disease (5), colon cancer (6), inflammatory bowel disease (7), and susceptibility to infection (8). Characterizing the ways microbial communities modulate their environments and the relationship between community structure and metabolic impact is therefore a major, timely, and complex challenge with promising implications for human health, as well as to environmental stewardship, agriculture, and industry.

When facing this challenge, perhaps the most important task is identifying specific community members that drive variation in metabolites of interest. Taxa responsible for observed metabolic differences across communities may be ideal targets for interventions aiming to modify metabolic phenotypes. Their identification, however, can be a daunting task. Complex microbial communities are often composed of hundreds or thousands of poorly characterized species, each with a unique and frequently unknown complement of metabolic capacities. Even when multiple species are known to possess the potential to synthesize or degrade a metabolite of interest, the metabolic activity of each species (and consequently, its contribution to metabolic variation) may be different (9). Moreover,

community ecology, interspecies interactions, and nutrient availability (e.g., via diet) can all regulate and influence the metabolic activity of each species, rendering the link between community members and metabolic products extremely complex and challenging to infer (10–12).

To address this challenge and to identify community members that play an important role in metabolic variation, a growing number of studies are now comprehensively assaying multiple facets of community structure across samples, including, most notably, taxonomic and metabolite compositions (13). For example, to investigate the links between taxonomic shifts and metabolic phenotypes in the healthy vaginal microbiome and in bacterial vaginosis, a recent study used a combination of 16S rRNA qPCR, sequencing, and both global and targeted metabolomics (14). Another study, aiming to identify taxonomic and metabolic features of resistance and susceptibility to *C. dificile* infection in the mouse gut similarly applied 16S rRNA sequencing and global metabolomics (15). In another example, researchers characterized metabolic and microbial features of periodontitis in the oral microbiome before and after treatment, combining 16S rRNA sequencing, shotgun metagenomic sequencing, and metabolomics (16). These are just a few examples of a plethora of recent microbiome-metabolome studies, investigating the metabolic effects of microbiome variation in the contexts of chronic and infectious disease, agriculture, precision medicine, nutrition, fermented food science, and more (17–24). Such multi-omic studies are also a major focus of several large-scale initiatives to study both host-associated and environmental microbiomes (25, 26).

Given the taxonomic and metabolomic profiles obtained via such microbiome-metabolome assays, the vast majority of studies rely on simple univariate correlation-based analyses to link variation in community ecology to variation in metabolic activity (11, 14, 15, 27–30). Such analyses specifically aim to identify species whose abundance across samples is correlated with the concentration of metabolites, often assuming that highly significant correlations reflect a direct mechanistic link between the taxon and metabolite in question. These studies further regularly assume that positive correlations imply synthesis and negative correlations imply degradation, or that targeting the microbe in question could be used to modulate the concentrations of the metabolites with which it is correlated. For example, a recent study characterizing the microbiome and metabolome in Spleen-yang-deficiency syndrome (29) concluded that a positive correlation between Bacteroides and mannose likely resulted from extracellular degradation of mannan into mannose by that taxon. Similarly, a study of antibiotic perturbations to the microbiome and metabolome stated that the presence of several weak positive and negative correlations between genera and arginine supported the conclusion that arginine levels may be affected by many community members with high functional redundancy (27).

Yet, to date, the extent to which a correlation-based analysis effectively detects direct metabolic relationships between taxa and metabolites is unclear. Obviously, a strong correlation between the abundance of a certain species and the concentration of a metabolite across samples *could* reflect direct synthesis or degradation of the metabolite by that species, but could also arise due to environmental effects, precursor availability,

115 selection, random chance, or co-occurrence between species. Similarly, cross-feeding,
116 external host processes, and varying enzymatic regulation can mask a correlation even
117 when this species does in fact contribute to observed metabolite variation. Indeed,
118 previous studies have suggested that microbe-metabolite correlations must have a high
119 rate of false positives (31), and a recent experimental study pairing microbiome-
120 metabolome correlation analysis with *in vitro* monoculture validations found anecdotally
121 that several observed correlations were in fact false positives (32). The limitations of
122 correlation analysis have also been discussed and well-characterized in other data types
123 (for example (33, 34)). Importantly, however, the extent of such limitations in the context
124 of microbiome-metabolome studies, the way they are shaped by microbial community
125 metabolism, and their impact on data interpretation in this context have not been
126 systematically evaluated.
127
128 Importantly, two crucial challenges hinder a comprehensive and systematic evaluation of
129 correlation-based analysis. The first is the lack of a rigorous general definition of a
130 microbe's contribution to metabolite variability. While establishing the main taxonomic
131 contributors to metabolite variation may be straightforward for specialized, well-
132 characterized metabolites that are synthesized by just a single taxon, it can be much less
133 clear for metabolites that can be synthesized (and/or degraded or modified) by many
134 different taxa in the community. The second challenge is the absence of ground truth data
135 on the nature of microbe-metabolite relationships. While limited data on the taxa driving
136 metabolite shifts can be obtained from comparative mono- and co-culture studies (32, 35,
137 36), large-scale and comprehensive datasets that link species and metabolite

138 abundances in the context of a complex community, for which the precise impact of each
139 species on observed metabolite variation is known, are currently not available.
140
141 In this study, we address these two challenges, combining a novel framework for
142 quantifying microbial contributions with a model-based simulated dataset. Specifically, we
143 first introduce a generalizable and rigorous mathematical framework for decomposing
144 observed metabolite variation and quantifying the contribution of each community
145 member to this variation based on uptake and secretion fluxes. Second, we use a
146 dynamic multi-species genome-scale metabolic model to simulate the metabolism of
147 microbial communities of varying complexity and to generate idealized datasets of paired
148 taxonomic and metabolomic abundances, with complete information on metabolite fluxes,
149 microbial growth, interspecies interactions, and environmental influences. Applying our
150 mathematical framework to these simulated datasets, we could then compare calculated
151 contribution values to observed taxon-metabolite correlations and evaluate the ability of
152 correlation-based analyses to identify key microbial contributors. We were additionally
153 able to investigate factors that shape the relationship between community composition
154 and metabolism in depth and to identify specific properties and mechanisms that impact
155 the performance of microbiome-metabolome correlation studies.
156
157 Notably, given the objectives of this study, we intentionally focus on characterizing
158 microbiome-metabolome relationship in a model-based, tractable, and well-defined
159 setting. Indeed, our metabolic model may not perfectly capture all the complex and
160 diverse mechanisms that are at play in host-associated communities; however,

161 considering the scope of this study, accurately modeling the metabolism of a specific
162 community may not be crucial. Rather, for our analysis, we want our simulated data to
163 recapitulate broad trends observed in naturally occurring microbial ecosystems, as indeed
164 has been observed in similar models (37–41). Moreover, utilizing this model-based
165 approach allows us to dissect the relationship between community composition and
166 metabolic phenotypes without the complexities inherent to *in vivo* communities (including
167 spatial heterogeneity, measurement error, inter-microbial signaling, or strain-level
168 variation), and with variation in the concentrations of environmental metabolites resulting
169 exclusively from microbial metabolic activity. Analyzing the ability of a correlation-based
170 analysis to detect true microbial drivers of metabolite variation in these simplified, best-
171 case settings provides a baseline for the expected performances of such analyses in real
172 microbiome-metabolome studies.
173

# Results

### *Quantifying the impact of individual microbial species on variation in metabolite concentrations*

In this study, we consider a microbial community as an idealized system, consisting of a population of multiple microbial species in a shared, well-mixed, biochemical environment. Each species uptakes necessary metabolites from the shared environment, performs a variety of metabolic processes to promote its growth, and secretes certain metabolites back into the shared environment. We additionally assume that certain nutrients flow into the environment and that microbial cells and metabolites are diluted over time. These processes can represent, for example, the inflow of dietary nutrients and the transit through the gut in the context of the gut microbiome. For simplicity, we primarily consider a constant inflow and dilution rate, as in a chemostat setting. Accordingly, a microbiome-metabolome study can be conceived as analyzing a set of several such communities (at a certain point in time), each with a different composition of microbial species and correspondingly variable environmental metabolite concentrations. We focus initially on a controlled setting with identical nutrient inflow across all microbiomes, but later examine the impacts of differences in nutrient inflow between communities.

Given this setting, we first sought to establish a rigorous and quantitative framework for defining the impact of each microbial species (or any taxonomic grouping) in the community on the variation observed in the concentration of a given metabolite across community samples. We focused on species that *directly* modulate the environmental

197 concentration of a given metabolite via synthesis or degradation, ignoring indirect effects
198 via, for example, the synthesis of a precursor substrate that could impact the metabolic
199 activity of other species. We noted that the total concentration of a metabolite in the
200 environment can be represented as the sum of cumulative synthesis or degradation fluxes
201 of this metabolite by each of the *n* species in the community, as well as cumulative
202 environmental fluxes (e.g., total nutrient inflow and dilution). Formally, the metabolite
203 concentration, *M,* can therefore be expressed as a sum of *n* dependent random variables
204 $m_{i,}$ where each $m_i$ denotes the overall synthesis or degradation of the metabolite by each
205 species, along with an additional random variable $m_{env}$, denoting the overall impact of
206 environmental processes.
207

208 $$M = \sum_{i=1}^{n} m_i + m_{env}$$

209

210 As discussed above, when analyzing microbiome-metabolome datasets, the goal is often
211 to identify taxa responsible for *changes* in the concentration of a metabolite of interest
212 across a set of samples. Accordingly, here we wish to quantify the *contribution* of each
213 species to the *variance* in the concentration of that metabolite across samples.
214 Specifically, in the formulation
215 above, *var(M)* depends on the variance in the constituent microbial and environmental
216 factors, as well as the covariance between these components. This variance can then be
217 linearly separated into *n+1* terms, representing the contribution of each species (denoted
218 $c_i$), and of any
219 environmental nutrient fluxes (denoted $c_{env}$) to the total variation in the metabolite:

$$var(M) = \sum_{i=1}^{n} c_i + c_{env}\ ;\ c_i = var(m_i) + \sum_{j \neq i} cov(m_i, m_j) + cov(m_i, m_{env})$$

If the nutrient inflow is constant across samples, its effect can be ignored and its contribution to the variance $c_{env}$ is 0. Additionally, in a chemostat setting, the dilution of each metabolite can be accounted for in the calculation of each contribution, as it depends strictly on the dilution rate and on previous metabolite concentrations (Methods). Finally, in order to compare species contributions across metabolites and to represents the relative share of the total variance of a given metabolite that is attributable to species *I*, we defined the *relative* contribution to variance $\hat{c}_i$ of each species *i* to metabolite *M* by normalizing contribution values by the metabolite's total variance:

$$\hat{c}_i = \frac{c_i}{var(M)}$$

This framework for calculating microbial contribution values provides a systematic measure of the causal impact of each taxon on observed variation in the environmental concentration of each metabolite, distilling the effect of complex ecological and metabolic interactions to a concise and interpretable set of quantities. Moreover, the obtained contribution profile is a linear decomposition of observed metabolic variation, wherein the sum of contributions of all species equals the observed variation in the metabolite. Notably, when a species' activity has large negative covariances with the activities of other community members, contribution values can be negative. Such negative contribution values indicate that a species' secretion or uptake of that metabolite varies

in a way that mitigates the activity of others. Correspondingly, contribution values can be greater than 1, reflecting scenarios in which a species in fact generates more variation of this metabolite than is ultimately observed, but that its impact is mitigated by other species.

It is also worth noting that our analytical decomposition of contributions to variance is mathematically equivalent to calculating the Shapley values for the variance in metabolite concentrations (see Methods and Figure S1). Shapley value analysis is a game theory technique that defines an individual's contribution to a collective outcome, and has been shown to be the only general definition that is efficient, linear, symmetric, and assigns zero values to null contributors (42). A similar, Shapley value-based approach was recently applied to address the related problem of identifying the primary taxonomic contributors to differential functional abundances in metagenomic data (43).

### *A multi-species metabolic model for generating complex microbiome-metabolome data*

We next set out to generate a large-scale dataset of microbiome-metabolome profiles with complete information about metabolite uptake and secretion fluxes. To this end, we used a multi-species metabolic model to simulate the growth, dynamics, metabolism, and environment of a simple microbial community. This model is based on a previously introduced genome-scale framework for modeling the metabolism of multi-species communities and for tracking the metabolic activity of each community member over time (44, 45). Briefly, this framework assumes that each species optimizes its growth selfishly

given available nutrients in the shared environment and predicts the metabolic activity for each species in short time increments using Flux Balance Analysis (46). After each increment, the model uses the predicted metabolic activities of the various species to update the biomass of each species and the concentration of metabolites in the shared environment (hence, potentially impacting the growth and metabolism of other species in subsequent time steps). Importantly, this model allows for the natural emergence of metabolic competition and exchange between species, as well as selection for taxa with the most efficient growth rate in a given nutrient environment. Full details of this model and simulation parameters can be found in the Methods.

We specifically modeled a simplified gut community that was previously explored experimentally (47). This community includes 10 representative gut species, spanning the major clades found in the human gut and collectively encoding the key metabolic processes taking place in this environment, including breakdown of complex dietary polysaccharides, amino acid fermentation, and removal of fermentation end products via sulfate reduction and acetogenesis. Genome-scale metabolic models of these 10 species were obtained from the AGORA collection (40) – a recently introduced set of high-quality gut-specific metabolic models. To mimic the experimental gnotobiotic mouse setting (47), we simulate growth in a chemostat, with a nutrient inflow mimicking the content of a standard corn-based mouse chow, and a dilution rate consistent with mouse transit time and gut volume. While maintaining this nutritional environment, we systematically explored the landscape of possible community compositions, varying the initial relative abundance of each species from 10% to 60% (with a consistent total abundance equal to

the community carrying capacity), resulting in a total of 61 different community compositions. For the analysis below, we simulated growth for 144 hours (as 576 15-minute time steps). For most community compositions considered, this simulation time consisted of an initial stabilization period followed by a transition to a near-steady-state equilibrium with little change in community composition (Figure 1A). Notably, across the various simulations, some species maintained high abundances throughout the course of the simulation, while others reverted to lower levels.

Throughout the course of each simulation, we recorded the abundances of each species, the secretion and uptake rate of each metabolite by each species (as well as internal reaction fluxes), and the concentration of each metabolite in the environment (Figure 1A-B), thereby obtaining a comprehensive dataset describing species composition, metabolic activities, and metabolite concentrations across 61 different communities. To mirror the typical structure of a microbiome-metabolome cross-sectional dataset, we specifically considered the abundances of species and the concentrations of metabolites in the environment at the end of each simulation (i.e., after the final time point; see Figure 1). 60 of the 68 metabolites present in the nutrient inflow exhibited at least some variation across communities, as did 18 additional microbially-produced metabolites. Metabolite variation was generally low (median coefficient of variation 0.021), reflecting a relatively stable nutrient environment, yet 25 metabolites (32%) did have a coefficient of variation greater than 0.1. For downstream analysis, we excluded metabolites without substantial measurable variance across samples, filtering those with variance at or below the 25$^{th}$ percentile. This resulted in a dataset of 52 variable metabolites, of which 14 are purely

311 microbially-produced metabolites, 9 are microbially-produced but also present in the
312 nutrient inflow, and 29 are introduced only through the nutrient inflow. Of these 52 variable
313 metabolites, 47 are utilized by any member of the community (including 18 that are cross-
314 fed in at least one simulation). The final species compositions and the final concentrations
315 of several key metabolites across all simulations are shown in Figure 2A-F, and ordination
316 plots of species and metabolite data are shown in Figure S2.

