Community structure and metabolic potentials for biofuel production of an ancient beer starter ‘emao' revealed by a whole-metagenome study

Diganta Narzary (d_narzary@gauhati.ac.in)
Gauhati University
Nitesh Boro
Gauhati University
Ashis Borah
Gauhati University
Okubo Takashi
Yokohama Institute for Earth Sciences, JAMSTEC, Yokohama
Hideto Takami
Yokohama Institute for Earth Sciences, JAMSTEC, Yokohama

Research Article

Keywords: Traditional beer starter, community structure, metagenomics, feasible metabolic modules, biofuel producing enzymes

Posted Date: December 15th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-124778/v1

Version of Record: A version of this preprint was published at Scientific Reports on July 16th, 2021. See the published version at https://doi.org/10.1038/s41598-021-94059-x.

Community structure and metabolic potentials for biofuel production of an ancient beer starter '*emao*' revealed by a whole-metagenome study

Diganta Narzary[ab*], Nitesh Boro[a], Ashis Borah[a], Takashi Okubo[b#] and Hideto Takami[b#]

[a]Microbiology and Molecular Systematics Lab, Department of Botany, Gauhati University, Guwahati, Assam, India

[b]Yokohama Institute for Earth Sciences, JAMSTEC, Yokohama 236-0001, Japan

*Corresponding author: d_narzary@gauhati.ac.in

[#]Current address: Atmosphere and Ocean Research Institute, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa 277-8564, Japan

Abstract

The community structure and functional metabolic modules occurring in '*emao*' – a traditional rice beer starter of ancient origin from north-east India has been unearthed for the first time using shot-gun whole-metagenome sequencing. *Emao* harbours potential microorganisms for saccharification, lignocellulose degradation, and biofuel production that correlate with probable metabolic functional modules. Eukaryotic microorganisms, especially moulds and yeasts, dominated over the prokaryotes in *emao* compared to previous metagenomic studies on such traditional starters where the relative abundance of prokaryotes was higher than the eukaryotes. The comprehensive microbial species diversity, community structure, and metabolic modules exhibited in *emao* would be practical in the formulation of mixed-microbial cultures for biofuel production from plant-based feedstocks.

Keywords: Traditional beer starter, community structure, metagenomics, feasible metabolic modules, biofuel producing enzymes

Introduction

The starter cultures used in traditional beer fermentations are generally composed of diverse microorganisms [1] that form a complex microbial consortium in each starter that gives us desirable end-products in fermentations. Understanding of community structures in various environmental samples, including traditional starters, is crucial for bioprospection. However, the estimation of community structure using culture-dependent methods is difficult due to various reasons such as non-cultivable of specific microbes in laboratory conditions [2], competition for nutrients and dominance of some microbes over the others in culture media, and the different growth rates among the species that hamper the community structure analysis. In contrast, culture-independent methods, such as metagenomics through Next Generation Sequencing (NGS), allow the microbial communities to study without the need for microbial isolation and cultivation [2]. Although recently PCR-amplicon sequencing of 16S rRNA genes and ITS rDNA regions have been used for microbial diversity study in many environmental samples including traditional starter cultures, this method is vulnerable to biases [3,4] due to copy number variation of the target marker at the species level (e.g., 16S rRNA genes in bacteria) and the inherent PCR problems (such as template competition, primer mismatch, biased amplification due to different template copy number) while using metagenomic DNAs. Therefore, the use of PCR amplicon-based NGS sequencing is also not reliable for microbial community structure analysis. Alternatively, the use of ribosomal proteins as markers for community structure estimation in the whole-metagenome approach is new hope in microbial ecology due to occurrence of almost a single-copy of each ribosomal protein in an individual [5,6].

The craft of brewing rice-beer (*zu* or *zou*) using *emao* (also written as *amao*) is an age-old tradition among the Bodo tribe – one of the aborigines of northeast India [7,8]. The origin of the Bodo tribe and thus their brewing tradition is still obscure. The Bodo, Dimasa, and Garo tribes of North East India are supposed to be segregated from a common ancestor about 1,500 years ago [9]. They have close similarities in their language, culture, and brewing traditions. Preparation of *emao* from non-sticky rice powder adding certain herbs and a small portion of *emao* from existing stock is the Bodos tradition (Table 1). The traditional belief of using the herbs in *emao* is to bring sweetness or hardness to the beer, depending on the herb used. Usually, 3–5 herbs are used in *emao* preparation, sometimes alternatively based on availability at the time of its preparation. Researchers studied the microbial diversity on several such traditional starters across the world using both culture-dependent [10-16] and

culture-independent methods [14,17-23]. Nonetheless, the microbial study on *emao* is limited, and only four fungal species (one mould and three non-*Saccharomyces* yeasts) have been reported to date using culture-dependent methods [27-27]. In the case of *emao*, bacterial diversity and culture-independent studies are not yet available.

The occurrence of three major groups of microorganisms *viz*. amylolytic, alcohol-producing, and lactic acid bacteria (LAB) are reported in similar kinds of traditional beer starter cultures by other researchers. Thus, several microorganisms work together in consortia mode in most traditional beer fermentation, which gives us unique products. The traditional way of starter culture preparations, maintenance, and the substrate used for fermentation varies among the people from the different locality as well as the community [1], so thus the microbial species diversity and the fermented products differ from starter to starter. More than two dozen traditional beer starters are in practice throughout northeast India [28], and only a few of them have been studied using culture-independent method [21,23,42]. Here, we unravelled the microbial community structure of *emao* based on ribosomal protein sequences derived from the whole-metagenome sequences and the metabolic potentials of *emao* for biofuel production for the first time.

Results

Whole-metagenome overview

The number of nucleotide pair sequences obtained from the pooled-sample of six *emao* representatives was 13,060,410, and the total contigs with ≥400 nt in each were 6,530,205 (Table 2). We obtained a total of 1,285,880 amino acid (AA) sequences with ≥50 AA in each, which was subsequently used as an input for taxonomic binning, enzyme identification, and metabolic module analyses.

Community structure and species diversity

The taxonomic binning of the non-redundant ribosomal protein (r-protein) sequences revealed 92% Eukarya and 8% Bacteria with a ratio of 9:2:1 for Mucoromycota (moulds), Ascomycota (yeasts), and Firmicutes (bacteria), respectively (Fig. 1a). The relative abundance of moulds was the highest (73.44%), followed by yeasts (18.02%) and lactic acid bacteria (LAB) (7.87%) (Fig. 1b), in contrast to four times higher relative species diversity in yeasts than bacteria and moulds, both showing equal numbers of species (12) in *emao* (Table S1). The relative abundance of the genus *Rhizopus* and its family Rhizopodaceae were the highest among the groups in *emao* (Fig. 1c, d). Among the yeasts, the relative abundance of the genus *Wickerhamomyces* and its family Phaffomycetaceae were the highest.

We identified a total of 74 microbial species in *emao* based on r-protein sequences (Fig. 2 and Table S1). *Rhizopus delemar* (syn. *R. oryzae*) was the highest with 56% overall relative abundance (ORA), followed by *R. microsporus* (7% ORA) and *Mucor circinelloides* (5% ORA) among the moulds. Among the yeasts, *Wickerhamomyces anomalus* was the most dominant (9% ORA), followed by *W. ciferrii* (1% ORA), *Ascoidea rubescens* (0.6% ORA), *Cyberlindnera fabiani* (0.6%), *Pachysolen tannophilus* (0.6% ORA), *Candida tropicalis* (0.5% ORA), *Saccharomyces cerevisiae* (0.4% ORA), and a few more ethanol-producing species with low ORA (Table S1). Among LAB, the most dominant species was *Leuconostoc mesenteroides* (1.9% ORA), followed by *Weissella confusa* (1.8% ORA) and *Lactococcus garvieae* (1% ORA).

