Microbiome Metabolome Integration Platform (MMIP): a web-based platform for microbiome and metabolome data integration and feature identification.

Anupam Gautam[1,2,3,$], Debaleena Bhowmik[4,5,$] , Sayantani Basu[6], Abhishake Lahiri[5,7], Wenhuan Zeng[1], Sandip Paul[8]*

$: Equal Contribution

[1] Algorithms in Bioinformatics, Institute for Bioinformatics and Medical Informatics, University of Tübingen, Tübingen, Germany.

[2] International Max Planck Research School "From Molecules to Organisms," Max Planck Institute for Biology Tübingen, Tübingen, Germany.

[3] Cluster of Excellence: EXC 2124: Controlling Microbes to Fight Infection, University of Tübingen, Tübingen, Germany.

[4] Cell Biology and Physiology Division, CSIR-Indian Institute of Chemical Biology, Kolkata, India.

[5] Academy of Scientific and Innovative Research (AcSIR), Ghaziabad- 201002, India.

[6] Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801, United States

[7] Infectious Diseases and Immunology Division, CSIR-Indian Institute of Chemical Biology, Kolkata, India.

[8] Centre for Health Science and Technology, JIS Institute of Advanced Studies and Research Kolkata, JIS University, West Bengal, India

To whom correspondence should be addressed: *Sandip Paul, JIS Institute of Advanced Studies and Research, Arch Waterfront, Kolkata 700091, JIS University, West Bengal, India, Phone: +91-8583916488, Email: websandip@gmail.com; sandipp@jisiasr.org

Abstract

Microbial community maintains its ecological dynamics via metabolites crosstalk. Hence knowledge of the metabolome, alongside its populace, would help us understand the functionality of that community and also predict how it alters in atypical conditions. The metabolic potential of a community from low-cost metagenomic sequencing data signifies the ability to produce or utilize each metabolite and can serve as potential markers of the differentially controlled biochemical pathways among different communities. We developed MMIP (Microbiome Metabolome Integration Platform), a web-based analytical and predictive tool that can describe the taxonomy, diversity variation and the metabolic potential between two sets of microbial communities from targeted amplicon sequencing data. MMIP is capable of highlighting statistically significant taxonomic, enzymatic and metabolic attributes as well as learning based features associated with one group in comparison with another. Further MMIP is capable of predicting the linkages indicating the relationship among species or groups of microbes in the community, a specific enzyme profile related to those organisms, and a specific compound or metabolite. Thus, MMIP can serve as a user-friendly, online web-server for performing most of the analyses of microbiome associated research from targeted amplicon sequencing data, and can provide the probable metabolite signature, along with learning-based linkage associations of any sample set, without the need for initial metabolomic analysis thereby helping in hypothesis generation.

Introduction

Microorganisms being omnipresent take part in the global bio-geochemical cycles as an important player. They also can be seen as an extensive symbiont for all major clades of life (1–4). The study of combined genetic information of microorganisms in any biome is the focus of microbiome research. This manifold and intensifying research field has been exploring the microbial community in an array of environments including ocean, soil, human health etc. In case of human health, the critical role of microbiome is rapidly getting recognized (5, 6). The microbial community structure and function is found to be governed by the environmental factors, host physiology and interactions among the microbes (7–9). Taxonomy and population-based community studies became easily accessible by the advancement of next generation sequencing techniques. This enables better understanding of the microbial community composition, function, dynamics and the inter-relationship with exposome.

For a long time, the primary focus of microbiome studies has been to explore the differences in diversities across different sample groups. These diversity analyses were mostly adapted from ecological studies along with the detailed taxonomic analyses focusing mainly on the differentially abundant taxonomic clades (10). On the other hand, metabolites derived by the microbes play an important role in overall community functionality as they are intertwined by metabolic crosstalk (11). Thus, integration of microbiome and metabolome data is particularly very important in order to understand the intricate interactions of metabolites either between two microbes or between the host and the microbes.

Following this approach, there has emerged a number of ingenious computational tools to decipher the metabolomic yield of a given microbial community (12–16). Motivated from all these approaches we attempt to integrate the two sets of data (microbiome and metabolome) into one single user-friendly online platform, MMIP (Microbiome Metabolome Integration Platform). It is aimed at delineating community-level information and the metabolic potential of microbial communities from targeted amplicon sequencing data.

In this web-platform, we introduce a user-friendly microbiome metabolome integration platform that was developed using Java Script, Python and R programming. This web-resource provides functionalities like diversity analyses, taxonomy profiling, along with metagenome prediction and metabolite potential measurement. A unique feature of our web-platform is a supervised learning-based approach to predict probable connections among the three layers of data (taxonomy abundance, predicted enzyme profiles and metabolite potential score); in other words, it predicts linkages indicating the relationship among species or groups of microbes in the community, a specific enzyme profile related to those organisms, and a specific compound or metabolite.

Methods and Materials

Overview of the Web Server

The Web Interface:
The web interface of MMIP has been designed to offer a simple, user-friendly interaction. The latest versions of bootstrap and HTML were used for the web-interface development and JavaScript further helped to provide the user with various interactive plots and high-quality, downloadable images. We implemented a "job queue" system wherein the job-based Job-ID will be created for accessing the analysis results. Two functionalities viz. MMIP and MMIP-MASS are now available in the platform, where the first one performs general comparative analyses for the metagenomic studies like different diversity indices, taxonomic abundance profile, metabolic pathway profile, etc. for a given set of samples representing two different communities or groups, based on the OTU table and associated metadata uploaded by the user. Different methods are implemented for calculating the statistical significance between the two groups of samples, and are all available as part of the analysis. Similar analyses are also carried out for Kyoto Encyclopaedia of Genes and Genomes (KEGG) database (17–19) compounds and pathways predicted from the taxonomic abundances of the samples between the two groups. While with the MMIP-MASS option, the user can correlate predicted metabolites with the microbial community in relation to their response to the environmental changes, with real-time metabolomic data.

