Predictive functionality of bacteria in naturally fermented milk products of India using PICRUSt2 and Piphillin pipelines

H. Nakibapher Jones Shingling and Jyoti Prakash Tamang*

DAICENTER (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

*Corresponding author: Professor Dr. Jyoti Prakash Tamang, Department of Microbiology, Sikkim University, Tadong 737102, Sikkim, India (e-mail: jyoti_tamang@hotmail.com; Mobile: +91-9832061073)

Running Title: Metagenome gene prediction of fermented milk

Abstract

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 MG-RAST/NCBI database server. PICRUSt2 and Piphillin tools were applied to study microbial functional gene prediction. MUSiCC-normalized KOs and mapped KEGG pathways from both PICRUSt2 and Piphillin resulted in higher percentage of the former in comparison to the latter. Though, functional features were compared from both the pipelines, however, there were significant differences between the predictions. Therefore, a consolidated presentation of both the algorithms presented an overall outlook into the predictive functional profiles associated with the microbiota of the NFM products of India.

Keywords**:** Metagenome gene prediction, PICRUSt2, Piphillin**,** naturally fermented milk products, lactic acid bacteria

Introduction

Naturally fermented milk (NFM) products are popular food items in daily diets of ethnic people of Arunachal Pradesh and Sikkim in India, which include *dahi*, *mohi*, *gheu*, soft-*chhurpi*, hard-*chhurpi*, *dudh-chhurpi*, *chhu*, *somar*, *maa*, *philu*, *shyow*, *mar*, *chhurpi*/*churapi*, *churkam* and *churtang*/*chhurpupu* (Rai et al. 2016; Tamang et al. 2021). Previously, taxonomic analysis using high-throughput sequencing (HTS) of NFM products of Arunachal Pradesh and Sikkim viz. *chhurpi*, *churkam mar*/*gheu* and *dahi*, have been studied (Shangpliang et al. 2018). We have recorded the abundance of phylum *Firmicutes* with predominated species of lactic acid bacteria (LAB) viz. *Lactococcus lactis* (19.7%) and *Lactobacillus helveticus* (9.6%) and *Leuconostoc mesenteroides* (4.5%) and acetic acid bacteria (AAB): *Acetobacter lovaniensis* (5.8%), *Acetobacter pasteurianus* (5.7%), *Gluconobacter oxydans* (5.3%), and *Acetobacter* syzygii (4.8%) (Shangpliang et al. 2018). Application of shotgun metagenomics is one of the commonly used methods for understanding the microbial-associated gene functional characteristics (Quince et al. 2017). However, alternately functional profiles of a microbial community can also be inferred indirectly by marker-gene surveys such as 16S rRNA gene (Ortiz-Estrada et al. 2019; Bokulich et al. 2020). Bioinformatics pipelines such as Phylogenetic Investigation of Communities by Reconstruction of Unobserved States version2 (PICRUSt2) (Douglas et al. 2020) and Piphillin (Narayan et al. 2020) among others are some of the well-known tools for microbial predictive functionality studies from various NGS-related metagenomic data (Ortiz-Estrada et al. 2019; Bokulich et al. 2020). These pipelines have also been applied in fermented milk products to infer the functional gene predictions (Zhang et al. 2017; Zhu et al. 2018; Chen et al. 2020; Choi et al. 2020a,b). Microbiota present in NFM products harbour probiotic properties and impart several health-promoting benefits to consumers (Bengoa et al. 2019; Tamang et al. 2020; García-Burgos et al. 2020). Predictive gene functionality in NFM

products of India has not been analysed yet. Hence, the present study is aimed to predict the microbial functional contents of 16S rRNA gene sequencing data of NFM products of India, previously analysed by high-throughput sequencing method (Shangpliang et al. 2018), using PICRUSt2 and Piphillin pipelines.

