MIMOSA2: A metabolic network-based tool for inferring mechanism-supported relationships in microbiome-metabolome data

Cecilia Noecker¹⁺, Alexander Eng¹, and Elhanan Borenstein^{2,3,4*}

¹Department of Genome Sciences, University of Washington, Seattle, WA, 98195, USA

²Blavatnik School of Computer Science, Tel Aviv University, 6997801, Tel Aviv, Israel

³Department of Clinical Microbiology and Immunology, Sackler Faculty of Medicine, Tel Aviv University, 6997801, Tel Aviv, Israel

⁴Santa Fe Institute, Santa Fe, NM, 87501, USA

⁺Present address: Department of Microbiology and Immunology, University of California San Francisco, San Francisco, CA, 94143, USA

*corresponding author, elbo@tauex.tau.ac.il

Abstract

Motivation: Recent technological developments have facilitated an expansion of microbiome-metabolome studies, in which a set of microbiome samples are assayed using both genomic and metabolomic technologies to characterize the composition of microbial taxa and the concentrations of various metabolites. A common goal of many of these studies is to identify microbial features (species or genes) that contribute to differences in metabolite levels across samples. Previous work indicated that integrating these datasets with reference knowledge on microbial metabolic capacities may enable more precise and confident inference of such microbe-metabolite links.

Results: We present MIMOSA2, an R package and web application for model-based integrative analysis of microbiome-metabolome datasets. MIMOSA2 uses reference databases to construct a community metabolic model based on microbiome data and uses this model to predict differences in metabolite levels across samples. These predictions are compared with metabolomics data to identify putative microbiome-governed metabolites and specific taxonomic contributors to metabolite variation. MIMOSA2 supports various input data types and can be customized to incorporate user-defined metabolic pathways. We demonstrate MIMOSA2's ability to identify ground truth microbial mechanisms in simulation datasets, and compare its results with experimentally inferred mechanisms in a dataset describing honeybee gut microbiota. Overall, MIMOSA2 combines reference databases, a validated statistical framework, and a user-friendly interface to facilitate modeling and evaluating relationships between members of the microbiota and their metabolic products.

35 **Availability and Implementation:** MIMOSA2 is implemented in R under the GNU
36 General Public License v3.0 and is freely available as a web server and R package from
37 www.borensteinlab.com/software_MIMOSA2.html.
38 **Keywords:** microbiome-metabolome, metabolic models, web application, metagenomics,
39 metabolomics

Background

Microbial community metabolism contributes to global nutrient cycling (1), metabolic dysregulation in human disease (2), and detoxification of pollutants (3), among other crucial processes. A growing number of studies investigate these processes by profiling the composition and metabolism of host-associated and environmental microbial communities using genomic and metabolomic technologies (4). Such studies commonly seek to answer two important questions: whether metabolic differences between environments (e.g. between healthy and diseased settings) can be attributed to differences in microbial composition and ecology, and if so, which specific microbial community members might be the key players generating such differences.

Broad surveys of microbial taxa and metabolites, which we refer to here as “microbiome-metabolome studies”, have great potential utility to answer these two questions by uncovering associations between taxa and metabolite abundances across samples. However, we have recently shown that calculating univariate correlations between taxa and metabolites may not identify mechanistic links between them with very high accuracy, depending on properties of the communities under study (5). Comparisons of taxon-metabolite associations with metabolite production in monoculture have also suggested a high false positive rate for this approach (6). Some recently introduced tools address this challenge by using machine learning to infer complex models predicting metabolite abundances from microbial taxa (7–9). However, these can be difficult to interpret in terms of possible mechanisms, and in general do not account for or incorporate prior knowledge of microbial metabolism.

An alternative or complementary approach is to use metabolic reference databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) (10), as well as collections of genome-scale metabolic reconstructions such as AGORA (11) and *embl_gems* (12), to generate

and evaluate more precise hypotheses on the relationships between microbes and metabolites. We previously demonstrated that this conceptual approach can lead to new insights and released a preliminary but widely used R package (13), MIMOSA (Model-based Integration of Metabolite Observations and Species Abundances), for performing these analyses (14–19). Other methods to integrate microbiome and metabolome data with metabolic reference databases have also recently been introduced (20–23). To date, however, these tools all have some limitations in usability and/or compatibility with common data types and reference databases. Furthermore, none have been systematically evaluated for their ability to accurately infer direct mechanistic links between microbes and metabolites.

Here, we introduce MIMOSA2, an R package and web application (http://borensteinlab.com/software_MIMOSA2.html) for model-based integration of microbiome and metabolome data. MIMOSA2 uses metabolic reference databases to analyze paired microbiome and metabolite profiles, identifying metabolite differences that can be explained by specific microbiome features. MIMOSA2 expands and improves on the original proof-of-concept version of MIMOSA in several ways: First, it applies a new statistical algorithm for quantifying links between taxa and metabolites, which we have validated using simulated microbiome-metabolome data and experimental comparisons. It also implements several distinct strategies for constructing community metabolic models from different types of microbiome data and reference databases. Finally, it can be run either locally or via a web server with greatly improved ease of use, flexibility, and documentation.

Implementation

MIMOSA2 integrates paired microbiome-metabolome datasets with reference reaction databases to generate specific mechanistic hypotheses. It constructs community metabolic

models using microbiome compositional data and a reaction database, assesses whether measured metabolite concentrations are consistent with estimated community metabolic potential (CMP) across a set of samples, and identifies specific taxa and reactions that can explain observed metabolite variation (Figure 1). Below, we describe the analysis workflow steps and how to run an analysis using the web application.

Figure 1. Summary of the MIMOSA2 analysis pipeline. In a MIMOSA2 analysis, microbiome data features are first linked to pre-processed reference databases to construct a community metabolic model describing the predicted metabolic reaction capabilities of each community member taxon (step 1). Next, this network model is combined with microbiome feature abundances to calculate community metabolic potential (CMP) scores for each metabolite, taxon, and sample, representing the approximate relative capacity to synthesize or utilize that metabolite (step 2). Total community metabolic potential scores for each metabolite are linked to metabolomics measurements using a regression model (step 3). For metabolites with a significant relationship between concentration and potential, the specific taxonomic contributors to each metabolite are then analyzed (step 4).

Data input options

The basic input requirement for MIMOSA2 is a pair of datasets from the same set of samples: one of microbiome measurements and the other of metabolites. Each of these datasets

may take a variety of forms, depending on the study's design, experimental assays, and processing techniques.

Users can provide microbiome data generated from 16S rRNA or shotgun metagenomic sequencing studies. 16S rRNA data can be provided as a feature table of amplicon sequence variants (ASVs), or of closed-reference operational taxonomic units (OTUs) using either the GreenGenes or SILVA databases (24,25). Data from a shotgun metagenomic study can be provided in the form of a table of KEGG Ortholog (10) functional abundances. Metagenomic data can be either unstratified (total KEGG Ortholog abundances in each sample) or stratified by microbial taxa (in the formats produced by either HUMAnN or PICRUSt2 software (26,27)).

