AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems™

RESEARCH ARTICLE
Novel Systems Biology Techniques

Metabolic Model-Based Integration of Microbiome Taxonomic and Metabolomic Profiles Elucidates Mechanistic Links between Ecological and Metabolic Variation

Cecilia Noecker,[a] Alexander Eng,[a] Sujatha Srinivasan,[b] Casey M. Theriot,[c] Vincent B. Young,[d,e] Janet K. Jansson,[f] David N. Fredricks,[b,g,h] Elhanan Borenstein[a,i,j]

Department of Genome Sciences, University of Washington, Seattle, Washington, USA[a]; Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA[b]; Department of Population Health and Pathobiology, North Carolina State University, Raleigh, North Carolina, USA[c]; Department of Internal Medicine, Division of Infectious Diseases, University of Michigan, Ann Arbor, Michigan, USA[d]; Department of Microbiology and Immunology, University of Michigan, Ann Arbor, Michigan, USA[e]; Biological Sciences Division, Pacific Northwest National Laboratory, Richland, Washington, USA[f]; Division of Allergy and Infectious Diseases, University of Washington, Seattle, Washington, USA[g]; Department of Microbiology, University of Washington, Seattle, Washington, USA[h]; Department of Computer Science and Engineering, University of Washington, Seattle, Washington, USA[i]; Santa Fe Institute, Santa Fe, New Mexico, USA[j]

ABSTRACT Multiple molecular assays now enable high-throughput profiling of the ecology, metabolic capacity, and activity of the human microbiome. However, to date, analyses of such multi-omic data typically focus on statistical associations, often ignoring extensive prior knowledge of the mechanisms linking these various facets of the microbiome. Here, we introduce a comprehensive framework to systematically link variation in metabolomic data with community composition by utilizing taxonomic, genomic, and metabolic information. Specifically, we integrate available and inferred genomic data, metabolic network modeling, and a method for predicting community-wide metabolite turnover to estimate the biosynthetic and degradation potential of a given community. Our framework then compares variation in predicted metabolic potential with variation in measured metabolites' abundances to evaluate whether community composition can explain observed shifts in the community metabolome, and to identify key taxa and genes contributing to the shifts. Focusing on two independent vaginal microbiome data sets, each pairing 16S community profiling with large-scale metabolomics, we demonstrate that our framework successfully recapitulates observed variation in 37% of metabolites. Well-predicted metabolite variation tends to result from disease-associated metabolism. We further identify several disease-enriched species that contribute significantly to these predictions. Interestingly, our analysis also detects metabolites for which the predicted variation negatively correlates with the measured variation, suggesting environmental control points of community metabolism. Applying this framework to gut microbiome data sets reveals similar trends, including prediction of bile acid metabolite shifts. This framework is an important first step toward a system-level multi-omic integration and an improved mechanistic understanding of the microbiome activity and dynamics in health and disease.

IMPORTANCE Studies characterizing both the taxonomic composition and metabolic profile of various microbial communities are becoming increasingly common, yet new computational methods are needed to integrate and interpret these data in

Received 6 November 2015 **Accepted** 1 December 2015 **Published** 19 January 2016

Citation Noecker C, Eng A, Srinivasan S, Theriot CM, Young VB, Jansson JK, Fredricks DN, Borenstein E. 2016. Metabolic model-based integration of microbiome taxonomic and metabolomic profiles elucidates mechanistic links between ecological and metabolic variation. mSystems 1(1):e00013-15. doi:10.1128/mSystems.00013-15.

Editor Laura M. Sanchez, University of Illinois at Chicago

Address correspondence to Elhanan Borenstein, elbo@uw.edu.

terms of known biological mechanisms. Here, we introduce an analytical framework to link species composition and metabolite measurements, using a simple model to predict the effects of community ecology on metabolite concentrations and evaluating whether these predictions agree with measured metabolomic profiles. We find that a surprisingly large proportion of metabolite variation in the vaginal microbiome can be predicted based on species composition (including dramatic shifts associated with disease), identify putative mechanisms underlying these predictions, and evaluate the roles of individual bacterial species and genes. Analysis of gut microbiome data using this framework recovers similar community metabolic trends. This framework lays the foundation for model-based multi-omic integrative studies, ultimately improving our understanding of microbial community metabolism.

KEYWORDS: microbiome, multi-omic, metabolic modeling, community composition, metabolomics

The human microbiome carries out a plethora of metabolic processes that are often vital to the health of the host. Microbiome metabolic activity can, for example, impact energy harvest, inflammation, and infection susceptibility (1–3), suggesting that alterations in community metabolism may be an important mechanism underlying an array of poorly understood associations between the composition of the microbiome and disease (4–6). Indeed, the metabolic capacity of the gut microbiome appears to be relatively constant across healthy individuals (7), and yet, it can vary dramatically in response to perturbations like antibiotic treatment or diet changes (8, 9) or in a variety of disease states (10, 11).

Understanding the relationship between the composition of the microbiome and its metabolic activity (and ultimately, the development of microbiome-associated diseases) is therefore an important task. To this end, numerous recent studies have paired comprehensive taxonomic characterization (based on, for example, 16S rRNA gene assays) with metabolomic profiling, aiming to reveal and evaluate the mechanisms underlying taxonomic and metabolic shifts in the microbiome across diverse environments and disease states (12–27). To date, however, methods for integrating taxonomic and metabolomic data are lacking, and consequently, the vast majority of these studies have analyzed community composition and metabolite profiles independently or focused on identifying statistical associations between these two data types.

While the discovery of such associations is clearly an important first step in describing the function and dynamics of the microbiome in health and disease, it ignores extensive prior knowledge of genomic capacities and metabolic mechanisms that link community ecology and metabolism and may accordingly fall short of gaining a systems-level mechanistic understanding of such complex ecosystems. For example, a strong correlation between a species and a metabolite may have very different interpretations depending on whether the species in question is known to degrade that metabolite or to synthesize it. Integrating the taxonomic and metabolomic profiles of the system under study therefore requires not only linking these two data sets but also the incorporation of prior reference information about the metabolic capacities of various community members and the way such capacities interact. Specifically, an integrated analysis could shed light on the extent to which variation in a metabolite of interest can be explained by observed shifts in community ecology and metabolic capacity, as opposed to alternative environmental factors. This is crucial for gaining a comprehensive understanding of the microbiome and for future efforts to modulate metabolic phenotypes via microbiome-based interventions.

Several recent studies have taken initial steps to address this challenge. One avenue of research aims to reconstruct predictive metabolic models of community metabolism in various settings (using, for example, constraint-based modeling), which can then potentially be validated by metabolomic profiling (28–30). This approach, however, depends on relatively complete and high-quality metabolic models of the species involved and, therefore, may not scale well to complex communities with partially

characterized taxa. Other studies have used information about enzymatic reactions to infer metabolic turnover potential from taxonomic composition and metagenome content (31, 32). However, these studies focused on comparing predicted metabolic potential to environmental parameters or community dysbiosis rather than to detailed, large-scale metabolomic phenotypes. McHardy et al. (33) used correlation network analysis to cluster metabolites and evaluated the correspondence between the resulting clusters and metabolically related pathway abundances, an approach that successfully quantified relationships between functional pathways and metabolites but that was still primarily association based and difficult to interpret. Sridharan et al. (34) similarly focused on a small subset of metabolism, constructing a reference genome-based supraorganism metabolic network model and applying a pathway construction algorithm to predict bioactive aromatic microbial metabolites likely to be found in the human gut. These studies all show the tremendous promise of linking microbial composition to metabolomic variation based on prior knowledge of the various metabolic processes, and yet, they are still limited in scale. Thus, the development of a systematic, mechanistic approach for evaluating the relationships between the community ecology and metabolite shifts is called for.

We therefore present here a comprehensive analytical multi-omic framework for integrating community structure and metabolic profile, aiming to elucidate mechanisms underlying metabolic variation in the human microbiome. Our framework first infers community gene content based on available and inferred genomic information and adapts a method originally developed to interpret environmental metagenomes (31) to approximate the potential effect of the microbiome on each metabolite. We systematically compare these estimates to measured metabolome variation and interpret the results in terms of metabolic mechanisms based on taxonomic shifts. We apply this framework to two data sets pairing community taxonomic composition and global metabolite profiles from the vaginal microbiome, as well as to data sets from the gut microbiomes of humans and mice. Using this framework, we identify a large number of metabolites whose variation across samples can be explained (or "predicted") by shifts in microbial community composition and the metabolic capacity of the various member species. We further use this approach to identify species and reactions that are key contributors to the calculated communitywide metabolic potential and highlight putative alternative mechanisms for poorly predicted metabolites. Importantly, our analysis detects broad trends in metabolite predictability across data sets and serves as a proof of concept of the use of systematic mechanism-based integration of multi-omic data to gain new insight into microbial community metabolism.

RESULTS

A metabolic model-based framework for integrating taxonomic and metabolomic data. We developed a computational framework to systematically link variation in community ecology with observed variation in its metabolic phenotype (Fig. 1). Our framework specifically assesses whether the measured between-sample variation in metabolite abundances can be explained by observed shifts in species composition and information about the metabolic capacity of each species.