Input Data Type:
MMIP requires a closed-reference otu-picked biom table using Greengenes 16S database (with option of versions13_5and 13_8) (20, 21) along with the associated metadata file in QIIME (22) format. For the MMIP-MASS option, users are required to upload the metabolic feature table (in a tab-delimited ".txt" file format) where the columns represent the sample names (IDs), while the rows contain the compound IDs.

The MMIP Algorithm

The overall MMIP working flowchart is depicted in figure 1. The two major microbial community feature predictions were incorporated in the MMIP; the details are described as follows.

Reconstruction of metagenome content from Amplicon (16S rRNA metagenomic) data:

A predictive metagenomic approach, PICRUSt (23) was utilized for this purpose. Wherein, the OTU abundance table is normalized by predicted 16S copy number of individual OTU followed by metagenomic function prediction, referred to as the "gene content inference step", and followed by the "metagenome inference" step. The first step provides an annotated table (represented as matrix A_{nk} where n-rows represent the predicted genes, in our case KEGG Orthology or KO identifiers, while k-columns represent all the samples) of the predicted gene family counts for individual sample. The second step provides the values representing the extent of contribution by all mapped OTUs to their respective gene families.

Estimation of community-wise metabolic potential (CMP):

The established method by Noecker et.al (13) was utilized for community-wise metabolite prediction, for which the predicted gene list (KO identifiers) is the input for the prediction of the metabolic potential of the community. The KO identifiers are mapped to the KEGG reaction map using a surjective function. Based on the directionality of the mapped reactions the compounds that are consumed or produced can be estimated. Further, these reactions are mapped to the metabolic pathways. Based on the reaction stoichiometry, we developed a matrix B_{mn} of dimensions m*n, where m (rows) holds the compound identifiers and n (columns) holds the KO identifiers (enzymes /genes) and each of the cells stores the stoichiometric information; it implies that if a compound is present at the reactant side its stoichiometry will be placed in the respective cell of the table with a negative sign and if a compound lies at the product side its stoichiometry will be placed with a positive sign. If the same compound is being used by the same KEGG enzyme in other pathways then its stoichiometry values will be added or subtracted in the table, depending on the position of the compound in that particular reaction (either on the reactant or the product side). Taking the stoichiometry of these reactions into account, the stoichiometric matrix B_{mn} is then normalized, such that the sum total of positive vectors in a row is +1 and that of negative vectors in each row is -1 leading to matrix C_{mn} where m (rows) again holds the compound identifiers and n (columns) holds the KO identifiers (enzymes /genes).

The cross product of generated matrix C_{mn} (m*n) with that of the predicted metagenome function matrix A_{nk} (n*k) results into a Community Metabolic Potential (CMP) score matrix D_{mk} having community-wise metabolic potential for each of the metabolites (or compounds) in the corresponding samples, where m (rows) and k (columns) represent the metabolites and the sample names respectively. The CMP score much like PRMT(12) is unit-less that only predicts the relative capacity of the microbial community to produce or to consume each metabolite. The mathematical expression of the CMP calculation is shown in supplementary figure 1.

Methodology of the Modules in MMIP

Basic Analyses:

The user uploaded OTU table and metadata file is processed through our Python pipeline which involves a sanity check step (removal of singletons OTUs observed fewer than 2 times, normalizing by copy number for PICRUSt, etc), followed by determination of the CMP score as described in the algorithm section; beside that, the same OTU table and metadata file passes through our in-house R pipeline which uses the Phyloseq package (24) for the determination of the taxonomic variation at different levels between two groups, along with the alpha-diversity analyses using Shannon and Simpson diversity indices and the beta-diversity analysis using Bray-Curtis dissimilarity. Further we designed our own R function to carry out the differential analysis at the KEGG functional levels.

We used R for all the statistical analysis in our framework including ANOVA, t-test etc at both, taxonomic and functional levels. Finally, all the results are accessible to the user in a visualization form (downloadable diagrams) with the help of Plotly (25), D3 (26), Krona Plot (27) and flat file on the webpage interface.

Feature prediction:

The CMP matrix that is being formed in the backend, as mentioned previously, is then further used along with the OTU table, to predict features that discriminate between the sample groups. This is one of the most innovative aspects of our web-server and is explained with figure 2.

We have used Support Vector Machine (SVM), which is a popular supervised learning approach that predicts the hyperplane separating the data points in a multi-dimensional space. For each dataset, we have generated a gene prediction table, a compound metabolic potential score table and a metagenome abundance table (OTU Table) with sample-specific data that were concatenated sample-wise in order to perform classification. The mathematical formulation of SVM can be stated according to Equation 1 with constraints as per Equation 2 where x represents the variables, y is the class outcome where y∈{-1,+1} (for an SVM model having two classes), C is a hyperparameter used for regularization, w represents the coefficient vector, b is a constant, ξ is the set of slack variables and φ represents the kernel used for data transformation.

$$min_{\mathbf{w},b,\xi_{1:N}} \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{N}\xi_i \quad (1)$$

$$y_i(\mathbf{w}^T\varphi(x_i) + b) \geq 1 - \xi_i \; for \; i = 1, \dots, N, \xi_i \geq 0 \quad (2)$$

Based on experimentation, we use the ‘linear’ kernel as part of our model and tune other hyperparameters using Grid Search. Leave-One-Out was applied for performing the cross-validation on the SVM model in order to eliminate the class imbalance problem that resulted from random sub-sampling of data in conventional k-fold cross validation when the value of k was chosen to be smaller than the number of samples in the dataset. We use the ROC-AUC metric in order to judge the performance of our classifier. Feature ranking was done based on the

coefficients of feature contributions extracted from our trained SVM model and ranking based on the values of these coefficients.

