Machine Learning-assisted Identification of Bioindicators Predicts Medium-chain Carboxylate Production Performance of an Anaerobic Mixed Culture

Bin Liu
Helmholtz-Zentrum fur Umweltforschung UFZ Abteilung Umweltmikrobiologie
Heike Sträuber
Helmholtz-Zentrum fur Umweltforschung UFZ Abteilung Umweltmikrobiologie
Joao Saraiva
Helmholtz-Zentrum für Umweltforschung UFZ Abteilung Umweltmikrobiologie: Helmholtz-Zentrum fur Umweltforschung UFZ Abteilung Umweltmikrobiologie
Hauke Harms
Helmholtz-Zentrum für Umweltforschung UFZ Abteilung Umweltmikrobiologie: Helmholtz-Zentrum fur Umweltforschung UFZ Abteilung Umweltmikrobiologie
Sandra Godinho Silva
Universidade de Lisboa Instituto Superior Tecnico
Sabine Kleinsteuber
Helmholtz-Zentrum für Umweltforschung UFZ Abteilung Umweltmikrobiologie: Helmholtz-Zentrum fur Umweltforschung UFZ Abteilung Umweltmikrobiologie
Ulisses Rocha (ulisses.rocha@ufz.de)
Department of Environmental Microbiology, Helmholtz Centre for Environmental Research – 8 UFZ, Leipzig, Germany https://orcid.org/0000-0001-6972-6692

Research

Keywords: Predictive biology, carboxylate platform, model ecosystems, reactor microbiota, microbial chain elongation

Posted Date: September 22nd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-78714/v1

Version of Record: A version of this preprint was published at International Chain Elongation Conference 2020 on November 2nd, 2020. See the published version at https://doi.org/10.18174/icec2020.18013.

Machine learning-assisted identification of bioindicators predicts medium-chain carboxylate production performance of an anaerobic mixed culture

Bin Liu[1], Heike Sträuber[1], João Saraiva[1], Hauke Harms[1], Sandra Godinho Silva[2], Sabine Kleinsteuber[1]* and Ulisses Nunes da Rocha[1]*

* Authors followed by an asterisk contributed equally to this work

[1]Department of Environmental Microbiology, Helmholtz Centre for Environmental Research – UFZ, Leipzig, Germany

[2]Institute for Bioengineering and Biosciences, Department of Bioengineering, Instituto Superior Técnico Universidade de Lisboa, Lisbon, Portugal

Corresponding Authors:

sabine.kleinsteuber@ufz.de / ulisses.rocha@ufz.de (ordered alphabetically according to last name).

Authors email addresses:

Bin Liu: liu.bin@ufz.de; Heike Sträuber: heike.straeuber@ufz.de; João Saraiva: joao.saraiva@ufz.de; Hauke Harms: hauke.harms@ufz.de; Sandra Godinho Silva: sandra.silva@ufz.de; Sabine Kleinsteuber: sabine.kleinsteuber@ufz.de; Ulisses Nunes da Rocha: ulisses.rocha@ufz.de

Abstract

Background: The ability to quantitatively predict ecophysiological functions of microbial communities provides an important step to engineer microbiota for desired functions related to specific biochemical conversions. Here, we present the quantitative prediction of medium-chain carboxylate production in two continuous anaerobic bioreactors from 16S rRNA gene dynamics in enrichment cultures.

Results: By progressively shortening the hydraulic retention time from 8 days to 2 days with different temporal schemes in both bioreactors operated for 211 days, we achieved higher productivities and yields of the target products *n*-caproate and *n*-caprylate. The datasets generated from each bioreactor were applied independently for training and testing in machine learning. A predictive model was generated by employing the random forest algorithm using 16S rRNA amplicon sequencing data. More than 90% accuracy in the prediction of *n*-caproate and *n*-caprylate productivities was achieved. Four inferred bioindicators belonging to the genera *Olsenella*, *Lactobacillus*, *Syntrophococcus* and *Clostridium* IV suggest their relevance to the higher carboxylate productivity at shorter hydraulic retention time. The recovery of metagenome-assembled genomes of these bioindicators confirmed their genetic potential to perform key steps of medium-chain carboxylate production.

Conclusions: Shortening the hydraulic retention time of the continuous bioreactor systems allows to shape the communities with desired chain elongation functions. Using machine-learning, we demonstrated that 16S rRNA amplicon sequencing data can be used to predict bioreactor process performance quantitatively and accurately. Characterising and harnessing bioindicators holds promise to manage reactor microbiota towards selection of the target processes. Our mathematical framework is transferrable to other ecosystem processes and

47 microbial systems where community dynamics is linked to key functions. The general
48 methodology can be adapted to data types of other functional categories such as genes,
49 transcripts, proteins or metabolites.
50
51 **Keywords:** Predictive biology, carboxylate platform, model ecosystems, reactor microbiota,
52 microbial chain elongation
53
54 **Background**
55
56 Microbes form complex communities that play essential roles in ecosystem functioning.
57 Identifying bioindicators derived from community analysis and using them to predict process
58 performance may delineate potential cause-effect relationships with ecosystem functioning [1,2].
59 The knowledge gained from prediction can be used to generate hypotheses on the role of key
60 species. At ecosystem level, designing effective control strategies for key species holds promise
61 to manage the community towards selection of the target processes, which is crucial for
62 microbiota-based biotechnologies [3–5].
63 Our goals were to investigate how environmental manipulations affect ecosystem
64 functioning and to predict performance metrics of the quantifiable biological processes by
65 following microbial community dynamics. Model ecosystems offer the opportunity to link
66 microbial diversity and ecosystem functioning in a quantifiable and predictable way [6–8]. Such
67 simplified ecosystems can be still complex regarding microbial interactions and involved
68 metabolic processes [6]. Here, we used anaerobic fermentation reactors as model ecosystems and
69 considered microbial chain elongation (CE) as the quantifiable model ecosystem process. CE is a

microbial process that produces medium-chain carboxylates (6 to 8 carbon atoms) through reverse β-oxidation [9]. Recently we enriched a mixed culture that produces *n*-butyrate (C4), *n*-caproate (C6) and *n*-caprylate (C8) from xylan and lactate in a daily-fed reactor system [10], to simulate the feedstock conditions of anaerobic fermentation of ensiled plant biomass [11]. For this bioprocess to be viable, it needs to include diverse functions such as xylan hydrolysis, xylose fermentation and CE with lactate as electron donor. Mixed culture fermentation is characterised by different trophic groups that may cooperate or compete with each other to metabolise complex substrates [9]. Species involved in these interactions can drive shifts in community structure and function [1]. During the long-term stable reactor operation, the community developed towards predominating C4 and biomass production at the cost of C6/C8 production [10]. We wanted to explore how process parameter changes shape the existing microbiota to optimise the process towards the target products C6 and C8. The current study was conducted on the enriched chain-elongating microbiota in two parallel bioreactors. To promote C6 and C8 production and enrich the functional groups relevant to process performance, we reduced the hydraulic retention time (HRT). HRT refers to the average time soluble compounds reside in the bioreactor. Shortening the HRT is a common operation-based strategy for increasing C6/C8 production [12–16] and a key factor influencing microbial diversity [17]. It is relevant to the microbial growth rate in reactors without biomass retention, and it affects biomass concentration and community composition [18]. Following variations in diversity induced by HRT reduction, we tested if productivity and yield of the target products (C6 and C8) could be predicted by using machine learning. To provide insight into the dynamics of community structure and function, we measured process performance and collected samples for community analysis using high-throughput sequencing of the 16S rRNA gene. Community analysis using 16S rRNA amplicon

sequencing data combined with environmental variables can reveal relationships between microbial communities and ecosystem functioning. For example, Werner et al. demonstrated strong relationships between the phylogenetic community structure, reflected by time-resolved 16S rRNA amplicon data, and the methanogenic activity in full-scale anaerobic digesters, by applying constrained ordination [19]. Predictive analytics using machine learning has shown promise in microbiota-based biotechnologies [6,20,21]. We chose the random forest algorithm because it runs efficiently and accurately on high-dimensional datasets with multi-features, and it avoids overfitting, particularly when using different training and test datasets [22]. Our random forest analysis consisted of two parts. First, we performed feature selection identifying Amplicon Sequence Variants (ASVs) that would be relevant to community dynamics caused by HRT reduction. Next, we trained the algorithm with these features (hereafter, HRT bioindicators) that later were used to predict the production of C6 and C8.

Methods

Reactor operation and process monitoring. The inoculum was initially taken from a continuous lab-scale bioreactor that produced C6 and C8 by anaerobic fermentation of lactate-rich corn silage [11]. Enrichment was performed in a reactor that was daily fed with mineral medium (pH 5.5; Additional file 1: Table S1) containing water-soluble xylan (more than 95% xylooligosaccharides, from corncob; Roth, Karlsruhe, Germany) and lactic acid (85%, FCC grade; Sigma Aldrich, St. Louis, USA) as defined carbon sources and produced C4, C6 and C8 over 150 days [10]. For the present study, two 1-L bioreactors (A and B; BIOSTAT® A plus, Sartorius AG, Göttingen, Germany) were filled up with 0.5 L of the enriched culture. Both

bioreactors were daily fed with 0.125 L medium containing 1.47 g lactic acid and 1.25 g xylan, without withdrawing effluent. After four days the contents of both bioreactors were mixed by pumping them three times from bioreactor A to B and back while keeping anoxic conditions. Eventually, they were equally distributed to both bioreactors, which is considered the starting point (day 0) of the experiment.

We employed semi-continuous stirred tank reactors for anaerobic fermentation, which were operated at 38 ± 1°C and constantly stirred at 150 rpm. The pH of the reactor broth was automatically controlled at 5.5 by addition of 1 M NaOH. For each bioreactor, the produced gas was collected in a coated aluminium foil bag that also served for compensating underpressure in the reactor system. The bag was connected after a MilliGascounter® (MGC-1; Ritter, Bochum, Germany) that measured on-line the volume of the produced gas. A gas-sample septum was placed in the gas pipe of each bioreactor.

In the beginning, both bioreactors were operated as replicates with an equal HRT of 8 days. For daily feeding, 1.47 g lactic acid and 1.25 g xylan were supplied in mineral medium. After 51 days, we gradually decreased the HRT of bioreactor A from 8 days to 6 days, and further to 4 days and 2 days as shown in Additional file 1: Table S2. Next, we shortened the HRT of bioreactor B from 8 days to 2 days in a fast transition mode and with the same substrate load as in bioreactor A. Considering the effect of time on community assembly, we conducted unequal HRT changes in two bioreactors and aimed to delineate the model prediction strength with the two different datasets. Finally, both bioreactors were operated at an HRT of 2 days until day 211.

Gas samples were taken through the septum twice per week. Samples for determining cell mass concentrations were collected from the reactor effluent. Concentrations of xylan,

carboxylates and alcohols were measured in the effluent supernatants [10]. In total, samples were collected on 59 time points for each bioreactor. At the beginning and the end of the experiment, pelleted biomass from the effluent was used to determine the cell dry mass as previously described [10]. For microbial community analysis, pelleted cells from 2 mL effluent were washed with 100 mM Tris-HCl pH 8.5 and stored at -20˚C until DNA extraction.

Analytical methods. Daily produced gas volume was monitored with the MGC-1 and normalised to standard pressure and temperature [23]. Gas composition (H_2, CO_2, N_2, O_2 and CH_4) was determined by gas chromatography in triplicate [24]. Concentrations of carboxylates and alcohols were analysed in triplicate by gas chromatography [10]. Concentration of xylan was measured by a modified dinitrosalicylic acid reagent method [10]. Cell mass concentration was calculated from optical density (OD) values that were correlated with the cell dry mass [10]. The calculated mean correlation coefficients were 1 OD_{600} = 0.548 g L^{-1} for bioreactor A and 1 OD_{600} = 0.537 g L^{-1} for bioreactor B.

Microbial community analysis. Total DNA was isolated from frozen cell pellets using the NucleoSpin® Microbial DNA Kit (Macherey-Nagel, Düren, Germany). Methods for DNA quantification and quality control were as described before [25]. For high-throughput amplicon sequencing, V3-V4 regions of the 16S rRNA genes were PCR-amplified using primers 341f and 785r [26]. Sequencing was performed on the Illumina Miseq platform (Miseq Reagent Kit v3; 2 × 300 bp). A total of 12,168,404 sequences ranging from 57,612 to 389,963 pairs of reads per sample (mean: 135,205; median: 122,367) were obtained.