Material and Methods

Pre-analysis prior to predictive functionality analysis

Raw sequences of NFM products of Arunachal Pradesh and Sikkim in India analysed by HTS method (Supplementary Table 1) were accessed from MG-RAST/NCBI database server and were used in this study. Raw reads were processed using QIIME2-2020.6 (https://docs.qiime2.org/2020.6/) (Bolyen et al. 2019). After importing into QIIME2 environment, Q-score based filtering and denoising was performed using Divisive Amplicon Denoising Algorithm (DADA2) (Callahan et al. 2016) via qiime dada2 denoise-paired plugin. Quality-filtered sequences were then clustered against SILVA v132 (Quast et al. 2012) databases and followed by taxonomic assignment using q2-vsearch-cluster-features-closed-reference (Rognes et al. 2016).

Predictive functionality analysis

PICRUSt2 analysis (https://github.com/picrust/picrust2/wiki)

Quality-filtered clustered sequences were feed into PICRUSt2 algorithm (Douglas et al. 2020) using via q2-vsearch-cluster-features-closed-reference (Rognes et al. 2016). PICRUSt2 deduced the predictive functionality of the marker genes by using a standard integrated genomes database. Firstly, multiple assignment of the exact sequence variants (ESVs) was performed using HMMER (http://www.hmmer.org/). Placements of ESVs in the reference tree with evolutionary placement-ng (EPA-ng) algorithm (Barbera et al. 2019) and Genesis

Applications for Phylogenetic Placement Analyses (GAPPA) omics (Czech and Stamatakis 2019) were applied. Prediction of gene families was run using a default castor R package (Louca and Doebeli 2018) with the default algorthim run (maximum parsimony) and metagenome prediction was acquired using metagenome_pipeline.py (Ye and Doak 2009).

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

Additionally, predictive functionality was also inferred using Piphillin (Narayan et al. 2020), a web-server analysis pipeline. DADA2-clustered representative sequences (.fasta) and abundance frequency table (.csv) were used as inputs for the analysis.

Statistical analysis and data visualization

Unnormalized Kyoto Encyclopaedia of Genes and Genomes (KEGG) ortholog (KO) profiles of PICRUSt2 and Piphillin predictive were normalized using Metagenomic Universal Single-Copy Correction (MUSiCC) (Manor and Borenstein 2015). The output features were then mapped to KEGG database for systematic analysis of gene functions (Kanehisa et al. 2012). Relative abundance at the category level was plotted as stacked bar-plot using MSEXCEL v365. Statistical analysis for significant features (pathways) was carried out using STAMP (Parks et al. 2014). Normalized predictive features were log-transformed and the differences between PICRUSt2 and Piphillin predictive features were calculated using White's non-parametric with Benjamini-Hochberg FDR (false discovery rate) (Parks et al. 2014). Non-parametric Spearman's correlation of the bacteria and functionality was analyzed through Statistical Package for the Social Sciences (SPSS) v20 and the heatmap representation was plotted using ClustVis (Metsalu and Vilo 2015).

Results

Microbial predictive gene functionality

A total of 1109 error-corrected ESVs was obtained from DADA2 analysis and about 268 SILVA-clustered sequences were used for the downstream predictive analysis. A total of 5995 MUSiCC-normalized KOs and 181 mapped KEGG pathways was obtained from PICRUSt2 analysis. Similarly, a total of 5245 MUSiCC-normalized KOs and 157 mapped KEGG pathways was obtained from Piphillin analysis. Overall, both PICRUSt2 and Piphillin pipelines showed a similar pattern (Fig. 1), except in the metabolism category where the PICRUSt2 was significantly higher in comparison to that predicted by Piphillin pipeline (Fig. 2). Additionally, at the super pathway level, PICRUSt2 prediction showed significantly high in amino acid metabolism, metabolism of cofactors and vitamins, energy metabolism, and biosynthesis of other secondary metabolites (Fig. 2). On the other hand, predictive super pathways which included carbohydrate metabolism, xenobiotics biodegradation and metabolism, metabolism of other amino acids, lipid metabolism, metabolism of terpenoids and polyketides, glycan biosynthesis and metabolism, and nucleotide metabolism were significantly higher through Piphillin prediction (Fig. 2). Significant metabolic-related pathways inferred by both PICRUSt2 and Piphillin tools were compared showing several functional features predicted by these two pipelines (Fig. 3).

