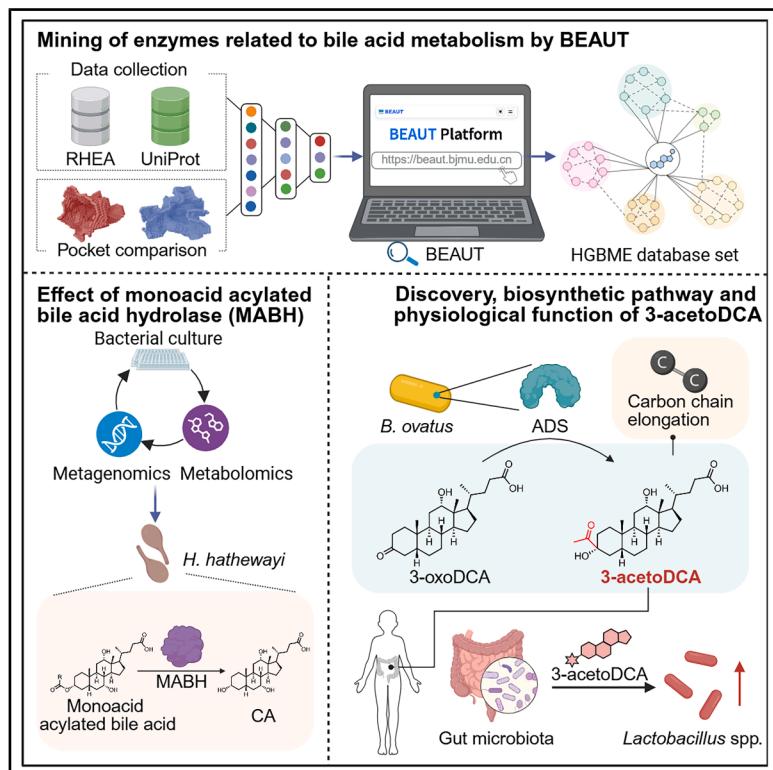


Identification of gut microbial bile acid metabolic enzymes via an AI-assisted pipeline

Graphical abstract



Authors

Yong Ding, Xi Luo, Jiasheng Guo, ..., Ming Ma, Luhua Lai, Changtao Jiang

Correspondence

wangkai@bjmu.edu.cn (K.W.),
mma@bjmu.edu.cn (M.M.),
lhla@pku.edu.cn (L.L.),
jiangchangtao@bjmu.edu.cn (C.J.)

In brief

Ding et al. report an AI-assisted workflow that enables top-down, cultivation-independent identification of human microbial bile acid metabolic enzymes.

Highlights

- AI-assisted pipeline BEAUT enables human microbial bile acid metabolic enzyme mining
- Gut bacteria degrade 3-acylated bile acids by MABH
- Bacterial enzyme ADS produces an unreported skeleton bile acid, 3-acetoDCA
- 3-acetoDCA is widely distributed and may mediate bacterial interactions



Article

Identification of gut microbial bile acid metabolic enzymes via an AI-assisted pipeline

Yong Ding,^{1,2,14} Xi Luo,^{3,14} Jiasheng Guo,^{4,14} Baiying Xing,^{5,14} Haoyu Lin,^{6,14} Haohan Ma,³ Yicun Wang,³ Meng Li,³ Chuan Ye,³ Sen Yan,^{7,8,9} Kangjie Lin,⁴ Jinxin Zhang,³ Yingying Zhuo,³ Qixing Nie,¹⁰ Donghui Yang,⁵ Zhipeng Zhang,¹¹ Yanli Pang,^{7,8,9} Kai Wang,^{1,2,*} Ming Ma,^{5,*} Luhua Lai,^{4,6,12,*} and Changtao Jiang^{1,2,7,13,15,*}

¹Department of Immunology, School of Basic Medical Sciences, State Key Laboratory of Vascular Homeostasis and Remodeling, Peking University, Beijing 100191, China

²NHC Key Laboratory of Medical Immunology, Peking University, Beijing, China

³Department of Physiology and Pathophysiology, Center for Obesity and Metabolic Disease Research, School of Basic Medical Sciences, Peking University, Beijing 100191, China

⁴Center for Life Sciences at BNIMS, College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China

⁵State Key Laboratory of Natural and Biomimetic Drugs, School of Pharmaceutical Sciences, Peking University, Beijing 100191, China

⁶Center for Quantitative Biology, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China

⁷Institute of Advanced Clinical Medicine, State Key Laboratory of Female Fertility Promotion, Center for Reproductive Medicine, Third Hospital, Peking University, Beijing, China

⁸National Clinical Research Center for Obstetrics and Gynecology, Peking University Third Hospital, Beijing, China

⁹Beijing Key Laboratory of Reproductive Endocrinology and Assisted Reproductive Technology, Beijing, China

¹⁰State Key Laboratory of Food Science and Resources, Key Laboratory of Bioactive Polysaccharides of Jiangxi Province, Nanchang University, Nanchang, China

¹¹Department of General Surgery, Peking University Third Hospital, Beijing, China

¹²Peking University Chengdu Academy for Advanced Interdisciplinary Biotechnologies, Chengdu 510100, Sichuan, China

¹³Center of Basic Medical Research, Institute of Medical Innovation and Research, Peking University Third Hospital, Beijing, China

¹⁴These authors contributed equally

¹⁵Lead contact

*Correspondence: wangkai@bjmu.edu.cn (K.W.), mma@bjmu.edu.cn (M.M.), lhlai@pku.edu.cn (L.L.), jiangchangtao@bjmu.edu.cn (C.J.)

<https://doi.org/10.1016/j.cell.2025.07.017>

SUMMARY

The modifications of bile acids (BAs) are fundamental to their role in host physiology and pathology. Identifying their synthetases is crucial for uncovering the diversity of BAs and developing targeted interventions, yet it remains a significant challenge. To address this hurdle, we developed an artificial intelligence (AI)-assisted workflow, bile acid enzyme announcer unit tool (BEAUT), which predicted over 600,000 candidate BA metabolic enzymes that we compiled into the human generalized microbial BA metabolic enzyme (HGBME) database (<https://beaut.bjmu.edu.cn>). We identified a series of uncharacterized BA enzymes, including monoacid acylated BA hydrolase (MABH) and 3-acetoDCA synthetase (ADS). Notably, ADS can produce an unreported skeleton BA, 3-acetoDCA, with a carbon-carbon bond extension. After determining its bacterial source and catalytic mechanism, we found that 3-acetoDCA is widely distributed among populations and regulates the microbial interactions in the gut. In conclusion, our work offers alternative insights into the relationship between microbial BAs and the host from an enzymatic perspective.

INTRODUCTION

The gut microbiota has a profound impact on host disease and health via its production of diverse active metabolites. An important class of these metabolites are bile acids (BAs), which act as pivotal messengers between the host and the gut microbiota. BAs, such as cholic acid (CA) and chenodeoxycholic acid (CDCA), are synthesized from cholesterol in the liver, after which they enter the gut and are further converted into various complex BA derivatives by gut microbial enzymes.^{1–6} Recent studies have shown that different modifications of BAs serve as microbial co-

des to regulate various host functions. Mining this diversity of microbial BAs can allow for the deciphering of this microbial code and may thus lay the foundation for utilizing it to maintain and improve human health.

With the development of metabolomics, hundreds of microbial BAs produced by different bacteria species have been identified.^{1,7–9} By contrast, the authentication of BA biosynthesis pathways in bacteria has significantly lagged behind, with less than 10 kinds of bacterial BA-related enzymes having been characterized to date.^{6,10–15} The large gap between the identification of microbial BAs and the enzymes responsible for their



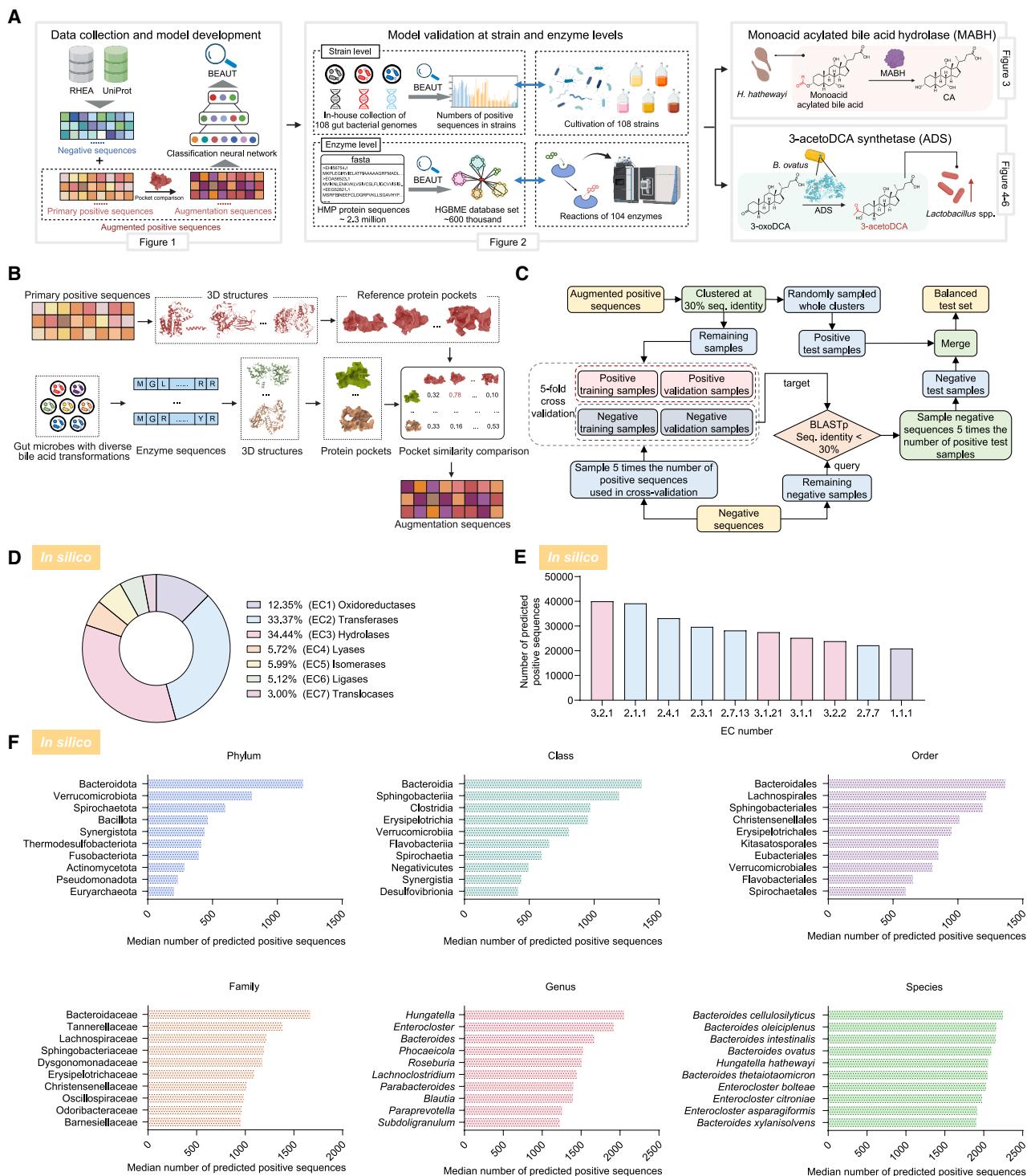


Figure 1. Model development of BEAUT

(A) Schematic representation of BA enzyme announcer unit tool (BEAUT) workflow. In the part of “data collection and model development,” the positive sequences (enzymes reacted with BAs as substrate) and negative sequences (enzymes reacted without BAs) were collected from the RHEA database and UniProt database. The augmented positive sequences were selected, and the ESM-2 model was used to encode function-aware representations for protein sequences and passed the representations through a dense neural network to predict whether the sequence from gut microbiota is a BA metabolic enzyme. Next, BEAUT model was validated at strain and enzyme levels. The predicted BA-related enzyme candidates were verified by enzymatic assays. In this study, two BA metabolic enzymes, as well as the reactions mediated by them, were confirmed.

(legend continued on next page)

biosynthesis greatly limits the development of engineered bacteria and related disease-targeted interventions. Though massive sequencing datasets of gut microbiota have been generated, effective mining methods to excavate BA metabolic enzymes from these data are still lacking. In addition, the microbial BAs identified to date are all modified at active functional groups, such as hydroxyl and carboxyl groups.^{1,6} Thus far, there has been a lack of investigation into whether BAs are modified along their carbon-carbon skeleton type, which may yield unique and unexplored biological activities.

In recent years, with the rapid development of computational biology and artificial intelligence (AI) methods, deciphering protein function based on algorithms has greatly accelerated the identification of so-called biological “dark matter.”^{16–18} For example, the AlphaFold2-assisted structure prediction and structure similarity analysis tools have been used to establish a structure-based protein classification strategy, and they have discovered a battery of unreported single-stranded deaminases, which improved the utility of base editors for therapeutic and agricultural applications.¹⁹ However, whether AI can mine enzymes that modify specific metabolites involved in host-microbe interactions remains largely unexplored.

Herein, we developed an AI-assisted workflow, which we termed bile acid enzyme announcer unit tool (BEAUT), to identify candidate microbial BA-modifying enzymes. BEAUT utilizes a protein language model, evolutionary scale modeling-2 (ESM-2), for the functional-aware representations of proteins and uses a dense neural network for the classification of BA metabolic enzymes from metagenomic data. BEAUT learns from the known sequences of BA metabolic enzymes and the augmented sequences featuring substrate-binding pockets similar to the known BA metabolic enzymes to enable its ability to screen for enzymes that modify BAs. Using this platform, we predicted more than 600,000 potential BA metabolic enzymes from human gut microbiota genomic sequences. Through functional verification, we identified more than 40 undocumented BA metabolic enzymes, such as monoacid 3-acylated BA hydrolase (MABH), which is responsible for the hydrolysis pathway of monoacid 3-acylcholic acid (3-acylCA). Moreover, we identified a previously unknown class of thiamine diphosphate (ThDP)-dependent BA metabolic enzymes, which we designated as 3-acetoDCA synthetases (ADSs) as they produce an unreported skeleton BA, 3-acetodeoxycholic acid (3-acetoDCA). We uncovered the structural basis and catalytic mechanism of ADSs in the formation of 3-acetoDCA, which includes a carboligation reaction. We further showed that 3-acetoDCA is commonly found in human fecal samples and can regulate the composition of gut

microbiota at physiological concentrations, mainly by increasing the abundance of *Lactobacillus* spp. The application of BEAUT and the identification of unreported BA metabolic enzymes greatly extend our knowledge of microbial BA metabolism and the chemical diversity of microbial BAs. Traditional research paradigm begins with the identification of metabolites followed by the excavation of relevant enzymes, while BEAUT enables a top-down approach by directly exploring BA metabolic enzymes, offering a scalable framework for investigating enzymes involved in the metabolism of other microbial-derived metabolites.

RESULTS

Development of the BEAUT platform

BEAUT exploits deep learning techniques to predict unexplored BA metabolic enzymes. We employed the ESM-2 model²⁰ to generate function-aware representations of protein sequences and passed the representations through a dense neural network to predict whether a given sequence encodes a BA metabolic enzyme (Figures 1A–1C, S1A, and S1B). This model takes only enzyme sequences as input and acts as a scoring system to evaluate whether the sequence could utilize BAs as substrates. We used the enzyme commission (EC) number prediction as an additional dimension in the screening process. This multi-dimensional screening strategy improves tolerance to data insufficiency and achieves acceptable performance when screening large-scale sequence datasets.

We collected metabolic reactions involving BAs as substrates from a curated subset of the reaction knowledgebase (RHEA) reactions.²¹ We collected enzyme sequences from UniProt²² annotated with these reactions as positive samples. Sequences annotated with other RHEA reactions were taken as negative samples. Several sequences were manually supplemented from the recent literature as positive samples.^{10,12,23–26} In total, we collected 1,032 positive samples and 108,319 negative samples (Figure S1A). After we removed fragments and redundant sequences at 90% sequence identity, 151 unique positive sequences remained, which were designated as the primary positive samples (Table S1).

We then filtered the negative samples based on sequence length to match the distribution of positive samples, resulting in 102,403 negative sequences. These unbalanced positive and negative data made it impossible to train an effective machine learning model. Thus, data augmentation was necessary. We first analyzed 8 structures of enzyme-BA complexes from the Protein Data Bank (PDB).²⁷ We extracted pockets in these proteins using

(B) Schematic representation of the augmentation sequence collection. The augmentation sequences with high binding pocket similarity to the primary positive sequences were selected from gut microorganisms with highly diverse BA transformations.

(C) Construction of the training sets, validation sets, and the test sets for BEAUT models. After the selection of positive test samples, the remaining positive samples and randomly selected negative samples were separately divided into 5-fold for the cross-validation. Negative test samples were randomly selected from the remaining negative samples that have <30% sequence identity with the negative samples in training and validation sets.

(D) The enzyme classification of BA-related enzyme candidates mined by BEAUT (*in silico*). The enzyme classification was predicted by CLEAN.

(E) The top 10 enzyme classes of BA-related enzyme candidates mined by BEAUT (*in silico*). The EC numbers were predicted by CLEAN.

(F) Enzyme distribution of BA-related enzyme candidates mined by BEAUT among phylum, class, order, family, genus, and species level (*in silico*). The numbers shown in the figure are medians. Species with unclear classification were not included in the statistical data.

See also Figure S1.

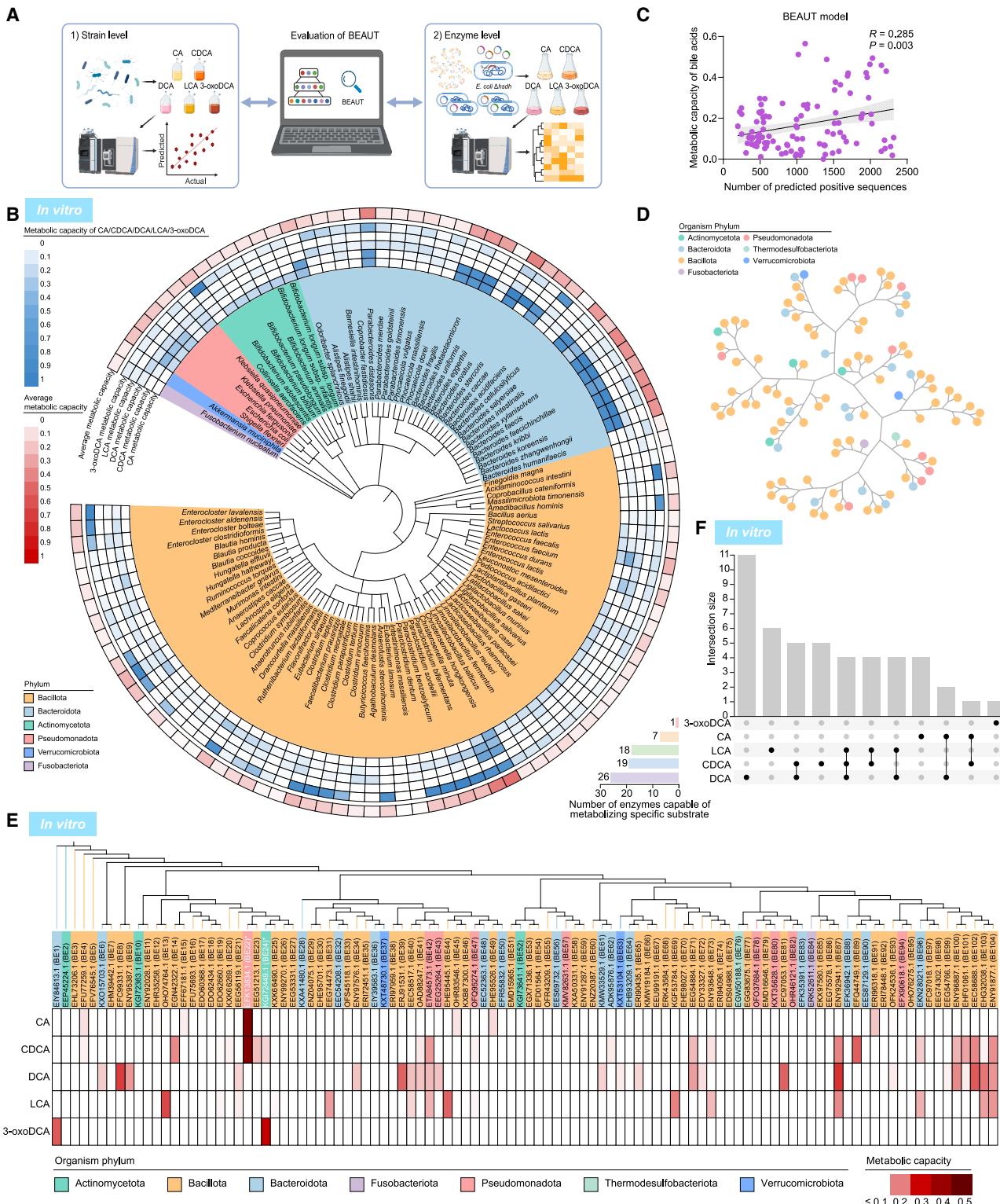


Figure 2. Evaluation of the BEAUT model at the strain and enzyme levels

(A) Schematic representation of model verification at the strain level and enzyme level.

(B) Phylogenetic tree of the 108 gut microbe species and the metabolic capacity of BA s (*in vitro*). The tree was colored in terms of the phylum of each node, and inner heatmap (blue) displayed metabolic capacity of CA/CDCA/DCA/LCA/3-oxoDCA, and the outer heatmap (red) displayed the average metabolic capacity of BA s.

(legend continued on next page)

Cavity²⁸ and aligned the BA ligands according to their common skeleton structure. Analysis of these structures (Data S1) demonstrated that BA molecules have relatively similar conformations and tend to bind pockets with volumes $\geq 1,000 \text{ \AA}^3$. We then calculated pairwise pocket similarities among these BA binding pockets in the protein complexes using PocketMatch,²⁹ and the result suggested that, despite low sequence and structural similarity, these proteins may share high pocket similarities (Data S1). For example, the sulfotransferase showed high pocket similarity to the other two bile salt hydrolases (BSHs), even though their sequence, structure, and function are different. These data suggest that enzymes with substrate-binding pockets similar to those in the primary positive samples may also be BA enzymes. To support this notion, we further collected protein structures of the primary positive samples and extracted their pockets using Cavity (Table S1). After filtering based on the qualities of the overall protein structures (predicted local distance difference test [$p\text{LDDT} \geq 70$]) and the pockets, we were able to generate 85 reference pockets. We conducted pairwise pocket similarity comparisons (see STAR Methods for details) and found that enzymes that use BAs as substrates exhibited higher pocket similarity than those that do not (Figure S1C). This result further validated our assumption that positive data can be augmented through pocket similarity analysis.