The demultiplexed sequence data were processed with the QIIME 2 v2019.7 pipeline

[27] using the DADA2 plugin [28]. The DADA2 parameters were set as follows: trim-left-f 0, trim-left-r 0, trunc-len-f 270, trunc-len-r 230, max-ee 2 and chimera-method consensus. A total of 4,194,700 sequences ranging from 13,518 to 138,498 reads per sample were retained, with a mean of 46,608 reads per sample. The generated feature table indicates the frequency of each ASV clustered at 100% identity. Taxonomic assignment was done with a naïve Bayes classifier trained on 16S rRNA gene sequences of the database MiDAS 2.1 [29], and curated using the RDP Classifier 2.2 with a confidence threshold of 80% [30]. For downstream analyses, ASVs of all samples were rarefied to a sequencing depth of 13,518 reads (rarefaction curve reached the plateau, Additional file 1: Figure S1). We obtained a total of 71 unique ASVs in 90 samples.

Alpha diversity based on rarefied ASV data was evaluated by the observed ASV counts and the Shannon index [31], which were determined using the R package phyloseq v1.30.0 [32]. Dissimilarities in bacterial community composition (beta-diversity) were calculated using Bray-Curtis distance [33] based on rarefied ASV abundances and visualised as nonmetric multidimensional scaling (NMDS) plots. Statistical analyses of beta-diversity results were performed using permutational multivariate analysis of variance (PERMANOVA) [34] in the R package "vegan" (v2.5.6, "adonis" function, Monto-Carlo test with 1000 permutations); *P* values were adjusted for multiple comparisons using the false discovery rate (FDR) method [35].

Network analysis. The co-occurrence network analysis was performed using the method described by Ju et al. [36]. Briefly, we constructed a correlation matrix by computing possible pairwise Spearman's rank correlations using the rarefied ASV abundances and abiotic parameters (HRT; concentrations of C4, C6, C8 and lactate; productivities and yields of C4, C6 and C8). Correlation coefficients below -0.7 or above 0.7 and adjusted *P*-values (FDR method)

lower than 0.05 were considered statistically robust. Network visualisation and topological features analysis were conducted in Gephi (v0.9.2) [37].

16S rRNA phylogenetic analysis. The 16S rRNA gene sequences of ASVs were aligned using the SINA alignment algorithm [38] via the SILVA web interface [39]. We additionally used SINA to search and classify the sequences with the least common ancestor method based on the SILVA taxonomy. For each query sequence, the minimum identity was set to 0.95 and the five nearest neighbours were considered. The tree was reconstructed based on the aligned sequences and their neighbours, with RAxML using the GTRCAT model of evolution. Later only ASV species of this study were kept in the generated tree for an easier viewing. The tree was visualised using iTOL [40].

Metagenomic analysis. Six samples were selected for whole-genome sequencing, which was performed by StarSEQ GmbH (Mainz, Germany), using the Illumina NextSeq 500 system (NEBNext Ultra II FS DNA library prep kit; 2 × 150 bp) with at a minimum of 20 million reads per library generated. Quality checking and reads trimming were performed using metaWRAP (v0.7, raw read QC module) [41] and TrimGalore (v0.4.3) [42]. Reads of human origin were discriminated from microbial reads using BMTagger (v3.101) [43]. All adapters were removed and the resulting reads were assembled using metaSPAdes (v3.11.1) [44]. Paired-end reads were aligned back to the assembly using BWA (v0.7.15, mem algorithm) [45]. Binning of assembled contigs was performed using the metaWRAP modules metaBAT (2.12.1) [46], MaxBin (2.2.4) [47] and CONCOCT (1.0.0) [48]. The metaWRAP-Bin_refinement module was applied to separate the overlaps between two bins. Quality of metagenome-assembled genomes (MAGs)

was checked using CheckM (v1.0.7) [49]. MAGs were classified in high or medium quality regarding completeness, contamination, quality score (completeness - 5 × contamination) and strain heterogeneity [50]. The following thresholds were used for high quality: quality score > 50, completeness > 80, contamination < 5 and strain heterogeneity < 50; and for medium quality: quality score > 50, completeness > 50 and contamination < 10. One bin with lower quality was removed from the analysis. The taxonomy was assigned using GTDB-Tk (v0.3.2) [51]. Genome metrics were calculated with the statswrapper tool in the BBTools suite [52]. A phylogenomic tree based on Mash distances was generated with Mashtree (V1.1.2) [53] and visualised in iTOL [40]. Miscellaneous visualisations of the dataset metrics were performed in R with the packages ggplot2 (v3.3.0) and DataExplorer (v0.8.1). Species differentiation was performed using fastANI [54] and aniSplitter.R (http://github.com/felipborim789/aniSplitter/). Genomes were annotated with Prokka (v1.14.6) [55]. Functional annotation of genes relevant to xylan hydrolysis, xylose fermentation and chain elongation was curated using Swiss-Prot, COG and GenBank [56–58].

Determining bioindicators of HRT changes. The HRT bioindicators were determined using the random forest algorithm (randomForest R package, v4.6-14) [59]. ASV relative abundances were used as features to train and test the random forest classifier. Considering how we replicated the HRT changing mode in both bioreactors (Additional file 1: Table S2), the whole operation period was divided into four sampling intervals: 0-50 days, 51-100 days, 101-140 days and 141-211 days. Based on the results of community analysis, we chose the ASV data of both bioreactors in the sampling intervals of 0-50 days and 141-211 days to determine the HRT bioindicators, and we used data of all samples in the four HRT phases as controls. To evaluate the robustness of the predictions, we trained the classifier with ASV data of one bioreactor and

231 tested in the other bioreactor and vice versa. For random forest classification analysis,
232 importance of the different features (ASVs) was measured by the Gini index (mean decrease in
233 Gini, default in randomForest R package; where larger values indicate a variable to be more
234 important for accurate classification [60]).

235 The random forest classifier was trained on the training set, with 2,000 trees and 40
236 variables (with lowest out-of-bag estimated error rates achieved) being selected randomly for
237 each tree. Explained variance (% Var. explained in R) was used to measure the model
238 performance on the training set [59]. We predicted the accuracy by measuring how well the
239 features can classify the HRT phases on the test set [60]. We first computed the feature
240 importance of all 71 ASVs. Then at each step, the ASVs having the smallest importance were
241 eliminated and a new forest was built with the remaining ASVs. For both bioreactors, the
242 features were selected when their Gini scores were higher than 1% of the sum of the Gini scores
243 of all ASVs (Additional file 1: Table S3). Finally, we selected the 15 top-ranked ASVs leading to
244 the model of smallest error rate for classifying the HRT phases of 8 days and 2 days. In each
245 bioreactor, the 15 ASVs that best discriminated between HRT phases were referred to as A-HRT
246 bioindicators or B-HRT bioindicators (bioreactors A and B, respectively). ASVs common to both
247 sets were defined as HRT bioindicators (workflow of random forest classification in Additional
248 file 1: Figure S2).
249
250 **Quantitative predictions based on HRT and non-HRT bioindicators.** The process parameters
251 specified as concentrations of lactate, C4, C6 and C8, and productivities as well as yields of C4,
252 C6 and C8 were the prediction objects. Here, the relevance of the different ASVs to the
253 prediction was determined by residual sum of squares (IncNodePurity, default in randomForest)

for the regressions. Explained variance (% Var. explained in R) was used to measure the model performance on the training set [59]. We predicted the accuracy by measuring how well the features can explain the variance of these process parameters on the test set [60].

We performed the quantitative prediction by applying a two-step regression analysis (workflow in Additional file 1: Figure S3). First, HRT bioindicators were used to predict the data of different process parameters in the sampling intervals of 0-50 days and 141-211 days. Data of all samples in the four HRT phases were considered as controls. Random forest regressors were trained as follows: relative abundance dataset of bioreactor A was used as training set and that of bioreactor B was used as test set and vice versa; 2,000 trees and four out of 11 features were selected randomly for each tree.

Considering community assembly caused by time, we determined the ASVs (non-HRT bioindicators) that could predict the numeric values of each process parameter, using data of samples in the intervals of 0-50 days and 141-211 days. For each process parameter, we started with computing the feature importance of all ASVs and further selected the 15 top-rated ASVs as the bioindicators of this non-HRT parameter. The model was trained as follows: datasets of bioreactors A and B were independently used for training and testing; 2,000 trees and five out of 15 features were selected randomly for each tree. As controls, we used the non-HRT bioindicators of each parameter to predict the corresponding data of all samples in the four HRT phases. The final set of ASVs presented in HRT bioindicators and not in non-HRT bioindicators were considered HRT bioindicators irrespective of time.

Evaluating prediction accuracy. When in both training sets the HRT bioindicators and non-HRT bioindicators explained more than 80% of the variance in a process parameter, we

proceeded only with those parameters. To compare the predicted and measured values for these process parameters, we considered the following performance metrics for reflecting the error of the model in predicting consecutive data: relative root mean square error (RRMSE, cutoff < 10%); R squared, slope and intercept of the least squares line of best fit. The final values of RRMSE were averaged among the 100 random forest replicates, with four ASVs for HRT bioindicators and five for non-HRT bioindicators randomly sampled at each replicate.

Results and discussion

Effects of HRT decrease on process performance and microbial diversity. The progressive HRT decrease from 8 to 2 days increased the C6 and C8 productivities and yields in two independent bioreactors (Figure 1). We first shortened the HRT to 6 days and then to 4 days in bioreactor A, which allowed the reactor microbiota to adapt to the new conditions and improved productivities of C4, C6 and C8 (Figure 1a). Further HRT decrease to 2 days confirmed the increasing trend in productivity. At the end of the 2-day HRT period in bioreactor A, we achieved the highest productivities (mmol C L^{-1} d^{-1}) of C4, C6 and C8 up to 115.0, 64.1 and 5.9, respectively. To confirm the observed effects of HRT shortening on the CE process and reactor microbiota, we executed a fast transition mode in bioreactor B and generated a different dataset from the parallel system. Comparable increases in productivity were observed (Figure 1b). We obtained maximum productivities (mmol C L^{-1} d^{-1}) of C4 up to 102.4, C6 up to 62.9 and C8 up to 7.0. The C6 and C8 yields (in terms of C mole product to transferred substrate ratio) increased along with decreasing HRT at the cost of C4 yield. Compared with yields at the 8-day HRT, C6 and C8 yields were higher and the C4 yield was lower in both bioreactors at the 2-day HRT

(Figure 1 and Additional file 1: Table S4). Our results suggest that the shorter HRT favoured lactate-based CE producing C6 and C8 over C4 production. C4 can be produced by CE of acetate but also from sugars by butyric acid fermentation [61]. Decreasing the HRT to 2 days led to the accumulation of lactate and fluctuations of the C4, C6 and C8 production, which lasted longer than 22 HRTs in bioreactor A (Figure 1a). Lactate concentrations were highly correlated with C4 fluctuations (Spearman Rho = -0.90, $P < 0.05$) and C6 concentrations (Rho = -0.89, $P < 0.05$), which reflects how lactate was produced and converted by the reactor microbiota. The HRT reduction resulted in higher gas production and hydrogen content (Additional file 1: Figure S4). Besides, an increase in cell mass production (Additional file 1: Figure S5) suggests a facilitating effect of short HRT on the growth of enriched populations with desirable activities, i.e. more biocatalysts were available in the high C6/C8 production phase.

Decreasing the HRT affected the composition and diversity of the reactor microbiota. Changes in relative abundance of ASVs categorised from phylum to genus between the HRT of 8 days and 2 days are shown in Additional file 1: Figure S6. Alpha diversity metrics showed significantly lower observed ASV counts (pairwise *t*-test, $P < 0.05$) and higher Shannon index values (pairwise *t*-test, $P < 0.05$) for HRT of 8 days compared with 2 days (Additional file 1: Figure S7). Beta diversity analysis revealed a significant difference between the communities at different HRTs (PERMANOVA; Pseudo-$F = 103.1$, $P < 0.001$) but no significant difference between the communities in both reactors at the same HRT (Pseudo-$F = 3.3$, $P > 0.05$) (Figure 2).

