Robust host source tracking building on the divergent and non-stochastic assembly of gut microbiome in wild and farmed large yellow croaker

Jun Zhu
Xiamen University School of Life Sciences

Hao Li
Xiamen University School of Life Sciences

Ze-Zhou Jing
Xiamen University College of Ocean and Earth Science

Wei Zheng
Xiamen University School of Life Sciences

Yuan-Rong Luo
Xiamen University

Shi-Xi Chen
Xiamen University College of Ocean and Earth Science

Feng Guo (fguo.bio@xmu.edu.cn)
Xiamen University https://orcid.org/0000-0001-7785-8388

Research

Keywords: non-stochastic pattern, diet, size, life-stage, age, and geographic origin, DNA

DOI: https://doi.org/10.21203/rs.3.rs-721308/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Robust host source tracking building on the divergent and non-stochastic assembly of gut microbiome in wild and farmed large yellow croaker

Jun Zhu¹, #, Hao Li¹, #, Ze-Zhou Jing², #, Wei Zheng¹, Yuan-Rong Luo³, Shi-Xi Chen²*, Feng Guo¹, 4, *

¹. School of Life Sciences, Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Xiamen University, Xiamen, China
². State-Provence Joint Engineering Laboratory of Marine Bioprodcts and Technology, College of Ocean and Earth Sciences, Xiamen University, Xiamen, China
³. Key Laboratory of the Ministry of Education for Coastal and Wetland Ecosystems, Xiamen University, Xiamen, China
⁴. Key Laboratory of Fujian Provinical University for Microorganism Resource, Xiamen, China

#These authors contribute equally
*Corresponding author.

Email: fguobio@xmu.edu.cn; chenshixi@xmu.edu.cn
Abstract

Background: Revealing the potential divergence of gut microbiome between farmed and wild fishes, and its underlying mechanism are informative to improve its mariculture, as well as establish the molecular marker of host source tracking, which is an alternative to the yet-to-be-established host genetic marker. A candidate for testing the feasibility is the large yellow croaker, *Larimichthys crocea*, which is carnivorous and ranking the top maricultural fish in China with depleted wild resource and frequently farmed individuals escaping and fry releasing for wild stock enhancement.

Results: The rectums of wild (*n*=212) and farmed (*n*=79) individuals from multiple batches were collected for the profiling of gut bacterial communities. The farmed individuals had a higher alpha diversity and lower bacterial loading than the wild individuals. The gut microbiota of the two sources exhibited divergence and high inter-batch variation, featured by the dominance of *Psychrobacter* spp. in the wild group. Predicted function of gut microbiome and representative isolates suggested that diet could be a key factor for the divergence, which was linked to the high ratio and diverse source of carbohydrate in formulated feed and low pH of rectum contents in farmed fishes. The non-stochastic distribution patterns of the core gut microbiota of the wild and farmed individuals indicated the feasibility of microbiota-based host source tracking through machine learning algorithm. Random forest classifier building on the divergence and non-stochastic assembly of gut microbiome was robust in host source tracking for individuals from all batches including a newly introduced batch.

Conclusions: Our study revealed the divergence of the gut microbiota between wild and farmed croakers and suggested that diet change is an underlying key factor for the
divergence. As the first time, we verified that with less biased datasets and non-stochastic pattern, gut microbiota can be robustly applied to the tracking of host source even in carnivorous fish.
Introduction

Microbes living in animal gastrointestinal tracts play important roles in the nutrition and health of their hosts through extensive metabolic and immune interactions [1-4]. With the development of the next generation sequencing techniques, diversity and function of the gastrointestinal microbiomes in many fish species have been unprecedentedly explored in the last decade [1, 5, 6]. Fish usually has much more dynamic and less diverse gut microbiomes than terrestrial vertebrates [7]. A great number of factors, such as source (i.e., domesticated or wild), diet, size, life-stage or age, and geographic origin, can have strong impacts on the microbiomes in the fish gut [5, 8-10]. Moreover, herbivorous and omnivorous fish usually possesses a higher selectivity in the intestine than carnivorous fish [11, 12], while host selectivity in certain microbial taxa has been reported in carnivorous Atlantic salmon [13].