Regarding the source of augmentation samples, we selected genomes of 7 intestinal strains with high diversity and a strong capacity for BA metabolism (namely, *Bacteroides xylosoxolvens* DSM 18836, *Bifidobacterium adolescentis* ATCC 15703, *Streptococcus infantarius* ATCC BAA-102, *Holdemania filiformis* DSM 12042, *Clostridium* sp. M62/1, *Mediterraneibacter gnavus* ATCC 29149 [previously termed *Ruminococcus gnavus*], and *Coprococcus comes* ATCC 27758) according to a previous report.³⁰ The protein sequences from the 7 selected strains were filtered by workflow I (see details in the STAR Methods). We then removed sequences that shared $\geq 50\%$ sequence identities with any negative samples (see STAR Methods). We predicted the structures of the remaining proteins and extracted potential binding pockets with volumes comparable to those of the primary positive samples. After filtering based on the qualities of the overall protein structures ($p\text{LDDT} \geq 70$) and the pockets (pocket index ≥ 0.7), we obtained 10,867 query pockets. We then calculated the similarity between the query pockets and the reference pockets from the primary positive samples using PocketMatch (Figures 1B and S1B). We then selected sequences whose pockets showed ≥ 0.7 similarity with the reference pockets for data augmentation. After merging them with the 151 primary positive samples and dereplicating at 90%

sequence similarity, we obtained 2,472 augmented positive samples that we used to construct the BEAUT model (Table S1).

We performed 5-fold cross-validation and trained 5 models using the augmented positive samples and negative sequences (Figure 1C). Analysis of the training loss and validation loss during model training demonstrated no signs of overfitting (Figure S1D). We evaluated the 5 models on a balanced test set, which showed an average area under the precision-recall curve (AUPRC) of 0.80, an average F1 score of 0.72, and an average recall of 0.75. A complete list of evaluation metrics can be found in Table S2. We selected the model with the best AUPRC on this balanced test set as the final BEAUT model.

Using the BEAUT model, we conducted a sequence-based virtual enzyme screening of the Human Microbiome Project (HMP) reference genomes.³¹ A total of 2,340,761 protein sequences within the range of positive sequence length were collected for screening. The predicted positive sequences were further annotated with eggNOG-mapper³² and filtered by workflow II (see details in the STAR Methods). The BEAUT model predicted 614,616 sequences as candidate BA metabolic enzymes (Figure S1E). We then used the enzyme function initiative (EFI) tool³³ to cluster the sequences into 118,599 clusters for further analysis. We further performed an EC prediction using contrastive learning-enabled enzyme annotation (CLEAN)¹⁸ to evaluate the possible reaction types. We found that the positive sequences were distributed across all 7 different enzyme classes, representing a variety of functions (Figures 1D and 1E). At the phylum level, Bacteroidota was predicted to harbor the highest number of BA metabolic enzymes (Figure 1F). At the genus level, *Hungatella* and *Enterocloster* showed a higher abundance of BA metabolic enzymes, with minimal variation in the number of predicted positive sequences among different strains (Figure 1F). We compiled all the predicted BA metabolic enzymes into the human generalized microbial BA metabolic enzyme (HGBME) dataset, which we have made publicly available via our web server (<https://beaut.bjmu.edu.cn>).

Evaluation of the BEAUT model at the strain and enzyme levels

We carried out experiments to validate the prediction accuracy of the BEAUT model at the strain and enzyme levels (Figure 2A). At the strain level, we incubated 108 intestinal strains (across 6 phyla) with 5 BA substrates, including CA, CDCA, deoxycholic acid (DCA), lithocholic acid (LCA), and 3-oxodeoxycholic acid (3-oxoDCA). After a 48-h incubation, we analyzed the BA metabolism profiles by liquid chromatography-mass spectrometry (LC-MS). The reduction of BA substrates acts as

(C) Correlative analysis of metabolic capacity of BAs (*in vitro*) with positive sequences number predicted by BEAUT (*in silico*). Correlation index R and p values were calculated.

(D) Phylogenetic tree of the chosen enzymes for validation. The enzymes were colored in terms of the phylum.

(E) Heatmap shows metabolic capacity of chosen validation enzymes toward different BAs substrates (*in vitro*). The origin of enzyme was colored in terms of the phylum. Enzymes BE22 and BE24 (colored in white) are positive controls of this experiment. The values of metabolic capacity > 0.1 were considered as positive metabolic enzyme, and the values of metabolic capacity ≤ 0.1 were considered as negative metabolic enzyme.

(F) UpSet plots demonstrate validated positive BA metabolic enzymes with different substrates (*in vitro*). Vertical columns in the UpSet plots represent the number of BA metabolic enzymes with different substrate scopes shown as black dots, while horizontal columns on the left represent the number of BA metabolic enzymes of different substrates (CA/CDCA/DCA/LCA/3-oxoDCA). Positive controls BE22 and BE24 were excluded for statistical analysis.

See also Figure S2.

an index of BA metabolic capacity of gut strains (Figure 2B). We found there was a positive correlation between the measured BA metabolic capacity and the numbers of predicted BA metabolic enzymes via the BEAUT model (Figures 2C and S2A).

To further evaluate the prediction ability of BEAUT for BA metabolic enzymes, we randomly selected a total of 102 enzymes from the HGBME dataset for experimental verification and 2 known BA enzymes as positive controls (Figures 2D–2F, S2B, and S2C). The chosen enzymes exhibited a wide functional diversity and were from a wide range of organisms covering 7 phyla and over 40 genera (Figures 2D and S2C; Table S3). Among these candidate enzymes, we found that 47 enzymes were able to degrade BAs to varying degrees, of which 26 enzymes used DCA as a substrate, 7 enzymes used CA as a substrate, and 20 enzymes recognized promiscuous substrates and could markedly consume 2 or more BAs (Figures 2E, 2F, and S2C).

Discovery of a hydrolase of monoacid 3-acylated BAs in the gut microbiota

Previous studies identified 3-acylation of BAs, including monoacid acylation and diacid succinylation.^{1,15,34–36} We recently showed that 3-succinylated CA (3-sucCA) is synthesized by BA acyl synthetase for succinyl (BAS-suc).¹⁵ Intriguingly, the BEAUT model was able to recall the enzyme BAS-suc (GenBank: EDO56041.1, BE105) without the corresponding positive sample training (Figure 3A). Further validation of the BAS-suc protein cluster revealed several previously unknown BAS-suc proteins (Figures 3B and 3C), which, like BE105, were able to catalyze the reactions between BAs and succinic acid (Figure S3A).¹⁵ These BAS-suc enzymes showed promiscuous substrate selectivity toward different BAs (Figure S3A).

To discover potential regulatory pathways of 3-acylated BAs using BEAUT, we recruited 30 volunteers for metagenomic sequencing and targeted BA profiling of their fecal samples. We observed significant negative correlations between the concentrations of monoacid 3-acylated BAs, especially 3-acetylcholic acid (3-aceCA), and the abundance of specific gut commensal bacteria (Figure 3D), indicating an underlying 3-acylated BA hydrolysis pathway in the gut microbiota. We conducted incubation of *Hungatella hathewayi*, which exhibits the strongest negative correlation, and found that it efficiently degraded 3-aceCA (Figure 3E). Next, we utilized BEAUT to predict potential BA metabolic enzymes in *H. hathewayi*. We annotated enzyme functions of the BEAUT-predicted candidates using CLEAN,¹⁸ and we selected the top 20 candidates classified as carboxylic-ester hydrolases (EC 3.1.1) for further analysis (Figures S3B and S3C). Through overexpression and activity validation of these candidates, we ultimately identified the functional enzyme (GenBank: EFD00665.1, BE129), which we termed MABH of *H. hathewayi* (HhMABH), as the one responsible for hydrolyzing 3-aceCA (Figure 3F). We then purified HhMABH (Figure S3D) and performed kinetic analysis. We found that HhMABH had a Michaelis constant (K_m) of 0.39 mM toward 3-aceCA (Figure 3G). By substrate adaptability analysis with various 3-acylCAs, we further found that HhMABH hydrolyzed 3-aceCA, 3-propionylcholic acid (3-proCA), and 3-butrylcholic acid (3-butCA), especially for the first two, while showing no activity toward 3-sucCA (Figures 3H and S3E).

ADS catalyzes the formation of the previously uncharacterized skeleton BA 3-acetoDCA

Bacterial modifications of BAs primarily occur at active functional groups, such as hydroxyl isomerization, dehydroxylation, acylation, and carboxyl amidation, rather than involving alterations to the carbon skeleton.^{6,9,15} However, modifications to the carbon skeleton may lead to large differences in the biological activities of the molecules.³⁷ To date, though, the existence of an alternative skeleton BA remains unknown. Among experimental verification results, a 3-oxoDCA metabolic enzyme (GenBank: E1Y84613.1, BE1), which belongs to a class of ThDP-dependent enzymes reported to be possibly involved in carbon chain elongation,³⁸ has not been reported in BA metabolism until now. Furthermore, we overexpressed 11 enzymes from the BE1-related cluster and found that the majority could consume 3-oxoDCA to varying degrees, with little effect on the other BA substrates (Figures 4A and 4B). In spite of the decreased level of 3-oxoDCA, we did not observe an increase of known 3-oxoDCA derivatives, such as DCA (Figure S4A).

During our investigation of the conversion products of 3-oxoDCA by enzyme BE1, we noticed that an LC-MS peak of m/z 433.2957 appeared after the reaction (Figure S4B). By sequential multiple chromatography and hyphenated techniques combined with LC-MS-based molecular weight tracing, we obtained a pure compound with the target mass and identified its molecular structure by nuclear magnetic resonance (NMR) spectrometry (Figures 4C and 4D), which showed that the 3-position carbon atom of substrate 3-oxoDCA was elongated by an ethyl ketone group (Figures 4E and S4C–S4I). Further, we determined its three-dimensional structure by single-crystal X-ray diffraction and confirmed it as 3-*R*-acetoDCA (termed 3-acetoDCA) (Figure 4F).

We reanalyzed the 11 enzymes from the enzyme BE1-related cluster and found that all these enzymes were able to produce 3-acetoDCA, which as a group, we termed ADS (Figures 4G and 4H). The production of 3-acetoDCA was observed in stool-derived ex vivo communities (SECs) of healthy volunteers using different media (Figure S4J). In addition, we screened for the bacterial source of 3-acetoDCA from 108 species in the laboratory gut microbiota strain library. We found that *Bacteroides ovatus* produced the highest level of 3-acetoDCA (Figure 4I), and different *B. ovatus* strains could generate 3-acetoDCA (Figures S4K and S4L). Based on the homology alignment, we identified the ADS homolog protein EDO13458.1 (BE147) in the genome of *B. ovatus* ATCC 8483 and constructed a targeted genomic mutant (*BoΔads*) by inactivating the *Boads* gene via homologous recombination (Figures S4M and S4N). The *BoΔads* strain showed a total deficiency in 3-acetoDCA production compared with wild-type *B. ovatus* ATCC 8483 (Figures 4J and 4K), indicating that BoADS was responsible for the production of 3-acetoDCA in *B. ovatus*.

The molecular structure of 3-acetoDCA is different from traditional BAs (Figures S4O–S4Q). Besides the strongly polar carboxylic acid side chain, 3-acetoDCA has an additional ethyl ketone side chain in the hydrophobic steroid skeleton part of its tetracyclic ring (Figure S4Q). Recent studies have identified a variety of BAs using untargeted metabolomics analysis of universal skeleton ions at m/z 319.24 and m/z 337.25 as BA diagnostic

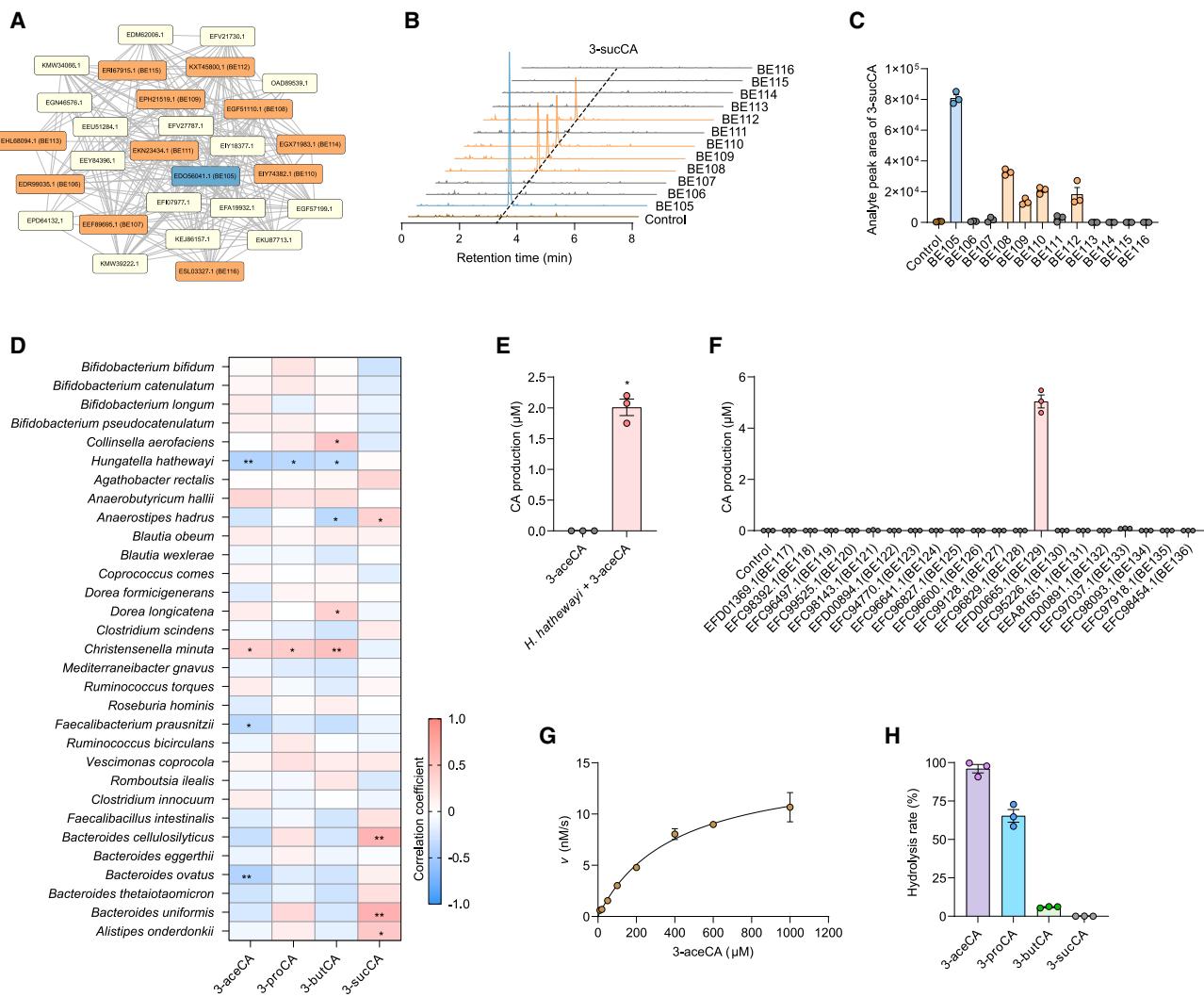


Figure 3. MABH is responsible for the hydrolysis of 3-acetCA

(A) The potential BAS-suc protein cluster generated by enzyme function initiative-enzyme similarity tool (EFI-EST) with an *E* value threshold of 10^{-5} and an alignment score threshold of 60 from HGBME database. The blue item is the reported BAS-suc, and the orange items are chosen for activity verification.

(B and C) Representative extracted ion chromatogram (B) and quantified production (C) of 3-sucCA of enzymatic assays catalyzed by different *E. coli* strains lysate harboring empty plasmid (control) or overexpression vectors (BE105–116) of different potential BAS-suc.

(D) Heatmap shows the correlation between the levels of acylated BAs and specific species of gut bacteria in the feces samples from 30 healthy volunteers.

* $Q < 0.05$, ** $Q < 0.01$, Q value (false discovery rate [FDR]-adjusted p value).

(E) Quantified production of CA hydrolyzed by *H. hathewayi* from 3-acetCA. $n = 3$.

(F) 3-acetCA hydrolytic activity of BEAUT-predicted enzymes from *H. hathewayi* DSM 13479.

(G) Michaelis-Menten curves for HhMABH (BE129). 3-AcetCA in different concentrations as substrate. $n = 3$.

(H) Hydrolytic rate of HhMABH incubating with different 3-acylated CAs.

All data are presented as the means \pm SEMs. In (E), the p values were determined by Mann-Whitney U test. * $p < 0.05$ versus the 3-acetCA group.

See also Figure S3.

fragments.⁹ We hypothesized that 3-acetoDCA may exhibit special molecular characteristics due to the extended skeleton. Indeed, non-target metabolomic data showed that 3-acetoDCA did not display characteristic ions of traditional BAs but rather exhibited a special 433.2957 *m/z* (Figures S4O and S4P). The optimization of a future metabolomics workflow will facilitate the identification of more BAs based on current discoveries.

The crystal structure and catalytic mechanism of BoADS

BoADS is a ThDP-dependent enzyme belonging to the 1-deoxy-D-xylulose-5-phosphate synthase (DXPS)-like subfamily. This subfamily of enzymes catalyzes the formation of 1-deoxy-D-xylulose-5-phosphate (DXP) from pyruvate and D-glyceraldehyde 3-phosphate (D-GAP), which is the crucial and rate-limiting step in the biosynthesis of the vitamins B1 and B6.^{39–41} By contrast, as an unreported enzyme for BA metabolism, BoADS

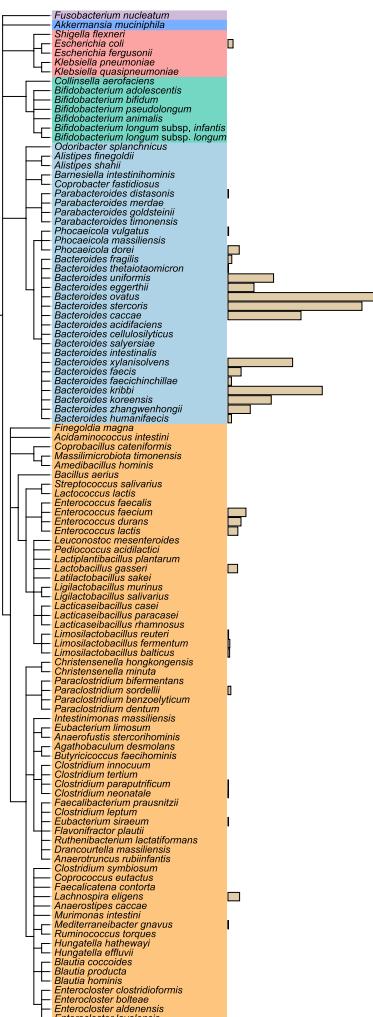
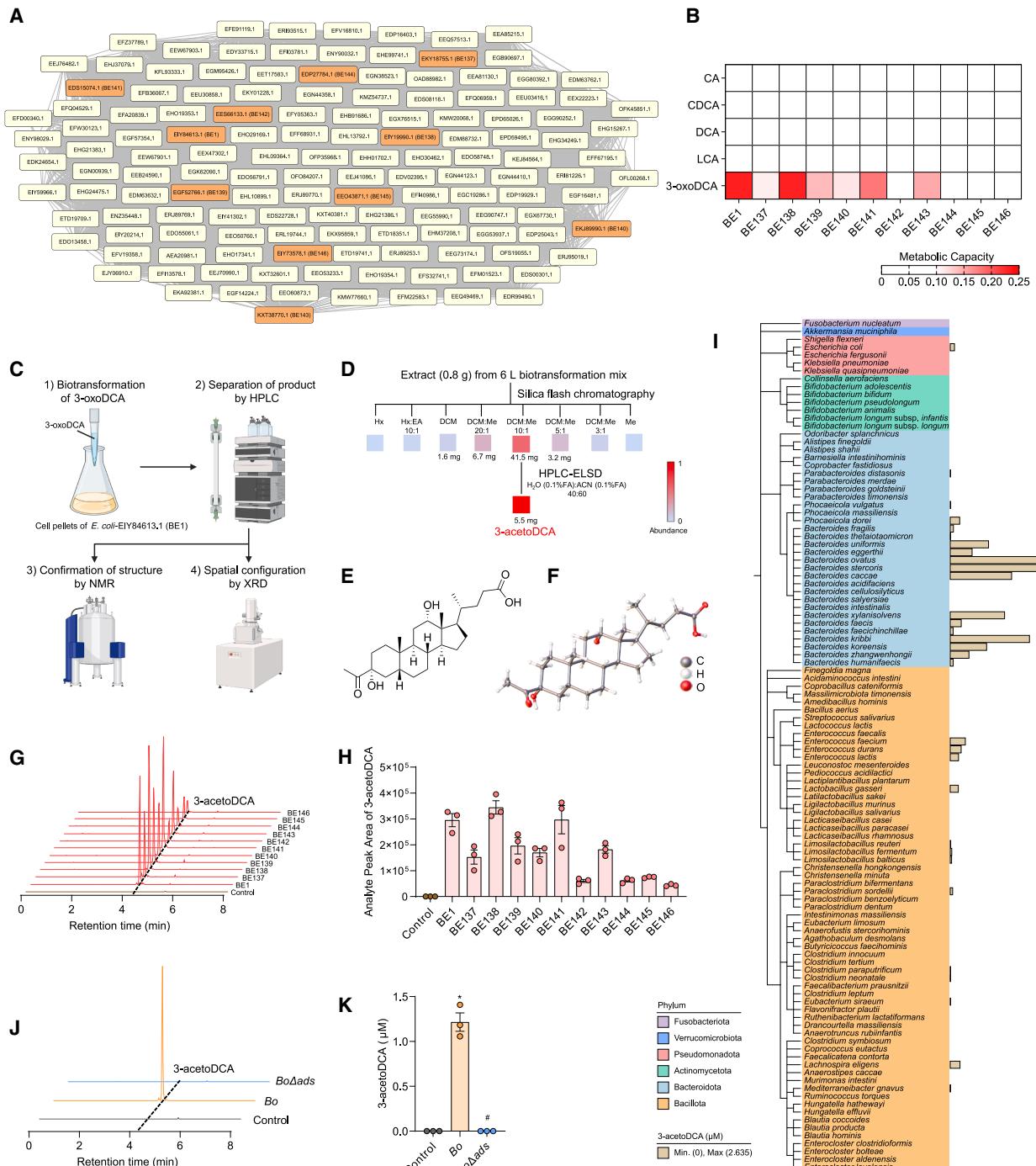


Figure 4. ADS catalyzes the formation of the unreported skeleton BA 3-acetoDCA

- (A) The BE-1-related cluster generated by EFI-EST with an E value threshold of 10^{-5} and an alignment score threshold of 60 and an alignment score threshold of 60 from HGBME database. The orange items are chosen for activity verification.
- (B) Heatmap shows metabolic capacity of chosen potential 3-oxoDCA metabolic enzymes from BE-1 related cluster toward different BA substrates.
- (C) Schematic diagram shows the isolation and identification flow of the unreported skeleton BA 3-acetoDCA. HPLC, high-performance liquid chromatography; NMR, nuclear magnetic resonance; XRD, X-ray diffraction.
- (D) Flow diagram for fractionation of 3-oxoDCA biotransformed by *E. coli*-EIY84613.1 (BE1). Hx, hexane; EA, ethyl acetate; DCM, dichloromethane; Me, methanol; FA, formic acid; ACN, acetonitrile; HPLC-ELSD, high-performance liquid chromatography-evaporative light scattering detection.
- (E) The chemical structure of 3-acetoDCA.
- (F) Thermal ellipsoid representation of 3-acetoDCA by X-ray crystallographic analysis.