We then extracted the ten most positive and ten most negative CMP ID features from our list of features. Using the contribution file, these features are then linked to the corresponding KEGG IDs and OTU IDs. Based on our experiments, we have set a threshold of 0.7 for presence of OTU IDs in a certain number of samples. This means that the corresponding mean OTU contributed would reflect a certain quantity only if the number of samples for that OTU ID exceeded the threshold value, otherwise the value in the mean columns will be zero. The features are then linked according to their feature ranking obtained earlier. The three-tuples that have positive ranks for all three features (the linked CMP ID, KEGG ID and OTU ID) have a greater probability of contributing to the positive side of the hyperplane and hence represent collective features for this particular group. Similarly, the three-tuples that have negative ranks for all three features have a greater probability of contributing to the negative side of the hyperplane and hence represent collective features for the other group considered for classification.

Added module in MMIP-MASS:
One can easily correlate predicted metabolic information with real-time metabolic data (LC-MS, GC-MS feature table) using the MMIP-MASS module. Along with OTU and metadata files, the user has to provide metabolic features where the column represents the sample names (ID) and the row contains the compound ID (KEGG Compound ID). The framework tries to look for the compounds or metabolites that are common between the predicted ones and the one obtained after real-time analyses. It then applies the mantel test on the distance matrices of the matched compounds and checks if the predicted ones are correlating positively or negatively with the real-time analysed one. If the correlation co-efficient, which lies within the range of -1 to +1, has leans more toward +1 then both predicted and matched are positively correlated and if it is more toward -1 then both predicted and matched are negatively correlated. This ultimately provides the user with the most probable microbial source of the compounds or metabolites.

Dataset selection for validation of our webtool

For the validation of our study, we selected two studies with publicly available amplicon sequencing data. The first study selected is by Srinivasan et. al. 2015, on Bacterial Vaginosis (BV) and its metabolome (28). We used the available 16S rRNA sequencing (V3-V4 region) data (accession no. PRJNA275907) along with the LC-MS (liquid chromatography mass spectrometry) and GC-MS (gas chromatography mass spectrometry) data. In the study there were two cohorts, of which we selected the cohort B containing 40 BV positive and 20 BV negative subjects. The condition that decides whether a female is BV+ or not, depends on the Nugent score and the Amsel criteria, which was also the basis of grouping the females in this study. The second study was by Dhakan et. al. 2019, who worked with gut microbiome of

omnivorous diet group and vegetarian diet group (accession no. PRJNA397112) and also checked the metabolome with GC-MS analysis (29).

The amplicon data was re-analysed using standard QIIME pipeline (22), and the final biom file or the OTU table was obtained after closed-reference OTU picking against Greengenes v13.8 database (20, 21). The OTU table and its associated metadata in QIIME format were then uploaded into our MMIP web server. All the analyses were also performed separately for validation of results with those obtained from MMIP. The predicted metabolites given by MMIP were matched with those reported in the study, after GC-MS and LC-MS of the actual samples, in order to check the accuracy of our prediction model

Result:

Overall community-wise profile analyses between Healthy and Bacterial Vaginosis groups:

The validation dataset obtained from the study of Srinivasan et. al. (2015) contained 40 Bacterial Vaginosis (BV+) samples and 20 BV- samples. We performed basic microbiome data analyses separately and uploaded the processed data in MMIP for comparison. Starting with the alpha diversity or the intra-sample diversity, we used both the Shannon and the Simpson diversity indices to show the diversity among samples belonging to both the groups and also calculated the ANOVA One-way significance test. As the results in table in Supplementary Figure 2(a) suggests that the values depicted by MMIP not only captures the significance level appropriately, but the Shannon Index value is quite close. Also, as the boxplot in Supplementary Figure 2(a) suggests that there is a significant increase in the median value of the BV+ (BV) samples compared to the BV- (H) ones, which clearly indicates the increase in the diversity of bacterial species in the vagina during bacterial vaginosis, which otherwise in the healthy condition is mostly dominated by population of *Lactobacillus*.

Moving to the beta diversity analysis, Supplementary Figure 2(b) depicts how the BV+ samples separate quite remarkably from that of the BV- samples MMIP. Moreover, after performing Adonis test the p-value is found to be significant (0.001). Similarly, when we compared this beta diversity results to the analysis we did separately, they gave equivalent results and most importantly the p-value was significant too (0.001). This significance is also quite valid as it shows that the Healthy (H) or BV- samples tend to be extensively different from the BV+ ones as far as bacterial population is concerned.

The taxonomic profile also demonstrated quite similar profiles from MMIP as found in our analysis, and reported in the study, but most importantly it can highlight the difference in the two groups appropriately (supplementary figure 3). As we know that *Lactobacillus* belonging to the Phylum Firmicutes, is the most abundant microbial genus found in a healthy vagina (30, 31),

maintaining the acidic environment by production of lactic acid or lactate, our results depict this scenario with significant relevance. The phylum-wise taxonomic plot showed abundance of Firmicutes while the genus-wise plot showed abundance of *Lactobacillus* both in the healthy (H or BV-) group; unlike the BV+ group, where in both the representation (phylum and genus-wise plots) it showed that the samples contained higher percentage of other bacteria, causing a dysbiotic vaginal environment.

The downloadable files showing the significant OTUs of both the groups also validate our findings as per the study, showing the dominating presence of *Lactobacillus* species (including *Lactobacillus iners*) in the healthy group while being quite negligible in BV group. Contrary to this the BV group shows predominance of bacteria genus like *Megasphaera*, *Gardnerella*, *Prevotella* and *Sneathia*, and also *Atopobium vaginae*, all of which were termed as BV-associated bacteria in the study.