HRT bioindicators predicting process performance. To determine HRT bioindicators, we used HRT of 8 days and 2 days as classes for the random forest classification model and relative

abundances of ASVs as the features. To delineate the model prediction strength, we used one reactor dataset to train the model while testing predictions with the other and vice versa. Feature selection based on the random forest classifier with its associated Gini index has shown abilities to identify optimal feature subsets in high-dimensional data [62]. Based on higher than 1% of the mean decrease in Gini scores for both reactors in the prediction accuracy of HRT phases, we selected 15 top-ranked ASVs that would give the best discrimination between HRT phases. The 15 ASVs most relevant to HRT changes were defined as "A- or B-HRT bioindicators", potentially reflecting the key species correlating with HRT changes in either bioreactor (feature importance in Figure 3). The two bioreactors shared 11 HRT bioindicators.

To answer the question whether HRT bioindicators can be used to predict process performance in terms of C6 and C8 production, we performed a random forest regression analysis in two steps. HRT bioindicators were first chosen as features to train the model. Considering community assembly caused by time, we then determined 15 ASVs most relevant to each non-HRT process parameter (i.e., concentrations of lactate, C4, C6 and C8; productivities and yields of C4, C6 and C8; hereafter, non-HRT bioindicators). Datasets from bioreactors A and B were trained and tested independently. When in both reactors the HRT and non-HRT bioindicators accounted for more than 80% of the variance in a process parameter, we proceeded only with those parameters. In our case, the model could explain more than 80% of the variance in C6 and C8 productivities (Additional file 1: Tables S5-S6).

We evaluated the prediction performance of the model by comparing the predicted and measured values of process parameters. RRMSE was used as the performance metric to reflect the model error in predicting quantitative data of C6/C8 productivity. Our results showed that the C6 and C8 productivities of both bioreactors at the HRT of 8 days and 2 days could be accurately

predicted (Figure 4 for HRT bioindicators and Additional file 1: Figure S8 for non-HRT bioindicators). We further tested samples in all HRT phases with HRT and non-HRT bioindicators. The C6 and C8 productivities were also accurately predicted (RRMSE < 6%, Additional file 1: Figures S9-S10). Therefore, we considered HRT bioindicators irrespective of time as the ASVs presented in HRT bioindicators and not in non-HRT bioindicators (feature importance in Additional file 1: Figures S11-S12). Interestingly, the same four ASVs assigned to the genera *Olsenella*, *Lactobacillus*, *Syntrophococcus* and *Clostridium* IV were identified for C6 and C8 productivity (Figure 5). We thus hypothesise that species represented by these four ASVs determined the increased C6/C8 productivities in the CE process manipulated by changing operational conditions – shortening the HRT.

Functional role of HRT bioindicators. Combined with metagenomics, species of HRT bioindicators irrespective of time indicated their roles in driving the catabolism of xylan and lactate to C6/C8 (Figure 6). Among 108 MAGs (dereplicated into 29 species; Figure 7 and Additional file 2), we recovered 12 species with similar phylogenies as the four genera (Table 1). In view of the fermentation process, we annotated the genetic potential for xylan hydrolysis, xylose fermentation and CE with lactate (Additional file 1: Figure S13 and Additional files 3-6). Specifically, *Clostridium* IV species were reported as lactate-based chain-elongating bacteria [63]. Our results suggest that four *Clostridium* IV species (*Acutalibacteraceae* spp. according to GTDB-Tk) can convert lactate to C6/C8. Two *Syntrophococcus* species (*Eubacterium*_H spp. according to EZBioCloud [64]) are potential C6/C8-producers as they hold complete gene sets encoding enzyme complexes that catalyse CE reactions. This genetic potential was also found in genomes of closely related *Syntrophococcus* species (*Eubacterium cellulosolvens* according to

EZBioCloud; Additional file 6), which was not described before. Lactate formation from xylose by lactic acid bacteria can enhance CE by providing additional electron donors [23,65–68]. A recent study reported an enriched community dominated by *Lactobacillus* and chain-elongating species, and their co-occurrence suggested lactate produced by *Lactobacillus* to be a key intermediate for C6/C8 production [69]. Network analysis of our previous study [10] revealed the co-occurrence of *Olsenella* with potential chain-elongating species. Species of *Lactobacillus* and *Olsenella* are potential xylose-consuming lactate producers (Figure 6b). Genes encoding xylanases were not found in *Lactobacillus* MAGs but in those assigned to other bioindicators (Figure 6a). Taken together, the delineated synergy effects between these bioindicator species suggest a division of labour with mutual benefits, converting xylan and lactate to C6/C8. A correlation network shows HRT, C6 and C8 productivity being the most highly connected nodes (Additional file 1: Figure S14). Their co-occurrence with ASVs assigned to *Clostridium* IV, *Olsenella* and *Syntrophococcus* indicates strong associations among these taxa, the changed environment and corresponding functions. The predictability of C6 and C8 productivities was relatively poor when using only the four HRT bioindicators irrespective of time (Additional file 1: Figure S15). Besides, we found redundancy in the main functions of catabolising xylan and lactate to C4, C6 and C8 (Figure 6), with the relevant HRT bioindicators increasing in relative abundances (Additional file 1: Figure S16). Thus, the involved metabolic pathways can be strongly coupled to HRT decreases. The genetic potential overlaps with other distinct taxa of the reactor microbiota, suggesting that HRT bioindicators might be key species of the process, but ecological interactions with other species are critical to ensure the C6/C8 production (functional annotations of xylose fermentation and chain elongation in Additional files 5-6).

Conclusions

Our approach enabled the quantitative prediction of process performance in the anaerobic bioreactor system (Figure 8). In artificial ecosystems with well-controlled conditions (temperature, pH and no immigration of other microbes; Figure 8a), HRT was the most influencing factor controlling community assembly (Figure 8b). However, we cannot exclude the impact of other deterministic factors like microbial interactions within temporal patterns, particularly for such a long-term reactor operation. Effects of compositional stochasticity on community assembly also need to be considered [70,71]. Further studies on these ecological principles will help manage reactor microbiota towards beneficial traits, such as high specificities for C6/C8 production.

The continuous reactor systems with enrichment cultures enabled to select communities with desired CE functions (i.e., high C6 and C8 productivities), and to demonstrate that 16S rRNA amplicon sequencing data can be used to predict CE process performance quantitatively (> 90% accuracy). The described machine learning framework (Figure 8c) may be suitable for other ecosystem processes and more complex communities. For that, it would be necessary to design experiments with (i) sufficient temporal and/or spatial resolution, (ii) parallel sampling for amplicon sequencing data and metadata from desired ecosystem processes, and (iii) correlation of phylogenetic diversity with the ecosystem processes. Our approach was based on phylogenetic diversity that in some ecosystems may correlate with ecosystem processes where microbiota perform key functions. Our general methodology can be adapted to other data types, such as metagenomes, metatranscriptomes, metaproteomes or metabolomes, and it opens new doors for prediction and hypothesis testing in microbial ecology.

List of abbreviations

ASVs: Amplicon Sequence Variants, C4: n butyrate, C6: n-caproate, C8: n-caprylate, CE: chain elongation, FDR: false discovery rate, GTDB: Genome Taxonomy Database, HRT: hydraulic retention time, MAGs: metagenome-assembled genomes, NMDS: nonmetric multidimensional scaling, OD: optical density, PERMANOVA: permutational multivariate analysis of variance, RRMSE: relative root mean square error.

Declarations

Availability of data and materials

All data described in this manuscript are present in the paper and/or the Supplementary material. Both amplicon sequencing data (ERR4158761 to ERR4158850) and metagenome sequencing data (ERR4183110 to ERR4183115) have been deposited in the European Nucleotide Archive (ENA) under study no. PRJEB38353. The MAGs are publicly available in the ENA under the sample accession no. ERS4594296 to ERS4594324.

Funding

The study was supported by the Initiative and Networking Fund of the Helmholtz Association. B.L. was supported by the China Scholarship Council (# 201606350010). J.S. and U.R. were financed by the Helmholtz Young Investigator grant VH-NG-1248 Micro 'Big Data'. H.S., H.H. and S.K. were financed by the BMBF – German Federal Ministry of Education and Research (# 031B0389B and # 01DQ17016) and the Helmholtz Association (Program Renewable Energies). S.G.S. was the recipient of a PhD scholarship conceded by FCT (PD/BD/143029/2018).

Authors' contributions

B.L., H.S., J.S., S.K. and U.R. designed the study and the experiments. B.L. performed the experiments and analysed the reactor data as well as sequencing data. B.L., J.S. and U.R. performed the machine learning analysis. B.L., H.S., J.S., S.G.S., S.K. and U.R. contributed to data analysis and interpretation. H.H. contributed to the discussion of the results. All authors critically contributed to the preparation of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors thank Ute Lohse for her technical assistance in molecular analyses, and the colleagues from DBFZ Deutsches Biomasseforschungszentrum GmbH for their technical support in analyses of abiotic parameters. We thank Rodolfo Brizola Toscan, Felipe Borim Corrêa and Jonas Coelho Kasmanas for their help with data analysis. We also thank Masun Nabhan Homsi for valuable discussions regarding our machine learning analysis.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

References

1. Banerjee S, Schlaeppi K, van der Heijden MGA. Keystone taxa as drivers of microbiome structure and functioning. Nat Rev Microbiol. 2018;16:567–576.

2. de los Reyes FL. Challenges in determining causation in structure-function studies using molecular biological techniques. Water Res. 2010;44:4948–57.

3. Harms H, Harnisch F, Koch C, Müller S. Microbiomes in bioenergy production: from analysis to management. Curr Opin Biotechnol. 2014;27:65–72.

4. Verstraete W, Wittebolle L, Heylen K, Vanparys B, de Vos P, van de Wiele T, et al. Microbial resource management: the road to go for environmental biotechnology. Eng Life Sci. 2007;2:117–26.

5. Kleerebezem R, van Loosdrecht MC. Mixed culture biotechnology for bioenergy production. Curr Opin Biotechnol. 2007;18:207–12.

6. Lawson CE, Harcombe WR, Hatzenpichler R. Common principles and best practices for engineering microbiomes. Nat Rev Microbiol. 2019;17:725–41.

7. Goldford JE, Lu N, Bajić D, Estrela S, Tikhonov M, Sanchez-Gorostiaga A, et al. Emergent simplicity in microbial community assembly. Science. 2018;361:469–74.

8. Zuñiga C, Li CT, Yu G, Al-Bassam MM, Li T, Jiang L, et al. Environmental stimuli drive a transition from cooperation to competition in synthetic phototrophic communities. Nat Microbiol. 2019;4:2184–91.

9. Angenent LT, Richter H, Buckel W, Spirito CM, Steinbusch KJJ, Plugge CM, et al. Chain elongation with reactor microbiomes: open-culture biotechnology to produce biochemicals. Environ Sci Technol. 2016;50:2796–810.

10. Liu B, Kleinsteuber S, Centler F, Harms H, Sträuber H. Competition between butyrate fermenters and chain-elongating bacteria limits the efficiency of medium-chain carboxylate production. Front Microbiol. 2020;11:336.

11. Lambrecht J, Cichocki N, Schattenberg F, Kleinsteuber S, Harms H, Müller S, et al. Key sub-community dynamics of medium-chain carboxylate production. Microb Cell Fact. 2019;18:92.

12. Kucek LA, Spirito CM, Angenent LT. High *n*-caprylate productivities and specificities from dilute ethanol and acetate: chain elongation with microbiomes to upgrade products from syngas fermentation. Energy Environ Sci. 2016;9:3482–94.

13. Kucek LA, Nguyen M, Angenent LT. Conversion of L-lactate into *n*-caproate by a continuously fed reactor microbiome. Water Res. 2016;93:163–71.

14. Duber A, Jaroszynski L, Zagrodnik R, Chwialkowska J, Juzwa W, Ciesielski S, et al. Exploiting the real wastewater potential for resource recovery – *n*-caproate production from acid whey. Green Chem. 2018;20:3790–803.

15. Grootscholten TIM, Steinbusch KJJ, Hamelers HVM, Buisman CJN. Improving medium chain fatty acid productivity using chain elongation by reducing the hydraulic retention time in an upflow anaerobic filter. Bioresour Technol. 2013;136:735–8.