Non-parametric correlation of bacteria with predictive functionality

Non-parametric Spearman's correlation analysis resulted in a complex bacterial-functions interaction. *Lactococcus* showed a significant negative correlation with glycerolipid metabolism and ubiquinone and other terpenoid-quinone biosynthesis. *Lactobacillus* showed significant negative correlation with tryptophan metabolism, galactose metabolism, and lipoic acid metabolism while it was observed to be positively significantly correlated with sulphur metabolism. On the other hand, valine, leucine and isoleucine degradation, arginine

biosynthesis and ubiquinone and other terpenoid-quinone biosynthesis was positively correlated with *Leuconostoc*, and negatively correlated with galactose metabolism. Furthermore, a significant negative correlation was observed between *Acetobacter* with pathways- tryptophan metabolism, valine, leucine and isoleucine biosynthesis, and lipoic acid metabolism. *Gluconobacter* also showed a significant negative correlation with phenylalanine metabolism, pentose and glucuronate interconversions, fructose and mannose metabolism, and nitrogen metabolism. Glycerolipid metabolism and ubiquinone and other terpenoid-quinone biosynthesis showed significant positive correlation with *Staphylococcus*, which significantly negatively correlated with propanoate metabolism. *Pseudomonas* showed significant negative correlation with fructose and mannose metabolism and significant positive correlation with tyrosine metabolism, valine, leucine and isoleucine degradation, arginine and proline metabolism, galactose metabolism, ubiquinone and other terpenoid-quinone biosynthesis and glutathione metabolism. Additionally, a significant positive correlation was observed between *Acinetobacter* with phenylalanine metabolism, streptomycin biosynthesis, ascorbate and aldarate metabolism, propanoate metabolism, nitrogen metabolism, and biosynthesis of ansamycins (Fig. 4).

Discussion

In this study, microbial predictive gene functional analysis from targetted-16S rRNA gene was explored using PICRUSt2 and Piphillin pipelines. Inference of predictive functionality using these two said pipelines showed a high metabolism rate, since most of these products are consortia of many metabolically active microbiota (Shangpliang et al. 2018). These findings are similar to recent studies reported from fermented dairy products (Zhang et al. 2017; Zhu et al. 2018; Chen et al. 2020; Choi et al, 2020a,b). The association of various metabolic pathways such as amino acid metabolism, carbohydrate metabolism, energy

159 metabolism, lipid metabolism, metabolism of cofactors and vitamins, and other secondary
160 metabolites with the bacterial genera indicated an active interaction of bacteria-function
161 complexity. LAB are predominant microbiota in many ethnic fermented milk products of
162 India followed by few AAB (Tamang et al. 2000; Dewan and Tamang 2006, 2007;
163 Shangpliang et al. 2018; Ghosh et al. 2019; Shangpliang and Tamang 2021). Spearman's
164 correlation of the predominant bacterial genera with the predictive functionality resulted in a
165 complex microbial-functions interaction in NFM products of Sikkim and Arunachal Pradesh.
166 Metabolic activity such as amino acid metabolism is important in dairy products as they
167 contribute in development of flavour (Yvon and Rijnen 2001). Similarly, carbohydrate
168 metabolism does also play a major role in flavour and aroma development in milk
169 fermentation (Pan et al. 2014). The abundance of functional pathways related to metabolism
170 of amino acids, lipid, energy and carbohydrates were earlier reported in fermented milk and
171 milk products (Zhang et al. 2017; Ramezani et al. 2017; Zhu et al. 2018; Yasir et al. 2020;
172 Chen et al. 2020). A high correlation of functional properties and LAB have also been
173 reported in cheeses (Yang et al. 2020), since LAB are the most predominant microorganisms
174 in fermented milk products (Rezac et al. 2018; Chen et al. 2020). We observed a positive
175 correlation of *Staphylococcus* with the predictive metabolic features of these NFM products,
176 and interestingly, *Staphylococcus* is metabolically active in dairy products playing functional
177 activities such as amino acid metabolism, carbohydrate metabolism, lipid metabolism and
178 nitrogen metabolism (Leroy et al. 2020). We also observed the presence of significant
179 correlation of bacteria with cofactors and vitamins metabolism such as ubiquinone and other
180 terpenoid-quinone biosynthesis and lipoic acid metabolism, which are essential for other
181 microbial metabolism (Yao et al. 2020). Apart from LAB, AAB have also contributed to
182 many functional features in NFM products; AAB involve in protein metabolism, production
183 of secondary metabolites and volatile compounds (Illeghems et al. 2015; Ai et al. 2019).