Metabolite data can be produced from any metabolomics platform, but it must contain putative metabolite identifications in the form of either metabolite names or KEGG compound IDs. If metabolite names are provided, they are mapped to KEGG IDs for the main analysis using the metabolite ID cross-referencing utility from the *MetabolAnalystR* package (28).

Reference data and metabolic model construction

MIMOSA2 uses reference data to estimate how community metabolic potential varies across a set of microbiome samples. It takes advantage of genome-scale metabolic model data from multiple alternative sources to use the most appropriate reference data for a given dataset. Specifically, MIMOSA2 can currently generate a metabolic model based on one of three sources. First, it can use a curated set of reactions from the KEGG database. If OTU data is provided, KEGG reactions are inferred for each OTU using PICRUSt 1 (29) pre-computed outputs. Alternatively, MIMOSA2 can link 16S rRNA amplicon data to one of two large collections of genome-scale metabolic reconstructions: the AGORA collection of genome-scale metabolic

models of gut microbial species (v1.0.2) (11), or the *embl_gems* library of genome-scale metabolic models for all 5,587 reference and representative bacterial genomes in RefSeq (12).

The method used to map microbiome taxa abundances to metabolic reactions depends on the input data type (described fully in Methods and illustrated in Figure 6). 16S rRNA sequence variants are mapped using *vsearch* to either RNA genes for the AGORA genome collection (see Methods), or to Greengenes 99% OTU representative sequences. Greengenes OTUs are mapped to KEGG using the precomputed genome inferences from PICRUSt 1.1.3 (29), and are linked to AGORA models using a pre-calculated alignment between the two databases. KEGG Ortholog abundances provided from a shotgun metagenomic dataset or another method (such as from PICRUSt2 (30)) can be directly linked to KEGG reactions.

Additionally, users can specify custom additions, subtractions, or modifications to any of the MIMOSA2 model templates. These can be provided at the gene, reaction, or taxon level. Example modification files are provided in the MIMOSA2 documentation. Adding or removing a particular reaction can be useful to assess the impact of suspected incomplete or incorrect annotations in the reference database, as we have observed in analyses using MIMOSA version 1 (15). Full details on the implementation of each of the metabolic model construction options are provided in the Methods section.

Core algorithm and identification of species-metabolite contributors

Using the constructed metabolic models, MIMOSA2 next calculates community metabolic potential (CMP) scores for each taxon, sample, and metabolite. These scores represent the predicted capacity of each taxon in each sample to synthesize or utilize the metabolite in question, facilitating downstream analysis of the links between individual taxa and metabolites. These metabolic potential scores are then aggregated at the community level and compared with

the relevant metabolite measurements. An example analysis for a toy metabolite and dataset is shown in Figure 2.

Figure 2. An example contribution analysis of hypothetical metabolite *M*. A) In this toy example, metabolite *M* is produced by two taxa and utilized by a third. B) Overview of CMP scores and metabolite measurements for metabolite *M* across a dataset of six samples. Taxon-level CMP scores are based on the estimated ability of each species to synthesize/utilize the metabolite and on the abundance of the taxon in the sample. C) Comparison of CMP scores and measurements of *M* across this dataset. The solution found by OLS regression is affected more strongly by the outlier samples D and F than that found by rank-based regression. D) Final summary contribution plot for metabolite *M* using the rank-based regression option. The bars represent the contributions to metabolite variance explained by each taxon.

Most of the variance is attributed to differences in the amount of utilization of *M* by Taxon 3 across samples, reflecting the larger variability in Taxon 3's CMP scores shown in panel B.

Specifically, metabolic potential scores are calculated as a linear combination of the abundances of genes predicted to contribute to synthesis or utilization of a metabolite multiplied by their expected effects given reaction stoichiometry (Figure 2A-B). This method is similar to the approach used by MIMOSA version 1 (13), but importantly, these scores are now calculated at the level of individual community members, rather than for the community as a whole. MIMOSA2 next fits a linear regression model relating the total CMP scores to the measured metabolite concentrations across samples (Figure 2C). By default, the regression model is fit using rank-based estimation (31) (referred to here as MIMOSA2-rank), although an option for ordinary least-squares (OLS) regression is also provided (MIMOSA2-OLS). The error function in rank-based estimation incorporates both the rank and the magnitude of the model residuals, resulting in a more robust fit to noisy data (31).

A metabolite is identified as putatively microbiome-governed if the overall regression model fit meets a significance threshold (*F* test or drop-in-deviance test $p < 0.1$ by default). MIMOSA2 then decomposes the share of metabolite variation explained by the model into linear contributions from each taxon (Figure 2D), and microbial taxa with large contributions to model variation are identified as potential contributors. The calculation of this decomposition is analogous to our previous approach for calculating taxonomic contributors based on simulated metabolic fluxes (5). This metric prioritizes taxa whose estimated metabolic potential is relatively abundant, variable, and associated with the measured metabolite concentrations (Figure 2 B and D). For analyses using OLS regression, contributions are calculated as the covariance of the metabolic potential scores from each taxon with metabolite concentrations. For

analyses using rank-based regression, this metric is calculated using a permutation-based analysis of the importance of each taxon's scores to the model fit. A full explanation of these contribution calculations can be found in the Methods section. Notably, this approach is different from the taxonomic contributor metric used in MIMOSA version 1, which prioritized taxa based on the correlation coefficient of their metabolic potential with metabolites across samples. It also differs from standard approaches for assessing covariate importance in a regression model in that it accounts for both the association and relative magnitude of contributions from each taxon, without any scaling.

Importantly, each of the steps in the MIMOSA2 workflow is modular. For instance, community metabolic potential is currently calculated from the community metabolic model using the gene abundance-based scoring approach described above, but these values could in the future be replaced with estimates of metabolic fluxes using Flux Balance Analysis or other simulation methods (32).

Interface, results, and visualization

The MIMOSA2 workflow can be run via either a web application or an R package. In both options, the input file names and analysis parameters are encoded in a configuration file, which can be used to reproduce the analysis.

MIMOSA2 produces several detailed results for each analyzed metabolite: the model fit between community metabolic potential and metabolite measurements, the primary taxa contributors to metabolite variation, and the specific reactions that formed the basis of the metabolic potential calculation. When an analysis is run using the web application, these results are summarized in an interactive table with a row for each metabolite (Figure 3). Processed result tables are also made available for download, along with the analysis configuration file.

MIMOSA2

MIMOSA2 is a tool for metabolic model-based evaluation of paired microbiome and metabolomics datasets. MIMOSA2 1) constructs community metabolic models, 2) assesses whether metabolite concentrations are consistent with estimated community metabolic potential, and 3) identifies specific taxa and reactions that can explain metabolite variation. For more information, see the [documentation](#).

