Metagenome gene prediction of naturally fermented milk products of India using PICRUSt2 and Piphillin

Jyoti Prakash Tamang (jyoti_tamang@hotmail.com)
Sikkim University
H. Nakibapher Jones Shangpliang
Sikkim University
Ranjita Rai
Sikkim University

Research article

Keywords: Metagenome gene prediction, PICRUSt2, Piphillin, naturally fermented milk products, DADA2, Deblur.

Posted Date: September 29th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-78771/v1

Abstract

Background: Naturally fermented milk (NFM) products are popular food delicacies in Indian states of Sikkim and Arunachal Pradesh. Bacterial communities in these NFM products of India were previously analysed by high-throughput sequence method. However, predictive gene functionality of NFM products of India has not been studied. In this study, raw sequences of NFM products of Sikkim and Arunachal Pradesh were accessed from NCBI database server. PICRUSt2 and Piphillin tools were applied to study microbial functional gene prediction in combination with the commonly used error-corrected denoising programs like DADA2 and Deblur.

Results: Significant functional hits were observed from the Piphillin analysis which included some interesting pathways including GABAergic synapse, glutamatergic synapse and serotonergic synapse, which are known to be probiotic-related, among others that are absent in PICRUSt2 analysis. This study also showed the negative correlation of lactic acid bacteria (LAB) members (*Lactococcus*, *Lactobacillus*, *Leuconostoc*) with most of the disease-related functions, which were on the other hand, positively correlated with unwanted contaminants like *Staphylococcus*, *Bacillus* and *Pseudomonas*.

Conclusion: The study explored the potential of microbial functional gene prediction using PICRUSt2 and Piphillin software, and indicated the significance of the presence of LAB in these NFM products of India. Since most LAB members are known to be potential health-promoting bacteria, their negative correlation to many of the disease-related functions also indicates their role in combatting unwanted potential contaminants.

Background

Microbial functional gene prediction has been popularly studied using metagenomic gene prediction tools in recent years [1]. Recently, PICRUSt2 (Phylogenetic Investigation of Communities by Reconstruction of Unobserved States version 2) has been improved for metagenomic gene prediction overcoming some of the limitations of its predecessor, PICRUSt and other similar pipelines [2]. Piphillin, another alternative tool, has also been improved to race with PICRUSt2 in gene functional prediction analysis using 16S rRNA marker gene [3]. PICRUSt and Piphillin software have been applied in studying functional gene prediction in fermented milk products [4, 5, 6, 7]. Introduction of error-corrected clustering methods such as Divisive Amplicon Denoising Algorithm (DADA2) [8] and Deblur [9] have also replaced the conventional clustering methods that define the traditional operational taxonomic units (OTUs). The outputs from DADA2 denoising analysis are known as amplicon sequence variants (ASVs) by their developers, whereas outputs from Deblur analysis are termed as sub-OTUs, and these are sometimes collectively known as exact sequence variants (ESVs) [10]. These two denoiser programs have been widely accepted in next-generation amplicon-sequencing analysis and are also implemented as alternatives to each other in the QIIME2 analysis pipelines [11].

Targeted gene amplicon sequencing has revolutionized the field of food microbial ecology [12, 13, 14, 15, 16]. High-throughput sequencing methods have been also widely used in many naturally fermented products of North-East India [17, 18, 19, 20]. Naturally fermented milk (NFM) products are popular food items in daily diets of ethnic people of Arunachal Pradesh and Sikkim in India which includes *dahi*, *mohi*, *gheu*, soft-*chhurpi*, hard-*chhurpi*, *dudh-chhurpi*, *chhu*, *somar*, *maa*, *philu*, *shyow*, *mar*, *chhurpi/ churapi*, *churkam* and *churtang/ chhurpupu* [21]. The 16S rRNA high-throughput sequence analysis of NFM products of Arunachal Pradesh and Sikkim revealed the dominance of phylum Firmicutes with predominated species of lactic acid bacteria *Lactococcus lactis* (19.7%) and *Lactobacillus helveticus* (9.6%) and *Leuconostoc mesenteroides* (4.5%) and acetic acid bacteria: *Acetobacter lovaniensis* (5.8%), *Acetobacter pasteurianus* (5.7%), *Gluconobacter oxydans* (5.3%), and *Acetobacter* syzygii (4.8%) [19]. Microbiota present in naturally fermented milk products harbour probiotic properties and impart several heath-promoting benefits to consumers [22, 23, 24]. Moreover, predictive gene functionality in NFM products of India has not been analysed yet. Hence, the present study is aimed to predict the microbial functional content of 16S rRNA gene sequencing data of NFM products of India previously analysed by high-throughput sequencing method [19], using PICRUSt2 and Piphillin software in addition to the commonly used denoisers i.e., DADA2 and Deblur that are implemented in QIIME2 environment analysis pipelines. We also compared the inferences tools PICRUSt2 and Piphillin in this study for functional metagenome contents in NFM products.

Methods

Pre-analysis prior to predictive functionality analysis

Raw sequences of naturally fermented milk products of Arunachal Pradesh and Sikkim analysed by high-throughput sequence method [19] were accessed from NCBI database server and were used in this study. Raw reads were processed using QIIME2-2020.6 (https://docs.qiime2.org/2020.6/) [11]. The analyses were divided into different categories based on the denoising methods and the bacterial databases used for taxonomic assignment. The paired-end reads were first paired using PEAR (Paired-End reAd mergeR, https://cme.h-its.org/exelixis/web/software/pear/) program [25] for Deblur analyis, prior to import of reads into QIIME2 environment. Denoising was performed by using both alternative denoising programs DADA2 [8] via qiime dada2 denoise-paired plugin and Deblur [9] via q2-Deblur denoise-16S plugin. Quality-filtered sequences were clustered against Greengenes v13_8 [26] and SILVA v132 [27] databases and followed by taxonomic assignment using q2-vsearch-cluster-features-closed-reference [28].