The microbial diversity compared to other traditional beer starters revealed that the *emao* exhibits 26% of its microbial species common to *Nuruk* from Korea [20,29], 18% to both *Marcha* [30] and *Xaj-pitha* [21] from India, and 16% to *Daku* from China [11,18]. Some of the dominant species recorded in *emao* were also common in several other traditional beer starters considered for comparison. Out of 24 other traditional starters (OTS) compared to *emao*, the dominant microbial species, *W. anomalus* was found common in 20 OTS, *S. cerevisiae* in 15 OTS, *R. delemar* in 13 OTS, *M. circinelloides* in 12 OTS, *Pediococcus pentosaceus* in 9 OTS, *R. microsporus* in 7 OTS, and *Torulaspora delbrueckii*, *L. mesenteroides* and *Candida tropicalis* in 6 OTS. However, the other dominant species (*R. stolonifer*, *Choanephora cucurbitarum*, *Parasitella parasitica*, *W. ciferrii*, *C. fabianii*, *A. rubescens*, *L. citreum*, *W. confusa*, and *L. garvieae*) as recorded in *emao*, was reported only in a few OTS (Table S1). The overall species similarity of microorganisms recorded in *emao* is only 43% compared to other traditional beer starters.

Identification of CAZymes and BPZymes in emao

A total of 19,702 CAZymes (out of which only 45% were assigned to a total of 123 CAZyme sub-families) and 34,493 BPZymes were recorded in *emao* (Fig. 3a-b, Fig. S1 & S2). Among the CAZymes, glycoside hydrolases (GH) were recorded the highest (21%), followed by glycosyltransferases (GT, 18%), whereas the rest four types of CAZymes were comparatively quite low (≤3%). We also emphasized identifying the lignocellulolytic enzymes occurring in *emao* as essential for mobilizing lignocellulosic substrates into useful products such as biofuel. There is no readymade database for lignocellulolytic enzymes as such; therefore, we compared the CAZymes and BPZymes of *emao* to the previous reports on lignocellulolytic enzymes [31,32] and identified a total of 1,929 lignocellulolytic CAZymes

and 5,576 lignocellulolytic BPZymes in *emao* (Fig. 3c-d). A comparison of lignocellulolytic enzymes of *emao* to pill bug (*Armadillidium vulgare*) gut microbiome [32] revealed seven times higher lignin modifying enzymes, four times higher hemicellulases, and four times higher hemicellulases and/or cellulases in *emao*, in contrast to four times higher lignocellulose-binding modules in pill bug. In BPZyme analysis, enzymes associated with alcohol production were found two times higher than the enzymes associated with diesel production and fuel cells. We could identify 44% ethanol-producing, 26% fuel-cell producing, 20% diesel producing, and 10% alternate-biofuel producing enzymes (Fig. 3b).

Metabolic and physiological potentials in emao

An analysis of amino acid sequences in Genomaple [33,34] revealed the most feasible functional modules with significant module completion ratio (MCR) and *Q*-values at an individual taxonomic rank (ITR) or whole microbial count (WC) level. Any module having 100% MCR and/or less than 0.5 *Q*-value is considered significant and feasible [33, 35]. A total of 489 metabolic KEGG (Kyoto Encyclopedia of Genes and Genomes) modules with > 0% MCR (WC) scores were recorded in *emao*, out of which 46% (*i.e.*, 28% out of total 804 existing modules in KEGG) had 100% MCR (WC), and that could be the crucial modules in determining the functionality and uniqueness of *emao* in beer fermentations (Table S2). The carbohydrate and lipid metabolisms are involved directly or indirectly in plant biomass degradation and biofuel production. Therefore, we focused on carbohydrate and lipid metabolic modules having significant MCR and *Q*-values (Fig. 4, Table 3).

The presence of 100% MCR for the Embden-Meyerhof pathway with zero *Q*-value in contrast to 75–80% MCR (0.5–0.85 *Q*-value) for the Entner-Doudoroff pathway which occurs mainly in some bacteria, revealed the presence of feasible alternative glycolytic pathways in *emao* (Table 3). Besides, 100% MCR with zero *Q*-value for pectin, galactose, D-galacturonate, and D-glucuronate degradation pathways are noteworthy, which signifies the possibility of metabolizing those substrates by the microorganisms associated with *emao* (Fig. 2). Acyl-CoA is necessary for the synthesis of fatty acids precursor, acetyl-CoA [36,37]. Thus, 100% MCR for both the beta-oxidation module and acyl-CoA synthesis module suffices the fatty acids and isoprenoid biosynthesis potentiality in *emao*. The presence of 100% MCR for mitochondrial and endoplasmic fatty acid biosynthesis modules and 100% MCR for lipid biosynthesis modules corroborate the involvement of eukaryotes (mainly fungi) in biodiesel production.

Discussion

We found a reasonable and justifiable community structure in *emao* using ribosomal protein-based taxonomic binning where the yeasts, moulds, and LAB were seen at a ratio of 9:2:1, respectively. We consider this community structure of *emao* ideal for producing traditional *zou* or *joubishi* containing 5–17% (w/v) ethanol that might have had garnered thousand years ago and is still being maintained in its pristine form by the Bodo peoples through their traditional practices. The findings of higher relative abundance of Firmicutes than Proteobacteria, and higher fungal species counts than bacteria in *emao* (Fig. 4) differ from the community structures as reported previously in the traditional beer starters *Xaj-pitha* [21], *Marcha* and *Thiat* [30] and from the region. Such a deviation is likely due to differences in methods and approaches followed.

The yeasts, moulds, and LAB play unique roles in alcoholic beverage fermentation, especially when starch is the feedstock. Moulds are mostly aerobic, do the saccharification, and sensitive to ethanol except for a few species that can produce a low level of ethanol [38]. The presence of a high number of saccharifying and/or lignocellulose degrading microorganisms is always advantageous and desirable at the beginning of fermentation while using plant-based substrates. Some non-*Saccharomyces* yeast does both saccharification and ethanol production simultaneously but is mostly non-tolerant to high ethanol concentrations [39-42]. Ethanol sensitive microbes are subsequently killed or arrested by increasing ethanol concentrations at the later stages of ethanol fermentation [43]. Ethanol production and tolerance level of *Saccharomyces* yeasts also vary from strain to strain, although they are generally more ethanol tolerant than the others [44-46]. Wild *S. cerevisiae* cannot utilize starch or other complex carbohydrates directly due to the lack of degrading enzymes for those substrates [47] for which they are dependent on other saccharifying microbes.

A low proportion of LAB is desirable in beer fermentation as they can perform some necessary functions despite knowing them as spoilage agents [48, 49]. They mainly produce lactic acids besides producing bacteriocins against some human pathogens, and they are responsible for maintaining low pH, keeping quality, and taste enhancement in beer [50-53]. We did not find any acetic acid-producing bacteria (Acetobacters) in *emao* – a good sign for this traditional starter of ancient origin. Acetic acid bacteria are responsible for beer defects [54,55], which often happens due to contamination of such spoilage agents if hygienic conditions are compromised during the preparation of starter culture or beer fermentation.

Our findings of a good number of saccharifying, lignocellulolytic (Table S3), and different biofuels producing microorganisms in corroboration with probable metabolic

functionomes (Fig. 4, Table 3) in *emao* is noteworthy, and the current community structure as unraveled in *emao* could help develop an effective lignocellulolytic bio-consortia necessary for second-generation biofuel production. Many simple carbohydrate degrading microorganisms do not easily break down the pentose sugars such as D-xylose and L-arabinose that comprise up to 20% of lignocellulosic biomass [56]. The breaking of lignocellulose into its subunits and subsequently mobilizing them as energy sources for biofuel production is complex. It requires various enzymes to catalyze the metabolic processes, and an organism bearing all the essential enzymes together is rare. However, further experimentation is required to validate the utility of *emao* as such or in combination with other lignocellulolytic microbes. As the species diversity and the metabolic potentials of *emao* as reflected from this study are diverse, a pyramiding of target microorganisms from such a natural bio-consortium towards achieving a target product from target substrates by necessary functional potentials could be a new avenue in tapping natural bioresources for bioprospection. Such an alternative approach can pave the way for bio-consortia formulation to produce biofuels from the lignocellulosic materials.