Functional profiles from both PICRUSt2 and Piphillin were normalized using MUSiCC (Manor and Borenstein 2015), which is a marker gene-based method which use universal single-copy genes for biasness correction of gene abundances (Noecker et al. 2017). Normalization using MUSiCC have proven necessary for gene functional studies (Vincent et al. 2017), rescaling the abundant predicted KOs to the actual average gene copy number, correcting several known biases (Manor and Borenstein 2017). Piphillin is usually applied in human clinical samples (Iwai et al. 2016); whereas PICRUSt2 is widely used for environmental samples (Douglas et al. 2020). However, these pipelines have also been widely used in dairy products (Choi et al. 2020a,b).

From our present analysis, PICRUSt2 analysis generated more predicted KOs and KEGG pathways in comparison to that of Piphillin. Though, significant differences were observed, however, there are functions which were predicted only from PICRUSt2 and missing in Piphillin and vice versa. Therefore, consolidated predictive functions from both these pipelines are necessary for a comprehensive outlook into the potential of bacteria associated with NFM products. Though predictive functionality study of the microbiota associated with NFM products at present is only speculations using bioinformatics tools, a general outlook into the potentiality of functions may be studied and compared. Nonetheless, in the absence of shotgun metagenomics data, using PICRUSt2 and Piphillin serves to be the reliable analysis for microbial predictive gene function.

Conclusion

Bacterial community in NFM products showed many functional features with many important health benefits to consumers. We applied PICRUSt2 and Piphillin tools to infer the predictive functional features of microbiota associated with the ethnic fermented milk

208 products of India. Therefore, such studies may be used for future comparison with detailed
209 gene functionality studies of other fermented foods elsewhere.
210
211 **Acknowledgements**
212 The authors are grateful to the Department of Biotechnology, Govt. of India through the
213 DAICENTER project. HNJS is grateful to DBT for Junior Research Fellowship.
214
215 **Funding**
216 This current research is supported by Department of Biotechnology, Govt. of India through
217 DBT-AIST International Centre for Translational and Environmental Research
218 (DAICENTER) project.
219
220 **Authors' contributions**
221 HNJS did analysis and bioinformatics analysis. JPT has supervised the bioinformatics
222 analysis and finalised the manuscript.
223
224 **Availability of data and materials**
225 Raw sequences were accessed from MG-RAST server having the MG-RAST ID number
226 4732361 to 4732414. The same were accessed from NCBI database server under the
227 BioProject No. PRJNA661385 with accession numbers SAMN16056817 to
228 SAMN16056870.
229
230 **Declaration of Competing Interest**
231 The authors declare that they have no competing interests.
232

References

Ai, M., Qiu, X., Huang, J., Wu, C., Jin, Y., & Zhou, R. (2019). Characterizing the microbial diversity and major metabolites of Sichuan bran vinegar augmented by *Monascus purpureus*. *International Journal of Food Microbiology* 292, 83–90.

Barbera, P., Kozlov, A.M., Czech, L., Morel, B., Darriba, D., Flouri, T. & Stamatakis, A. (2019). EPA-ng: massively parallel evolutionary placement of genetic sequences. *Systematic Biology* 68(2), 365–369.

Bengoa, A.A., Iraporda, C., Garrote, G.L., & Abraham, A.G. (2019). Kefir micro-organisms: their role in grain assembly and health properties of fermented milk. *Journal of Applied Microbiology* 126(3), 686–700.

Bokulich, N.A., Ziemski, M., Robeson, M. & Kaehler, B. (2020). Measuring the microbiome: best practices for developing and benchmarking microbiomics methods. *Computational and Structural Biotechnology Journal* 18, 4048–4062.

Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., Alexander, H., Alm, E.J., Arumugam, M. & Asnicar, F. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology* 37(8), 852–857.

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A. & Holmes, S.P. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. *Nature Methods* 13(7), 581–583.

Chen, X., Zheng, R., Liu, R. & Li, L. (2020). Goat milk fermented by lactic acid bacteria modulates small intestinal microbiota and immune responses. *Journal of Functional Foods* 65, 103744. https://doi.org/https://doi.org/10.1016/j.jff.2019.103744

Choi, J., Lee, I., Rackerby, B., Frojen, R., Goddik, L., Ha, S.D. & Park, S.H. (2020a). Assessment of overall microbial community shift during Cheddar cheese production

from raw milk to aging. *Applied Microbiology and Biotechnology* 104, 6249–6260.

Choi, J., Lee, S.I., Rackerby, B., Goddik, L., Frojen, R., Ha, S.D., Kim, J.H. & Park, S.H. (2020b). Microbial communities of a variety of cheeses and comparison between core and rind region of cheeses. *Journal of Dairy Science* 103(5), 4026–4042.

Czech, L. & Stamatakis, A. (2019). Scalable methods for analyzing and visualizing phylogenetic placement of metagenomic samples. *PloS One* 14(5), e0217050–e0217050. https://doi.org/https://doi.org/10.1371/journal.pone.0217050

Dewan, S. & Tamang, J.P. (2006). Microbial and analytical characterization of Chhu-A traditional fermented milk product of the Sikkim Himalayas. *Journal of Scientific and Industrial Research* 65, 747–752.

Dewan, S. & Tamang, J.P. (2007). Dominant lactic acid bacteria and their technological properties isolated from the Himalayan ethnic fermented milk products. *Antonie van Leeuwenhoek* 92(3), 343–352.

Douglas, G.M., Maffei, V.J., Zaneveld, J.R., Yurgel, S.N., Brown, J.R. Taylor, C.M., Huttenhower, C., & Langille, M.G.I. (2020). PICRUSt2 for prediction of metagenome functions. *Nature Biotechnology* 38, 685–688.

García-Burgos, M., Moreno-Fernández, J., Alférez, M.J.M., Díaz-Castro, J. & López-Aliaga, I. (2020). New perspectives in fermented dairy products and their health relevance. *Journal of Functional Foods* 72, 104059. https://doi.org/https://doi.org/10.1016/j.jff.2020.104059

Ghosh, T., Beniwal, A., Semwal, A. & Navani, N.K. (2019). Mechanistic insights into probiotic properties of lactic acid bacteria associated with ethnic fermented dairy products. *Frontiers in Microbiology* 10, 502. https://doi.org/https://doi.org/10.3389/fmicb.2019.00502

Illeghems, K., Weckx, S. & De Vuyst, L. (2015). Applying meta-pathway analyses through

metagenomics to identify the functional properties of the major bacterial communities of a single spontaneous cocoa bean fermentation process sample. *Food Microbiology* 50, 54–63.

Iwai, S., Weinmaier, T., Schmidt, B.L., Albertson, D.G., Poloso, N.J., Dabbagh, K. & DeSantis, T.Z. (2016). Piphillin: improved prediction of metagenomic content by direct inference from human microbiomes. *PloS One* 11(11), e0166104. https://doi.org/https://doi.org/10.1371/journal.pone.0166104

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research* 40(D1), D109–D114.

Leroy, S., Even, S., Micheau, P., de La Foye, A., Laroute, V., Le Loir, Y. & Talon, R. (2020). Transcriptomic analysis of *Staphylococcus xylosus* in solid dairy matrix reveals an aerobic lifestyle adapted to rind. *Microorganisms* 8(11), 1807. https://doi.org/https://doi.org/10.3390/microorganisms8111807

Louca, S. & Doebeli, M. (2018). Efficient comparative phylogenetics on large trees. *Bioinformatics* 34(6), 1053–1055.