Predictive Functionality Analysis

(a) ICRUSt2 analysis (https://github.com/picrust/picrust2/wiki):

Quality-filtered clustered sequences were feed into PICRUSt2 algorithm [2], using via q2-vsearch-cluster-features-closed-reference [28]. PICRUSt2 deduced the predictive functionality of the marker gene by using a standard integrated genomes databases. Firstly, multiple assignment of the exact sequence variants (ESVs) was done using HMMER (http://www.hmmer.org/). Placements of ESVs in the reference tree with

evolutionary placement-ng (EPA-ng) algorithm [29] and Genesis Applications for Phylogenetic Placement Analyses (GAPPA) omics [30] were applied. Prediction of gene families was run using a default castor R package [31] with the default algorthim run (maximum parsimony). Metagenome prediction was run using metagenome_pipeline.py [32] and the output features were compared to KEGG (Kyoto Encyclopaedia of Genes and Genomes) database for systematic analysis of gene functions [33].

(b) Piphillin analysis (https://piphillin.secondgenome.com/):

Alternatively, predictive functionality was also inferred using Piphillin [3], a web-server analysis pipeline. Clustered representative sequences (.fasta) and clustered abundance frequency table (.csv) denoised by both DADA2 and Deblur outputs were used in the analysis. The workplan used in this study is represented in Fig. 1.

Statistical Analysis And Data Visualization

Statistical analysis for significant features (pathways) was carried out using STAMP [34] and ANOVA was used for testing of significance. Two sample groups comparison was also computed using two-sided Welch's t-test in STAMP. Relative abundance representation of the top 45 predicted pathways of both PICRUSt2 and Piphillin-generated pathways was plotted using MSEXCEL v365. Spearman's correlation of the bacteria and functionality was analyzed through Statistical Package for the Social Sciences (SPSS) v20.

Results

Reliability of the metagenome predictions for PICRUSt2 analysis was based on the Nearest Sequenced Taxon Index (NSTI). Average of 0.083 ± 0.07 NSTI scores observed in NFM products were analyzed. We acquired about 1109 error-corrected ESVs from DADA2 method which was almost twice the number resulted from deblur method *i.e.*, 650. DADA2-clustered reads with SILVA database resulted in a significant number of total ASVs (*i.e.*, 268), whereas about 201 ASVs were resulted from clustering with Greengenes database. On the other hand, a total of 231 sOTUs resulted from Deblur method clustered with SILVA database and 188 sOTUs were acquired after clustered with Greengenes.

Predictive Pathways Inferred By Picrust2 And Piphillin

We observed the difference between PICRUSt2 algorithm and Piphillin pipeline only in the predictive analysis results, where denoising (DADA2/Deblur) and clustering (Greengenes/SILVA) steps did not have significant effect in the overall predictive gene function findings. The PICRUSt2 algorithm predicted about 151–152 pathways, whereas Piphillin predicted about 265–270 pathways (Table 1). Details of different predictive metabolic pathways acquired from PICRUSt2 and Piphillin analysis are described in Supplementary Table 1. Interestingly among others, Piphillin analysis showed some important pathways that included GABAergic synapse, dopaminergic synapse, glutamatergic synapse, serotonergic synapse which are absent in PICRUSt2 analysis (Supplementary Table 1).

Table 1
Total number of predictive metabolic pathways analyzed by PICRUSt2 and Piphillin.

Methods	Predictive metabolic pathways
DADA2-Greengenes-PICRUSt2	151
DADA2-SILVA-PICRUSt2	152
Deblur-Greengenes-PICRUSt2	151
Deblur- SILVA-PICRUSt2	151
DADA2-SILVA-Piphillin	265
DADA2-Greengenes-Piphillin	265
Deblur-SILVA-Piphillin	270
Deblur-Greengenes-Piphillin	270

Note: The methods are combined based on the denoisers, reference database (used for clustering) and the predictive functionality analysis tools used.

Categorical Distribution Of The Overall Predictive Pathways

We observed a high active metabolic pathways harboured by the NFM products. PICRUSt2-predictive categories involved metabolism (78.11% – 79.96%), genetic information processing (11.8% – 13.59%), cellular processes (4.4% – 5.77%), environmental information processing (2.39% – 3.29%), human diseases (0.25% – 0.43%), and organismal systems (0.18% – 0.28%) (Fig. 2a). Piphillin analysis showed a similar pattern of predictive categories which included metabolism (73.57%-73.86%), genetic information processing (11.62%-12.13%), environmental information processing (6.7%-7.21%), cellular processes (3.17%-3.39%), human diseases (2.97%-2.99%), and organismal systems (1.15%-1.25%) (Fig. 2b).

Correlation Of Predominant Bacterial Genera With Predictive Pathways

As already mentioned, the highest number of ASVs retained was observed in DADA2-denoised reads clustered with SILVA databases, which also resulted in a significant number of predictive pathways (Table 1). Top 45 predictive pathways of both PICRUSt2 and Piphillin predictive pathways are represented in Supplementary Figs. 1 and 2. DADA2-SILVA combination outputs were used for further hypothesis testing of significant pathways. Significant PICRUSt2-predictive pathways compared among different test groups are shown in Supplementary Tables 2,3,4 and 5 and that of Piphillin-significant predictive pathways are shown in Supplementary Tables 6 and 7. A total of 104 significant pathways from PICRUSt2-predictive analysis were recorded; and 81 predicted pathways were found to be significant through Piphillin analysis using STAMP. Spearman's rank correlation analysed through SPSS showed a significant correlation of predominant bacterial genera with significant pathways. Overall, complex positive and negative correlations were observed among the predominant bacteria with the significant

pathways (Supplementary Tables 8 and 9). However, most of the pathways were related to common metabolic pathways.