[Download All Results Tables and Plots](#) [Analysis Summary Statistics](#) [Model Summaries](#) [Contribution Results](#) [Mapped Taxa Abundances](#) [Community Metabolic Network Models](#)
[Community Metabolic Potential Scores](#) [Record of Configuration Settings](#)

After this session terminates, a zipped directory of all results files can be downloaded from the following link for 30 days: http://elbo-spice.gs.washington.edu/shiny/MIMOSA2shiny/analysisResults/example_allResults.zip

Each row of the table below summarizes the MIMOSA2 results for a given metabolite, including the overall concordance between metabolic potential and concentration, and the contributing taxa. Mouse over the names of each column of the table for a more detailed description. Specific plots can be downloaded using the buttons below the table.

Show entries

Search:

Compound ID	Name	R-squared	P-value	Slope	Comparison Plot	Contribution Plot	Top Producing Taxa and Genes/Rxns	Top Utilizing Taxa and Genes/Rxns	Intercept
All					All	All	All	All	/
1	C00123	L-Leucine	0.144	0.00001	0.008			Butyrivibrio_crossotus_DSM_2876/LEUabc Mitsuokella_jalaludinii_DSM_13811/LEUabc Mobiluncus_curtisii_ATCC_43063/LEUabc Fusobacterium_nucleatum_subsp_animalis_D11/LEUabc Peptoniphilus_lacrimalis_DSM_7455/LEUabc	16.682
2	C00407	L-Isoleucine	0.141	0.00006	0.005			Prevotella_timonensis_CRIS_5C_B1/ILEabc Mitsuokella_jalaludinii_DSM_13811/ILEabc Butyrivibrio_crossotus_DSM_2876/ILEabc Mobiluncus_curtisii_ATCC_43063/ILEabc Fusobacterium_nucleatum_subsp_animalis_D11/ILEabc	16.207

Figure 3. **Interactive results interface for the MIMOSA2 web application.** In addition to making all processed results available for download, the application displays an interactive table in which each row summarizes the results for a single metabolite. The best-predicted metabolites are shown first, along with associated information including model statistics, plots of the data and top contributors, and lists of the top contributing taxa and reactions predicted via both synthesis and utilization.

Results

Application of MIMOSA2 to simulated datasets with known microbe-metabolite interactions

We validated the MIMOSA2 method by applying it first to two simulated microbiome-metabolome datasets (previously described in (5)). We used these simulation datasets to evaluate the ability of MIMOSA2 to achieve two main objectives: 1) to classify whether metabolite levels are likely determined by microbiome metabolism, and 2) to identify *key contributors* for those metabolites and the relevant genes and reactions. In our previous study, we defined “key contributors” as the set of microbial taxa (or exogenous environmental influences) that are responsible for a substantial share of the variation in the levels of a metabolite across samples,

based on their quantitative metabolite fluxes. Notably, these two objectives differ from the objectives of other recently published microbiome-metabolome analysis methods (7,8), which are designed primarily to predict metabolite levels from microbiome data with high accuracy without incorporating additional information or generating mechanistic hypotheses. In this study, we compared the performance of MIMOSA2 using both rank-based and OLS regression estimation with three alternative approaches: a species-metabolite pairwise Spearman correlation analysis, a modified correlation analysis in which significant taxa-metabolite correlations are only retained if the taxon is known to possess reactions linked to the metabolite (see Methods for additional details), as well as the original implementation of MIMOSA.

The two simulation datasets describe divergent sets of hypothetical microbiome samples (Figure 4A). They were generated using a multi-species dynamic Flux Balance Analysis framework to simulate gut bacterial metabolism in a fixed nutrient environment, using metabolic network reconstructions from the AGORA collection (5,11). In this framework, simulated microbial uptake and secretion dynamically impact concentrations of metabolites in the shared environment. Both datasets consisted of species and metabolite concentrations at the final time point of several 144-hour dynamic multi-species simulations. Both datasets included small random variations in the simulated environmental metabolite concentrations available to the simulated species across samples (see Methods), but microbial fluxes are the primary drivers of final concentrations for a large share of compounds. Using this method, Dataset 1 consists of communities with varying compositions of 10 representative gut species and 3% variation in environmental metabolite concentrations. Dataset 2 is more complex, consisting of 57 samples whose initial compositions were designed to emulate Human Microbiome Project (HMP) gut samples (33), along with 1% variation in nutrient inflow. These species compositions were

determined by aligning HMP 16S rRNA sequencing variants against genomes linked to AGORA reconstructions, resulting in a total of 131 species unevenly distributed across the dataset. To construct community metabolic network models for MIMOSA2, we used the unconstrained community metabolic reconstructions which formed the basis for each set of simulations.

Figure 4. Identification of key microbe-metabolite contributions by MIMOSA2 from simulated datasets. A) Descriptive summary statistics for the two simulated datasets analyzed, including number of samples, species, and metabolites included. B) Precision, specificity, and sensitivity of MIMOSA2 analysis for recovering true key microbial contributors to metabolite variation from metabolites identified as potentially microbiome-governed by MIMOSA2 in two simulated datasets, compared with correlation-based approaches and with MIMOSA version 1. Standard thresholds were used to designate contributors for each method (Correlation: 0.01 q-value, MIMOSA2: 3% contribution). C) Precision-recall curves for identifying key contributors for the same set of metabolites in simulated Datasets 1 and 2 as in panel B, by MIMOSA2 and the same alternative correlation-based methods. Colors for each method are the same as labeled in panel B. D) Precision, specificity, and sensitivity of MIMOSA2 analysis for recovering true

key microbial contributors to metabolite variation (as in panel B), but for all metabolites. All thresholds are the same as above. E) Precision-recall curves for identifying key taxon-metabolite contributors in simulated Datasets 1 and 2, as in panel C, but for all metabolites.

We first evaluated the ability of MIMOSA2 to identify true microbiome-governed metabolites in these datasets. We found that MIMOSA2 identifies these metabolites with somewhat low sensitivity but very high precision, especially in the higher-complexity dataset. Specifically, in Dataset 1, 50 of 85 simulated metabolites were true microbiome-governed metabolites (i.e., at least 10% of variation in their concentrations is determined by the microbiome as opposed to external fluctuations). Using a significance cutoff of $p < 0.1$, MIMOSA2-OLS recovered 19 of these metabolites with only 3 false positives (86% precision), and MIMOSA2-rank recovered 17 with only 2 false positives (89% precision). A large share of the metabolites missed by either MIMOSA2 method ($n=19$) were not able to be analyzed at all by MIMOSA2 due to a lack of linked non-reversible reactions. MIMOSA version 1 had a higher false positive rate, as it identified 32 metabolites as consistent with metabolic potential of which 21 were true microbiome-governed metabolites (65.6% precision). Using significant taxon-metabolite correlations (q -value < 0.01) as a basis for inferring microbiome-governed metabolites resulted in similar performance, correctly identifying 30 of the 50 metabolites as microbiome-governed, with 2 false positives. In Dataset 2, nearly all simulated metabolites were classified as true microbiome-governed compounds: 193 of 221. MIMOSA2 again identified these with high precision: 73 metabolites were detected by MIMOSA2-rank, with 0 false positives (MIMOSA2-OLS, 68 with 0 false positives), out of 156 analyzed metabolites, while MIMOSA 1 identified 93 microbiome-governed compounds and 6 false positives (93.9% precision). The larger share of metabolites identified as microbiome-governed by MIMOSA2-

rank suggests that it may be able to detect microbial contributions with higher sensitivity than MIMOSA2-OLS in complex datasets. Overall, MIMOSA2 displayed high precision and lower sensitivity than correlation analysis in identifying microbiome-governed metabolites. This result is to be expected to some extent, since MIMOSA2's sensitivity is constrained by the quality of reference reaction information used to calculate metabolic potential.