Among the gut microbiota, core microbial taxa, which can be detected in most or all host individuals, has been proposed to elucidate the patterns of physiological interactions and evolutionary relationships between microbes and hosts [6, 12, 14, 15]. In most cases, the core taxa were proposed based on limited sampling batches despite the high dynamics of fish gut microbiota. From an ecological point of view, the assembly of microbiome are underpinned by two major categories of factors: deterministic and neutral processes [16-18]. Reasonably, core taxa are more likely to be determined by deterministic factors than by stochastic processes. Validation of deterministic or neutral process is a basis for the further examination on the specific host–microbe relationships or potential applications.

Understanding the divergence of gut microbiota between conspecific wild and farmed
fishes has been considered in improving diet efficiency, farming mode and developing probiotics [19-23]. Moreover, the divergence of gut microbiota between wild and farmed fishes may be informative for host source tracking. The continued increase in production of mariculture fish species and interaction between farmed and wild fishes [24, 25], coupled with divergent selling price [26, 27], has accentuated the need of host source tracking. The application scenarios include, but are not limited to determine the source of unknown individuals and discriminating fishes escaping from farmed cages and seedling release from true wild individuals. For fish species with sufficient historical specimens and genetic background, such as salmon, the genetic marker from host showed well performance in source tracking [28]. However, for most mariculture fish species with poor genetic marker for identification of wild and farmed populations, gut microbiome, as the secondary genome of corresponding host, can serve as an alternative biomarker of host source.

The large yellow croaker *Larimichthys crocea* (referred to as “croaker” hereafter) is an economically important marine carnivorous fish species in China [29]. Long-term overfishing since 1950s has resulted in the severe depletion (>95%) of wild stocks. Currently, the majority of sales is from mariculture, exceeding 220,000 tons in 2019; ranking the top maricultural fish in China [30]. Wild stock enhancement through the release of tens of millions of fries has been performed annually for over a decade. However, the yield from the wild stock has shown little increase possibly because of fishing pressure, human interference on habitats, niche occupation by mariculture, loss of genetic diversity, and poor adaptation for released juveniles [31]. Moreover, wild resources can be overestimated given that croakers captured in coastal regions, which overlap with or are adjacent to the mariculture region of the species, may be directly
derived from escaped farmed individuals. Owing to the insufficient accumulation of “true” wild individuals, host genetic biomarkers for distinguishing wild croakers from domesticated ones has not been well established [32].

To the best of our knowledge, the discrepancy of the gut microbiome between wild and farmed croakers has not been characterized. As a carnivorous species with multiple geographic populations [33], it may have high inter-batch variations in the gut microbiota. Therefore, this species seems a suitable candidate to test the feasibility of microbiota-based host source tracking. In this study, we profiled bacterial community assembly and core taxa inhabiting the rectums of farmed and wild croakers sampled from various geographical populations and batches. By evaluating the fitness of the neutral model and establishing a classifier using random forest model, a machine learning approach that is suitable for unbalance distribution data with noise feature and less prone to overfitting [34, 35], the feasibility of using the divergence of the microbiota between captive and wild individuals in source tracking can be verified. Moreover, farmed croakers have been exhibiting reduced genetic diversity and disease resistance for decades [32]. Understanding the gut microbiome divergence between wild and farmed individuals may provide new insights into the sustainable development of croaker mariculture.