Predicted metabolites and associated pathways distinguishing the two health states of vagina as mentioned in the study:

The metabolite list from the study by Srinivasan et. al. (2015) reported 108 metabolites with their KEGG annotations and was used for our comparison (28). From the MMIP predicted list of significant metabolites, 52 overlapped with their list and 21 of those 52 metabolites depicted the same trend as reported in the study. We provide a relative value that signifies the capacity of that community to consume or accumulate a particular metabolite. Like for example cadaverine, in the study they showed that the BV+ samples had a higher amount of this compound; in our prediction we showed that compared to that of BV- (or H) group, the relative capacity of cadaverine being accumulated in the BV+ samples is significantly higher than that of BV- samples.

With regards to amino acids, we correctly predicted the lowering of amino acids in BV+ samples. The study reported that 18 out 19 amino acids that they found in the samples, were lower in amount in BV+ samples compared to that of BV- or healthy samples; there is an increase in amino acid catabolites in BV+ samples. The amino acids that lowered in our prediction included alanine, cysteine, histidine, tryptophan, phenylalanine etc. and the dipeptide, alanylalanine. Also, the compounds that are involved in amino acid metabolism pathways like spermine and glutathione, were also found to be low in BV+ and matched with that of the study. While drawing a biological inference from this result we can say, an increase in the number of bacterial taxa and not just *Lactobacillus*, probably cause an increased usage of amino acids since they are essential during cell growth; correlating to this pattern we can also say that this should also increase the amount of amino acid catabolites, and this too is confirmed by our prediction. Thus, the increase in amino acid catabolites and decrease of amino acids in BV+ samples with increased bacterial diversity.

Focusing on the carbohydrate metabolism associated metabolites, the study reported lowering of simple sugars in BV+ environment, and the same is observed in our results too; maltose being one of them, that overlap with them and follow the correct trend. Also, glucose-6-phosphate, and sugar alcohols like sorbitol and mannitol are other reported carbohydrate metabolism metabolites that are correctly predicted and maintain the trend of being present in lower amounts in BV+ cases. The same explanation that justified our amino acids metabolite prediction could also justify the case for carbohydrate metabolism metabolites; while the simple sugars and intermediate metabolites are being used up, the carbohydrate catabolites tend to accumulate in BV+ samples with the growing diversity of microbes.

They also reported lowering of glycerol, ethanolamine and glycerophosphorylcholine (GPC), all of which are associated with lipid metabolism, in BV+ samples and they too were correctly predicted.

Prediction of features at three different levels and the links among these three levels of features could be extrapolated to hypothesize biological implications:

To interpret the biological correlation and provide a relevant linkages, we have predicted among the three levels of data – OTU IDs giving information on community structures, enzyme profiles (KEGG IDs) and metabolites – again keeping in mind that when talking about the CMP scores of each metabolite, it does not predict the net usage or synthesis nor the net amount present in a samples; a positive score implies relatively increased utilization while a negative score implies relatively more accumulation while comparing between the two concerned groups of samples. Thus, a positive score could possibly imply a relatively low number of certain metabolites, and quite opposite in case of a negative score, when compared between the two groups.

One of the positive feature links depicted in supplementary figure 4(a) for this dataset is predicted for the compound C00089 (sucrose), where it is associated with the enzyme sucrose phosphorylase (K00690, part of carbohydrate metabolism reaction map) and further to *Lactobacillus* species. Being abundant in the healthy or BV- group, it can thus be expected that high *Lactobacillus* marks an increased utilization of the metabolite via the above-mentioned enzyme. We also inspected a negative feature link for this dataset, as shown in supplementary figure 4(b), which is predicted for the compound C00436 (N-Carbamoylputrescine). As per MMIP linkage, this is associated with the enzyme Agmatine deiminase (K10536) and in turn with the bacterial family *Coriobacteriaceae*, harbouring genus like *Eggerthella* sp. and *Atopobium* sp. specific to the BV+ group according to the study. The link thus estimates that increase of these species in BV+ group would mean increased action of the enzyme yielding more of the metabolite mentioned.

Community-wide differences observed in groups due to different dietary habits:

The study by Dhakan et. al. (2019) recruited a total of 110 individuals of different food habits, mainly omnivorous and vegetarian diet, respectively belonging to the Southern and the North-Central states of India. We processed the data and filtered out the samples with extremely low count of OTUs, we uploaded in MMIP a sample set of 57 omnivorous diet and 35 vegetarian diet samples. The alpha diversity indices (both Shannon and Simpson) gave comparable results with that of our analysis and so was the beta diversity, which also gave a significant p-value; and thus, point out the notable difference amongst the two sample groups, with respect to bacterial compositions reported in the study (Supplementary figure 5).

Looking at the taxonomic profiles, the composition obviously looks different for the two groups with different dietary habits, at both phylum and species level, but to pin-point the significant OTUs that standout in both, omnivorous and vegetarian diet groups, the significant OTUs files provide a more precise picture. Correlating with the findings of the study, we too found few of the bacterial genus that were associated with either of the diet groups; omnivorous group was enriched in *Ruminococcus* and *Faecalibacterium* (including *F. prausnitzii*), and even though in smaller proportions, *Roseburia* and *Akkermensia muciniphila* also had a significant occurrence; while that of the vegetarian group had the unique residence of the genus *Prevotella* and *Megasphaera*as reported in the study (Supplementary figure 5).

Pathways and metabolites overlapping with that of the study:

Dhakan et. al highlighted a total of 5928 KEGG pathways in their study which were found enriched forming two clusters, each associating with a particular diet group. Our KEGG pathway level feature prediction resulted in a set of feature pathways (K-Ids) that could serve as distinguishing factors between the two groups, omnivorous and vegetarian; and 4918 of those overlapped with those from that of the study. As far as metabolites are concerned, the study identified 13 annotated metabolite clusters, 4 of which overlapped with our results and showed similar enrichment trends between the two groups (Supplementary table 1).