A central feature of MIMOSA2 is its ability to link metabolites with microbial taxa and mechanisms that appear to be responsible for the observed variation. Notably, MIMOSA2 only aims to identify contributors for metabolites that are predicted well by its model, and this prioritization of metabolites that are meaningfully associated with microbial profiles is an important advantage of the framework. Therefore, we focused our evaluation on each method's performance in identifying key microbial contributors to metabolites for only the set of metabolites that were detected as microbiome-governed by either MIMOSA2-rank and/or MIMOSA2-OLS (22 from Dataset 1, 78 from Dataset 2), although results of the same evaluation across all metabolites are shown in Figure 4D-E. Key microbial contributors were identified with comparable sensitivity and higher specificity and precision overall by MIMOSA2 than any alternative approach at standard significance cutoffs (Figure 4B,D). Both MIMOSA2 methods outperform correlation analysis and MIMOSA version 1 across a range of significance thresholds, particularly in the noisier and more diverse Dataset 2 (Figure 4C,E). These results, including MIMOSA2's lower sensitivity overall, are expected in light of its approximate metabolic model, which acts as a filter for spurious associations, but cannot detect true mechanisms that are not described well by the model.

MIMOSA2 results compare favorably with experimentally inferred microbe-metabolite links in insect microbiomes

We next applied MIMOSA2 to re-analyze a publicly available dataset of taxonomic and metabolomics measurements from the gut microbiota of the honeybee *Apis mellifera* (34). This dataset consists of samples from medium-complexity natural communities as well as monocolonization experiments, which allowed us to compare MIMOSA2's inferences of taxon-metabolite links with independent experimental data. We applied MIMOSA2 to a set of 18 samples from gnotobiotic bees, each colonized with a synthetic community of resident gut microbial strains. We then compared MIMOSA2's inference with metabolomics data from microbiota-depleted (i.e. nearly germ-free) and monocolonized samples (Figure 5A). Specifically, we used the metabolomics data from microbiota-depleted and monocolonized samples to define *experimentally inferred* microbiome-governed metabolites and key contributing strains. We defined experimentally inferred microbiome-governed metabolites (EIMM) as those that differed in colonized bee samples compared with microbiota-depleted control bee samples. Similarly, experimentally inferred key contributor strains (EIKC) were defined by comparing metabolite levels in bees monocolonized with each individual gut strain with levels from community-colonized bees, following the approach used in the original study (see Methods). We compared these experimentally inferred metabolites and contributions with the results of MIMOSA2-rank applied to just the set of 18 community samples (Figure 5A, full details in Methods).

Figure 5. **Partial concordance between MIMOSA2 results and experimental inferences in honeybee**

gut microbiota. A) Summary of experiments from (34) reanalyzed here. MIMOSA2 was applied to a

dataset of 18 samples from community-colonized bees, and its results were compared with inferences

from metabolomics of microbiota-depleted (germ-free) bees and bees monocolonized with individual

bacterial strains. B) Experimentally-inferred microbial metabolites in 11-strain communities are

significantly better predicted by MIMOSA2 than other metabolites. C) MIMOSA2 identifies

experimentally-inferred microbial contributors with higher precision and recall than microbe-metabolite

correlation analysis. D) Metabolite-level comparison of experimentally-inferred and MIMOSA2-inferred

microbial key contributors. Cell color indicates a microbe's contribution to variance in a metabolite as

inferred by MIMOSA2; black dots indicate experimentally-inferred contributors based on metabolomics

of monocolonized samples. Metabolites shown are microbiome-governed as determined by the

MIMOSA2 model. Ba, *Bartonella apis*; Bi, *Bifidobacterium asteroides*; F4, *Lactobacillus* Firm-4; F5,

Lactobacillus Firm-5; Fp, *Frischella perrara*; Ga, *Gilliamella apicola*; Sa, *Snodgrassella alvi*.

MIMOSA2 results were largely consistent with experimental findings. The abundances

of EIMM were predicted with higher accuracy by MIMOSA2-rank than other metabolites

(Figure 5B, $p=0.04$, Wilcoxon rank-sum test), suggesting that MIMOSA2 correctly identified

metabolites affected by microbial composition. Strikingly, MIMOSA2-inferred key contributors

were 6.02 times more likely than expected by chance to also be EIKC (Fisher exact test, p

$=0.0099$). MIMOSA2-identified contributors were also more predictive of EIKC than

correlation-based contributors: MIMOSA2 predicted EIKC with 23% precision and an area

under the ROC curve (AUC) of 0.76 among microbiome-governed metabolites, compared with

20.6% precision and AUC of 0.68 for correlation analysis in the same set of metabolites. Across

all EIMM (i.e. including those not identified as microbiome-governed by MIMOSA2), microbe-

metabolite correlations were strikingly not predictive of EIKC, with a precision of 18.8% and

AUC of 0.52 for correlation magnitude across all EIMM (Figure 5C), suggesting that metabolite shifts in this dataset are not well described by univariate associations. Filtering correlations based on genomic potential of the associated strain only slightly improved these values (precision 21.7%, AUC 0.53).

The contributors identified by both methods are shown for microbiome-governed metabolites in Figure 5D. These span a variety of metabolic categories including vitamins, amino acid metabolites, fatty acids, and nucleoside metabolites, which were noted as a particularly enriched category in the original study (34). The metabolite best predicted by MIMOSA2, anthranilate, is an aromatic compound in the tryptophan biosynthesis pathway, and variation in its levels is attributed to the microbes *Lactobacillus* Firm-5 and *Snodgrassella alvi*, coinciding exactly with the contributors identified in the monocolonization data. Additionally, discrepancies observed between experimental inferences and MIMOSA2 findings for other metabolites are not necessarily due to inaccurate inference by MIMOSA2, but could alternatively indicate context-dependent metabolism, in which the metabolic effects of a microbe differ between a community setting and monocolonized samples. For example, the strong contribution to *sn*-glycerol-3-phosphate by *Lactobacillus* Firm-5 inferred by MIMOSA2 could indicate a true effect, but one that occurs at lower levels in isolation than in a complex community.

Conclusions

Here, we have described MIMOSA2, a comprehensive software framework for generating and evaluating mechanistic metabolic hypotheses from microbiome-metabolome datasets. MIMOSA2 evaluates whether metabolite levels across a set of microbial communities are consistent with the communities' estimated metabolic capacities. The software has multiple interfaces, is compatible with several microbiome data formats, and can provide evidence of

metabolic mechanisms from varied microbiomes measured in their natural context. The MIMOSA2 analysis framework can assess possible microbial causes of differing metabolite phenotypes between health and disease (15), and can evaluate the metabolic consequences of microbiome shifts across lifestyle or environmental factors (14).