We observed a negative correlation of *Lactococcus* (most dominant genus in NFM products of Arunachal Pradesh and Sikkim) with some of the disease-related pathways including biofilm formation - *Vibrio cholerae*, Chagas disease, pathogenic *Escherichia coli* infection, pathways in cancer and *Staphylococcus aureus* infection by PICTRUSt2 analysis (Fig. 3). *Lactobacillus* and *Leuconostoc* were also negatively correlated to some of the aforementioned pathways (Fig. 3). On the other hand, we observed a positive significant correlation of primary bile acid biosynthesis with *Lactobacillus*, *Leuconostoc*, *Staphylococcus*, *Bacillus* and *Enterococcus* (Fig. 3). Contrastingly, *Pseudomonas* and *Staphylococcus* showed a positive correlation to some disease-related pathways: Chagas disease. By Piphillin-analysis, we observed a significant positive correlation of African trypanosomiasis, Alzheimer disease, Chagas disease, Legionellosis, Parkinson disease, Pertussis with *Acinetobacter* and *Pseudomonas* with pathways in cancer (Fig. 4). Similarly, we observed a negative correlation of *Lactococcus* (most dominant genera) with some of the disease-related pathways including African trypanosomiasis, Alzheimer disease, biofilm formation - *Pseudomonas aeruginosa*, Chagas disease, chemical carcinogenesis, choline metabolism in cancer, Parkinson disease, pathways in cancer, pertussis, hepatocellular carcinoma etc. *Lactobacillus* and *Leuconostoc* also showed negative correlation on some of the disease-related pathways (Fig. 4). A significant positive correlation of *Lactobacillus* with glutamatergic synapse was observed. *Staphylococcus*, *Bacillus* and *Pseudomonas* showed predictive regulation on almost all the disease-related pathways (Fig. 4).

Discussion

It was observed that deblur-denoised reads were lesser compared to that of DADA2-denoised reads, since deblur requires the length of the sequences should be of equal lengths and paired-end reads be merged before the denoising process [9]. Short sequences below the desired cut-off length were automatically discarded during deblur analysis, whereas raw reads of DADA2 were directly merged, denoised with chimera-free variable length ASVs. However, we could not find significant differences between DADA2 or deblur-denoised reads used for both PICRUSt2 and Piphillin in this study. As per the predictive functionality analysis, there were significant numbers of Piphillin-predictive pathways in respect to that of PICRUSt2. The Piphillin pipeline outperforms PICRUSt2 with 19% greater balanced accuracy and 54% greater precision [3]. This difference may be due to the methods implemented by both the algorithms. PICRUSt2 uses the hidden state prediction to infer genomic content based on genome position in a reference phylogenetic tree [2]. Piphillin, on the other hand, predicts metagenomic content via nearest-neighbour matching between 16S rRNA gene amplicons and genomes from reference databases (i.e., KEGG or BioCyc) [3]. Till date, there is no specific metagenome database for NFM products for such analysis. Moreover, 16S rRNA marker gene serves as a structural gene, functional inferences from such genes are mere prediction based on the algorithm and the databases used. Hence, the combination of different omics technologies in food-related microbiome studies is mandatory for species/strain resolution of microbial diversity and functional characterization of complex community structures [35].

PICRUSt2 analysis also heavily relies on the Nearest Sequenced Taxon Index (NSTI) score, an evaluation measure describing the novelty of organism within an OTU table with respect to previously sequenced genomes [36], for accurate metagenome prediction of which 2 is the highest value [2]. We also observed a very low NSTI scores for NFM products indicating low accuracy of metagenomic gene prediction which also suggests that there remains a large number of metagenomes to be sequenced and classified.

Predictive gene functionality analysis of NFM products as inferred by PICRUSt2 and Piphillin showed a high metabolism rate as expected since most of these products are consortia of many metabolically active microbiota. These findings are also similar to recent studies reported from milk and dairy products [4, 5, 6, 7]. Interestingly, few signal pathways with well-known probiotic functional features were also inferred from Piphillin-predictive pathways that included GABAergic synapse, glutamatergic synapse and serotonergic synapse, which were absent in PICRUSt2 analysis. These predictive functional hits were related to cognitive function. Probiotics are well-known for enhancing brain function and have been demonstrated to have various mechanisms in alleviating depression and enhancing brain function [37, 38]. *Lactobacillus* is one of the main LAB genera reported from NFM products of India (Shangpliang et al. 2018) and we also observed a significant positive correlation of this genus with glutamatergic synapse as per our findings. This suggests the presence of *Lactobacillus* to be in tandem with this feature. As reported earlier [19], these NFM products of India are dominated by lactic acid bacterial groups, which in general are considered as beneficial microorganisms [39]. Naturally fermented milk products are known to a good source of health-promoting probiotic bacteria [23, 40]. It is important to note that predictive gene functionality also suggests the importance of LAB members in these products.

We observed the negative correlation of LAB groups– *Lactococcus*, *Lactobacillus*, *Leuconostoc* on many of the disease-related pathways, like biofilm formation - *Vibrio cholerae*, Chagas disease, pathogenic *Escherichia coli* infection, pathways in cancer and *Staphylococcus aureus* infection, African trypanosomiasis, Alzheimer disease, biofilm formation - *Pseudomonas aeruginosa*, chemical carcinogenesis, choline metabolism in cancer, Parkinson disease, pathways in cancer, pertussis, and hepatocellular carcinoma. These findings suggest that the presence of LAB has a negative impact on the disease-related functions, perhaps, controlling the population of those disease-causing bacteria like *Staphylococcus*, *Bacillus* and *Pseudomonas*, which are commonly present in contaminated raw milk and fermented milk products [41, 42, 43]. LAB isolated from fermented foods are known to be health-promoting bacteria acting against many pathogenic bacteria [44, 45, 46]. Based on our present findings on predictive functionality, correlation studies highly suggest the significance of LAB present in NFM products of India to be a good source of potential probiotic bacteria and health-promoting agents.

Conclusion

Our present study explores the potential of microbial functional gene prediction using PICRUSt2 and Piphillin pipelines, and indicates the significance of the presence of LAB in these naturally fermented milk products of India. Since most LAB members are known to be potential health-promoting bacteria, their negative correlation to many of the disease-related functions also indicates their role in combatting

unwanted potential contaminants. This present study also provides valuable insight into the potential of LAB-predominated products in many health-promoting applications. Additionally, more advanced methods are yet to be applied for in-depth functional profiling and safety evaluation.