(legend continued on next page)

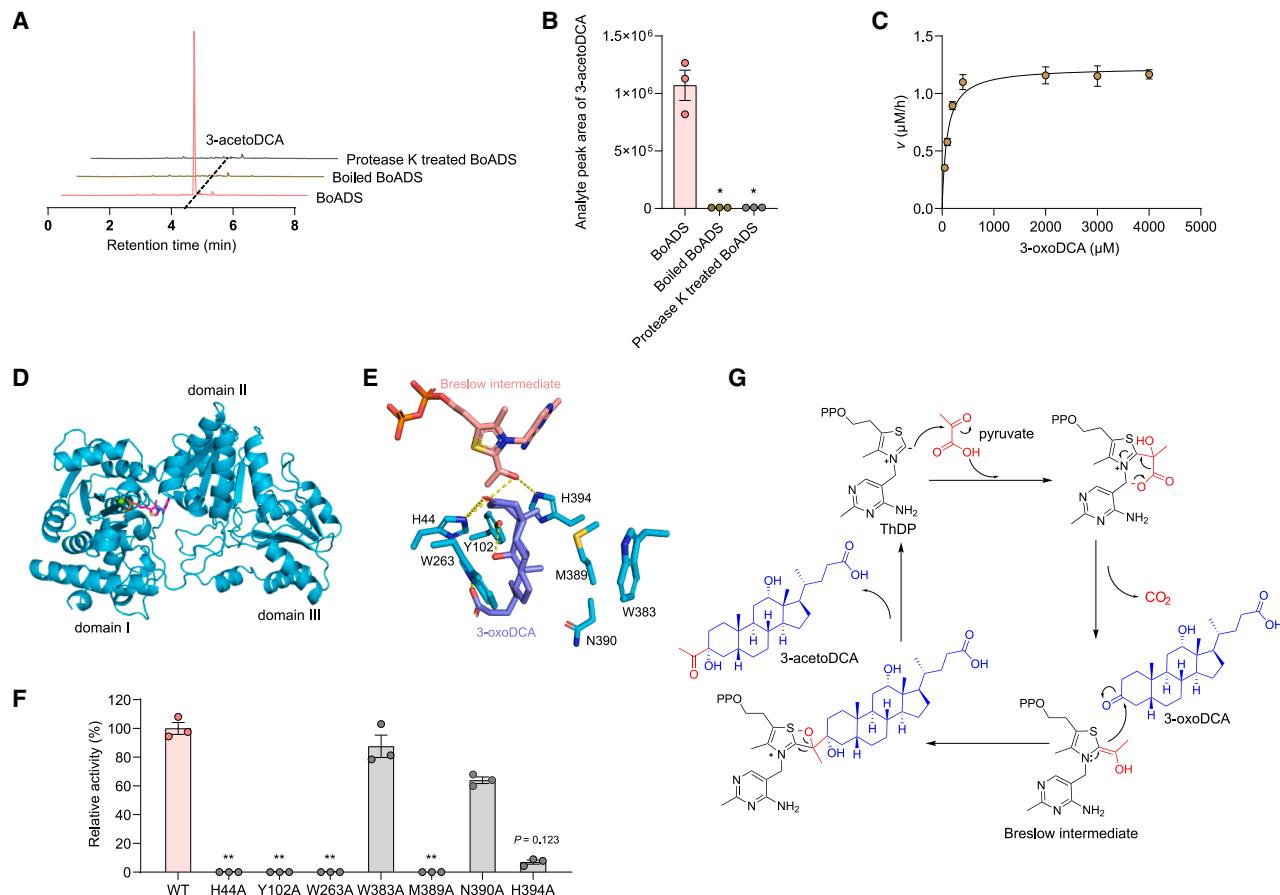


Figure 5. The crystal structure of BoADS-ThDP and its catalytic mechanism

(A and B) Representative extracted ion chromatogram (A) and quantified production (B) of 3-acetoDCA in enzymatic assays with BoADS protein under different conditions. $n = 3$.

(C) Michaelis-Menten curves for BoADS with 3-oxoDCA as the substrate in different concentrations. $n = 3$.

(D) The overall structure of BoADS-ThDP consists of three domains: I, II, and III. The ThDP was shown in magenta sticks, and the Mg²⁺ ion was shown as a green sphere.

(E) The modeled complex structure of BoADS with Breslow intermediate (salmon sticks) and 3-oxoDCA (slate sticks). Hydrogen bonds were shown as yellow dashed lines.

(F) The relative catalytic activities of BoADS and its variants are related to the 3-oxoDCA binding sites. $n = 3$.

(G) The proposed catalytic mechanism of BoADS.

All data are presented as the means \pm SEMs. In (B), the p values were determined by one-way ANOVA with Dunnett's T3 test. * $p < 0.05$ versus the BoADS group. In (F), the p values were determined by Kruskal-Wallis test followed by Dunn's post hoc test. ** $p < 0.01$ versus WT group.

See also Figure S5.

represents the special ThDP-dependent enzyme using a BA as the substrate (Figures 5A–5C and S5A–S5D). Kinetic analysis was performed under optimum pH and temperature, showing that BoADS had a Michaelis constant (K_m) of 95.03 μM (Figures 5C, S5B, and S5C).

To elucidate the catalytic mechanism of BoADS, the crystal structure of BoADS complexed with ThDP (termed BoADS-ThDP) was determined at 2.42 \AA with two molecules in one asymmetric unit (Table S4). The overall structure of BoADS-ThDP showed a similar structural fold to DXPSs and contained

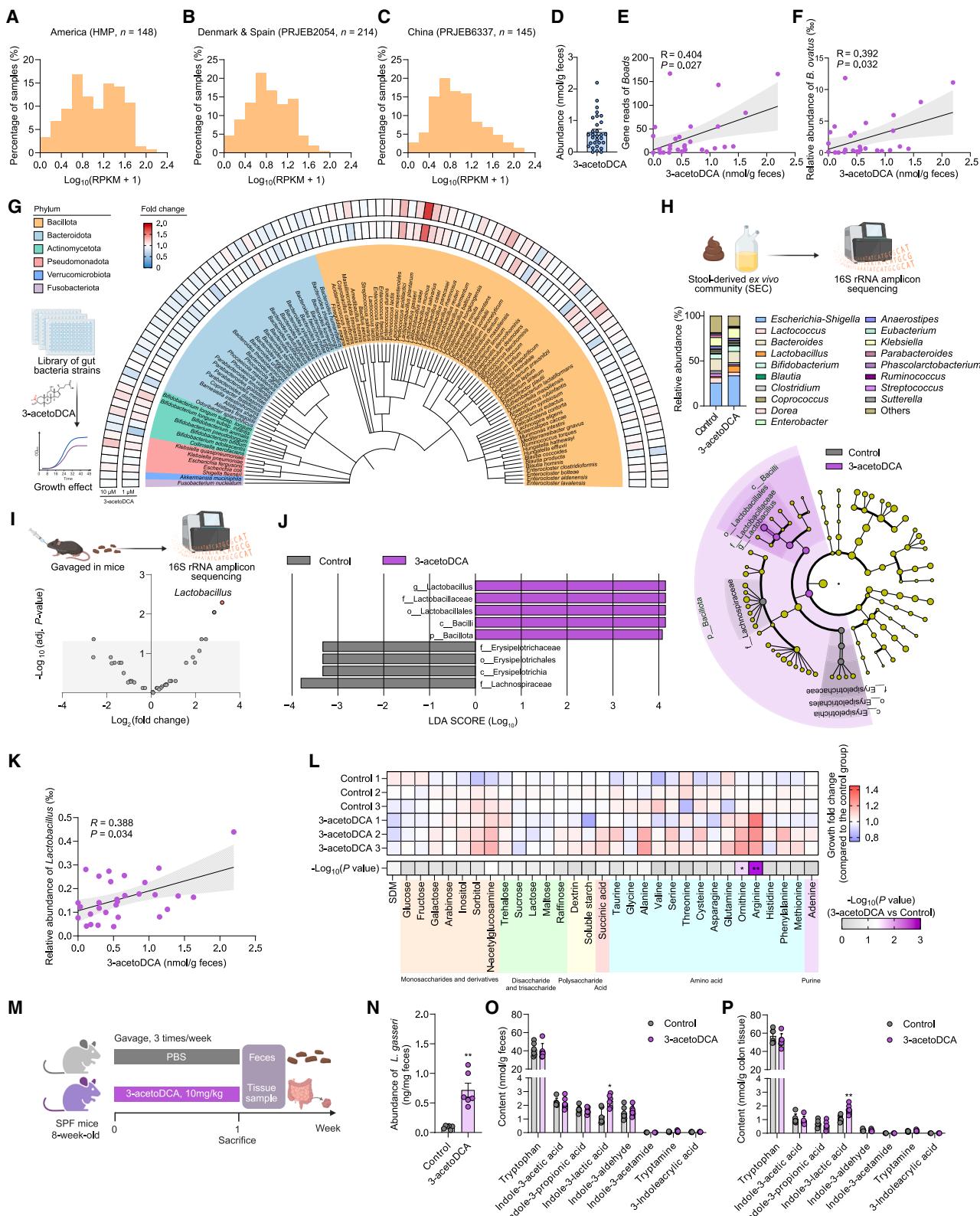
(G and H) Representative extracted ion chromatogram (G) and quantified production (H) of 3-acetoDCA production catalyzed by different *E. coli* strains harboring empty plasmid (control) or overexpression vectors with different potential ADS (BE1, BE137–146).

(I) Phylogenetic tree of the 108 gut microbe species and the yield of 3-acetoDCA. The tree was colored in terms of the phylum of each node, and the histogram displayed the yield of 3-acetoDCA.

(J and K) Representative extracted ion chromatogram (J) and quantified production (K) of 3-acetoDCA by medium control, culture of *B. ovatus* ATCC 8483 (*Bo*), or *B. ovatus* ATCC 8483 *ads* deficiency strain (*BoΔads*).

All data are presented as the means \pm SEMs. In (K), the p values were determined by Kruskal-Wallis test followed by Dunn's post hoc test. * $p < 0.05$ versus the control. # $p < 0.05$ versus the *Bo*.

See also Figure S4.



(legend on next page)

three distinct domains (I, II, and III) (Figure 5D). Each domain was composed of a central five- or six-stranded parallel β -sheet surrounded by varying numbers of α -helices. We found that the ThDP-binding pocket is located at the interface between domains I and II. The diphosphate moiety of ThDP and a Mg²⁺ ion interacted with the residues H75, D147, N174, N176, K247, and H262 (Figure S5E). We further found that the pyrimidine ring of ThDP is π - π stacked with F358 and forms hydrogen bonds with G116 and E333 (Figure S5E). The point mutations of above residues showed a sharp decrease or loss of catalytic activity, confirming the significant roles of these residues in the binding to ThDP (Figure S5F).

In the catalysis of BoADS, the cofactor ThDP first reacts with the substrate pyruvate to form a Breslow intermediate, which then reacts with another substrate, 3-oxoDCA. To investigate the substrate-binding mode, we modeled both the Breslow intermediate and 3-oxoDCA into the BoADS active site. We found that the hydroxy group of the Breslow intermediate forms hydrogen bonds with H44 and H394 and that the C3-carbonyl, C12-hydroxyl, and C24-carboxylic groups of 3-oxoDCA form hydrogen bonds with H44, Y102, and W263, respectively (Figure 5E). More importantly, we found that the hydrocarbon rings of 3-oxoDCA form a “sandwiched” hydrophobic interaction style with W263 from one side and M389 from the other side (Figure 5E). The enzyme variants of H44A, H394A, Y102A, W263A, and M389A showed decrease or loss of catalytic activity, confirming the roles of these residues in substrate binding (Figure 5F).

Through DALI analysis, we found that BoADS showed the highest structural similarity to DrDXPS (the DXPS from *Deinococcus radiodurans*, PDB: 6OUV) with a root-mean-square deviation (RMSD) of 1.19 for C α atoms. The substrates of BoADS and DrDXPS were different: 3-oxoDCA for BoADS and d-GAP for DrDXPS (Figure S5G). Structurally, BoADS showed similar struc-

tural folds and a ThDP-binding pocket to those of DrDXPS (Figures S5H and S5I), but with distinct substrate-binding pockets, resulting in their different substrate recognition activity (Figures S5J and S5K). Moreover, the key residues W263 and M389 involved in the sandwiched interactions with 3-oxoDCA in BoADS are replaced by G305 and G428 in DrDXPS, highlighting the special binding mode of BoADS toward 3-oxoDCA.

Based on the crystal structure and site-directed mutagenesis analyses outlined above, we proposed a catalytic mechanism for BoADS (Figure 5G). Namely, the ThDP cofactor binds tightly between the domains I and II of BoADS and attacks the α -keto group of pyruvate, followed by decarboxylation to generate a Breslow intermediate, which then attacks the 3-keto group of 3-oxoDCA, followed by the leaving of the ThDP to produce 3-acetoDCA. This mechanism expands our understanding of how nature utilizes enzymes to diversify the secondary BA pools.

3-acetoDCA is widespread among the human population with a special effect on the gut microbiota

We next investigated whether 3-acetoDCA exists naturally in the human gut. By analyzing metagenomic data from three independent human cohorts from public databases (HMP-America, PRJEB2054-Europe, and PRJEB6337-China), we found that the gene *Boads* is widely distributed across different regions (Figures 6A–6C). By further metabolomics analysis, we confirmed the existence of 3-acetoDCA in the fecal samples from healthy volunteers (Figure 6D). Meanwhile, metagenomic analysis of feces from the volunteers showed that the concentration of 3-acetoDCA was positively correlated with the reads of *Boads* and the abundance of *B. ovatus* (Figures 6E and 6F).

We further examined the effect of 3-acetoDCA on traditional BA receptors and found that 3-acetoDCA had no effect on farnesoid X receptor (FXR), vitamin D receptor (VDR), pregnane X

Figure 6. 3-acetoDCA is widespread in humans and promotes the expansion of *Lactobacillus*

- (A–C) 3-acetoDCA synthesis gene *ads* of *B. ovatus* is widespread among three cohorts: HMP (A), PRJEB2054 (B), and PRJEB6337 (C). RPKM, reads per kilobase of transcript per million reads mapped.
- (D) Levels of 3-acetoDCA in feces of 30 healthy volunteers.
- (E and F) Correlative analysis of 3-acetoDCA with the reads of gene *Boads* (E) and *B. ovatus* abundance (F). Correlation index R and p values were calculated.
- (G) Phylogenetic tree of the 108 gut microbe species and the effect of 3-acetoDCA at different concentrations on the growth of gut bacteria at the single-strain level. The tree was colored in terms of the phylum of each node, and the inner heatmap shows the effect of 3-acetoDCA (1 μ M) incubation on the growth of gut bacteria, and the outer heatmap shows the effect of 3-acetoDCA (10 μ M) incubation on the growth of gut bacteria. Each value represents the ratio of OD₆₀₀ value between the 3-acetoDCA and control groups.
- (H) Genus-level compositions of human stool-derived ex vivo communities (SECs) cultured in mGAM media supplemented with 3-acetoDCA (10 μ M) or PBS control anaerobically for 48 h.
- (I and J) SPF mice were treated with PBS (control) or 3-acetoDCA (10 mg/kg) 3 times per week for 1 week. The fecal samples in the endpoint were collected, and 16S rRNA amplicon sequencing was performed. n = 6 mice/group.
- (K) Volcano plot of 16S rRNA gene sequencing data of control and 3-acetoDCA groups. The adjusted p value threshold was calculated using moderated Student's t test followed by Benjamini-Hochberg multiple-test-correction FDR.
- (L) Taxonomic cladogram generated from linear discriminant analysis effect size (LEfSe) analysis. Gray and purple represent enriched taxa in the control and 3-acetoDCA groups, respectively.
- (M) Correlative analysis of 3-acetoDCA with *Lactobacillus* abundance of 30 healthy volunteers. Correlation index R and p values were calculated.
- (N) *L. gasseri* was cultured in semi-defined medium (SDM) supplemented with 3-acetoDCA or DMSO (control) in the presence of different nutrients (0.5 mg/mL). Each value represents the ratio of OD₆₀₀ value between the 3-acetoDCA and control groups. n = 3.
- (O–P) Abundance of *Lactobacillus gasseri* (N) and levels of tryptophan metabolites in feces (O) and colon (P) of control and 3-acetoDCA-treated SPF mice, as described in the schematic diagram (M). n = 6 mice per group.
- All data are presented as the means \pm SEMs. In (L) and (N), the p values were determined by a two-tailed Student's t test. In (O) and (P), the p values were determined by Mann-Whitney U test. *p < 0.05, **p < 0.01 versus the control group.
- See also Figure S6.

receptor (PXR), and Takeda G-protein-coupled receptor 5 (TGR5) activity in luciferase assays (Figures S6A–S6D). Considering that certain BAs can regulate host functions by influencing the composition of the gut microbiota,^{15,42–45} we then explored the effects of 3-acetoDCA on the growth of gut microbiota members and found it could facilitate the growth of *Lactobacillus* spp. (Figures 6G–6J and S6E–S6N). In detail, 3-acetoDCA could specifically promote the growth of *Lactobacillus gasseri* via a dose-dependent manner *in vitro* (Figure 6G), whereas other well-studied BAs, such as CA, CDCA, DCA, LCA, and 3-oxoDCA, showed no promoting effect on *L. gasseri* growth (Figure S6E). Different *L. gasseri* strains exhibited similar growth-promoting effects under 3-acetoDCA treatment (Figures S6H–S6J). Through the characterization of various *Lactobacillus* species, we found that 3-acetoDCA promoted the growth of *Lactobacillus* species, especially *L. gasseri* (Figures S6K–S6N). In SEC culture and in mice, 3-acetoDCA treatment also significantly increased the abundance of *Lactobacillus* both *ex vivo* and *in vivo* (Figures 6H–6J, S6F, and S6G). Furthermore, we observed a significant positive correlation between the concentration of 3-acetoDCA and the abundance of *Lactobacillus* in fecal samples (Figure 6K).

Next, we sought to investigate the mechanism by which 3-acetoDCA promotes the growth of *L. gasseri*. By examining the growth of *L. gasseri* in the presence or absence of 3-acetoDCA treatment in a semi-defined medium (SDM)⁴⁶ supplemented with different nutrients, such as monosaccharides, disaccharides, polysaccharides, and amino acids, we observed that 3-acetoDCA exhibited almost no growth-promoting effect on *L. gasseri* in the absence of added nutrients. Notably, when arginine or ornithine was supplemented, the growth of *L. gasseri* was significantly improved after 3-acetoDCA treatment (Figures 6L, S6O, and S6P). Previous studies have reported that arginine serves as an essential nutrient for *Lactobacillus*, and its deficiency inhibits the growth of *Lactobacillus*.^{47–49}

Considering that previous studies have reported that *Lactobacillus* species, such as *L. gasseri*, produce various tryptophan derivatives with the potential to improve metabolic diseases,^{50–53} we further evaluated the levels of tryptophan derivatives under the treatment of 3-acetoDCA *in vivo* (Figure 6M). We found that administration of 3-acetoDCA in mice by gavage significantly increased the abundance of *L. gasseri*, as verified by qPCR (Figure 6N), as well as the concentrations of indole-3-lactic acid in both the feces and colon tissue (Figures 6O and 6P).

DISCUSSION

Metabolites serve as messengers between the gut microbiota and the host.^{1,54} As co-metabolic products of the host and gut microbiota, BAs play a crucial role in lipid metabolism, immune regulation, and bacterial infections.^{1–5,55–59} The gut microbiota can utilize host-derived BAs to produce a diverse array of microbial BAs (also referred to as secondary BAs), including DCA, LCA, amino acid-conjugated BAs, 3-acylated BAs, oxidative BAs (3-oxo, 6-oxo, 7-oxo, and 12-oxo), and others.^{6–9,15,34} BAs with different structures can regulate host homeostasis and diseases by activating distinct receptors or influencing the

growth and/or activity of specific gut microbiota.^{3–5,55–57} Microbial modifications of BAs serve as a microbial code for the regulation of physiological processes. Identifying the relevant enzymes involved in such BA modification is crucial for developing targeted strategies, constructing engineered strains, and understanding the role of BA modifications in health and diseases. However, pinpointing the specific microbial enzymes involved in BA metabolism remains a major challenge.¹ For example, the elucidation of 7 α -dehydroxylation pathway from CA to DCA took approximately 40 years.¹⁰ Likewise, though it has long been known that amino acid-conjugated BAs are synthesized by the BA-coenzyme A (CoA):amino acid N-acyltransferase (BAAT) enzymes in the host, only recently was it discovered that bacterial BSH/transferase (BSH/T) can also synthesize microbially conjugated BAs.^{13,14} Herein, we utilized AI to expand the discovery of BA metabolic enzymes by developing the BEAUT workflow to predict BA metabolic enzymes from large-scale datasets. We validated more than 100 candidate enzymes and identified two unreported enzymes—a previously uncharacterized 3-aceCA hydrolase and an unreported skeleton BA synthetase. To enhance user accessibility, we developed an interactive web server—BEAUT (accessible at <https://beaut.bjmu.edu.cn>)—for BA metabolic enzyme prediction. The HGBME dataset that contains potential BA metabolic enzymes in the human microbiome genomes was also deposited on the web server. Users can search the dataset and download their proteins of interest freely.