16. Nzeteu CO, Trego AC, Abram F, O'Flaherty V. Reproducible, high-yielding, biological caproate production from food waste using a single-phase anaerobic reactor system. Biotechnol Biofuels. 2018;11:108.

17. Mansfeldt C, Achermann S, Men Y, Walser JC, Villez K, Joss A, et al. Microbial residence time is a controlling parameter of the taxonomic composition and functional profile of microbial communities. ISME J. 2019;13:1589–601.

18. Bonk F, Popp D, Weinrich S, Sträuber H, Becker D, Kleinsteuber S, et al. Determination of

507 microbial maintenance in acetogenesis and methanogenesis by experimental and modelling
508 techniques. Front Microbiol. 2019;10:166.
509 19. Werner JJ, Knights D, Garcia ML, Scalfone NB, Smith S, Yarasheski K, et al. Bacterial
510 community structures are unique and resilient in full-scale bioenergy systems. Proc Natl Acad
511 Sci USA. 2011;108:4158–63.
512 20. Oyetunde T, Bao FS, Chen JW, Martin HG, Tang YJ. Leveraging knowledge engineering
513 and machine learning for microbial bio-manufacturing. Biotechnol Adv. 2018;36:1308–15.
514 21. Lopatkin AJ, Collins JJ. Predictive biology: modelling, understanding and harnessing
515 microbial complexity. Nat Rev Microbiol. 2020;18:507-20.
516 22. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
517 23. Sträuber H, Bühligen F, Kleinsteuber S, Dittrich-Zechendorf M. Carboxylic acid production
518 from ensiled crops in anaerobic solid-state fermentation – trace elements as pH controlling
519 agents support microbial chain elongation with lactic acid. Eng Life Sci. 2018;0:447–58.
520 24. Urban C, Xu J, Sträuber H, dos Santos Dantas TR, Mühlenberg J, Härtig C, et al. Production
521 of drop-in fuel from biomass by combined microbial and electrochemical conversions. Energy
522 Environ Sci. 2017;10:2231–44.
523 25. Lucas R, Kuchenbuch A, Fetzer I, Harms H, Kleinsteuber S. Long-term monitoring reveals
524 stable and remarkably similar microbial communities in parallel full-scale biogas reactors
525 digesting energy crops. FEMS Microbiol Ecol. 2015;91:fiv004.
526 26. Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, et al. Evaluation of
527 general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-
528 based diversity studies. Nucleic Acids Res. 2013;41:e1.
529 27. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Chase J, Cope EK, et al. Reproducible,

530 interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol.
531 2019;37:852–7.
532 28. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: high-
533 resolution sample inference from Illumina amplicon data. Nat Methods. 2016;13:581–3.
534 29. McIlroy SJ, Kirkegaard RH, McIlroy B, Nierychlo M, Kristensen JM, Karst SM, et al.
535 MiDAS 2.0: an ecosystem-specific taxonomy and online database for the organisms of
536 wastewater treatment systems expanded for anaerobic digester groups. Database.
537 2017;2017:bax016.
538 30. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naïve Bayesian classifier for rapid assignment of
539 rRNA sequences. Appl Environ Microbiol. 2007;73:5261–7.
540 31. Shannon CE. A mathematical theory of communication. Bell Syst Tech J. 1948;27:379–423.
541 32. McMurdie PJ, Holmes S. Phyloseq: an R package for reproducible interactive analysis and
542 graphics of microbiome census data. PLoS One. 2013;8:e61217.
543 33. Bray JR, Curtis JT. An ordination of the upland forest communities of southern Wisconsin.
544 Ecol Monogr. 1957;27:325–49.
545 34. Anderson MJ. A new method for non-parametric multivariate analysis of variance. Austral
546 Ecol. 2001;26:32–46.
547 35. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful
548 approach to multiple Testing. J R Stat Soc B. 1995;57:289–300.
549 36. Ju F, Xia Y, Guo F, Wang Z, Zhang T. Taxonomic relatedness shapes bacterial assembly in
550 activated sludge of globally distributed wastewater treatment plants. Environ Microbiol.
551 2014;16:2421–32.
552 37. Bastian M, Heymann S, Jacomy M. Gephi: an open source software for exploring and

553 manipulating networks. Int AAAI Conf Weblogs Soc Media. 2009;8:361–2.
554 38. Pruesse E, Peplies J, Glöckner FO. SINA: accurate high-throughput multiple sequence
555 alignment of ribosomal RNA genes. Bioinformatics. 2012;28:1823–9.
556 39. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, et al. SILVA: a
557 comprehensive online resource for quality checked and aligned ribosomal RNA sequence data
558 compatible with ARB. Nucleic Acids Res. 2007;35:7188–96.
559 40. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and
560 annotation of phylogenetic and other trees. Nucleic Acids Res. 2016;44:W242–5.
561 41. Uritskiy G V., Diruggiero J, Taylor J. MetaWRAP – a flexible pipeline for genome-resolved
562 metagenomic data analysis. Microbiome. 2018;6:158.
563 42. Galore K. Trim Galore!: a wrapper tool around Cutadapt and FastQC to consistently apply
564 quality and adapter trimming to FastQ files. 2015. Available from:
565 https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
566 43. Rotmistrovsky, K. Agarwala R. BMTagger: best match tagger for removing human reads
567 from metagenomics datasets. 2010. Available from:
568 ftp://ftp.ncbi.nlm.nih.gov/pub/agarwala/bmtagger/
569 44. Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. MetaSPAdes: a new versatile
570 metagenomic assembler. Genome Res. 2017;27:824–34.
571 45. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform.
572 Bioinformatics. 2009;25:1754–60.
573 46. Kang DD, Froula J, Egan R, Wang Z. MetaBAT, an efficient tool for accurately
574 reconstructing single genomes from complex microbial communities. PeerJ. 2015;3:e1165.
575 47. Wu YW, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover

576 genomes from multiple metagenomic datasets. Bioinformatics. 2016;32:605–7.
577 48. Alneberg J, Bjarnason BS, De Bruijn I, Schirmer M, Quick J, Ijaz UZ, et al. Binning
578 metagenomic contigs by coverage and composition. Nat Methods. 2014;11:1144–6.
579 49. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the
580 quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome
581 Res. 2015;25:1043–55.
582 50. Parks DH, Rinke C, Chuvochina M, Chaumeil PA, Woodcroft BJ, Evans PN, et al. Recovery
583 of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat
584 Microbiol. 2017;2:1533–42.
585 51. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify
586 genomes with the Genome Taxonomy Database. Bioinformatics. 2019;36:1925–7.
587 52. Bushnell B. BBMap short read aligner, and other bioinformatic tools. Available from:
588 http://sourceforge.net/projects/bbmap
589 53. Katz L, Griswold T, Morrison S, Caravas J, Zhang S, Bakker H, et al. Mashtree: a rapid
590 comparison of whole genome sequence files. J Open Source Softw. 2019;4:1762.
591 54. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI
592 analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun.
593 2018;9:5114.
594 55. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30:2068–9.
595 56. Bateman A. UnitProt: a worldwide hub of protein knowledge. Nucleic Acids Res.
596 2019;47:D506–15.
597 57. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, et al. The COG
598 database: an updated version includes eukaryotes. BMC Bioinformatics. 2003;4:41.

599 58. Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res.
600 2016;44:D67-72.
601 59. Liaw A, Wiener M. Classification and regression with random forest. R News. 2002;2:18–
602 22.
603 60. Huang BFF, Boutros PC. The parameter sensitivity of random forests. BMC Bioinformatics.
604 2016;17:331.
605 61. Temudo MF, Mato T, Kleerebezem R, Van Loosdrecht MCM. Xylose anaerobic conversion
606 by open-mixed cultures. Appl Microbiol Biotechnol. 2009;82:231–9.
607 62. Menze BH, Kelm BM, Masuch R, Himmelreich U, Bachert P, Petrich W, et al. A comparison
608 of random forest and its Gini importance with standard chemometric methods for the feature
609 selection and classification of spectral data. BMC Bioinformatics. 2009;10:213.
610 63. Zhu X, Zhou Y, Wang Y, Wu T, Li X, Li D, et al. Production of high-concentration *n*-
611 caproic acid from lactate through fermentation using a newly isolated *Ruminococcaceae*
612 bacterium CPB6. Biotechnol Biofuels. 2017;10:102.
613 64. Yoon SH, Ha SM, Kwon S, Lim J, Kim Y, Seo H, et al. Introducing EzBioCloud: a
614 taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. Int J
615 Syst Evol Microbiol. 2017;67:1613–7.
616 65. Xu J, Hao J, Guzman JJL, Spirito CM, Harroff LA, Angenent LT. Temperature-phased
617 conversion of acid whey waste into medium-chain carboxylic acids via lactic acid: no external e-
618 donor. Joule. 2018;2:1–16.
619 66. Scarborough MJ, Lynch G, Dickson M, McGee M, Donohue TJ, Noguera DR. Increasing the
620 economic value of lignocellulosic stillage through medium-chain fatty acid production.
621 Biotechnol Biofuels. 2018;11:200.

67. Khor WC, Andersen S, Vervaeren H, Rabaey K. Electricity-assisted production of caproic acid from grass. Biotechnol Biofuels. 2017;10:180.

68. Andersen SJ, de Groof V, Khor WC, Roume H, Props R, Coma M, et al. A *Clostridium* group IV species dominates and suppresses a mixed culture fermentation by tolerance to medium chain fatty acids products. Front Bioeng Biotechnol. 2017;5:8.

69. Contreras-Dávila CA, Carrión VJ, Vonk VR, Buisman CNJ, Strik DPBTB. Consecutive lactate formation and chain elongation to reduce exogenous chemicals input in repeated-batch food waste fermentation. Water Res. 2020;1:115215.

70. Chase JM. Stochastic community assembly causes higher biodiversity in more productive environments. Science. 2010;328:1388–91.

71. Ofiţeru ID, Lunn M, Curtis TP, Wells GF, Criddle CS, Francis CA, et al. Combined niche and neutral effects in a microbial wastewater treatment community. Proc Natl Acad Sci USA. 2010;107:15345–50.

Additional files

Additional file 1: Figure S1. Alpha rarefaction curves. **Figure S2.** Workflow of the random forest classification analysis. **Figure S3.** Workflow of a two-step random forest regression analysis. **Figure S4.** Gas production of bioreactors. **Figure S5.** Biomass production of bioreactors. **Figure S6.** Microbial community composition profiles of bioreactors. **Figure S7.** Alpha diversity metrics of bioreactor communities. **Figure S8.** Prediction results of C6 and C8 productivities using non-HRT bioindicators for considering community assembly caused by time. **Figure S9.** Prediction results of C6 and C8 productivities for all samples in the four HRT phases using HRT bioindicators. **Figure S10.** Prediction results of C6 and C8 productivities for all samples in the four HRT phases using non-HRT bioindicators for considering community assembly caused by time. **Figure S11.** Random forest feature importance of A-HRT bioindicators and B-HRT bioindicators used to predict C6 and C8 productivities. **Figure S12.** Random forest feature importance of the non-HRT bioindicators used to predict C6 and C8 productivities. **Figure S13.** Metabolic pathways involved in converting lactate and xylan to *n*-caproate and *n*-caprylate. **Figure S14.** Correlation network of environmental factors, process performance and microbial community. **Figure S15.** Prediction results of C6 and C8 productivities for all samples in the four HRT phases using the four ASVs of HRT bioindicators irrespective of time. **Figure S16.** Reducing HRT increases abundances of HRT bioindicators driving the catabolism of xylan and lactate to *n*-caproate and *n*-caprylate. **Table S1.** Growth medium used for the reactor operation. **Table S2.** Daily feeding of bioreactors A and B during the four HRT phases. **Table S3.** Gini scores of all ASVs in the classification-based prediction of HRT phases. **Table S4.** Mean carboxylate yields (i.e. C mole product to substrate ratios) at

HRTs of 8 days and 2 days (stable production period). **Table S5.** Explained variances of the training set in the regression-based prediction of process parameters using A-HRT bioindicators and B-HRT bioindicators. **Table S6.** Explained variances of the training set in the regression-based prediction of process parameters using non-HRT bioindicators for considering community assembly caused by time.

Additional file 2: Dataset S1. MAGs taxonomy and genome metrics.