MIMOSA2 supports a variety of analysis options with distinct benefits and caveats. For example, 16S rRNA amplicon sequencing datasets are easy to generate and link to microbial reference databases, but sequence divergence in this gene is imperfectly linked to metabolic and phenotypic divergence (35). Nevertheless, a large share of functional information can be accurately predicted from 16S rRNA amplicon datasets (8,29), supporting the value of this analysis. Similarly, the AGORA database (11) contains high-quality metabolic network reconstructions, but only for 818 microbial species commonly found in the human gut, so other database options are likely preferable for the analysis of datasets from other environments.

As microbiome-metabolome studies continue to grow, a diverse set of analysis tools for these datasets is also emerging, with different focuses and options depending on the research goals and study design. These tools include machine learning methods to predict the metabolome from the microbiome (7,8), methods to construct detailed metabolic models of particular communities (36), and methods to annotate and categorize metabolites based on the presence or absence of relevant microbial genes (23). MIMOSA2 fills a unique niche, performing a generalizable and user-friendly analysis that generates hypotheses about metabolic mechanisms by quantitatively synthesizing microbiome and metabolome data with reference knowledge.

Methods

MIMOSA2 Database Construction

MIMOSA2 can rapidly link multiple types of microbiome datasets to multiple databases of metabolic network models. All currently implemented linking options and their corresponding analysis steps are shown in Figure 6. Briefly, 16S rRNA sequence variants are mapped to one of 3 reference databases, which contain pre-processed links to metabolic reactions for each sequence variant (pre-processing described below). Metagenomic datasets are linked to reactions in the KEGG database based on KEGG Orthology gene family annotations.

Figure 6. **Illustration of the various network model reconstruction options performed by MIMOSA2.** As shown, microbiome data can be provided in various input formats, and linked to various databases of metabolic reactions to construct community metabolic models for each sample.

16S rRNA pre-processing and mapping

To facilitate mapping of 16S rRNA amplicon datasets to the AGORA database of metabolic reconstructions, the biomartR R package (37) was used to download RNA genes for AGORA species from NCBI, based on NCBI Taxonomy IDs provided in Table S5 of (11). To establish a mapping between Greengenes OTUs and the AGORA genome collection, the representative set of sequences for Greengenes 13_8 99% OTUs were aligned optimally to the downloaded AGORA RNA database using *vsearch* 2.8.1, returning the single best alignment for each OTU. An alignment with the same parameters was performed for SILVA v132 99% OTUs

to find their closest match in both the Greengenes database (for construction of KEGG-based models), and the AGORA RNA database (for construction of AGORA-based models).

To facilitate mapping to the *embl_gems* database of metabolic reconstructions, RNA genes for all RefSeq genomes were similarly downloaded using biomartR and aligned with *vsearch*.

The sets of RNA genes linked to reference models are stored as *vsearch* .udb databases for fast alignment on the MIMOSA2 server, and are also available for download from the MIMOSA2 website for local analyses.

Metabolic network construction

For KEGG-based metabolic models, MIMOSA2 uses a generic KEGG model template constructed using 3 files from the KEGG database (February 2018 KEGG release), following (13): *reaction_mapformula.lst* (streamlined set of core metabolic KOs, pathways, reactions, and metabolites), *reaction_ko.list* (KO-reaction links), and *reaction* (reaction annotations).

To generate a set of reactions for each Greengenes 99% OTU, the generic KEGG model described above was merged with the list of KOs inferred to be present in each OTU according to PICRUSt (29). The copy number for each reaction is also normalized according to the 16S rRNA copy number inferred by PICRUSt for each OTU.

To use the AGORA genome-scale metabolic reconstructions, we first downloaded the AGORA model collection version 1.0.2 from the Virtual Metabolic Human database (38), including the set of 818 unconstrained models as well as the set constrained based on an average European diet. We converted the models to a format usable by MIMOSA2 using the R.matlab package. A pre-processed reaction file was generated from the stoichiometry matrix and bound constraints for each model. A normalized copy number was included for each reaction, which

was assumed to be 1 divided by the number of 16S rRNA genes annotated in the relevant reference genome. Metabolite IDs were mapped to KEGG compounds using the mappings provided with each model.

Fitting of the CMP-Metabolite model and calculation of taxa contribution values

When provided with a taxon-specific microbiome dataset (i.e. any input option except for unstratified functional metagenomic abundances), MIMOSA2 calculates the contribution of each taxon to the overall model fit for each metabolite. The initial analysis finds a solution to the linear model relating total CMP scores ($\mathbf{X}$) to the concentrations of a single metabolite ($\mathbf{y}$) in the form:

$$\mathbf{y} = \beta_1 \mathbf{X} + \beta_0 + \epsilon,$$

where ϵ is the residual error for each sample. This structure is the same regardless of whether OLS or rank-based fitting is used. For OLS models, the total model error S over all n samples is equal to:

$$S_{OLS} = \sum_{i=1}^n (y_i - \beta_1 X_i - \beta_0)^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

For rank-based models, the model error is instead calculated as Jaeckel's dispersion (39):

$$S_{Rank} = \frac{1}{n+1} \sum_{i=1}^n \varphi(R(y_i - \beta_1 X_i))(y_i - \beta_1 X_i)$$

where φ is the Wilcoxon score function and R is the rank function. (The value of the model intercept β_0 has no effect on the calculation of rank-based dispersion.) Under this model, the model error is a function of the product of the relative ranking of a residual and its value.

In both models, the obtained solution is chosen to minimize the total model error. The share of variance or dispersion in metabolite concentrations explained by the obtained model (i.e., the unadjusted R^2 value) can be calculated as:

$$R^2 = 1 - \frac{S}{nV} = \frac{nV - S}{nV}$$

where S is the sum of squared error and V is the metabolite variance for OLS, and S is the model dispersion and V is the null dispersion (i.e. dispersion from the sample mean) for rank-based regression.

Importantly, every total CMP score (represented as X_i in the equations above) is the sum of the CMP scores from all m taxa in a given sample:

$$X_i = \sum_{j=1}^m x_j$$

where x_j is the CMP score for taxon j . Therefore, the R^2 expression above can be separated into the share of explained variation attributable to each taxon's scaled CMP scores. This is shown step-by-step below for OLS, using the fact that the variance of a sum of correlated random variables can be defined as a function of their covariances, and also that any observation y_i is equal to $y_i = \hat{y}_i + \epsilon = \beta_1 X_i + \beta_0 + \epsilon$.

$$R^2 = \frac{nV - S}{nV} = \frac{1}{nV} \left(nV - \sum_{i=1}^n (y_i - \hat{y}_i)^2 \right) = \frac{1}{nV} \left(n \sum_{i=1}^m \sum_{j=1}^n cov(x_i, x_j) - \sum_{i=1}^n (y_i - \hat{y}_i)^2 \right) \dots$$

For rank-based regression, because the model dispersion depends on the rank of the residuals, the explained variance cannot be analytically separated into linear contributions from each taxon. Therefore, for this model we calculate linear contributions using a permutation-based method based on the game theory concept of Shapley values (40). In this method, the scores

from each taxon are removed from the total score in random orderings and the effect of each taxon is calculated as the average marginal contribution across all orderings.