Manor, O. & Borenstein, E. (2015). MUSiCC: a marker genes based framework for metagenomic normalization and accurate profiling of gene abundances in the microbiome. *Genome Biology* 16(1), 53. https://doi.org/https://doi.org/10.1093/nar/gkr988

Manor, O. & Borenstein, E. (2017). Systematic characterization and analysis of the taxonomic drivers of functional shifts in the human microbiome. *Cell Host & Microbe* 21(2), 254–267.

Metsalu, T. & Vilo, J. (2015). ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. *Nucleic Acids Research* 43(W1),

W566–W570.

Narayan, N.R., Weinmaier, T., Laserna-Mendieta, E.J., Claesson, M.J., Shanahan, F., Dabbagh, K., Iwai, S. & DeSantis, T.Z. (2020). Piphillin predicts metagenomic composition and dynamics from DADA2-corrected 16S rDNA sequences. *BMC Genomics* 21(1), 1–12. https://doi.org/https://doi.org/10.1186/s12864-020-6537-9

Noecker, C., McNally, C.P., Eng, A. & Borenstein, E. (2017). High-resolution characterization of the human microbiome. *Translational Research* 179, 7–23.

Ortiz-Estrada, Á.M., Gollas-Galván, T., Martínez-Córdova, L.R. & Martínez-Porchas, M. (2019). Predictive functional profiles using metagenomic 16S rRNA data: a novel approach to understanding the microbial ecology of aquaculture systems. *Reviews in Aquaculture* 11(1), 234–245.

Pan, D.D., Wu, Z., Peng, T., Zeng, X.Q. & Li, H. (2014). Volatile organic compounds profile during milk fermentation by *Lactobacillus pentosus* and correlations between volatiles flavor and carbohydrate metabolism. *Journal of Dairy Science* 97(2), 624–631.

Parks, D.H., Tyson, G.W., Hugenholtz, P. & Beiko, R.G. (2014). STAMP: statistical analysis of taxonomic and functional profiles. *Bioinformatics* 30(21), 3123–3124.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J. & Glöckner, F. O. (2012). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research* 41(D1), D590–D596.

Quince, C., Walker, A., Simpson, J., Loman, N.J., Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology 35,* 833–844.

Rai, R., Shangpliang, H.N.J. & Tamang, J.P. (2016). Naturally fermented milk products of the Eastern Himalayas. *Journal of Ethnic Foods* 3(4), 270–275.

Ramezani, M., Hosseini, S.M., Ferrocino, I., Amoozegar, M.A. & Cocolin, L. (2017). Molecular investigation of bacterial communities during the manufacturing and ripening

of semi-hard Iranian Liqvan cheese. *Food Microbiology* 66, 64–71.

Rezac, S., Kok, C.R., Heermann, M. & Hutkins, R. (2018). Fermented foods as a dietary source of live organisms. *Frontiers in Microbiology* 9, 1785. https://doi.org/https://doi.org/10.3389/fmicb.2018.01785

Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 4, e2584. https://doi.org/https://doi.org/10.7717/peerj.2584

Shangpliang, H.N.J., Rai, R., Keisam, S., Jeyaram, K. & Tamang, J.P. (2018). Bacterial community in naturally fermented milk products of Arunachal Pradesh and Sikkim of India analysed by high-throughput amplicon sequencing. *Scientific Reports* 8(1), 1532. https://doi.org/https://doi.org/10.1038/s41598-018-19524-6

Shangpliang, H.N.K. and Tamang, J.P. (2021). Phenotypic and genotypic characterizations of lactic acid bacteria isolated from exotic naturally fermented milk (cow and yak) products of Arunachal Pradesh, India. *International Dairy Journal* 118: 105038.doi.org/ 10.1016/j.idairyj.2021.105038

Tamang, J.P., Cotter, P.D., Endo, A., Han, N.S., Kort, R., Liu, S.Q., Mayo, B., Westerik, N. & Hutkins, R. (2020). Fermented foods in a global age: East meets West. *Comprehensive Reviews in Food Science and Food Safety* 19(1), 184–217.