Additional file 3: Dataset S2. Functional annotations of xylose fermentation for MAGs with the same taxonomy as HRT bioindicators.

Additional file 4: Dataset S3. Functional annotations of chain elongation for MAGs with the same taxonomy as HRT bioindicators.

Additional file 5: Dataset S4. Functional annotations of xylose fermentation for all MAGs.

Additional file 6: Dataset S5. Functional annotations of chain elongation for all MAGs.

Figure legends

Figure 1. Performance of bioreactors. Concentrations of chain elongation products and lactate, as well as productivities and yields of chain elongation products in bioreactors A (**a**) and B (**b**) during the four HRT phases. Chain elongation products: C4, *n*-butyrate; C6, *n*-caproate; C8, *n*-caprylate.

Figure 2. Dissimilarities in bacterial community composition (beta-diversity). Non-metric multidimensional scaling (NMDS) based on Bray-Curtis dissimilarities of microbial community composition in bioreactors. **a,** All samples in the four HRT phases were considered for dissimilarity calculation. **b,** Samples in the 8-day HRT phase classified to the sampling interval 0-50 days and in the 2-day HRT phase classified to the interval 141-211 days were included.

Figure 3. Random forest feature importance of ASVs used to classify the HRT phases (A-HRT bioindicators and B-HRT bioindicators). The top-ranked 15 ASVs reducing the uncertainty in the prediction of HRT phases (HRT of 8 days and 2 days). The order of features (from top to bottom) was based on their mean decrease in Gini scores, according to their ASV abundances distribution, with HRT as the response variable. **a,** Feature importance of A-HRT bioindicators. The ASV importance was calculated using the relative abundance data of bioreactor A as a training set and data of bioreactor B as a test set. **b,** Feature importance of B-HRT bioindicators. Similar to A-HRT bioindicators, ASV importance of B-HRT was calculated using the relative abundance data of bioreactor B as a training set and data of bioreactor A as a test set. The taxonomic classification of ASVs assigned at the genus level is provided in parentheses.

Figure 4. Prediction results of C6 and C8 productivities using HRT bioindicators. a,b, Prediction performance of C6 productivity. **c,d,** Prediction performance of C8 productivity. Results in **a** and **c** were obtained by using relative abundance data of bioreactor A for training the model and data of bioreactor B for testing. Results using the data of bioreactor B for training and bioreactor A for testing are shown in **b** and **d**. The red lines and grey shaded areas depict the best-fit trendline and the 95% confidence interval of the least-squares regression, respectively. C6, *n*-caproate; C8, *n*-caprylate; %Var., explains the variance (%) in C6/C8 productivity of the training set; RRMSE, relative root mean square error.

Figure 5. Phylogeny of HRT bioindicators and non-HRT bioindicators for considering community assembly caused by time. a,b, A maximum likelihood 16S rRNA gene tree showing the ASV species based on the rarefied sequencing data. ASVs are coloured according to the class (**a**, first inner ring) and family (**b**, second inner ring). **c**, The third inner ring shows the 11 HRT bioindicators identified in both reactors for the prediction of HRT phases of 8 days and 2 days. The ASVs identified as HRT bioindicators are shown in bold. Their taxonomic assignments at the genus level are provided in the legend. **d**, The four ASVs of HRT bioindicators irrespective of time are shown in red in the outer ring. The ASVs only present in non-HRT bioindicators of C6/C8 productivity are shown in pink in the outer ring. **e**, Relative abundance dynamics of HRT bioindicators during the whole reactor operation period. In the legend, A and B stand for bioreactors A and B, respectively. The four ASVs (in bold) of HRT bioindicators, irrespective of time, assigned at the genus level are indicated in parentheses. C6, *n*-caproate; C8, *n*-caprylate.

Figure 6. Genetic potential of metagenome-assembled genomes (MAGs) with the same taxonomy as HRT bioindicators driving the catabolism of xylan and lactate to *n*-caproate and *n*-caprylate. These catabolic steps were categorised into four main functions of the anaerobic mixed culture fermentation. **a,** Hydrolysis of xylan. **b,** Xylose fermentation producing acetate and lactate. **c,** Butyrate formation from lactate and acetate. **d,** Chain elongation with lactate as electron donor producing *n*-butyrate, *n*-caproate and *n*-caprylate. Numbers represent the 18 different MAGs with similar phylogenies as the HRT bioindicators at the genus level (details in Table 1). The enzyme abbreviations are provided in red letters next to the pathways (solid lines). Dashed lines represent multi-enzyme reactions between the two indicated molecules. In **(d)**, "cycle" refers to the reverse β-oxidation cycle. The complete metabolic pathways are depicted in Additional file 1: Figure S13. un., unclassified; XL, xylanase (EC 3.2.1.8); XylT, xylose transporter (EC 7.5.2.10, EC 7.5.2.13); LacP, lactate permease (TC 2.A.14); CoAT, butyryl-CoA:acetate CoA-transferase (EC 2.8.3.-); PTB, phosphate butyryltransferase (EC 2.3.1.19); BUK, butyrate kinase (EC 2.7.2.7); ACT, acyl-CoA thioesterase (EC 3.1.2.20).

Figure 7. Phylogenetic tree of the recovered metagenome-assembled genomes (MAGs). a,b, A phylogenomic tree based on mash distances showing the MAGs taxonomy determined by GTDB-Tk at phylum (**a**) and family (**b**) levels. A total of 108 MAGs were recovered and differentiated into 29 species based on the ANI values. We defined the representative MAG for each species as that showing high quality. Only the representative MAG for each species is depicted in the tree. The ENA accession numbers of the representative MAGs are shown in

parentheses. MAGs with similar phylogenies as HRT bioindicators are indicated by a star.

Figure 8. Overview on the quantitative prediction of process performance in the anaerobic bioreactor system. a, Anaerobic mixed culture fermentation of lactate and xylan for the production of *n*-caproate (C6) and *n*-caprylate (C8) by lactate-based chain elongation. Based on the recovery of metagenome-assembled genomes, the left panel shows the bioindicators capable of performing key steps of the fermentation. **b,** Reducing the hydraulic retention time (HRT) as an operation-based strategy to optimise the process performance and to manage the reactor microbiota towards desired functions. Shortening the HRT from 8 days to 2 days enhanced productivities of C4, C6 and C8. The enriched reactor microbiota comprised functional groups involved in xylan hydrolysis, xylose fermentation and chain elongation with lactate, presented by a co-occurrence network of environmental factors (controlled conditions with only reducing the HRT), ecosystem functioning (process performance) and microbial community. The full network is shown in Additional file: Figure S14. **c**, Predicting performance of ecosystem processes with random forest analysis. We developed a random forest two-step workflow to qualitatively predict the HRT phases and to quantitatively predict carboxylate production by using relative abundance data of the 16S rRNA-derived species (ASVs, Amplicon Sequence Variants).

783 **Table 1. Summary of metagenome-assembled genomes (MAGs) with the same taxonomy as HRT bioindicators.**

HRT bioindicators	Number of MAGs		Taxonomic classification						Representative MAG
	High quality	Medium quality	Phylum	Class	Order	Family	Genus	Species	
Olsenella sp. ASV034	2	3	*Actinobacteriota*	*Coriobacteriia*	*Coriobacteriales*	*Atopobiaceae*	*Olsenella*_B	*Olsenella*_B sp000752675	UMB00010
Olsenella sp. ASV057	4	2	*Actinobacteriota*	*Coriobacteriia*	*Coriobacteriales*	*Atopobiaceae*	*Olsenella*_C	unclassified	UMB00003
***Olsenella* sp. ASV058**	1	0	*Actinobacteriota*	*Coriobacteriia*	*Coriobacteriales*	*Atopobiaceae*	*Olsenella*	unclassified	UMB00074
unclassified *Erysipelotrichaceae* sp. ASV002	4	1	*Firmicutes*	*Bacilli*	*Erysipelotrichales*	*Erysipelotrichaceae*	*unclassified*	unclassified	UMB00059
Bulleidia sp. ASV010	0	1	*Firmicutes*	*Bacilli*	*Erysipelotrichales*	*Erysipelotrichaceae*	*Solobacterium*	unclassified	UMB00050
	6	0	*Firmicutes*	*Bacilli*	*Erysipelotrichales*	*Erysipelotrichaceae*	*Solobacterium*	*Solobacterium* sp900343155	UMB00007
	5	1	*Firmicutes*	*Bacilli*	*Erysipelotrichales*	*Erysipelotrichaceae*	*Solobacterium*	*Solobacterium* sp900290205	UMB00011
Lachnospiracea incertae sedis ASV053	3	0	*Firmicutes*_A	*Clostridia*	*Lachnospirales*	*Lachnospiraceae*	UBA4285	unclassified	UMB00063
***Syntrophococcus* sp. ASV060**	*Eubacterium cellulosolvens* 6		*Firmicutes*_A	*Clostridia*	*Lachnospirales*	*Lachnospiraceae*	*Eubacterium*_H	*Eubacterium*_H *cellulosolvens*	
	Eubacterium cellulosolvens LD2006		*Firmicutes*_A	*Clostridia*	*Lachnospirales*	*Lachnospiraceae*	*Eubacterium*_H	*Eubacterium*_H *cellulosolvens*_A	
	5	0	*Firmicutes*_A	*Clostridia*	*Lachnospirales*	*Lachnospiraceae*	*Eubacterium*_H	unclassified	UMB00012
	6	0	*Firmicutes*_A	*Clostridia*	*Lachnospirales*	*Lachnospiraceae*	*Eubacterium*_H	unclassified	UMB00020
***Clostridium* IV sp. ASV073**	*Caproiciproducens galactitolivorans* BS-1		*Firmicutes*_A	*Clostridia*	*Oscillospirales*	*Acutalibacteraceae*	MS4	unclassified	
	5	0	*Firmicutes*_A	*Clostridia*	*Oscillospirales*	*Acutalibacteraceae*	UBA1033	UBA1033 sp002399935	UMB00014
	1	0	*Firmicutes*_A	*Clostridia*	*Oscillospirales*	*Acutalibacteraceae*	UBA1033	UBA1033 sp002407675	UMB00060
	3	0	*Firmicutes*_A	*Clostridia*	*Oscillospirales*	*Acutalibacteraceae*	UBA1033	UBA1033 sp002409675	UMB00097
	Ruminococcaceae bacterium CPB6		*Firmicutes*_A	*Clostridia*	*Oscillospirales*	*Acutalibacteraceae*	UBA4871	UBA4871 sp002119605	

	6	0	*Firmicutes_A*	*Clostridia*	*Oscillospirales*	*Acutalibacteraceae*	UBA4871	UBA4871 sp002399445	UMB00016
Clostridium sensu stricto sp. ASV008	*Clostridium luticellarii* DSM29923		*Firmicutes_A*	*Clostridia*	*Clostridiales*	*Clostridiaceae*	*Clostridium_B*	*Clostridium_B luticellarii*	
	3	0	*Firmicutes_A*	*Clostridia*	*Clostridiales*	*Clostridiaceae*	*Clostridium_B*	*Clostridium_B* sp003497125	UMB00080
***Lactobacillus* sp. ASV074**	6	0	*Firmicutes*	*Bacilli*	*Lactobacillales*	*Lactobacillaceae*	*Lactobacillus_H*	*Lactobacillus_H mucosae*	UMB00017
	0	1	*Firmicutes*	*Bacilli*	*Lactobacillales*	*Lactobacillaceae*	*Lactobacillus*	unclassified	UMB00041
	2	2	*Firmicutes*	*Bacilli*	*Lactobacillales*	*Lactobacillaceae*	*Lactobacillus*	*Lactobacillus amylovorus*	UMB00015
unclassified *Coriobacteriaceae* sp. ASV082	0	0							

Taxonomy refers to the GTDB (Genome Taxonomy Database) phylogenomic classification. ASVs in bold represent the four HRT bioindicators irrespective of time. Sequence datasets of genomes in red letters were taken from the databases of NCBI and EzBioCloud. These genomes (in red) were used to affiliate the MAGs of *Syntrophococcus*, *Clostridium* IV and *Clostridium sensu stricto*, since their genomes are not available in GTDB. See details of MAGs in Additional file 2: Dataset S1. ASV: amplicon sequencing variant.