**Materials and Methods**

**Sample collection and preparation**

Ten batches (designated as A, B, C, F, H, N, S, T, W, and X; $n=291$) of croakers from diverse locations and sizes were collected from wild catching (C, F, S, and W; $n=212$)
and raft farming sites (A, B, H, N, T, and X; \(n=79\)). Information on the sampling sites and basic information on the batches are shown in Figure S1. Batch S, a unique wild batch that was collected in a bay with a high-density mariculture of croakers, was only used in testing the host source tracking classifier (see below). Notably, obtaining large wild individuals (>300 g) is difficult because their natural stocks were depleted since 1980s. Farmed fishes were mainly fed with artificial formulated feed and occasionally with fresh fish meals during the sampling period. All individuals were frozen at \(-20^\circ\text{C}\) immediately after they were removed from seawater. Approximately 1–2 cm of the rectum was aseptically removed from the abdominal cavity with sterile scissors and tweezers (see Figure S2 for the details of the digestive organs and the typical sampled section). Then, the tissue samples were transferred to centrifuge tubes and stored at \(-20^\circ\text{C}\) until DNA extraction.

**DNA extraction**

DNA extraction was performed with a commercial kit (QIAamp PowerFecal DNA Kit, QIAGEN, Germany). Before extraction, the tissues containing rectum contents were aseptically homogenized with a tissue homogenizer after the addition of 200 μL of Solution CD1 (a buffer of the kit). The rectum and the content were processed together for DNA extraction. Then, using MilliQ water as extraction blank, DNA extraction was performed according to the kit’s instructions. To minimize DNA contamination from the extraction buffers, we used freshly prepared MilliQ water to elute the DNA in the final step. After extraction, the quantity and quality of the yielded DNA were examined with a micro-spectrophotometer (NanoDrop ND-1000, Thermo Scientific, US). The \(\text{OD}_{260\text{nm}}/\text{OD}_{280\text{nm}}\) ratios ranged from 1.7 to 2.0 for all samples.
PCR and high-throughput sequencing

PCR targeting on the V4 region of the bacterial 16S rRNA gene was conducted following the previously described method [36], except that adaptor sequences were added during library construction rather than PCR. To minimize potential cross-talking contamination, as suggested by a previous study [37], we applied unique barcodes linking to forward and reverse primers during multiplexing (i.e., no barcode was share by any sample in a library). The number of PCR cycles was set as 30, under which the DNA extraction blank and PCR blank (using MilliQ water as the template) produced no visible band during electrophoresis. Purified PCR products were pooled together at equal mass before sequencing library construction (TruSeq DNA PCR-Free Library Preparation Kit, Illumina, USA). High-throughput sequencing was performed in the Illumina Hiseq2500 sequencing platform with PE250 strategy (commercial service provided by Novogen, China).

Quantitative PCR for determining bacterial loading

For the determination of the bacterial loading in the rectum, randomly picked 20 samples from both wild (batch C and F) and farmed samples (batch H and T) were analyzed with quantitative PCR (qPCR). For each sample, 5 ng of DNA template was added to 25 μL of PCR solution (final volume, SYBR GreenER™ qPCR SuperMix Universal, ThermoFisher Scientific, USA). A standard curve (R²>0.99) generated through the 10-fold dilutions of a plasmid DNA containing a full-length 16S rRNA gene from *Escherichia coli* was used in absolute quantification. The V3 region was amplified by using the primer set of 341F and 534R [38]. qPCR was performed in triplicate for each sample. To calculate bacterial loading per unit of host tissue, by referring to the standard curve, we quantified 16S rRNA gene copy number per ng of DNA given that
most of the extracted DNA was derived from the host tissue (indicated by the low copy number of bacterial 16S rRNA gene per ng DNA).

Analysis of 16S rRNA gene high-throughput sequencing data

Raw high-throughput sequencing data were cleaned using TRIMMOMATIC [39]. USEARCH v10 was used in removing suspicious sequences (i.e., chimeras and rare sequences with frequency less than 8 across all samples) and determining 0.97-level operational taxonomic units (OTUs) with the UNOISE algorithm and UPARSE, respectively [40, 41]. The table of OTU abundance generated in USEARCH platform was then introduced into the Mothur v1.39.5 for alpha-diversity and beta-diversity analyses and taxonomic classification [36]. Data normalization was performed by subsampling 10,000 valid reads for each sample [36]. For the beta-diversity analysis, weighted Unifrac distance was calculated during running the Mothur Miseq SOP [42]. We used the EzBioCloud 16S database as the taxonomic reference [43]. The effect of four factors (body weight, season, source, and batch) on the bacterial community was estimated using partial canonical ordination analysis [44]. Analysis of molecular variance (AMOVA) was conducted to show the significance of inter-group difference among community structures in the Mothur. Heatmap with sample and OTU-level clustering was realized in R using the pheatmap and vegan packages [45, 46]. Due to the high inter-batch variation, the core microbiota at OTU-level was defined as the taxa detected in >70% individuals for wild or farmed samples.