Briefly, our framework first infers the metagenome content for each sample based on taxonomic composition and available or inferred reference genome information (35). Inferred metagenomes are then normalized using a previously introduced method (MUSiCC) (36), resulting in an estimate of the average copy number of each gene across microbiome genomes. Next, our framework applies a method for predicting relative metabolic turnover (31), using a metabolic network model to translate the resulting enzymatic gene abundance estimates into community-based metabolite potential (CMP) scores. These scores represent the relative capacity of the community in a given sample to generate or deplete each metabolite, based on metabolic reference information that links enzymes to their substrates and products (37). To evaluate these scores, our framework then compares for each metabolite the differences in CMP scores between all pairs of samples with the differences in the corresponding measured

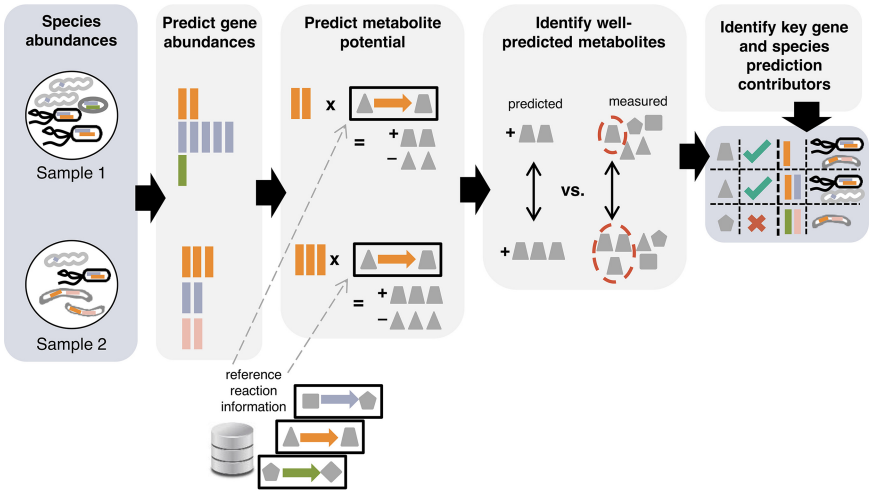

FIG 1 Framework for integrating taxonomic and metabolomic data. Species composition is first used to predict the metagenome's gene content, which is then paired with reaction information to estimate the community metabolic potential (CMP) for each sample and metabolite. Variation in predicted CMP scores is compared to variation in measured metabolite abundances (using pairwise differences) to identify well-predicted metabolites. A perturbation-based approach is used to additionally identify key species, gene, and reaction contributors to CMP scores.

metabolite abundance. Using these pairwise comparisons and a statistical test for correlation between two distance matrices, our framework evaluates whether there is an agreement between variation in predicted CMP scores and variation in measured metabolite abundances. We term those metabolites for which this agreement is statistically significant "well-predicted." Finally, our framework uses a perturbation-based approach to identify the bacterial species, genes, and reactions that are the key mechanistic contributors to calculated CMP scores. A more detailed description of this framework can be found in Materials and Methods.

Metabolic model-based prediction explains metabolite variation in the vaginal microbiome based on taxonomic shifts. We first applied our framework to data sets pairing bacterial community and metabolomic profiles from the vaginal microbiome, a relatively simple community typically dominated by a limited number of species. We specifically analyzed two independently obtained data sets (each consisting of ~70 samples; see Table S1 in the supplemental material), characterizing the vaginal microbiomes and metabolomes of healthy women and women with bacterial vaginosis (BV) (22). Samples from the first data set (data set 1) were analyzed for taxonomic composition using quantitative PCR (qPCR) for 14 vaginal bacterial species and for metabolites using global liquid chromatography-mass spectrometry (LC-MS) and gas chromatography (GC)-MS, whereas samples from the second data set (data set 2) were analyzed using broad-range 16S rRNA gene sequencing and targeted LC-MS (see Materials and Methods).

In each of these data sets, we used our framework to calculate the CMP score for each metabolite and in each sample. Of the metabolites assayed in each data set, roughly 50% could not be associated with a CMP score due to missing or noninformative annotated metabolic data (see Materials and Methods; see also Table S1 in the supplemental material) and were accordingly discarded from downstream analysis. The CMP scores of the remaining metabolites were compared to measured metabolite abundances as described above to examine whether the observed variation in the metabolite abundances across samples can be explained mechanistically by variation in the set of species comprising the community. Surprisingly, we found that 40.2% of the metabolites analyzed in data set 1 and 34.5% of metabolites analyzed in data set 2 were well-predicted (see Table S2), suggesting that for a substantial fraction of metabolites, information about the metabolic capabilities for the member species is sufficient to explain observed differences in metabolite abundance. We further confirmed that the

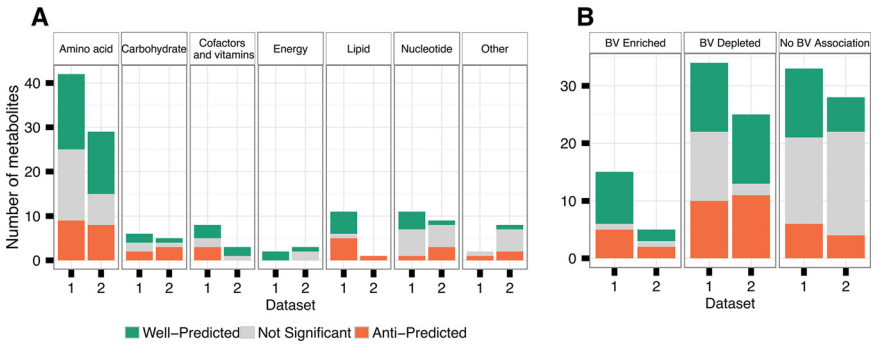

FIG 2 **Metabolite predictability across metabolic categories (A) and disease states (B) in the vaginal microbiome. Well-predicted metabolites are defined as those for which variation in CMP scores is significantly correlated (using a Mantel test) with variation in measured metabolite abundance at a false discovery rate (FDR) of 0.01. Anti-predicted metabolites are similarly defined as those for which variation in CMP scores is significantly negatively correlated with variation in measured metabolite abundances (FDR 0.01). Metabolic categorization is based on KEGG data, and disease enrichment is based on a Wilcoxon rank sum test for association with bacterial vaginosis (BV) with a Bonferroni-corrected *P* value of <0.1.**

identification of well-predicted metabolites and the correlations observed between calculated CMP scores and measured abundances are not artifacts of the data covariance structure, using randomized metabolic networks to generate a predictability null model (see Materials and Methods). We found that randomized networks produced a consistently lower proportion of well-predicted metabolites than the real network ($P <$ 0.01 for both data sets). Metabolites analyzed in both data sets were generally predictable at similar levels ($\rho = 0.63$, Spearman correlation test) (see Fig. S1). Finally, we also observed a significant overlap between metabolites for which variation in CMP scores was significantly correlated with variation in measured metabolite abundance in both data sets 1 and 2 and in a simple monoculture-based *Escherichia coli* data set ($P =$ 0.04; Fisher exact text) (see Text S1 and Fig. S2). This finding suggests that our framework may identify consistent control points in microbial metabolism.

We next examined whether well-predicted metabolites tend to be associated with specific metabolic categories or host state. We found that well-predicted metabolites spanned a range of metabolic categories (Fig. 2A). Specifically, well-predicted metabolites represent all major metabolic categories, with many well-predicted metabolites being associated with amino acid metabolism, an important category of microbe-mediated processes in this environment. Additionally, 60% and 40% of the strongly BV-enriched metabolites, including known metabolic markers of BV (38), such as the amino acid catabolites *N*-acetylputrescine, spermidine, and citrulline, were predicted well in each data set (Fig. 2B).

Interestingly, we also observed a substantial portion of metabolites for which variation in the CMP scores was strongly negatively correlated with variation in measured abundances (25.6% in data set 1 and 29.3% in data set 2; see Table S2 in the supplemental material). These "anti-predicted" metabolites were often linked to a well-predicted metabolite either by a reversible reaction (which is not factored into CMP score calculation) (7 and 4 metabolite pairs in data set 1 and 2, respectively) or by a reaction synthesizing the anti-predicted metabolite from a well-predicted metabolite (6 and 2 metabolite pairs). For example, in data set 1, glutamate is well-predicted, while glutamine, a metabolite that can be synthesized from glutamate, is anti-predicted, suggesting that other, unaccounted-for factors influence its abundance in this environment. Overall, anti-predicted metabolites were adjacent to well-predicted metabolites more frequently than expected by chance (15 and 8 metabolite pairs, $P < 0.005$ and $P < 0.03$ in data set 1 and 2, respectively, by a permutation-based test; see Materials and Methods). Such anti-predicted metabolites may be the result of missing information about community composition or genomic capacities. However, they may also point to environmentally regulated points in metabolism (as opposed to

microbiome-controlled metabolites), where an environmental change in metabolite abundance and nutrient availability give rise to taxonomic shifts in the microbiome. Put differently, in contrast to well-predicted metabolites that are likely produced by the microbiome, so that an increase in their abundance correlates with an increase in the abundance of species that have the capacity to synthesize them, an increase in the abundance of anti-predicted metabolites can potentially be introduced by the environment and selects for species that have the capacity to degrade them (see Discussion).