Previous studies have identified monoacid and diacid 3-O-acylation modifications of BAs.^{15,34} Herein, we utilized BEAUT to successfully recall BAS-suc that catalyzes succinylation without the prior incorporation of relevant positive sequences. Whether the composition of 3-acylated BAs is regulated by specific acylated BA hydrolases, similar to how conjugated BAs are modulated by BSHs, remains unclear. To address the identification of a specific BA metabolic enzyme, we developed a systematic workflow using BEAUT that involved (1) employing an omics approach to identify strains associated with specific BA metabolism, (2) obtaining these relevant strains and experimentally validating whether the candidate strains can convert the specific BA into an expected product, and (3) utilizing BEAUT to predict all genes in the genome of candidate strains related to BA metabolism, followed by prioritizing and validating the potential candidate enzymes based on predicted scores and enzymatic reaction types for their ability to catalyze the specific BA transformation. Through the above workflow, we identified a previously unknown microbial hydrolase for monoacid acylated BAs in *H. hathewayi*. Prior studies have shown that 3-acylated BAs can regulate type 2 diabetes through the gut FXR-ceramide axis.^{15,34} Our research suggests that MABH serves as a potential target for developing a treatment for diabetes, warranting further clinical investigation.

Identification of BA molecules is an intricate and time-consuming task. Recent advancements have employed MS and metabolomics, introducing strategies such as reverse metabolomics and the mass spectrometry query language (MassQL) algorithm, leading to the discovery of unreported BAs, including polyamine-conjugated BAs, and the recently discovered bacterial BA amides (BBAAs).^{7–9} BEAUT correlates

massive sequencing data with specific molecules and bypasses the traditional strategy of strain isolation and compound purification, thus accelerating the discovery of BA metabolic enzymes. The gut microbiota can modify the hydroxyl, carboxyl, and other reactive functional groups of BAs to produce BAs with different biological activities, which play important roles in immune cell differentiation and microbial regulation.⁶⁰ Compared with common functional group modifications, reactions that alter the carbon-carbon skeleton of BAs are rare and limited in variety, yet such modifications can often significantly affect the activity of metabolites (for example, the microbial ring-opening metabolism of cholesterol yields metabolites with entirely new scaffolds, initially termed calcifediol, with the elucidation of its special physiological and pathophysiological functions, which later evolved into a distinct metabolite class now known as vitamin D⁶¹). Using the BEAUT microbiota enzyme mining system, we identified a microbial enzyme, ADS, that catalyzes a C3 side-chain formation of BAs, generating an undocumented skeleton with a double sidechain structure, which represents an unsearched class of ditail-BA (DTB). Through in-depth analysis of the catalytic mechanism, we discovered that BoADS, the enzyme responsible for 3-acetoDCA synthesis, belongs to the DXPS family.⁶² However, compared with the other member's substrate-binding pocket in primary metabolism, the pocket of BoADS is larger and provides a special sandwiched interaction with its BA substrate. Subsequent analysis of this pocket can aid in the discovery and engineering of BA metabolic enzymes.

Our analysis of clinical cohorts revealed that the *BoADS* gene is widespread among different human populations. Furthermore, we found that 3-acetoDCA can promote the growth of the probiotic bacteria *Lactobacillus* species and increase the level of indole derivatives, which are known for their potential metabolic protective role.^{51–53} Future research should delve into the roles and mechanisms of 3-acetoDCA under physiological and pathological conditions.

In conclusion, our approach for the mining of microbial BA synthetase led to the discovery of the hydrolysis mechanisms of gut microbiota on monoacid acylated BAs, providing insights into how the gut microbiota maintains glucose homeostasis through subtle regulation of monoacid acylated BAs. More importantly, starting from enzymes with unknown functions, we identified an unreported type of a widespread BA with an extended carbon chain skeleton, 3-acetoDCA, which can specifically promote the growth of gut commensal bacteria *L. gasseri*, enrich tryptophan metabolites in mice, and play potential protective roles in metabolic diseases. Our findings offer alternative avenues into the exploration of microbial BAs and their biosynthetic pathways while providing a “nature-mimicking top-down” research strategy for exploring BAs with physiological regulatory functions.

Limitations of the study

In this study, we identified more than 40 enzymes that may take part in the transformation of various BA substrates through functional research. However, further investigation into the reaction types and the BA derivatives produced by these enzymes is needed. Furthermore, the prediction accuracy and application of BEAUT are restricted by the currently limited number of func-

tionally validated BA metabolic enzymes. Further optimization and iterations of BEAUT are needed to improve prediction accuracy and to enhance our understanding of microbial BA metabolism.

RESOURCE AVAILABILITY

Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Changtao Jiang (jiangchangtao@bjmu.edu.cn).

Materials availability

All unique/stable reagents generated in this study are available from the [lead contact](#) with a completed materials transfer agreement.

Data and code availability

- Metagenome data for healthy volunteers have been deposited at NMDC (<https://nmdc.cn/>) at NMDC10019873 and are publicly available as of the date of publication.
- Non-targeted metabolome data for 3-acetoDCA identification have been deposited at NMDC (<https://nmdc.cn/>) at NMDCX0002114 and are publicly available as of the date of publication.
- All original code and necessary data for running the code have been deposited on Github (<https://github.com/skystreet8/BEAUT>) and Zenodo (<https://zenodo.org/records/15388149>) and are publicly available as of the date of publication.
- Any additional information required to reanalyze the data reported in this paper is available from the [lead contact](#) upon request.

ACKNOWLEDGMENTS

We thank Jinwei Ren (State Key Laboratory of Mycology, Institute of Microbiology, and Chinese Academy of Sciences) for his help in obtaining the NMR data. Schematic diagrams were generated by [Biorender.com](#).

This work was supported by the National Natural Science Foundation of China (nos. 82341226, 92357305, 82130022, T2321001, 8288102, 82422017, 82370812, 32300989, and 82325046), the National Key Research and Development Program of China (no. 2022YFA0806400), and the Beijing Outstanding Young Scientist Program (no. JWZQ20240102003). C.J. acknowledges the support from the Tencent Foundation through the Xplorer Prize.

AUTHOR CONTRIBUTIONS

C.J. conceptualized and designed the study. Y.D., X.L., J.G., B.X., H.L., H.M., Y.W., M.L., C.Y., S.Y., K.L., J.Z., Y.Z., Q.N., D.Y., Z.Z., Y.P., K.W., M.M., L.L., and C.J. performed the experiments and analyzed the data. C.J., L.L., M.M., and K.W. supervised the study. Y.D., X.L., J.G., B.X., and C.J. wrote the manuscript with input from all authors. Y.D., X.L., J.G., B.X., and H.L. contributed equally to this work. All authors edited the manuscript and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR METHODS

Detailed methods are provided in the online version of this paper and include the following:

- [KEY RESOURCES TABLE](#)
- [EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS](#)
 - Human study
 - Mice and Treatments
 - Bacteria

● METHOD DETAILS

- Development of BEAUT workflow
- Targeted metabolomics analysis
- Detection of bile acid metabolic capacity in different strains
- Construction of *Escherichia coli* BL21 (DE3) knockout strain
- Activity test of bile acid related enzyme candidates
- Heterologous expression and activity of MABH candidates
- Extraction and isolation of 3-acetoDCA
- Construction of the *ads* knockout mutant of *Bacteroides ovatus*
- Gene expression and protein purification of BoADS
- Activity assay, kinetics analyses and substrate scope of BoADS
- Crystallization, data collection and structural elucidation of BoADS-ThDP
- Site-directed mutagenesis of BoADS
- Substrate modeling of BoADS
- Growth curves of tested strains
- 16S rRNA gene amplicon sequencing and analyses
- Metagenomic sequencing
- Luciferase reporter gene assays
- Stool-derived *ex vivo* communities
- Collection of public datasets and analyses of metagenomic sequencing data

● QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at <https://doi.org/10.1016/j.cell.2025.07.017>.

Received: November 23, 2024

Revised: May 19, 2025

Accepted: July 13, 2025

Published: August 7, 2025

REFERENCES

1. Mohanty, I., Allaband, C., Mannochio-Russo, H., El Abiead, Y., Hagey, L.R., Knight, R., and Dorrestein, P.C. (2024). The changing metabolic landscape of bile acids - keys to metabolism and immune regulation. *Nat. Rev. Gastroenterol. Hepatol.* 21, 493–516. <https://doi.org/10.1038/s41575-024-00914-3>.
2. de Aguiar Vallim, T.Q., Tarling, E.J., and Edwards, P.A. (2013). Pleiotropic roles of bile acids in metabolism. *Cell Metab.* 17, 657–669. <https://doi.org/10.1016/j.cmet.2013.03.013>.
3. Jia, W., Xie, G., and Jia, W. (2018). Bile acid-microbiota crosstalk in gastrointestinal inflammation and carcinogenesis. *Nat. Rev. Gastroenterol. Hepatol.* 15, 111–128. <https://doi.org/10.1038/nrgastro.2017.119>.
4. Wahlström, A., Sayin, S.I., Marschall, H.U., and Bäckhed, F. (2016). Intestinal Crosstalk between Bile Acids and Microbiota and Its Impact on Host Metabolism. *Cell Metab.* 24, 41–50. <https://doi.org/10.1016/j.cmet.2016.05.005>.
5. Chávez-Talavera, O., Tailleux, A., Lefebvre, P., and Staels, B. (2017). Bile Acid Control of Metabolism and Inflammation in Obesity, Type 2 Diabetes, Dyslipidemia, and Nonalcoholic Fatty Liver Disease. *Gastroenterology* 152, 1679–1694.e3. <https://doi.org/10.1053/j.gastro.2017.01.055>.
6. Guzior, D.V., and Quinn, R.A. (2021). Review: microbial transformations of human bile acids. *Microbiome* 9, 140. <https://doi.org/10.1186/s40168-021-01101-1>.
7. Quinn, R.A., Melnik, A.V., Vrbanac, A., Fu, T., Patras, K.A., Christy, M.P., Bodai, Z., Belda-Ferre, P., Tripathi, A., Chung, L.K., et al. (2020). Global chemical effects of the microbiome include new bile-acid conjugations. *Nature* 579, 123–129. <https://doi.org/10.1038/s41586-020-2047-9>.
8. Gentry, E.C., Collins, S.L., Panitchpakdi, M., Belda-Ferre, P., Stewart, A. K., Carrillo Terrazas, M., Lu, H.H., Zuffa, S., Yan, T., Avila-Pacheco, J., et al. (2024). Reverse metabolomics for the discovery of chemical structures from humans. *Nature* 626, 419–426. <https://doi.org/10.1038/s41586-023-06906-8>.
9. Mohanty, I., Mannochio-Russo, H., Schweer, J.V., El Abiead, Y., Bittremieux, W., Xing, S., Schmid, R., Zuffa, S., Vasquez, F., Muti, V.B., et al. (2024). The underappreciated diversity of bile acid modifications. *Cell* 187, 1801–1818.e20. <https://doi.org/10.1016/j.cell.2024.02.019>.
10. Funabashi, M., Grove, T.L., Wang, M., Varma, Y., McFadden, M.E., Brown, L.C., Guo, C., Higginbottom, S., Almo, S.C., and Fischbach, M.A. (2020). A metabolic pathway for bile acid dehydroxylation by the gut microbiome. *Nature* 582, 566–570. <https://doi.org/10.1038/s41586-020-2396-4>.
11. Paik, D., Yao, L., Zhang, Y., Bae, S., D'Agostino, G.D., Zhang, M., Kim, E., Franzosa, E.A., Avila-Pacheco, J., Bisanz, J.E., et al. (2022). Human gut bacteria produce T_H17-modulating bile acid metabolites. *Nature* 603, 907–912. <https://doi.org/10.1038/s41586-022-04480-z>.
12. Sato, Y., Atarashi, K., Plichta, D.R., Arai, Y., Sasajima, S., Kearney, S.M., Suda, W., Takeshita, K., Sasaki, T., Okamoto, S., et al. (2021). Novel bile acid biosynthetic pathways are enriched in the microbiome of centenarians. *Nature* 599, 458–464. <https://doi.org/10.1038/s41586-021-03832-5>.
13. Rimal, B., Collins, S.L., Tanes, C.E., Rocha, E.R., Granda, M.A., Solanki, S., Hoque, N.J., Gentry, E.C., Koo, I., Reilly, E.R., et al. (2024). Bile salt hydrolase catalyses formation of amine-conjugated bile acids. *Nature* 626, 859–863. <https://doi.org/10.1038/s41586-023-06990-w>.
14. Guzior, D.V., Okros, M., Shivel, M., Armwald, B., Bridges, C., Fu, Y., Martin, C., Schilmiller, A.L., Miller, W.M., Ziegler, K.M., et al. (2024). Bile salt hydrolase acyltransferase activity expands bile acid diversity. *Nature* 626, 852–858. <https://doi.org/10.1038/s41586-024-07017-8>.
15. Nie, Q., Luo, X., Wang, K., Ding, Y., Jia, S., Zhao, Q., Li, M., Zhang, J., Zhuo, Y., Lin, J., et al. (2024). Gut symbionts alleviate MASH through a secondary bile acid biosynthetic pathway. *Cell* 187, 2717–2734.e33. <https://doi.org/10.1016/j.cell.2024.03.034>.
16. Gerber, G.K. (2024). AI in microbiome research: Where have we been, where are we going? *Cell Host Microbe* 32, 1230–1234. <https://doi.org/10.1016/j.chom.2024.07.021>.
17. Wang, Z., Koirala, B., Hernandez, Y., Zimmerman, M., and Brady, S.F. (2022). Bioinformatic prospecting and synthesis of a bifunctional lipopeptide antibiotic that evades resistance. *Science* 376, 991–996. <https://doi.org/10.1126/science.abn4213>.
18. Yu, T., Cui, H., Li, J.C., Luo, Y., Jiang, G., and Zhao, H. (2023). Enzyme function prediction using contrastive learning. *Science* 379, 1358–1363. <https://doi.org/10.1126/science.adf2465>.
19. Huang, J., Lin, Q., Fei, H., He, Z., Xu, H., Li, Y., Qu, K., Han, P., Gao, Q., Li, B., et al. (2023). Discovery of deaminase functions by structure-based protein clustering. *Cell* 186, 3182–3195.e14. <https://doi.org/10.1016/j.cell.2023.05.041>.
20. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 379, 1123–1130. <https://doi.org/10.1126/science.adc2574>.
21. Bansal, P., Morgat, A., Axelsen, K.B., Muthukrishnan, V., Coudert, E., Aimo, L., Hyka-Nouspikel, N., Gasteiger, E., Kerhornou, A., Neto, T.B., et al. (2022). Rhea, the reaction knowledgebase in 2022. *Nucleic Acids Res.* 50, D693–D700. <https://doi.org/10.1093/nar/gkab1016>.
22. UniProt Consortium (2023). UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* 51, D523–D531. <https://doi.org/10.1093/nar/gkac1052>.
23. Yao, L., D'Agostino, G.D., Park, J., Hang, S., Adhikari, A.A., Zhang, Y., Li, W., Avila-Pacheco, J., Bae, S., Clish, C.B., et al. (2022). A biosynthetic pathway for the selective sulfonation of steroid metabolites by human gut bacteria. *Nat. Microbiol.* 7, 1404–1418. <https://doi.org/10.1038/s41564-022-01176-y>.
24. Harris, S.C., Devendran, S., Alves, J.M.P., Mythen, S.M., Hylemon, P.B., and Ridlon, J.M. (2018). Identification of a gene encoding a flavoprotein

- involved in bile acid metabolism by the human gut bacterium *Clostridium scindens* ATCC 35704. *Biochim. Biophys. Acta Mol. Cell Biol. Lipids* 1863, 276–283. <https://doi.org/10.1016/j.bbaliip.2017.12.001>.
25. Ridlon, J.M., and Hylemon, P.B. (2012). Identification and characterization of two bile acid coenzyme A transferases from *Clostridium scindens*, a bile acid 7alpha-dehydroxylating intestinal bacterium. *J. Lipid Res.* 53, 66–76. <https://doi.org/10.1194/jlr.M020313>.
26. Doden, H.L., Wolf, P.G., Gaskins, H.R., Anantharaman, K., Alves, J.M.P., and Ridlon, J.M. (2021). Completion of the gut microbial epi-bile acid pathway. *Gut Microbes* 13, 1–20. <https://doi.org/10.1080/19490976.2021.1907271>.
27. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. *Nucleic Acids Res.* 28, 235–242. <https://doi.org/10.1093/nar/28.1.235>.
28. Wang, S., Xie, J., Pei, J., and Lai, L. (2023). CavityPlus 2022 Update: An Integrated Platform for Comprehensive Protein Cavity Detection and Property Analyses with User-friendly Tools and Cavity Databases. *J. Mol. Biol.* 435, 168141. <https://doi.org/10.1016/j.jmb.2023.168141>.
29. Yeturu, K., and Chandra, N. (2008). PocketMatch: A new algorithm to compare binding sites in protein structures. *BMC Bioinformatics* 9, 543. <https://doi.org/10.1186/1471-2105-9-543>.
30. Lucas, L.N., Barrett, K., Kerby, R.L., Zhang, Q., Cattaneo, L.E., Stevenson, D., Rey, F.E., and Amador-Noguez, D. (2021). Dominant Bacterial Phyla from the Human Gut Show Widespread Ability To Transform and Conjugate Bile Acids. Published online August 31, 2021. *mSystems*, e0080521. <https://doi.org/10.1128/mSystems.00805-21>.
31. O'Leary, N.A., Cox, E., Holmes, J.B., Anderson, W.R., Falk, R., Hem, V., Tsuchiya, M.T.N., Schuler, G.D., Zhang, X., Torcivia, J., et al. (2024). Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets. *Sci. Data* 11, 732. <https://doi.org/10.1038/s41597-024-03571-y>.
32. Cantalapiedra, C.P., Hernández-Plaza, A., Letunic, I., Bork, P., and Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol. Biol. Evol.* 38, 5825–5829. <https://doi.org/10.1093/molbev/msab293>.
33. Oberg, N., Zallot, R., and Gerlt, J.A. (2023). EFI-EST, EFI-GNT, and EFI-CGFP: Enzyme Function Initiative (EFI) Web Resource for Genomic Enzymology Tools. *J. Mol. Biol.* 435, 168018. <https://doi.org/10.1016/j.jmb.2023.168018>.
34. Liu, C., Du, M.X., Xie, L.S., Wang, W.Z., Chen, B.S., Yun, C.Y., Sun, X.W., Luo, X., Jiang, Y., Wang, K., et al. (2024). Gut commensal *Christensenella minuta* modulates host metabolism via acylated secondary bile acids. *Nat. Microbiol.* 9, 434–450. <https://doi.org/10.1038/s41564-023-01570-0>.
35. Takei, H., Narushima, S., Suzuki, M., Kakiyama, G., Sasaki, T., Murai, T., Yamashiro, Y., and Nittono, H. (2022). Characterization of long-chain fatty acid-linked bile acids: a major conjugation form of 3beta-hydroxy bile acids in feces. *J. Lipid Res.* 63, 100275. <https://doi.org/10.1016/j.jlr.2022.100275>.
36. Garcia, C.J., Kosek, V., Beltrán, D., Tomás-Barberán, F.A., and Hajslava, J. (2022). Production of New Microbially Conjugated Bile Acids by Human Gut Microbiota. *Biomolecules* 12, 687. <https://doi.org/10.3390/biom12050687>.
37. Gonçalinho, G.H.F., Sampaio, G.R., Soares-Freitas, R.A.M., and Damasceno, N.R.T. (2023). Stearic Acid, but not Palmitic Acid, is Associated with Inflammatory and Endothelial Dysfunction Biomarkers in Individuals at Cardiovascular Risk. *Arq. Bras. Cardiol.* 120, e20220598. <https://doi.org/10.36660/abc.20220598>.
38. Müller, M., Sprenger, G.A., and Pohl, M. (2013). CC bond formation using ThDP-dependent lyases. *Curr. Opin. Chem. Biol.* 17, 261–270. <https://doi.org/10.1016/j.cbpa.2013.02.017>.
39. Laber, B., Maurer, W., Scharf, S., Stepusin, K., and Schmidt, F.S. (1999). Vitamin B6 biosynthesis: formation of pyridoxine 5'-phosphate from 4-(phosphohydroxy)-L-threonine and 1-deoxy-D-xylulose-5-phosphate by PdxA and PdxJ protein. *FEBS Lett.* 449, 45–48. [https://doi.org/10.1016/s0014-5793\(99\)00393-2](https://doi.org/10.1016/s0014-5793(99)00393-2).
40. Xiang, S., Usunow, G., Lange, G., Busch, M., and Tong, L. (2007). Crystal structure of 1-deoxy-D-xylulose 5-phosphate synthase, a crucial enzyme for isoprenoids biosynthesis. *J. Biol. Chem.* 282, 2676–2682. <https://doi.org/10.1074/jbc.M610235200>.
41. Du, Q., Wang, H., and Xie, J. (2011). Thiamin (vitamin B1) biosynthesis and regulation: a rich source of antimicrobial drug targets? *Int. J. Biol. Sci.* 7, 41–52. <https://doi.org/10.7150/ijbs.7.41>.
42. Tian, Y., Gui, W., Koo, I., Smith, P.B., Allman, E.L., Nichols, R.G., Rimal, B., Cai, J., Liu, Q., and Patterson, A.D. (2020). The microbiome modulating activity of bile acids. *Gut Microbes* 11, 979–996. <https://doi.org/10.1080/19490976.2020.1732268>.
43. Kuang, J., Wang, J., Li, Y., Li, M., Zhao, M., Ge, K., Zheng, D., Cheung, K.C.P., Liao, B., Wang, S., et al. (2023). Hyodeoxycholic acid alleviates non-alcoholic fatty liver disease through modulating the gut-liver axis. *Cell Metab.* 35, 1752–1766.e8. <https://doi.org/10.1016/j.cmet.2023.07.011>.
44. Staley, C., Weingarden, A.R., Khoruts, A., and Sadowsky, M.J. (2017). Interaction of gut microbiota with bile acid metabolism and its influence on disease states. *Appl. Microbiol. Biotechnol.* 101, 47–64. <https://doi.org/10.1007/s00253-016-8006-6>.
45. Buffie, C.G., Bucci, V., Stein, R.R., McKenney, P.T., Ling, L., Gobourne, A., No, D., Liu, H., Kinnebrew, M., Viale, A., et al. (2015). Precision microbiome reconstitution restores bile acid mediated resistance to *Clostridium difficile*. *Nature* 517, 205–208. <https://doi.org/10.1038/nature13828>.
46. Kimmel, S.A., and Roberts, R.F. (1998). Development of a growth medium suitable for exopolysaccharide production by *Lactobacillus delbrueckii* ssp. *bulgaricus* RR. *Int. J. Food Microbiol.* 40, 87–92. [https://doi.org/10.1016/s0168-1605\(98\)00023-3](https://doi.org/10.1016/s0168-1605(98)00023-3).
47. Arena, M.E., and Manca De Nadra, M.C. (2002). Comparative survey in *Lactobacillus plantarum* of the growth and metabolism of arginine and citrulline in different media. *J. Agric. Food Chem.* 50, 6497–6500. <https://doi.org/10.1021/jf020383f>.
48. Shen, J., Du, Y., Shen, Y., Kang, N., Fan, Z., Fang, Z., Yang, B., Wang, J., and Li, B. (2025). Integrated Metabolomics and Transcriptomics Analyses Identify Key Amino Acid Metabolic Mechanisms in *Lacticaseibacillus paracasei* SMN-LBK. *Foods* 14, 730. <https://doi.org/10.3390/foods14050730>.
49. Xu, N., Liu, J., Ai, L., and Liu, L. (2015). Reconstruction and analysis of the genome-scale metabolic model of *Lactobacillus casei* LC2W. *Gene* 554, 140–147. <https://doi.org/10.1016/j.gene.2014.10.034>.
50. Hendrikx, T., Duan, Y., Wang, Y., Oh, J.H., Alexander, L.M., Huang, W., Stärkel, P., Ho, S.B., Gao, B., Fiehn, O., et al. (2019). Bacteria engineered to produce IL-22 in intestine induce expression of REG3G to reduce ethanol-induced liver disease in mice. *Gut* 68, 1504–1515. <https://doi.org/10.1136/gutjnl-2018-317232>.
51. Xia, J., Guo, W., Hu, M., Jin, X., Zhang, S., Liu, B., Qiu, H., Wang, K., Zhuge, A., Li, S., et al. (2023). Resynchronized rhythmic oscillations of gut microbiota drive time-restricted feeding induced nonalcoholic steatohepatitis alleviation. *Gut Microbes* 15, 2221450. <https://doi.org/10.1080/19490976.2023.2221450>.
52. Yu, J.S., Youn, G.S., Choi, J., Kim, C.H., Kim, B.Y., Yang, S.J., Lee, J.H., Park, T.S., Kim, B.K., Kim, Y.B., et al. (2021). *Lactobacillus lactis* and *Pediococcus pentosaceus*-driven reprogramming of gut microbiome and metabolome ameliorates the progression of non-alcoholic fatty liver disease. *Clin. Transl. Med.* 11, e634. <https://doi.org/10.1002/ctm2.634>.
53. Li, M., Ding, Y., Wei, J., Dong, Y., Wang, J., Dai, X., Yan, J., Chu, F., Zhang, K., Meng, F., et al. (2024). Gut microbiota metabolite indole-3-acetic acid maintains intestinal epithelial homeostasis through mucin sulfation. *Gut Microbes* 16, 2377576. <https://doi.org/10.1080/19490976.2024.2377576>.
54. Chen, B., Zeng, G.Y., Sun, L.L., and Jiang, C.T. (2024). When smoke meets gut: deciphering the interactions between tobacco smoking and gut microbiota in disease development. *Sci. China Life Sci.* 67, 854–864. <https://doi.org/10.1007/s11427-023-2446-y>.