To determine the importance of the stochastic process in the assembly of gut microbiome community, the Sloan neutral model was tested using the R code [16, 47]. Additionally, the relative importance of stochastic and deterministic processes in
community assembly, the nearest taxon index (NTI) and beta nearest taxon index (βNTI) were also calculated Picante and MicEco R package (for OTUs abundance >0.01% and abundance.weighted =True) [48]. The 16S rRNA gene-based MetaCyc pathway profiling was inferred using PICRUST2 and the identification of differently abundance MetaCyc pathways in farmed and wild croakers were using ALDEx2 R package [49, 50]

Isolation of typical gut bacteria and genome analysis

Three farmed (batch T) and three wild (batch C) individuals were used in gut bacterial isolation. Microorganisms in freshly prepared rectums (~0.5 g) were rigorously washed off before serial 10-fold dilution in sterile 0.9% NaCl. Then, the dilutions were spread on 2216E agar plates (Hope Bio-Technology Co., Ltd., Qingdao, China) and cultivated for 48 h at 20°C. Colonies with different morphology and colors were selected, and their taxonomy was determined through full-length 16S rRNA gene sequencing. Only isolates affiliated with Photobacterium (n=7) and Psychrobacter (n=7), which were the representative taxa for farmed and wild croakers, respectively, were kept for downstream analysis.

Genomic DNA of isolates was extracted and sequenced using Illumina HiSeq X Ten platform. Assembly was performed with SPAdes v3.9.0 (parameters: -t 50, -k 55, 77, 99, -careful) [51]. Only scaffolds of >1000 bp were used in predicting open reading frames with Prodigal v2.6.3 [52]. Species-level identification was based on genome-to-genome average nucleotide identity values against the genomes from type strains [53]. Carbohydrate-active enzyme (CAZyme) families were annotated by dbCAN2 v2.0.11 under the default parameters and the signal peptides were predicted using SignalP v4.0 [54, 55]. The optimal pH for the CAZymes was predicted using the AcalPred online
server [56]. The capability of organic acid production of *Photobacterium* and *Psychobacter* isolates were annotated using DRAM [57].

**Assays of pH measurement for rectum content and bacterial biofilm formation**

To measure the pH in the rectums of the croakers, each freshly prepared rectum (~0.5 g; containing content) from wild (*n* = 27) and farmed (*n* = 15) individuals were transferred into a 15 mL centrifuge tube and gently and thoroughly washed in 5 mL ddH₂O. Then, the pH values of the suspensions were measured with a pH measurer.

To test the capability of biofilm formation for the isolates, we inoculated each of the isolates into replicate wells (*n* = 6) of a 96-well plate containing 200 μL of 2216E broth (approximately 10⁶ cells per well). After growth for 48 h, the OD₆₀₀ of the cell suspension was measured with a microplate reader (CLARIO star® Plus, BMG LABTECH Inc, USA). Then, the biofilm was stained with 0.1% crystal violet, and the OD₅₅₀ of the ethanol elution was measured [58]. The OD₅₅₀/OD₆₀₀ ratio was referred to as the capability of biofilm formation. Images of the stained biofilms were recorded under an inverted light microscope.