A small set of BV-enriched bacterial species explains a large portion of the metabolome variation. We next examined the contribution of individual species to the calculated CMP scores of each metabolite. We quantified each species' contribution as the correlation between a CMP score that is calculated based on that species alone (e.g., ignoring all other species in the community) and the community-wide CMP score described above (Materials and Methods). We defined species for which this correlation was above 0.5 as key contributors. We first focused on data set 1, in which only a small number of species was assayed but the availability of absolute concentration data (owing to the use of qPCR) may better distinguish key species in the community. In total, we found that 10 of the 11 species analyzed in data set 1 were key contributors to at least one metabolite. Importantly, the vast majority of metabolites (93.9%) had 4 or fewer key contributors, yet the particular combination of species varied widely across metabolites. This suggests that shifts in the abundance of each metabolite (and in particular shifts associated with the BV state) may be attributed to a small number of species rather than to community-wide dysbiosis. For instance, although both *N*-acetylputrescine and citrulline are BV-enriched polyamine metabolites, the increased abundance of *N*-acetylputrescine in BV is driven in both data sets mostly by the genomic capacities and variation in the abundance of *Prevotella* species, while citrulline's enrichment is driven primarily by *Atopobium vaginae* and *Eggerthella*. Species contributing to the CMP scores of anti-predicted metabolites also recover known processes: for example, *Lactobacillus iners* is the only key species contributor driving the anti-prediction of glycerol in data set 1 (due to *L. iners*' genome encoding glycerol utilization genes). A recent metatranscriptomic study of vaginal *L. iners* found evidence that this species is largely the only member of this community that uses glycerol as a carbon source (39), which combined with our results, suggests that a vaginal environment with glycerol availability may promote *L. iners* growth.

We further examined the number of metabolites (and specifically, well-predicted metabolites) for which each species was a key contributor to CMP score calculation. We found that in data set 1, *Eggerthella* sp. 1 and *Megasphaera* type 1 were key contributors to a particularly high number of metabolites relative to the contributions of other species (Fig. 3). BV-enriched metabolites that were well-predicted primarily by these two species alone include *N*-acetylneuraminate, ethanolamine, and the lipid metabolites 4-trimethylaminobutanoate/gamma-butyrobetaine and 3-methyl-2-oxobutanoate (see Fig. S3 in the supplemental material). Notably, these are neither the most abundant nor the most variable species in this data set, although *Eggerthella* sp. 1 is the most differentially abundant species between healthy and BV samples based on Wilcoxon rank-sum tests ($P < 10^{-8}$), whereas *Megasphaera* is fifth most differentially abundant ($P < 10^{-9}$). *Eggerthella* also has the largest genome of any of the analyzed species in terms of the number of protein-coding genes (2,936 genes). Combined, these findings illustrate that the species contributing most significantly to potential shifts in disease-associated metabolic phenotypes may not necessarily be the most abundant or most variable species and that observed metabolic shifts are the product of complex dependencies between ecological dynamics and metabolic capacity.

These trends are partially recapitulated in data set 2 (see Fig. S4 in the supplemental material). Specifically, 31 of the 171 operational taxonomic units (OTUs) in this data set were key contributors to at least one metabolite. Again, most metabolites (64%) had 4 or fewer key contributors, but the combination of OTUs varied across metabolites. Of the 42 metabolites analyzed in both data sets, 26 share at least one key contributing

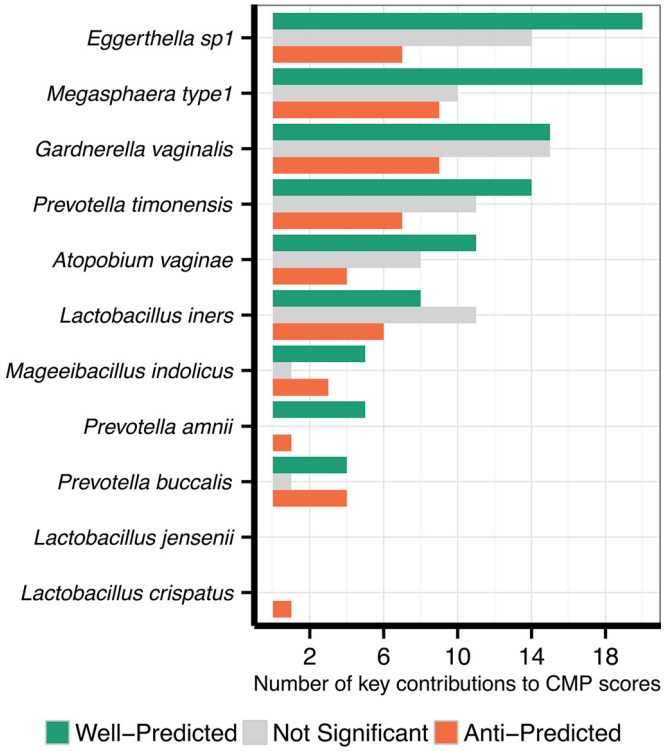

FIG 3 Key species contributors to metabolites in the vaginal microbiome. Each species that participated in the calculation of CMP scores in data set 1 is shown along the *y* axis. The *x* axis indicates the numbers of well-predicted and anti-predicted metabolites (as well as those with nonsignificant predictions) for which that species was a key contributor (see Materials and Methods).

genus, including 9 of the 11 metabolites that were well-predicted in both data sets (see Fig. S4B). Interestingly, however, the OTUs that contributed to the CMP scores of the most well-predicted metabolites in data set 2 included an OTU (227000) identified as BVAB1, an OTU (4377809) corresponding to the bacterium *Mageeibacillus indolicus* (previously known as BVAB3), an OTU corresponding to *Prevotella amnii* (663885), another *Prevotella* OTU (403822), and an OTU in the genus *Parvimonas* (132546) (of which only *M. indolicus* and *P. amnii* were analyzed in data set 1). An OTU corresponding to the *Eggerthella* species noted in data set 1 was also a key contributor to many well-predicted metabolites. Relatively low contributions to CMP scores by *Lactobacillus crispatus* (typically associated with health) and *Atopobium vaginae* were consistent between the two data sets. Given the difference in taxonomic profiling methods between the two data sets (qPCR versus 16S rRNA gene sequencing), the difference in the way genomic content was inferred (reference genomes versus PICRUSt-based predictions), missing reference genomic information for three species assayed in data set 1, and the focus on selected metabolites of interest in data set 2, the variation in the key contributors obtained is perhaps not surprising. For example, the increased importance of *M. indolicus* in data set 2 could be a function of differences in the features of metabolites assayed and analyzed between the two data sets, and/or it could be from differences in reference information; a total of 88 of 732 Kyoto Encyclopedia of Genes and Genomes (KEGG) orthology groups (KOs) differed in copy number between the reference genome used for prediction in data set 1 and the predicted genome content for the corresponding OTU in data set 2. The full list of key species contributors can be found in Table S2.

Well-predicted metabolites tend to be involved in condition-specific metabolism. We next set out to identify key gene contributors to each metabolite's CMP score, by calculating the correlation between the original CMP scores and a CMP score calculated when the link between the gene in question and the metabolite was deleted

from the metabolic model (Materials and Methods). Genes for which this correlation was <0.5 were considered key contributors for that metabolite, and any reaction catalyzed by the enzyme encoded by that gene was considered a key reaction contributor. This analysis relates specific combinations of reaction information and genomic shifts to the predicted potential for metabolite variation, allowing us to examine whether our approach recovers known metabolic mechanisms (see Fig. S3 in the supplemental material). For example, the CMP scores for well-predicted amino acid derivatives, including *N*-acetylputrescine and citrulline, were driven by synthesis enzymes forming part of amino acid catabolism pathways and encoded by BV-associated bacteria (see Fig. S3A and B). A subset of amino acids, including glutamate and phenylalanine, were well-predicted on the basis of a combination of available biosynthesis pathways and the predicted abundance of tRNA synthetase genes and degradation pathways. Pyruvate levels were slightly lower in BV samples and well-predicted primarily by acetolactate synthase, which catalyzes the first step diverting pyruvate to branched-chain amino acid synthesis. This mechanism is consistent with the overall shift from carbohydrate-based to amino acid-based metabolism that is typical of the BV state. In another example, Srinivasan et al. (22) have noted that the depletion of reduced glutathione in BV samples is surprising, as the BV vagina is a relatively reduced environment (40). Our framework predicts this shift in glutathione well in both data sets (prediction levels of 0.49 and 0.30 in data sets 1 and 2, respectively) and attributes it to a lack of glutathione peroxidase genes in *Lactobacillus* species that predominate in healthy vaginal samples (see Fig. S3C). Genes in cofactor synthesis pathways also tended to contribute to predictive CMP scores for metabolites in these pathways, including nicotinate, NAD^+, and FAD^+.

We further characterized the set of key gene contributors of each metabolite and explored their relationships with metabolite predictability (see Fig. S5 in the supplemental material). Most metabolites had only a small number of genes with the potential to enzymatically impact them, and of these, most were identified as key contributors. Interestingly, well-predicted metabolites tended to have a higher proportion of the set of relevant genes as key contributors in both data sets ($P = 0.002$ and $P = 0.09$ in data sets 1 and 2, respectively; Wilcoxon rank-sum test). Surprisingly, the key genes for well-predicted metabolites were less variable across samples than those for other metabolites ($P = 0.08$ and $P = 0.002$ in data sets 1 and 2, respectively; Wilcoxon rank-sum test). We also examined whether the key gene contributors for each metabolite encoded enzymes solely catalyzing reactions synthesizing the metabolite in question, degrading it, or both (see Materials and Methods). We found that BV-enriched metabolites with key gene contributors that are associated only with synthesis enzymes were almost always well-predicted (11 of 13 metabolites across both data sets) (Fig. 4; see also Fig. S6). Conversely, metabolites depleted in the BV state and with key gene contributors encoding only degradation enzymes also tended to be well-predicted (18 of 31 metabolites across both data sets) (Fig. 4; see also Fig. S6). These trends suggest that the most predictable variation resulted from the transition between these two conditions, in particular, the impact of the presence or absence of novel metabolic synthesis and degradation capacities in BV, rather than shifts in the abundance of more widely found metabolic pathways.