List Of Abbreviations

NFM: Naturally Fermented Milk

rRNA: Ribosomal ribonucleic acid

PICRUSt: Phylogenetic investigation of communities by reconstruction of unobserved states

LAB: Lactic acid bacteria

QIIME: Quantitative Insights Into Microbial Ecology

DADA: Divisive Amplicon Denoising Algorithm

OTUs: Operational taxonomic units

ASVs: Amplicon sequence variants

ESVs: Exact sequence variants

sOTUs: sub-Operational taxonomic units

NGS: Next-Generation Sequencing

PEAR: Paired-End reAd mergeR

HMM: Hidden Markov models

GAPPA: Genesis Applications for Phylogenetic Placement Analyses

EPA-ng: Evolutionary placement-ng algorithm

KEGG: Kyoto Encyclopedia of Genes and Genomes

STAMP: STatistical Analysis of Metagenomic Profiles

ANOVA: Analysis of variance

SPSS: Statistical Package for the Social Sciences

NSTI: Nearest Sequenced Taxon Index

Declarations

Authors' contributions

HNJS and RR did all the analysis and experiments. JPT has supervised the bioinformatics analysis and finalised the manuscript.

Funding

This current research is supported by Department of Biotechnology, Govt. of India through DBT-AIST International Centre for Translational and Environmental Research (DAICENTRE) project.

Availability of data and materials

In this present study, analysis was performed on previously published data [19]. Raw sequences were accessed from MG-RAST server having the MG-RAST ID number 4732361 to 4732414. The same were accessed from NCBI database server under the BioProject No. PRJNA661385 with accession numbers SAMN16056817 to SAMN16056870.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Authors' details

[1]DAICENTRE (DBT-AIST International Centre for Translational and Environmental Research) and Bioinformatics Centre, Department of Microbiology, School of Life Sciences, Sikkim University, Gangtok 737102, Sikkim, India.

References

1. Ortiz-Estrada ÁM, Gollas-Galván T, Martínez-Córdova LR, Martínez-Porchas M. Predictive functional profiles using metagenomic 16S rRNA data: a novel approach to understanding the microbial ecology of aquaculture systems. Reviews in Aquaculture. 2019;11(1):234-45. https://doi.org/10.1111/raq.12237

2. Douglas GM, Maffei VJ, Zaneveld JR, Yurgel SN, Brown JR, Taylor CM, Huttenhower C, Langille MG. PICRUSt2 for prediction of metagenome functions. Nature Biotechnology. 2020;1:1-5. https://doi.org/10.1038/s41587-020-0548-6
3. Narayan NR, Weinmaier T, Laserna-Mendieta EJ, Claesson MJ, Shanahan F, Dabbagh K, Iwai S, DeSantis TZ. Piphillin predicts metagenomic composition and dynamics from DADA2-corrected 16S rDNA sequences. BMC Genomics. 2020;1:1-2. https://doi.org/10.1186/s12864-019-6427-1
4. Zhang F, Wang Z, Lei F, Wang B, Jiang S, Peng Q, Zhang J, Shao Y. Bacterial diversity in goat milk from the Guanzhong area of China. Journal of Dairy Science. 2017;100(10):7812-24. https://doi.org/10.3168/jds.2017-13244
5. Zhu Y, Cao Y, Yang M, Wen P, Cao L, Ma J, Zhang Z, Zhang W. Bacterial diversity and community in Qula from the Qinghai–Tibetan Plateau in China. PeerJ. 2018;6:e6044. https://doi.org/10.7717/peerj.6044
6. Chen X, Zheng R, Liu R, Li L. Goat milk fermented by lactic acid bacteria modulates small intestinal microbiota and immune responses. Journal of Functional Foods. 2020;65:103744. https://doi.org/10.1016/j.jff.2019.103744
7. Choi J, Lee SI, Rackerby B, Goddik L, Frojen R, Ha SD, Kim JH, Park SH. Microbial communities of a variety of cheeses and comparison between core and rind region of cheeses. Journal of Dairy Science. 2020;103(5):4026-42. https://doi.org/10.3168/jds.2019-17455
8. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. DADA2: high-resolution sample inference from Illumina amplicon data. Nature Methods. 2016;13(7):581-3. https://doi.org/10.1038/nmeth.3869
9. Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Xu ZZ, Kightley EP, Thompson LR, Hyde ER, Gonzalez A, Knight R. Deblur rapidly resolves single-nucleotide community sequence patterns. MSystems. 2017;21:2(2). https://doi.org/10.1128/mSystems.00191-16
10. Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. The ISME Journal. 2017;11(12):2639-43. https://doi.org/10.1038/ismej.2017.119
11. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature Biotechnology. 2019;37(8):852-7. https://doi.org/10.1038/s41587-019-0209-9
12. Bybee SM, Bracken-Grissom H, Haynes BD, Hermansen RA, Byers RL, Clement MJ, Udall JA, Wilcox ER, Crandall KA. Targeted amplicon sequencing (TAS): a scalable next-gen approach to multilocus, multitaxa phylogenetics. Genome Biology and Evolution. 2011;3:1312-23. https://doi.org/10.1093/gbe/evr106
13. Ercolini D. High-throughput sequencing and metagenomics: moving forward in the culture-independent analysis of food microbial ecology. Applied and Environmental Microbiology. 2013;79(10):3148-55. https://doi.org/10.1128/AEM.00256-13

