Supplemental Information

Assessment of enzyme diversity in the fermented food microbiome

Peng Li[1, #], Jingyu Sun[1, #], Yu Geng[1], Yiru Jiang[1], Yue-zhong Li[1, *], Zheng Zhang[1, *]

[1] State Key Laboratory of Microbial Technology, Institute of Microbial Technology, Shandong University, Qingdao 266237, China

[#] These authors contributed equally: Peng Li, Jingyu Sun.

[*] Address correspondence to Yue-zhong Li (E-mail: lilab@sdu.edu.cn, ORCID: 0000-0001-8336-6638) or Zheng Zhang (E-mail: zhangzheng@sdu.edu.cn, ORCID: 0000-0001-9971-6006)

This PDF file includes the following:

Figures S1-S6

Legends for Data S1-S4

Other supplementary materials for this manuscript include the following:

Data S1-S4

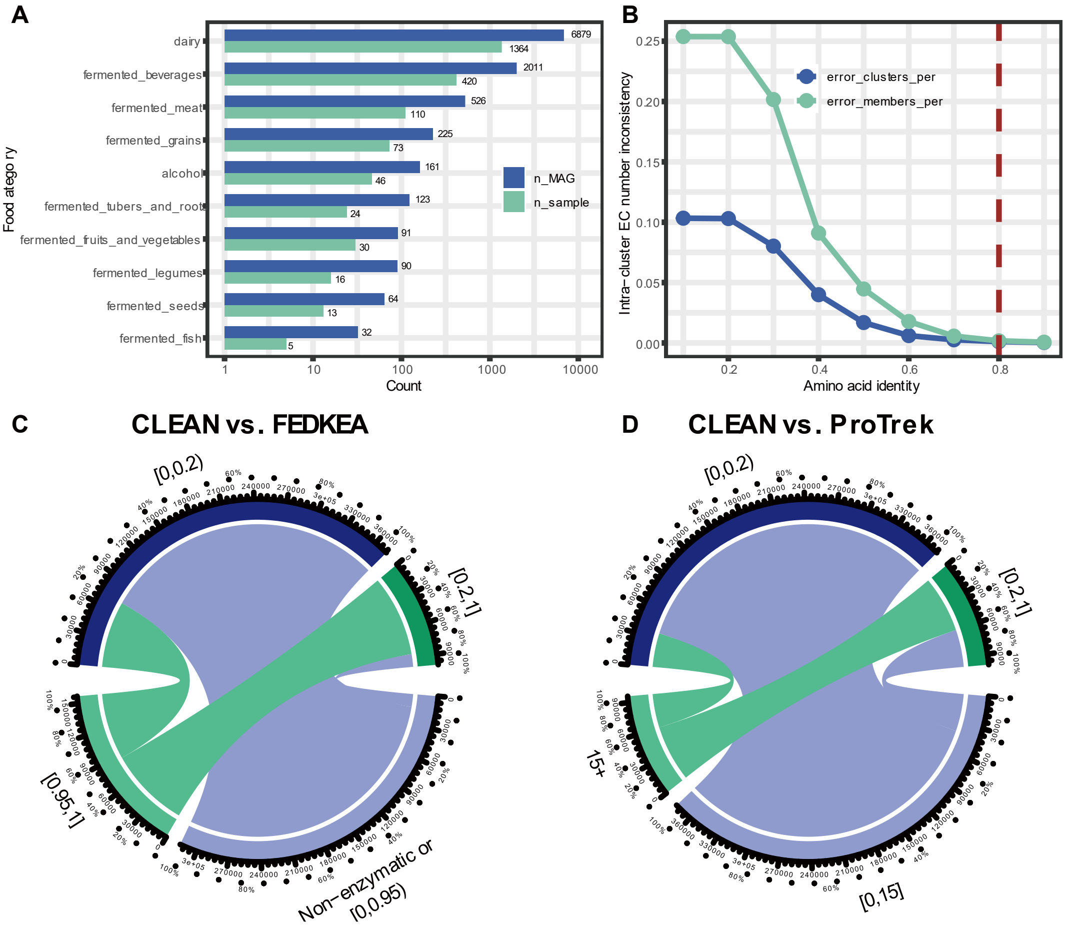

Figure S1: Sources of fermented food metagenome-assembled genomes (MAGs) and validation of enzyme annotation methods.

(A) Distribution of MAGs across different sample sources and food categories (n = number of MAGs).