Application to gut microbiome communities reiterates metabolic trends and highlights community complexity. Finally, we explored the application of this framework to samples from gut microbial communities, bearing in mind the caveats of increased environmental influences resulting from diet, as well as increased community complexity. Specifically, we applied this framework to two additional data sets, one evaluating the impact of antibiotic treatment with cefoperazone on the cecal contents of specific-pathogen-free mice (data set 3) (23) and another profiling the microbiome and metabolome of humans with inflammatory bowel disease and healthy controls (data set 4) (15, 41). Because the second study used shotgun metagenomic sequencing, in its analysis, we did not need to predict metagenome content from community composition and instead estimated gene abundances directly (see Materials and

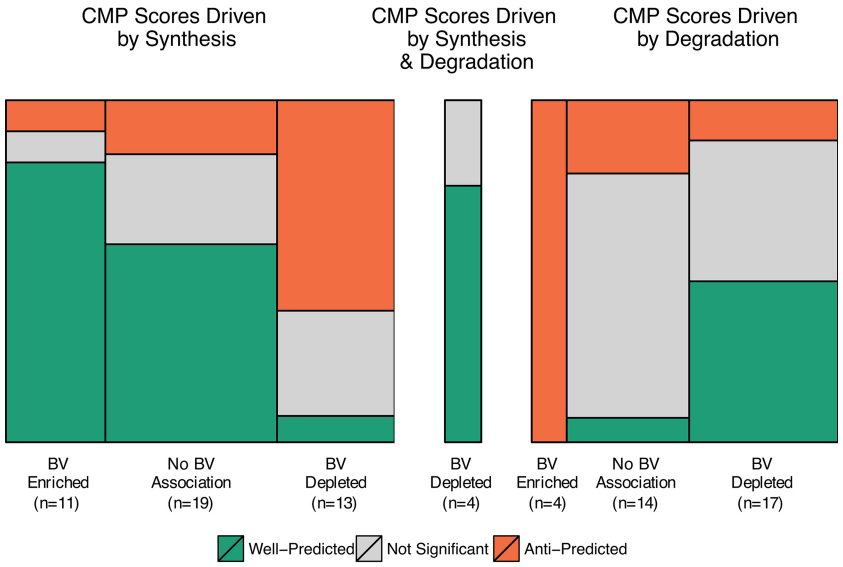

FIG 4 **Trends in metabolite predictability in terms of key gene contributors. Area plots depict the numbers of metabolites in data set 1 whose CMP scores are driven by synthesis, by degradation, or by both in relation to their association with the host state and their predictability. The width of each box corresponds to the number of metabolites associated with each host disease state (enriched in BV samples, depleted in BV samples, or neither), and the height corresponds to the number of metabolites that are well-predicted, anti-predicted, or not significantly predicted (also indicated by color). See Fig. S6 in the supplemental material for a similar plot describing metabolite prediction in data set 2.**

Methods) (42). As expected given the more complex and unstable community environment, we observed lower proportions of well-predicted metabolites in these data sets (see Fig. S7 in the supplemental material), but we also identified interesting patterns in the relationships between variation in community composition and metabolism in these settings.

Data set 3 recapitulated many metabolite predictability trends observed in our analysis of the vaginal microbiota, including successful prediction of metabolite variation and the effects of perturbation on community ecology and metabolism. In total, 39 of the 116 metabolites (33.6%) assayed and analyzed in this data set were well-predicted. Interestingly, we observed substantial overlap in the identities of the metabolites that were well-predicted and anti-predicted in this data set with those predicted similarly in data sets 1 and 2, as well as a general positive correlation between prediction levels across data sets (Fig. 5). One well-predicted metabolite of interest is gamma-aminobutyrate (GABA), which was enriched in the subset of samples from mice 6 weeks after antibiotic treatment. Key contributor analysis indicated that increased synthesis from 4-aminobutanal by an OTU in the genus *Oscillospira* and a *Clostridiales* OTU drove the CMP score variation for this metabolite. Several products of carbohydrate metabolism were also well-predicted, including the sugars stachyose and mannose. Analysis of key contributors revealed that the oligosaccharide stachyose is predicted on the basis of its depletion by glycosidases from diverse *Firmicutes* taxa, including *Ruminococcus* and *Turicibacter*, while mannose is predicted on the basis of increased production via glycan degradation from mannoglycans by several OTUs in the *Clostridiales* order in healthy samples. These shifts reflect the impacts of increased glycan degradation potential in the microbiome of mice from the control cohort compared to those treated with antibiotics. As in the BV data sets, synthesis products found to be more abundant in the more diverse microbiome of the control cohort were most likely to be predictable (48% of 29 such metabolites were well-predicted).

In data set 4, only a very low proportion of the metabolites analyzed were well-predicted (6 of 31), which is likely due to a markedly smaller sample size and potentially noisier metabolomic data and identifications. Interestingly, however, four of these six

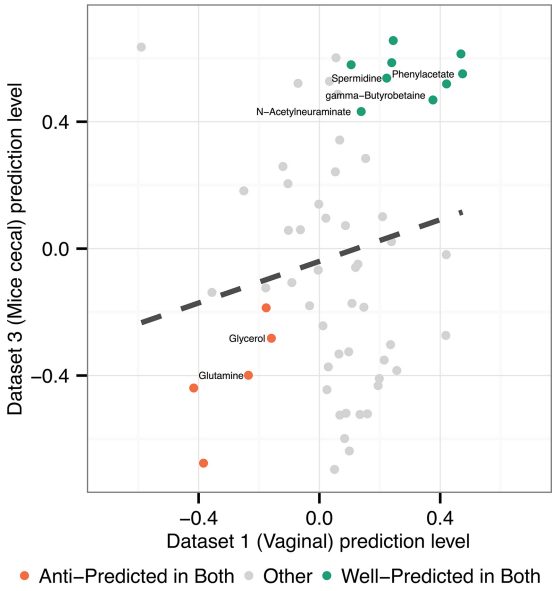

FIG 5 **Metabolite predictability is consistent between vaginal and mouse cecal data sets. The plot shows the relationship between the level of predictability for each metabolite (measured as the Spearman correlation between pairwise differences in calculated CMP scores and pairwise differences in measured metabolite abundances) in data set 1 (human vaginal microbiome samples) and data set 3 (mouse gut samples). Colors indicate metabolites that are well-predicted in both data sets or anti-predicted in both data sets. Metabolites that are well-predicted in both data sets are enriched for amino acid catabolites, including phenylacetate, spermidine, and beta-alanine.**

metabolites (chenodeoxycholate, glycochenodeoxycholate, glycocholate, and taurocholate) are primary conjugated or unconjugated bile acids, which form part of a tightly regulated pathway of host-microbial cometabolism with hormonal signaling functions. This enrichment of bile acid-associated products among the well-predicted metabolites ($P = 0.03$; Fisher exact test) highlights the important role of microbiome ecology in microbial metabolism of bile acids in the gut. Specifically, higher levels of bile acid metabolites in irritable bowel disease cases have been noted previously in this data set (41). Our results show that this shift in bile acids is concordant with variation in the abundances of microbial bile salt hydrolase genes.

DISCUSSION

Above, we have introduced a novel analytical framework that represents an important step toward a principled systematic and mechanistic integrative analysis of microbial community composition and metabolomic data. Our framework goes beyond *ad hoc* correlation-based analysis and aims to assess the correspondence between ecology and metabolic phenotype based on the existing body of knowledge about microbial genomes and metabolic capacities. By evaluating metabolite variation in terms of the functional implications of ecological shifts, we identified a large share of the vaginal metabolome whose variation can be explained by shifts in ecology-based and community-wide enzymatic potential. This high predictability is somewhat surprising, as our framework ignores many factors that could potentially impact this link, including strain variation, gene and protein expression, and metabolic fluxes (14, 43, 44). This finding suggests that ecological dynamics and their impact on community metabolic capacities likely play a major role in mediating broad metabolic differences between microbiomes.

Furthermore, our characterization of key species and gene contributors to calculated CMP scores and, consequently, to the predictability of each metabolite provides evidence that particular BV-associated species have substantial effects on the metabolome. By comprehensively identifying species whose enzymatic capacity and variation across samples are consistent with the observed shift in the abundance of a particular

metabolite, we were able to gain deeper insight into the drivers of species-metabolite dynamics in the vaginal microbiota and bacterial vaginosis. Specifically, our analysis of key species contributors identified a subset of BV-associated species (*Eggerthella* sp. 1, *Megasphaera* type 1, and *Mageeibacillus indolicus*) as particularly likely to be important drivers of metabolic variation in this environment. The low contributions of lactobacilli suggest that their importance in the vaginal microbial ecosystem is not described well by current reference knowledge of their role in canonical pathways. Alternatively, this can be attributed to having twice as many women with BV as without BV in our data sets, of which only some women without BV had abundant *L. crispatus*. More generally, we observed that the abundance of metabolic capacities (based on taxonomic composition) is often sufficient to explain measured variability in the abundance of many BV-associated metabolites. This intriguing result suggests that while information about ecological shifts may not necessarily provide a comprehensive understanding of changes in flux in core metabolic reactions, it is often sufficient to account for the accumulation or depletion of many more peripheral metabolites that vary most dramatically between health and dysbiosis.