Figures

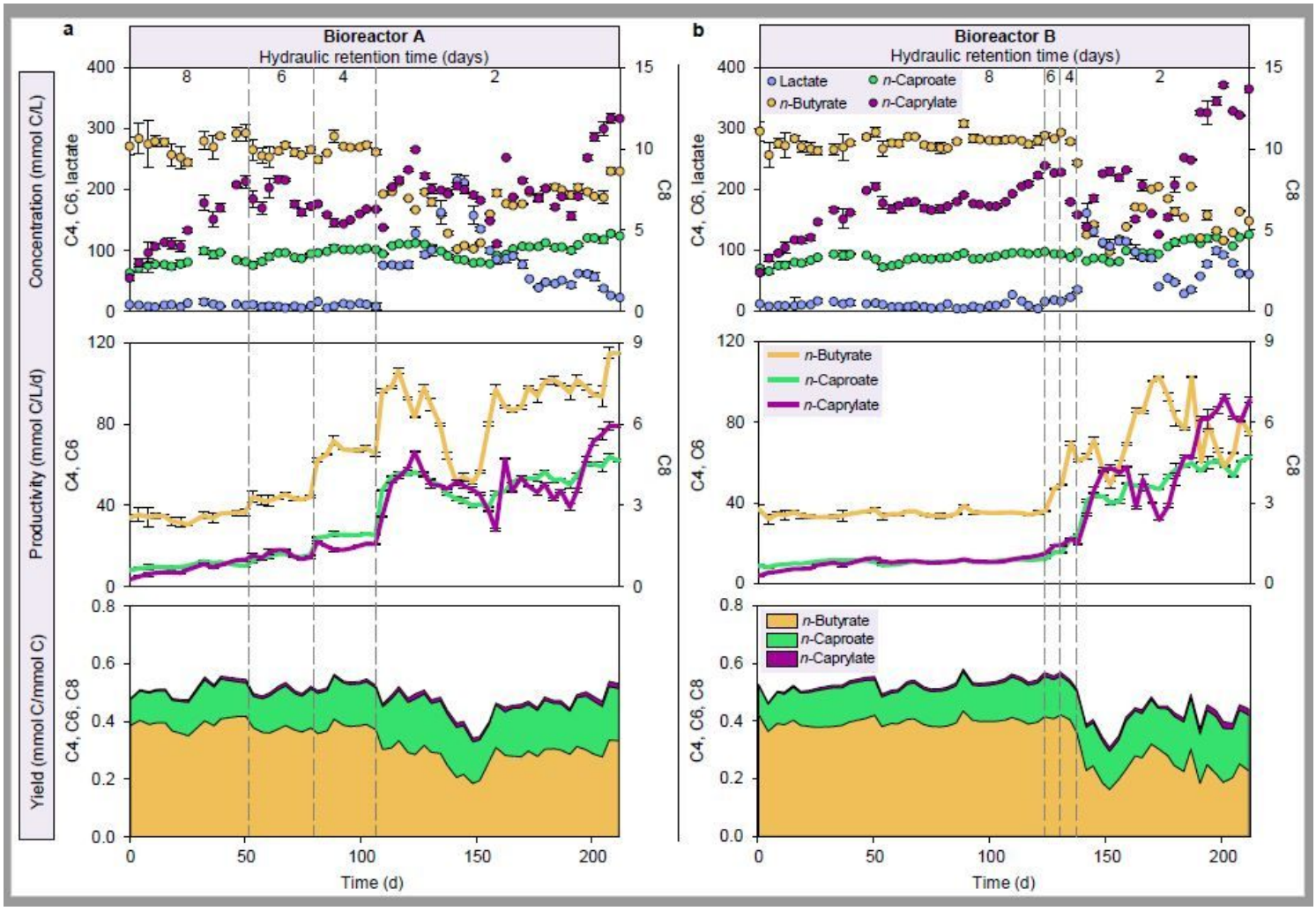

Figure 1

Performance of bioreactors. Concentrations of chain elongation products and lactate, as well as productivities and yields of chain elongation products in bioreactors A (a) and B (b) during the four HRT phases. Chain elongation products: C4, n-butyrate; C6, n-caproate; C8, n-caprylate.

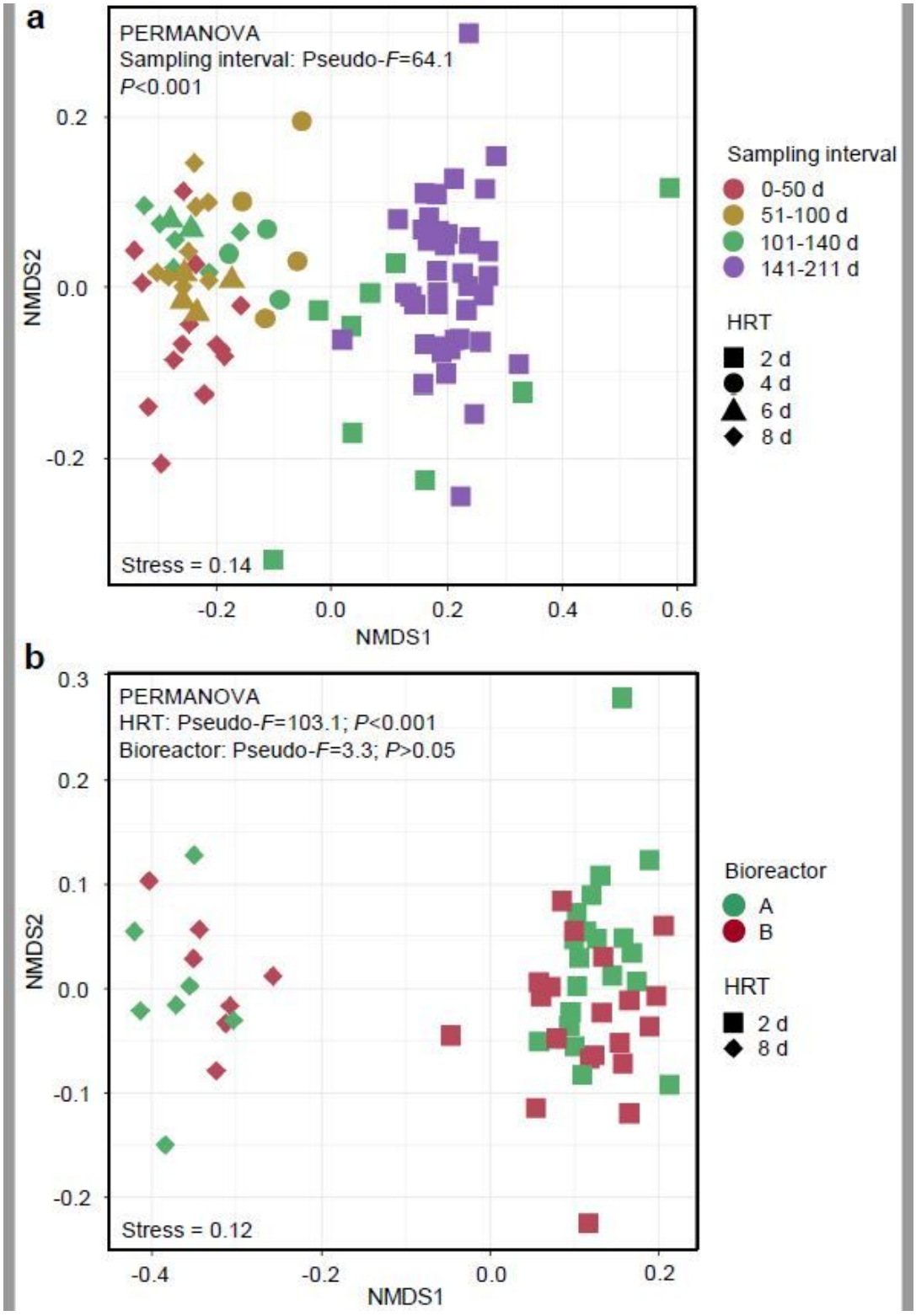

Figure 2

Dissimilarities in bacterial community composition (beta-diversity). Non-metric multidimensional scaling (NMDS) based on Bray-Curtis dissimilarities of microbial community composition in bioreactors. a, All samples in the four HRT phases were considered for dissimilarity calculation. b, Samples in the 8-day HRT phase classified to the sampling interval 0-50 days and in the 2-day HRT phase classified to the interval 141-211 days were included.

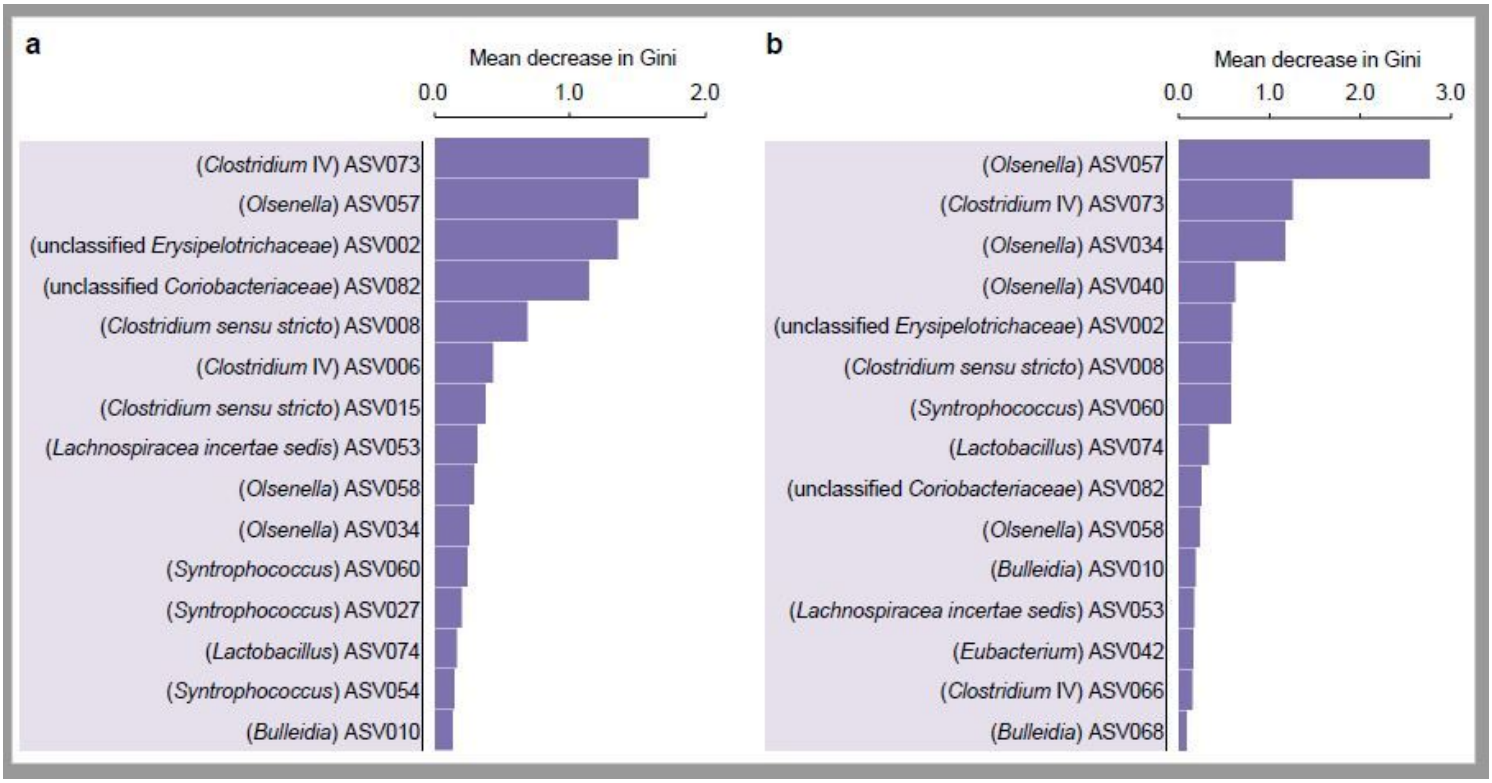

Figure 3

Random forest feature importance of ASVs used to classify the HRT phases (A-HRT bioindicators and B-HRT bioindicators). The top-ranked 15 ASVs reducing the uncertainty in the prediction of HRT phases (HRT of 8 days and 2 days). The order of features (from top to bottom) was based on their mean decrease in Gini scores, according to their ASV abundances distribution, with HRT as the response variable. a, Feature importance of A-HRT bioindicators. The ASV importance was calculated using the relative abundance data of bioreactor A as a training set and data of bioreactor B as a test set. b, Feature importance of B-HRT bioindicators. Similar to A-HRT bioindicators, ASV importance of B-HRT was calculated using the relative abundance data of bioreactor B as a training set and data of bioreactor A as a test set. The taxonomic classification of ASVs assigned at the genus level is provided in parentheses.