(B) Error rates of EC assignment for enzyme cluster level and sequence level under varying protein clustering thresholds.

(C, D) Performance comparison of CLEAN, FEDKEA, and ProTrek for annotating 472,428 representative sequences of protein clusters.

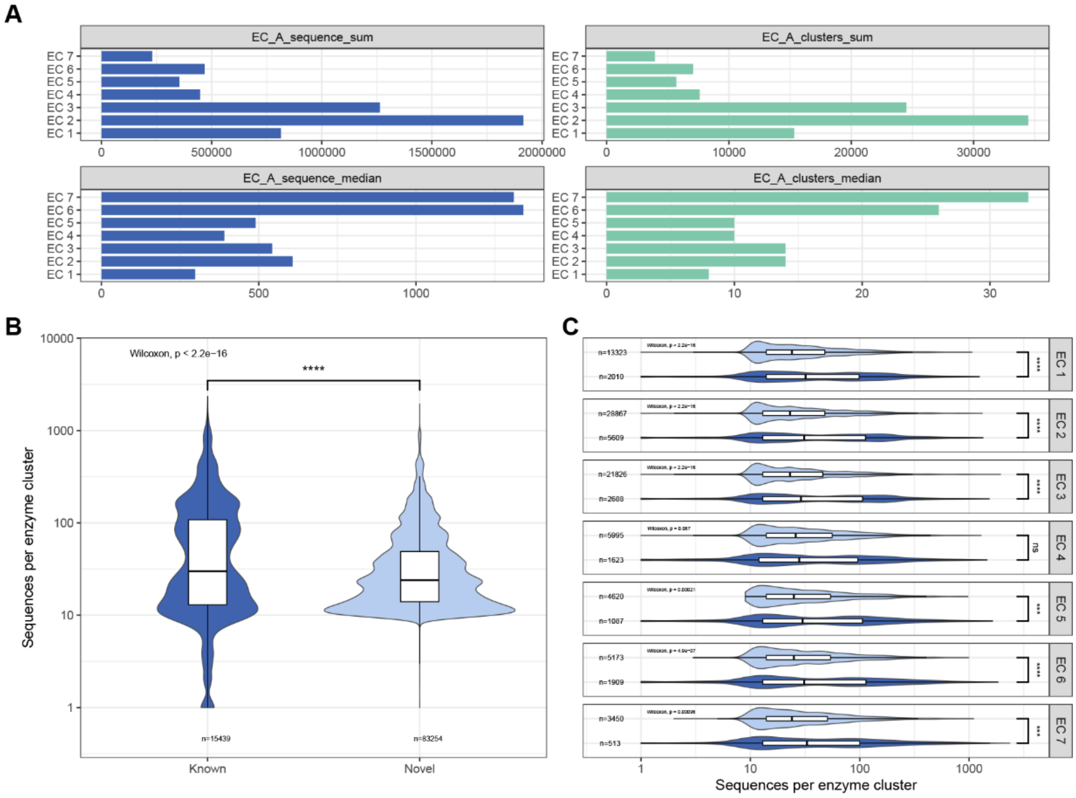

Figure S2: Distribution of enzyme cluster diversity and novelty across enzyme classes.

(A) Total cluster/sequence counts and median size per four-digit EC number within primary enzyme categories.

(B) Cluster size distribution by novelty status (known vs. novel) (n = number of enzyme clusters).

(C) Cluster size distribution stratified by primary EC category and novelty status (n = number of enzyme clusters). Statistical significance was assessed using the Wilcoxon rank-sum test. NS, $P > 0.05$; * $P \leq 0.05$; ** $P \leq 0.01$; *** $P \leq 0.001$; **** $P \leq 0.0001$.