55. Perino, A., and Schoonjans, K. (2022). Metabolic Messengers: bile acids. *Nat. Metab.* 4, 416–423. <https://doi.org/10.1038/s42255-022-00559-z>.
56. Collins, S.L., Stine, J.G., Bisanz, J.E., Okafor, C.D., and Patterson, A.D. (2023). Bile acids and the gut microbiota: metabolic interactions and impacts on disease. *Nat. Rev. Microbiol.* 21, 236–247. <https://doi.org/10.1038/s41579-022-00805-x>.
57. Fuchs, C.D., and Trauner, M. (2022). Role of bile acids and their receptors in gastrointestinal and hepatic pathophysiology. *Nat. Rev. Gastroenterol. Hepatol.* 19, 432–450. <https://doi.org/10.1038/s41575-021-00566-7>.
58. Ridlon, J.M., and Gaskins, H.R. (2024). Another renaissance for bile acid gastrointestinal microbiology. *Nat. Rev. Gastroenterol. Hepatol.* 21, 348–364. <https://doi.org/10.1038/s41575-024-00896-2>.
59. Luo, X., Wang, K., and Jiang, C. (2025). Gut microbial enzymes and metabolic dysfunction-associated steatohepatitis: Function, mechanism, and therapeutic prospects. *Cell Host Microbe* 33, 836–853. <https://doi.org/10.1016/j.chom.2025.04.020>.
60. Lee, M.H., Nuccio, S.P., Mohanty, I., Hagey, L.R., Dorresteijn, P.C., Chu, H., and Raffatellu, M. (2024). How bile acids and the microbiota interact to shape host immunity. *Nat. Rev. Immunol.* 24, 798–809. <https://doi.org/10.1038/s41577-024-01057-x>.
61. Holick, M.F. (2023). The One-Hundred-Year Anniversary of the Discovery of the Sunshine Vitamin D3: Historical, Personal Experience and Evidence-Based Perspectives. *Nutrients* 15, 593. <https://doi.org/10.3390/nut15030593>.
62. Hamid, R., Adam, S., Lacour, A., Monjas, L., Köhnke, J., and Hirsch, A.K. H. (2023). 1-deoxy-D-xylulose-5-phosphate synthase from *Pseudomonas aeruginosa* and *Klebsiella pneumoniae* reveals conformational changes upon cofactor binding. *J. Biol. Chem.* 299, 105152. <https://doi.org/10.1016/j.jbc.2023.105152>.
63. Li, Y., Wang, S., Umarov, R., Xie, B., Fan, M., Li, L., and Gao, X. (2018). DEEPRe: sequence-based enzyme EC number prediction by deep learning. *Bioinformatics Oxf. Engl.* 34, 760–769. <https://doi.org/10.1093/bioinformatics/btx680>.
64. Hekkelman, M.L., de Vries, I., Joosten, R.P., and Perrakis, A. (2023). AlphaFill: enriching AlphaFold models with ligands and cofactors. *Nat. Methods* 20, 205–213. <https://doi.org/10.1038/s41592-022-01685-y>.
65. Wang, X., Yin, X., Jiang, D., Zhao, H., Wu, Z., Zhang, O., Wang, J., Li, Y., Deng, Y., Liu, H., et al. (2024). Multi-modal deep learning enables efficient and accurate annotation of enzymatic active sites. *Nat. Commun.* 15, 7348. <https://doi.org/10.1038/s41467-024-51511-6>.
66. Ribeiro, A.J.M., Holliday, G.L., Furnham, N., Tyzack, J.D., Ferris, K., and Thornton, J.M. (2018). Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. *Nucleic Acids Res.* 46, D618–D623. <https://doi.org/10.1093/nar/gkx1012>.
67. Ryu, J.Y., Kim, H.U., and Lee, S.Y. (2019). Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. *Proc. Natl. Acad. Sci. USA* 116, 13996–14001. <https://doi.org/10.1073/pnas.1821905116>.
68. Liu, Y., H., C., Zeng, T., Rao, J., Zhang, Z., Wu, R., Coley, C.W., and Zheng, S. (2024). EnzymeCAGE: A Geometric Foundation Model for Enzyme Retrieval with Evolutionary Insights. Preprint at bioRxiv. <https://doi.org/10.1101/2024.12.15.628585>.
69. Li, Q.i., Sun, B., Chen, J., Zhang, Y., Jiang, Y.u., and Yang, S. (2021). A modified pCas/pTargetF system for CRISPR-Cas9-assisted genome editing in *Escherichia coli*. *Acta Biochim. Biophys. Sinica* 53, 620–627. <https://doi.org/10.1093/abbs/gmab036>.
70. Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W.S., and Huttenhower, C. (2011). Metagenomic biomarker discovery and explanation. *Genome Biol.* 12, R60. <https://doi.org/10.1186/gb-2011-12-6-r60>.
71. Wu, Q., Liang, X., Wang, K., Lin, J., Wang, X., Wang, P., Zhang, Y., Nie, Q., Liu, H., Zhang, Z., et al. (2021). Intestinal hypoxia-inducible factor 2 α regulates lactate levels to shape the gut microbiome and alter thermogenesis. *Cell Metab.* 33, 1988–2003.e7. <https://doi.org/10.1016/j.cmet.2021.07.007>.
72. McIver, L.J., Abu-Ali, G., Franzosa, E.A., Schwager, R., Morgan, X.C., Waldron, L., Segata, N., and Huttenhower, C. (2018). bioBakery: a meta'omic analysis environment. *Bioinformatics Oxf. Engl.* 34, 1235–1237. <https://doi.org/10.1093/bioinformatics/btx754>.
73. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. <https://doi.org/10.1038/nmeth.1923>.
74. Lin, J., Nie, Q., Cheng, J., Zhong, Y.N., Zhang, T., Zhang, X., Ge, X., Ding, Y., Niu, C., Gao, Y., et al. (2025). A microbial amino-acid-conjugated bile acid, tryptophan-cholic acid, improves glucose homeostasis via the orphan receptor MRGPRE. *Cell*. Published online May 22, 2025. <https://doi.org/10.1016/j.cell.2025.05.010>.
75. Human Microbiome Project Consortium (2012). A framework for human microbiome research. *Nature* 486, 215–221. <https://doi.org/10.1038/nature11209>.
76. Wood, D.E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biol.* 20, 257. <https://doi.org/10.1186/s13059-019-1891-0>.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial strains		
<i>E. coli</i> BL21 (DE3)	Thermo Scientific	Cat# EC0114
<i>E. coli</i> BL21 (DE3) Δ hdhA	This study	N/A
<i>E. coli</i> DH5 α	Thermo Scientific	Cat# EC0112
<i>E. coli</i> S17-1 λ pir	Beyotime	Cat# D1075S
Other strains used in this study	This study	Table S3
Chemicals, peptides, and recombinant proteins		
Ampicillin	Sigma-Aldrich	Cat# BP021
Kanamycin	Sigma-Aldrich	Cat# D403
Methanol	Sigma-Aldrich	Cat# 439193
Acetonitrile	Sigma-Aldrich	Cat# 34851
Formic acid	Sigma-Aldrich	Cat# F0507
Fluoromethalone	MedChemExpress	Cat# HY-B1893
EtOAc	Beijing FreeMore bioscience	Cat# E116131
DCM	AcmeC	Cat# D12641
Hexane	Beijing FreeMore bioscience	Cat# H109656
Cholic acid	Aladdin	Cat# C103690
Deoxycholic acid	Aladdin	Cat# D103697
Chenodeoxycholic acid	Aladdin	Cat# C104902
Lithocholic acid	Aladdin	Cat# L106779
3-oxodeoxycholic acid	Aladdin	Cat# O333250
Sodium taurocholate hydrate	Aladdin	Cat# T134625
Cholic acid-2,2,4,4-d4	Sigma-Aldrich	Cat# 614149
3-acetodeoxycholic acid	This study	N/A
3-acetylcholic acid	Lab stock	N/A
3-propionylcholic acid	Lab stock	N/A
3-butyrylcholic acid	Lab stock	N/A
3-succinylated cholic acid	Lab stock	N/A
GAM broth, modified	Solarbio	Cat# LA4490
Defibrinated sheep blood	Solarbio	Cat# TX0030
Clarified rumen fluid	PERFEMIKER	Cat# PZ01010
BHI medium	Solarbio	Cat# B8130
TSB medium	Solarbio	Cat# RL100111
YCFA medium	Solarbio	Cat# LA4040
MRS medium	Solarbio	Cat# M8540
Yeast extract	Oxoid	Cat# LP0021
Tryptone	Oxoid	Cat# LP0042
ThDP	Aladdin	Cat# T111202
MgCl ₂	Sigma-Aldrich	Cat# 208337
Tris	Solarbio	Cat# T8060
Imidazole	Solarbio	Cat# I8090
IPTG	Solarbio	Cat# I8070
Thrombin	Solarbio	Cat# T8021
NaCl	Macklin	Cat# S805275

(Continued on next page)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Glycerol	Macklin	Cat# G810582
FastDigest <i>DpnI</i>	ThermoFisher Scientific	Cat#FD1703
Critical commercial assays		
MultiF Seamless Assembly Mix	Abclonal	Cat# RK21020
2 × Phanta Flash Master Mix (Dye Plus)	Vazyme	Cat# P520-01
FastPure Gel DNA Extraction Mini Kit	Vazyme	Cat# DC301-01
Lipofectamine 3000 transfection reagent	ThermoFisher Scientific	Cat# L3000015
Dual-luciferase assay system	Promega	Cat# E1910
Deposited data		
Code	This paper	Github or Zenodo
Experimental models: Cell lines		
HEK293T	ATCC	Cat# CRL-3216
Experimental models: Organisms/strains		
Mouse: C57BL/6J	GemPharmatech	N/A
Oligonucleotides		
Primers used in this study	This study	Table S3
Recombinant DNA		
pET-28a	Lab stock	N/A
pEcCas	Addgene	Cat# 73227
pEcRNA	Addgene	Cat# 166581
pGL4-Shp-TK firefly luciferase construct	Lab stock	N/A
Human FXR expression plasmid	Lab stock	N/A
Human VDR expression plasmid	Lab stock	N/A
Human PXR expression plasmid	Lab stock	N/A
Human RXR expression plasmid	Lab stock	N/A
Human ASBT expression plasmid	Lab stock	N/A
<i>Renilla</i> luciferase control vector	Lab stock	N/A
pCMVSPORT6/hTGR5	Lab stock	N/A
cAMP response element-driven luciferase reporter plasmids	Lab stock	N/A
Software and algorithms		
GraphPad Prism version 9.0	GraphPad	https://www.graphpad-prism.cn/
Mega 7	Mega	http://www.megasoftware.net/
SPSS 27.0	SPSS	https://www.ibm.com/spss/
Cytoscape 3.10.3	Cytoscape	https://cytoscape.org/
SnapGene Viewer	SnapGene	http://www.snapgene.com/
Blast	NCBI	https://blast.ncbi.nlm.nih.gov/
NCBI Dataset tool	NCBI	https://www.ncbi.nlm.nih.gov/datasets/docs/v2/download-and-install/
Python 3.9	Python Software Foundation	https://www.python.org/
Cavity	Cavity	http://162.105.160.28:2017/cavityplus/#/computation
DIAMOND	DIAMOND	https://github.com/bbuchfink/diamond
eggNOG-mapper 2.1.12	eggNOG-mapper	http://eggnog-mapper.embl.de/
EFI-EST	EFI-EST	https://efi.igb.illinois.edu/efi-est/
CLEAN	CLEAN	https://github.com/ttianhao/CLEAN
AlphaFold2	ColabFold	https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb

(Continued on next page)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
ESMFold	ESMFold	https://github.com/facebookresearch/esm
AutoDock Vina 1.1.2.9.	AutoDock Vina	https://vina.scripps.edu/
XDS	XDS	https://xds.mr.mpg.de/
PHENIX 1.19	PHENIX	https://phenix-online.org/
WinCoot 0.9.8.95	WinCoot	https://bernhardcl.github.io/coot/
Pymol 2.5.5	Pymol	https://pymol.org/
R-4.4.1	R	https://cran.r-project.org/
RStudio 2024.04.2+764	RStudio	https://posit.co/download/rstudio-desktop/
Deposited data		
Metagenome data for healthy volunteers	This paper	NMDC: NMDC10019873
Non-targeted metabolome data for 3-acetoDCA identification	This paper	NMDC: NMDCX0002114
Other		
NMR system	Bruker	Cat# Avance-500
Quaternary HPLC System	Agilent	Cat# 1200 Infinity
ACQUITY UPLC CSH C18 column	Waters	Cat# 186005297
QTRAP 5500	AB SCIEX	N/A
HisTrap HP	Cytiva	Cat# 17524801
PD10 column	Cytiva	Cat# 17085101
RESOURCE Q	Cytiva	Cat# 17117901
ÄKTA FPLC system	Cytiva	Cat# 29383015

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS**Human study**

The human study was approved by the Ethical Committee of Peking University Third Hospital according to the Council for International Organization of Medical Sciences. The exclusion criteria were taking antibiotics, probiotics, prebiotics, proton pump inhibitors, and laxatives taken during the previous 3 months. A total of 30 healthy volunteers, comprising 15 males (aged 24-49 years) and 15 females (aged 24-60 years), were enrolled in the study. All individuals included in the study are of Han Chinese ethnicity. And their clinical information was collected according to standard procedures. Stool samples from the healthy volunteers were collected in collection cups and immediately frozen at -80 °C until analysis. Informed written consent was obtained from all participants.

Mice and Treatments

All animal protocols were approved by the Animal Care and Use Committee of Peking University (permit: PUIRB-LA2023297). 8-week-old male specific-pathogen-free (SPF) mice of C57BL/6J were purchased from GemPharmatech and maintained under standard laboratory conditions (a strict 12-hour light/dark cycle, a controlled temperature of $23 \pm 2^\circ\text{C}$ and a humidity at 40–70%) with free access to food and water at the Department of Experimental Animal Science, Peking University Health Science Center. All mice were randomly assigned to experimental groups, and no mice were excluded from analysis.

To test the role of 3-acetoDCA on the composition of gut microbiota, 8-week-old SPF male mice were administered with phosphate-buffered saline (PBS), 10 mg/kg 3-acetoDCA by oral gavage 3 times per week for one week, feces were collected for 16S rRNA gene amplicon sequencing and detection of the abundance of *L. gasseri* and level of tryptophan derivatives.

Bacteria

All strains used in this study are listed in **Table S3**. These strains were isolated from fecal sample of health volunteers and cultivated in modified Gifu anaerobic medium (mGAM) at 37°C under anaerobic conditions (gas mix: 70% N₂, 25% CO₂, 5% H₂) unless otherwise stated. Individual strains were inoculated from glycerol stocks into 5 mL mGAM liquid media and incubated for 48 h to reach the stationary phase, which served as starter cultures.

For the screening of gut microbes functioning in biosynthesis of 3-acetoDCA, the starter cultures were diluted 1:100 into 5 mL of fresh mGAM medium containing 10 μM 3-oxoDCA. Cultures were grown at 37°C for 48 h and harvested for further detection of 3-acetoDCA formation by LC-MS. All experiments were performed in triplicate unless otherwise stated.

METHOD DETAILS

Development of BEAUT workflow

Collection of primary positive samples and negative samples

We collected 120 metabolic reactions from RHEA that contain at least one reaction participant identified as a BA or its derivative. All UniProt sequences associated with these reactions are collected as positive sequences. We also added 12 sequences collected from literature till 2024,^{10,12,23–26} arriving at a total of 1032 positive samples. Fragments were filtered out, leaving 469 complete positive sequences.⁶³ Redundant sequences were removed at 90% sequence identity through an all-by-all BLAST using DIAMOND, which gave 151 unique primary positive sequences. The ultra-sensitive mode was used in the BLAST calculation. The SwissProt sequences associated with the remaining RHEA reactions are used as negative samples. We kept sequences which have 157 to 1,074 amino acids to fit with the minimum and maximum lengths found in primary positive samples, which gave 102,403 negative sequences. Note that all sequences collected are enzymes, whose substrates are small molecules, which indicates that the model alone cannot discriminate enzyme/non-enzyme sequences or identify enzymes acting on biomacromolecules like proteins or nucleic acids. Separate filter workflows are applied to avoid these sequences in our prediction results.

Selection of potential substrate binding pockets from primary positive samples

119 structures of 151 primary positive sequences were obtained from AlphaFold protein structure database. The remaining 32 structures were predicted by ESMFold. The 151 structures were filter by their average pLDDT, keeping those with an average pLDDT ≥ 70 . 147 structures passed the filtering. Cavity was then used to extract pockets from the structures. To avoid missing some large pockets, Cavity was run with the default settings and with the “rescue large pockets” option. Next, 137 potential substrate binding pockets were successfully identified and manually selected (see the section below and Table S1). We calculated the ratio of amino acid residues with pLDDT ≥ 90 among all residues in the pocket to discriminate high quality pockets. For the ESMFold predicted structures, the pLDDT for any residue was calculated as the mean pLDDT among all atoms of the residue. We required the extracted pockets to have the ratio ≥ 0.7 . Finally, 85 substrate binding pockets were kept. These pockets are referred to as the reference pockets.

Manual selection of substrate binding pockets from structures of primary positive samples

The structures were queried in UniProt to retrieve their corresponding reactions. UniProt also provided information about possible binding co-factors and annotated binding and catalytic sites. If the enzyme binds a co-factor, in most cases the pocket where it binds was selected as the substrate binding pocket because the substrate should locate closely to the co-factor for the reaction to occur. AlphaFill⁶⁴ was used through the selection process to locate the co-factors in the structures. If the enzyme has annotated catalytic sites, the substrate binding pocket was selected to fully cover these sites. If the enzyme needs no co-factor and there are no annotated binding sites, we used EasIFIA⁶⁵ for enzymes shorter than 600 amino acids to predict the catalytic residues. For a few such enzymes that are longer than 600 amino acids, we made assumptions about the catalytic mechanisms of the enzymes and compared the residues in the pockets with enzymes bearing similar functions based on the M-CSA⁶⁶ database. For a few structures, the above process still could not determine the substrate binding pocket. No pocket was selected from these structures.

Collection of augmentation sequences for BEAUT model

We downloaded the corresponding translated protein sequences of the 7 bacterial genomes using the NCBI Dataset tool. First, we filtered the sequences by workflow I (described below) to keep potential enzymes with small molecules as substrates only. Then we ran BLAST using the enzyme sequences as query and all negative samples as target to filtered out any potential negative sequences. Sequences with maximum identity $\geq 50\%$ were removed. Protein structures were predicted with ESMFold. Structures with an average pLDDT ≥ 70 were subject to pocket detection by the Cavity program. Here, we define the pocket index to be the ratio of C α atoms with pLDDT ≥ 80 among all C α atoms in the pocket. Cavity was used with the same two settings as in extracting pockets from primary positive samples to extract pockets from these structures. Pockets with volumes between 1000–5500 Å³ which fit with the volumes of substrate binding pockets from primary positive samples and an index ≥ 0.7 were used as the query pockets for similarity calculations. Sequences from which no such pocket could be extracted were discarded.

We used PocketMatch to compute the similarity scores between the query pockets and the reference pockets. Sequences whose pockets have a maximum similarity score ≥ 0.7 compared with all reference pockets were selected as augmentation sequences.

Parameter searching for removing possible negative samples from the genomes

We followed the idea adopted by DeepEC⁶⁷ and divided the negative samples into two sets of approximately equal sizes. One set was merged with the unique primary positive sequences and used as queries in the BLAST calculation against the other set of negative sequences. Two parameters were optimized in the search: the minimum sequence identity and query coverage. Both parameters took values from 0 to 90% with an interval of 10%. The objective was higher recall of the negative samples and fewer matches of positive sequences. We finally chose the minimum sequence identity to be 50% and the query coverage to be 0. This setting could recall about 90% of the negative samples and keep the matched positive samples as few as 20 sequences.