317
318 Exploring this dataset, we found that species composition and metabolite concentrations
319 exhibited complex patterns and biologically reasonable distributions (Figure S3) (49).
320 Several metabolic processes known to occur in the mammalian gut were replicated by
321 our simulations, including, for example, conversion of acetate to butyrate by *E. rectale*
322 (48), and production of key microbial metabolites such as 4-aminobutyric acid (GABA),
323 indole, and succinate. Cross-feeding relationships were observed frequently (18
324 metabolites), including cross-feeding of 6 amino acids, whose exchange is widespread in
325 host-associated microbiota (50). Additionally, we ran several sets of simulations with
326 introduced fluctuations in the nutrient inflow concentrations, and found that the resulting
327 species compositions partially recapitulated the diet responses observed by Faith *et al.*
328 (47) (Supplementary Results).

329
330 Clearly, the model and simulations described above represent a gross simplification of
331 the microbiome's structure, dynamics, and function. Importantly, however, this
332 simplification is also an important strength. Specifically, the data obtained from these
333 simulations provide a unique opportunity to examine the relationship between community
334 dynamics and metabolic activity in a realistic, yet tractable model of community

335 metabolism where complete information about the activity and fluxes of each microbial
336 species is available (Figure S4). Indeed, our multi-species model captures many of the
337 intricacies of bacterial genome-scale metabolism and the interconnectedness (both within
338 and between species) of multiple metabolic processes, yet without additional complexities
339 inherent to *in vivo* communities. Furthermore, in our simulations, variation in the
340 concentrations of environmental metabolites results exclusively from microbial metabolic
341 activity, with no variation in nutrient inflow or other non-microbial sources, providing a
342 controlled setting for evaluating the relationship between community members and
343 metabolite concentrations.
344
345 ***Metabolite variation is driven by diverse microbial mechanisms***
346 Given the simulated dataset described above (for which uptake and secretion fluxes are
347 known), we applied our contribution framework to calculate the contribution of each
348 species to the variation observed in each of the 52 variable metabolites (Figure S5). The
349 resulting contribution values can be used as ground-truth information about the link
350 between microbial activity and environmental metabolites.
351
352 To highlight the nature and utility of such contribution values, and to demonstrate how
353 metabolic fluxes translate into contribution profiles, we first describe our results for several
354 example metabolites (Figure 2). Putrescine, an amino acid fermentation product, is an
355 example of the simplest case, in which one microbial species – *E. coli* – synthesizes a
356 metabolite that is not utilized or modified by other community members. Variation in the
357 environmental concentration of putrescine was hence fully determined by the level of

secretion from *E. coli*, which is therefore assigned a relative contribution of 1 (Figure 2B). Tetradecanoic acid, in contrast, was introduced (at a constant rate) via the nutrient inflow and utilized by the three *Bacteroides* species in the community to varying degree (primarily by *B. ovatus* and to a slightly lesser extent by *B. thetaiotaomicron*). The calculated contribution values successfully attributed variation in the environmental concentration of this metabolite to these three species, and correctly captured the difference in the magnitude between their effects (Figure 2C). Variation in uracil, another metabolite introduced via the nutrient inflow, was mainly driven by large shifts in its uptake by *B. ovatus,* but this effect is partially masked by *E. rectale,* which reduced its uptake when *B. ovatus*' flux was high and vice versa. Other species also utilized uracil, but at relatively similar levels across samples, and accordingly with relatively little impact on its variation. These complex patterns were all captured by the contribution profile obtained by our framework, with *B. ovatus* assigned a high positive contribution, *E. rectale* assigned an intermediate *negative* contribution, and other species assigned relatively negligible contribution values (Figure 2D). More complex species-metabolite relationships were also accurately and effectively summarized. Contribution values for acetate, for example, reflected the cross-feeding interactions that underlie variation in its concentration (Figure 2E). It was introduced to the shared environment by several species (primarily *C. symbiosum*), but most of its variation ultimately depended on the level of uptake by *E. rectale.* Finally, the contribution profile of succinate demonstrates how extremely strong interspecies interactions can produce contribution values much greater than the observed variance (Figure 2F). In the simulated data, this metabolite was synthesized by *B. hydrogenotrophica,* but was almost always fully utilized by other

community members. The calculated contributions suggest that if the synthesis of succinate by *B. hydrogenotrophica* would not have been offset by uptake from other species, the variance in succinate concentration across samples would have been 71.7 times higher than is actually observed. (Note that the difference between positive and negative is always 1.)

Examining the complete set of variable metabolites and calculated contribution values revealed similar patterns of interactions (Figure S5). Specifically, as for the metabolites discussed above, negative contributions and/or contribution values greater than 1 were widespread. Nearly all metabolites (50 out of 52) had at least one species with a negative contribution value, and 36 had at least one species with a contribution value greater than 1. Of the 32 other metabolites with negative contributions, 29 were present in the nutrient inflow and their negative contributions result from competition between species for their uptake. This prevalence of negative and extreme values suggests that strong negative interspecies interactions have substantial impacts on metabolite concentrations, and that often, observed variation in a given metabolite's concentration is the complex outcome of multiple species generating and offsetting much higher variation.

It is also important to note that while the average metabolic uptake/secretion flux of each species and the magnitude of its contribution to a given metabolite were generally significantly correlated (Spearman, $p < 0.01$ for 49 of the 52 metabolites), the species with the highest flux was often *not* the largest contributor to variation (26 of the 52 metabolites). Similarly, the variance in a species' flux was significantly correlated with its contribution

for 48 of the metabolites, but for 9 metabolites the species with the most variable flux was still not the largest contributor (due to differences in whether variable flux generated by one species is compensated by variation in the flux of another). These findings suggest that even if the magnitude and variation of species uptake and secretion fluxes across a set of microbiome samples are known (rather than just the abundances of species, which is the only measure usually assayed), metabolic interdependence between species would still make true contributor species challenging to identify.

Combined, the observations above highlight the complex relationship between species activity and measured metabolite concentrations, demonstrating the important role of both direct and indirect species interactions. This complex relationship, observed even in the idealized settings of our simulation model, is potentially markedly more complex than what is assumed by many microbiome-metabolite association-based analyses.

***Correlation analysis fails to detect true microbial contributors to metabolite variation***

Given our observations above, we next set out to comprehensively assess how well pairwise correlation analysis (commonly used for analyzing microbiome-metabolome data) can detect true taxonomic contributors to metabolite variance. Put differently, we evaluated the extent to which a correlation between species abundance and metabolite concentration across samples captures the true causative contribution of a species' metabolic activity to observed metabolite variation.

Following numerous microbiome-metabolome studies (14, 23, 28, 51), we considered identifying species-metabolite relationships as a classification task, aiming to identify for each metabolite the set of species that are primarily responsible for the variation observed in its concentration across samples. To this end, we defined *key contributor* species for each metabolite as those with a contribution value greater than 10% of the total positive contribution values. This resulted in a set of 83 species-metabolite key contributor pairs, representing true links between species activity and metabolite variation. On average, each metabolite had only 1.6 contributors (Figure S6), although 7.5 species on average had utilized or synthesized each metabolite at any point. 31.3% of these contributions occurred via synthesis reactions, 66.3% via utilization, and 2.4% (2 instances) via both processes. We then calculated the Spearman rank correlations between species abundances and metabolite concentrations across samples, and used a *p*-value threshold of 0.01 to define significant correlation between species and metabolites. This produced a set of 191 significant species-metabolite correlations, representing putative species-metabolite links. Scatter plots of these species-metabolite abundance relationships are shown for several example pairs in Figure S7.

Comparing this set of significant species-metabolite correlations to the set of species-metabolite key contributors clearly illustrated the difficulty of using univariate associations to infer mechanistic contributions (Figure 3). Indeed, of the 191 significant species-metabolite correlations, the vast majority (141) were false positives (corresponding to a positive predictive value of only 26.2%), and did not represent true contributor relationships (Figure 3A). Moreover, more than a third of these false positive species-metabolite pairs (51 out of 141) had *no* mechanistic connection; i.e., the species did not

ever use or produce the metabolite in question. Furthermore, for 12 variable metabolites (out of 52), none of the key contributors were successfully detected by a correlation analysis. The overall accuracy was somewhat higher (66.5%), reflecting the high number of non-contributors that are also not correlated. Using a stricter cutoff ($p < 0.0001$, equivalent to a Bonferroni-corrected value of 0.05) only improved the positive predictive value to 33% and the accuracy to 77.1%. Indeed, a ROC curve analysis (Figure 3B) produced an area under the curve of 0.72, and overall correlations and scaled contribution values were only weakly associated (Figure 3C), suggesting that these findings can only be partially mitigated by changing classification thresholds. Metabolites of different classes had generally similar correspondence between correlations and contributions (Figure 3D).

Notably, key contributors for purely microbially-produced metabolites were not identified more accurately than those for metabolites in the nutrient inflow (66% versus 67%), which is perhaps not surprising since we used a constant inflow across samples (but see also our analysis below with variable inflow). Moreover, the total variance in a metabolite was not associated with the accuracy or predictive value with which key contributors for that metabolite were identified (Spearman rho, $p > 0.1$). Across species, contributions were identified most accurately for *D. piger*, which had a relatively low number of contributions (Figures 3E and S5C), but the positive predictive value was nonetheless <50% for all species.

We obtained similar results across several variants of this analysis (Supplementary

Results, Figures S6, S8, and S9). To assess the impact of dynamic shifts over the duration of each simulation, we calculated an alternative set of contribution values based on the net steady-state metabolite flux rates at the final time point of each simulation, finding extremely similar results as for contributions to cumulative variation in concentration. We also evaluated the use of an alternative classification task, aiming to detect all microbes that affect variation in a given metabolite across samples regardless of whether their effects are ultimately reflected in the observed concentrations (i.e. those with large positive or negative contributions), again resulting in similar findings (Supplemental Results, Figure S6). Finally, we profiled the effects of model simulation parameters on correlation results, including the simulation length and the maximum enzymatic rate *V_max*, again finding minimal effects on contribution and correlation results (Supplementary Results, Figures S8-9).

***Species and metabolite properties explain discrepancies between correlations and contributions***

Our analysis above demonstrated that correlations between species abundances and metabolite concentrations can often be only poorly associated with true contribution of species to metabolite variation. We therefore next investigated the origins of such discrepancies. We examined whether individual metabolites or species are predisposed to produce a significant species-metabolite correlation when the species in fact does *not* contribute to that metabolite variation (i.e., false positives), or to mask such correlation when the species *does* in fact contribute to this metabolite variation (i.e., false negatives), and if so, what species and metabolite properties are linked to those outcomes.

To determine whether the identity of the species or metabolite in question can explain inaccurate identifications of key contributors, we used a regression-based analysis. Specifically, we considered all species-metabolite non-contributor pairs, and fitted a logistic regression model to predict whether a species-metabolite pair exhibited significant correlation (false positive), based on either species identities, metabolite identities, or both (Methods). We then compared these three models using a likelihood ratio test to assess whether species and/or metabolite identities are informative. We similarly considered all species-metabolite key contributor pairs separately, again fitting a logistic regression model based on species identities, metabolite identities, or both to predict whether a pair failed to exhibit significant correlation (false negative).

For non-contributors, we found that false positives can be explained largely by species identity (likelihood ratio test (LRT) for inclusion of species terms $p < 10^{-13}$). Incorporating both species and metabolite identities did not significantly improve the model (LRT for metabolite terms $p$=0.72). This finding suggests that false positives – correlations observed between species and metabolites to which they in fact did not contribute – are the outcome of interactions at the species level, regardless of the metabolite in question. This impact of strong interactions between dataset features on association test results has been described extensively in other data types (33, 34). Indeed, examining the 141 false positives identified above, we found that many can be explained by the relationships between the three dominant species in this community: *E. rectale*, *B. thetaiotaomicron*, and *B. ovatus*. These species competed strongly for carbon sources (and utilized their maximum allocation of sucrose, glucose, and fructose at nearly every step of the

simulation), and their abundances were therefore negatively correlated. As a result, metabolites that varied due to the activity of one of these species were also frequently correlated with the other two. In total, 32 false positive correlations paired one of these species with a metabolite for which another species in this trio was a key contributor. More generally, we found that the probability of a false positive correlation for a particular species and metabolite depended on the species' correlation with the true key contributors for that metabolite ($p$=0.006, Spearman rho between share of false positives and interspecies correlation; Figure 4A). Moreover, the maximum correlation each species had with any other species is a strong predictor of its overall specificity, which varies widely from 33.3% for *E. rectale* to 92% for *D. piger* (Spearman rho=-0.84, $p$=0.002). We also found that species identity was similarly predictive of whether a significantly correlated metabolite-species pair represented a true contributor versus a false positive (Supplementary Results).

In the case of key contributors, we found that false negative correlations can be explained largely by metabolite identity (LRT for metabolite terms $p$=0.002; although the species involved was also somewhat informative with LRT $p$=0.08). Put differently, a lack of correlation between the abundance of a key contributor species and the concentration of the metabolite to which it contributed was determined mainly by the nature of the metabolite in question. This lack of correlation between a given metabolite and its contributors could have resulted from competition or exchange of a metabolite between multiple species, such that none of the involved species end up strongly associated with the final outcome on their own. Indeed, across all metabolites, the average correlation

544 between a metabolite and its key contributors is negatively associated with its number of
545 key contributors (Spearman rho=-0.45, *p*=0.0008). The number of key contributors for
546 any metabolite was also thus negatively associated with the sensitivity of contributor
547 detection for that metabolite (Spearman rho=-0.48, *p*=0.0004; Figure 4B). We further
548 hypothesized that false negative outcomes might be more common for metabolites with
549 more or larger negative species contributions, since these, by definition, mask or
550 compensate for the activity of key contributor species. While all metabolites with a false
551 negative outcome did have at least one species with a negative contribution value, as
552 mentioned above, this was true for nearly all analyzed metabolites (50/52), and the
553 number of negative contributing species was not associated with the occurrence of a false
554 negative correlation (*p*=0.86, Wilcoxon rank sum test). Moreover, we also did not observe
555 any effect of the average concentration of a metabolite on the sensitivity and accuracy of
556 its detection via correlation analysis, nor of whether it is secreted, utilized, or cross-fed
557 (Figure 4C). In summary, our analysis suggests that the largest factor explaining whether
558 a metabolite's key contributor can be detected by a correlation analysis is simply whether
559 there are other community members (key contributors) that also impact the observed
560 concentration of that metabolite.