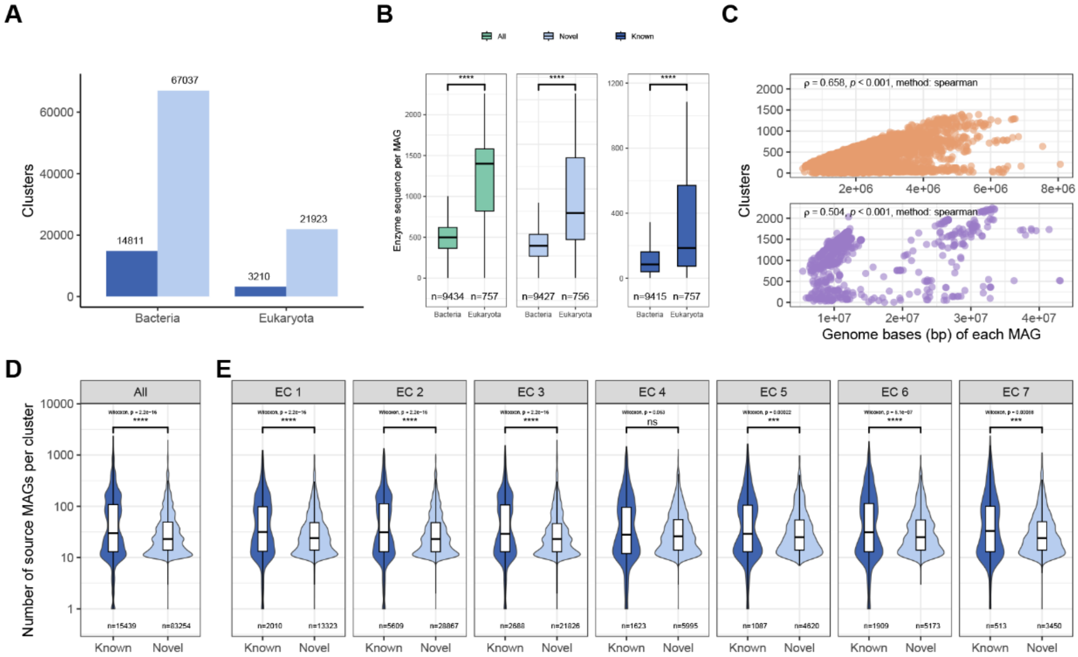

Figure S3: Taxonomic distribution of known and novel enzyme clusters and their relationship with genome size.

(A) Number of known and novel enzyme clusters in bacteria and fungi (n = number of enzyme clusters).

(B) Enzyme coding density of MAGs in bacteria and fungi (n = number of enzyme clusters).

(C) Correlation between genome size (bp) of each MAG and the type of enzyme cluster encoded.

(D) Difference in the number of known or novel MAGs assigned to different species.

(E) Difference in the number of known or novel MAGs assigned to different species for different primary EC categories (n = number of enzyme clusters). Statistical significance was assessed using the Wilcoxon rank-sum test. NS, *P* > 0.05; * *P* ≤ 0.05; ** *P* ≤ 0.01; *** *P* ≤ 0.001; **** *P* ≤ 0.0001.

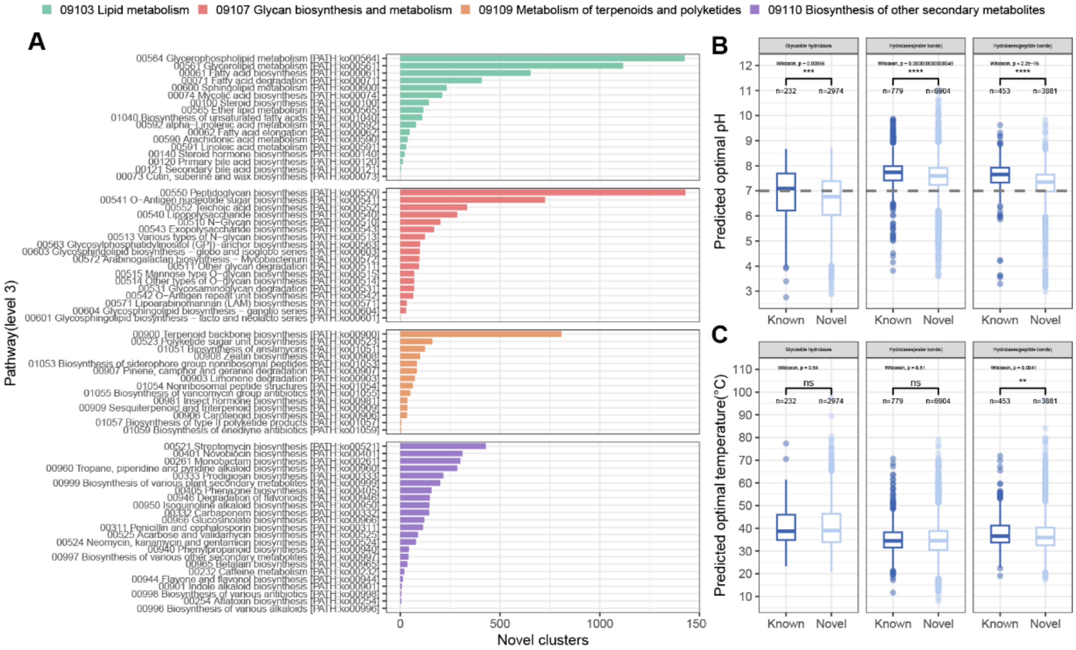

Figure S4: Characterization of known and novel enzyme clusters.