14. Mayo B, TCC Rachid C, Alegría Á, MO Leite A, S Peixoto R, Delgado S. Impact of next generation sequencing techniques in food microbiology. Current Genomics. 2014;15(4):293-309. http://dx.doi.org/10.2174/1389202915666140616233211
15. De Filippis F, Parente E, Ercolini D. Metagenomics insights into food fermentations. Microbial Biotechnology. 2017;1:91-102. https://doi.org/10.1111/1751-7915.12421
16. De Filippis F, Parente E, Ercolini D. Recent past, present, and future of the food microbiome. Annual Review of Food Science and Technology. 2018; 9:589-608. https://doi.org/10.1146/annurev-food-030117-012312
17. Romi W, Ahmed G, Jeyaram K. Three-phase succession of autochthonous lactic acid bacteria to reach a stable ecosystem within 7 days of natural bamboo shoot fermentation as revealed by different molecular approaches. Molecular Ecology. 2015;13:3372-89. https://doi.org/10.1111/mec.13237
18. Keisam S, Romi W, Ahmed G, Jeyaram K. Quantifying the biases in metagenome mining for realistic assessment of microbial ecology of naturally fermented foods. Scientific Reports. 2016;6(1):1-2. https://doi.org/10.1038/srep34155
19. Shangpliang HNJ, Rai R, Keisam S, Jeyaram K, Tamang JP. Bacterial community in naturally fermented milk products of Arunachal Pradesh and Sikkim of India analysed by high-throughput amplicon sequencing. Scientific Reports. 2018;8(1):1532. https://doi.org/10.1038/s41598-018-19524-6
20. Sha SP, Suryavanshi MV, Tamang JP. Mycobiome diversity in traditionally prepared starters for alcoholic beverages in India by high-throughput sequencing method. Frontiers in Microbiology. 2019;10:348. https://doi.org/10.3389/fmicb.2019.00348
21. Rai R, Shangpliang HN, Tamang JP. Naturally fermented milk products of the Eastern Himalayas. Journal of Ethnic Foods. 2016;3(4):270-275. https://doi.org/10.1016/j.jef.2016.11.006
22. Bengoa AA, Iraporda C, Garrote GL, Abraham AG. Kefir micro-organisms: their role in grain assembly and health properties of fermented milk. Journal of Applied Microbiology. 2019;126(3):686-700. https://doi.org/10.1111/jam.14107
23. Tamang JP, Cotter PD, Endo A, Han NS, Kort R, Liu SQ, Mayo B, Westerik N, Hutkins R. Fermented foods in a global age: East meets West. Comprehensive Reviews in Food Science and Food Safety. 2020;1:184-217. https://doi.org/10.1111/1541-4337.12520
24. García-Burgos M, Moreno-Fernández J, Alférez MJ, Díaz-Castro J, López-Aliaga I. New perspectives in fermented dairy products and their health relevance. Journal of Functional Foods. 2020;72:104059. https://doi.org/10.1016/j.jff.2020.104059
25. Zhang J, Kobert K, Flouri T, Stamatakis A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics. 2014;30(5):614-20. https://doi.org/10.1093/bioinformatics/btt593
26. McDonald D, Price MN, Goodrich J, Nawrocki EP, DeSantis TZ, Probst A, Andersen GL, Knight R, Hugenholtz P. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary

analyses of bacteria and archaea. The ISME Journal. 2012;6(3):610-8. https://doi.org/10.1038/ismej.2011.139

27. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research. 2012;41(D1):D590-6. https://doi.org/10.1093/nar/gks1219
28. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016;4:e2584. https://doi.org/10.7717/peerj.2584
29. Barbera P, Kozlov AM, Czech L, Morel B, Darriba D, Flouri T, Stamatakis A. EPA-ng: massively parallel evolutionary placement of genetic sequences. Systematic Biology. 2019;68(2):365-9. https://doi.org/10.1093/sysbio/syy054
30. Czech L, Stamatakis A. Scalable methods for analyzing and visualizing phylogenetic placement of metagenomic samples. PLoS One. 2019;14(5):e0217050. https://doi.org/10.1371/journal.pone.0217050
31. Louca S, Doebeli M. Efficient comparative phylogenetics on large trees. Bioinformatics. 2018;34(6):1053-5. https://doi.org/10.1093/bioinformatics/btx701
32. Ye Y, Doak TG. A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. PLoS Computational Biology. 2009;5(8):e1000465. https://doi.org/10.1371/journal.pcbi.1000465
33. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Research. 2012;40(D1):D109-14. https://doi.org/10.1093/nar/gkr988
34. Parks DH, Tyson GW, Hugenholtz P, Beiko RG. STAMP: statistical analysis of taxonomic and functional profiles. Bioinformatics. 2014; 30(21): 3123-4. https://doi.org/10.1093/bioinformatics/btu494
35. Cao Y, Fanning S, Proos S, Jordan K, Srikumar S. A review on the applications of next generation sequencing technologies as applied to food-related microbiome studies. Frontiers in Microbiology. 2017;8:1829. https://doi.org/10.3389/fmicb.2017.01829
36. Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, *et al*. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nature Biotechnol 2013; 31: 814-821.
37. Mohajeri MH, La Fata G, Steinert RE, Weber P. Relationship between the gut microbiome and brain function. Nutrition Reviews. 2018;76(7):481-96. https://doi.org/10.1093/nutrit/nuy009
38. Yong SJ, Tong T, Chew J, Lim WL. Antidepressive Mechanisms of Probiotics and Their Therapeutic Potential. Frontiers in Neuroscience. 2020;13:1361. https://doi.org/10.3389/fnins.2019.01361
39. Mathur H, Beresford TP, Cotter PD (2020) Health benefits of lactic acid bacteria (LAB) fermentates.Nutrients 12(6):1679. doi:10.3390/nu12061679
40. Rezac S, Kok CR, Heermann M, Hutkins R. Fermented foods as a dietary source of live organisms. Frontiers in Microbiology. 2018;9:1785. https://doi.org/10.3389/fmicb.2018.01785

41. Verraes C, Vlaemynck G, Van Weyenberg S, De Zutter L, Daube G, Sindic M, Uyttendaele M, Herman L. A review of the microbiological hazards of dairy products made from raw milk. International Dairy Journal. 2015;50:32-44. https://doi.org/10.1016/j.idairyj.2015.05.011
42. Lan X, Wang J, Zheng N, Zhao S, Li S, Li F. Prevalence and risk factors for *Bacillus cereus* in raw milk in Inner Mongolia, Northern China. International Journal of Dairy Technology. 2018;71(1):269-73. https://doi.org/10.1111/1471-0307.12416.
43. Reichler SJ, Trmčić A, Martin NH, Boor KJ, Wiedmann M. *Pseudomonas fluorescens* group bacterial strains are responsible for repeat and sporadic post-pasteurization contamination and reduced fluid milk shelf life. Journal of Dairy Science. 2018;101(9):7780-800. https://doi.org/10.3168/jds.2018-14438
44. Kumar H, Franzetti L, Kaushal A, Kumar D. *Pseudomonas fluorescens*: a potential food spoiler and challenges and advances in its detection. Annals of Microbiology. 2019;1:1-1. https://doi.org/10.1007/s13213-019-01501-7.
45. Karami S, Roayaei M, Hamzavi H, Bahmani M, Hassanzad-Azar H, Leila M, Rafieian-Kopaei M. Isolation and identification of probiotic *Lactobacillus* from local dairy and evaluating their antagonistic effect on pathogens. International Journal of Pharmaceutical Investigation. 2017;7(3):137.doi:10.4103/jphi.JPHI_8_17.
46. Elbanna K, El Hadad S, Assaeedi A, Aldahlawi A, Khider M, Alhebshi A. In vitro and in vivo evidences for innate immune stimulators lactic acid bacterial starters isolated from fermented camel dairy products. Scientific Reports. 2018;8(1):1-5. https://doi.org/10.1038/s41598-018-31006-3.