Tamang, J.P., Dewan, S., Thapa, S., Olasupo, N.A., Schillinger, U., Wijaya, A. & Holzapfel, W. H. (2000). Identification and enzymatic profiles of the predominant lactic acid bacteria isolated from soft-variety Chhurpi, a traditional cheese typical of the Sikkim Himalayas. *Food Biotechnology* 14(1–2), 99–112.

Tamang, J.P., Jeyaram, K., Rai, A.K. & Mukherjee, P.K. (2021). Diversity of beneficial microorganisms and their functionalities in community-specific ethnic fermented foods of the Eastern Himalayas. *Food Research International* 148,

110633.doi.org/10.1016/j.foodres.2021.110633.

Vincent, A.T., Derome, N., Boyle, B., Culley, A.I. & Charette, S.J. (2017). Next-generation sequencing (NGS) in the microbiological world: How to make the most of your money. *Journal of Microbiological Methods* 138, 60–71.

Yang, C., Zhao, F., Hou, Q., Wang, J., Li M. & Sun, Z. (2020). PacBio sequencing reveals bacterial community diversity in cheeses collected from different regions. *Journal of Dairy Science* 103(2) 1238–1249.

Yao, Y., Zhou, X., Hadiatullah, H., Zhang, J. & Zhao, G. (2020). Determination of microbial diversities and aroma characteristics of Beitang shrimp paste. *Food Chemistry* 128695. https://doi.org/https://doi.org/10.1016/j.foodchem.2020.128695

Yasir, M., Bibi, F., Hashem, A.M. & Azhar, E.I. (2020). Comparative metagenomics and characterization of antimicrobial resistance genes in pasteurized and homemade fermented Arabian laban. *Food Research International* 137, 109639. https://doi.org/https://doi.org/10.1016/j.foodres.2020.109639

Ye, Y. & Doak, T.G. (2009). A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. *PLoS Computational Biology* 5(8). https://doi.org/https://doi.org/10.1371/journal.pcbi.1000465

Yvon, M. & Rijnen, L. (2001). Cheese flavour formation by amino acid catabolism. *International Dairy Journal* 11(4–7), 185–201. Zhang, F., Wang, Z., Lei, F., Wang, B., Jiang, S., Peng, Q., Zhang, J. & Shao, Y. (2017). Bacterial diversity in goat milk from the Guanzhong area of China. *Journal of Dairy Science* 100(10), 7812–7824.

Zhu, Y., Cao, Y., Yang, M., Wen, P., Cao, L., Ma, J., Zhang, Z. & Zhang, W. (2018). Bacterial diversity and community in Qula from the Qinghai–Tibetan plateau in China. *PeerJ 6*, e6044. https://doi.org/https://doi.org/10.7717/peerj.6044

384 **Legends for Figures:**

385 **Figure 1:** An overall categorical representation of the MUSiCC-normalized predictive
386 microbial functions as inferred by (a) PICRUSt2 and (b) Piphillin.

387

388 **Figure 2:** Extended error bar chart representation of the significant predictive functionalities
389 as inferred by both PICRUSt2 and Piphillin. (a) Overall, metabolism is significantly higher in
390 PICRUSt2 analysis as compared to that of Piphillin, however, (b) a shared difference was
391 observed at the super-pathway level. Significance (q-value>0.05) was calculated using
392 White's non-parametric test with Benjamini-Hochberg FDR (false discovery rate) in
393 STAMP.