Analysis of simulation datasets

Simulation datasets were generated using a dynamic Flux Balance Analysis (FBA) co-culture algorithm described elsewhere (5,41,42). Briefly, at each time point of the simulation, the growth rate of each microbe in the population is optimized using FBA, and biomass and environmental nutrient constraints are iteratively updated based on the resulting solutions. To generate these simulated datasets, we initialized simulated co-cultures with either systematically varying microbial abundances (Dataset 1) or abundances based on human microbiome datasets (Dataset 2), ran the corresponding simulations for 144 hours, and “sampled” the concentrations of microbes and metabolites at the final time point.

To apply MIMOSA2 to these datasets, the original AGORA reconstructions were used as the basis for the community metabolic network models. Similarly, these reconstructions were also used to filter significant correlations based on whether each species was known to be capable of modifying each metabolite. All reactions were assumed to proceed in a forward direction. True microbiome-governed metabolites were defined as those for which the scaled contribution to variation from nutrient inflow fluxes was less than 90% (i.e., the microbial contribution was at least 10%). For the comparisons with MIMOSA version 1, a q -value cutoff of 0.1 was used to identify microbiome-governed or consistent metabolites, and key contributors for microbiome-governed metabolites were identified on the basis of the Pearson correlation between each species' CMP scores and metabolite abundances. True key microbial contributors were defined as any taxon with a scaled positive variance contribution greater than 10%. Putative

microbiome-governed metabolites inferred by MIMOSA2 were defined as those for which the community metabolic potential model explained greater than 10% of metabolite variation. Putative taxonomic contributors were those whose contributions to the model were greater than 10%. Precision, recall, and ROC curve statistics were calculated using the pROC package in R (43).

Analysis of honeybee validation dataset

We obtained the qPCR and untargeted metabolomics datasets from Supplementary Data S1 and S2 of the publication describing these experiments (34). To run MIMOSA2 on this dataset, we constructed a KEGG-based metabolic network model of each microbial strain using annotations from the IMG database (44). We downloaded KO annotations of the genomes of each of the colonized strains from IMG, and estimated taxon-stratified KO abundances in each sample by multiplying the respective KO profiles by the qPCR-measured taxon abundances for the subset of samples from colonized bees (CL group), a total of 18 samples. The resulting table was used as input for MIMOSA2 along with the KEGG-identified metabolite data from the same samples. Metabolite abundances were not log-transformed, as most metabolites displayed approximately normally distributed abundances.

To compare MIMOSA2 results with the findings from monocolonized bees, we used the differential abundance analysis available from Supplementary Data S7 of the same publication. The authors evaluated whether metabolite features display similar shifts in monocolonization samples compared to community samples using an ANOVA followed by Tukey post-hoc test at 99% confidence, $p < 0.05$. We classified metabolites as microbiome-governed using a less conservative significance threshold than the original analysis ($p < 0.2$), in order to not exclude effects potentially detectable by MIMOSA2.

527 **Availability and requirements**

528 **Project name:** MIMOSA2

529 **Project homepage:** www.borensteinlab.com/software_MIMOSA2.html

530 **Operating system:** Platform independent

531 **Programming language:** R

532 **Other requirements:** The web application has been tested with Safari (versions 11.0 and
533 higher), Firefox (versions 53.0 and higher), Chrome (versions 61.0 and higher), and Opera
534 (versions 47.0 and higher) browsers. The R package requires a working installation of R 3.6.0 or
535 higher.

536 **License:** GNU General Public License v3.0

537 **Any restrictions to use by non-academics:** None.

538 **List of abbreviations**

539 KEGG, Kyoto Encyclopedia of Genes and Genomes

540 MIMOSA, Model-based Integration of Metabolite Observations and Species Abundances

541 CMP, community metabolic potential

542 rRNA, ribosomal ribonucleic acid

543 HUMANn, Human Microbiome Project Unified Microbiome Analysis Network

544 PICRUSt,

545 Phylogenetic Investigation of Communities by Reconstruction of Unobserved States

546 ASV, amplicon sequence variant

547 OTU, operational taxonomic unit

548 OLS, ordinary least squares regression

549 HMP, Human Microbiome Project

550 EIMM, experimentally inferred microbiome-governed metabolites

551 EIKC, experimentally inferred key contributors

552 **Declarations**

553 **Ethics approval and consent to participate**

554 Not applicable

555 **Consent for publication**

556 Not applicable

557 **Availability of data and materials**

558 The simulated microbiome-metabolome datasets analyzed in this study are available from

559 www.borensteinlab.com/download.html. The honeybee microbiota data is available from the

560 Supplementary Data S1 and S2 of (34). The processed versions of the AGORA and *embl_gems*

561 databases generated for use in the MIMOSA2 software are available at

562 www.github.com/borenstein-lab/MIMOSA2shiny. The processed version of KEGG data can be

563 provided upon request with confirmation of a KEGG subscription. The code for processing raw

564 KEGG files is available at www.github.com/borenstein-lab/MIMOSA2shiny.

565 **Competing interests**

566 The authors declare they have no competing interests.

567 **Funding**

568 This project was supported in part by National Institutes of Health Grant

569 1R01GM124312 to EB, National Institutes of Health Grant R01DK095869, and Israel Science

570 Foundation Grant 2435/19 to EB. EB is a Faculty Fellow of the Edmond J. Safra Center for

571 Bioinformatics at Tel Aviv University.

572 **Authors' contributions**

CN and EB conceived and designed the study. CN implemented the software and performed the analysis. AE assisted with development, testing, and maintenance of the web application. CN and EB wrote the paper. All authors read and approved the final manuscript.

Acknowledgements

We thank members of the Borenstein lab for suggestions and for software testing and support.

References

1. McGuire KL, Treseder KK. Microbial communities and their relevance for ecosystem models: Decomposition as a case study. *Soil Biology and Biochemistry*. 2010 Apr;42(4):529–35.
2. Kasubuchi M, Hasegawa S, Hiramatsu T, Ichimura A, Kimura I. Dietary Gut Microbial Metabolites, Short-chain Fatty Acids, and Host Metabolic Regulation. *Nutrients*. 2015 Apr 14;7(4):2839–49.
3. Hazen TC, Dubinsky EA, DeSantis TZ, Andersen GL, Piceno YM, Singh N, et al. Deep-Sea Oil Plume Enriches Indigenous Oil-Degrading Bacteria. *Science*. 2010 Oct 8;330(6001):204–8.
4. Shaffer M, Armstrong AJS, Phelan VV, Reisdorph N, Lozupone CA. Microbiome and metabolome data integration provides insight into health and disease. *Translational Research* [Internet]. 2017 Jul [cited 2017 Oct 9]; Available from: <http://linkinghub.elsevier.com/retrieve/pii/S1931524417302323>
5. Noecker C, Chiu H-C, McNally CP, Borenstein E. Defining and Evaluating Microbial Contributions to Metabolite Variation in Microbiome-Metabolome Association Studies. Chia N, editor. *mSystems*. 2019 Dec 17;4(6):e00579-19, /msystems/4/6/msys.00579-19.atom.
6. Hoyles L, Jiménez-Pranteda ML, Chilloux J, Brial F, Myridakis A, Aranas T, et al. Metabolic retroconversion of trimethylamine N-oxide and the gut microbiota. *Microbiome* [Internet]. 2018 Dec [cited 2018 Apr 29];6(1). Available from: <https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-018-0461-0>