561

562 ***Environmental fluctuations in metabolite concentrations impact detection of key***
563 ***contributors***

564 Our analyses above all focused on a single simulated dataset in which the nutrient inflow
565 was constant across all samples, meaning that metabolite variation was fully governed
566 by microbial activity. However, in reality, metabolite variation can and does arise also from
567 non-microbial sources, potentially affecting both the landscape of key microbial

568 contributors and our ability to detect them via correlation-based analyses. To explore the
569 impact of environmental fluctuations, we therefore ran several sets of additional
570 simulations with varying degrees of nutrient fluctuation, designed to emulate a range of
571 levels of experimental diet control and variation in host absorption across the simulated
572 mouse gut communities. In these simulations, we maintained the same set of 61 initial
573 species compositions but added small amounts of stochastic noise to the nutrient inflow,
574 sampling inflow concentrations for each compound in each simulation from a normal
575 distribution with a mean equal to the compound's original inflow rate and a standard
576 deviation ranging from 0.5% to 10% of the mean in 8 increments (Methods). For each of
577 the resulting 8 datasets, we again calculated contribution values (with the added element
578 of the nutrient inflow as a potential contributor to variance), identified key contributors,
579 and compared them with the results of a correlation analysis.
580
581 Examining the obtained contribution values, we found, as expected, that variation in inflow
582 quantities can outweigh the variation in microbial fluxes, and that as the variation in inflow
583 increases, its contribution to metabolite variation increased at the expense of the
584 contributions of community members (Figure 5A). As a result, the number of key
585 contributions attributed to each species decreased for metabolites in the nutrient inflow
586 (Figure 5B). Interestingly, however, some species lost their contributions more gradually
587 than others, and in some cases even became key contributors for additional metabolites
588 (Figure 5B). For most metabolites, the relative ranking of species with the highest
589 contribution values was unchanged with increasing fluctuations (Supplementary Results).
590

We next examined how correlation-based detection of key microbial contributors was affected by these inflow fluctuations. We assigned each of the 52 metabolites in each of the 9 datasets (the original dataset with no inflow fluctuations and the 8 datasets with varying degree of fluctuations) to bins according to the level of contribution attributed to the inflow for this metabolite at that degree of fluctuation (see Methods). We then evaluated the performance of correlation analysis for each bin separately. The share of true key contributors naturally decreased rapidly with increasing environmental contribution, as did the number of significantly correlated species-metabolite pairs (Figure 5C). Importantly, however, the sensitivity of correlations decreased substantially with the level of contribution attributed to the inflow, but the specificity in fact increased from 67.7% to 92.3% (Figure 5D). This suggests that while environmental fluctuations disrupted the signal linking microbial species with the metabolites they impact, they also disrupted indirect associations between species and metabolites (false positives). Overall, however, the AUC did not change significantly with increasing environmental contribution (Figure S10A), and the positive predictive value is similarly relatively stable (and never rose higher than 37%). Interestingly, the detection of some metabolites not present in the inflow was also affected by inflow fluctuations in a similar manner (Supplementary Text, Figure S10B).

***Correlation analysis is similarly limited in simulations of more complex and diverse human gut microbiota***

Our results have illustrated consistent discrepancies between microbe-metabolite correlations and microbial contributions to metabolite variation in a model ten-species

614 community. We lastly addressed the question of whether these findings generalize to
615 more complex mammalian gut microbiota, communities with many times more taxa and
616 a more uneven distribution across individuals. To do so, we ran an additional set of
617 simulations emulating human gut microbiota transplanted into gnotobiotic mice. We first
618 mapped 16S rRNA sequence variants from the Human Microbiome Project (52) to the
619 genomes of the AGORA model collection at 97% sequence identity (40), and selected 57
620 samples with a successful mapping rate greater than 25% relative abundance. The total
621 share of mapped reads averaged 36.7% across these samples, with a maximum of
622 73.5%. Despite this variation, mapped reads displayed features typical of Western gut
623 microbiomes, including a predominance of Bacteroidetes and Firmicutes phyla along with
624 varying lower abundances of Actinobacteria and Proteobacteria (Figure 6A). The number
625 of species identified in each sample ranged from 23 to 62, with a median of 42. We ran a
626 simulation based on each sample by setting the initial species relative abundances
627 according to the relative abundances of mapped reads, while maintaining the same
628 physical parameters as previous simulations (see Methods for additional details). We
629 used nutrient inflow quantities with 1% standard deviation between samples. Initial
630 species compositions displayed characteristic shifts in abundance over the simulation
631 time course (Figure S11A). Metabolites were also highly variable, with a median
632 coefficient of variation of 71% across 222 metabolites (Figure S11B). We calculated
633 contribution values for this dataset, finding a smaller share of key contributions (only 392
634 out of 29,082 possible species-metabolite pairs). Only 35.1% of species (46 out of 131)
635 were identified as key contributors to any metabolite. The genera with the most
636 contributions were *Bacteroides*, *Ruminococcus*, and *Enterobacter*, which were also three

637 of the four most abundant genera in the final dataset (Figure 6B).

638

639 In this noisier and more layered dataset, only a small share of species-metabolite pairs
640 was significantly correlated. In order to fairly compare with the previous dataset while
641 accounting for the larger number of hypothesis tests, we defined significance based on
642 an equivalent Benjamini-Hochberg estimated false discovery rate (0.027) as the $p < 0.01$
643 cutoff used for the previous dataset. 2.2% of species-metabolite pairs displayed
644 significant correlations at this cutoff ($p < 0.00058$). This level of correlation is comparable
645 to a recent microbiome-metabolome study of the colon of healthy humans (51), in which
646 1.4% of OTU-metabolite pairs displayed Spearman correlation coefficients of the same
647 effect size. In our dataset, correlation analysis detected contributors with high specificity
648 (98.4%), and an area under the ROC curve of 0.89. However, the positive predictive value
649 was still only 29.0%, rising as high as 57% with a significance cutoff of $p < 10^{-10}$. We
650 compared these classification results with the original dataset, finding that despite the
651 difference in overall AUC, sensitivity, sensitivity, and predictive value are similar or worse
652 for the two datasets at commonly used FDR thresholds between 0.1 and 0.01 (Figure
653 6C), and sensitivity and predictive value are both highly dependent on the choice of
654 significance threshold. As in the ten-species dataset, a large share of false positive
655 species-metabolite pairs (65.4%, 291 out of 445) also involved species with no capacity
656 to impact the metabolite in question.

657

658 The outcomes of correlation analysis were influenced by the same factors as observed in
659 the model community dataset, but also by several additional characteristics. False

660 positive classifications were, again, driven by interspecies covariance: Species
661 significantly correlated (at 10% FDR) with a true key contributor for a metabolite were
662 13.6 times more likely to have a false positive correlation with that metabolite than species
663 with no such link ($p < 10^{-16}$). Notably, the false positive rate of a given species was also
664 substantially affected by its prevalence: the number of samples in which a species was
665 present was negatively associated with its specificity (Spearman rho = -0.57, $p$=0.002,
666 Figure S11C), among species with at least 3 key contributions. In other words, widely
667 prevalent species were more prone to false positive correlations than rarer species.
668
669 False negative contributions were again influenced by properties of both metabolites and
670 species. As in the ten-species dataset, species contributions to metabolites with more
671 than one key contributor were 5.2 times more likely to not be correlated than those that
672 were the sole key contribution for a metabolite ($p < 10^{-10}$, Fisher exact test). In this dataset,
673 an elevated share of these metabolites with multiple key contributors were cross-fed
674 between different species ($p$=0.00007, Fisher exact test), and correspondingly, key
675 contributors for cross-fed metabolites were also 1.6 times less likely to be significantly
676 correlated ($p$=0.02). Both cross-feeding and false negative outcomes occur variably
677 across metabolite classes, with nucleotide metabolites having the highest rates of both
678 phenomena (Figure S11D). Taken overall, our simulations and analysis of this realistic
679 microbiota simulation demonstrates that correlation analysis can have greater utility in a
680 microbial community dataset with greater complexity and variability, but the results are
681 again strongly influenced by properties of individual metabolites and species.
682

# Discussion: Insights and implications for microbiome-metabolome analyses

Above, we have investigated the ability of correlation-based analyses to detect key microbial contributors responsible for variation in metabolite concentrations across samples. Our findings suggest that microbe-metabolite correlation analysis may be a useful approach for exploratory analyses, but they highlight some of the limitations and caveats of such microbiome-metabolome studies and identify several factors that impact the relationship between community composition and metabolite concentrations. Below, we elaborate on a set of practical conclusions and their implications for the analysis and interpretation of microbiome-metabolome studies.

**Association-based analyses of microbiome-metabolome assays have low predictive value for detecting direct species-metabolite relationships and require conservative interpretation.** Microbiome-metabolome association studies have been previously proposed as a powerful tool for the identification of causal mechanisms of microbiome metabolism (53), and indeed, such studies often present detected associations as evidence for mechanistic relationships (11, 27, 29). However, our analysis suggested that the positive predictive value of significant species-metabolite correlations for identifying true microbial contributors can be extremely low: less than 50% across all settings, as low as 10% in the context of large environmental fluctuations, and 29% in simulations based directly on human gut composition. Recent experimental

studies pairing microbiome-metabolite correlation analysis with *in vitro* monoculture validations have similarly anecdotally observed many false positive correlations (32). Additionally, given the somewhat low sensitivity observed in our analysis, a lack of association is not necessarily sufficient to reject a hypothesis that a particular microbial taxon impacts a particular metabolite. The choice of correlation threshold should therefore be chosen carefully, taking into account the complexity of the community and the environmental context. In general, identified correlations between microbial taxa and metabolites should be interpreted very conservatively and used mostly to prioritize microbe-metabolite relationships for follow-up validation studies (e.g., via culture-based studies or germ-free model organism colonization). One potential approach for improving the predictive value of such correlation-based analyses is to examine whether they replicate across multiple conditions. Indeed, we found that a correlation does provide stronger evidence for a contributor relationship if it persists across different contexts. Across our 9 simulated datasets with varied environmental fluctuations, the 43 species-metabolite pairs that were significantly correlated in every dataset were 2.1 times more likely to denote true key contributor relationships than other significant correlations (Fisher exact test, $p$=0.05), although their positive predictive value was still relatively low (39.5%). Of the limited number of significant correlations shared between our original and HMP-based datasets ($n$=5), all were false positives in both datasets, reiterating the need for caution.

**The predictive power of correlation-based analysis is species-, metabolite-, and context- dependent.** In our datasets, metabolites varied widely in both contribution profiles and in their detectability via correlation analysis. In particular, the key contributors

729 for metabolites acted upon by fewer species, and potentially those that are not exchanged
730 between different species, were identified more readily. Moreover, in our simulations of
731 human gut communities, contributions by less prevalent species were identified much
732 more accurately than those by widely-found species, indicating that hypotheses based on
733 associations of rarer species should potentially be prioritized. Correlation analysis may
734 thus identify microbes involved in specialized secondary metabolic processes (e.g.
735 products of complex biosynthetic pathways) more readily than those involved in more
736 widespread processes.. Therefore, correlation-based approaches may be more
737 informative for analyzing compounds that are specific to a small number of rare taxa, but
738 accurate dissection of the taxa controlling variation in widely-trafficked metabolites may
739 require more detailed analysis and experimentation. Similarly, we found that species-
740 metabolite correlations for species that are strongly associated with other taxa (e.g., those
741 with tight interactions with other community members) are often spurious, suggesting that
742 such correlations should be regarded less confidently.

743

744 **External metabolic fluctuations can strongly impact the detection of microbial**
745 **contributions.** Our analysis of the impact of environmental fluctuations suggested that
746 the presence of environmental variability from a diverse set of samples could in fact
747 increase correlation specificity. We also found that the sensitivity of correlation analysis
748 rapidly decreased with increasing environmental fluctuations (from 60% to 9%). These
749 observations suggest that while a tightly controlled environment (e.g., identical diets) is
750 intuitively expected to increase the strength of microbiome-metabolome studies, its value
751 depends on the study priorities. Specifically, if the goal is to identify clear-cut microbial
752 drivers of healthy- and disease-associated metabolite shifts, stochastic variation in

753 nutrient availability could be beneficial as it may reduce the rate of false positive
754 associations. In contrast, for studies searching for a particular microbial taxon's
755 involvement in a particular process (e.g. aiming to determine whether an ingested
756 probiotic impacts aspects of gut metabolism), a more controlled environment may be
757 favorable. It should, however, be noted that our findings were based on environmental
758 fluctuations that were uniform and independent, which may not hold for real-life
759 environmental fluctuations such as diet variation. It is also worth noting that in our
760 simulations, microbial fluxes for some environmental metabolites could be drowned out
761 by as little as 0.5% variation in nutrient inflow quantities, while others still had substantial
762 microbial contributions even with 10% variation in inflow. When interpreting an observed
763 association, the scale of possible microbial variation relative to external variation should
764 therefore be taken into account.
765
766 **Mechanistic reference information can improve the predictive power of**
767 **microbiome-metabolome studies.** In our simulated dataset, 36% of the false positive
768 correlations occurred between a metabolite and a species that was in fact not capable of
769 uptaking or secreting that metabolite. Ruling out such falsely detected links would
770 substantially improve the positive predictive value of a correlation-based analysis. One
771 approach for doing so is by utilizing genomic information, which can be obtained or
772 predicted for many microbial taxa (54). By coupling such genomic information with
773 metabolic databases such as KEGG or MetaCyc (55, 56), researchers can filter out
774 correlation-based links that are likely not feasible causative relationships. Further
775 improvement can be obtained by integrating such reference information directly into the

776 analysis. Indeed, we previously introduced a computational framework, termed MIMOSA
777 (57), that utilizes a simple community-wide metabolic model to assess whether measured
778 metabolite variation is consistent with shifts in community metabolic potential, and to
779 identify potential contributing taxa. MIMOSA has been applied to varied host-associated
780 microbiomes from varied body sites and from human and mouse hosts (12, 58, 59).
781 Applying MIMOSA to the simulated ten-species dataset analyzed above (Methods), we
782 found that it indeed identified key contributors significantly more accurately than a
783 correlation-based analysis, with an AUC of 0.89 (Figure 6). Notably, in this analysis, we
784 assumed MIMOSA has access to the correct set of metabolic reactions possessed by
785 each species. Using standard less-complete information obtained directly from the KEGG
786 database (as done regularly when using this tool) reduced the number of metabolites that
787 could be analyzed from 52 to 39, with improved specificity (96%) and positive predictive
788 value (61%) and an ultimately comparable AUC (0.74). Combined, these findings suggest
789 that reference model-based approaches can provide stronger evidence for mechanistic
790 relationships than strictly correlation-based methods, but their use depends on complete
791 and high-quality metabolic reference databases.
792

# 793 Future opportunities and challenges

794 Microbiome-metabolome studies have an important role in microbial ecology research.
795 They specifically have great potential to dissect the metabolic interactions of complex
796 microbial communities, and to unify “top down” and “bottom up” microbiome research
797 approaches by providing mechanistic information at a systems level. Moreover, from a

798 translational perspective, microbiome-metabolome studies can inform efforts to design
799 targeted therapies to alter specific microbial or metabolic features of a community (13).
800 Such interventions require first identifying putative targets, which in many cases may
801 entail identifying the key contributor species that drive observed shifts in a particular
802 beneficial or detrimental metabolic phenotype.
803
804 Importantly, while we show here that a correlation-based analysis may be limited in its
805 ability to identify these key microbe-metabolite links, this does not necessarily imply an
806 inherent limitation of microbiome-metabolome data. For example, analyzing our data, we
807 found that species abundance is in fact a very good proxy for metabolic activity (median
808 correlation of 0.996 between abundance and flux for all species-metabolite pairs),
809 meaning that the variance in total species abundance drastically outweighs the individual-
810 level variance in flux rates. When we further examined whether false negative
811 associations in our original dataset stem from a disconnect between the abundance of a
812 species and its metabolite uptake or secretion rates, we identified only 2 undetected key
813 contributor pairs that could be explained by such a discrepancy. This analysis suggests
814 that taxonomic abundance data is sufficient to explain and model community metabolic
815 variation to great extent, despite common concerns about potential discrepancies
816 between community composition and function. It also suggests that metatranscriptomic
817 expression data may not provide much additional value for this purpose, as other studies
818 have indicated (54, 60, 61).
819

Given the increasing prevalence of microbiome-metabolome studies, their promise, and the caveats of association-based research discussed above, further development of computational and statistical methods for analyzing such datasets is clearly needed. Possible directions include the use of multi-species dynamic metabolic models that can replicate experimental observations (62), multivariate approaches for deconvolving interactions between species and the environment (63, 64), and probabilistic methods that can integrate prior information while allowing for other unknown mechanisms (31, 65). The conceptual framework of taxon-metabolite contributions, and the use of dynamic simulations demonstrated here, can both inform the future development and evaluation of such methods.