While analysing the individual trends of the predicted metabolites, MMIP speculates, quite correctly, many of the metabolites as mentioned in the study, including serine (C00065), butyrate (C00246) and valine (C00183), all of which are found to be associated with that of the omnivorous diet group. For serine, MMIP reveals a higher negative CMP score in the omnivorous group, implying that the omnivorous group associated microbes have a relatively higher potential of serine synthesis as compared to that of the vegetarian group associated microbes; the same holds true for valine as well. The study also correlates SFCAs (short chain fatty acids) like butyrate, propionate, with that of omnivorous diet, and we too observe this in our

webtool; butyrate is found to have a negligible CMP score in vegetarian group, while many of the samples of the omnivorous group show a negative CMP score concluding that again this group has a higher potential of synthesizing butyrate.

Predicted links between the three levels supporting biological relevance:

As in the previous case, we tried to validate and provide biologically relevant explanation for the links predicted by MMIP in this study too, and showed that our webtool can not only distinguish between two groups of "healthy" and "non-healthy" states, but also any two sample groups discriminated by any parameters (like dietary habits as in this case). The positive link is predicted for the compound Manninotriose (C05404) and is linked with that of enzyme alpha-galactosidase (K07406) and members of *Ruminococcaceae* family and of genus *Ruminococcus*, given in supplementary figure 6(a). These microbes are related to the omnivorous diet group and could imply the increased utilization of the associated metabolite. Similarly, the other link was predicted for the compound sucrose-6-phosphate (C02591) associated with the enzyme sucrose-6F-phosphate phosphohydrolase (K07024) and in turn with microbial genus *Ruminococcus* and *Faecalibacterium*, found in omnivorous diet group, shown in supplementary figure 6(b). This linkage too shows increased utilization of the metabolite via the enzyme in the given set of microbes.

Discussion

For a genuine understanding of the microbial community functional dynamics, taxonomic profiling alone wouldn't suffice. To understand why a certain microbe is present in far greater proportions than another, but also not completely eliminating the existence of the latter, lies in their metabolites interchange; these small molecules efficiently orchestrate the homeostasis of microbial community (32). In short, to understand the microbiome in depth, we need to understand the metabolome aspect of it as well. Thus, our attempt to connect the two, developing MMIP. Our user-friendly web-server incorporates a number of features that might help the user speculate overall metabolomic functionality from targeted amplicon sequencing data. This server performs all the basic microbiome associated analysis, including α-diversity, β-diversity, taxonomic profiles, and represents them in well-informative diagrams; not only that it exhibits pathway analysis as well as metabolite prediction when comparing between two groups (healthy and diseased cohorts). If the user has metabolomic data (mass spectrometry data) of the associated samples, then MMIP also results in a comparative analysis between the actual metabolomic data and the predicted metabolites, of the samples. In addition to all of the above-mentioned features, the highlight of our server is not just predicting the significant metabolites but also forming biologically pertinent feature linkages between the predicted metabolites, associated pathways and microbes. Analysing these linkages may show that how these microbes

could be associated with the consumption or secretion of certain metabolites, which might not always be the end products, but could also be intermediates of a much extensive pathway.

We have validated the functionality and efficacy of our web-server tool using two datasets from two different studies, both from amplicon data of microbiome from different body niche. This has enabled us to check the versatile nature of MMIP. The first study was by Srinivasan et. al. (28) working with vaginal microbiome data of healthy and individuals affected with Bacterial vaginosis. As represented in detail in the results section, the basic microbiome analysis shows high level of accuracy with ones analysed separately, and also correlating really well with the findings of the study itself. The taxonomic data clearly depicts the dominating abundance of the genus *Lactobacillus* in healthy group, which is quite expected in as in a healthy vaginal environment; while in the BV group bacterial genus like *Megasphaera*, *Gardnerella*, *Prevotella* and *Sneathia*, and also *Atopobium vaginae*, show a remarkable increase, and as demonstrated in the study itself. 52 out of the 108 reported metabolites overlapped with the MMIP predicted list of significant metabolites. Finally, the positive and negative feature linkage illustrates how our predicted linkages could actually hold biological validity, which is explained as follows.

Supplementary figure 4(a) tries to explain how the positive link could be interpreted biologically and biochemically. Sucrose (C00089) is a disaccharide made of fructose and glucose and when acted upon by the enzyme sucrose phosphorylase (K00690, part of carbohydrate metabolism reaction map), it is broken down to form fructose and glucose-1-phosphate (KEGG reaction ID R00803). This enzyme or pathway is in turn associated with organisms including *Lactobacillus iners* and other *Lactobacillus* species, according to MMIP prediction. Now according to the study, these are the microbes which are associated with the BV- group, or the healthy group. Increase in the utilization of sucrose (C00089) is shown by the high positive CMP score in healthy group (H or BV-), suggesting that the forward reaction is taking place (reaction in the diagram above) and hence there would be an increased formation of fructose, which again according to the study is less in BV+ compared to that of BV- or healthy. Thus, this confirms the validity of one of our positive feature links.

According to the reaction in the supplementary figure 4(b), N-Carbamoylputrescine (C00436) is formed as the product (KEGG reaction ID R01416) and the reaction is catalysed by the enzyme Agmatine deiminase (K10536); also, this is associated with bacterial family *Coriobacteriaceae* as predicted by MMIP. This bacterial family harbours genus like *Eggerthella* sp. and *Atopobium* sp. that are associated with BV+ samples according to the study. Also, agmatine (reactant in above reaction) is a metabolite formed after Arginine catabolism, which would justify lowering this amino acid the BV+ samples. Thus, interpretation of the link suggests, increase in the population of these BV+ associated bacteria might be associated with an increase in the enzyme action of Agmatine deiminase, along with the entire Arginine degradation pathway, allowing the forward reaction and hence more production of the compound N-Carbamoylputrescine, a

precursor for putrescine, a compound which again is found to be higher in BV+ samples as per the study. The CMP score of N-Carbamoylputrescine shows that this compound is highly accumulated in BV+ samples and thus favouring our linkage prediction that this might be an important metabolic reaction and metabolite for the BV+ associated bacteria in BV+ samples.