(A) Number of novel enzyme clusters in the biosynthetic pathways of lipids, sugars, terpenes and other secondary metabolisms.

(B) Predicted optimum pH of glycosidic bond, peptide bond and ester bond hydrolases (n = number of enzyme clusters).

(C) Predicted optimum temperature of glycosidic bond, peptide bond and ester bond hydrolases (n = number of enzyme clusters). Statistical significance was assessed using the Wilcoxon rank-sum test. NS, $P > 0.05$; * $P \leq 0.05$; ** $P \leq 0.01$; *** $P \leq 0.001$; **** $P \leq 0.0001$.

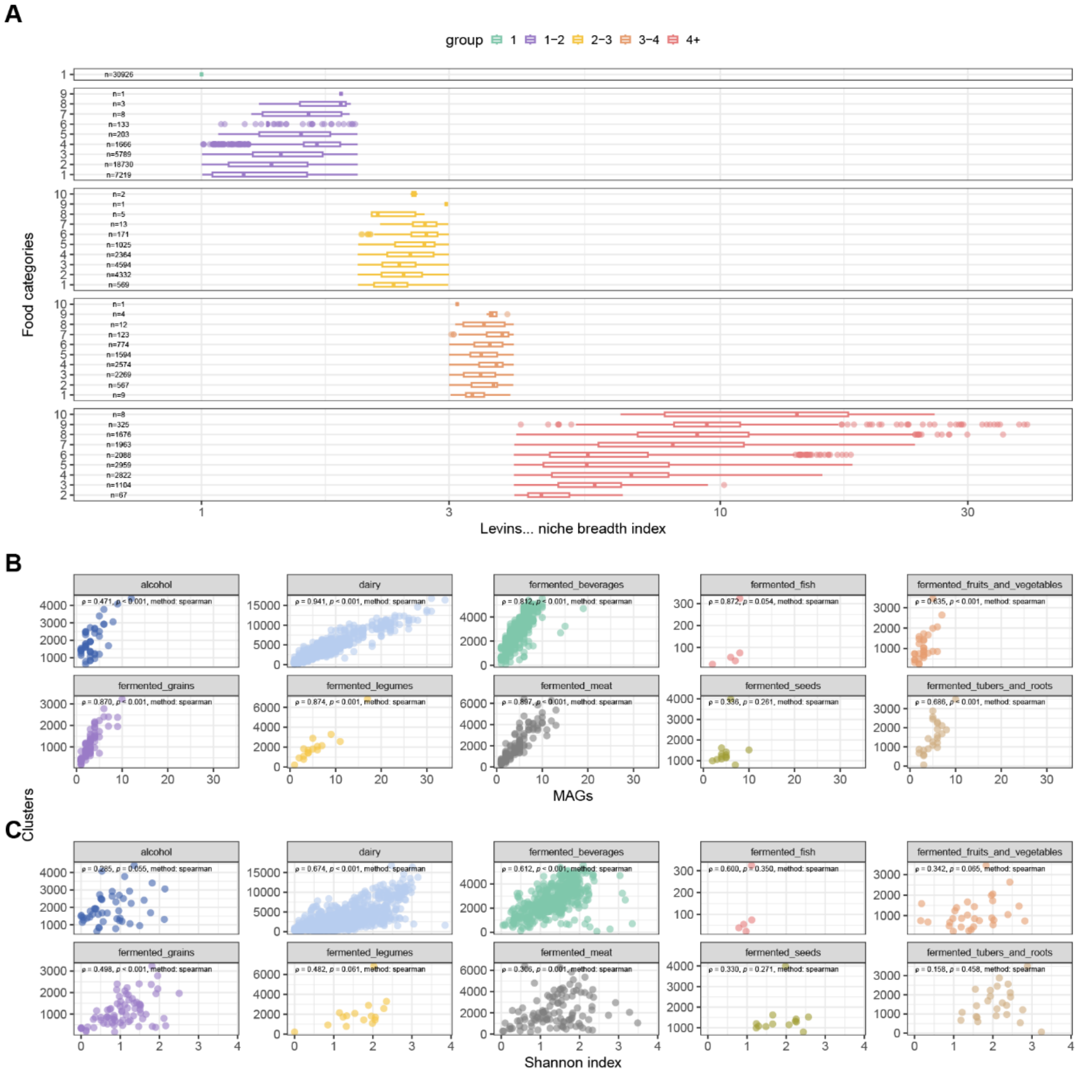

Figure S5: Distribution of microbial enzymes across food categories.