Extraction of substrate binding pockets from negative samples

Proteins structures of the negative samples were downloaded from the AlphaFold Protein Structure Database and only structures with average pLDDT ≥ 70 were kept. We used Cavity with default settings to extract all pockets in the structures. We utilized the substrate pocket data set developed by EnzymeCAGE⁶⁸ as reference which includes one potential substrate pocket for one structure. Because the methods for pocket extraction are different between Cavity and EnzymeCAGE, we calculated the pairwise Dice

similarity between each Cavity pocket for a given structure and the EnzymeCAGE pocket for the structure. The Dice similarity is defined between two sets of amino acid residues, A and B, each including those residues present in the pocket and the residues adjacent to the former:

$$\text{Dice similarity} = \frac{|A \cap B|}{|A \cup B|}$$

We selected the most similar pocket that had a similarity ≥ 0.5 among all Cavity pockets for a given structure as the substrate pocket. The pockets were then filtered to keep those whose volumes were between 1000-5500 Å³ and with the ratio of amino acid residues with pLDDT ≥ 90 among all residues in the pocket ≥ 0.7 . Finally, 14,068 pockets were extracted from the negative samples.

Comparing substrate binding pockets from enzymes with and without bile acid substrates

We computed all pairwise substrate pocket similarity using PocketMatch for the group of 85 reference pockets from the positive samples and for the group of 14068 pockets from the negative samples, respectively. We plotted the distributions of the pairwise similarity within each group.

Workflows to exclude non-enzymes and enzymes with biomacromolecules as substrates

We designed two workflows to exclude non-enzymes and enzymes involving biomacromolecules as substrates, Workflow I for filtering genome-derived sequences used in data augmentation and Workflow II for filtering the prediction results obtained by the BEAUT model. The workflows used eggNOG-mapper to acquire functional annotations first and to analyze the annotations to identify enzyme sequences utilizing small molecular substrates.

Workflow I first kept sequences with annotated EC numbers. The rest of the sequences were filtered by the annotated COG category. Sequences belonging to the C, E, F, G, I, Q, R, S categories were kept for further analysis. Then the workflow checked if there was at least one '-ase' descriptor in the sequence annotation, and sequences with descriptions containing -ase words related to biomacromolecules, like proteins, nucleic acids, polysaccharides, etc., were filtered out. Sequences passing the filter were kept. Sequences which had no descriptions or could not be annotated were kept without further consideration.

Workflow II first keeps sequences with annotated EC numbers. The rest of the sequences were checked if there was at least one '-ase' descriptor present in the sequence annotation, and those containing -ase words that are related to biomacromolecules, like proteins, nucleic acids, polysaccharides, etc. were filtered out. Sequences that passed the filter were kept. Sequences which have no descriptions or could not be annotated were kept without further consideration.

Data set partition strategy

The augmentation sequences were merged with the primary positive sequences and deduplicated at 90% sequence identity by performing an all-by-all BLAST with DIAMOND, leaving 2,472 sequences (augmented positive sequences) for model development of BEAUT. The ultra-sensitive mode was used in the BLAST calculation. These sequences were then clustered at 30% sequence identity. We used a graph-based clustering method, treating each sequence as one node and adding edges between nodes if the nodes have sequence identity $\geq 30\%$. We used the connected components in the graph as clusters, which guaranteed that we could sample sequences of the test set with maximum sequence identity $< 30\%$ to the training set in cross-validation. We randomly selected whole clusters which contained 247 positive samples for testing. For the remaining 2,225 positive samples, we calculated the distribution of their lengths and divided them into five sets. For each set, randomly sampled negative samples 5 times the size of the set and with nearly identical distribution of sequence length of the set were added. We performed a stratified 5-fold cross-validation. For each round of cross-validation, one set of samples was used as validation set and the remaining 4 sets were combined and used as the training set. [Table S2](#) shows the number of positive and negative samples in the training and validation sets used in cross-validation.

91,278 negative samples not used in the training or validation set were considered for the test set. We performed BLAST with DIAMOND using these sequences as queries and the 11,125 negative sequences used in cross-validation as target. The ultra-sensitive mode was used and all matches were reported (parameter -k 0). We sampled 1,235 negative sequences with $< 30\%$ sequence identity to the 11,125 sequences. The 1,235 sequences were merged with the 247 positive sequences for testing to form a balanced test set.

Model architecture and hyperparameters

The BEAUT model is a dense neural network. The model takes the ESM-2 embedding of a given sequence with 1280 dimensions as input, passes through a hidden layer with 256 dimensions and a hidden layer of 32 dimensions and predicts the probability that the sequence is a positive or negative with a softmax function. ReLU non-linearities are used for the hidden layers. The model was trained to minimize the cross-entropy loss function as commonly used by classification models. The model was implemented in PyTorch. Adam optimizer with 0.0001 weight decay was used. The learning rate was set to be 0.0002. We used ReduceLROnPlateau as the learning rate scheduler with a factor of 0.2, a patience of 3 epochs and a minimum learning rate of 5e-6. Early stopping was used when the F1-score on the validation set does not improve over 5 epochs. The training batch size is 64.

Model evaluation metrics

Area under the precision-recall curve (AUPRC), precision, recall, F1 score and Matthew correlation coefficient (MCC) were used to evaluate different models on the corresponding test set. All metrics were calculated with the scikit-learn package in Python. AUPRC was calculated using the predicted probabilities from the models. The other metrics were calculated as:

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$F1 \text{ score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision}+\text{Recall}}$$

$$MCC = \frac{TN \times TP - FN \times FP}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

where TN, TP, FN, FP are the true negatives, true positives, false negatives, false positives in the model predictions, respectively.

Human gut microbe metagenomic dataset screening

The Genbank protein sequences from NCBI BioProject PRJNA28331 were downloaded using the NCBI Dataset tool. Only genomes whose assembly levels are above the scaffold level were considered. 1923 genomes were kept for further processing. A few sequences were removed because they were annotated partial in Genbank. We kept sequences whose lengths lie between 157 to 1074, which was the same as the positive sequences. Finally, we arrived at 2,340,761 unique sequences. We predicted the probabilities using the BEAUT model. Sequences with a predicted score of the positive class higher than that of the negative class were considered positive. The positive predictions were subjected to workflow II to filter out non-enzymes and enzymes with biomacromolecules as substrates. Finally, 614,616 sequences were predicted by the BEAUT model.

Targeted metabolomics analysis

Quantification of different bile acids was performed by a liquid chromatography-tandem mass spectrometry (LC-MS/MS) system that composed of a Acuity Ultra-Performance Liquid Chromatography (UPLC) system (Waters Corporation, Milford, USA) coupled to a SCIEX 5500 triple quadrupole linear ion trap mass spectrometer (AB SCIEX, Framingham, MA, USA.). Chromatographic separation was employed with an ACQUITY UPLC CSH C18 column (2.1×100 mm, $1.7 \mu\text{m}$, Waters) at 40°C and a flow rate of 0.25 mL/min . The injection volume was $5 \mu\text{L}$, Mobile phase A was 0.1% formic acid (FA) in water and mobile phase B was 0.1% FA in acetonitrile. All analytes were detected in negative ion multiple reaction monitoring (MRM) mode. Chromatographic separation was performed using a linear gradient as follows: 0-1 min, 40% B; 1-6 min, 40-100% B; 6-7 min, 100% B; 7-8 min, 100-40% B. The detailed MRM parameters was listed in [Table S5](#). The standard curves with internal calibration were constructed using mixed working standard solutions with gradient dilution (0.1, 1, 10, 100, 1000 nM), CA-d4 ($0.2 \mu\text{M}$) was used as internal standard for each concentration. Operational control of the LC-MS/MS was performed with Analyst version 1.6.2, and quantitative analysis was performed using MultiQuant software (version 3.0.1).

For the quantification of bile acids from feces, accurately weighed samples (20 mg) were placed in tube and add $400 \mu\text{L}$ of ice-cold methanol containing $0.2 \mu\text{M}$ internal standards. The samples were thoroughly homogenized with a tissue homogenizer. Samples were subsequently ultrasonic extraction for 20 min for metabolites extraction. All samples were subsequently centrifuged at 14,000 rpm for 20 min at 4°C . The supernatant was filtered through a membrane filter (pore size, $0.22 \mu\text{m}$) for LC-MS analysis.

For the quantification of bile acids from culture supernatant from bacteria, samples were diluted with ice-cold methanol (1:4) containing $0.2 \mu\text{M}$ internal standards (CA-d4). Then the samples were vortex for 10 min and were incubated at -20°C for 1 h. All samples were subsequently centrifuged at 14,000 rpm for 20 min at 4°C . The supernatant was filtered through a membrane filter (pore size, $0.22 \mu\text{m}$) for LC-MS/MS analysis.

Detection of bile acid metabolic capacity in different strains

Bile acid metabolic capacity refers to the ability of microorganisms to consume bile acids. Here, we provide the following mathematical formula for quantitative calculation:

$$\text{Bile acid metabolic capacity} = 1 - \frac{RC}{IC}$$

where RC is the rest bile acids content, IC is the initial bile acid content.

To detect bile acid metabolic capacity of different gut microbes, 10 μ L starter cultures were inoculated individually into 96-well deep plates added with 200 μ L fresh appropriate medium (mGAM/MRS/BHI) supplemented with different substrates at indicated concentrations (100 μ M for CA/CDCA/DCA/LCA and 10 μ M for 3-oxoDCA) and incubated anaerobically at 37°C for 48 h. The cultures were harvested through dilution with ice-cold methanol, vortex for 10 min, and centrifugation at 14,000 rpm for 20 min at 4°C. The supernatant was filtered through a membrane filter (pore size, 0.22 μ m) for LC-MS/MS analysis to detect bile acid metabolic capacity.

Construction of *Escherichia coli* BL21 (DE3) knockout strain

CRISPR-Cas9-assisted strategy based on two-plasmid system pEcCas/pEcRNA was applied for *hdhA* (annotated as 7-HSDH) knockout of *E. coli* BL21 (DE3).⁶⁹ In brief, the 400 bp upstream and downstream of *hdhA* fragments were respectively amplified through primer pairs *hdhA*-UF/*hdhA*-UR and *hdhA*-DF/*hdhA*-DR from *E. coli* BL21 (DE3), then fused into dDNA-*hdhA* through overlap extension PCR. The N20-specific sequence was designed via chopchop (chopchop.cbu.uib.no), and cloned into pEcRNA to yield pEcRNA-*hdhA*, the cassette *ccdB* of which was replaced by designed N20-specific sequence.

For *hdhA* knockout, pEcCas was firstly introduced into *E. coli* BL21 (DE3) via electroporation. Subsequently, the cells were cultured in lysogeny broth (LB) medium containing 50 μ g/mL kanamycin for 4 h, with the addition of 10 mM arabinose to induce the expression of λ -Red system for competent cell preparation. Then, the mixture containing 100 μ L competent cells, 100 ng of pEcRNA-*hdhA* plasmids and 400 ng of dDNA-*hdhA* was electroporated in a precooled 2 mm Gene Pulser cuvette (Bio-Rad, Hercules, USA) at 2.5 kV. Cells were recovered at 37°C for 1 h, spread on LB plates supplemented with 50 μ g/mL kanamycin and spectinomycin, and incubated overnight at 37°C. The *hdhA*-knockout colonies were verified by PCR and sequencing.

To cure plasmids pEcCas and pEcRNA-*hdhA*, the positive mutants were cultured in LB medium containing 10 mM rhamnose and 50 μ g/mL kanamycin for 16 h, transferred to fresh LB for 2 hours, and streaked on LB plates containing 50 g/L sucrose. Single colonies were assessed for their sensitivity to kanamycin and spectinomycin to verify plasmids elimination.

Activity test of bile acid related enzyme candidates

The bile acid related enzyme candidates (BE1 - BE104) were synthesized by the Sangon Biotech and cloned into pET28a. Confirmed plasmids were transformed into *E. coli* BL21 (DE3) Δ *hdhA* strains. Single colony for each gene was incubated in 5 mL LB medium containing 50 μ g/mL kanamycin at 37°C, 220 rpm overnight. And then 40 μ L of the cultures were inoculated respectively into 4 mL fresh LB medium with 50 μ g/mL kanamycin. When the OD₆₀₀ reached 0.6, 0.25 mM isopropyl β -d-1-thiogalactopyranoside (IPTG) was added to induce the protein expression. Meanwhile, bile acids (100 μ M for CA/CDCA/DCA/LCA and 10 μ M for 3-oxoDCA) were supplemented into the cultures individually. Cultures were cultured for additional 16 hours before being harvested for detection of bile acids consumption by LC-MS.

Heterologous expression and activity of MABH candidates

The candidates (BE117 - BE136) were amplified from genome of *H. hathewayi* with primers listed in Table S3 and cloned into pET28a. Confirmed plasmids were transformed into *E. coli* BL21 (DE3) Δ *hdhA* strains. Single colony for each aimed gene was incubated in 5 mL LB medium containing 50 μ g/mL kanamycin at 37°C, 220 rpm overnight. And then 40 μ L of the cultures were inoculated respectively into 4 mL fresh LB medium with 50 μ g/mL kanamycin, 10 μ M 3-acylated bile acids. When the OD₆₀₀ reached 0.6, 0.25 mM IPTG was added to induce the protein expression. Cultures were cultured for additional 16 h before being harvested for detection of 3-acylated bile acids hydrolysis by LC-MS.

Extraction and isolation of 3-acetoDCA

The *E. coli* BL21 (DE3) harboring pET28a-BE1 was used as 3-acetoDCA producer. The cell pellet of 6 L culture broth of IPTG-induced BL21 (DE3) pET28a-BE1 was collected by centrifugation and resuspended by 100 mL PBS containing 10 mM 3-oxoDCA and 20 mM pyruvate. Then the resuspended biotransformation mix was incubated at 37°C 220 rpm for another 10 hours. The biotransformation mix was extracted repeatedly with isometric ethyl acetate three times, and the organic solvent was evaporated under vacuum to yield the crude extract (0.8 g). The crude extract was subsequently separated on a silica gel column (10.0 g of silica gel) eluted with hexane-ethyl acetate via gradient elution (v/v, 1:0, 10:1, each 100 mL) and then with dichloromethane-methanol (v/v, 1:0, 20:1, 10:1, 5:1, 3:1, 0:1, each 50 mL). The presence of 3-acetoDCA was detected in the above fractions using LC-MS. Fraction eluted with dichloromethane-methanol (v/v, 10:1) which contained 3-acetoDCA was further separated by HPLC-ELSD. The pure 3-acetoDCA was eluted by H₂O (0.1%FA)-ACN (0.1%FA) (v/v, 40:60).

X-ray Crystallographic Analysis of 3-acetoDCA: Crystal of 3-acetoDCA was obtained by slow evaporation from 50% aqueous MeOH in a cold room (4°C). Crystallographic data were collected with CuK α radiation (1.54184 Å, 2 θ range from 8.326 to 155.21°) at 100 K on a ROD, Synergy Custom DW system, HyPix diffractometer. Using Olex2, the structure was solved with the SHELXT structure solution program using Intrinsic Phasing and refined with the SHELXL refinement package using Least Squares minimisation. Crystal Data for 3-acetoDCA: C₂₆H₄₂O₅ (M = 434.59 g/mol): monoclinic, space group P21 (no. 4), a = 7.24230(10) Å, b = 15.43920 (10) Å, c = 10.64120(10) Å, β = 93.6350(10)°, V = 1187.46(2) Å³, Z = 2, T = 100.00(10) K, μ (Cu K α) = 0.654 mm⁻¹, D_{calc} = 1.215 g/cm³, 26282 reflections measured (8.326° ≤ 2 θ ≤ 155.21°), 4765 unique (R_{int} = 0.0209, R_{sigma} = 0.0121) which were used in all calculations. The final R 1 was 0.0249 ($I > 2\sigma(I)$) and wR 2 was 0.0661 (all data). The goodness of fit on F2 was 1.028. Flack parameter = 0.01(3).

Construction of the *ads* knockout mutant of *Bacteroides ovatus*

An allelic replacement inactivation strategy based on double-crossover was used to inactivate the *B. ovatus* gene *ads*. The 798 bp upstream fragment and the 778 bp downstream fragment of *ads* were amplified using primer pairs boAds-inUF/boAds -inUR and boAds-inDF/boAds-inDR from *B. ovatus* and the erythromycin resistance gene (*erm*) amplified using primer pair erm-inF/erm-inR from plasmid pGERM. The Ads-Up/erm/Ads-Down fragments were cloned into plasmid pUC19 to afford the *ads* disruption plasmid, which was then introduced into *B. ovatus* via *E. coli*-*Bacteroides* conjugation. After 24 h anaerobic incubation at 37°C, the transconjugants were selected on gentamicin (200 µg/mL) and erythromycin (25 µg/mL) GAM plates. The colonies of *ads* disrupted strains which were then verified by PCR and sequencing.

Gene expression and protein purification of BoADS

The BoADS were synthesized by the Sangon Biotech and cloned into pET28a forming pET28a-Boads. The plasmid pET28a-Boads was transformed into *E. coli* BL21 (DE3). Overexpression was performed using LB medium with 50 µg/mL kanamycin at 37°C, 220 rpm until an OD₆₀₀ of 0.6 was reached, then the medium was cooled to 18°C with the addition of IPTG (final concentration 0.25 mM) for 18 h. The cells were harvested by centrifugation at 4,000 rpm for 10 min at 4°C. The pellet was resuspended in lysis buffer (100 mM Tris, pH 8.0, containing 300 mM NaCl, 15 mM imidazole, and 10% glycerol), lysed by sonication and centrifuged at 10,000 rpm for 30 min at 4°C. The supernatant containing BoADS was filtered by 0.45 µm membrane and loaded onto a 5 mL HiStrap HP column (Cytiva) with the ÄKTA FPLC system (Cytiva), then BoADS protein was eluted with buffer A (50 mM Tris, 100 mM imidazole, and 300 mM NaCl, pH 8.0), and the collected BoADS was incubated with thrombin protease in 20 mM Tris, 100 mM NaCl, pH 8.0, at 16 °C for 10 h to remove the His₆ tag. After the digestion by thrombin protease, the BoADS was purified with the 5 mL HiStrap HP column again (protein in the flow-through part). The fractions containing BoADS protein were exchanged to buffer B (20 mM Tris and 10 mM NaCl, pH 8.0) using PD10 column (Cytiva) and loaded onto an anion-exchange Resource Q column (Cytiva), which was eluted with an increased gradient of buffer B in buffer C (20 mM Tris and 1 M NaCl, pH 8.0). Purified proteins were analyzed by SDS-PAGE and the concentration was determined at 280 nm with a NanoDrop 2000C spectrophotometer (Thermo Scientific, Waltham, MA, USA). The homogeneity of BoADS was analyzed by SDS-PAGE.

Activity assay, kinetics analyses and substrate scope of BoADS

The BoADS enzymatic activity was tested in a 50 µL mixture containing 50 mM PB buffer (pH 7), 100 µM 3-oxoDCA, 100 µM pyruvate, 10 µM FAD, 10 µM ThDP, 2 mM MgCl₂ and 10 µg BoADS at 37°C for 2 h. Then, the reaction was quenched by addition of an equal volume of methanol. The samples were vortexed for 10 min and were incubated at -20°C for 1 h. All samples were subsequently centrifuged at 14,000 rpm for 20 min at 4°C. The supernatant was filtered through a membrane filter (pore size, 0.22 µm) for LC-MS/MS analysis. Besides, corresponding negative controls were undergone in the same conditions except for boiled protein.

Reaction condition optimization

the pH value optimization of BoADS was carried out at 37°C in buffers of pH ranging from 3 to 8 (50 mM phosphate buffer for pH 6-8, and 50 mM citric acid-sodium citrate buffer for pH 3-6) for 60 min. Temperature optimization was performed in a 50 µL reaction mixture in 50 mM citric acid-sodium citrate buffer (pH 6.0) for 60 min from 20°C to 50°C.

Kinetic analysis

The steady-state kinetic analysis of BoADS was carried out in 50 mM citric acid-sodium citrate buffer (pH 6.0) at 30°C. To investigate the kinetic values of 3-oxoDCA, various concentrations of 3-oxoDCA (from 0.05 to 4 mM) was incubated with the saturating pyruvate (5 mM) and 10 µg purified enzyme in the reaction buffer in a total volume of 50 µL for 60 min at 30°C. The catalytic reactions were stopped by addition of 50 µL methanol and prepared for LC-MS/MS.

Crystallization, data collection and structural elucidation of BoADS-ThDP

The purified BoADS and ThDP complex (incubated with 5 mM ThDP/MgCl₂ for 3 h on ice) were concentrated to 20 mg/mL for crystallization. With hanging-drop vapor-diffusion method at 16 °C, complex crystals of BoADS-ThDP were observed after 3 weeks in 0.2 M Sodium malonate pH 6.0, 20% w/v Polyethylene glycol 3,350.

The X-ray diffraction data of BoADS-ThDP complex was collected at beamline BL19U1 at the Shanghai Synchrotron Radiation Facility (SSRF) with a wavelength of 0.97954 Å. The diffraction data were processed and scaled with XDS. The crystal structure of the complex was solved by molecular replacement using the predicted structure from AlphaFold2 as the searching model. Molecular replacement was performed with Phaser in PHENIX, and the structure was further optimized manually with Coot and PHENIX. Refine.

Site-directed mutagenesis of BoADS

The site-directed mutagenesis experiments of Boads were carried out by PCR using two overlapping primers containing the mutated sequences with pET28a-Boads as the template, to generate a series of mutated plasmids. Subsequently, the parental vector is digested by *Dpn*I and the mixture is transformed into *E. coli* DH5α. The resulting plasmids were confirmed by sequencing.

Substrate modeling of BoADS

Substrate modeling was performed using AutoDock Vina 1.1.2.9. Before the docking, water molecules were removed from complex structure BoADS-ThDP and manually remove the ThDP. AutoDockTools 1.5.6 was used to prepare the Breslow intermediate,

3-oxoDCA and grid map files. The default parameters were used to set the torsion constraints for the Breslow intermediate and 3-oxoDCA. The charges and hydrogen atoms were added to BoADS.

For the docking parameters of BoADS with the Breslow intermediate, a grid box of $13.5 \times 6.75 \times 14.25 \text{ nm}^3$ ($x \times y \times z$) with the search space center of $x = 22.987$, $y = -12.091$, and $z = 15.001$ was created. For the parameters of BoADS with 3-oxoDCA, a grid box of $17.25 \times 14.25 \times 14.25 \text{ nm}^3$ ($x \times y \times z$) with the search space center of $x = 32.802$, $y = -11.141$, and $z = 12.044$ was created. The search parameters were used with default values.