Figures

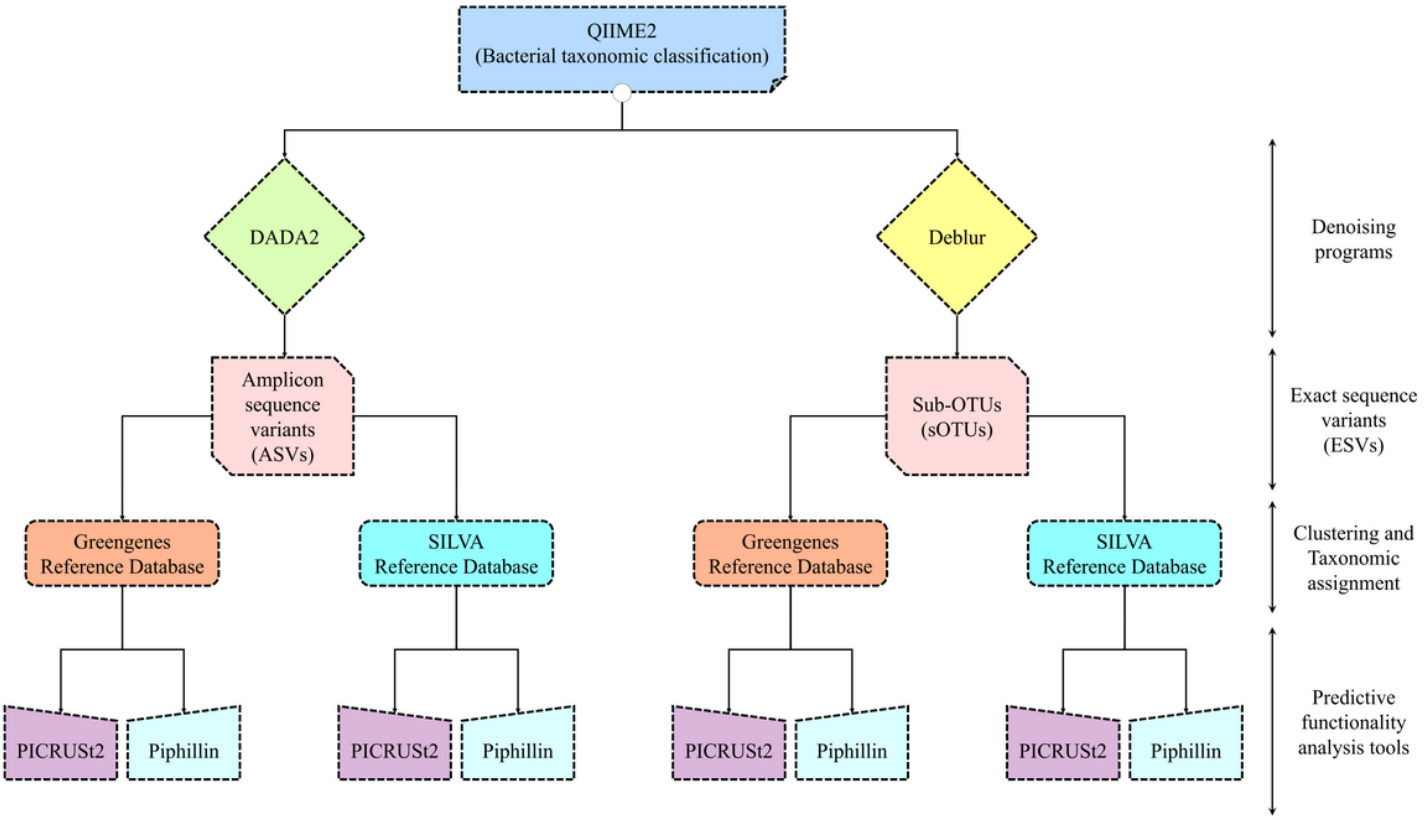

Figure 1

Workplan used in this study. Illumina-based amplicon gene sequencing analysis of naturally fermented milk (NFM) products [19] was performed using QIIME2. Denoising was performed using DADA2 and Deblur and the respective error-corrected exact sequence variants (ASVs/sOTUs) were then clustered using Greengenes and SILVA reference databases of which the taxonomic assignments were also performed. Predictive functionality was inferred using PICRUSt2 and Piphillin.