**Random-forest classification for wild and captive individuals**

To distinguish wild and farmed individuals using a machine learning approach, we used the random-forest algorithm in constructing a classifier. The dataset (nine batches except for batch S, *n* = 276) was pre-processed by removing rare OTUs (<20% frequency). The samples were split into two partitions as training and testing dataset under different proportions with 10 iterations. We built the classifier with the *random forest* R package with 5001 trees and default *mtry* number on the training samples and then validated it on the test samples [59]. The receiver operating characteristic curve
was implemented in \textit{pROC} R packages [60]. The top 15 OTUs ranked in terms of mean
decrease in accuracy were used in rebuilding an optimized classifier, and the accuracy
of the model was accessed by leave-one-out cross-validation using the \textit{caret} R package
[61]. We also tested the reliability of the classifier by using new wild samples (batch S,
\( n = 15 \)) that were caught in a bay where the farmed individuals of A, B, H, T, and X were
collected (Figure S1).

**Results**

**Divergence of bacterial alpha diversity, abundance, and high-rank taxa in the**
**rectums of wild and farmed croakers**

Under the sampling depth of 10,000 sequences, Shannon index for the bacterial
community in the rectums of croakers ranged from 1.17 to 4.96, with a median value
of 3.80. Intriguingly, wild individuals had lower diversity than the farmed ones (Figure
1A, \( P = 4.06 \times 10^{-14} \), Wilcoxon test). As the copy number of 16S rRNA gene per ng DNA for
a putative bacterium with 4M genome and four copies of 16S rRNA gene was \( 9.2 \times 10^5 \),
the determined 16S rRNA gene copy number (only 9–17,409 copies per ng DNA,
Figure 1B) indicated a low bacterial loading in the rectums of wild and farmed
individuals. However, the wild individuals contained more gut bacterial inhabitants
than the farmed ones (Figure 1B, \( P = 0.03 \), Wilcoxon test). The major detected bacterial
phyla (or classes of \textit{Proteobacteria}) were \textit{Gammaproteobacteria}, \textit{Firmicutes},
\textit{Fusobacteria}, \textit{Alphaproteobacteria}, \textit{Betaproteobacteria}, \textit{Actinobacteria}, \textit{Bacteroidetes},
and \textit{Deltaproteobacteria}. Significant difference (\( P < 0.05 \), Wilcoxon test, FDR-corrected
\( P \) value) between wild and captive individuals was observed for the relative abundance
rates of nearly all above high-rank taxa, which are apparently related to the domination
of \textit{Gammaproteobacteria} in most wild fishes (Figure 1C).
Beta-diversity indicated batch- and source-associated variation of gut microbiota

Nonmetric multidimensional scaling (NMDS) showed that wild and farmed groups were separated from each other to a great extent with a few exceptions (Figure 2A). The AMOVA indicated significant divergence between the two sources, as well as among most batches (Figure 2A). The beta-diversity divergence followed the order: between sources > between batches > within batches (Figure S3). Divergences among the farmed batches were higher than that among the wild batches (Figure S3). Batch, source and sampling season can explain 0.397, 0.148, and 0.093 of microbiota variation (all $P<0.01$, Analysis of Variance), whereas body weight can barely explain the variation (0.020, $P>0.05$). In fact, all above explainable variation is explained by batch since source and season has no independent contribution (Figure 2B, individuals in one batch had consistent source and season property). High inter-batch divergence suggested the unpredictable overall microbiota variation of newly introduced batch, which may have a negative impact on reliable host source tracking practice.

As shown in the heatmap (Figure 2C), 24 major gut bacterial OTUs that were present in over 20% samples with over 0.2% mean relative abundance were used in the analysis of the divergence of microbiota. These OTUs accounted for 1.4%–99.9% (83.4% in median value) of the total bacterial community in all the samples. Similar to the NMDS results, although most individuals from each source tended to cluster together, some samples in wild batch C ($n=7$), F ($n=1$), and all individuals of batch W ($n=18$) were clustered with the most (>95%) farmed individuals. After clustering based on Pearson correlation, the 24 OTUs fell into two clusters, namely, Cluster I (9 OTUs, which were enriched in the wild individuals) and Cluster II (15 OTUs, which were generally
enriched in the farmed ones). In Cluster I, five *Psychrobacter* OTUs constituted 93.0% (median value) of total bacteria in the wild samples, whereas OTUs affiliated with *Photobacterium, Vibrio, Streptococcus, Fusobacterium*, and *Clostridium*, are representative taxa in the farmed individuals.