- 603 7. Morton JT, Aksenov AA, Nothias LF, Foulds JR, Quinn RA, Badri MH, et al. Learning
604 representations of microbe–metabolite interactions. *Nat Methods*. 2019 Dec;16(12):1306–
605 14.
- 606 8. Mallick H, Franzosa EA, McIver LJ, Banerjee S, Sirota-Madi A, Kostic AD, et al.
607 Predictive metabolomic profiling of microbial communities using amplicon or
608 metagenomic sequences. *Nat Commun*. 2019 Dec;10(1):3136.
- 609 9. Reiman D, Layden BT, Dai Y. MiMeNet: Exploring microbiome-metabolome relationships
610 using neural networks. Mendes P, editor. *PLoS Comput Biol*. 2021 May
611 17;17(5):e1009021.
- 612 10. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids*
613 *Research*. 2000 Jan 1;28(1):27–30.
- 614 11. Magnúsdóttir S, Heinken A, Kutt L, Ravcheev DA, Bauer E, Noronha A, et al. Generation
615 of genome-scale metabolic reconstructions for 773 members of the human gut microbiota.
616 *Nature Biotechnology* [Internet]. 2016 Nov 28 [cited 2016 Dec 2]; Available from:
617 <http://www.nature.com/doifinder/10.1038/nbt.3703>
- 618 12. Machado D, Andrejev S, Tramontano M, Patil KR. Fast automated reconstruction of
619 genome-scale metabolic models for microbial species and communities. *Nucleic Acids*
620 *Research*. 2018 Sep 6;46(15):7542–53.
- 621 13. Noecker C, Eng A, Srinivasan S, Theriot CM, Young VB, Jansson JK, et al. Metabolic
622 Model-Based Integration of Microbiome Taxonomic and Metabolomic Profiles Elucidates
623 Mechanistic Links between Ecological and Metabolic Variation. *mSystems*. 2016 Feb
624 23;1(1):e00013-15.
- 625 14. Snijders AM, Langley SA, Kim Y-M, Brislawn CJ, Noecker C, Zink EM, et al. Influence of
626 early life exposure, host genetics and diet on the mouse gut microbiome and metabolome.
627 *Nature Microbiology*. 2016 Nov 28;2:16221.
- 628 15. Sharon G, Cruz NJ, Kang D-W, Gandal MJ, Wang B, Kim Y-M, et al. Human Gut
629 Microbiota from Autism Spectrum Disorder Promote Behavioral Symptoms in Mice. *Cell*.
630 2019 May;177(6):1600-1618.e17.
- 631 16. Stewart CJ, Mansbach JM, Wong MC, Ajami NJ, Petrosino JF, Camargo CA, et al.
632 Associations of Nasopharyngeal Metabolome and Microbiome with Severity among Infants
633 with Bronchiolitis. A Multiomic Analysis. *American Journal of Respiratory and Critical*
634 *Care Medicine*. 2017 Oct;196(7):882–91.
- 635 17. Ilhan ZE, Łaniewski P, Thomas N, Roe DJ, Chase DM, Herbst-Kralovetz MM. Deciphering
636 the complex interplay between microbiota, HPV, inflammation and cancer through
637 cervicovaginal metabolic profiling. *EBioMedicine*. 2019 Apr;S2352396419302671.
- 638 18. Casero D, Gill K, Sridharan V, Koturbash I, Nelson G, Hauer-Jensen M, et al. Space-type
639 radiation induces multimodal responses in the mouse gut microbiome and metabolome.

- 640 Microbiome [Internet]. 2017 Dec [cited 2017 Sep 12];5(1). Available from:
641 <http://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-017-0325-z>
- 642 19. Adamovsky O, Buerger AN, Vespalcova H, Sohag SR, Hanlon AT, Ginn PE, et al.
643 Evaluation of Microbiome-Host Relationships in the Zebrafish Gastrointestinal System
644 Reveals Adaptive Immunity Is a Target of Bis(2-ethylhexyl) Phthalate (DEHP) Exposure.
645 *Environ Sci Technol*. 2020 May 5;54(9):5719–28.
- 646 20. Garza DR, van Verk MC, Huynen MA, Dutilh BE. Towards predicting the environmental
647 metabolome from metagenomics with a mechanistic model. *Nature Microbiology*
648 [Internet]. 2018 Mar 12 [cited 2018 Mar 14]; Available from:
649 <http://www.nature.com/articles/s41564-018-0124-8>
- 650 21. McHardy IH, Goudarzi M, Tong M, Ruegger PM, Schwager E, Weger JR, et al. Integrative
651 analysis of the microbiome and metabolome of the human intestinal mucosal surface
652 reveals exquisite inter-relationships. *Microbiome*. 2013;1(1):17.
- 653 22. Pedersen HK, Forslund SK, Gudmundsdottir V, Petersen AØ, Hildebrand F, Hyötyläinen T,
654 et al. A computational framework to integrate high-throughput ‘-omics’ datasets for the
655 identification of potential mechanistic links. *Nature Protocols* [Internet]. 2018 Oct 31 [cited
656 2018 Nov 21]; Available from: <http://www.nature.com/articles/s41596-018-0064-z>
- 657 23. Shaffer M, Thurimella K, Quinn K, Doenges K, Zhang X, Bokatzian S, et al. AMON:
658 annotation of metabolite origins via networks to integrate microbiome and metabolome
659 data. *BMC Bioinformatics*. 2019 Dec;20(1):614.
- 660 24. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal
661 RNA gene database project: improved data processing and web-based tools. *Nucl Acids*
662 *Res*. 2013 Jan 1;41(D1):D590–6.
- 663 25. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, et al. An
664 improved Greengenes taxonomy with explicit ranks for ecological and evolutionary
665 analyses of bacteria and archaea. *The ISME Journal*. 2012 Mar;6(3):610–8.
- 666 26. Beghini F, McIver LJ, Blanco-Míguez A, Dubois L, Asnicar F, Maharjan S, et al.
667 Integrating taxonomic, functional, and strain-level profiling of diverse microbial
668 communities with bioBakery 3. *eLife*. 2021 May 4;10:e65088.
- 669 27. Douglas GM, Maffei VJ, Zaneveld JR, Yurgel SN, Brown JR, Taylor CM, et al. PICRUSt2
670 for prediction of metagenome functions. *Nat Biotechnol*. 2020 Jun;38(6):685–8.
- 671 28. Chong J, Soufan O, Li C, Caraus I, Li S, Bourque G, et al. MetaboAnalyst 4.0: towards
672 more transparent and integrative metabolomics analysis. *Nucleic Acids Research* [Internet].
673 2018 May 14 [cited 2018 May 20]; Available from: [https://academic.oup.com/nar/advance-](https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gky310/4995686)
674 [article/doi/10.1093/nar/gky310/4995686](https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gky310/4995686)