There is also a continued need for gold standards to evaluate new methods. This study is only a first step in that direction and has analyzed one specific type of research question: identifying microbial taxa directly responsible for variation in metabolite concentrations between samples in a cross-sectional study design. Although this focus describes many recent microbiome-metabolome studies, other studies may address a wide range of complementary research questions, and correspondingly, the desired "ground truth" can take different forms. Moreover, depending on the objective, an alternative definition of a taxon-metabolite relationship may be required. For example, it may be valuable to identify key contributors that act via alternative mechanisms, such as by modifying substrate availability or environmental conditions (for example (66)), or to distinguish metabolite variation arising in response to a perturbation from variation due to differences in steady-state metabolism between communities. Additionally, our findings

rely on an *in silico* system that may not capture many aspects of community ecology and metabolism, and it is possible that the predictive value of correlation analysis, as well as of other analytical methods, differs fundamentally in this system as compared to true biological systems. Further studies should also consider additional variables such as community diversity, sample size, measurement error, and other types of environmental variation. Ongoing technology developments in mass spectrometry and stable isotope probing will ideally enable future evaluation analyses using experimental, quantitative, species-specific community flux data to define key microbial contributors (67, 68). Such evaluations can also take advantage of datasets comparing community microbiome-metabolome data with *in vitro* monoculture or mono-colonization data (32, 35, 36).

Ultimately, much remains to be learned about the many processes through which complex microbial communities shape their environment. The first major call for the application of metabolomics to microbiome research, published 10 years ago (69), noted that new methods will be necessary to integrate genomic and metabolic data and inform the prediction of community metabolic properties from metagenomes. Now that microbiome-metabolome datasets are widely available, ongoing development of analysis methods for these studies has great potential to generate new knowledge. Moreover, future work in this area stands to benefit from the utility of dynamic, multiscale metabolic modeling. Detailed mechanistic simulations are used widely in astronomy, climate science, and other fields to make methodological choices and assess possible experimental outcomes when ground truth measurements are unavailable or difficult to obtain (70, 71). An analogous strategy in microbiome research may be similarly fruitful.

# Methods

### *Derivation of species contributors to variation*

We derived an expression representing the contribution of each species to the variance in the concentration of each metabolite. While we describe this calculation in terms of species, a similar calculation could be done at the level of phyla, strains, or any grouping of the community for which metabolite secretion and uptake fluxes are available.

The concentration of a given metabolite *M* at the end of a single simulation run is a function of the uptake and secretion fluxes (responding to the species' degradation and synthesis activities) of the *n* species, the environmental inflow over all time steps $m_{in}$, and the dilution $m_{out}$ out of the chemostat over all time steps:

$$M = \sum_{i=1}^{n} m_i + m_{in} - m_{out}$$

The value of $m_{out}$ at a given time step *t* is the product of the dilution rate *D* and the metabolite concentration at the previous time point (see above). This fact can be used to express $m_{out}$ in terms of all the previously recorded environmental inflow and microbial activities. The metabolite concentration at any time point *t*, *M(t)*, is then equal to:

886 $$M(t) = \sum_{k=1}^{t-1}\left[(1-D)^{t-k-1}\sum_{i=1}^{n} m_{ik}\right] + m_{in}\sum_{k=1}^{t-1}(1-D)^{k},$$

887

888 where $m_{ik}$ represents the activity of species *i* at a single time point *k.* We can then ignore
889 dilution outflow by replacing each activity value $m_i$ in the final concentration calculation
890 above with a value corrected for the mitigating effect of chemostat dilution over the course
891 of the simulation up to time *t*, defined here as $m_i$*. $m_i$* represents the total amount of a
892 compound secreted or uptaken by species *i*, minus the share of that quantity that is
893 eventually diluted out over the course of the simulation.

894

895 $$m_i^* = \sum_{k=1}^{t-1}(1-D)^{t-k-1} m_{ik},$$

896 and thus,

897 $$M = m_{in} + \sum_{i=1}^{n} m_i^*$$

898

899 In this work, we refer to “environmental fluctuations” as the effect of the independently
900 parameterized nutrient inflow, $m_{in}$, and where not otherwise specified we use $m_i$ to imply
901 $m_i$*, a species activity quantity that accounts for the corresponding subsequent dilution
902 out of the system.

903

904 Using the expression above, *var(M)* can then be clearly expressed as a sum of correlated
905 environmental and microbial random variables:

906 $$var(M) = \sum_{i=1}^{n}\sum_{j=1}^{n} cov(m_i, m_j) + \sum_{i=1}^{n} cov(m_i, m_{env})$$

907 $$= \sum_{j=1}^{n} var(m_j) + var(m_{env}) + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} cov(m_i, m_j) + 2\sum_{i=1}^{n} cov(m_i, m_{env})$$

908

909 This expression can then be partitioned additively into *n+1* terms representing the
910 contribution of each microbial species and of fluctuations in the environmental nutrient
911 inflow.

912

913 $$c_i = \sum_{j=1}^{n} cov(m_i, m_j) + cov(m_i, m_{env}) = var(m_i) + \sum_{j\neq i} cov(m_i, m_j) + cov(m_i, m_{env})$$

914

915 ***Multi-species Dynamic Flux Balance Analysis modeling***

916 In this study, we simulated the growth and metabolism of a community of 10
917 representative gut species that was previously explored experimentally (47). We
918 specifically utilized a previously introduced multi-scale framework for modeling the
919 dynamics and metabolism of multiple microbial species in a well-mixed shared nutrient
920 environment (44, 72). This framework assumes that each species in the community aims
921 to maximize its own growth on a short time scale given available nutrients, and uses Flux
922 Balance Analysis to predict the growth and metabolic activity of each species at this short
923 time scale (46). The shared environment is then iteratively updated based on the species'
924 predicted growth, uptake, and secretion rates, such that metabolic interactions are

925 mediated via the environment as a natural byproduct of species activities, rather than
926 being explicitly modeled (45).
927
928 We used genome-scale metabolic model reconstructions of the 10 community members
929 from the AGORA collection version 1.01 (40), which have been consistently curated to
930 remove or modify thermodynamically unfavorable reactions, remove futile cycles, and
931 confirm growth in anaerobic environments on expected carbon sources, with additional
932 curation for several biosynthesis pathways. The COBRA toolbox was used to convert
933 each AGORA model to MATLAB format (73). The growth and metabolism of the 10-
934 species community were simulated in a chemostat setting in 15-minute time intervals. We
935 set the chemostat volume to be approximately equal to a mouse gut (0.00134 liter (74)).
936 We similarly set metabolite inflows to emulate the macronutrient and micronutrient
937 quantities in a corn-based mouse chow (47) (provided in Supplementary Data 1).
938
939 The simulations were performed following a previously introduced procedure (44),
940 repeated for each time step $t_n$: First, the maximum uptake rate for all metabolites by all
941 species, denoted as $v_{jk}$ for metabolite $j$ and species $k$, were calculated based on
942 Michaelis-Menten single-substrate kinetics, with assumed universal values for maximum
943 rate $V_{max}$ and transporter affinity $K_m$ for all metabolites (provided in Supplementary Data
944 1). $v_{jk}$ was further constrained based on an allocation of the metabolite's environmental
945 concentration to each species in proportion with its biomass. Then, the steady state
946 reaction fluxes for each species $k$ at time point $t_n$ were determined by maximizing the
947 growth rate $\mu_k$, within the obtained constraints on environmental metabolite uptake. To

obtain a single and consistent flux solution for each species, the total flux activity for each species (i.e., the sum of absolute fluxes given the predicted optimal growth rate) was minimized, under the assumption that organisms prefer to operate their metabolism with minimal enzymatic cost (75). The optimal flux solutions were solved using linear programming with GLPK (www.gnu.org/software/glpk). With the resulting flux and growth rate information, the total biomass of each species $k$, $bio_k(t_n)$, was updated for the next time point $t_{n+1}$, using a standard exponential growth function incorporating dilution:

$$bio_k(t_{n+1}) = bio_k(t_n)e^{\mu_k \Delta t} - bio_k(t_n)D\Delta t,$$

where $D$ is the dilution rate. We set $D$ to 0.0472 per hour, in order to obtain community growth rates consistent with the observed average growth rate of the three most abundant species growing under 47 different carbon conditions (76). The total amount of uptake or secretion for each species $k$ and metabolite $j$ over a single time step was then calculated as previously derived (44):

$$m_{FBA}^{jk}(t_n) = \frac{v_{jk}}{\mu_k} * \ bio_k(t_n)(e^{\mu_k \Delta t} - 1),$$

where $v_{jk}$ is the rate of uptake or secretion specified by the FBA solution for that species and metabolite at that time point, $\mu_k$ is the species growth rate, $bio_k(t_n)$ is the species abundance, and $\Delta t$ is the size of the time step. Finally, combining the flux solutions of all species, nutrient inflow, and dilution, along with the steady state assumption of no

intracellular metabolite accumulation, the concentration of a given metabolite in the shared nutrient environment at the next time point, $M_j(t_{n+1})$ can be updated as:

$$M_j(t_{n+1}) = M_j(t_n) + m^j_{FBA}(t_n) + m^j_{in}\Delta t - M_j(t_n)D\Delta t,$$

where $m^j_{FBA}(t_n)$ is the metabolic impact from all species considering their abundance and their uptake and secretion rates of metabolite *j*, and $m^j_{in}$ is the inflow rate of metabolite *j*. This process of calculating uptake rates, Flux Balance Analysis solutions, and updated metabolite concentrations was then repeated iteratively for the duration of the simulation.

Each simulation was run for a period of 144 hours or 576 time steps. This time period was long enough for most simulation runs to approach a steady state composition: specifically, in >65% of the simulations analyzed in our study, the change in abundance in any species over the final 3 hours was less than 0.01% of the carrying capacity (see below), and all had no changes greater than 0.3% of the capacity over that period. The concentrations of species and metabolites, the species growth rates, and the solved rates of all reactions for each species (including uptake and secretion) were recorded in each step of each simulation and used for subsequent analyses (Supplementary Data 1 and 2).

***Simulation initialization parameters***

We fixed the initial total abundances of microbes to the carrying capacity for this system and media, which was estimated to be 0.433 units of biomass. This capacity was calculated as the average final total abundance from a set of simulations with varying

compositions and low initial abundances. We then varied the relative abundances, increasing the abundance of one species at a time at the expense of all other species equally. Specifically, for each species, we ran simulations in which the ratio of that species' initial abundance relative to all other species was 2, 3, 4.5, 6, 9, and 13 times (equating to a range in relative abundance of 10% to 60% for each species). This resulted in a total of 61 simulation runs (one with all species starting at equal abundance and 6 with increased abundance of each species). We chose this sample size to approximately represent the sample sizes of published cross-sectional microbiome-metabolome association studies (14, 16). We set the initial inflow concentrations to the amount that would dilute in over one hour under the calculated inflow rates.

***Calculation of contribution values for variable metabolites***

We calculated contribution values for all metabolites with variance in concentration above the 25$^{th}$ percentile. We chose this threshold in order to include as many metabolites as possible while excluding those that only varied at all in fewer than half of the simulation runs, or whose variation would be subject to potential numerical errors.

***Comparison with Shapley values***

We implemented an approximate Shapley value algorithm (43) as an alternative strategy to calculate contributions for the simulated dataset. Briefly, 15,000 random orderings of the 10 species were randomly generated. For each ordering, the variance in metabolite activity is calculated for subsets of size 1 to 10, adding in species according to the specified ordering. The difference in variance as a given species is added to the subset,

1016 denoting the *marginal* contribution of that species to variation, is recorded. The average
1017 marginal contribution across all orderings for each species is then defined as its
1018 contribution to variance.

1019

1020 ***Species-metabolite correlation analysis***

1021 We calculated Spearman correlations between absolute species abundances (quantified
1022 as total biomass) and concentrations of variable metabolites. We used absolute
1023 abundances in order to evaluate the relationships between species and metabolites under
1024 the hypothetically best possible measurements of both data types. We also compared
1025 correlation results using relative abundances and found very minimal differences in the
1026 main simulation dataset: only 7 species-metabolite pairs (1.3%) are significantly
1027 correlated using absolute abundances but not relative, and only 4 pairs (0.8%) are
1028 correlated using relative abundances but not absolute.

1029

1030 We used a *p*-value threshold of 0.01 to classify "significant" associations for binary
1031 comparisons. For interpretability, we refer to *p*-values not corrected for multiple
1032 hypothesis testing, since the number of tests remained constant across nearly all of our
1033 analyses (520 possible species-metabolite pairs). The 0.01 threshold we use to define
1034 significantly correlated pairs is equivalent to a Benjamini-Hochberg corrected false
1035 discovery threshold of 0.027, calculated using the R function *p.adjust* (77).

1036

1037 ***Logistic regression modeling of correlation outcomes***

1038 We used logistic regression models to identify factors that can be used to predict whether

a non-contributing species-metabolite pair displays a significant correlation (false positive), and whether a key contributor species-metabolite pair fails to be correlated (false negative). We used the *glm* function in R to fit models of the log odds of whether a non-contributing species is correlated with its corresponding metabolite (false positive or true negative), using as predictors grouped indicator values for species and metabolite identities. We separately fit another set of logistic regression models to predict whether a key contributor species is correlated (true positive or false negative), with the same predictors. Models were compared using likelihood ratio tests using the *anova* function in R.

***Simulations with varied inflow quantities***

We ran 8 additional sets of simulations with the same set of 61 different initial species compositions but with varying degrees of inflow fluctuations. Specifically, the nutrient inflow quantities were sampled independently from a normal distribution, with a mean of the original inflow concentration and the standard deviation equal to a set percent of the mean. The 8 levels of deviation were 0.5%, 1%, 2%, 3%, 4%, 5%, 8%, or 10%. In the comparison of correlation results across samples, we evaluated the same set of 52 variable metabolites as for the original dataset for consistency, although given the added noise, additional metabolites met the same variance cutoff we used to define variable metabolites.

To evaluate correlation performance as a function of increasing environmental contribution, we binned the 38 analyzed inflow metabolites across the 8 datasets based

on the size of the environmental contribution to variance for the metabolite in that dataset. In other words, metabolites in any dataset with an environmental contribution greater than 0 but less than 10% of the total positive variance contributions were binned into a single category, those with an environmental contribution between 10% and 20% were binned into the next category, and so on. We analyzed the 52 metabolites in the original constant-environment dataset as a separate category, and did the same for the 14 non-inflow metabolites in each of the 8 environmentally-varying datasets.

Confidence intervals for AUC values were calculated using the *pROC* package in R (78), using a bootstrap method with 500 resamplings.