(A) Category distribution of known and novel enzyme clusters in different niche groups (n = number of enzyme clusters).

(B) Correlation between MAGs and the type of enzyme cluster encoded within samples.

(C) Correlation between Shannon index and the type of enzyme cluster encoded within samples.

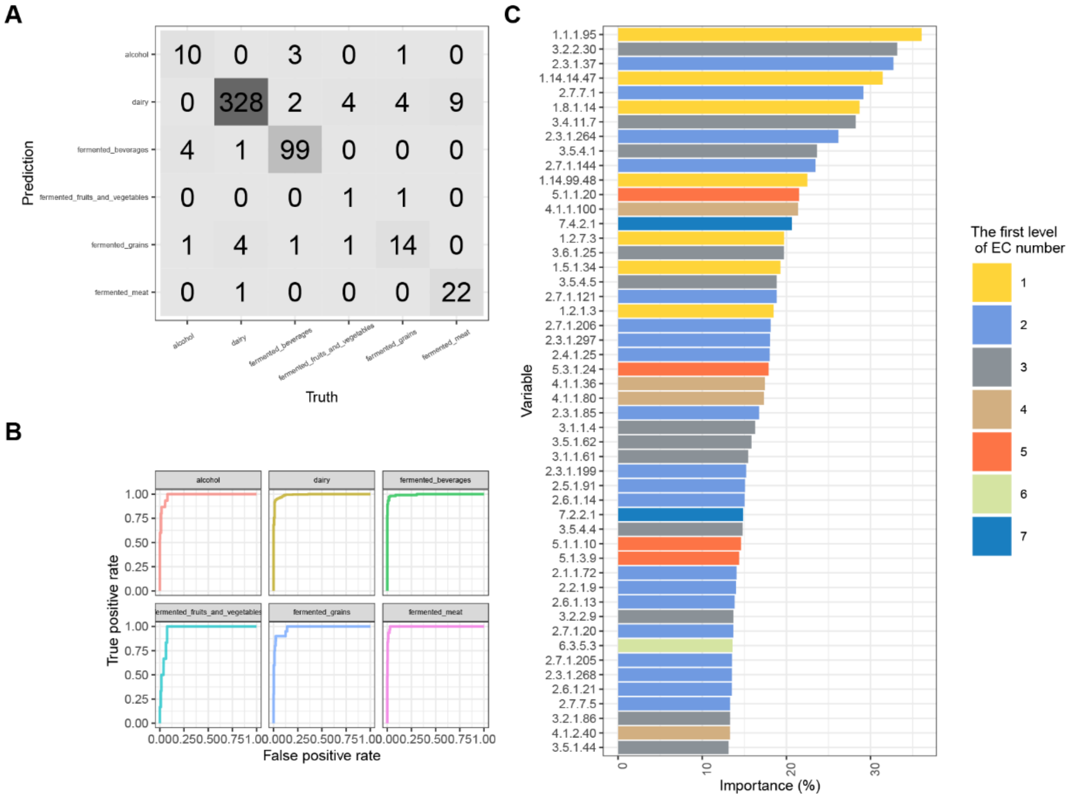

Figure S6: Model performance and variable importance of random forest classification.

(A) Evaluation of the predictive ability of a classifier for food matrix types constructed via the random forest model.

(B) Receiver operating characteristic (ROC) curve area under the curve (AUC) of 0.986 (hand-till method), indicating the relative reliability of the model.

(C) List of model variable importance, classified according to the first level of the number of ECs.

Supplementary Data

Data S1. Information on clusters identified as enzymes.

Data S2. Information on the identification of 3,017 different types of enzymes.

Data S3. Information on the sequence diversity of each enzyme within individual samples.

Data S4. List of variable importance in the machine learning model.