The Genomaple system provides an effective platform to visualize the module completion ratio (MCR) along with the taxonomic information at Phylum or Class level that reflects their functional activity in completing different metabolic modules in any environmental sample considered for an investigation [33,34]. However, it is dependent on the KEGG database, which includes only the species with complete genome sequence information. Unfortunately, not a single representative of moulds (Zygomycota) is available in the KEGG database due to the absence of complete genome information, in contrast to several draft genome information on this particular group fungus to date. For the same reason, we cannot specify and segregate the metabolic modules among the moulds and yeasts in this functional metagenome study on *emao*. Moulds are ubiquitous in distribution and play critical ecological roles like other fungal groups [57]. Complete genome information on moulds is needed in the public domain to understand better their roles in natural environments and brewing.

Conclusions

The present study is the first to unravel the comprehensive community structure and probable metabolic potentials of the microorganisms associated with the traditional starter culture *emao*. The presence of diverse groups of microorganisms in corroboration with amylolytic, lignocellulolytic, biofuels producing enzymes as recorded in *emao* is noteworthy.

231 It could be a pathfinder in the field of microbial consortia bioformulation for biofuel
232 production from otherwise recalcitrant plant biomasses. We found the ribosomal protein-
233 based community structure enumeration a suitable approach in metagenome study. Complete
234 genome information on moulds is equally essential in the line of other fungal groups to better
235 understand their roles in traditional brewing and other natural environments. Some dominant
236 microbial species recorded in *emao* are familiar to some other traditional starters reported
237 earlier from the region. Therefore, a comparative study of age-old traditional starters that
238 carry microbial genetic information could be of paramount significance in understanding the
239 history of human population migration and civilization in ancient times, like archaeology and
240 philology.

241 **Materials and methods**

242 *Sample collection*

243 Six representative *emao* samples traditionally prepared by the Bodo people were
244 collected from different Bodo-dominated Assam (Table 1). Information on the method of
245 preparation and herbs used in starter culture were also recorded. We ensured that no beer
246 defect had been experienced during sample collection in using *emao* from the same stock that
247 we brought for scientific investigation. Traditionally, more than one-year-old *emao* samples
248 are generally not used for rice beer fermentation; instead, the starter culture is revived in the
249 fresh rice-based medium before completing one year. Therefore, only the active samples,
250 *i.e.,* less than one year from the date of preparation, were considered for this study. (since
251 starter culture is revived within a month among the active brewers)

252 *Total DNA isolation*

253 Two grams of each of six *emao* samples were taken and mixed by grinding in pre-
254 sterilized mortar and pestle, from which 10 g sample was taken to isolate the total DNA. To
255 recover quality metagenomic DNA, we modified the method of Zhou *et al*. [58] where the
256 extraction buffer was supplemented with 1% activated charcoal and 10 mM $MgCl_2$ as
257 recommended by Sharma *et al*. [59], and the DNA obtained from the modified method of
258 Zhou *et al*. [58] was again purified with the MoBio DNA isolation kit (QIAGEN, Cat. No.
259 12888-100). The steps we followed are described below. Liquid nitrogen was used for
260 effective sample grinding and the sample was transferred into 18 ml of extraction buffer [100
261 mM Tris-HCl (pH 8.0), 100 mM sodium EDTA (pH 8.0), 100 mM sodium phosphate (pH
262 8.0), 1.5 M NaCl, 1% CTAB, 10mM $MgCl_2$ and 1% activated charcoal] before thawing. A

volume of 100 µl proteinase K (10 mg/ml) was added to the tube, mixed properly by vortexing, and incubated at 37°C for 30 minutes in a water bath with gentle end-over-end inversions every 5 min. After that, 2 ml of 20% SDS was added, mixed by vortexing, and incubated at 65°C for 2 hours with gentle end-over-end inversions every 30 min. The sample was allowed to cool up to room temperature (RT), and the supernatant was collected in a fresh centrifuge tube after centrifugation at 6000 × g for 10 min at RT. The supernatant was mixed with an equal volume of chloroform: isoamyl alcohol (24:1, vol/vol) by inverting the tube gently. The aqueous phase was recovered by centrifugation at 10000 × g for 10 minutes at RT and then precipitated with 0.6 volume of pre-chilled isopropanol at RT for one h. The pellet of crude nucleic acids was obtained by centrifugation at 12,000 × g for 20 min at RT, washed with chilled 70% ethanol, and resuspended in sterile deionized water to make the final volume 700 µl. The DNA solution so obtained was purified with MoBio DNA isolation kit, and the steps from the treatment of C4 solution onwards were followed according to the procedure of the kit. DNA was stored at –20°C till sending for whole-metagenome sequencing to the service provider.

Whole-metagenome sequencing

The whole-genome sequencing of our metagenomic DNA was outsourced to the AgriGenome Labs Pvt. Ltd., Kerala, India, and sequencing was done in Illumina HiSeq 2500 Platform. The quality of the DNA was confirmed in Qubit Fluorometer and agarose gel electrophoresis before the library preparation. The Genomic DNA was fragmented using Covaris M220 for 500bp, and the library was prepared using NEBNextUltra DNA Library Prep Kit. The library quality was checked using Agilent Tapestation 2200. The quantity was estimated using Qubit 2.0. The libraries were sequenced in the HiSeq 2500 platform for 2x250bp read length generating the required data. The FASTQ files generated by the Illumina HiSeq platform were trimmed with MetaSPAdes (v 3.10.1) [60] (Version 1.8.1) to remove the adapters.

Sequence assembly, annotation, and evaluation of potential metabolic modules

The forward and reverse DNA sequences in FASTQ format were submitted to the MAPLE Submission Data Maker (MSDM) pipeline where the nucleotide sequences with a minimum base quality score of Q20, minimum 80% of quality bases in each sequence, and minimum 400 bp in each contig length were set to get high-quality amino acid (AA) sequences with a minimum cut-off length of 50 AA in FASTA format [35]. The AA sequence file so obtained was then analyzed in Genomaple ver. 2.3.2 (formerly MAPLE) server opting

for the GHOST X search engine with the single-direction best hit annotation for all organisms in KEGG [33,34]. The module completion ratio (MCR) and *Q*-value at the individual taxonomic rank (ITR) and the whole microbial community (WC) level were retrieved from MAPLE results, and the KEGG Orthology (KO) genes assigned by Genomaple were used for subsequent taxonomic binning, CAZyme and BPZyme analyses. The module information generated by MAPLE was used to create the coarse-grained metabolic maps of KEGG modules using the KEGG Atlas map as a reference. The metabolic map for carbohydrate and lipid metabolism was created separately for the modules having less than 0.5 *Q*-values to understand the biomass degradation and biofuel production potentiality in *emao*.

Taxonomic binning

The KOs assigned to the ribosomal protein module for all organisms (M91000) was extracted back from the MAPLE input file using NCBI-blast dbcmd command, which was then subjected to homology search against the non-redundant NCBI-nr protein database to assign the taxonomic identity for each sequence in GHOSTX .program [61] using the top hit option. GHOSTX result was manually curated to parse the species name against each sequence, which was then meganized and visualized in MEGAN [62] Community Version (V6.12.5). The identified species names were uploaded to the NCBI Tree Viewer (https://www.ncbi.nlm.nih.gov/projects/treeview/) to generate the circular phylogenetic tree.

CAZyme identification

The AA sequences created by the MSDM pipeline (version 1.0) [35] were used as query files in HMMR hmmscan program (version 3.2.1) [63] against dbCAN database [64] as a reference with an E-value threshold of 1e-5 to predict CAZymes. The AA sequences detected as CAZymes [65] were retrieved back from the input file, and the duplicates were removed using some basic perl and shell commands to get the non-redundant FASTA sequences. Then the non-redundant sequences so obtained were assigned to different CAZyme categories using the Hotpep program [66].