***Simulations of Human Microbiome Project-based microbiota***

To simulate more complex gut microbiota, we downloaded the 16S rRNA sequence variant abundance tables from the Human Microbiome Project (52), processed with deblur (79), from Qiita (80). We also downloaded ribosomal RNA sequences for all of the 818 genomes corresponding with AGORA v1.0.2 models from NCBI RefSeq and GenBank using the biomartr R package (81). We used *vsearch* version 2.8.1 (82) to map the HMP sequences to the AGORA ribosomal sequences with 97% identity, with the max_rejects parameter set to 0 in order to obtain the highest identity match for each sequence variant. We chose to model a subset of 57 samples for which at least 25% of their total read counts successfully mapped to an AGORA genome. We normalized species abundances based on the 16S rRNA copy number of the corresponding genome, and initialized 57 simulations with the starting relative abundances determined based on

the AGORA-mapped relative abundances of these samples. We updated the nutrient inflow to enable growth by most models. We assessed whether the additional of each individual metabolite to the original nutrient inflow had a growth-promoting effect on any species, specifying proportions similar to the average European diet in the Virtual Metabolic Human database where possible (83). Metabolites that promoted growth in at least one species were retained in the revised nutrient inflow, and the process of testing for increased growth with the addition of any single metabolite was repeated. After two rounds of adding metabolites to the inflow, 15 models, representing 3.4% of the total normalized abundance across all samples, still displayed zero growth. We removed these from the simulations and used the final updated nutrient inflow with the 131 remaining models. All other simulation parameters were the same as for the original 10-species community simulations. When analyzing the role of interspecies correlation in this dataset, we excluded species that appear in fewer than 4 samples.

***Application of MIMOSA to simulated data and comparison with correlation analysis***

We applied MIMOSA v1.0.2 (github.com/borenstein-lab/MIMOSA) (57) to the obtained set of metabolite and species abundances. To construct the community metabolic network model required by MIMOSA, we merged the 10 species-level models used in the simulations into a single stoichiometric matrix. If a reversible reaction only ever proceeded in a single direction in any simulation, we encoded it as non-reversible. To apply the KEGG-based version of MIMOSA, we converted the model metabolite IDs to KEGG IDs (56), downloaded KEGG Orthology gene annotations for the 10 modeled species from

1108 the IMG/M database (84), and ran a MIMOSA analysis using the KEGG metabolic
1109 network model encoded in *reaction_mapformula.lst* (KEGG version downloaded 2-2018).
1110
1111 ***Code and data availability***
1112 Code for all the analyses presented in this study is available online in the form of R
1113 notebooks at https://github.com/borenstein-lab/microbiome-metabolome-evaluation. The
1114 code and media files for performing dynamic FBA co-culture simulations is available from
1115 https://borensteinlab.com/download.html. All data generated and analyzed in this study
1116 and displayed in the figures are included in Supplementary Data 1 through 4.
1117
1118

# Author contributions

C.N. and E.B. designed the study and wrote the paper. C.N. performed the analysis. H.C.C. and C.P.M. contributed to the multi-species metabolic modeling simulations. All authors read and approved the paper.

# Acknowledgements

C.N. was supported in part by a National Science Foundation (NSF) IGERT DGE-1258485 fellowship. C.P.M. was funded by NHGRI grant T32 HG000035. This work was supported in part by NIH New Innovator Award DP2 AT007802–01 and NIH grant 1R01GM124312–01 to E.B.

# References

1. Hazen TC, Dubinsky EA, DeSantis TZ, Andersen GL, Piceno YM, Singh N, Jansson JK, Probst A, Borglin SE, Fortney JL, Stringfellow WT, Bill M, Conrad ME, Tom LM, Chavarria KL, Alusi TR, Lamendella R, Joyner DC, Spier C, Baelum J, Auer M, Zemla ML, Chakraborty R, Sonnenthal EL, D'haeseleer P, Holman H-YN, Osman S, Lu Z, Van Nostrand JD, Deng Y, Zhou J, Mason OU. 2010. Deep-Sea Oil Plume Enriches Indigenous Oil-Degrading Bacteria. Science 330:204–208.

2. Shi W, Moon C, Leahy S, Kang D, Froula J, Kittelmann S, Fan C, Deutsch S, Gagic D, Seedorf H, Kelly W, Atua R, Sang C, Soni P, Li D, Pinares-Patiño C, McEwan J, Janssen P, Chen F, Visel A, Wang Z, Attwood G, Rubin E. 2014. Methane yield phenotypes linked to differential gene expression in the sheep rumen microbiome. Genome Res gr.168245.113.

3. Koeth RA, Wang Z, Levison BS, Buffa JA, Org E, Sheehy BT, Britt EB, Fu X, Wu Y, Li L, Smith JD, DiDonato JA, Chen J, Li H, Wu GD, Lewis JD, Warrier M, Brown JM, Krauss RM, Tang WHW, Bushman FD, Lusis AJ, Hazen SL. 2013. Intestinal microbiota metabolism of l-carnitine, a nutrient in red meat, promotes atherosclerosis. Nat Med 19:576–585.

4. Hsiao EY, McBride SW, Hsien S, Sharon G, Hyde ER, McCue T, Codelli JA, Chow J, Reisman SE, Petrosino JF, Patterson PH, Mazmanian SK. 2013. Microbiota Modulate Behavioral and Physiological Abnormalities Associated with Neurodevelopmental Disorders. Cell 155:1451–1463.

5. Dumas M-E, Barton RH, Toye A, Cloarec O, Blancher C, Rothwell A, Fearnside J, Tatoud R, Blanc V, Lindon JC, Mitchell SC, Holmes E, McCarthy MI, Scott J, Gauguier D, Nicholson JK. 2006. Metabolic profiling reveals a contribution of gut microbiota to fatty liver phenotype in insulin-resistant mice. Proc Natl Acad Sci 103:12511–12516.

6. Louis P, Hold GL, Flint HJ. 2014. The gut microbiota, bacterial metabolites and colorectal cancer. Nat Rev Microbiol 12:661–672.

7. Wlodarska M, Luo C, Kolde R, d'Hennezel E, Annand JW, Heim CE, Krastel P, Schmitt EK, Omar AS, Creasey EA, Garner AL, Mohammadi S, O'Connell DJ, Abubucker S, Arthur TD, Franzosa EA, Huttenhower C, Murphy LO, Haiser HJ, Vlamakis H, Porter JA, Xavier RJ. 2017. Indoleacrylic Acid Produced by Commensal Peptostreptococcus Species Suppresses Inflammation. Cell Host Microbe 22:25-37.e6.

8. Ferreyra JA, Wu KJ, Hryckowian AJ, Bouley DM, Weimer BC, Sonnenburg JL. 2014. Gut Microbiota-Produced Succinate Promotes C. difficile Infection after Antibiotic Treatment or Motility Disturbance. Cell Host Microbe 16:770–777.

9. Rath S, Heidrich B, Pieper DH, Vital M. 2017. Uncovering the trimethylamine-producing bacteria of the human gut microbiota. Microbiome 5.

10. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, Ling AV, Devlin AS, Varma Y, Fischbach MA, Biddinger SB, Dutton RJ, Turnbaugh PJ.

2013. Diet rapidly and reproducibly alters the human gut microbiome. Nature 505:559–563.

11. De Filippis F, Pellegrini N, Vannini L, Jeffery IB, La Storia A, Laghi L, Serrazanetti DI, Di Cagno R, Ferrocino I, Lazzi C, Turroni S, Cocolin L, Brigidi P, Neviani E, Gobbetti M, O'Toole PW, Ercolini D. 2015. High-level adherence to a Mediterranean diet beneficially impacts the gut microbiota and associated metabolome. Gut gutjnl-2015-309957.

12. Snijders AM, Langley SA, Kim Y-M, Brislawn CJ, Noecker C, Zink EM, Fansler SJ, Casey CP, Miller DR, Huang Y, Karpen GH, Celniker SE, Brown JB, Borenstein E, Jansson JK, Metz TO, Mao J-H. 2016. Influence of early life exposure, host genetics and diet on the mouse gut microbiome and metabolome. Nat Microbiol 2:16221.

13. Shaffer M, Armstrong AJS, Phelan VV, Reisdorph N, Lozupone CA. 2017. Microbiome and metabolome data integration provides insight into health and disease. Transl Res.

14. Srinivasan S, Morgan MT, Fiedler TL, Djukovic D, Hoffman NG, Raftery D, Marrazzo JM, Fredricks DN. 2015. Metabolic Signatures of Bacterial Vaginosis. mBio 6:e00204-15.

15. Theriot CM, Koenigsknecht MJ, Carlson Jr PE, Hatton GE, Nelson AM, Li B, Huffnagle GB, Z. Li J, Young VB. 2014. Antibiotic-induced shifts in the mouse gut

1191 microbiome and metabolome increase susceptibility to Clostridium difficile
1192 infection. Nat Commun 5.

1193 16. Califf KJ, Schwarzberg-Lipson K, Garg N, Gibbons SM, Caporaso JG, Slots J,
1194 Cohen C, Dorrestein PC, Kelley ST. 2017. Multi-omics Analysis of Periodontal
1195 Pocket Microbial Communities Pre- and Posttreatment. mSystems 2:e00016-17.

1196 17. Garg N, Wang M, Hyde E, da Silva RR, Melnik AV, Protsyuk I, Bouslimani A, Lim
1197 YW, Wong R, Humphrey G, Ackermann G, Spivey T, Brouha SS, Bandeira N, Lin
1198 GY, Rohwer F, Conrad DJ, Alexandrov T, Knight R, Dorrestein PC. 2017. Three-
1199 Dimensional Microbiome and Metabolome Cartography of a Diseased Human
1200 Lung. Cell Host Microbe.

1201 18. Antharam VC, McEwen DC, Garrett TJ, Dossey AT, Li EC, Kozlov AN, Mesbah Z,
1202 Wang GP. 2016. An Integrated Metabolomic and Microbiome Analysis Identified
1203 Specific Gut Microbiota Associated with Fecal Cholesterol and Coprostanol in
1204 Clostridium difficile Infection. PLoS ONE 11:e0148824.

1205 19. Heintz-Buschart A, May P, Laczny CC, Lebrun LA, Bellora C, Krishna A, Wampach
1206 L, Schneider JG, Hogan A, de Beaufort C, Wilmes P. 2016. Integrated multi-omics
1207 of the human gut microbiome in a case study of familial type 1 diabetes. Nat
1208 Microbiol 2:16180.

1209 20. Hua C, Tian J, Tian P, Cong R, Luo Y, Geng Y, Tao S, Ni Y, Zhao R. 2017.
1210 Feeding a High Concentration Diet Induces Unhealthy Alterations in the

Composition and Metabolism of Ruminal Microbiota and Host Response in a Goat Model. Front Microbiol 8.

21. Price ND, Magis AT, Earls JC, Glusman G, Levy R, Lausted C, McDonald DT, Kusebauch U, Moss CL, Zhou Y, Qin S, Moritz RL, Brogaard K, Omenn GS, Lovejoy JC, Hood L. 2017. A wellness study of 108 individuals using personal, dense, dynamic data clouds. Nat Biotechnol.

22. Vandeputte D, Falony G, Vieira-Silva S, Wang J, Sailer M, Theis S, Verbeke K, Raes J. 2017. Prebiotic inulin-type fructans induce specific changes in the human gut microbiota. Gut 66:1968–1974.

23. Walsh AM, Crispie F, Kilcawley K, O'Sullivan O, O'Sullivan MG, Claesson MJ, Cotter PD. 2016. Microbial Succession and Flavor Production in the Fermented Dairy Beverage Kefir. mSystems 1:e00052-16.

24. Weir TL, Manter DK, Sheflin AM, Barnett BA, Heuberger AL, Ryan EP. 2013. Stool Microbiome and Metabolome Differences between Colorectal Cancer Patients and Healthy Adults. PLoS ONE 8:e70803.

25. Alivisatos AP, Blaser MJ, Brodie EL, Chun M, Dangl JL, Donohue TJ, Dorrestein PC, Gilbert JA, Green JL, Jansson JK, Knight R, Maxon ME, McFall-Ngai MJ, Miller JF, Pollard KS, Ruby EG, Taha SA, Unified Microbiome Initiative Consortium. 2015. A unified initiative to harness Earth's microbiomes. Science 350:507–508.

1230 26. iHMP Research Network Consortium. 2014. The Integrative Human Microbiome
1231 Project: Dynamic Analysis of Microbiome-Host Omics Profiles during Periods of
1232 Human Health and Disease. Cell Host Microbe 16:276–289.

1233 27. Choo JM, Kanno T, Zain NMM, Leong LEX, Abell GCJ, Keeble JE, Bruce KD,
1234 Mason AJ, Rogers GB. 2017. Divergent Relationships between Fecal Microbiota
1235 and Metabolome following Distinct Antibiotic-Induced Disruptions. mSphere
1236 2:e00005-17.

1237 28. Kang D-W, Ilhan ZE, Isern NG, Hoyt DW, Howsmon DP, Shaffer M, Lozupone CA,
1238 Hahn J, Adams JB, Krajmalnik-Brown R. 2018. Differences in fecal microbial
1239 metabolites and microbiota of children with autism spectrum disorders. Anaerobe
1240 49:121–131.

1241 29. Lin Z, Ye W, Zu X, Xie H, Li H, Li Y, Zhang W. 2018. Integrative metabolic and
1242 microbial profiling on patients with Spleen-yang-deficiency syndrome. Sci Rep 8.

1243 30. Melnik AV, da Silva RR, Hyde ER, Aksenov AA, Vargas F, Bouslimani A, Protsyuk
1244 I, Jarmusch AK, Tripathi A, Alexandrov T, Knight R, Dorrestein PC. 2017. Coupling
1245 Targeted and Untargeted Mass Spectrometry for Metabolome-Microbiome-Wide
1246 Association Studies of Human Fecal Samples. Anal Chem 89:7549–7559.

1247 31. Chong J, Xia J. 2017. Computational Approaches for Integrative Analysis of the
1248 Metabolome and Microbiome. Metabolites 7:62.

1249 32. Hoyles L, Jiménez-Pranteda ML, Chilloux J, Brial F, Myridakis A, Aranias T,
1250 Magnan C, Gibson GR, Sanderson JD, Nicholson JK, Gauguier D, McCartney AL,

1251 Dumas M-E. 2018. Metabolic retroconversion of trimethylamine N-oxide and the
1252 gut microbiota. Microbiome 6.

1253 33. Weiss S, Van Treuren W, Lozupone C, Faust K, Friedman J, Deng Y, Xia LC, Xu
1254 ZZ, Ursell L, Alm EJ, Birmingham A, Cram JA, Fuhrman JA, Raes J, Sun F, Zhou
1255 J, Knight R. 2016. Correlation detection strategies in microbial data sets vary
1256 widely in sensitivity and precision. ISME J 10:1669–1681.

1257 34. Werhli AV, Grzegorczyk M, Husmeier D. 2006. Comparative evaluation of reverse
1258 engineering gene regulatory networks with relevance networks, graphical gaussian
1259 models and bayesian networks. Bioinformatics 22:2523–2531.

1260 35. Biggs MB, Medlock GL, Moutinho TJ, Lees HJ, Swann JR, Kolling GL, Papin JA.
1261 2016. Systems-level metabolism of the altered Schaedler flora, a complete gut
1262 microbiota. ISME J.

1263 36. Kešnerová L, Mars RAT, Ellegaard KM, Troilo M, Sauer U, Engel P. 2017.
1264 Disentangling metabolic functions of bacteria in the honey bee gut. PLOS Biol
1265 15:e2003467.

1266 37. Bauer E, Zimmermann J, Baldini F, Thiele I, Kaleta C. 2017. BacArena: Individual-
1267 based metabolic modeling of heterogeneous microbes in complex communities.
1268 PLOS Comput Biol 13:e1005544.

1269 38. Garza DR, van Verk MC, Huynen MA, Dutilh BE. 2018. Towards predicting the
1270 environmental metabolome from metagenomics with a mechanistic model. Nat
1271 Microbiol.

1272 39. Heinken A, Thiele I. 2015. Anoxic conditions promote species-specific mutualism
1273 between gut microbes in silico. Appl Environ Microbiol AEM.00101-15.