We also extended this method to analyze data sets from the gut microbiota of mice and humans and identified preliminary mechanistic links in these complex environments. The lower predictability in this context likely reflected the greater complexity of these communities and the plethora of factors, both external and internal to the community, that can potentially affect metabolite abundances. Studying the impact of such factors on various metabolic processes is an important direction for future research. Nevertheless, the overlap observed here in the set of metabolites that are well-predicted across a single organism in culture (*E. coli*), a simple community (the vaginal microbiome), and a complex host-associated community (gut microbiome) may represent shared control points in microbial metabolic networks. This consistency indicates that across multiple environments, the limiting factor for accumulation or depletion of these metabolites is the presence or abundance of microbial enzymatic potential that can directly act on them. This finding further reinforces the credibility of our framework and the shared features of microbial metabolic regulation across all of these settings. In addition, the predictability of the metabolic shifts associated with major ecological perturbations across data sets is consistent with previous metabolic regulation findings that core reactions tend to be regulated by a precise balance between precursor metabolite concentrations and enzyme concentrations, and that intracellular concentrations of core metabolites are generally robust in response to perturbations (45, 46).

One obvious caveat of our framework and of the resulting findings is the inability to distinguish between failure to predict due to missing reference information (e.g., incomplete genome annotation) and failure due to a range of alternative mechanisms regulating metabolite shifts and environmental inputs and outputs. For example, our framework currently does not capture host metabolism, and future work may extend our model to include human gut metabolic processes. Similarly, our model does not consider signaling processes, transcriptional regulation, or bounds on metabolic fluxes. This limitation is further compounded by the use of a broad reaction database, such as KEGG. For example, our model only assigns effects for enzymes catalyzing nonreversible reactions. This approach presumably captures major metabolic fluxes for well-characterized microbes, but the information lost from reversible reactions may hinder our ability to predict metabolites in other pathways. An extended framework could, for example, infer reaction directionality from pathway context or constraint-based modeling, or directly from metabolomic data using a machine learning approach.

Such improvements could also help clarify the interpretation of anti-predicted metabolites, which spanned roughly a third of all predictions across data sets and can be explained by several potential mechanisms. Anti-predicted metabolites whose CMP scores are driven by degradation reactions, especially with downstream well-predicted metabolites, are suggestive of environmentally regulated metabolite changes that cause taxonomic shifts based on nutrient availability, such as the example of glycerol

anti-predicted by *L. iners*. Other anti-predicted metabolites may be explained by missing reaction information. For example, putrescine and cadaverine are both anti-predicted based on a high correlation with the abundance of genes coding for enzymes that utilize these metabolites to synthesize further polyamine derivatives (*N*-acetylputrescine and aminopropylcadaverine, respectively). This finding suggests that other enzymes that are currently not incorporated into our predictions (including synthesis reactions that are present but lacking information on reaction directionality) or enzymes from other, unmeasured microbes in these samples are likely important for driving the accumulation of these metabolites. In other cases, anti-prediction may suggest alternative metabolic mechanisms controlling metabolite variation, beyond direct enzymatic regulation.

Finally, the trends revealed by our analysis highlight the tight coordination of various metabolic processes even in the context of complex communities. Our framework evaluated each metabolite independently, but the resulting predictability trends, as well as evidence from other studies (14, 22), show that dramatic shifts in metabolite abundances occur in a strongly coordinated fashion, through a combination of changes in substrate and enzyme concentrations mediated by a variable range of taxa. The analysis framework presented here is an important first step toward deconstructing and interpreting these relationships in mechanistic detail from comprehensive multi-omic data. In turn, this mechanistic understanding will be vital to ultimately enable the rational design of strategies to modify the microbiome and its metabolic phenotype (47, 48).

MATERIALS AND METHODS

Assembling and processing data sets. We obtained several previously published data sets (15, 22, 23, 41, 45) from publicly available databases or through a collaboration, each pairing 16S rRNA gene-based taxonomic data with metabolomic profiles. For vaginal samples, DNA was extracted for 16S rRNA gene analysis from vaginal swabs, and cervicovaginal lavage fluid was collected for metabolomic analysis. Samples from the first data set (data set 1) were analyzed for taxonomic composition using quantitative PCR (qPCR) with primers and probes specific for 14 vaginal bacterial species and for metabolites using global liquid and gas chromatography coupled with mass spectrometry for metabolomics. Samples from the second data set (data set 2) were analyzed by using broad-range 16S rRNA gene PCR coupled with high-throughput 454 sequencing of the 16S rRNA gene (Roche) and targeted metabolomics using LC-MS with multiple-reaction monitoring for 180 compounds, chosen partially based on findings from data set 1. In data set 3, taxonomic composition was assayed using 454 FLX Titanium sequencing of V3-V5 regions of the 16S rRNA gene, and metabolites were measured using global LC-MS and GC-MS metabolomics. Data set 4 paired Roche 454 shotgun sequencing of sample DNA with Fourier transform ion cyclotron resonance mass spectrometry metabolomics. See Table S1 for details about each data set. Metabolite and transcript (microarray) data for the *E. coli* data set were downloaded from the supplementary material of reference 45 and from the NCBI GEO database, profiling *E. coli* grown in culture, treated to cause five different stress-based perturbations, and assayed before and after perturbation. We included only one time point before perturbation and one immediately after for each biological replicate in our analysis. We mapped identified metabolite names to KEGG identifiers (IDs) following the same approach as for data set 2.

We reprocessed 16S rRNA taxonomic sequencing data sets via a standard closed-reference OTU-picking pipeline using QIIME version 1.8.0 (49–53) and rarefied the resulting OTU tables to the number of reads in the lowest-coverage sample. For data set 2, we confirmed that this pipeline produced taxonomic profiles similar to the ones from the *pplacer* method used in the original publication (the Pearson correlation across samples of all genera quantified by both methods was 0.97). We did not normalize the 16S rRNA gene qPCR data of data set 1. A subset of samples in data set 1 also had associated 16S rRNA gene sequencing data, on the basis of which we removed one outlier sample whose sequencing results were dominated by a species not profiled by qPCR. For the *E. coli* gene expression data, we used the KEGG application programming interface (API) to map gene IDs to KEGG orthology groups (KOs) for consistency with the other data sets and used the normalized microarray intensities provided.

Processing of metabolomic data varied depending on the technology used. For data sets 1 and 3, in which metabolomic profiles were produced by Metabolon, Inc., and included detailed metabolite identifications, we filtered out compounds without KEGG identifications and used the raw peak area values. For data set 2 (generated at the Northwest Metabolomics Research Center) and the *E. coli* data set, we used the KEGG API (37) to associate named and measured compounds with KEGG metabolite IDs. For data set 4, which lacked confident library-based metabolite identifications, we used MetaboSearch (54) to perform a mass-based search against the Metlin, MMCD, LipidMaps, and HMDB databases (55–58), with a matching threshold of 1 ppm. The KEGG identifications with the smallest mass difference were assigned as the putative identification, following Tong et al. (24, 33). When multiple putative identifi-

cations had the same difference in mass from a peak, preference was given to metabolites in the metabolic network generated based on species abundances, using genomic information as additional support for the presence of that metabolite. When multiple putative KEGG identifications remained, one was randomly assigned to that peak. If multiple peaks mapped to the same KEGG metabolite ID, their abundances were summed. Metabolites with nonzero abundance in <5 samples were discarded from downstream analysis.

Predicting metagenome content from taxonomic composition. For data sets 2 and 3, we used PICRUSt (35) to predict metagenome content across samples, based on taxonomic composition, and normalized the resulting predictions using MUSiCC (release 1.0) (36). To predict genome content from the qPCR species abundances in data set 1, we searched IMG for available reference genomes. For 11 of the 14 species profiled, at least one reference genome was available. When multiple reference genomes were available for a given species, we selected either the highest quality genome or a genome from a vaginal isolate (in consultation with the researchers that generated this data set). See Table S2 in the supplemental material for details about the genomes used. We downloaded KEGG orthology (KO) annotations for these genomes from IMG (January 2014) and predicted the metagenome as a product of the reference genome KO annotations and species abundances. For data set 4, orthology group abundances were estimated directly from shotgun sequencing reads using a BLAST-based annotation pipeline (42).

Metabolic network reconstruction and CMP score calculation. We adapted the predicted relative metabolomic turnover (PRMT) method developed by Larsen et al. (31) to estimate the metabolic potential of a microbial community based on measurements of gene content. This method does not predict metabolite fluxes or concentrations directly; instead, it synthesizes and integrates information about gene abundances in terms of KEGG orthology groups and a stoichiometric matrix describing the quantitative relationship between genes and metabolites to provide an estimate of the way the community composition may impact each metabolite's abundance.