BPZyme identification

The AA sequences generated by MSDM in FASTA format were used to identify the enzymes involved in biofuel production as a query file in the HMMR phmmer program [63] against BioFuelDB [67] as a reference with an E-value threshold of 1e-5. The sequence homology and the corresponding EC number of non-redundant AA sequences identified as the biofuel producing enzymes were reconfirmed using the GHOSTX homology search (top

hit only) against the BioFuelDB. All the ECs were then segregated into different biofuel categories as classified by Chaudhary *et al*. [67], and the total enzyme counts for each category were obtained.

References

1. Tamang, J.P. Diversity of fermented beverages and alcoholic drinks in *Fermented Foods and Beverages of the World* (ed Tamang J.P. & Kailasapathy, K.) 85–126 (CRC Press, Taylor & Francis Group 2010).
2. Cason, E.D. *et al*. Bacterial and fungal dynamics during the fermentation processes of Sesotho, a traditional beer of Southern Africa. *Frontiers in Microbiology.* **11**, 1451. https://doi.org/10.3389/fmicb.2020.01451 (2020).
3. McLaren, M. R., Willis, A. D. & Callahan, B. J. Consistent and correctable bias in metagenomic sequencing experiments. *Elife.* **8**, e46923. https://doi.org/10.7554/eLife.46923 (2019).
4. Rausch, P. *et al.* Comparative analysis of amplicon and metagenomic sequencing methods reveals key features in the evolution of animal metaorganisms. *Microbiome.* **7**, 133. https://doi.org/10.1186/s40168-019-0743-1 (2019).
5. Mende, D. R., Sunagawa, S., Zeller, G. & Bork, P. Accurate and universal delineation of prokaryotic species. *Nat. Methods.* **10**, 881–884. https://doi.org/10.1038/nmeth.2575 (2013).
6. Martiny, C., High proportions of bacteria are culturable across major biomes. *ISME.* **13**, 2125–2128. https://doi.org/10.1038/s41396-019-0410-3 (2019).
7. Hodgson, H., Kocch, Bodo and Dhimal Tribes (J. Thomas, Baptist Mission Press, Calcutta). https://archive.org/details/in.ernet.dli.2015.93469 (1847)
8. R. G. Latham, *The Natural History of the Varieties of Man* (London: John van Voorst, Paternoster Row). https://archive.org/details/naturalhistoryof00lathuoft (1850)
9. Zhang, M., Yan, S., Pan, W. & Jin, L. Phylogenetic evidence for Sino-Tibetan origin in northern China in the Late Neolithic. *Nature.* **569**, 112–115. https://doi.org/10.1038/s41586-019-1153-z (2019).
10. Taechavasonyoo, A., Thaniyavarn, J., & Yompakdee, C. Identification of the moulds and yeasts characteristic of a superior *Loogpang*, starter of Thai rice-based alcoholic beverage Sato. *As. J. Food Ag-Ind.* **6(01)**, 24–38 (2013).
11. Chen, B., Wu, Q. & Xu, Y. Filamentous fungal diversity and community structure associated with the solid state fermentation of Chinese Maotai-flavor liquor. *Int. J. Food Microbiol.* **179**, 80–84. https://doi.org/10.1016/j.ijfoodmicro.2014.03.011 (2014).
12. Bhardwaj, K. N., Jain, K. K., Kumar, S. & Kuhad, R. C. Microbiological Analyses of Traditional Alcoholic Beverage (Chhang) and its Starter (Balma) Prepared by Bhotiya Tribe of Uttarakhand, India. *Indian J. Microbiol.* **56(1)**, 28–34. https://doi.org/10.1007/s12088-015-0560-6 (2016)
13. Lv, X. C., Weng, X., Zhang, W., Rao, P. F. & Ni, L. Microbial diversity of traditional fermentation starters for Hong Qu glutinous rice wine as determined by PCR-

mediated DGGE. *Food Control* **28**, 426–434. https://doi.org/10.1016/j.foodcont.2012.05.025 (2012)

14. Ahmadsah, L. S. F., Kim, E., Jung, Y. S. & Kim, H. Y. Identification of LAB and fungi in Laru, a fermentation starter, by PCR-DGGE, SDS-PAGE, and MALDI-TOF MS. *J. Microbiol. Biotechnol.* **28(1)**, 32–39. https://doi.org/10.4014/jmb.1705.05044 (2018)
15. Pradhan, P. & Tamang, J. P. Phenotypic and Genotypic Identification of Bacteria Isolated from Traditionally Prepared Dry Starters of the Eastern Himalayas. *Front. Microbiol.* **10**, 1–15. https://doi.org/10.3389/fmicb.2019.02526 (2019).
16. Anupma, A. & Tamang, J. P. Diversity of Filamentous Fungi Isolated From Some Amylase and Alcohol-Producing Starters of India. *Front. Microbiol.* **11**, 1–16. https://doi.org/10.3389/fmicb.2020.00905 (2020)
17. Thanh, V. N., Mai, L. T. & Tuan, D. A. Microbial diversity of traditional Vietnamese alcohol fermentation starters (banh men) as determined by PCR-mediated DGGE. *Int. J. Food Microbiol.* **128**, 268–273. https://doi.org/10.1016/j.ijfoodmicro.2008.08.020 (2008)
18. Zheng, X. W. *et al.* Complex microbiota of a Chinese 'Fen' liquor fermentation starter (Fen-Daqu), revealed by culture-dependent and culture-independent methods. *Food Microbiol.* ***31***, 293–300. https://doi.org/10.1016/j.fm.2012.03.008 (2012).
19. Lv, X. C. *et al.* Microbial community structure and dynamics during the traditional brewing of Fuzhou Hong Qu glutinous rice wine as determined by culture-dependent and culture-independent techniques. *Food Control.* **57**, 216–224. https://doi.org/10.1016/j.foodcont.2015.03.054 (2015)
20. Bal, J., Yun, S. H., Yeo, S. H., Kim, J. M. & Kim, D. H. Metagenomic analysis of fungal diversity in Korean traditional wheat-based fermentation starter nuruk. *Food Microbiol.* **60**, 73–83. https://doi.org/10.1016/j.fm.2016.07.002 (2016).
21. Bora, S. S., Keot, J., Das, S., Sarma, K. & Barooah, M. Metagenomics analysis of microbial communities associated with a traditional rice wine starter culture (Xaj-pitha) of Assam, India. *3 Biotech.* **6**, 153. https://doi.org/10.1007/s13205-016-0471-1 (2016).
22. Hui, W. *et al.* Identification of Microbial Profile of Koji Using Single Molecule, Real-Time Sequencing Technology. *J. Food Sci.* **82(5),** 1193-1199. https://doi.org/10.1111/1750-3841.13699 (2017).
23. Sha, S. P. *et al.* Diversity of yeasts and molds by culture-dependent and culture-independent methods for mycobiome surveillance of traditionally prepared dried starters for the production of Indian alcoholic beverages. *Front. Microbiol.* **9**, 2237 https://doi.org/10.3389/fmicb.2018.02237 (2018).
24. Buragohain, A. K., Tanti, B., Sarma, H. K., Barman, P. & Das, K. Characterization of yeast starter cultures used in household alcoholic beverage preparation by a few ethnic communities of Northeast India. *Ann. Microbiol.* **63**, 863–869. https://doi.org/10.1007/s13213-012-0537-1 (2013).
25. Song, S. H. *et al.* Analysis of microflora profile in Korean traditional Nuruk. *J. Microbiol. Biotechnol.* **23(1)**, 40–46. https://doi.org/10.4014/jmb.1210.10001 (2013).