The results of NMDS and clustering showed that the overall profiles of bacterial community of some wild samples was close to those of major farmed fishes. Thus, we further examined the relative abundance of the OTUs of Clusters I and II in these samples (*n*=26, from C, F, and W) and clustered farmed samples (*n*=76, excluding other six that were clustered with wild samples). As shown in Figure 2D, in the Cluster I OTUs (sum together), the wild samples showed higher abundance than the farmed samples (*P*<0.001, Wilcoxon test), whereas no difference was observed for the Cluster II OTUs. This result indicated that these wild samples still enriched Cluster I taxa despite their similar microbiota with farmed ones. These taxa may serve as indicators for host source tracking.

**Functional prediction of microbiome and representative isolates suggested diet can be a key factor for the divergent microbiomes**

To discover the underlying mechanism responsible for the divergence of gut microbiota between wild and farmed croakers, the function of microbiome was predicted by PICRUSt2. As expected, the functional microbiome of wild samples is significantly different from the farmed ones (*P*<0.001, AMOVA, Figure S4). Dissimilarity between pathway and microbiota was highly correlated (Figure S4). In the top-level functional catalogue, relative abundance of Degradation/Utilization/Assimilation-related pathways showed significant difference between wild and captive groups (Figure S5).
Relative abundances of the second-level catalogues in this category, also exhibited high
divergence between two groups (Figure 3A). Remarkably, wild samples were enriched
by fatty acid and lipid degradation pathways, while farmed samples contained higher
proportion of carbohydrate and polymeric compound degradation pathways. Pathway-
level profiling indicated that the functions enriched in farmed samples were related to
degradation of starch, glycogen, chitin, mannan, glucose, galactose, etc. (Figure 3B).

Apparently, the functional prediction of microbiome suggested that a potential
causation for the divergence of gut microbiota between farmed and wild croakers was
diet. Farmed croakers are typical fed with formulated food containing high proportions
and diverse sources of carbohydrate (~30% in dry weight from starch, soybean meal,
shrimp meal, and yeast; see Figure S6 for a typical diet content). However, in natural
habitats, croakers usually prey on zooplanktons (mostly crustaceans) and small fishes
[62], which contain few carbohydrates, although they may obtain high-level chitin-like
materials from crustaceans.

To further examine the hypothesis that diet is a deterministic factor for the microbiota
divergence of wild and farmed croaker, we analyzed the genomes of 14 representative
isolates affiliated with Photobacterium (Ph1 to Ph7) and Psychrobacter (Ps1 to Ps7),
which were obtained from the farmed and wild samples, respectively. Their
phylogenetic information was shown in Figure S7. Firstly, we compared the CAZymes
in the isolates of Photobacterium and Psychrobacter. As expected, Photobacterium
genomes encoded more CAZyme families and genes than each Psychrobacter genomes
(Figure 3C). However, chitinase was widely detected in the Photobacterium strains but
are absent in the Psychrobacter strains (Figure S8). It should be noticed that the genome
of croaker encodes three chitinases [63], which may minimize niche selection for chitin-utilizing microorganisms. Secondly, producing fatty acid was also predicted in the genomes of all the isolates (Figure S8). A reasonable assumption is that a high-carbohydrate diet may decrease the pH value in the rectum by producing short-chain fatty acids [48]. Thus, we tested the pH values of the rectum contents obtained from farmed and wild individuals and found that the assumption is positively supported (Figure 3D). Thirdly, the prediction of optimal pH for the CAZymes indicated higher proportions of acidic glycoside hydrolases (GHs) in Photobacterium than in Psychrobacter (Figure 3E). The difference of the proportions is consistent with their dominant distributions in the guts of farmed and wild croakers, respectively, although glycosyltransferases (GTs), which are usually involved in polysaccharide biosynthesis, have no such signal (Figure 3E). Meanwhile, signal peptides were predicted in 50%-64.8% acidic CAZymes of Photobacterium genomes, suggesting that most of these enzymes are secreted or bound to the cell surface and may partially function at extracellular environment.