To this end, we first created a modified stoichiometric matrix M in which each row represents a particular metabolite and each column represents a particular gene (KO), such that each cell M_{ij} represents the combined relative capacity for enzymatic gene j to modify metabolite i (see reference 31). To create this matrix, we utilized pathway reaction information and stoichiometric coefficients from KEGG (37). Specifically, for each reaction catalyzed by an enzyme coded by gene x that transforms metabolite A into metabolite B with stoichiometric coefficients c and d, respectively, we subtract c from M_{Ax} and add d to M_{Bx}. To focus our analysis on the primary transformations catalyzed by each enzyme, we only linked genes to reactions and metabolites that are annotated in KEGG metabolic pathways, using the *reaction_mapformula.lst* file from the KEGG database (2013 version). We then filtered this matrix to only include reactions annotated as occurring in a single direction, ignoring all reversible reactions that cannot contribute to metabolite predictions and all metabolites involved in only reversible reactions. Lastly, we performed two additional modifications: first, following previous studies (59, 60), we excluded "currency" metabolites that are involved in reactions associated with 30 or more genes from the final matrix, and second, following Larsen et al. (31), we normalized each row of M such that all negative elements sum to -1 and all positive elements sum to 1.

The resulting matrix accordingly estimates the relative contribution of each gene to the accumulation or depletion of each metabolite. We then multiply this matrix M with a matrix G that represents the abundance of each gene in each sample to obtain communitywide metabolic potential (CMP) scores, capturing the relative capacity of the metagenome content of each sample to create or deplete each metabolite.

Comparing CMP scores with metabolomic data. Notably, since CMP scores represent relative predictions, they can only be interpreted in the context of comparisons between samples (assuming some baseline metabolite profile across samples). Accordingly, to assess how the CMP scores obtained compare to the metabolomic variation measured, we performed a Mantel test for each metabolite, assessing the correlation between pairwise differences (across all pairs of samples) in CMP scores and the corresponding pairwise differences in measured metabolite abundances. We further corrected for multiple hypothesis testing using a local false discovery rate (FDR) approach implemented in the R package *qvalue* (61) and classified metabolites with both a Mantel P value and FDR q value of less than 0.01 as well-predicted. We evaluated the significance of negative pairwise correlations in a similar manner, classifying metabolites as anti-predicted based on the same significance cutoffs.

Testing significance with randomly shuffled networks and metabolite labels. Given the covariance structure of the data set, we also wished to quantify whether our framework identified more well-predicted metabolites than expected by chance. To this end, we repeatedly generated randomized metabolic networks, ran our framework as detailed above using the randomized network to link genes to metabolites, and compared the number of well-predicted metabolites obtained with these randomized networks to the number of well-predicted metabolites obtained with the original network. To preserve the core structural characteristics of the original network, random networks were generated following the edge-shuffling approach outlined in reference 62 (exchanging edges 5,000 times to produce each network).

We also used a permutation-based approach to evaluate whether anti-predicted metabolites are linked by a metabolic reaction to well-predicted metabolites more frequently than expected by chance. To this end, we repeatedly permuted the labels of every metabolite in the network while maintaining a fixed network topology. We counted the number of times an anti-predicted metabolite was connected to a well-predicted metabolite by a synthesizing reaction, a depleting reaction, or a reversible reaction

in these permuted networks and compared the resulting distribution with the numbers obtained using the original data.

Identifying key species and gene contributors. To quantify the contribution of each species to the calculated CMP score of each metabolite and to identify key species contributors, we examined the Pearson correlation between the CMP scores obtained for a given metabolite across samples using the entire community and the CMP score calculated based on each species by itself (i.e., recalculating the metagenome content and CMP scores based solely on the abundance of each species separately). Species for which this correlation coefficient for a given metabolite was >0.5 were considered key species contributors for that metabolite.

To compare key species contributors between data sets 1 and 2, we identified corresponding species across the two data sets by searching the Greengenes 97% OTU representative sequence set for exact matches with the PCR primers used by Srinivasan et al. (22) to generate data set 1. Notably, this mapping identified OTU 4377809 as *Mageeibacillus indolicus* (previously known as BVAB3), OTU 227000 (mistakenly characterized as *Shuttleworthia*) as BVAB1, and OTU 133178 as *Eggerthella* sp. 1.

To identify key gene contributors to the calculated CMP scores for each metabolite, we examined the Pearson correlation between the CMP scores obtained for a given metabolite across samples using the original stoichiometric matrix and the CMP scores calculated when using a matrix in which the link between the gene in question and the metabolite was deleted (i.e., zeroing the corresponding entry in the matrix). Genes for which this correlation was <0.5 were considered key contributors for that metabolite. We further defined any reaction catalyzed by the enzyme coded by that gene as a key reaction contributor. In addition, if all of the key reaction contributors of a given metabolite produce that metabolite, we classified that metabolite's CMP scores as driven primarily by synthesis. We similarly classified a metabolite whose key reaction contributors all deplete that metabolite as driven primarily by degradation.

Data availability. All data sets analyzed in this paper are from published work, and the relevant sequence data can be found at NCBI under accession numbers SRA051298 (data set 1), SRP056030 (data set 2), SRP033403 (data set 3), and PRJNA46321 (data set 4). The *E. coli* data analyzed are available from the supporting information of the pertaining publication (45) and from NCBI GEO series GSE20305. Taxonomic and metabolomic profiles for each of the data sets analyzed in this work, as well as the code for our framework and analysis, are available at: http://elbo.gs.washington.edu/download.html.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at http://dx.doi.org/10.1128/mSystems.00013-15.

Figure S1, TIFF file, 0.5 MB.
Figure S2, TIFF file, 1.2 MB.
Figure S3, TIFF file, 1.0 MB.
Figure S4, TIFF file, 4.9 MB.
Figure S5, TIFF file, 4.0 MB.
Figure S6, TIFF file, 0.8 MB.
Figure S7, TIFF file, 0.8 MB.
Table S1, PDF file, 0.4 MB.
Table S2, XLSX file, 0.1 MB.
Text S1, DOCX file, 0.1 MB.

ACKNOWLEDGMENTS

C.N. and E.B. conceived and designed the research. S.S., C.M.T., J.K.J., V.B.Y., and D.N.F. provided data and advice on data processing and interpretation. A.E. helped process the data. C.N. performed the research. C.N. and E.B. wrote the paper. All authors reviewed the paper and provided comments.

We thank Keith Bayer and Colin Brislawn for technical support in obtaining various data sets. We are grateful to all members of the Borenstein lab for helpful discussions.

FUNDING INFORMATION

National Institutes of Health (NIH) provided funding to Elhanan Borenstein under grant number DP2 AT007802-01. National Institutes of Health (NIH) provided funding to David N. Fredricks under grant number R01 HG-005816. National Institutes of Health (NIH) provided funding to David N. Fredricks under grant number R01 AI-061628. National Institutes of Health (NIH) provided funding to Casey M. Theriot under grant number K01GM109236. National Science Foundation (NSF) provided funding to Cecilia Noecker under grant number IGERT DGE-1258485.