26. Das, A. J., Miyaji, T. & Deka, S. C. Amylolytic fungi in starter cakes for rice beer production. *J. Gen. Appl. Microbiol.* **63(4)**, 236–245. https://doi.org/10.2323/jgam.2016.11.004 (2017).
27. Parasar, D. P., Sarma, H. K. & Kotoky, J. Exploring the genealogy and phenomic divergences of indigenous domesticated yeasts cultivated by six ethnic communities of Assam, India. *J. Biol. Sci.* **17**, 91–105. https://doi.org/10.3923/jbs.2017.91.105 (2017).
28. Nath, N., Ghosh, S., Rahaman, L., Kaipeng, D. L. & Sharma, B. K. An overview of traditional rice beer of north-east india: Ethnic preparation, challenges and prospects. *Indian J. Tradit. Knowl.* **18(4)**, 744–757 (2019).
29. Kwon S.J. & Sohn J.H., Analysis of microbial diversity in Nuruk using PCR-DGGE. *Journal of Life Science.* 22(1), 110–116. https://doi.org/10.5352/jls.2012.22.1.110 (2012).
30. Sha, S. P. *et al.* Analysis of bacterial and fungal communities in Marcha and Thiat, traditionally prepared amylolytic starters of India. *Sci. Rep.* 10.1038/s41598-017-11609-y. https://doi.org/10.1038/s41598-017-11609-y (2017).
31. Janusz, G. *et al.* Lignin degradation: Microorganisms, enzymes involved, genomes analysis and evolution. *FEMS Microbiology Reviews.* **41(6)**, 941–962 https://doi.org/10.1093/femsre/fux049G (2017).
32. Bredon, M., Dittmer, J., Noël, C., Moumen, B. & Bouchon, D. Lignocellulose degradation at the holobiont level: teamwork in a keystone soil invertebrate. *Microbiome.* **6**, 162. https://doi.org/10.1186/s40168-018-0536-y (2018).
33. Takami, H. *et al.* An automated system for evaluation of the potential functionome: MAPLE version 2.1.0. *DNA Res.* **23**, 467–475. https://doi.org/10.1093/dnares/dsw030 (2016).
34. Arai, W. *et al.* Maple 2.3.0: An improved system for evaluating the functionomes of genomes and metagenomes. *Biosci. Biotechnol. Biochem.* **82**, 1515–1517. https://doi.org/10.1080/09168451.2018.1476122 (2018)
35. Takami, H. MAPLE enables functional assessment of microbiota in various environments in *Marine metagenomics-Technological aspects and applications.* (ed Gojobori, T, Wada, T, Kobayashi, T, Mineta, K.) 85–119 (Springer, Singapore, 2019).
36. Courchesne, N. M. D., Parisien, A., Wang, B. & Lan, C. Q. Enhancement of lipid production using biochemical, genetic and transcription factor engineering approaches. *Journal of Biotechnology* **141**, 31–41. https://doi.org/10.1016/j.jbiotec.2009.02.018 (2009).
37. Phulara, S. C., Chaturvedi, P. & Gupta, P. Isoprenoid-based biofuels: Homologous expression and heterologous expression in prokaryotes. *Applied and Environmental Microbiology* . **82(19)**, 5730–5740. https://doi.org/10.1128/AEM.01192-16 (2016).
38. Dung, N. T. P., Rombouts, F. M. & Nout, M. J. R. Functionality of selected strains of moulds and yeasts from Vietnamese rice wine starters. *Food Microbiol.* **23(4),** 331–340. https://doi.org/10.1016/j.fm.2005.05.002 (2006).
39. Limtong, S., Sintara, S. & Suwannarit, P. Yeast Diversity in Thai Traditional Alcoholic Starter (Loog-Pang). *Kasetsart J. (Nat. Sci.)* **36,** 149–158 (2002).

40. Pina, C., Antonio, J. & Hogg, T. Inferring ethanol tolerance of Saccharomyces and non-Saccharomyces yeasts by progressive inactivation. *Biotechnol. Lett.* **26**, 1521–1527. https://doi.org/10.1007/s10529-005-1787-9 (2005).
41. Pina, C., Santos, C., Couto, J. A. & Hogg, T. Ethanol tolerance of five non-*Saccharomyces* wine yeasts in comparison with a strain of *Saccharomyces cerevisiae* - Influence of different culture conditions. *Food Microbiol.* **21(4)**, 439–447. https://doi.org/10.1016/j.fm.2003.10.009 (2004).
42. Jamai, L., Ettayebi, K., Yamani, J. El & Ettayebi, M. Production of ethanol from starch by free and immobilized Candida tropicalis in the presence of α-amylase. *Bioresour. Technol.* **98(14)**, 2765–2770. https://doi.org/10.1016/j.biortech.2006.09.057 (2007).
43. Steensels, J. & Verstrepen, K. J. Taming wild yeast: Potential of conventional and nonconventional yeasts in industrial fermentations. *Annual Review of Microbiology* **68**, 61–80. https://doi.org/10.1146/annurev-micro-091213-113025 (2014).
44. Casey, G. P. & Ingledew, W. M. M. Ethanol tolerance in yeasts. *Crit. Rev. Microbiol.* (1986). **13(3)**, 219–280. https://doi.org/10.3109/10408418609108739 (1986).
45. Ghareib, M., Youssef, K. A. & Khalil, A. A. Ethanol tolerance of *Saccharomyces cerevisiae* and its relationship to lipid content and composition. *Folia Microbiol. (Praha).* **33**, 447–452. https://doi.org/10.1007/BF02925769 (1988).
46. Riles, L. & Fay, J. C. Genetic basis of variation in heat and ethanol tolerance in *Saccharomyces cerevisiae*. *G3 Genes, Genomes, Gene* **9(1),** 179–188. https://doi.org/10.1534/g3.118.200566 (2019).
47. Yamada, R. *et al.* Direct and efficient ethanol production from high-yielding rice using a Saccharomyces cerevisiae strain that express amylases. *Enzyme Microb. Technol.* https://doi.org/10.1016/j.enzmictec.2011.01.002 (2011)
48. Hollerová, I. & Kubizniaková, P. Monitoring Gram positive bacterial contamination in Czech breweries. *J. Inst. Brew*. **107**, 355–358. https://doi.org/10.1002/j.2050-0416.2001.tb00104.x (2001).
49. Bokulich, N. A. & Bamforth, C. W. The Microbiology of Malting and Brewing. *Microbiol. Mol. Biol. Rev.* **77(2)**, 157–172. https://doi.org/10.1016/B978-0-12-809633-8.13014-6 (2013).
50. Ennahar, S., Sonomoto, K. & Ishizaki, A. Class IIa bacteriocins from lactic acid bacteria: Antibacterial activity and food preservation. *Journal of Bioscience and Bioengineering* **87**, 705–716. https://doi.org/10.1016/S1389-1723(99)80142-X (1999).
51. Mokoena, M. P. Lactic acid bacteria and their bacteriocins: Classification, biosynthesis and applications against uropathogens: A mini-review. *Molecules* **22(8)**, 1255. https://doi.org/10.3390/molecules22081255 (2017)
52. Ghosh, S. *et al.* Community-wise evaluation of rice beer prepared by some ethnic tribes of Tripura. *J. Ethn. Foods.* https://doi.org/10.1016/j.jef.2016.12.001 (2016).
53. M.Miller, *Monitoring acids and pH in wine making*. eBook. http://www.gencowinemakers.com/docs/Acids%20Presentation.pdf (2019)