1274 40. Magnúsdóttir S, Heinken A, Kutt L, Ravcheev DA, Bauer E, Noronha A,
1275 Greenhalgh K, Jäger C, Baginska J, Wilmes P, Fleming RMT, Thiele I. 2016.
1276 Generation of genome-scale metabolic reconstructions for 773 members of the
1277 human gut microbiota. Nat Biotechnol.

1278 41. Shoaie S, Ghaffari P, Kovatcheva-Datchary P, Mardinoglu A, Sen P, Pujos-Guillot
1279 E, de Wouters T, Juste C, Rizkalla S, Chilloux J, Hoyles L, Nicholson JK, Dore J,
1280 Dumas ME, Clement K, Bäckhed F, Nielsen J. 2015. Quantifying Diet-Induced
1281 Metabolic Changes of the Human Gut Microbiome. Cell Metab 22:320–331.

1282 42. Shapley LS. 1953. 17. A Value for n-Person Games, p. . *In* Kuhn, HW, Tucker, AW
1283 (eds.), Contributions to the Theory of Games (AM-28), Volume II. Princeton
1284 University Press, Princeton.

1285 43. Manor O, Borenstein E. 2017. Systematic Characterization and Analysis of the
1286 Taxonomic Drivers of Functional Shifts in the Human Microbiome. Cell Host
1287 Microbe 21:254–267.

1288 44. Chiu H-C, Levy R, Borenstein E. 2014. Emergent Biosynthetic Capacity in Simple
1289 Microbial Communities. PLoS Comput Biol 10:e1003695.

1290 45. Manor O, Levy R, Borenstein E. 2014. Mapping the Inner Workings of the
1291 Microbiome: Genomic- and Metagenomic-Based Study of Metabolism and
1292 Metabolic Interactions in the Human Microbiome. Cell Metab 20:742–752.

1293 46. Varma A, Palsson BO. 1994. Metabolic Flux Balancing: Basic Concepts, Scientific
1294 and Practical Use. Bio/Technology 12:994–998.

1295 47. Faith JJ, McNulty NP, Rey FE, Gordon JI. 2011. Predicting a Human Gut
1296 Microbiota's Response to Diet in Gnotobiotic Mice. Science 333:101–104.

1297 48. Rivière A, Gagnon M, Weckx S, Roy D, De Vuyst L. 2015. Mutual Cross-Feeding
1298 Interactions between Bifidobacterium longum subsp. longum NCC2705 and
1299 Eubacterium rectale ATCC 33656 Explain the Bifidogenic and Butyrogenic Effects
1300 of Arabinoxylan Oligosaccharides. Appl Environ Microbiol 81:7767–7781.

1301 49. Unterseher M, Jumpponen A, öPik M, Tedersoo L, Moora M, Dormann CF,
1302 Schnittler M. 2011. Species abundance distributions and richness estimations in
1303 fungal metagenomics - lessons learned from community ecology: COMMUNITY
1304 ECOLOGY IN FUNGAL METAGENOMICS. Mol Ecol 20:275–285.

1305 50. Mee MT, Collins JJ, Church GM, Wang HH. 2014. Syntrophic exchange in
1306 synthetic microbial communities. Proc Natl Acad Sci 111:E2149–E2156.

1307 51. McHardy IH, Goudarzi M, Tong M, Ruegger PM, Schwager E, Weger JR, Graeber
1308 TG, Sonnenburg JL, Horvath S, Huttenhower C, McGovern DP, Fornace AJ,
1309 Borneman J. 2013. Integrative analysis of the microbiome and metabolome of the
1310 human intestinal mucosal surface reveals exquisite inter-relationships. Microbiome
1311 1:17.

1312 52. Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT,
1313 Creasy HH, Earl AM, FitzGerald MG, Fulton RS, Giglio MG, Hallsworth-Pepin K,

1314 Lobos EA, Madupu R, Magrini V, Martin JC, Mitreva M, Muzny DM, Sodergren EJ,
1315 Versalovic J, Wollam AM, Worley KC, Wortman JR, Young SK, Zeng Q, Aagaard
1316 KM, Abolude OO, Allen-Vercoe E, Alm EJ, Alvarado L, Andersen GL, Anderson S,
1317 Appelbaum E, Arachchi HM, Armitage G, Arze CA, Ayvaz T, Baker CC, Begg L,
1318 Belachew T, Bhonagiri V, Bihan M, Blaser MJ, Bloom T, Bonazzi V, Paul Brooks J,
1319 Buck GA, Buhay CJ, Busam DA, Campbell JL, Canon SR, Cantarel BL, Chain
1320 PSG, Chen I-MA, Chen L, Chhibba S, Chu K, Ciulla DM, Clemente JC, Clifton SW,
1321 Conlan S, Crabtree J, Cutting MA, Davidovics NJ, Davis CC, DeSantis TZ, Deal C,
1322 Delehaunty KD, Dewhirst FE, Deych E, Ding Y, Dooling DJ, Dugan SP, Michael
1323 Dunne W, Scott Durkin A, Edgar RC, Erlich RL, Farmer CN, Farrell RM, Faust K,
1324 Feldgarden M, Felix VM, Fisher S, Fodor AA, Forney LJ, Foster L, Di Francesco V,
1325 Friedman J, Friedrich DC, Fronick CC, Fulton LL, Gao H, Garcia N, Giannoukos G,
1326 Giblin C, Giovanni MY, Goldberg JM, Goll J, Gonzalez A, Griggs A, Gujja S, Kinder
1327 Haake S, Haas BJ, Hamilton HA, Harris EL, Hepburn TA, Herter B, Hoffmann DE,
1328 Holder ME, Howarth C, Huang KH, Huse SM, Izard J, Jansson JK, Jiang H, Jordan
1329 C, Joshi V, Katancik JA, Keitel WA, Kelley ST, Kells C, King NB, Knights D, Kong
1330 HH, Koren O, Koren S, Kota KC, Kovar CL, Kyrpides NC, La Rosa PS, Lee SL,
1331 Lemon KP, Lennon N, Lewis CM, Lewis L, Ley RE, Li K, Liolios K, Liu B, Liu Y, Lo
1332 C-C, Lozupone CA, Dwayne Lunsford R, Madden T, Mahurkar AA, Mannon PJ,
1333 Mardis ER, Markowitz VM, Mavromatis K, McCorrison JM, McDonald D, McEwen
1334 J, McGuire AL, McInnes P, Mehta T, Mihindukulasuriya KA, Miller JR, Minx PJ,
1335 Newsham I, Nusbaum C, O'Laughlin M, Orvis J, Pagani I, Palaniappan K, Patel
1336 SM, Pearson M, Peterson J, Podar M, Pohl C, Pollard KS, Pop M, Priest ME,

1337 Proctor LM, Qin X, Raes J, Ravel J, Reid JG, Rho M, Rhodes R, Riehle KP, Rivera
1338 MC, Rodriguez-Mueller B, Rogers Y-H, Ross MC, Russ C, Sanka RK, Sankar P,
1339 Fah Sathirapongsasuti J, Schloss JA, Schloss PD, Schmidt TM, Scholz M, Schriml
1340 L, Schubert AM, Segata N, Segre JA, Shannon WD, Sharp RR, Sharpton TJ,
1341 Shenoy N, Sheth NU, Simone GA, Singh I, Smillie CS, Sobel JD, Sommer DD,
1342 Spicer P, Sutton GG, Sykes SM, Tabbaa DG, Thiagarajan M, Tomlinson CM,
1343 Torralba M, Treangen TJ, Truty RM, Vishnivetskaya TA, Walker J, Wang L, Wang
1344 Z, Ward DV, Warren W, Watson MA, Wellington C, Wetterstrand KA, White JR,
1345 Wilczek-Boney K, Wu Y, Wylie KM, Wylie T, Yandava C, Ye L, Ye Y, Yooseph S,
1346 Youmans BP, Zhang L, Zhou Y, Zhu Y, Zoloth L, Zucker JD, Birren BW, Gibbs RA,
1347 Highlander SK, Methé BA, Nelson KE, Petrosino JF, Weinstock GM, Wilson RK,
1348 White O. 2012. Structure, function and diversity of the healthy human microbiome.
1349 Nature 486:207–214.

1350 53. Gilbert JA, Quinn RA, Debelius J, Xu ZZ, Morton J, Garg N, Jansson JK, Dorrestein
1351 PC, Knight R. 2016. Microbiome-wide association studies link dynamic microbial
1352 consortia to disease. Nature 535:94–103.

1353 54. Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA,
1354 Clemente JC, Burkepile DE, Vega Thurber RL, Knight R, Beiko RG, Huttenhower
1355 C. 2013. Predictive functional profiling of microbial communities using 16S rRNA
1356 marker gene sequences. Nat Biotechnol 31:814–821.

1357 55. Caspi R, Altman T, Billington R, Dreher K, Foerster H, Fulcher CA, Holland TA,
1358 Keseler IM, Kothari A, Kubo A, Krummenacker M, Latendresse M, Mueller LA, Ong

1359 Q, Paley S, Subhraveti P, Weaver DS, Weerasinghe D, Zhang P, Karp PD. 2014.
1360 The MetaCyc database of metabolic pathways and enzymes and the BioCyc
1361 collection of Pathway/Genome Databases. Nucleic Acids Res 42:D459–D471.

1362 56. Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes.
1363 Nucleic Acids Res 28:27–30.

1364 57. Noecker C, Eng A, Srinivasan S, Theriot CM, Young VB, Jansson JK, Fredricks
1365 DN, Borenstein E. 2016. Metabolic Model-Based Integration of Microbiome
1366 Taxonomic and Metabolomic Profiles Elucidates Mechanistic Links between
1367 Ecological and Metabolic Variation. mSystems 1:e00013-15.

1368 58. Casero D, Gill K, Sridharan V, Koturbash I, Nelson G, Hauer-Jensen M, Boerma M,
1369 Braun J, Cheema AK. 2017. Space-type radiation induces multimodal responses in
1370 the mouse gut microbiome and metabolome. Microbiome 5.

1371 59. Stewart CJ, Mansbach JM, Wong MC, Ajami NJ, Petrosino JF, Camargo CA,
1372 Hasegawa K. 2017. Associations of Nasopharyngeal Metabolome and Microbiome
1373 with Severity among Infants with Bronchiolitis. A Multiomic Analysis. Am J Respir
1374 Crit Care Med 196:882–891.

1375 60. Franzosa EA, Morgan XC, Segata N, Waldron L, Reyes J, Earl AM, Giannoukos G,
1376 Boylan MR, Ciulla D, Gevers D, Izard J, Garrett WS, Chan AT, Huttenhower C.
1377 2014. Relating the metatranscriptome and metagenome of the human gut. Proc
1378 Natl Acad Sci 111:E2329–E2338.

1379 61. Iwai S, Weinmaier T, Schmidt BL, Albertson DG, Poloso NJ, Dabbagh K, DeSantis
1380 TZ. 2016. Piphillin: Improved Prediction of Metagenomic Content by Direct
1381 Inference from Human Microbiomes. PLOS ONE 11:e0166104.

1382 62. Magnúsdóttir S, Thiele I. 2018. Modeling metabolism of the human gut microbiome.
1383 Curr Opin Biotechnol 51:90–96.

1384 63. Doledec S, Chessel D. 1994. Co-inertia analysis: an alternative method for
1385 studying species-environment relationships. Freshw Biol 31:277–294.

1386 64. Randolph TW, Zhao S, Copeland W, Hullar M, Shojaie A. 2015. Kernel-Penalized
1387 Regression for Analysis of Microbiome Data. ArXiv151100297 Stat.

1388 65. Zhu J, Sova P, Xu Q, Dombek KM, Xu EY, Vu H, Tu Z, Brem RB, Bumgarner RE,
1389 Schadt EE. 2012. Stitching together Multiple Data Dimensions Reveals Interacting
1390 Metabolomic and Transcriptomic Networks That Modulate Cell Regulation. PLoS
1391 Biol 10:e1001301.

1392 66. Keren N, Konikoff FM, Paitan Y, Gabay G, Reshef L, Naftali T, Gophna U. 2015.
1393 Interactions between the intestinal microbiota and bile acids in gallstones patients:
1394 Bile acid and microbiota in gallstones patients. Environ Microbiol Rep 7:874–880.

1395 67. Berry D, Stecher B, Schintlmeister A, Reichert J, Brugiroux S, Wild B, Wanek W,
1396 Richter A, Rauch I, Decker T, Loy A, Wagner M. 2013. Host-compound foraging by
1397 intestinal microbiota revealed by single-cell stable isotope probing. Proc Natl Acad
1398 Sci 110:4720–4725.

68. Kurczy ME, Forsberg EM, Thorgersen MP, Poole FL, Benton HP, Ivanisevic J, Tran ML, Wall JD, Elias DA, Adams MWW, Siuzdak G. 2016. Global Isotope Metabolomics Reveals Adaptive Strategies for Nitrogen Assimilation. ACS Chem Biol 11:1677–1685.

69. Turnbaugh PJ, Gordon JI. 2008. An Invitation to the Marriage of Metagenomics and Metabolomics. Cell 134:708–713.

70. Collins WD, Bitz CM, Blackmon ML, Bonan GB, Bretherton CS, Carton JA, Chang P, Doney SC, Hack JJ, Henderson TB, Kiehl JT, Large WG, McKenna DS, Santer BD, Smith RD. 2006. The Community Climate System Model Version 3 (CCSM3). J Clim 19:2122–2143.

71. Connolly AJ, Angeli GZ, Chandrasekharan S, Claver CF, Cook K, Ivezic Z, Jones RL, Krughoff KS, Peng E-H, Peterson J, Petry C, Rasmussen AP, Ridgway ST, Saha A, Sembroski G, vanderPlas J, Yoachim P. 2014. An end-to-end simulation framework for the Large Synoptic Survey Telescope, p. 915014. *In* Angeli, GZ, Dierickx, P (eds.), .

72. McNally CP, Borenstein E. 2018. Metabolic model-based analysis of the emergence of bacterial cross-feeding via extensive gene loss. BMC Syst Biol 12.

73. Schellenberger J, Que R, Fleming RMT, Thiele I, Orth JD, Feist AM, Zielinski DC, Bordbar A, Lewis NE, Rahmanian S, Kang J, Hyduke DR, Palsson BØ. 2011. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. Nat Protoc 6:1290–1307.

74. Casteleyn C, Rekecki A, Van der Aa A, Simoens P, Van den Broeck W. 2010. Surface area assessment of the murine intestinal tract as a prerequisite for oral dose translation from mouse to man. Lab Anim 44:176–183.

75. Holzhütter H-G. 2004. The principle of flux minimization and its application to estimate stationary fluxes in metabolic networks: Flux minimization. Eur J Biochem 271:2905–2922.

76. McNulty NP, Wu M, Erickson AR, Pan C, Erickson BK, Martens EC, Pudlo N a, Muegge BD, Henrissat B, Hettich RL, Gordon JI. 2013. Effects of diet on resource utilization by a model human gut microbiota containing Bacteroides cellulosilyticus WH2, a symbiont with an extensive glycobiome. PLoS Biol 11:e1001637.

77. Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B Methodol 57:289–300.

78. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, Müller M. 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 12:77.

79. Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Zech Xu Z, Kightley EP, Thompson LR, Hyde ER, Gonzalez A, Knight R. 2017. Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. mSystems 2:e00191-16.

80. Gonzalez A, Navas-Molina JA, Kosciolek T, McDonald D, Vázquez-Baeza Y, Ackermann G, DeReus J, Janssen S, Swafford AD, Orchanian SB, Sanders JG, Shorenstein J, Holste H, Petrus S, Robbins-Pianka A, Brislawn CJ, Wang M, Rideout JR, Bolyen E, Dillon M, Caporaso JG, Dorrestein PC, Knight R. 2018. Qiita: rapid, web-enabled microbiome meta-analysis. Nat Methods 15:796–798.