REFERENCES

1. **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. http://dx.doi.org/10.1016/j.chom.2014.11.003.
2. **Smith PM, Howitt MR, Panikov N, Michaud M, Gallini CA, Bohlooly-Y M, Glickman JN, Garrett WS**. 2013. The microbial metabolites, short-chain fatty acids, regulate colonic Treg cell homeostasis. Science **341:**569–573. http://dx.doi.org/10.1126/science.1241165.
3. **Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI**. 2006. An obesity-associated gut microbiome with increased capacity for energy harvest. Nature **444:**1027–1131. http://dx.doi.org/10.1038/nature05414.
4. **Cox LM, Yamanishi S, Sohn J, Alekseyenko AV, Leung JM, Cho I, Kim SG, Li H, Gao Z, Mahana D, Zárate Rodriguez JG, Rogers AB, Robine N, Loke P, Blaser MJ**. 2014. Altering the intestinal microbiota during a critical developmental window has lasting metabolic consequences. Cell **158:**705–721. http://dx.doi.org/10.1016/j.cell.2014.05.052.
5. **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 WH, 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. http://dx.doi.org/10.1038/nm.3145.
6. **Stefka AT, Feehley T, Tripathi P, Qiu J, McCoy K, Mazmanian SK, Tjota MY, Seo G-Y, Cao S, Theriault BR, Antonopoulos DA, Zhou L, Chang EB, Fu Y-X, Nagler CR**. 2014. Commensal bacteria protect against food allergen sensitization. Proc Natl Acad Sci U S A **111:**13145–13150. http://dx.doi.org/10.1073/pnas.1412008111.
7. **Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, Creasy HH, Earl AM, FitzGerald MG, Fulton RS, Giglio MG, Hallsworth-Pepin K, Lobos EA, Madupu R, Magrini V, Martin JC, Mitreva M, Muzny DM, Sodergren EJ, Versalovic J, Wollam AM, Worley KC, Wortman JR, Young SK, Zeng Q, Aagaard KM, Abolude OO, Allen-Vercoe E, Alm EJ, Alvarado L, Andersen GL, Anderson S, Appelbaum E, Arachchi HM, Armitage G, Arze CA, Ayvaz T, Baker CC, Begg L, Belachew T, Bhonagiri V, Bihan M, Blaser MJ, Bloom T, Bonazzi V, Brooks J, Buck GA, Buhay CJ, Busam DA, Campbell JL, Canon SR, et al.** 2012. Structure, function and diversity of the healthy human microbiome. Nature **486:**207–214. http://dx.doi.org/10.1038/nature11234.
8. **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**. 2014. Diet rapidly and reproducibly alters the human gut microbiome. Nature **505:**559–563. http://dx.doi.org/10.1038/nature12820.
9. **Pérez-Cobas AE, Gosalbes MJ, Friedrichs A, Knecht H, Artacho A, Eismann K, Otto W, Rojo D, Bargiela R, von Bergen M, Neulinger SC, Däumer C, Heinsen F-A, Latorre A, Barbas C, Seifert J, dos Santos VM, Ott SJ, Ferrer M, Moya A**. 2013. Gut microbiota disturbance during antibiotic therapy: a multi-omic approach. Gut **62:**1591–1601. http://dx.doi.org/10.1136/gutjnl-2012-303184.
10. **Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, Liang S, Zhang W, Guan Y, Shen D, Peng Y, Zhang D, Jie Z, Wu W, Qin Y, Xue W, Li J, Han L, Lu D, Wu P, Dai Y, Sun X, Li Z, Tang A, Zhong S, Li X, Chen W, Xu R, Wang M, Feng Q, Gong M, Yu J, Zhang Y, Zhang M, Hansen T, Sanchez G, Raes J, Falony G, Okuda S, Almeida M, LeChatelier E, Renault P, Pons N, Batto J-M, Zhang Z, Chen H, Yang R, Zheng W, Li S, Yang H, Wang J, Ehrlich SD, Nielsen R, Pedersen O, Kristiansen K, Wang J**. 2012. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature **490:**55–60. http://dx.doi.org/10.1038/nature11450.
11. **Morgan XC, Tickle TL, Sokol H, Gevers D, Devaney KL, Ward DV, Reyes JA, Shah SA, LeLeiko N, Snapper SB, Bousvaros A, Korzenik J, Sands BE, Xavier RJ, Huttenhower C**. 2012. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biol **13:**R79. http://dx.doi.org/10.1186/gb-2012-13-9-r79.
12. **Cowan TE, Palmnäs MS, Yang J, Bomhof MR, Ardell KL, Reimer RA, Vogel HJ, Shearer J**. 2014. Chronic coffee consumption in the diet-induced obese rat: impact on gut microbiota and serum metabolomics. J Nutr Biochem **25:**489–495. http://dx.doi.org/10.1016/j.jnutbio.2013.12.009.
13. **Daniel H, Moghaddas Gholami A, Berry D, Desmarchelier C, Hahne H, Loh G, Mondot S, Lepage P, Rothballer M, Walker A, Böhm C, Wenning M, Wagner M, Blaut M, Schmitt-Kopplin P, Kuster B, Haller D, Clavel T**. 2014. High-fat diet alters gut microbiota physiology in mice. ISME J **8:**295–308. http://dx.doi.org/10.1038/ismej.2013.155.
14. **Edlund A, Yang Y, Yooseph S, Hall AP, Nguyen DD, Dorrestein PC, Nelson KE, He X, Lux R, Shi W, McLean JS**. 2015. Meta-omics uncover temporal regulation of pathways across oral microbiome genera during in vitro sugar metabolism. ISME J **9:**2605–2619 http://dx.doi.org/10.1038/ismej.2015.72.
15. **Erickson AR, Cantarel BL, Lamendella R, Darzi Y, Mongodin EF, Pan C, Shah M, Halfvarson J, Tysk C, Henrissat B, Raes J, Verberkmoes NC, Fraser CM, Hettich RL, Jansson JK**. 2012. Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's disease. PLoS One **7:**e49138. http://dx.doi.org/10.1371/journal.pone.0049138.
16. **He B, Nohara K, Ajami NJ, Michalek RD, Tian X, Wong M, Losee-Olson SH, Petrosino JF, Yoo S-H, Shimomura K, Chen Z**. 2015. Transmissible microbial and metabolomic remodeling by soluble dietary fiber improves metabolic homeostasis. Sci Rep **5:**10604. http://dx.doi.org/10.1038/srep10604.
17. **Kim Y-M, Nowack S, Olsen MT, Becraft ED, Wood JM, Thiel V, Klapper I, Kühl M, Fredrickson JK, Bryant DA, Ward DM, Metz TO**. 2015. Diel metabolomics analysis of a hot spring chlorophototrophic microbial mat leads to new hypotheses of community member metabolisms. Microb Physiol Metab **6:**209. http://dx.doi.org/10.3389/fmicb.2015.00209.
18. **Lu K, Abo RP, Schlieper KA, Graffam ME, Levine S, Wishnok JS, Swenberg JA, Tannenbaum SR, Fox JG**. 2014. Arsenic exposure perturbs the gut microbiome and its metabolic profile in mice: an integrated metagenomics and metabolomics analysis. Environ Health Perspect **122:**284–291 http://dx.doi.org/10.1289/ehp.1307429.
19. **Mao S-Y, Huo W-J, Zhu WY**. 4 December 2014. Microbiome-metabolome analysis reveals unhealthy alterations in the composition and metabolism of ruminal microbiota with increasing dietary grain in a goat model. Environ Microbiol http://dx.doi.org/10.1111/1462-2920.12724.
20. **Marcobal A, Kashyap PC, Nelson TA, Aronov PA, Donia MS, Spormann A, Fischbach MA, Sonnenburg JL**. 2013. A metabolomic view of how the human gut microbiota impacts the host metabolome using humanized and gnotobiotic mice. ISME J **7:**1933–1943. http://dx.doi.org/10.1038/ismej.2013.89.
21. **Shankar V, Homer D, Rigsbee L, Khamis HJ, Michail S, Raymer M, Reo NV, Paliy O**. 2015. The networks of human gut microbe-metabolite associations are different between health and irritable bowel syndrome. ISME J **9:**1899–1903. http://dx.doi.org/10.1038/ismej.2014.258.
22. **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. http://dx.doi.org/10.1128/mBio.00204-15.
23. **Theriot CM, Koenigsknecht MJ, Carlson PE, Jr, Hatton GE, Nelson AM, Li B, Huffnagle GB, Z Li J, Young VB**. 2014. Antibiotic-induced shifts in the mouse gut microbiome and metabolome increase susceptibility to *Clostridium difficile* infection. Nat Commun **5:**3114. http://dx.doi.org/10.1038/ncomms4114.
24. **Tong M, McHardy I, Ruegger P, Goudarzi M, Kashyap PC, Haritunians T, Li X, Graeber TG, Schwager E, Huttenhower C, Fornace AJ, Sonnenburg JL, McGovern DP, Borneman J, Braun J**. 2014. Reprograming of gut microbiome energy metabolism by the FUT2 Crohn's disease risk polymorphism. ISME J **8:**2193–2206. http://dx.doi.org/10.1038/ismej.2014.64.
25. **Walker A, Pfitzner B, Neschen S, Kahle M, Harir M, Lucio M, Moritz F, Tziotis D, Witting M, Rothballer M, Engel M, Schmid M, Endesfelder D, Klingenspor M, Rattei T, Castell WZ, de Angelis MH, Hartmann A, Schmitt-Kopplin P**. 2014. Distinct signatures of host-microbial meta-metabolome and gut microbiome in two C57BL/6 strains under high-fat diet. ISME J **8:**2380–2396. http://dx.doi.org/10.1038/ismej.2014.79.
26. **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. http://dx.doi.org/10.1371/journal.pone.0070803.