54. Sakamoto, K. & Konings, W. N. Beer spoilage bacteria and hop resistance. *Int. J. Food Microbiol.* **89(2)**, 105–124. https://doi.org/10.1016/S0168-1605(03)00153-3 (2003).
55. Vriesekoop, F., Krahl, M., Hucker, B. & Menz, G. 125th Anniversary review: Bacteria in brewing: The good, the bad and the ugly. *J. Inst. Brew.* **118**, 335–345. https://doi.org/10.1002/jib.49 (2012).
56. Cadete, R. M. & Rosa, C. A. The yeasts of the genus Spathaspora: potential candidates for second-generation biofuel production. *Yeast* **35**, 191–199. https://doi.org/10.1002/yea.3279 (2018).
57. Gryganskyi, A. P. & Muszewska, A. Whole genome sequencing and the Zygomycota. *Fungal Genom. Biol.* **4**, 1. https://doi.org/10.4172/2165-8056.1000e116 (2014).
58. Zhou, J., Bruns M. A & Tiedje, J. M. DNA recovery from soils of diverse composition. *Appl. Environ. Microbiol.* **62**, 316–322. https://doi.org/10.1128/aem.62.2.316-322.1996 (1996)
59. Sharma, S., Sharma K. K., & Kuhad, R. C. An efficient and economical method for extraction of DNA amenable to biotechnological manipulations, from diverse soils and sediments. *J. Appl. Microbiol*. **116**, 923–933. https://doi.org/10.1111/jam.12420 (2013)
60. Nurk, S., Meleshko, D., Korobeynikov, A. &. Pevzner, P.A. metaSPAdes: a new versatile metagenomic assembler. *Genome Research.* **27(5)**, 824–834. https://doi.org/10.1101/gr.213959.116 (2017).
61. Suzuki, S., Kakuta, M. T. Ishida & Y. Akiyama, GHOSTX: An Improved Sequence Homology Search Algorithm Using a Query Suffix Array and a Database Suffix Array. *PLoS ONE* **9(8)**, e103833. https://doi.org/10.1371/journal.pone.0103833 (2014).
62. Huson, D.H. *et al.* MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data, *PLoS Comput. Biol.* https://doi.org/10.1371/journal.pcbi.1004957 (2016).
63. Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A. & Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* **41(12)**, e121. https://doi.org/10.1093/nar/gkt263 (2013).
64. Yin, Y. *et al.* DbCAN: A web resource for automated carbohydrate-active enzyme annotation, *Nucleic Acids Res.* https://doi.org/10.1093/nar/gks479 (2012).
65. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* **42**, D490–D495. https://doi.org/10.1093/nar/gkt1178 (2014).
66. Busk, P. K., Pilgaard, B., Lezyk, M. J., Meyer, A. S. & Lange, L. Homology to peptide pattern for annotation of carbohydrate-active enzymes and prediction of function. *BMC Bioinformatics* **18**, 214. https://doi.org/10.1186/s12859-017-1625-9 (2017).
67. Chaudhary, N., Gupta, A., Gupta, S. & Sharma, V. K. BioFuelDB: A database and prediction server of enzymes involved in biofuels production. *PeerJ* **5**, e3497. https://doi.org/10.7717/peerj.3497 (2017).

Acknowledgments

The authors thank T.S. Ran and J.S. Chandrani for their valuable feedback and suggestions. This research was supported by the DST, Govt of India under the SERB scheme (SB/EMEQ-443/2014) to D.N. We thank the Department of Botany, Gauhati University, Guwahati and the Yokohama Institute (JAMSTEC), Japan for research facilities. DN is also thankful to the DBT, Govt of India for financial support to set up a sophisticated research laboratory under the Unit of Excellence scheme (BT/408/NE/U-Excel/2013), and also for overseas fellowship to visit JAMSTEC for data analysis under the DBT-Associateship program. The authors are thankful to the local Bodo women who shared their traditional material and information for our research.

Authors' contributions

D.N. conceived and designed experiments. N.B. and A.B. contributed to sample collection and DNA isolation. H.T. directed metagenomic data analysis. D.N. and O.T. performed computational work. D.N. wrote and H.T. revised the manuscript.

Competing interests The authors declare no competing interests.

Supplementary data

Table S1: Species diversity and comparison of *emao* to other traditional starters

Table S2: Metabolic modules predicted from *emao* metagenome in MAPLE system

Table S3: List of potential lignocellulose/pentose metabolizing microbial species identified in *emao* as inferred from cross-references

Fig. S1: CAZyme sub-families as recorded in *emao*

Fig. S2: BPZymes as recorded in '*emao*'

Data availability

Whole-metagenome amino acid sequence data (DNS1), MAPLE assigned KOs (DNS2), KEGG metabolic modules (DNS3), and ribosomal protein-based community structure information in RMA format (DNS4) are available through figshare (DOI: 10.6084/m9.figshare.8868689). Any other relevant data are available from the corresponding author upon reasonable request.

Figures

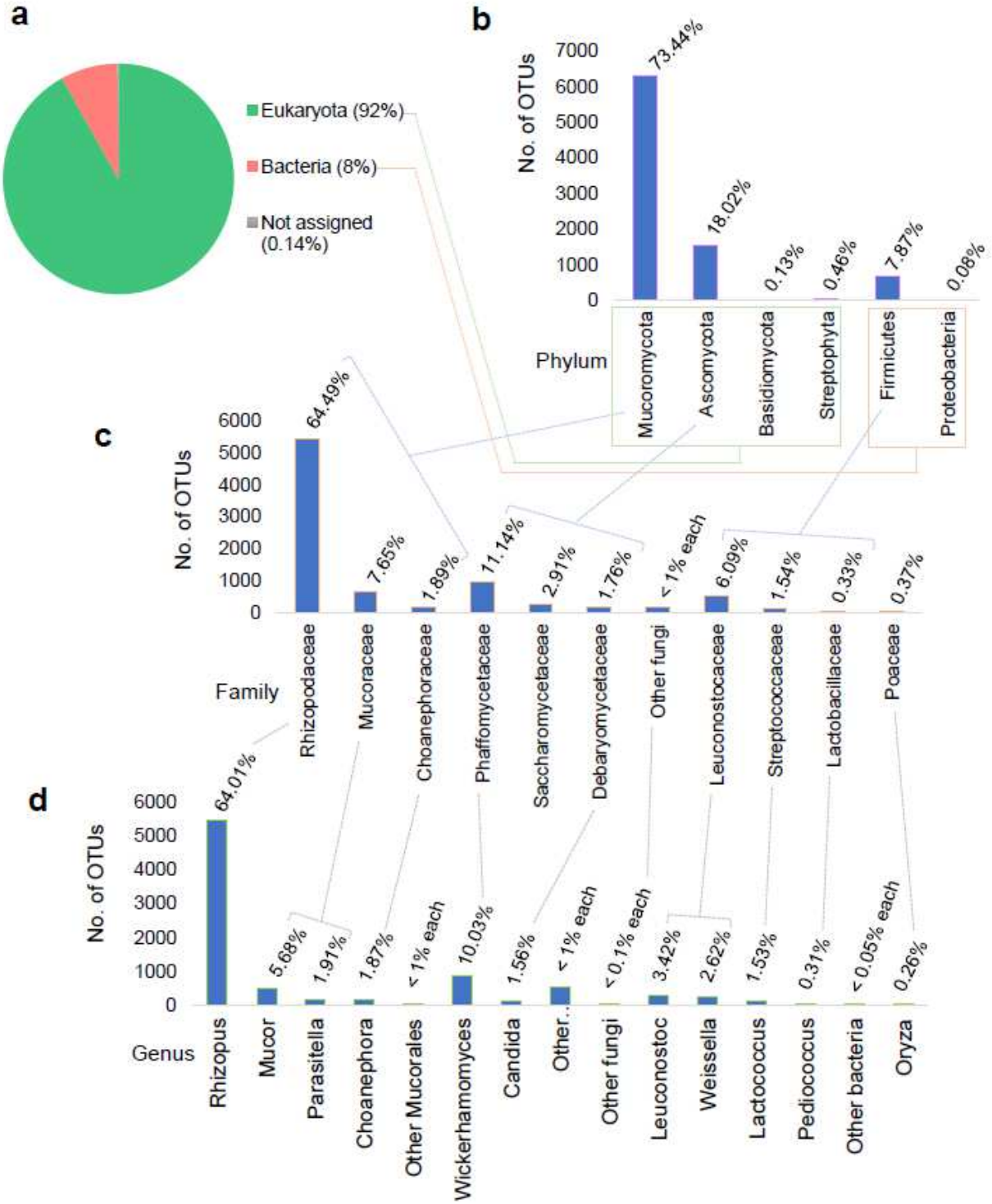

Figure 1

Community structure estimation in emao based on ribosomal protein sequence analysis. The ratio of molds, yeasts and bacteria was 9.2:2:1. a, At Domain level; b, At Phylum level; c, At Family level; d, At Genus level