In our second validation, we selected the gut samples of two groups, one with omnivorous diet and the other with vegetarian diet, from the study by Dhakan et. al. (29). The taxonomic distribution between the two groups run at par with that of the study, showing significant presence of bacteria *Roseburia* and *Akkermensia muciniphila* in the omnivorous group and that of *Prevotella* and *Megasphaera* in case of vegetarian group. Here in the study, they additionally checked for pathways associated with the two forms of diet and when compared with MMIP results there was a remarkable overlap with our pathway analysis; and MMIP was able to predict the correct trend of three of them (serine, butyrate and valine) which was critically highlighted in the omnivorous group in the study. If not all, it is evident that MMIP is able to predict many of the crucial metabolites that differentiate between any two groups. Eventually the last section of validation study 2 shows the positive and negative predicted feature linkages, elaborated ahead.
Describing a positive feature link as portrayed in supplementary figure 6(a), we found that the compound Manninotriose (C05404) has a relatively higher positive CMP score which corresponds to a high utilization potential by the associated microbes. This sugar-metabolite is metabolized with the help of the enzyme alpha-galactosidase (K07406) as per KEGG reaction R05549 to form D-Galactose (C00124) and Melibiose (C05402). Now this particular enzyme was reported to be high in the omnivorous group in the study and so was D-Galactose. This link also connects OTUs with taxonomies associated with the omnivorous diet group and includes microbes from the *Ruminococcaceae* family, phylum Firmicutes and few from genus *Ruminococcus*. Thus, the interconnection could be summed up as follows – increase in the omnivorous diet specific microbes also implies increased activity of the associated alpha-galactosidase enzyme, favouring forward reaction, depleting manninotriose while increasing D-Galactose in the system.

Supplementary figure 6(b) describes the link from the negative feature section of MMIP. Sucrose-6-phosphate (C02591) is a reaction intermediate compound that has a relatively higher negative CMP score associated with omnivorous diet implying higher chance is synthesis. This compound is broken down by the enzyme sucrose-6F-phosphate phosphohydrolase (K07024) which is correlated with omnivorous diet as per the study, to form sucrose (C00089) (KEGG reaction ID R00803). The sucrose then in turn could be broken down to D-Fructose (C00095) and D-Glucose (C00031) via KEGG reaction R00801 using any of the following enzymes – oligo-1,6-glucosidase (K01182), alpha-glucosidase (K01187) or beta-fructofuranosidase (K01193) – all of which were found to be associated with omnivorous group by the study, thus proving the validity of the prediction. Another proof of concept lies in the fact that this particular link of MMIP connects the OTUs belonging to taxonomies genus *Ruminococcus*,

Faecalibacterium prausnitzii and of phylum Firmicutes, all of which in turn were shown to be characteristic of the omnivorous group. Thus, connecting the dots, we can infer the omnivorous group is enriched in the above-mentioned bacteria, implying an increased action of the associated enzymes. Thus, an elevated conversion from sucrose–6-phosphate to sucrose and eventually to D-Glucose and D-Fructose, both of which were reported to be increased in omnivorous diet.

There are certain key points that the user should always be aware of while using MMIP, the most essential being, it gives the potential value of the community's utilization or consumption of any metabolite concerned, and not the absolute content or metabolism rate of that metabolite. One must also keep in mind that the metabolites are "predicted" which means they are highly confident assumptions, and could only be validated experimentally or extensive analysis of the gene content obtained from whole genome sequencing of the samples. Like many computational tools, MMIP too has a few limitations, including that the reference database of the OTU table should be Greengenes only; also, not all the predicted features and linkages are statistically significant, but since machine leaning based models are predictive models while significance test is probabilistic, it might not always find features that show a significant p-value. Further, the feature links formed in MMIP is useful only to build a preliminary hypothesis about the types of reactions that goes on in that microbial community; they aren't absolute data, and needs literature-based validity checking and proper experimental design for confirmation.

In conclusion, MMIP can serve as a quick, easy-to-use, online web-server which performs most of the analyses of microbiome associated research from targeted amplicon sequencing data, and can provide the user with a hassle-free basic initial idea of the probable metabolite signature of any sample set, without the need for metabolomic analysis. Alongside, MMIP also could help the user form a preliminary idea of the metabolic potential or the enriched pathways of the microbial communities; and further help the user form a biologically valid hypothesis to get an overall contour of the two groups of samples.

Data and code availability:

The web server is available at https://github.com/websandip/mmip and data for validation was used from the already published study which can be accessed by NCBI BioProject through Project ID mentioned in the main text. Further detailed manual on how to use MMIP is available at https://github.com/websandip/mmip/blob/main/manual.pdf

Acknowledgment:
DB would like to thank ICMR for providing her with the fellowship. AG and WZ acknowledge support by the High Performance and Cloud Computing Group at the Zentrum für Datenverarbeitung of the University of Tübingen, the state of Baden-Württemberg through bwHPC, the German Research Foundation (DFG) through grant no. INST 37/935-1 FUGG. AG

also acknowledge support of the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A532B, 031A533A, 031A533B, 031A534A, 031A535A, 031A537A, 031A537B, 031A537C, 031A537D, 031A538A)

Conflict of Interest:

None declared.