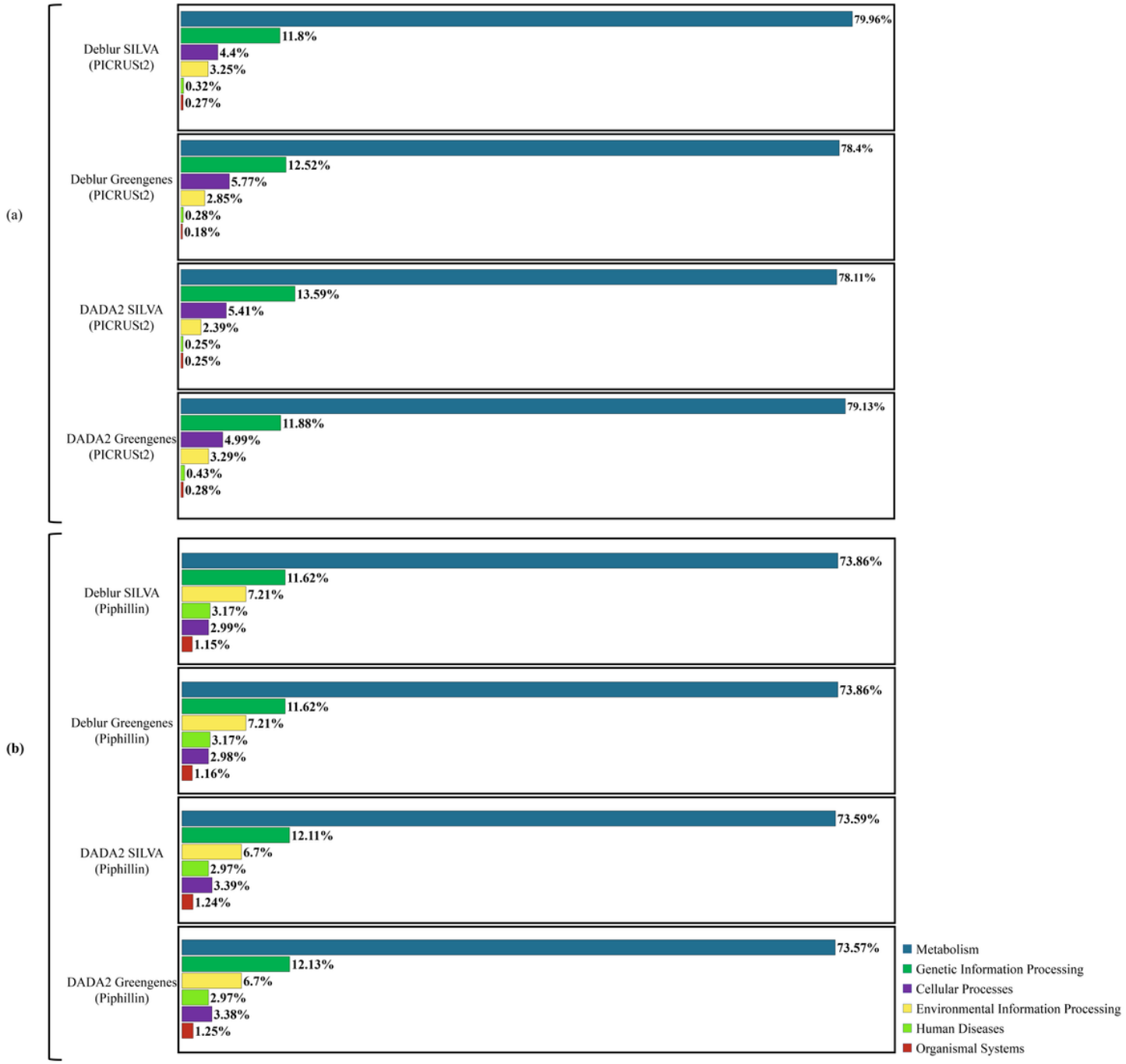

Figure 2

(a) PICRUSt2-predicted metabolic pathways represented at Level-1 (Category). (b) Piphillin-predicted metabolic pathways represented at Level-1 (Category). Overall, there are no significant differences from DADA2 or Deblur denoised reads, nor from clustered by Greengenes or SILVA database, prior to predictive analysis using PICRUSt2/Piphillin. Dominance of predictive features under metabolism category was observed, followed by genetic information processing, cellular processes, environmental information processing, human diseases ad organismal system.