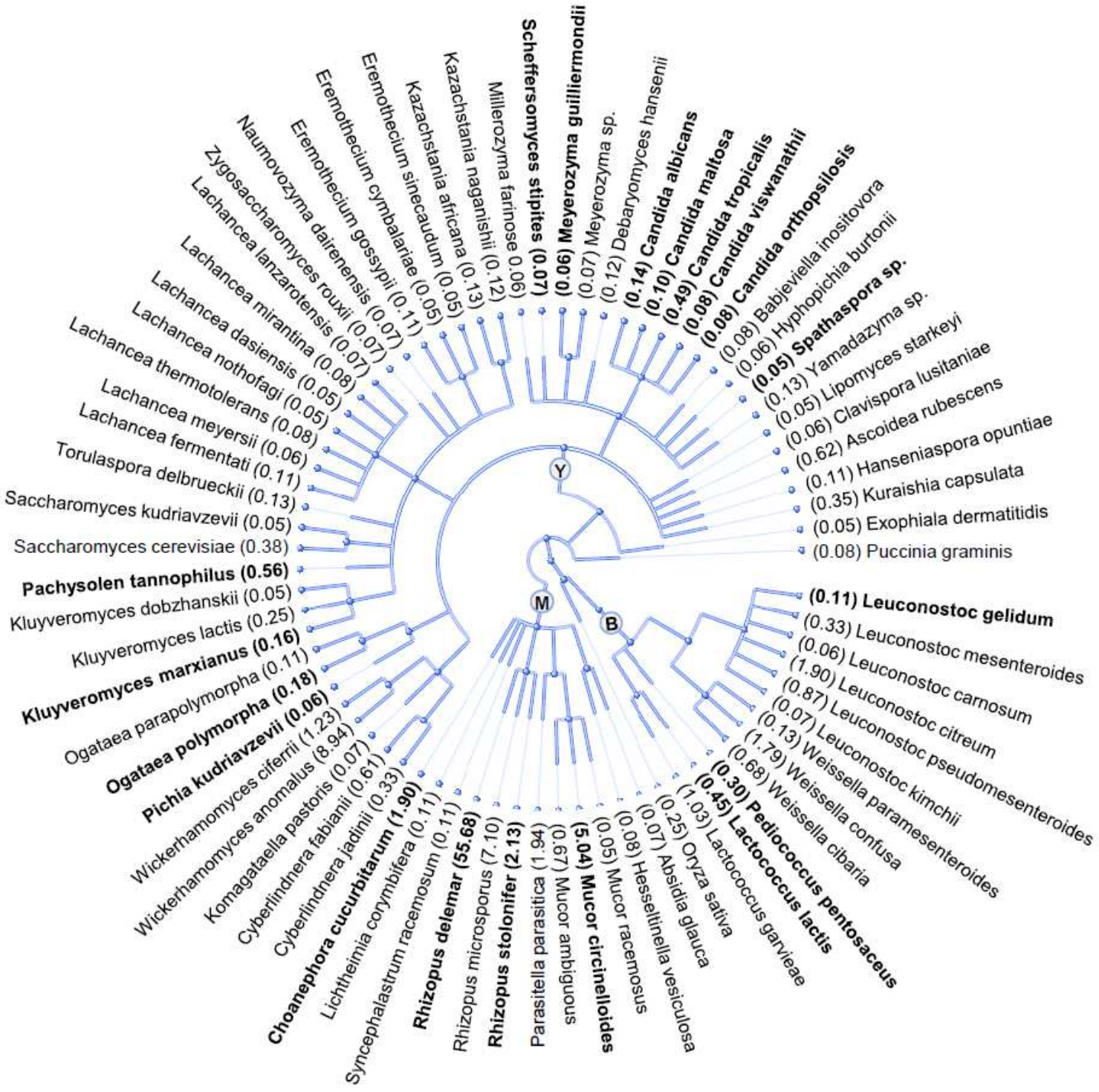

Figure 2

Phylogenetic tree of species as recorded in emao. Total 74 microbial species and 1 rice species were identified based on ribosomal protein sequence homology search against NCBI-nr database. Values within bracket indicate relative abundance in percent. Bold-faced species (total 19) indicate lignocellulose/pentose metabolizer (for cross references see Extended Data Table S3). B, Bacteria; M, Molds; Y, Yeasts.