References

1. Falkowski PG, Fenchel T, Delong EF. 2008. The Microbial Engines That Drive Earth's Biogeochemical Cycles. Science 320:1034–1039.
2. Oh D-C, Poulsen M, Currie CR, Clardy J. 2009. Dentigerumycin: a bacterial mediator of an ant-fungus symbiosis. Nat Chem Biol 5:391–393.
3. McFall-Ngai M. 2014. Divining the Essence of Symbiosis: Insights from the Squid-Vibrio Model. PLOS Biol 12:e1001783.
4. Kawaguchi M, Minamisawa K. 2010. Plant–Microbe Communications for Symbiosis. Plant Cell Physiol 51:1377–1380.
5. Carabotti M, Scirocco A, Maselli MA, Severi C. 2015. The gut-brain axis: interactions between enteric microbiota, central and enteric nervous systems. Ann Gastroenterol Q Publ Hell Soc Gastroenterol 28:203–209.
6. Thaiss CA, Zmora N, Levy M, Elinav E. 2016. The microbiome and innate immunity. Nature 535:65–74.
7. Bonder MJ, Kurilshikov A, Tigchelaar EF, Mujagic Z, Imhann F, Vila AV, Deelen P, Vatanen T, Schirmer M, Smeekens SP, Zhernakova DV, Jankipersadsing SA, Jaeger M, Oosting M, Cenit MC, Masclee AAM, Swertz MA, Li Y, Kumar V, Joosten L, Harmsen H, Weersma RK, Franke L, Hofker MH, Xavier RJ, Jonkers D, Netea MG, Wijmenga C, Fu J, Zhernakova A. 2016. The effect of host genetics on the gut microbiome. Nat Genet 48:1407–1412.
8. Dyhrman S, Ammerman J, Van Mooy B. 2007. Microbes and the Marine Phosphorus Cycle. Oceanography 20:110–116.

9. Krause E, Wichels A, Giménez L, Lunau M, Schilhabel MB, Gerdts G. 2012. Small Changes in pH Have Direct Effects on Marine Bacterial Community Composition: A Microcosm Approach. PLOS ONE 7:e47035.
10. Birtel J, Walser J-C, Pichon S, Bürgmann H, Matthews B. 2015. Estimating Bacterial Diversity for Ecological Studies: Methods, Metrics, and Assumptions. PLOS ONE 10:e0125356.
11. Lee-Sarwar KA, Lasky-Su J, Kelly RS, Litonjua AA, Weiss ST. 2020. Metabolome–Microbiome Crosstalk and Human Disease. Metabolites 10:181.
12. 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 dataset. Microb Inform Exp 1:4.
13. 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:e00013-15.
14. Noecker C, Eng A, Muller E, Borenstein E. 2022. MIMOSA2: a metabolic network-based tool for inferring mechanism-supported relationships in microbiome-metabolome data. Bioinforma Oxf Engl 38:1615–1623.
15. Mallick H, Franzosa EA, Mclver LJ, Banerjee S, Sirota-Madi A, Kostic AD, Clish CB, Vlamakis H, Xavier RJ, Huttenhower C. 2019. Predictive metabolomic profiling of microbial communities using amplicon or metagenomic sequences. Nat Commun 10:3136.
16. Shaffer M, Thurimella K, Quinn K, Doenges K, Zhang X, Bokatzian S, Reisdorph N, Lozupone CA. 2019. AMON: annotation of metabolite origins via networks to integrate microbiome and metabolome data. BMC Bioinformatics 20:614.
17. Kanehisa M. 2019. Toward understanding the origin and evolution of cellular organisms. Protein Sci Publ Protein Soc 28:1947–1951.
18. Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27–30.
19. Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. 2023. KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Res 51:D587–D592.

20. 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.
21. 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. 3. ISME J 6:610–618.
22) 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.

23. Langille MGI, 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.
24. McMurdie PJ, Holmes S. 2013. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLOS ONE 8:e61217.
25. Inc PT. 2015. Collaborative data science. Plotly Technologies Inc., Montreal, QC. https://plot.ly.
26. Bostock M, Ogievetsky V, Heer J. 2011. D3: Data-Driven Documents. IEEE Trans Vis Comp Graph Proc InfoVis.
27. Ondov BD, Bergman NH, Phillippy AM. 2011. Interactive metagenomic visualization in a Web browser. BMC Bioinformatics 12:385.
28. 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.
29. Dhakan DB, Maji A, Sharma AK, Saxena R, Pulikkan J, Grace T, Gomez A, Scaria J, Amato KR, Sharma VK. 2019. The unique composition of Indian gut microbiome, gene catalogue, and associated fecal metabolome deciphered using multi-omics approaches. GigaScience 8:giz004.

30. Antonio MAD, Hawes SE, Hillier SL. 1999. The Identification of Vaginal Lactobacillus Species and the Demographic and Microbiologic Characteristics of Women Colonized by These Species. J Infect Dis 180:1950–1956.
31. Atassi F, Brassart D, Grob P, Graf F, Servin AL. 2006. Lactobacillus strains isolated from the vaginal microbiota of healthy women inhibit Prevotella bivia and Gardnerella vaginalis in coculture and cell culture. FEMS Immunol Med Microbiol 48:424–432.
32. Rooks MG, Garrett WS. 2016. Gut microbiota, metabolites and host immunity. 6. Nat Rev Immunol 16:341–352.

Figure Legends:

Figure 1. The entire workflow that runs at the backend of MMIP. Keys define different categories for each section of workflow.

Figure 2. Flowchart for the classification and feature extraction of the two groups for MMIP.