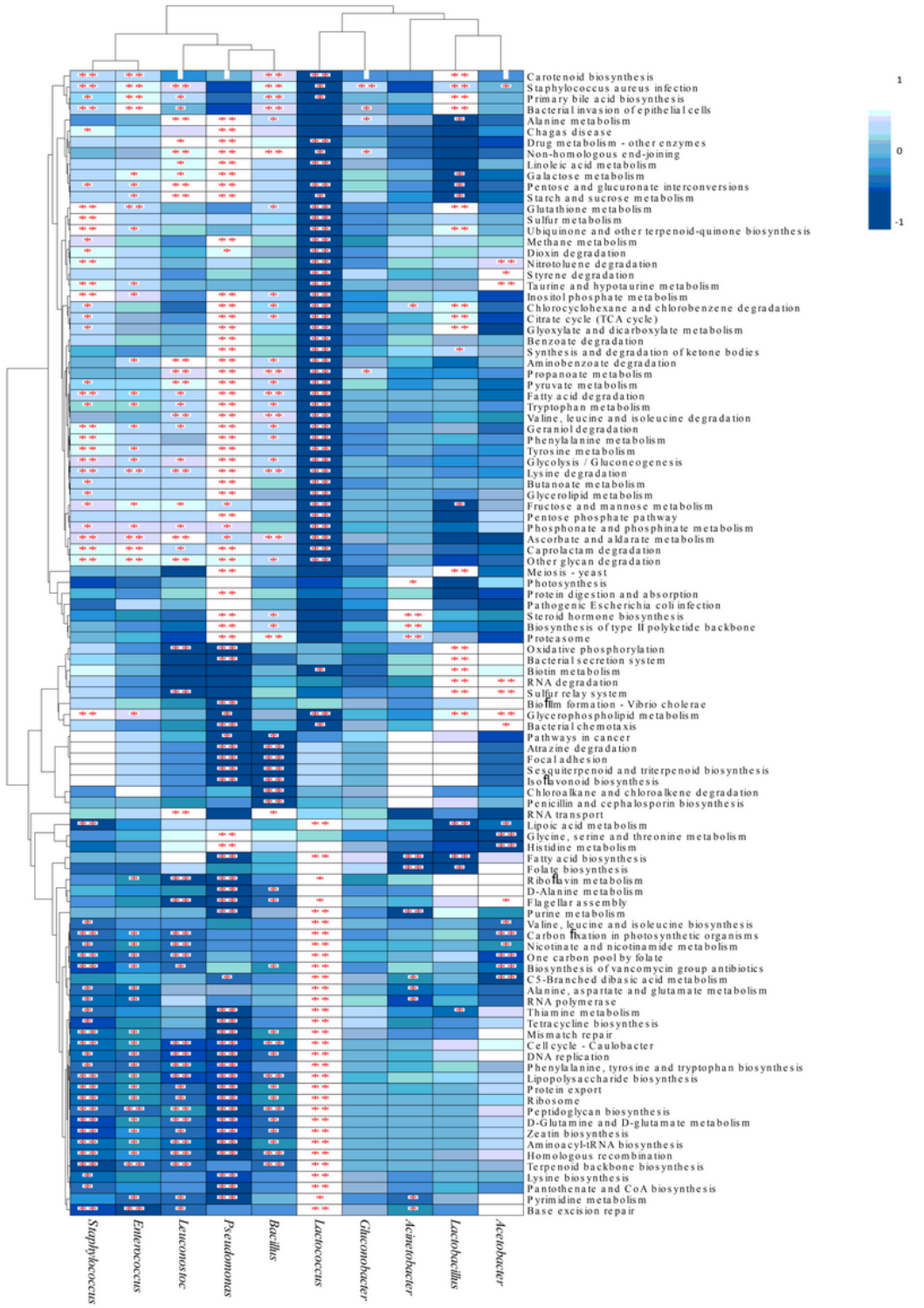

Figure 3

Heatmap representation of Spearman's rank correlation of bacterial genera and the PICRUSt2-predictive significant pathways calculated using Statistical Package for the Social Sciences (SPSS) v20. Negative values denote the negative correlation and positive values denotes the positive correlation of genera and predictive pathways. All significant correlation pairs are denoted by * (*<0.05 and **<0.01).

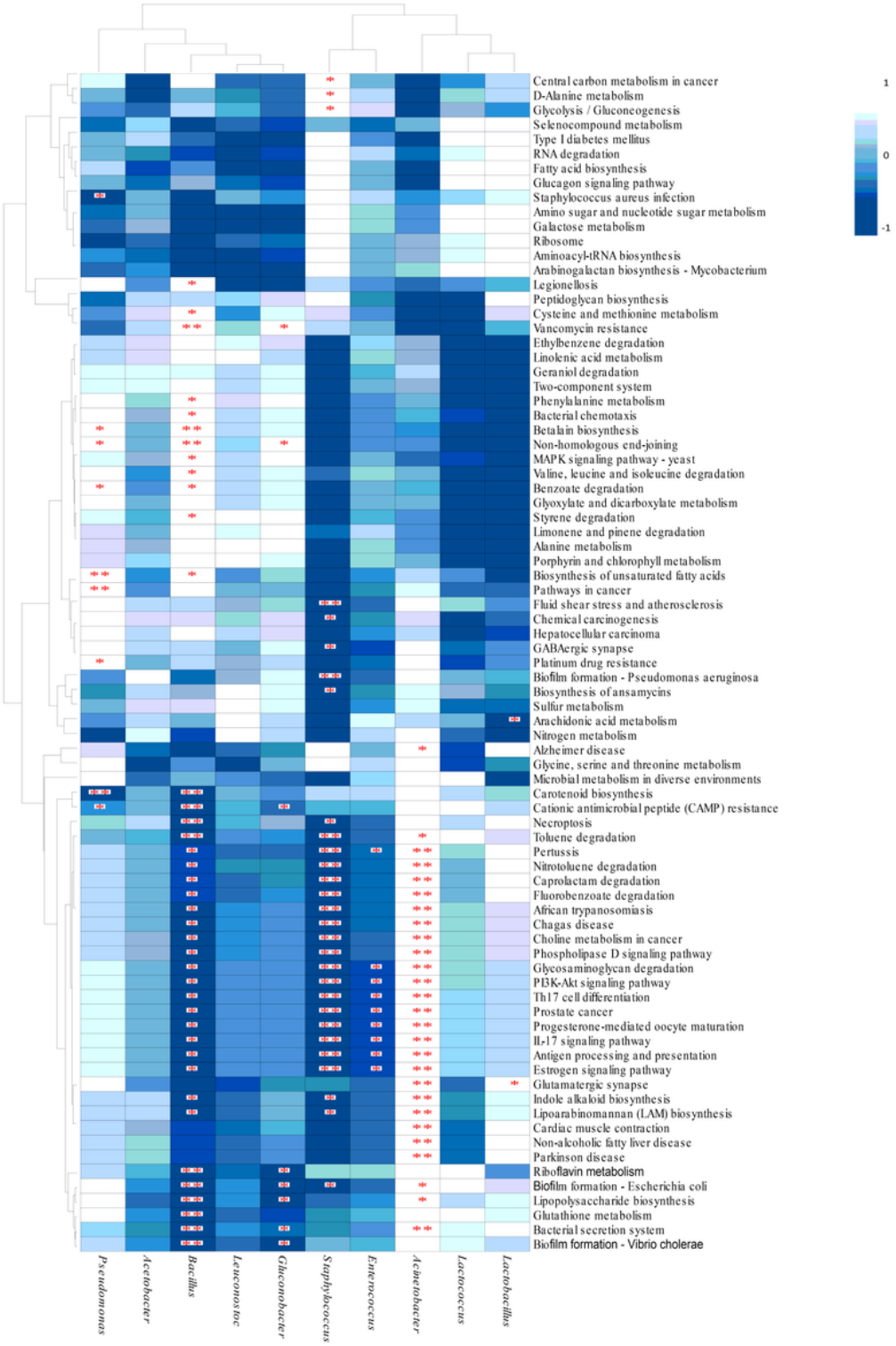

Figure 4

Heatmap representation of Spearman's rank correlation of bacterial genera and the Piphillin-predictive significant pathways calculated using Statistical Package for the Social Sciences (SPSS) v20. Negative values denote the negative correlation and positive values denotes the positive correlation of genera and predictive pathways. All significant correlation pairs are denoted by * (*<0.05 and **<0.01).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTable1.docx
- SupplementaryTable2.docx
- SupplementaryTable3.docx
- SupplementaryTable4.docx
- SupplementaryTable5.docx
- SupplementaryTable6.docx
- SupplementaryTable7.docx
- SupplementaryTable8.docx
- SupplementaryTable9.docx
- SupplementaryFigure.1.tif
- SupplementaryFigure.2.tif