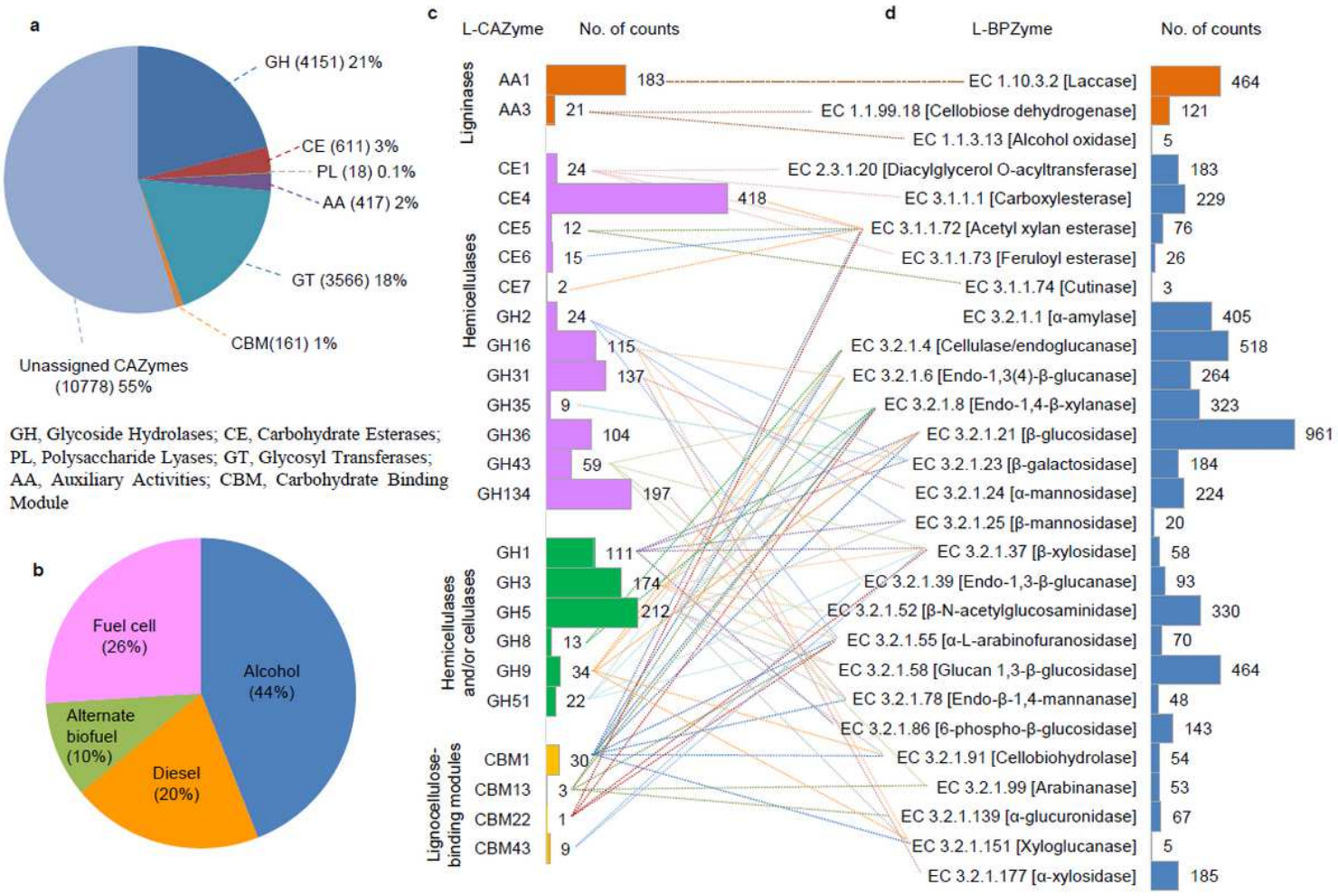

Figure 3

Carbohydrate active enzymes (CAZymes) and biofuel producing enzymes (BPZymes) predicted in emao. Complete list of CAZymes and BPZymes recorded in emao is available in Extended Data Fig. S1 & S2 respectively. a, Pie chart indicating percentage of CAZyme families assigned by Hotpep. Value within brackets is total count. b, Relative percentage of BPZyme categories as classified in BioFuelDB. c, Lignocellulolytic CAZyme sub-families segregated activity-wise. d, Lignocellulolytic BPZymes with Enzyme Commission (EC) number extracted from b and correlated to c (dotted lines).

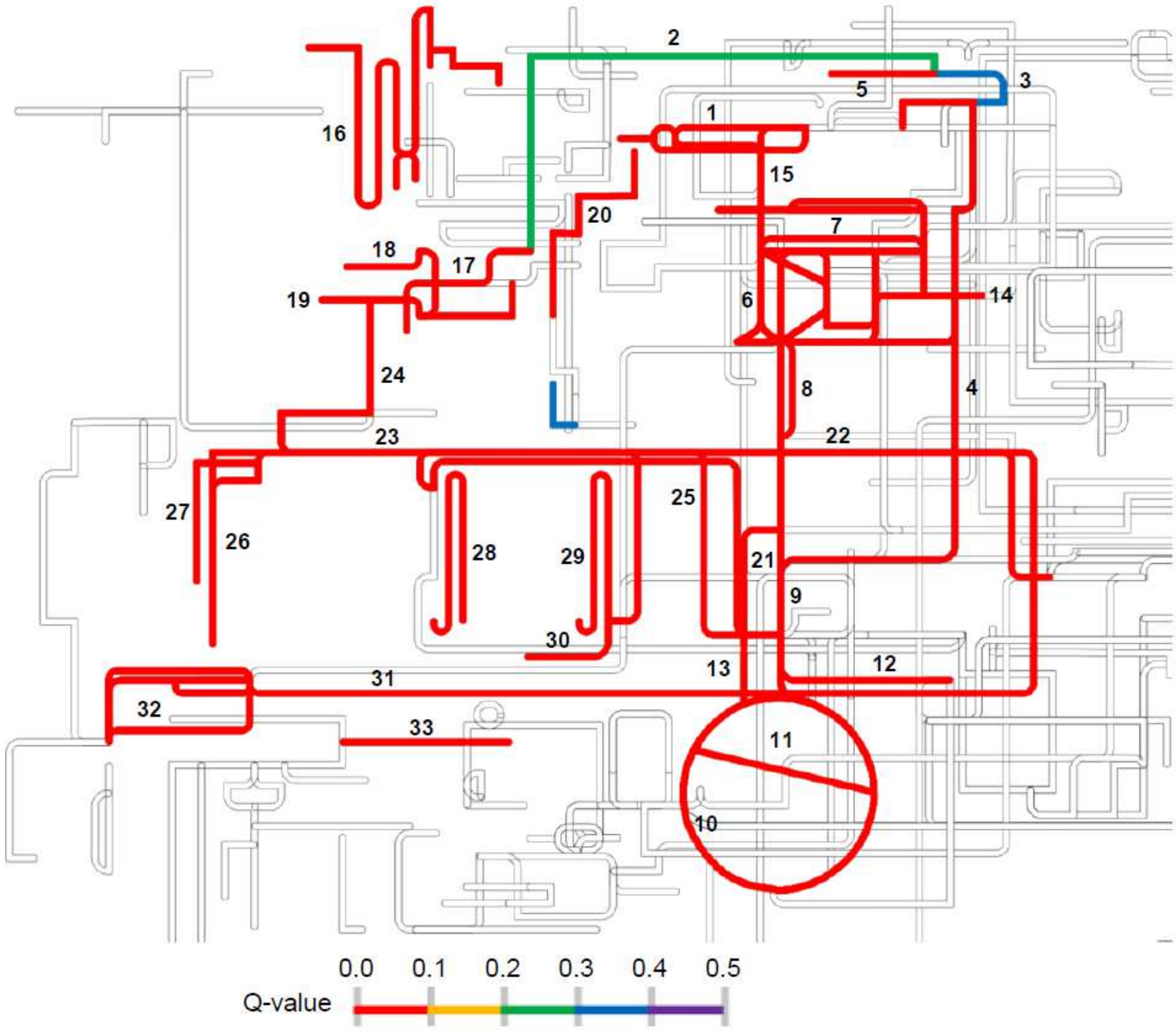

Figure 4

Carbohydrate and lipid metabolic maps predicted in emao. Metabolic modules with Q-value less than 0.5 for whole microbial communities (WC) were shown in the map (Extended Data Table 3). The probability of occurrence of a metabolic module increases with decreasing Q-value as the latter is zero if all the genes necessary for a module is complete19. This coarse-grained map was created using the KEGG Atlas map. 1, Galactose degradation; 2, D-galacturonate degradation (fungi); 3, D-galacturonate degradation (bacteria); 4, D-glucuronate degradation; 5, Pectin degradation; 6, Embden-Meyerhof Pathway (glycolysis); 7, Pentose phosphate pathway; 8, Glycolysis core module; 9, Pyruvate oxidation; 10, Citrate cycle; 11, Glyoxylate cycle; 12, Melanoate semialdehyde pathway; 13, Gluconeogenesis; 14, PRPP biosynthesis; 15, Nucleotide sugar biosynthesis; 16, N-glycan metabolism; 17, Acyl glycerol degradation; 18, Phosphatidylcholine biosynthesis; 19, Phosphatidylethanolamine biosynthesis; 20, GPI-anchor

biosynthesis, core oligosaccharide; 21, Inositol phosphate metabolism; 22, Ceramide biosynthesis; 23, Sphingosine biosynthesis; 24, Sphingosine degradation; 25, Fatty acid biosynthesis, initiation; 26, Fatty acid biosynthesis, elongation; 27, Fatty acid biosynthesis, elongation (ER); 28, Fatty acid biosynthesis, elongation (mitochondria); 29, Beta-oxidation; 30, Beta-oxidation acyl-CoA synthesis; 31, C5 isoprenoid biosynthesis (Mevalonate); 32, C10-20 isoprenoid biosynthesis; 33, Ergocalciferol biosynthesis.

Supplementary Files

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

- Table1Narzaryetal..pdf
- Table2Narzaryetal..pdf
- Table3Narzaryetal..pdf
- SupplementaryDataFig.S1Narzaryetal..pdf
- SupplementaryDataFig.S2Narzaryetal..pdf
- SupplementaryDataTableS1Narzaryetal..xlsx
- SupplementaryDataTableS2Narzaryetal..xls
- SupplementaryDataTableS3Narzaryetal..pdf
- SupplementaryDataFig.S1Narzaryetal..pdf
- SupplementaryDataFig.S2Narzaryetal..pdf
- SupplementaryDataTableS1Narzaryetal..xlsx
- SupplementaryDataTableS2Narzaryetal..xls
- SupplementaryDataTableS3Narzaryetal..pdf