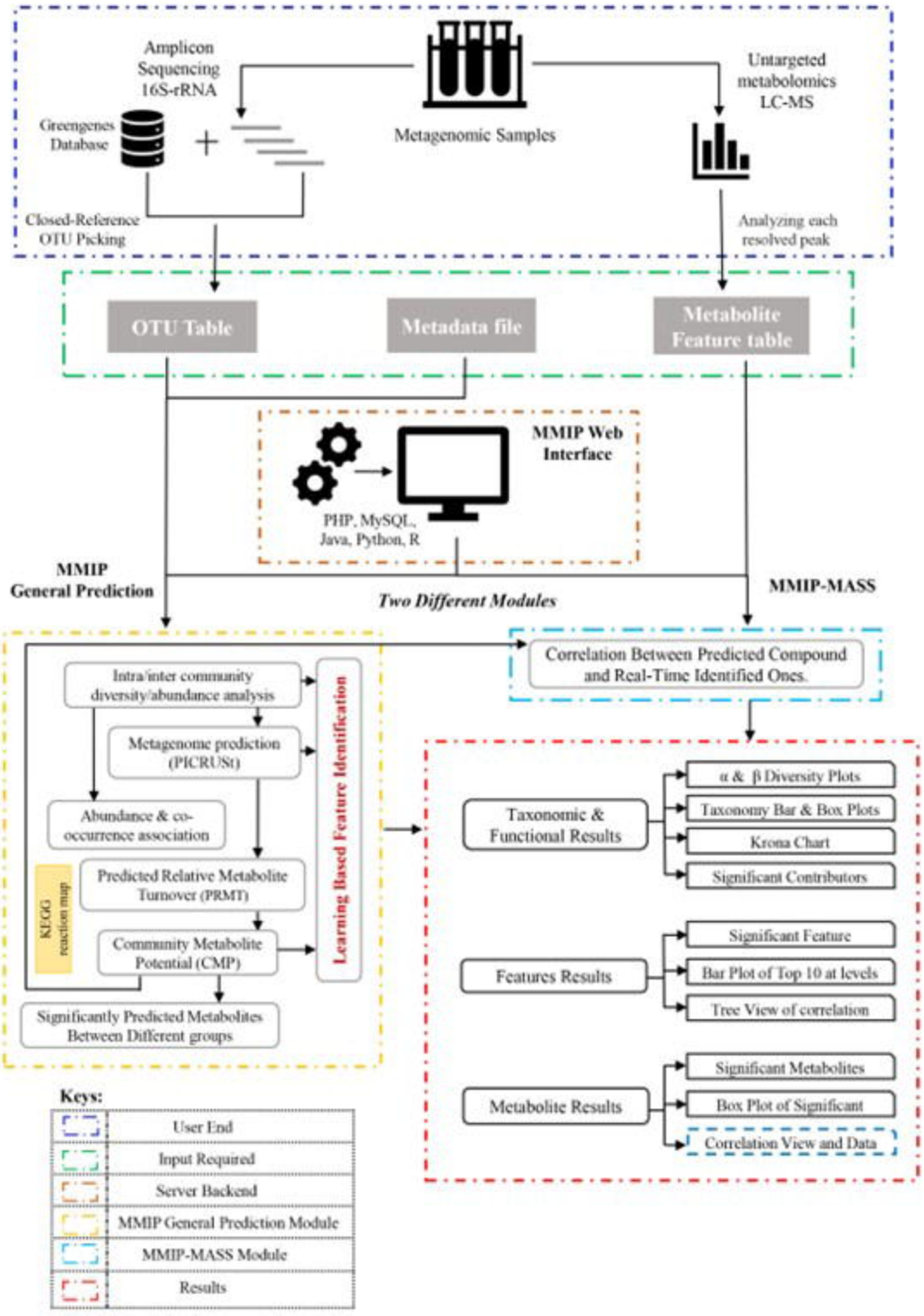

Figure 1. The entire workflow that runs at the backend of MMIP. Keys define different categories for each section of workflow.

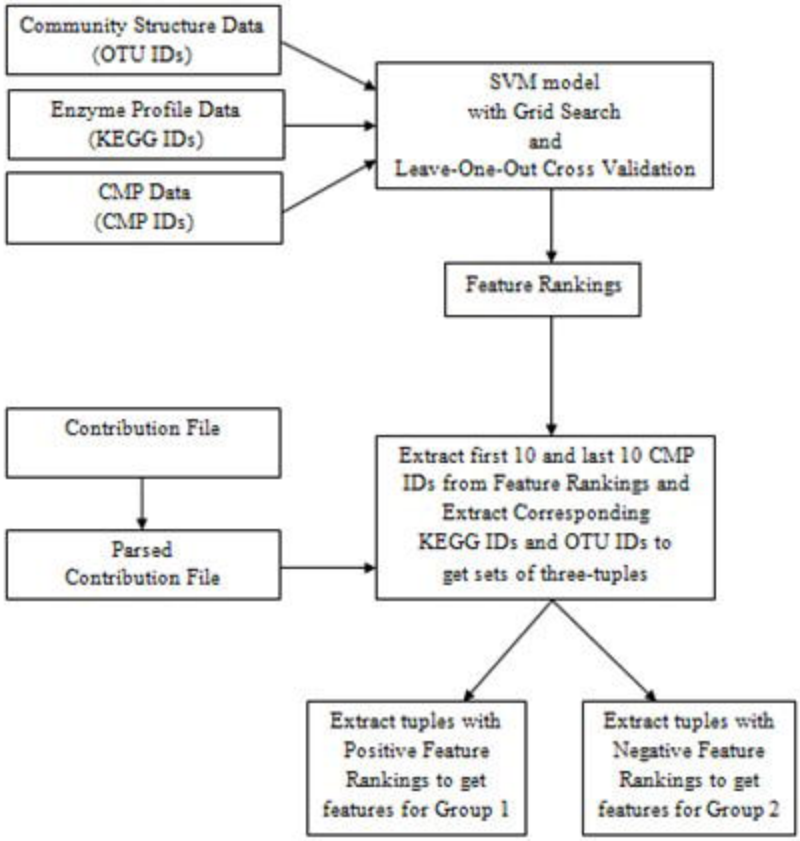

Figure 2. Flowchart for the classification and feature extraction of the two groups for MMIP.