Growth curves of tested strains

The recovered strains (2×10^7 CFU/mL) were inoculated anaerobically into freshly prepared 96-well plates containing $100 \mu\text{L}$ medium at the ratio of 5%. The medium for bacteria growth was listed in [Table S3](#). The bacterial growth kinetics were analyzed by measuring the optical density (OD_{600}).

16S rRNA gene amplicon sequencing and analyses

16S rRNA gene amplicon sequencing and analyses were performed as previously described.¹⁵ DNA samples were extracted from *ex vivo* human fecal culture or mice feces and applied to 16S rRNA gene high-throughput sequencing using the Illumina NovaSeq PE250 platform. The products of amplification of 16S rRNA gene was obtained by using the primers for V3-V4 region (F341: CCTACGGGRSGCAGCAG and R806: GGACTACVVGGGTATCTAATC). The quality of the sequencing data was evaluated by FastQC and analyzed using the DADA2 software package. The DADA2 output sequence table was converted into biom format using biomformat software and this data used to assess sequence variant abundances, producing counts for each sample. Full analysis workflows were performed by microbiomeanalyst (<https://www.microbiomeanalyst.ca/>). The α -diversity was estimated based on the profile of each sample according to the ACE, Chao1 and Shannon indices. The LEfSe analysis was performed with Galaxy.⁷⁰

Metagenomic sequencing

Metagenomic sequencing was performed according to the methods as previously described.⁷¹ The sequencing libraries were established from fecal samples of the healthy volunteers by the Ultra DNA Library Prep Kit for Illumina; the manufacturer's recommendations and index codes were adopted to attribute sequences to each sample. The fragmented DNA ends were repaired, polyA-tailed, and ligated with a sequencing adaptor for Illumina sequencing. PCR amplification and purification (AMPure XP system) were performed. DNA concentrations were measured and diluted to $2 \text{ ng}/\mu\text{L}$. The insert size of the library was assessed using an Agilent Bioanalyzer 2100 system. qPCR was performed to ensure an accurate concentration ($> 3 \text{ nM}$) of the library. The clustering of the index-coded samples was performed on a cBot Cluster Generation System using a HiSeq 4000 PE Cluster Kit according to the manufacturer's instructions. After cluster generation, the library preparations were sequenced on an Illumina HiSeq 4000 platform, and 150-bp paired-end reads were generated.

bioBakery⁷² tools were used to process metagenomic reads (in fastq format) by trimming the reads to PHRED quality 30 and removing Illumina adapters. Following trimming, the KneadData integrated Bowtie2 tool⁷³ was used to remove human-associated reads. Filtered clean data were used for further analysis.

Luciferase reporter gene assays

The luciferase reporter gene assays of FXR, PXR and VDR were carried out according to the protocol described previously.⁷⁴ HEK293T cells were co-transfected with human expression vectors of nuclear receptors, human apical sodium-dependent bile acid transporter (ASBT) expression vector, human retinoid X receptor (RXR) expression vector, an individual firefly luciferase reporter vector and a Renilla luciferase control vector. After 24 h post-transfection, the cells were exposed to different concentrations of various bile acids for 16 hours. Luciferase assays were performed using the dual-luciferase assay system. For TGR5 luciferase assay, HEK293T cells were co-transfected with pCMVSPORT6/hTGR5 and cAMP response element-driven luciferase reporter plasmids before lysed for detection using assay buffer with coelenterazine.

Stool-derived *ex vivo* communities

For stool-derived *ex vivo* communities, 10 mg human fecal pellet was suspended in 1 mL anaerobic sterilized PBS, which was then homogenized by vortexed for 1 min. The fecal suspension was obtained by centrifugation at $500 \times g$ for 5 min to remove large particles, which was ten-fold diluted for incubation in appropriate medium with the inoculum size at 1% and cultured anaerobically at 37°C .

Collection of public datasets and analyses of metagenomic sequencing data

The analysis related to *Boads* used following public datasets: HMP,⁷⁵ NCBI BioProject PRJEB2054 and NCBI BioProject PRJEB6337. For each cohort, we selected samples obtained from healthy individuals.

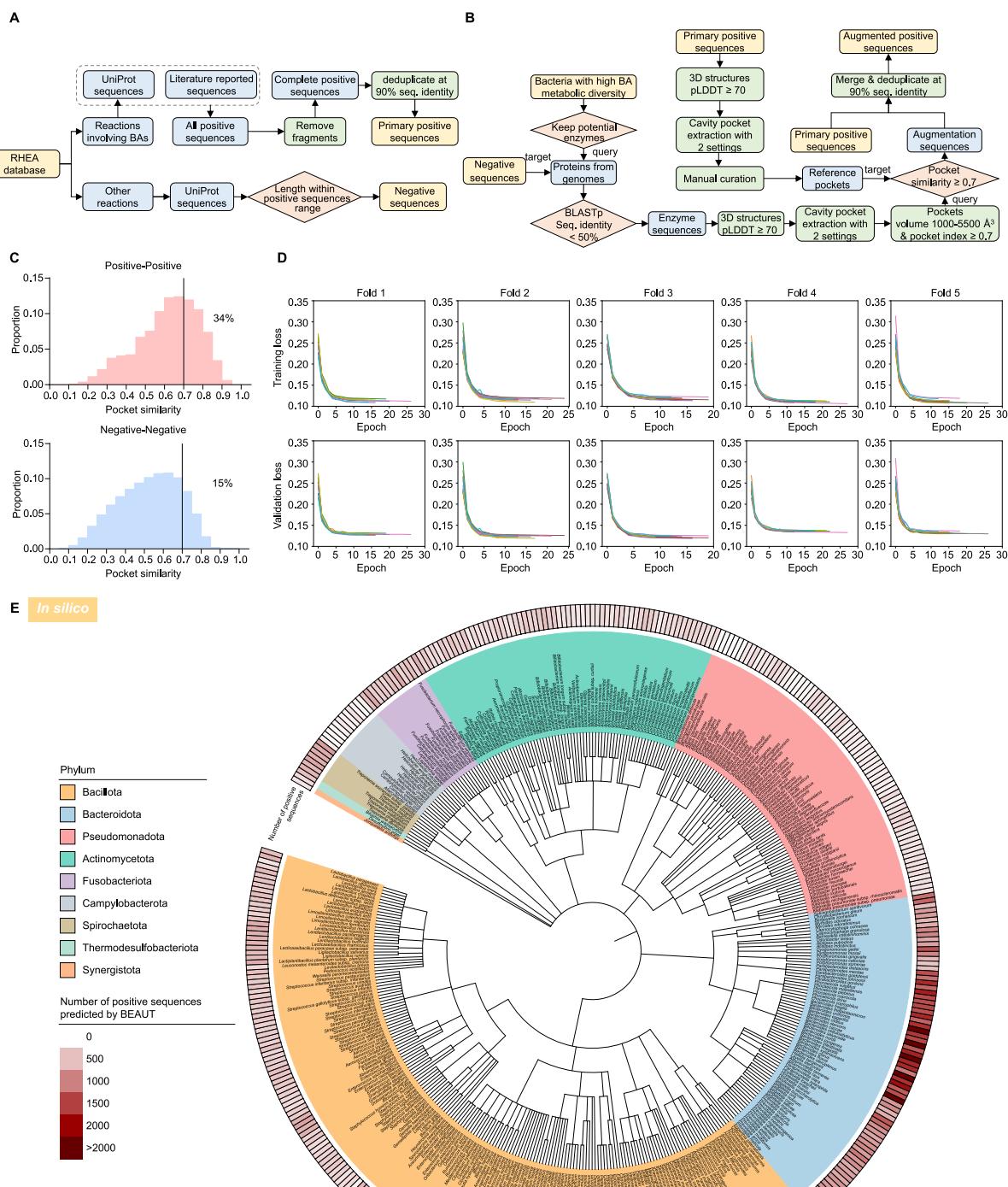
All the raw whole-metagenome shotgun sequencing datasets were downloaded from the NCBI Sequence Read Archive (SRA, <https://www.ncbi.nlm.nih.gov/sra>) and NIH Human Microbiome Project (<http://www.hmpdacc.org>). Metadata of datasets were obtained from original articles and sample information on websites.

If necessary, an aforementioned pipeline for metagenomic sequencing data was processed to conduct quality control and eliminate human contaminant. The clean sequences were mapped into target gene using Bowtie2 (-end-to-end, -very-sensitive). Taxonomic composition of metagenomes was profiled by Kraken2⁷⁶ using Kraken2 standard reference genomes database. All the parameters of Kraken2 utilized default settings.

QUANTIFICATION AND STATISTICAL ANALYSIS

GraphPad Prism version 9.0 and SPSS version 27.0 were used for statistical analysis. The experimental data were shown as the mean \pm SEMs. The sample size was estimated according to previous experience, sample availability and previously reported studies. No data were excluded from the data analysis. The normal distribution of the data was determined by the Shapiro-Wilk normality test. For statistical comparisons, Student's *t*-test (normally distributed data) or Mann-Whitney *U* test (nonnormally distributed data) was applied to variables between two groups. One-way ANOVA with Tukey's (the same standard deviation) or with Dunnett's T3 test (different standard deviation) analysis was used to compare normally distributed variables, while nonnormally distributed data were compared by the or the Kruskal-Wallis test for variables among multiple groups. $P < 0.05$ was considered significant.

Supplemental figures

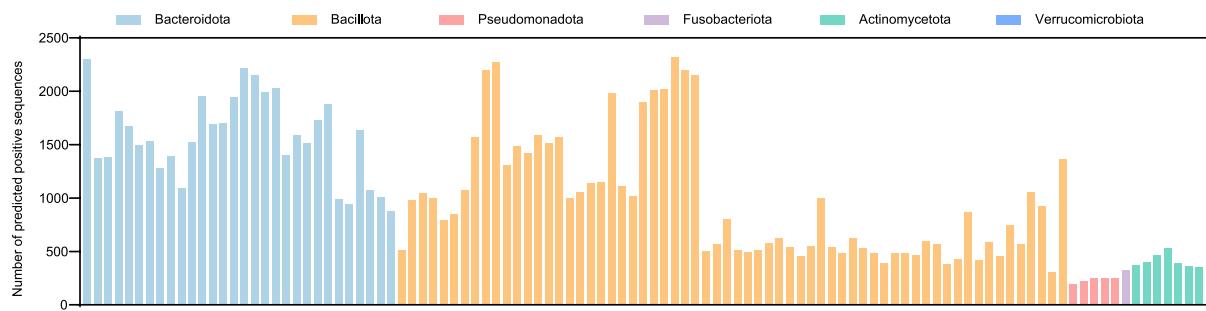


(legend on next page)

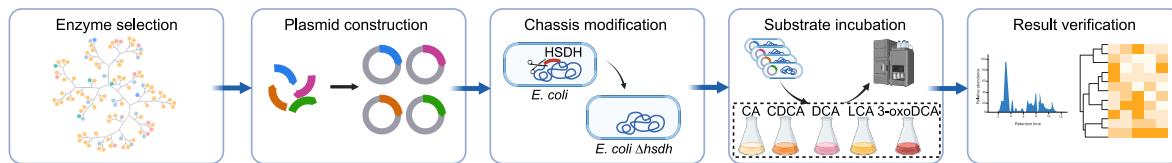
Figure S1. BEAUT development and query results of gut microbiota, related to Figure 1

- (A) The data collection flow of primary positive sequences and negative sequences for BEAUT training.
- (B) The data collection flow of augmented positive sequences for BEAUT training. pLDDT, predicted local distance difference test; BA, bile acid.
- (C) Pockets similarity among different enzyme groups. Up panel shows the distribution of primary positive sequence pocket pairs similarities. Down panel shows the distribution of negative sequence pocket pairs similarities. The number on the upper-right corner shows the percentage of enzyme pairs with pocket similarity ≥ 0.7 .
- (D) Analysis of the training loss and validation loss of the BEAUT model.
- (E) Phylogenetic tree of the gut microbe species from the HGBME dataset and the median number of positive sequences predicted by the BEAUT model (*in silico*). The tree was colored in terms of the phylum of each node, and the heatmap displays the number of positive sequences predicted by BEAUT.

A *In silico*



B



C

ID#	Accession number	Organism	EC number annotated by CLEAN	Metabolic capacity > 0.1?	ID#	Accession number	Organism	EC number annotated by CLEAN	Metabolic capacity > 0.1?
BE1	EIY84613.1	<i>Bacteroides xylosovens</i>	2.2.1.13	Yes	BE57	KMV82631.1	<i>Klebsiella oxytoca</i>	3.2.1.86	No
BE2	EEP45224.1	<i>Collinsella intestinalis</i>	2.4.1.212	No	BE58	KXA50331.1	<i>Streptococcus agalactiae</i>	1.1.1.14 / 3.6.4.12	No
BE3	EHL7206.1	<i>Bacillus sp. 7_6_55CFAA_CT2</i>	2.7.1.190	Yes	BE59	ENY91287.1	<i>Enterocloster clostridiiformis</i>	2.3.1.48	No
BE4	EFU77259.1	<i>Lachnoanaerobaculum saburreum</i>	1.1.1.77	Yes	BE60	ENZ20387.1	<i>Enterocloster clostridiiformis</i>	3.1.1.5	No
BE5	EFV76545.1	<i>Bacillus sp. 2_A_57_CT2</i>	2.7.1.95	No	BE61	KMW33529.1	<i>Parabacteroides sp. D25</i>	1.1.1.176	Yes
BE6	KXO15703.1	<i>Prevotella bivia</i>	3.1.1.72	Yes	BE62	ADK95876.1	<i>Prevotella melanogenica</i>	2.6.1.42	Yes
BE7	EHM39442.1	<i>Anaeroglobus geminatus</i>	4.4.1.13	No	BE63	KXT53104.1	<i>Akkermansia sp. KLE1605</i>	2.3.1.183	Yes
BE8	EFC99331.1	<i>Hungatella hathewayi</i>	2.3.1.53 / 2.3.1.57	Yes	BE64	EHB93209.1	<i>Alistipes indistinctus</i>	4.1.3.38	No
BE9	ENY95387.1	<i>Hungatella hathewayi</i>	2.7.1.33	Yes	BE65	ERI90435.1	<i>Blautia sp. KLE 1732</i>	4.1.1.81	Yes
BE10	KG172363.1	<i>Eggerthella lenta</i>	2.3.1.85	No	BE66	KMW19184.1	<i>Enterocloster citroniae</i>	2.1.1.189	No
BE11	ENY92028.1	<i>Hungatella hathewayi</i>	2.3.1.85	No	BE67	EEU99191.1	<i>Roseburia intestinalis</i>	2.1.1.190	No
BE12	EHR35258.1	<i>Dolosigranulum pigrum</i>	1.1.1.301	Yes	BE68	ERK43584.1	<i>Eubacterium ramulus</i>	1.8.5.5 / 1.14.19.51	No
BE13	OH074764.1	<i>Bacillus sp. HMSC036E02</i>	3.1.1.32	Yes	BE69	KGF53784.1	<i>Flavonifractor plautii</i>	2.3.1.48	Yes
BE14	EGN42322.1	<i>Lachnospiraceae bacterium</i>	2.5.1.58	Yes	BE70	EHE98027.1	<i>Enterocloster citroniae</i>	6.3.5.11	No
BE15	EFF67181.1	<i>Selenomonas noxidea</i>	6.2.1.14	Yes	BE71	EEG54884.1	<i>Enterocloster asparagiformis</i>	6.3.5.11	Yes
BE16	EFU75693.1	<i>Lachnoanaerobaculum saburreum</i>	2.7.1.130	No	BE72	EDY33327.1	<i>Ruminococcus lactaris</i>	6.3.5.11	Yes
BE17	EDO63686.1	<i>Clostridium leptum</i>	2.4.1.321 / 3.2.1.37	No	BE73	ENY93648.1	<i>Hungatella hathewayi</i>	4.1.1.11	Yes
BE18	EDO62434.1	<i>Clostridium leptum</i>	2.4.1.321	No	BE74	ERI94096.1	<i>Blautia sp. KLE 1732</i>	6.3.5.11	No
BE19	EDO62660.1	<i>Clostridium leptum</i>	2.4.1.321	Yes	BE75	EDS04988.1	<i>Clostridium scindens</i>	6.3.5.11	No
BE20	KKK66269.1	<i>Christensenellina minuta</i>	1.1.1.391	Yes	BE76	EGW50188.1	<i>Desulfovibrio sp. 6_1_46AFAA</i>	2.3.1.18 / 2.3.1.79	No
BE21	EEG56119.1	<i>Enterocloster asparagiformis</i>	1.1.1.127	Yes	BE77	EGG83675.1	<i>Lachnospiraceae bacterium</i>	5.3.1.8	No
BE22	EFK13834.1	<i>Escherichia coli</i>	1.1.1.159	Yes	BE78	OFO37684.1	<i>Neisseria sp. HMSC075C10</i>	2.4.1.308 / 2.5.1.98 /	No
BE23	EEG51213.1	<i>Enterocloster asparagiformis</i>	1.1.1.391	Yes	BE79	EMD16646.1	<i>Eggerthia catenaformis</i>	3.1.1.72	No
BE24	KG172449.1	<i>Eggerthella lenta</i>	1.1.1.391	Yes	BE80	KXT35628.1	<i>Sutterella sp. KLE1602</i>	1.97.1.1	No
BE25	KXK66490.1	<i>Christensenellina minuta</i>	1.1.1.393	Yes	BE81	EFC97053.1	<i>Hungatella hathewayi</i>	2.3.1.202	Yes
BE26	ENY99270.1	<i>Hungatella hathewayi</i>	1.1.1.10	No	BE82	OHR46121.1	<i>Neisseria sp. HMSC071C03</i>	3.4.16.4	No
BE27	EEG53331.1	<i>Enterocloster asparagiformis</i>	1.3.1.25	No	BE83	EFK35391.1	<i>Chryseobacterium gleum</i>	1.14.16.5	No
BE28	KXA41480.1	<i>Prevotella corporis</i>	3.2.2.19	No	BE84	ERK52611.1	<i>Leptotrichia wadei</i>	1.1.1.195 / 2.4.1.111 / 2.4.1.115 / 2.4.1.126	No
BE29	ENZ04075.1	<i>Enterocloster clostridiiformis</i>	2.3.1.48	No	BE85	EKX97580.1	<i>Selenomonas sp. oral taxon 138</i>	2.7.8.42	Yes
BE30	EHE95701.1	<i>Enterocloster citroniae</i>	3.1.6.1	No	BE86	EEG75578.1	<i>Clostridium hylemonae</i>	3.2.1.78	No
BE31	EEG74473.1	<i>Clostridium hylemonae</i>	4.1.1.81	Yes	BE87	ENY92941.1	<i>Hungatella hathewayi</i>	2.7.13.3	Yes
BE32	EEC54208.1	<i>Bacteroides eggerthii</i>	6.1.1.9	No	BE88	EFK36942.1	<i>Chryseobacterium gleum</i>	1.1.3.6 / 5.3.3.1	No
BE33	OF545118.1	<i>Staphylococcus sp. HMSC07A08</i>	6.1.1.5	No	BE89	EFQ44749.1	<i>Lactobacillus crispatus</i>	1.1.1.146	Yes
BE34	ENY97576.1	<i>Hungatella hathewayi</i>	4.4.1.13	Yes	BE90	EEF87129.1	<i>Bacteroides fragilis</i>	1.1.1.313 / 1.1.1.381	No
BE35	ERI72451.1	<i>Clostridium sp. KLE 1755</i>	4.4.1.13	No	BE91	ERI6318.1	<i>Blautia sp. KLE 1732</i>	5.4.2.7	Yes
BE36	EIY39583.1	<i>Bacteroides celulosilyticus</i>	4.4.1.13	No	BE92	ERI78448.1	<i>Clostridium symbiosum</i>	5.4.2.7	No
BE37	KXT48730.1	<i>Akkermansia sp. KLE1605</i>	4.4.1.13	No	BE93	OFK24536.1	<i>Staphylococcus sp. HMSC068H12</i>	2.1.1.103 / 2.4.1.311 / 2.4.1.384	Yes
BE38	ERI97959.1	<i>Blautia sp. KLE 1732</i>	3.5.1.24	No	BE94	EFX90618.1	<i>Actinobacillus ureae</i>	3.2.2.22	No
BE39	ERJ1531.1	<i>Ruminococcus callidus</i>	3.5.1.24	Yes	BE95	OH076275.1	<i>Bacillus sp. HMSC036E02</i>	2.4.1.384	No
BE40	EEC58517.1	<i>Bacteroides pectinophilus</i>	3.1.1.5 / 3.1.3.12	Yes	BE96	EKN28021.1	<i>Parabacteroides distasonis</i>	2.7.13.3	Yes
BE41	OAD88247.1	<i>Clostridiales bacterium KLE1615</i>	3.1.1.5	Yes	BE97	EFC97018.1	<i>Hungatella hathewayi</i>	3.1.1.32	No
BE42	ETA84573.1	<i>Eikenella corrodens</i>	3.1.1.1	Yes	BE98	EEG74302.1	<i>Clostridium hylemonae</i>	2.4.1.198 / 4.2.3.25 / 4.2.3.71 / 4.2.3.104	No
BE43	EEG25264.1	<i>Eikenella corrodens</i>	3.1.1.1	Yes	BE99	EEG54766.1	<i>Enterocloster asparagiformis</i>	3.2.2.22	Yes
BE44	EHE95445.1	<i>Enterocloster citroniae</i>	1.16.1.7	Yes	BE100	ENY96887.1	<i>Hungatella hathewayi</i>	3.4.22.71	Yes
BE45	OHR83546.1	<i>Staphylococcus sp. HMSC34C02</i>	3.5.1.104	Yes	BE101	EHF01061.1	<i>Enterocloster citroniae</i>	3.4.22.71	Yes
BE46	KXB87300.1	<i>Veillonella sp. DNFO0869</i>	3.5.1.104	No	BE102	EEC58688.1	<i>Bacteroides pectinophilus</i>	3.4.22.70	Yes
BE47	OFO95274.1	<i>Pseudomonas sp. HMSC067G02</i>	6.1.1.4	Yes	BE103	EHG32038.1	<i>Enterocloster clostridiiformis</i>	3.4.22.71	Yes
BE48	EEC52363.1	<i>Bacteroides eggerthii</i>	6.1.1.4	No	BE104	ENY91877.1	<i>Hungatella hathewayi</i>	3.4.22.71	Yes
BE49	EHE95526.1	<i>Enterocloster citroniae</i>	2.3.1.48	Yes					
BE50	EFR55832.1	<i>Bacteroides fragilis</i>	2.7.1.33	No					
BE51	EMD15965.1	<i>Eggerthella catenaformis</i>	3.1.1.32	No					
BE52	KG17364.1	<i>Eggerthella lenta</i>	1.3.1.116	No					
BE53	EGX71336.1	<i>Erysipelotrichaceae bacterium</i>	2.3.1.48 / 2.4.1.198	No					
BE54	EF0D1564.1	<i>Hungatella hathewayi</i>	2.1.1.151	No					
BE55	OFR43209.1	<i>Streptococcus sp. HMSC062H02</i>	2.5.1.16 / 4.1.3.40 / 6.2.1.22	No					
BE56	EES69732.1	<i>Bacteroides thetaiotaomicron</i>	6.2.1.14	No					

(legend on next page)

Figure S2. BEAUT model validation at the strain and enzyme levels, related to Figure 2

- (A) Number of positive sequences predicted by BEAUT of 108 gut microbial strains (*in silico*). The histograms were colored in terms of the phylum of each strain. The genomes of the strains were searched through the NCBI Genome database, and the genomes with the highest assembly quality were selected for BEAUT prediction.
- (B) Schematic representation of potential BA metabolic enzyme validation chosen from BEAUT.
- (C) The list of 104 chosen potential BA metabolic enzymes. Enzymes BE22 and BE24 are positive controls of this experiment.