27. **Zhang Y, Zhao F, Deng Y, Zhao Y, Ren H**. 2015. Metagenomic and metabolomic analysis of the toxic effects of trichloroacetamide-induced gut microbiome and urine metabolome perturbations in mice. J Proteome Res **14:**1752–1761 http://dx.doi.org/10.1021/pr5011263.
28. **Heinken A, Thiele I**. 2015. Systematic prediction of health-relevant human-microbial co-metabolism through a computational framework. Gut Microbes **6:**120–130. http://dx.doi.org/10.1080/19490976.2015.1023494.
29. **Shoaie S, Ghaffari P, Kovatcheva-Datchary P, Mardinoglu A, Sen P, Pujos-Guillot E, de Wouters T, Juste C, Rizkalla S, Chilloux J, Hoyles L, Nicholson JK, MICRO-Obes Consortium, DoreJ, Dumas ME, Clement K, Bäckhed F, Nielsen J**. 2015. Quantifying diet-induced metabolic changes of the human gut microbiome. Cell Metab **22:**320–331. http://dx.doi.org/10.1016/j.cmet.2015.07.001.
30. **Chiu H-C, Levy R, Borenstein E**. 2014. Emergent biosynthetic capacity in simple microbial communities. PLoS Comput Biol **10:**e1003695. http://dx.doi.org/10.1371/journal.pcbi.1003695.
31. **Larsen PE, Collart FR, Field D, Meyer F, Keegan KP, Henry CS, McGrath J, Quinn J, Gilbert JA**. 2011. Predicted relative metabolomic turnover (PRMT): determining metabolic turnover from a coastal marine metagenomic d*ataset. Microb Inform Exp **1:**4. http://dx.doi.org/10.1186/2042-5783-1-4.
32. **Larsen PE, Dai Y**. 2015. Metabolome of human gut microbiome is predictive of host dysbiosis. Gigascience **4:**42. http://dx.doi.org/10.1186/s13742-015-0084-3.
33. **McHardy IH, Goudarzi M, Tong M, Ruegger PM, Schwager E, Weger JR, Graeber TG, Sonnenburg JL, Horvath S, Huttenhower C, McGovern DP, Fornace AJ, Borneman J, Braun J**. 2013. Integrative analysis of the microbiome and metabolome of the human intestinal mucosal surface reveals exquisite inter-relationships. Microbiome **1:**17. http://dx.doi.org/10.1186/2049-2618-1-17.
34. **Sridharan GV, Choi K, Klemashevich C, Wu C, Prabakaran D, Pan LB, Steinmeyer S, Mueller C, Yousofshahi M, Alaniz RC, Lee K, Jayaraman A**. 2014. Prediction and quantification of bioactive microbiota metabolites in the mouse gut. Nat Commun **5:**5492. http://dx.doi.org/10.1038/ncomms6492.
35. **Langille MG, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, Clemente JC, Burkepile DE, Vega Thurber RL, Knight R, Beiko RG, Huttenhower C**. 2013. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol **31:**814–821. http://dx.doi.org/10.1038/nbt.2676.
36. **Manor O, Borenstein E**. 2015. MUSiCC: a marker genes based framework for metagenomic normalization and accurate profiling of gene abundances in the microbiome. Genome Biol **16:**53. http://dx.doi.org/10.1186/s13059-015-0610-8.
37. **Kanehisa M, Goto S**. 2000. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res **28:**27–30. http://dx.doi.org/10.1093/nar/28.1.27.
38. **Wolrath H, Forsum U, Larsson PG, Borén H**. 2001. Analysis of bacterial vaginosis-related amines in vaginal fluid by gas chromatography and mass spectrometry. J Clin Microbiol **39:**4026–4031. http://dx.doi.org/10.1128/JCM.39.11.4026-4031.2001.
39. **Macklaim JM, Fernandes AD, Di Bella JM, Hammond J-A, Reid G, Gloor GB**. 2013. Comparative meta-RNA-seq of the vaginal microbiota and differential expression by Lactobacillus iners in health and dysbiosis. Microbiome **1:**12. http://dx.doi.org/10.1186/2049-2618-1-12.
40. **Holmes KK, Chen KC, Lipinski CM, Eschenbach DA**. 1985. Vaginal redox potential in bacterial vaginosis (nonspecific vaginitis). J Infect Dis **152:**379–382. http://dx.doi.org/10.1093/infdis/152.2.379.
41. **Jansson J, Willing B, Lucio M, Fekete A, Dicksved J, Halfvarson J, Tysk C, Schmitt-Kopplin P**. 2009. Metabolomics reveals metabolic biomarkers of Crohn's disease. PLoS One **4:**e6386. http://dx.doi.org/10.1371/journal.pone.0006386.
42. **Carr R, Borenstein E**. 2014. Comparative analysis of functional metagenomic annotation and the mappability of short reads. PLoS One **9:**e105776. http://dx.doi.org/10.1371/journal.pone.0105776.
43. **Sonnenburg JL, Xu J, Leip DD, Chen C-H, Westover BP, Weatherford J, Buhler JD, Gordon JI**. 2005. Glycan foraging in vivo by an intestine-adapted bacterial symbiont. Science **307:**1955–1959. http://dx.doi.org/10.1126/science.1109051.
44. **Shi W, Moon CD, Leahy SC, Kang D, Froula J, Kittelmann S, Fan C, Deutsch S, Gagic D, Seedorf H, Kelly WJ, Atua R, Sang C, Soni P, Li D, Pinares-Patiño CS, McEwan JC, Janssen PH, Chen F, Visel A, Wang Z, Attwood G, Rubin E**. 2014. Methane yield phenotypes linked to differential gene expression in the sheep rumen microbiome. **24:**1517–1525. Genome Res. http://dx.doi.org/10.1101/gr.168245.113.
45. **Jozefczuk S, Klie S, Catchpole G, Szymanski J, Cuadros-Inostroza A, Steinhauser D, Selbig J, Willmitzer L**. 2010. Metabolomic and transcriptomic stress response of *Escherichia coli*. Mol Syst Biol **6:**364. http://dx.doi.org/10.1038/msb.2010.18.
46. **Ishii N, Nakahigashi K, Baba T, Robert M, Soga T, Kanai A, Hirasawa T, Naba M, Hirai K, Hoque A, Ho PY, Kakazu Y, Sugawara K, Igarashi S, Harada S, Masuda T, Sugiyama N, Togashi T, Hasegawa M, Takai Y, Yugi K, Arakawa K, Iwata N, Toya Y, Nakayama Y, Nishioka T, Shimizu K, Mori H, Tomita M**. 2007. Multiple high-throughput analyses monitor the response of *E. coli* to perturbations. Science **316:**593–597. http://dx.doi.org/10.1126/science.1132067.
47. **Waldor MK, Tyson G, Borenstein E, Ochman H, Moeller A, Finlay BB, Kong HH, Gordon JI, Nelson KE, Dabbagh K, Smith H**. 2015. Where next for microbiome research? PLoS Biol **13:**e1002050. http://dx.doi.org/10.1371/journal.pbio.1002050.
48. **Greenblum S, Chiu H-C, Levy R, Carr R, Borenstein E**. 2013. Towards a predictive systems-level model of the human microbiome: progress, challenges, and opportunities. Curr Opin Biotechnol **24:**810–820. http://dx.doi.org/10.1016/j.copbio.2013.04.001.
49. **Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R**. 2010. QIIME allows analysis of high-throughput community sequencing data. Nat Methods **7:**335–336. http://dx.doi.org/10.1038/nmeth.f.303.
50. **Caporaso JG, Bittinger K, Bushman FD, DeSantis TZ, Andersen GL, Knight R**. 2010. PyNAST: a flexible tool for aligning sequences to a template alignment. Bioinformatics **26:**266–267. http://dx.doi.org/10.1093/bioinformatics/btp636.
51. **Edgar RC**. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics **26:**2460–2461. http://dx.doi.org/10.1093/bioinformatics/btq461.
52. **McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, Hugenholtz P**. 2012. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J **6:**610–618. http://dx.doi.org/10.1038/ismej.2011.139.
53. **DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL**. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol **72:**5069–5072. http://dx.doi.org/10.1128/AEM.03006-05.
54. **Zhou B, Wang J, Ressom HW**. 2012. MetaboSearch: tool for mass-based metabolite identification using multiple databases. PLoS One **7:**e40096. http://dx.doi.org/10.1371/journal.pone.0040096.
55. **Smith CA, O'Maille G, Want EJ, Qin C, Trauger SA, Brandon TR, Custodio DE, Abagyan R, Siuzdak G**. 2005. METLIN: a metabolite mass spectral database. Ther Drug Monit **27:**747–751. http://dx.doi.org/10.1097/01.ftd.0000179845.53213.39.
56. **Sud M, Fahy E, Cotter D, Brown A, Dennis EA, Glass CK, Merrill AH, Murphy RC, Raetz CR, Russell DW, Subramaniam S**. 2007. LMSD: LIPID MAPS Structure Database. Nucleic Acids Res **35:**D527–D532. http://dx.doi.org/10.1093/nar/gkl838.
57. **Wishart DS, Knox C, Guo AC, Eisner R, Young N, Gautam B, Hau DD, Psychogios N, Dong E, Bouatra S, Mandal R, Sinelnikov I, Xia J, Jia L, Cruz JA, Lim E, Sobsey CA, Shrivastava S, Huang P, Liu P, Fang L, Peng J, Fradette R, Cheng D, Tzur D, Clements M, Lewis A, De Souza A, Zuniga A, Dawe M, Xiong Y, Clive D, Greiner R, Nazyrova A, Shaykhutdinov R, Li L, Vogel HJ, Forsythe I**. 2009. HMDB: a knowledgebase for the human metabolome. Nucleic Acids Res **37:**D603–D610. http://dx.doi.org/10.1093/nar/gkn810.
58. **Cui Q, Lewis IA, Hegeman AD, Anderson ME, Li J, Schulte CF, Westler WM, Eghbalnia HR, Sussman MR, Markley JL**. 2008. Metabolite identification via the Madison Metabolomics Consortium Database. Nat Biotechnol **26:**162–164. http://dx.doi.org/10.1038/nbt0208-162.
59. **Greenblum S, Turnbaugh PJ, Borenstein E**. 2012. Metagenomic systems biology of the human gut microbiome reveals topological shifts

associated with obesity and inflammatory bowel disease. Proc Natl Acad Sci U S A **109:**594–599. http://dx.doi.org/10.1073/pnas.1116053109.

60. **Taxis TM, Wolff S, Gregg SJ, Minton NO, Zhang C, Dai J, Schnabel RD, Taylor JF, Kerley MS, Pires JC, Lamberson WR, Conant GC**. 2015. The players may change but the game remains: network analyses of ruminal microbiomes suggest taxonomic differences mask functional similarity. Nucleic Acids Res **43:**9600–9612. http://dx.doi.org/10.1093/nar/gkv973.
61. **Storey JD, Bass AJ, Dabney A, Robinson D**. 2015. qvalue: Q-value estimation for false discovery rate control. R package version 2.2.0. Bioconductor version 3.2. http://www.bioconductor.org/packages/3.2/bioc/html/qvalue.html.
62. **Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U**. 2002. Network motifs: simple building blocks of complex networks. Science **298:**824–827. http://dx.doi.org/10.1126/science.298.5594.824.