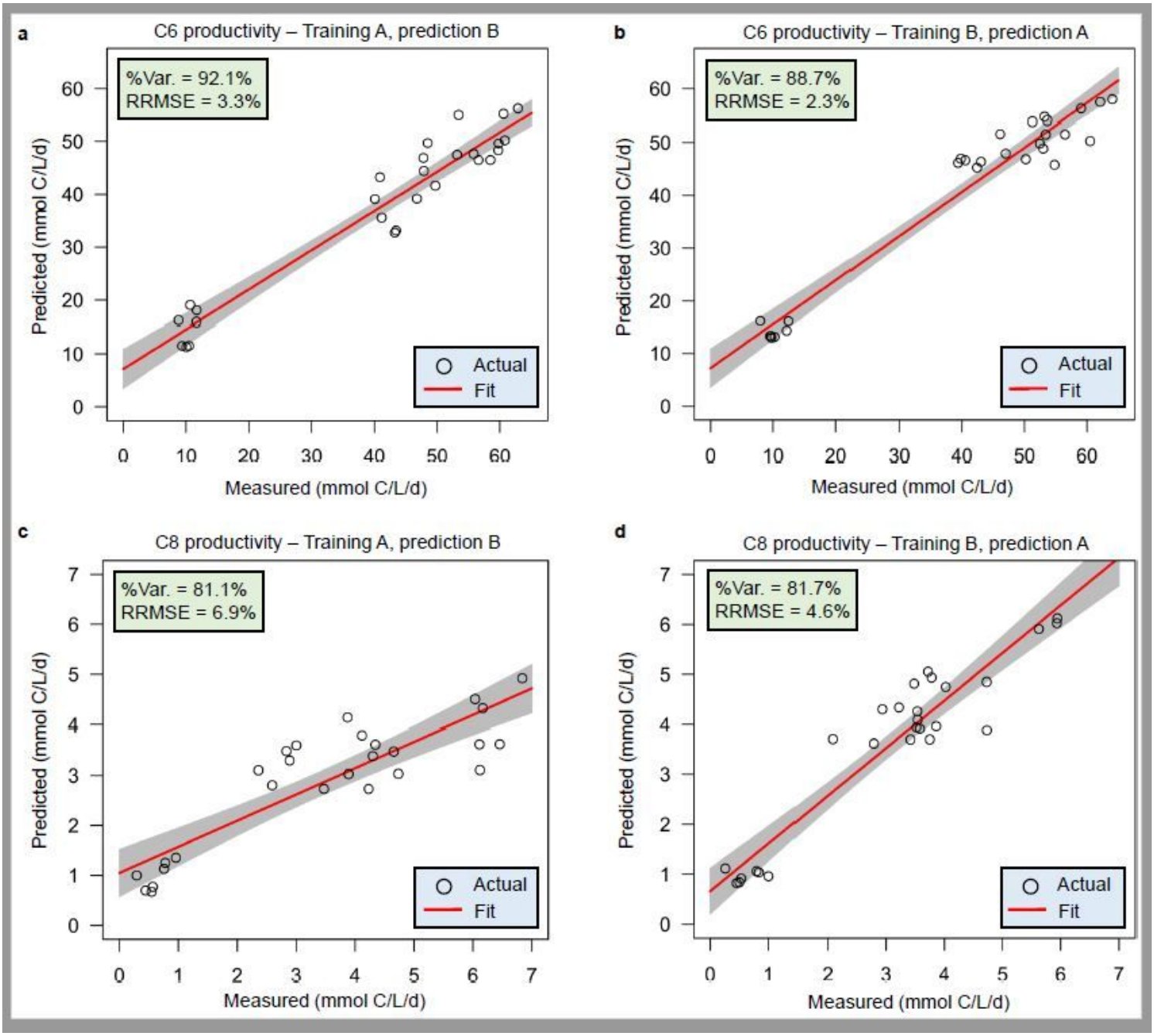

Figure 4

Prediction results of C6 and C8 productivities using HRT bioindicators. a,b, Prediction performance of C6 productivity. c,d, Prediction performance of C8 productivity. Results in a and c were obtained by using relative abundance data of bioreactor A for training the model and data of bioreactor B for testing. Results using the data of bioreactor B for training and bioreactor A for testing are shown in b and d. The red lines and grey shaded areas depict the best-fit trendline and the 95% confidence interval of the least-squares regression, respectively. C6, n-caproate; C8, n-caprylate; %Var., explains the variance (%) in C6/C8 productivity of the training set; RRMSE, relative root mean square error.

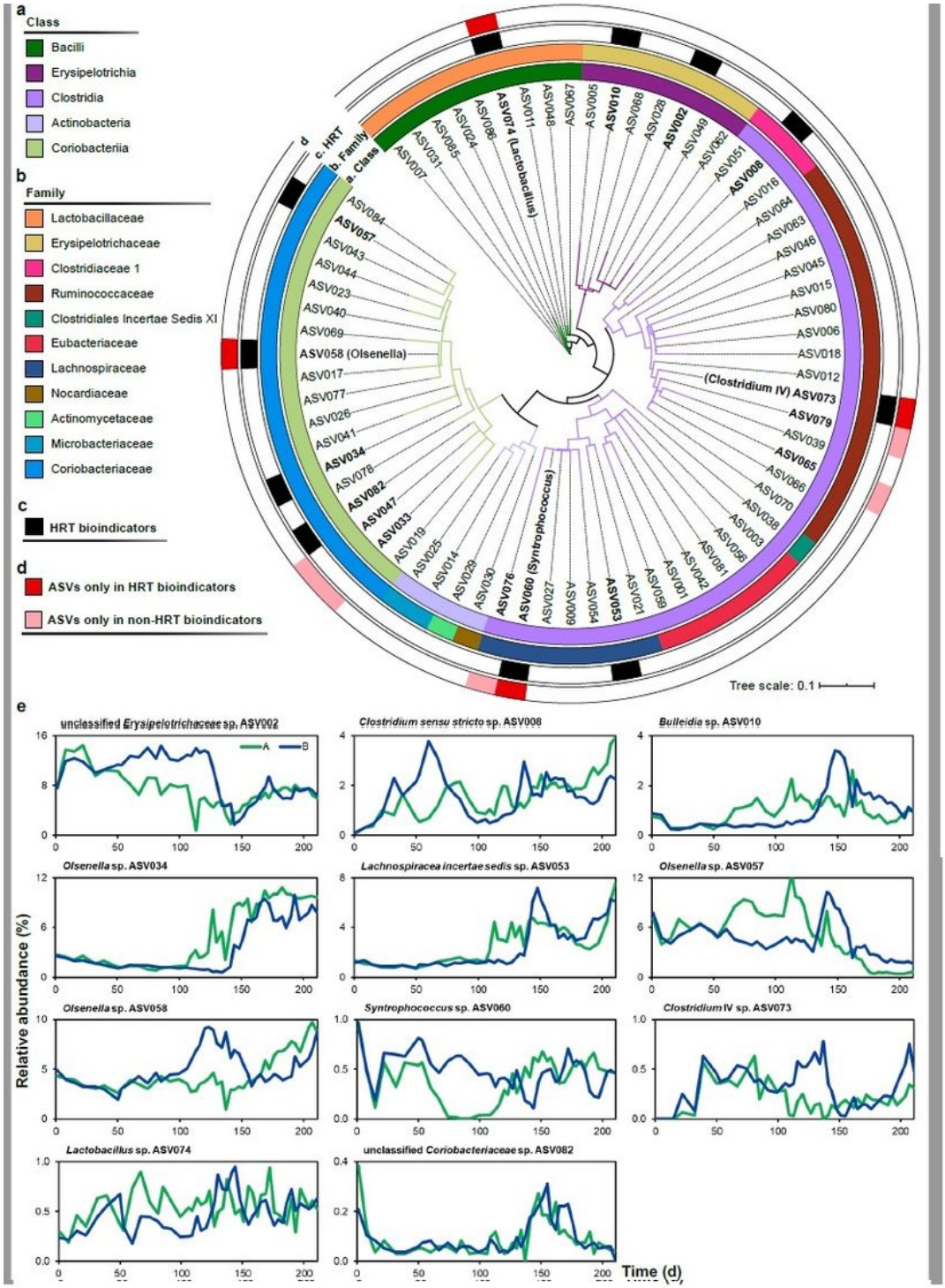

Figure 5

Phylogeny of HRT bioindicators and non-HRT bioindicators for considering community assembly caused by time. a,b, A maximum likelihood 16S rRNA gene tree showing the ASV species based on the rarefied sequencing data. ASVs are coloured according to the class (a, first inner ring) and family (b, second inner ring). c, The third inner ring shows the 11 HRT bioindicators identified in both reactors for the prediction of HRT phases of 8 days and 2 days. The ASVs identified as HRT bioindicators are shown in bold. Their

taxonomic assignments at the genus level are provided in the legend. d, The four ASVs of HRT bioindicators irrespective of time are shown in red in the outer ring. The ASVs only present in non-HRT bioindicators of C6/C8 productivity are shown in pink in the outer ring. e, Relative abundance dynamics of HRT bioindicators during the whole reactor operation period. In the legend, A and B stand for bioreactors A and B, respectively. The four ASVs (in bold) of HRT bioindicators, irrespective of time, assigned at the genus level are indicated in parentheses. C6, n-caproate; C8, n-caprylate.