81. Drost H-G, Paszkowski J. 2017. Biomartr: genomic data retrieval with R. Bioinformatics btw821.

82. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016. VSEARCH: a versatile open source tool for metagenomics. PeerJ 4:e2584.

83. Noronha A, Modamio J, Jarosz Y, Guerard E, Sompairac N, Preciat G, Daníelsdóttir AD, Krecke M, Merten D, Haraldsdóttir HS, Heinken A, Heirendt L, Magnúsdóttir S, Ravcheev DA, Sahoo S, Gawron P, Friscioni L, Garcia B, Prendergast M, Puente A, Rodrigues M, Roy A, Rouquaya M, Wiltgen L, Žagare A, John E, Krueger M, Kuperstein I, Zinovyev A, Schneider R, Fleming RMT, Thiele I. 2018. The Virtual Metabolic Human database: integrating human and gut microbiome metabolism with nutrition and disease. Nucleic Acids Res gky992–gky992.

84. Markowitz VM, Chen I-MA, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P, Huntemann M, Anderson I, Mavromatis K, Ivanova NN, Kyrpides NC. 2012. IMG: the integrated microbial genomes database and comparative analysis system. Nucleic Acids Res 40:D115–D122.

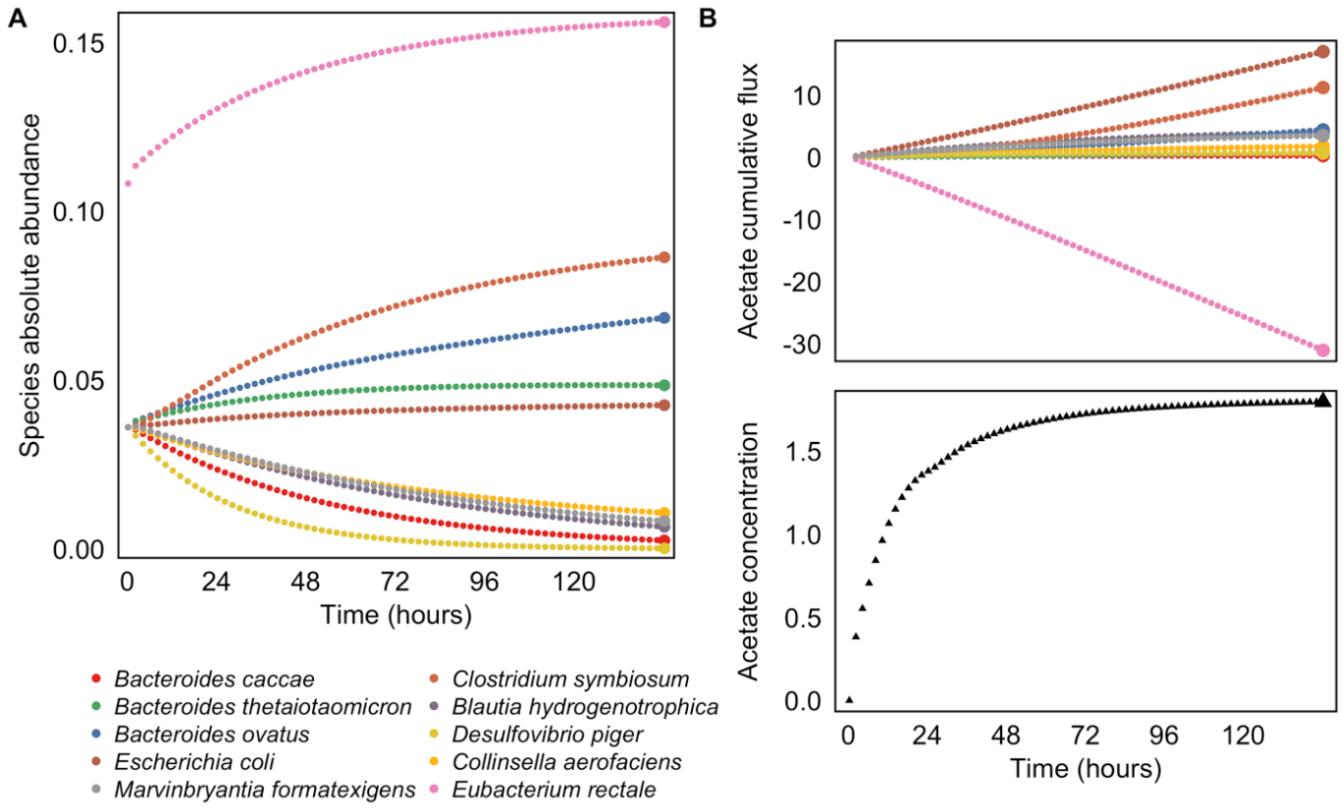


**Figure 1. Simulating multi-omic data with a dynamic multi-species genome-scale framework. (A)** Community species abundances throughout a single simulation run. Abundances were quantified in units of microbial biomass. In this simulation, community composition was initialized with a high relative abundance of *Eubacterium rectale*. For visual clarity, only every eighth time step is illustrated. Species abundances at the final time point (highlighted with larger colored circles) were used for calculating species-metabolite correlations. **(B)** Cumulative secretion and uptake of acetate by each community member, throughout the same simulation run illustrated in panel A. Acetate was synthesized by several species and consumed by *E. rectale* over the course of the simulation. Total cumulative fluxes (highlighted with larger colored circles) were used for calculating species contributions to metabolite variation. The bottom plot illustrates the resulting environmental concentration of acetate at each time point. The metabolite concentration at the final time point (highlighted with a larger black triangle) was used for calculating species-metabolite correlations.

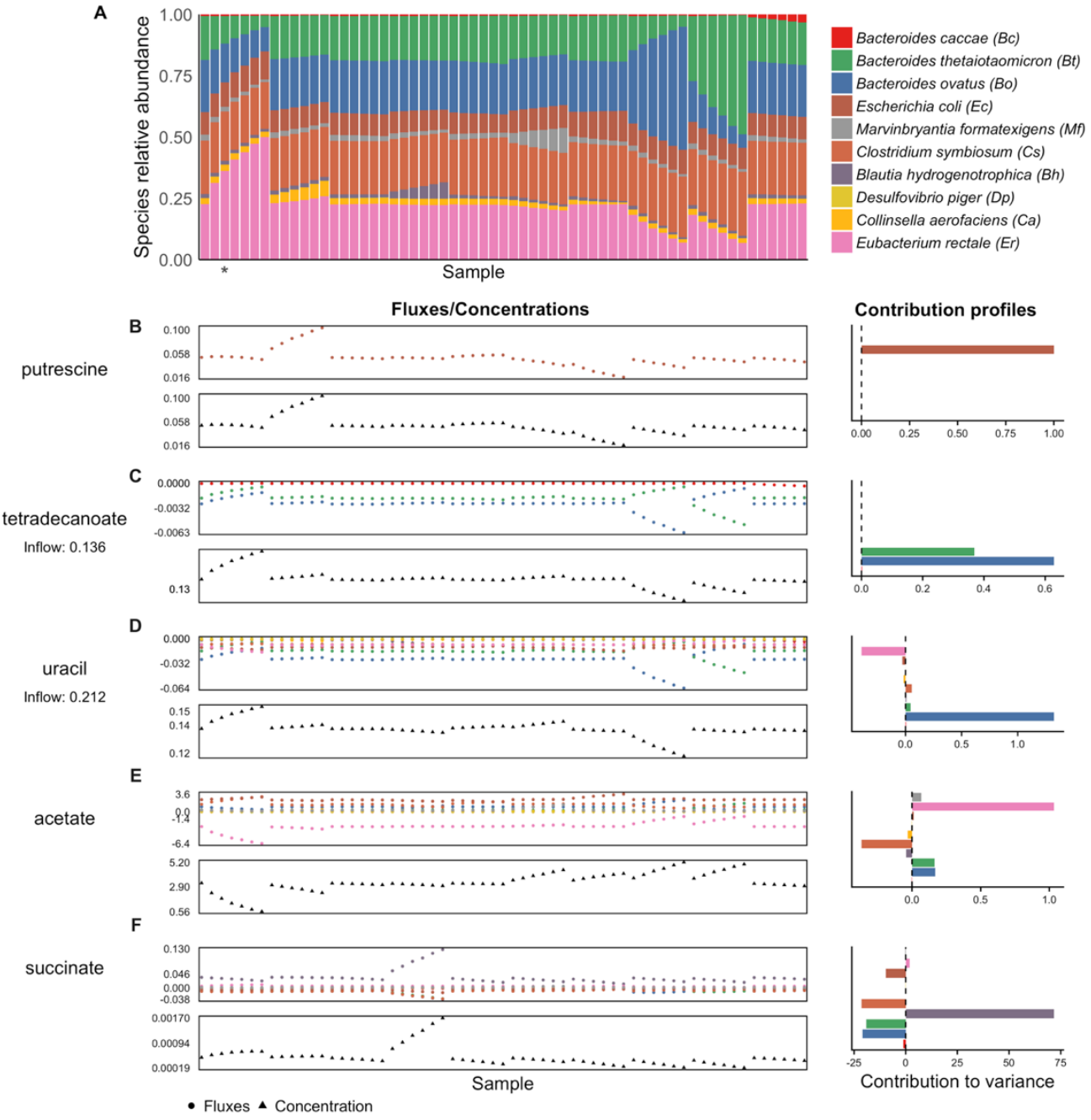


**Figure 2. Species abundances, cumulative fluxes, and contributions to variance in metabolite concentrations in our simulated dataset. (A)** The dataset of species abundances at the final time point of 61 simulation runs. Each bar represents a simulation run, with the colors indicating relative abundance of each species. The abundance profile from the simulation runs highlighted in Figure 1 is indicated with an asterisk. **(B-F)** For five example metabolites, the upper plot shows the total cumulative secretion or uptake of that metabolite by each species across all 61 simulation runs (or samples). The lower plot shows the corresponding environmental concentration at the final time point. The bar plot on the right shows the contribution values for each species and metabolite, calculated from the flux values and describing each species' linear contribution to the overall metabolite variance.

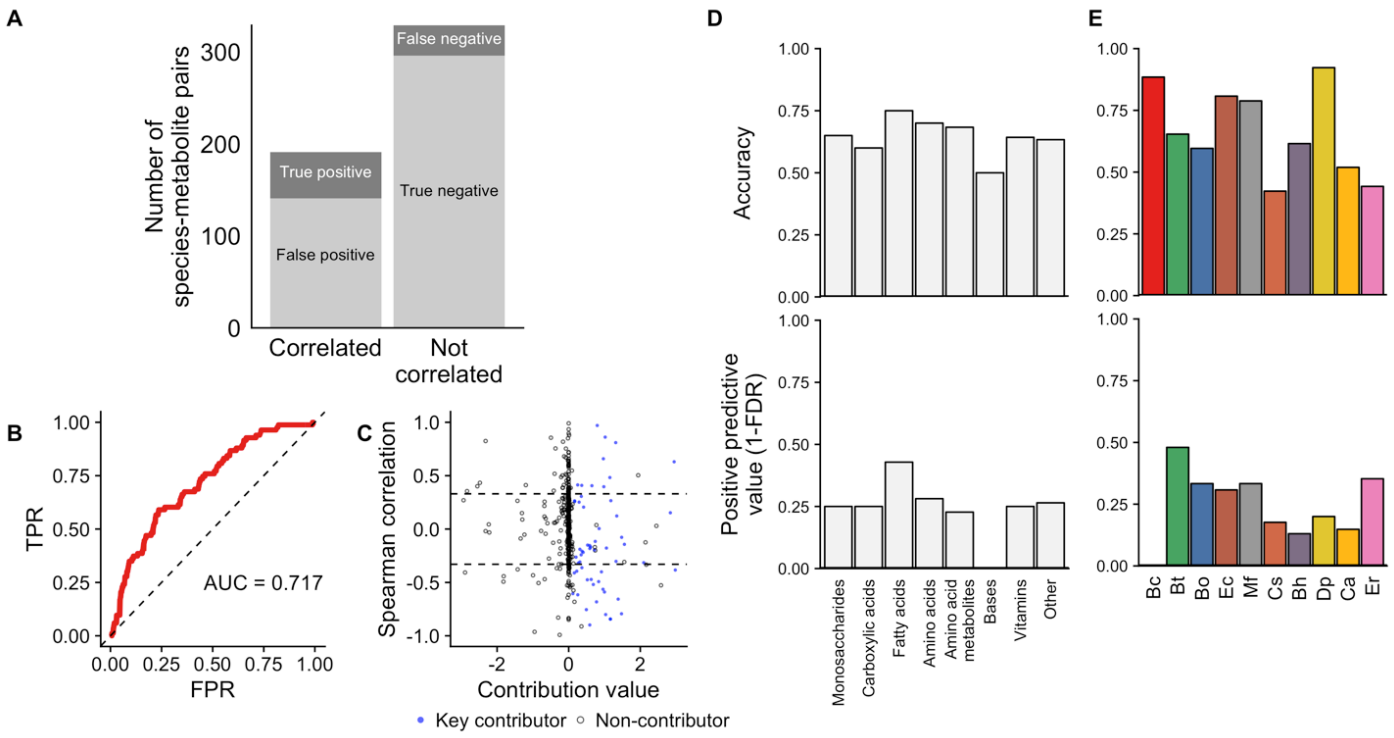


**Figure 3**. **Species-metabolite correlations poorly predict species contributions to metabolite variation. (A)** The number of species-metabolites pairs that were significantly correlated (left bar) or not-correlated (right bar) and its correspondence with true species-metabolite key contributors (indicated by shade of gray). **(B)** Receiver operating characteristic (ROC) plot, showing the ability of absolute Spearman correlation values to classify key contributors among all species-metabolite pairs. **(C)** Scatter plot of species-metabolite pairs, showing the poor correspondence between true contribution values (x-axis) and Spearman correlation (y-axis). Key contributors are plotted as blue points, others as hollow circles. Dashed lines show significant correlations ($p$<0.01). There are 65 species-metabolite pairs with a contribution value greater than 3 in magnitude whose values are not shown. **(D-E)** Accuracy and positive predictive value of Spearman correlation analysis for detecting true key contributors across metabolite classes (Panel D) and for each of the 10 species (Panel E).