- 675 29. Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, et al.
676 Predictive functional profiling of microbial communities using 16S rRNA marker gene
677 sequences. *Nat Biotech.* 2013 Sep;31(9):814–21.
- 678 30. Douglas GM, Maffei VJ, Zaneveld J, Yurgel SN, Brown JR, Taylor CM, et al. PICRUSt2:
679 An improved and extensible approach for metagenome inference [Internet]. *Bioinformatics*;
680 2019 Jun [cited 2020 Feb 9]. Available from: <http://biorxiv.org/lookup/doi/10.1101/672295>
- 681 31. Kloeke JD, McKean JW. Rfit: Rank-based estimation for linear models,. *The R Journal*.
682 2012;4(2):57–64.
- 683 32. Varma A, Palsson BO. Metabolic Flux Balancing: Basic Concepts, Scientific and Practical
684 Use. *Bio/Technology.* 1994 Oct;12(10):994–8.
- 685 33. Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, et al.
686 Structure, function and diversity of the healthy human microbiome. *Nature.* 2012 Jun
687 13;486(7402):207–14.
- 688 34. Kešnerová L, Mars RAT, Ellegaard KM, Troilo M, Sauer U, Engel P. Disentangling
689 metabolic functions of bacteria in the honey bee gut. Relman D, editor. *PLOS Biology*.
690 2017 Dec 12;15(12):e2003467.
- 691 35. Bauer E, Laczny CC, Magnusdottir S, Wilmes P, Thiele I. Phenotypic differentiation of
692 gastrointestinal microbes is reflected in their encoded metabolic repertoires. *Microbiome*.
693 2015 Dec;3(1):55.
- 694 36. Baldini F, Heinken A, Heirendt L, Magnusdottir S, Fleming RMT, Thiele I. The
695 Microbiome Modeling Toolbox: from microbial interactions to personalized microbial
696 communities. Wren J, editor. *Bioinformatics* [Internet]. 2018 Nov 21 [cited 2018 Nov 28];
697 Available from: [https://academic.oup.com/bioinformatics/advance-](https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/bty941/5194337)
698 [article/doi/10.1093/bioinformatics/bty941/5194337](https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/bty941/5194337)
- 699 37. Drost H-G, Paszkowski J. Biomart: genomic data retrieval with R. *Bioinformatics.* 2017
700 Jan 20;btw821.
- 701 38. Noronha A, Modamio J, Jarosz Y, Guerard E, Sompairac N, Preciat G, et al. The Virtual
702 Metabolic Human database: integrating human and gut microbiome metabolism with
703 nutrition and disease. *Nucleic Acids Research.* 2018 Oct 29;gky992–gky992.
- 704 39. Jaeckel LA. Estimating Regression Coefficients by Minimizing the Dispersion of the
705 Residuals. *Ann Math Statist.* 1972 Oct;43(5):1449–58.
- 706 40. Shapley LS. 17. A Value for n-Person Games. In: Kuhn HW, Tucker AW, editors.
707 Contributions to the Theory of Games (AM-28), Volume II [Internet]. Princeton: Princeton
708 University Press; 1953 [cited 2018 Mar 23]. Available from:
709 [https://www.degruyter.com/view/books/9781400881970/9781400881970-](https://www.degruyter.com/view/books/9781400881970/9781400881970-018/9781400881970-018.xml)
710 [018/9781400881970-018.xml](https://www.degruyter.com/view/books/9781400881970/9781400881970-018/9781400881970-018.xml)

- 711 41. McNally CP, Borenstein E. Metabolic model-based analysis of the emergence of bacterial
712 cross-feeding via extensive gene loss. BMC Systems Biology [Internet]. 2018 Dec [cited
713 2018 Jul 16];12(1). Available from:
714 <https://bmcsystbiol.biomedcentral.com/articles/10.1186/s12918-018-0588-4>
- 715 42. Chiu H-C, Levy R, Borenstein E. Emergent Biosynthetic Capacity in Simple Microbial
716 Communities. PLoS Comput Biol. 2014 Jul 3;10(7):e1003695.
- 717 43. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-
718 source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics.
719 2011;12:77.
- 720 44. Markowitz VM, Chen I-MA, Palaniappan K, Chu K, Szeto E, Grechkin Y, et al. IMG: the
721 integrated microbial genomes database and comparative analysis system. Nucleic Acids
722 Research. 2012 Jan 1;40(D1):D115–22.

723

724

Microbiome data

	#	#	#
#	#	#

1) Assemble taxon-specific community metabolic model

2) Calculate taxon-level and community-level metabolic potential (CMP)

+

-

3) Fit regression model relating metabolic potential and metabolite measurements

Metabolome data

	#	#	#
#	#	#

4) Decompose overall model fit (R^2) into contributions from each taxon

Reference data

KEGG
NCBI
VMH
EMBL_GEMs

A) Reaction network **B) Taxon-level CMP scores** **C) CMP-Metabolite comparison** **D) Taxon contributions**

WILSON, J. D. 1994. *Journal of Great Lakes Research* 20:103-114.

bioRxiv preprint doi: <https://doi.org/10.1101/152111>; this version posted May 10, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1000

1000

100

W

[illegible]

1000

1000

1000

© 2011 Blackwell Publishing Ltd *Journal of Internal Medicine* 270: 103–110

Small size of the study population and cross-sectional study design may be a general limitation in study of the prevalence of dental caries in children. Further, the use of a single dental radiograph may not be sufficient for the detection of all carious lesions. However, the use of a single dental radiograph is a common practice in clinical settings, and the results of this study may be generalizable to other studies using a similar methodology.

1000

--	--

100

100

100

1000

1. **Introduction**
 2. **Background**
 3. **Methodology**
 4. **Results**
 5. **Conclusion**

100

1. *Staphylococcus aureus* (Gram positive)
 2. *Streptococcus pneumoniae* (Gram positive)
 3. *Escherichia coli* (Gram negative)
 4. *Salmonella enteritidis* (Gram negative)
 5. *Shigella flexneri* (Gram negative)
 6. *Neisseria meningitidis* (Gram negative)
 7. *Haemophilus influenzae* (Gram negative)
 8. *Legionella pneumophila* (Gram negative)
 9. *Campylobacter jejuni* (Gram negative)
 10. *Yersinia enterocolitica* (Gram negative)

© 2000 Blackwell Science Ltd, *Journal of Internal Medicine* 247: 395–402

Processing Steps

- Map to AGORA
- Map to RefSeq (embl_gems)
- Map to Greengenes OTUs
- PICRUST
- KEGG gene-reaction mapping
- Precomputed mapping to AGORA genomes
- Precomputed mapping to RefSeq/Embl_Gems
- Custom model modifications

Input options

Community metabolic network model construction

**MIMOSA2
analysis**