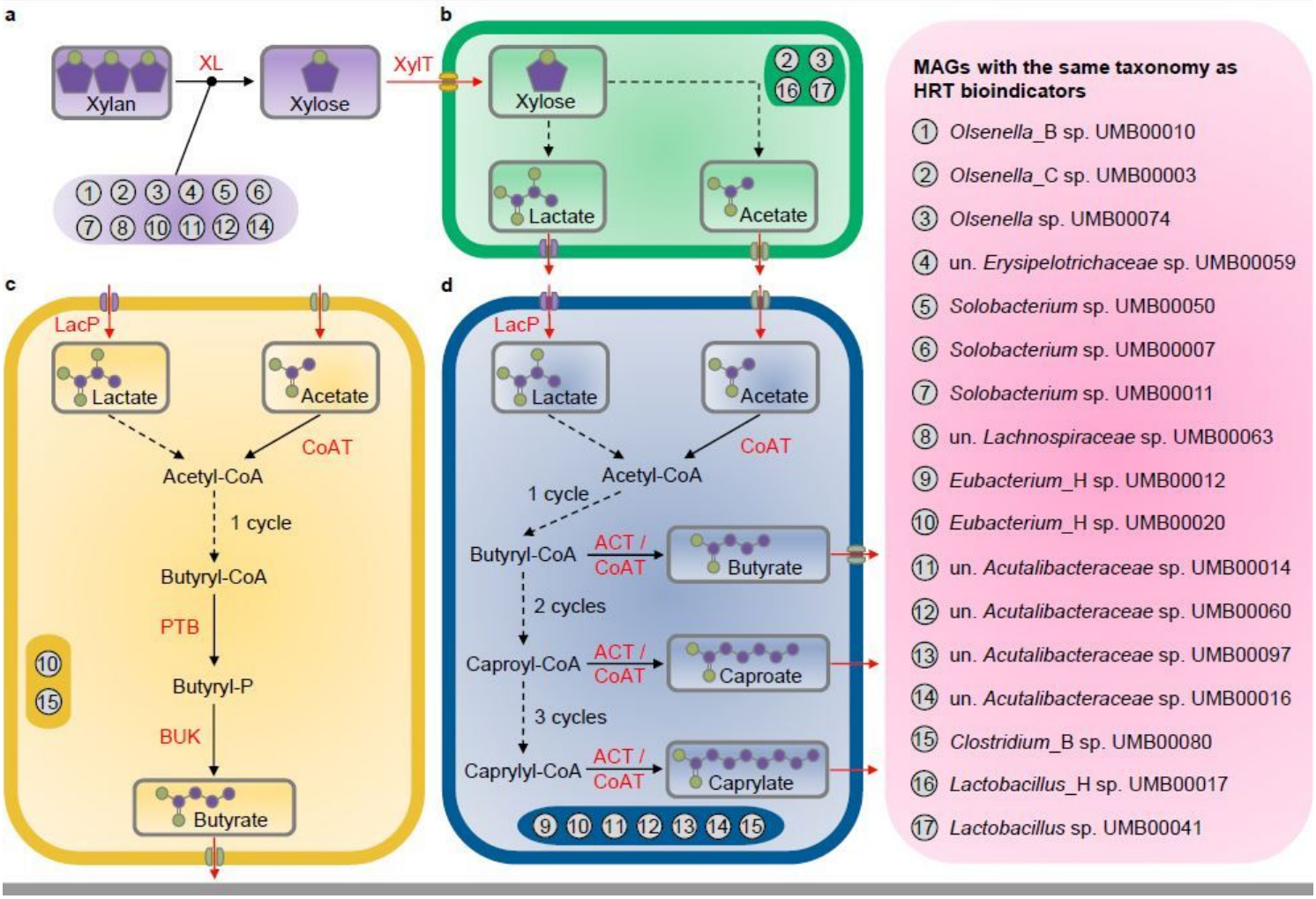

Figure 6

Genetic potential of metagenome-assembled genomes (MAGs) with the same taxonomy as HRT bioindicators driving the catabolism of xylan and lactate to n-caproate and n-caprylate. These catabolic steps were categorised into four main functions of the anaerobic mixed culture fermentation. a, Hydrolysis of xylan. b, Xylose fermentation producing acetate and lactate. c, Butyrate formation from lactate and acetate. d, Chain elongation with lactate as electron donor producing n-butyrate, n-caproate and n-caprylate. Numbers represent the 18 different MAGs with similar phylogenies as the HRT bioindicators at the genus level (details in Table 1). The enzyme abbreviations are provided in red letters next to the pathways (solid lines). Dashed lines represent multi-enzyme reactions between the two indicated molecules. In (d), “cycle” refers to the reverse β-oxidation cycle. The complete metabolic pathways are depicted in Additional file 1: Figure S13. un., unclassified; XL, xylanase (EC 3.2.1.8); XylT,

xylose transporter (EC 7.5.2.10, EC 7.5.2.13); LacP, lactate permease (TC 2.A.14); CoAT, butyryl-CoA:acetate CoA-transferase (EC 2.8.3.-); PTB, phosphate 750 butyryltransferase (EC 2.3.1.19); BUK, butyrate kinase (EC 2.7.2.7); ACT, acyl-CoA thioesterase (EC 3.1.2.20).

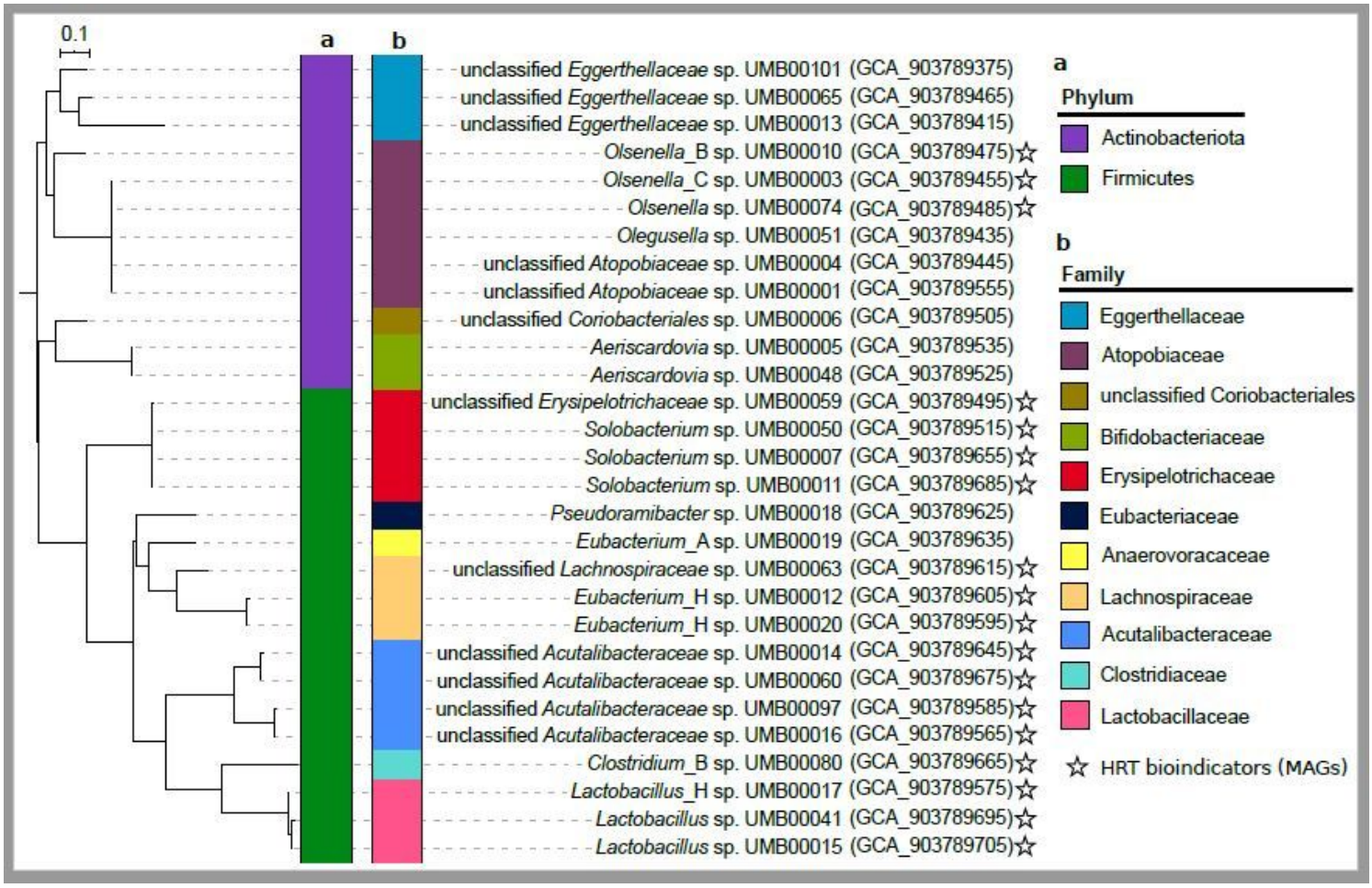

Figure 7

Phylogenetic tree of the recovered metagenome-assembled genomes (MAGs). a,b, A phylogenomic tree based on mash distances showing the MAGs taxonomy determined by GTDB-Tk at phylum (a) and family (b) levels. A total of 108 MAGs were recovered and differentiated into 29 species based on the ANI values. We defined the representative MAG for each species as that showing high quality. Only the representative MAG for each species is depicted in the tree. The ENA accession numbers of the representative MAGs are shown in 34 parentheses. MAGs with similar phylogenies as HRT bioindicators are indicated by a star.

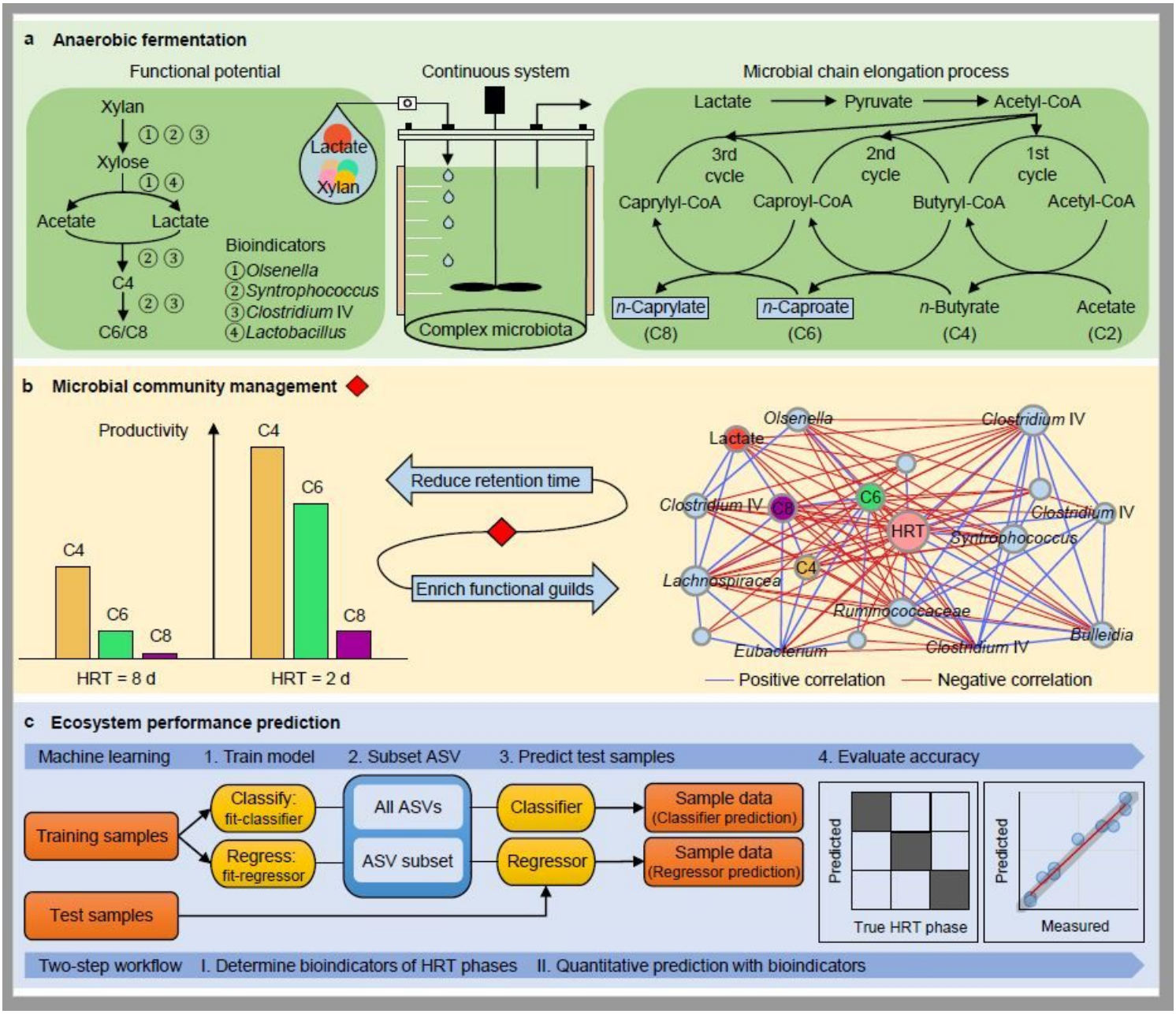

Figure 8

Overview on the quantitative prediction of process performance in the anaerobic bioreactor system. a, Anaerobic mixed culture fermentation of lactate and xylan for the production of n-caproate (C6) and n-caprylate (C8) by lactate-based chain elongation. Based on the recovery of metagenome-assembled genomes, the left panel shows the bioindicators capable of performing key steps of the fermentation. b, Reducing the hydraulic retention time (HRT) as an operation-based strategy to optimise the process performance and to manage the reactor microbiota towards desired functions. Shortening the HRT from 8 days to 2 days enhanced productivities of C4, C6 and C8. The enriched reactor microbiota comprised functional groups involved in xylan hydrolysis, xylose fermentation and chain elongation with lactate, presented by a co-occurrence network of environmental factors (controlled conditions with only reducing the HRT), ecosystem functioning (process performance) and microbial community. The full network is

shown in Additional file: Figure S14. c, Predicting performance of ecosystem processes with random forest analysis. We developed a random forest two-step workflow to qualitatively predict the HRT phases and to quantitatively predict carboxylate production by using relative abundance data of the 16S rRNA-derived species (ASVs, Amplicon Sequence Variants).

Supplementary Files

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

- Additionalfile1.FigureS1S16andTableS1S6.pdf
- Additionalfile2.MAGstaxonomyandgenomemetrics.xlsx
- Additionalfile3.xlsx
- Additionalfile4.xlsx
- Additionalfile5.xlsx
- Additionalfile6.xlsx