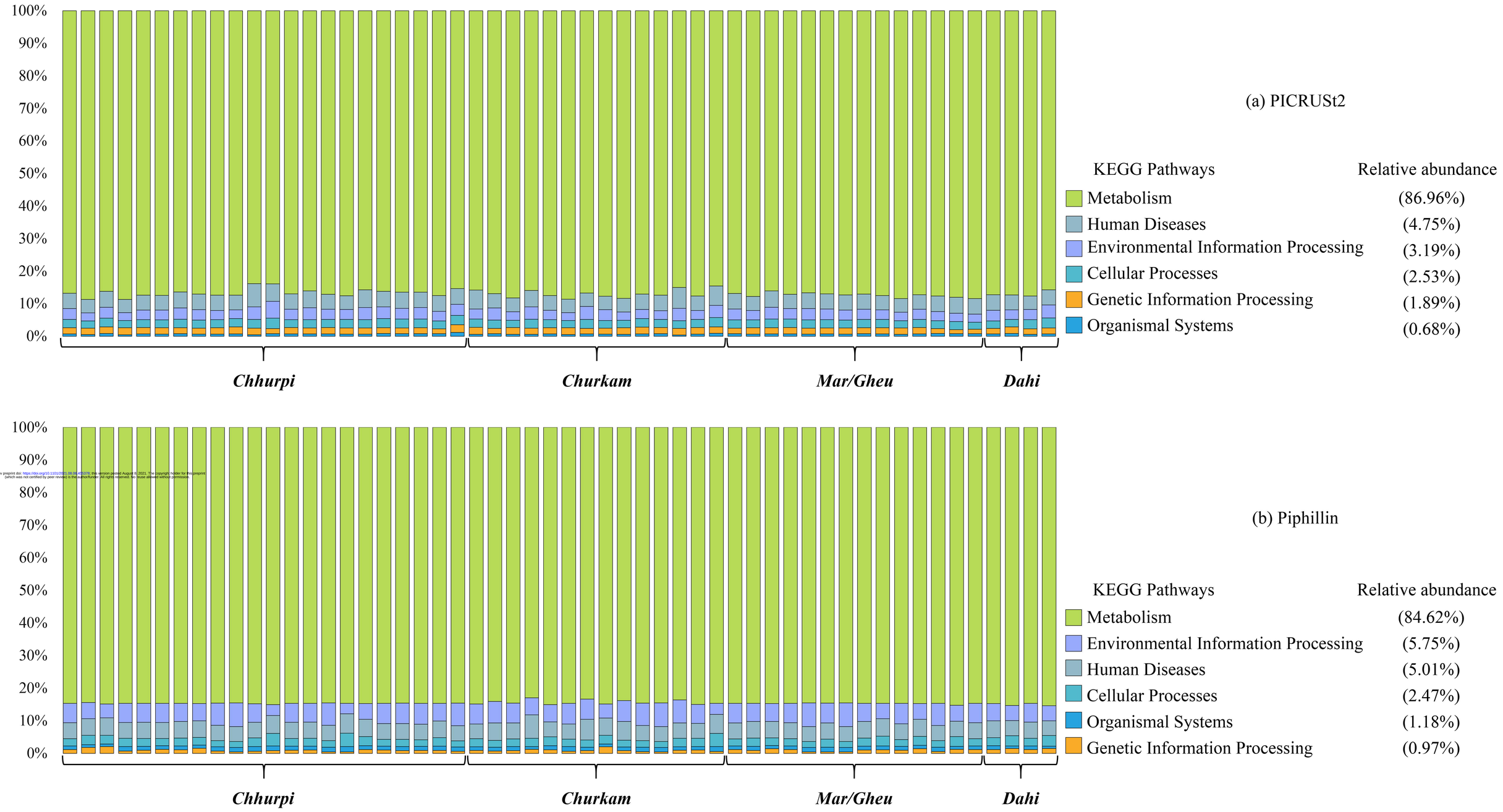
394

395 **Figure 3:** An overall comparison of the significant metabolic pathways as inferred by
396 PICRUSt2 and Piphillin depicting a significant number of functional features predicted by
397 these two pipelines. Significance (q-value>0.05) was calculated using White's non-
398 parametric test with Benjamini-Hochberg FDR (false discovery rate) in STAMP**.**

399

400 **Figure 4:** Non-parametric Spearman's correlation of the ASV-associated predominant
401 bacterial genera of the NFM products with a consolidated functional feature as inferred by
402 both PICRUSt2 and Piphillin. Here, calculation was carried out using Statistical Package for
403 the Social Sciences (SPSS) v20 and heatmap was generated using ClustVis. All significant
404 correlation pairs are denoted by * (*<0.05 and **<0.01). LAB-lactic acid bacteria; AAB-
405 acetic acid bacteria.

406



(a)

(b)