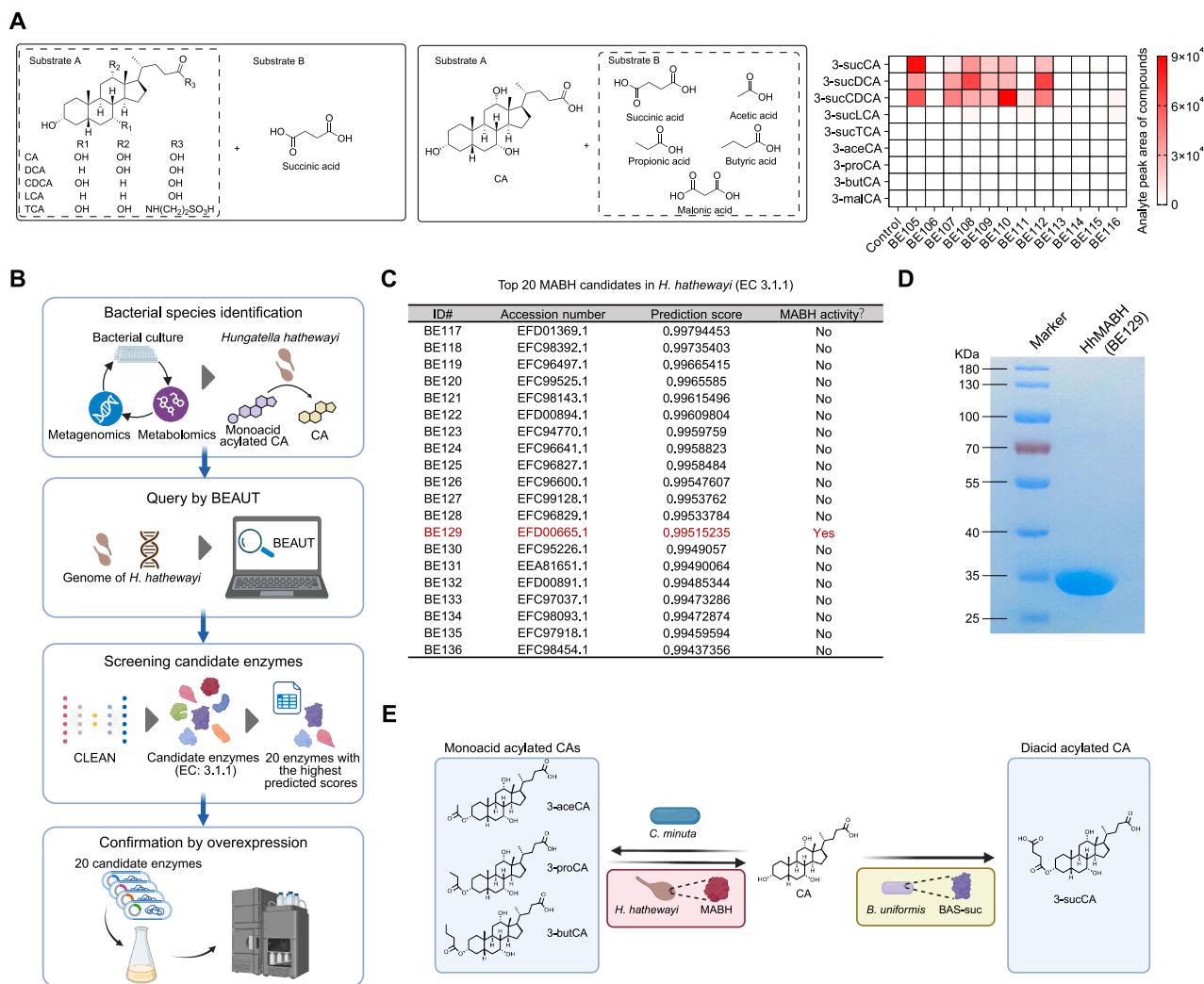


Figure S3. MABH is the 3-aceCA hydrolase, related to Figure 3

- (A) Schematic diagram of BAS-suc reaction with various substrates. Heatmap shows different BA substrates and short-chain carboxylic acid substrates catalyzed by potential BAS-suc enzymes.
- (B) Schematic diagram shows identification flow of MABH of *H. hathewayi*.
- (C) The top 20 candidate MABH list predicted by BEAUT of *H. hathewayi* DSM 13479.
- (D) Coomassie-stained SDS-PAGE of purified HhMABH.
- (E) Schematic diagram shows MABH as a regulator in 3-acylCAs.
- All data are presented as the means \pm SEMs.

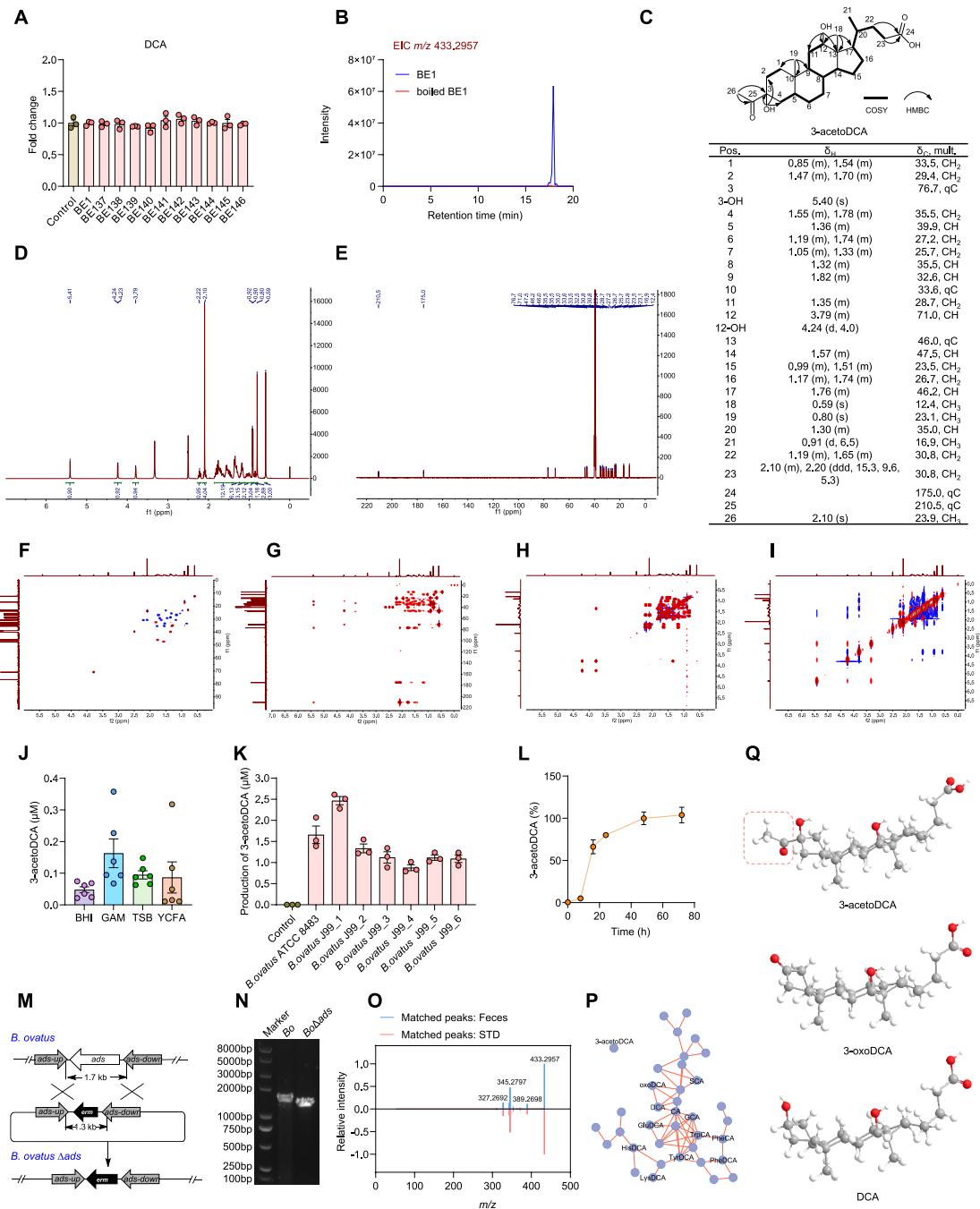


Figure S4. Production and authentication of 3-acetoDCA, related to Figure 4

(A) The yields of possible product DCA generated from predicted 3-oxoDCA metabolic enzyme reactions.

(B) Representative extracted ion chromatogram shows a peak at m/z 433.2957 in the biotransformation with EIY84613.1 (BE1).

(3) Representative extracted ion chromatogram shows a peak at m/z 182.183.355.7 in the biotransformation with L-715, 18°C, 5 d (DE).

(D), ^1H - ^1H correlated spectroscopy (COSY) and heteronuclear multiple bond correlation (HMBC) of isolated 3-acetoDCA in DMSO-d₆.

(J) 3-acetoDCA production in the SECs anaerobically cultured

(b) 3-acetylCoA production in the CEC5 anaerobically cultured in casein hydrolysate-fatty acids broth (YCFA) medium. $n = 6$ individuals.

(K) The production of 3-acetoDCA of different *B. ovatus* isolates.

(L) Production ratio of 3-acetoDCA by *B. ovatus* J99 1 at

(legend continued on next page)

-
- (O) Comparison of liquid chromatography-tandem mass spectrometry (LC-MS/MS) data of 3-acetoDCA from standards (STD) and fecal samples.
- (P) Molecular networks of CA (right) and 3-acetoDCA (left). Each node represents a clustered tandem mass spectrum, and connections between the nodes indicate relationships through the cosine score (cutoff minimum of 0.7).
- (Q) The 3D structure of 3-acetoDCA (up), 3-oxoDCA (middle), and DCA (down) presented by the ball-and-stick model. The gray ball represents C atom, the white ball represents H atom, the red ball represents O atom, and red dashed box is an additional ethyl ketone side chain of 3-acetoDCA.
- All data are presented as the means \pm SEMs.

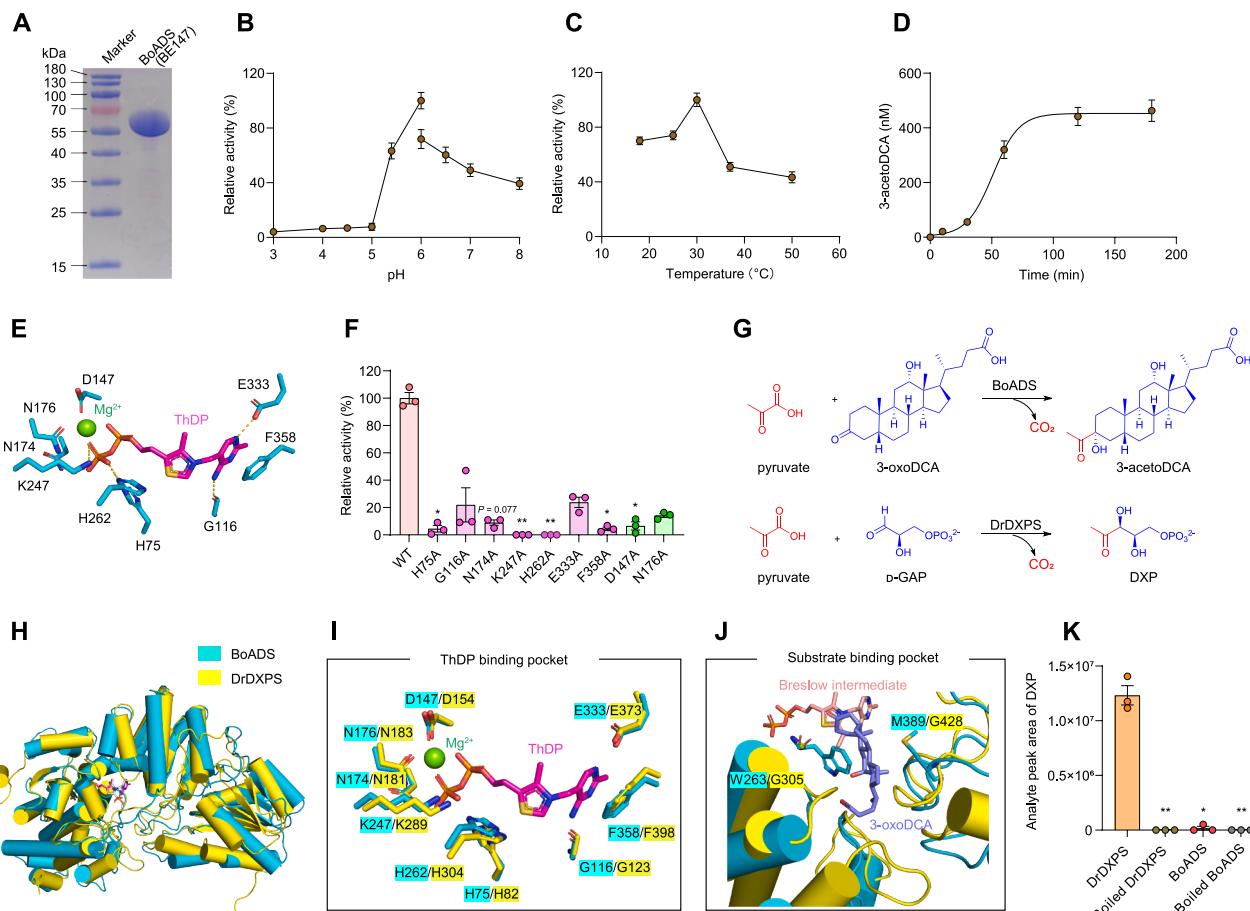


Figure S5. The catalytic reaction condition optimization of BoADS and its comparison with DrDXPS, related to Figure 5

- (A) Coomassie-stained SDS-PAGE of purified BoADS.
- (B) The optimization of BoADS relative activity using different pH values. 50 mM citric acid-sodium citrate buffer for pH 3–6 and 50 mM phosphate buffer for pH 6–8.
- (C) The optimization of BoADS relative activity under different temperatures.
- (D) Time-dependence of 3-acetoDCA production catalyzed by BoADS.
- (E) The ThDP-binding site of BoADS-ThDP complex structure. ThDP was shown as magenta sticks, and Mg²⁺ ion was shown as a green sphere.
- (F) The relative catalytic activities of BoADS and its variants are related to the ThDP/Mg²⁺-binding sites. $n = 3$.
- (G) The catalytic reactions of BoADS and DrDXPS (1-deoxy-D-xylulose 5-phosphate synthase from *Deinococcus radiodurans*). d-GAP, D-glyceraldehyde 3-phosphate; DXP, 1-deoxy-D-xylulose 5-phosphate.
- (H) The superimposition of BoADS (cyan) with DrDXPS (yellow).
- (I) Structural comparison of the ThDP-binding pockets in BoADS (residues in cyan) and DrDXPS (residues in yellow). The ThDP molecules in BoADS and DrDXPS were shown as magenta sticks.
- (J) Structural comparison of the substrate-binding pockets in modeled complex structure BoADS (cyan) with Breslow intermediate and 3-oxoDCA and DrDXPS (yellow).
- (K) The quantified production of DXP catalyzed by DrDXPS and BoADS. $n = 3$.

All data are presented as the means \pm SEMs. In (F) and (K), the p values were determined by Kruskal-Wallis test followed by Dunn's post hoc test. In (F), * $p < 0.05$ and ** $p < 0.01$ versus the WT group. In (K), * $p < 0.05$ and ** $p < 0.01$ versus DrDXPS group.

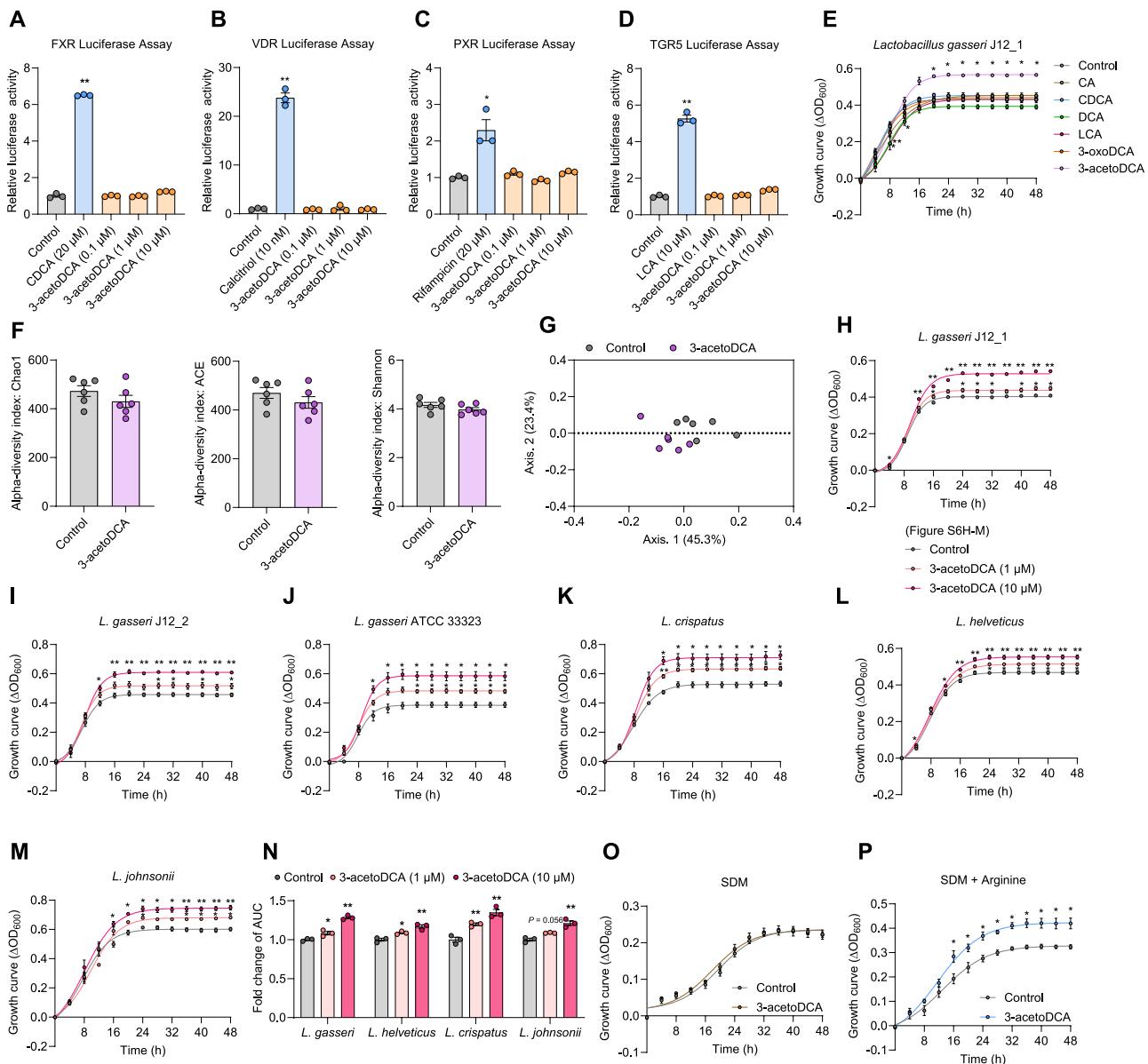


Figure S6. Prevalence of the ADS gene *ads* and the function of 3-acetoDCA, related to Figure 6

- (A) Luciferase activity was detected in HEK293T cells, which were treated with control, farnesoid X receptor (FXR) agonist CDCA, and 3-acetoDCA at the indicated concentrations. $n = 3$.
- (B) Luciferase activity was detected in HEK293T cells, which were treated with control, vitamin D receptor (VDR) agonist calcitriol, and 3-acetoDCA at the indicated concentrations. $n = 3$.
- (C) Luciferase activity was detected in HEK293T cells, which were treated with control, pregnane X receptor (PXR) agonist rifampicin, and 3-acetoDCA at the indicated concentrations. $n = 3$.
- (D) Luciferase activity was detected in HEK293T cells, which were treated with control, Takeda G-protein-coupled receptor 5 (TGR5) agonist LCA, and 3-acetoDCA at the indicated concentrations. $n = 3$.
- (E) The effect of different BA incubations on the growth of *L. gasseri*. $n = 3$.
- (F) Alpha-diversity indexes of the gut microbiota between the control and 3-acetoDCA groups as indicated by the Chao1, ACE, and Shannon indexes. $n = 6$.
- (G) Principal coordinates analysis (PCoA) of the gut microbiota between the control and 3-acetoDCA groups using Jensen-Shannon divergence distance. $n = 6$.
- (H–J) The effect of 3-acetoDCA incubation on the growth of different *L. gasseri* strains. $n = 3$.
- (K–N) The effect of 3-acetoDCA incubation on the growth of different *Lactobacillus* species (K–M) and the area under curve (AUC) of the growth curve for 48 h (N). $n = 3$.
- (O) The growth of *L. gasseri* strains cultured in SDM with 3-acetoDCA or without 3-acetoDCA. $n = 3$.

(legend continued on next page)

(P) The effect of 3-acetoDCA incubation on the growth of *L. gasseri* strains cultured in SDM supplemented with arginine. $n = 3$. All data are presented as the means \pm SEMs. In (A), (B), (D), (J), (K), and (M), the p values were determined by one-way ANOVA with Dunnett's T3 test. In (C) and (E), the p values were determined by Kruskal-Wallis test followed by Dunn's post hoc test. In (F), the p values were determined by a two-tailed Student's t test. In (H), (I), (L), and (N), the p values were determined by one-way ANOVA with Tukey's post hoc test. In (O) and (P), the p values were determined by Mann-Whitney U test.
* $p < 0.05$ and ** $p < 0.01$ versus the control group.