4

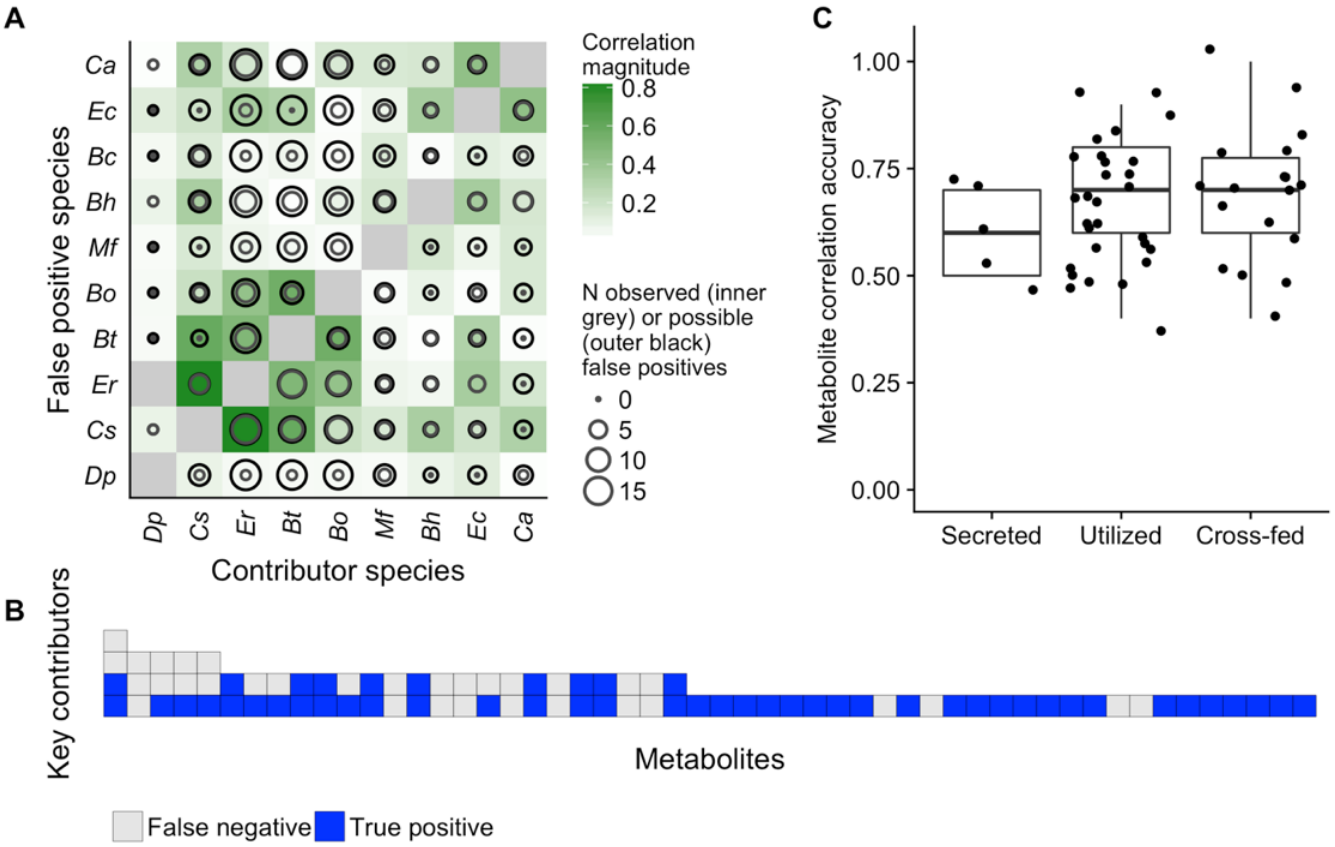


**Figure 4. Metabolite and species properties explain correlation-contribution discrepancies. (A)** Strongly correlated species pairs produced more false positive metabolite correlations. In this plot, the color of each tile indicates the strength of correlation in the abundances of each pair of species. The size of the outer black circle in each cell represents the number of metabolites for which the species on the x-axis is a key contributor and the species on the y-axis is not. The size of the inner circle represents the share of those metabolites for which a false positive is observed for the species on the y-axis. It can be seen that many false positive correlations involve the taxa with the strongest interspecies associations: *E. rectale, B. ovatus,* and *B. thetaiotaomicron.* **(B)** Metabolites with more microbial key contributors were more prone to false negative correlations. Each column represents an analyzed metabolite, ordered by its number of key microbial contributors, which are represented by each tile. The tiles are coded by the correlation outcome for each contributor. **(C)** Correlations detected key contributors equally accurately regardless of whether a metabolite is secreted, utilized, or cross-fed by the species. Each point represents the accuracy of correlations for a single metabolite across its comparisons with all 10 species.

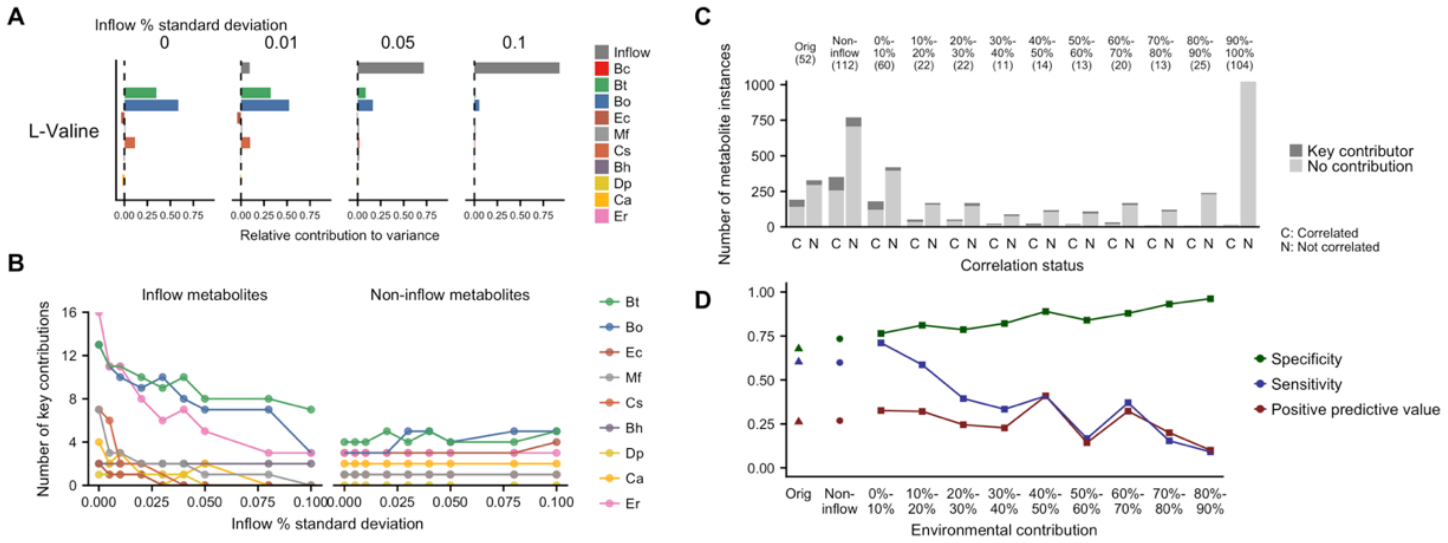


**Figure 5. Environmental fluctuations impact correlation-contributor sensitivity and specificity.** (A) Example set of contribution profiles for a single inflow metabolite, L-valine, with increasing fluctuations in its inflow. The relative contribution values for each species and for the inflow are shown for 4 sets simulation runs, each with a different degree of fluctuation. The label on each plot describes the relative standard deviation (coefficient of variation) of inflow metabolite concentrations for that set of simulations. The microbial contributions to variance in L-valine concentrations became relatively smaller with increasing variation from the external environment. **(B)** Shifts in key microbial contributors with increasing environmental inflow fluctuations. The number of key contributions of each species to the 52 analyzed metabolites is shown, separately for metabolites present in and absent from the nutrient inflow. Microbial contributors to inflow metabolites decreased as environmental contributions increased, but this effect varied between taxa. **(C)** Correlation analysis failed to detect key microbial contributors regardless of the size of contribution from external inflow variation. Across all sets of simulations, metabolites were binned based on the percent of total positive contribution from the external inflow. The bar plots shown have the same format as Figure 3A, showing the number of species-metabolites pairs that were significantly correlated (left bar) or not-correlated (right bar) and its correspondence with true species-metabolite key contributors (indicated by shade of gray). The first two bars, labeled “Orig” describe the original set of simulations (replicating Figure 3A). The next two show the results for non-inflow metabolites across all levels of inflow fluctuations. The remaining bars show the results for metabolites with increasing levels of environmental contribution. **(D)** Correlation analysis detected key microbial contributors with increased specificity, decreased sensitivity, and generally consistent positive predictive value with increasing contribution from the external inflow. Sensitivity, specificity, and positive predictive value are shown for same environmental contribution bins as in Panel C.

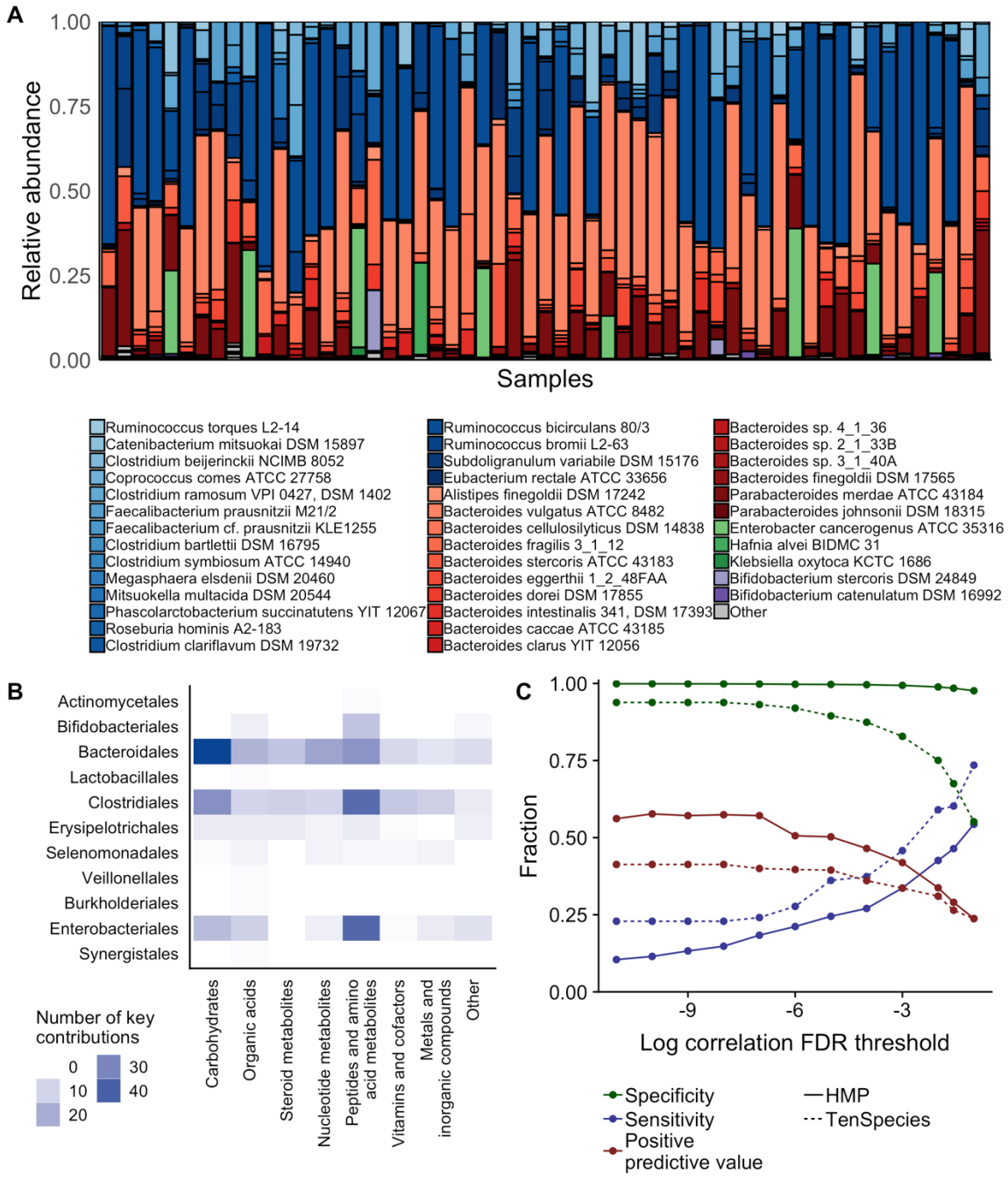


**Figure 6. Correlation-contribution discrepancies persist in simulations of complex human gut-based microbiota. (A)** Species abundances of the 57 Human Microbiome Project (HMP) based-simulations at the 144 hour time point. Shades of blue indicate species in the phylum Firmicutes; red, Bacteroidetes; green, Proteobacteria; and purple, Actinobacteria. **(B)** Key contributions to metabolite variation across the HMP-based dataset, summarized at the level of taxonomic orders and metabolite categories. **(C)** Performance of correlation analysis for identifying key species-metabolite contributors in the HMP-based dataset (solid lines) compared with the original 10-species dataset (dashed lines) across varying significance levels, using Benjamini-Hochberg false discovery rate (FDR) corrected *p*-values.

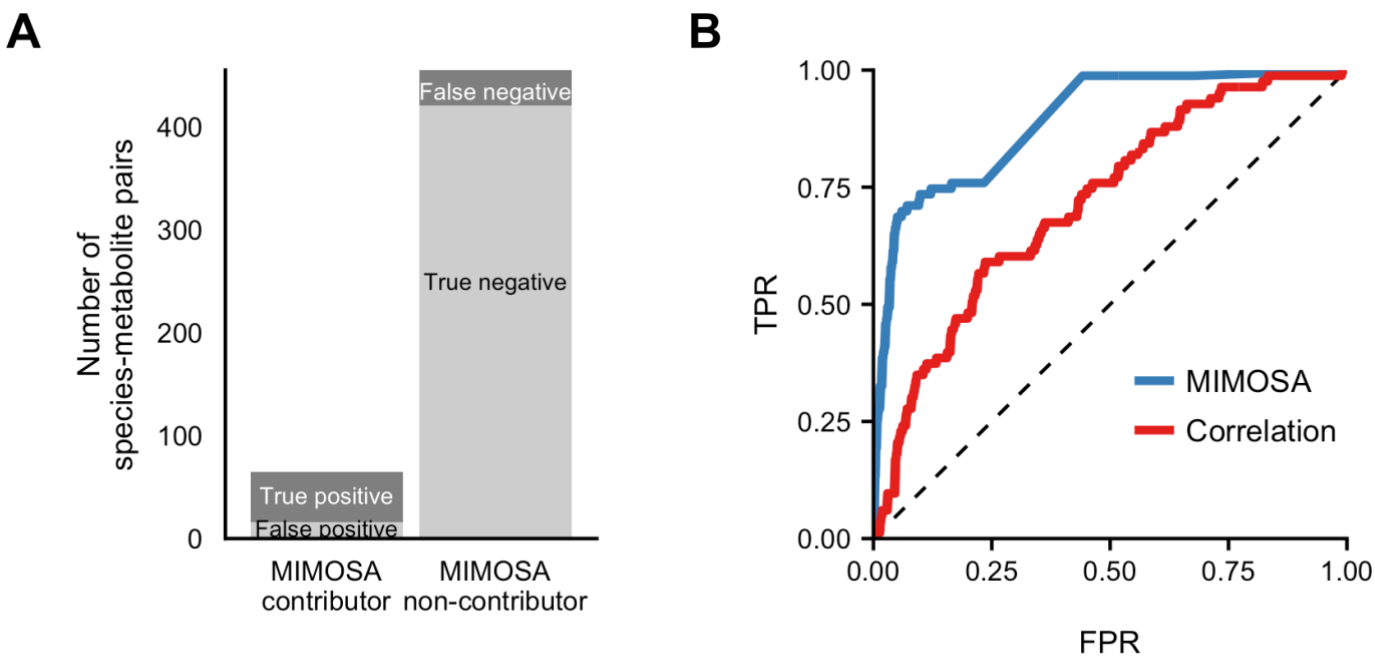


**Figure 7. MIMOSA identified key microbial contributors more accurately than correlation analysis. (A)** The number of species-metabolite pairs that were identified as potential contributors (left bar) or not (right bar) by MIMOSA, and its correspondence with true key contributors. **(B)** Receiver operating characteristic (ROC) plot, showing the ability of both MIMOSA and absolute Spearman correlation values to classify key contributors among all species-metabolite pairs.