In addition, we tested on the biofilm formation capability of the 14 isolates. The results indicated that the Psychrobacter isolates usually formed denser biofilms than the Photobacterium strains (Figure 3F), as confirmed microscopically (Figure S9). Although the experiment was performed in vitro, this result could be a reasonable explanation to the higher gut bacterial loading of wild croakers than that of the farmed ones (Figure 1B).

**Major and core taxa follow non-stochastic pattern in farmed and wild croakers**

Although high beta-diversity among batches and sources were detected, it may result
from neutral processes other than deterministic factors. To evaluate weight of non-stochastic process in the assembly of gut microbiota in farmed and wild croakers, core taxa, which we defined as detected in at least 70% samples, are listed at first. As shown in Figure 4A, among batches of C, F, N, and H, which have enough individuals, there is no overlapping core taxa, indicating the high dynamics between sources and among batches as aforementioned. After combining batches, seven and five core OTUs passed the frequency criterion for wild and farmed sources, respectively (Figure 4B). *Psychrobacter* OTUs, as major differential taxa for farmed and wild samples (Figure 2C), were the major core taxa for the wild group. By contrast, the core taxa for farmed fishes are affiliated with *Vibrio*, *Streptococcus*, *Photobacterium*, etc., without any *Psychrobacter* OTU.

Then, all OTUs were examined for their goodness-of-fit to the neutral model for farmed and wild individuals, respectively (Figure 4C and D). The values indicated low goodness-of-fit to the model for both groups ($R^2=0.378$ and 0.259 for farmed and wild group, respectively). For the OTUs with high relative abundance (>0.1%, mean value), 41.7% in the farmed group fell into the 99% confidence interval, whereas only 8.2% in the wild groups, respectively, were within the region. Moreover, most core OTUs deviated from the 99% confidence interval, except one in the farmed group and one in the wild group (Figure 4C and D). The mean NTIs were higher than zero in farmed and wild individuals ($P<0.05$), indicating that the phylogenetic relatedness of microbial taxa in the two communities are more related than expected by chance (Fig 4E). The βNTI values of 79.8% and 46.1% samples were lower than −2 in wild and farmed group, respectively. It indicates that the deterministic processes (homogeneous selection) are important in gut microbiome assembly of both wild and farmed croakers [64], although
stochastic process may also play a major role in the community assembly of the framed individuals (Fig 4E). These results indicate that non-stochastic processes dominate the assembly of major and core taxa in the gut of croaker, especially for the wild ones, which is fundamental to apply the microbiota-based host source tracking, because stochastically assembled community may introduce more unpredictable noises for newly introduced samples.

Robust microbiota-based host source tracking based on random-forest classification

Although the gut microbiota of wild and farmed croakers displayed high inter-batch variation, the overall divergence and non-stochastic distribution of most abundant OTUs suggested distinguishable and majorly deterministic microbial assembly patterns. We then tested the performance of random forest classification under different ratios of training and test sets. As shown in Figure 5A, the average area-under-curve (AUC) value increased from 0.898 in the 5:5 sets to 0.943 in the 9:1 sets. From the 5:5 set to the 8:2 set, the accuracy for the farmed group was consistently lower than that for the wild group possibly because of the high inter-batch divergence in the farmed group.

To validate whether stochastic assembly has negative effect on the random forest classification, by bootstrapping \( n=100 \), designating 5:5 of training: test for each batch, we calculated the average probability (>50% for a correct assignment) of each sample. Meanwhile, the accumulated relative abundance of neutral OTUs and below-prediction OTUs were also determined for farmed and wild samples, respectively (see Figure 4C and D for the definition of the OTUs). As shown in Figure 5B, we found that samples with high relative abundance of neutral OTUs and below prediction OTUs (higher than
the upper boundary of 90% confidence interval) are more likely to be poorly assigned (average probability <50%) comparing with other samples ($P<0.05$, Fisher’s exact test).

Reasonably, samples with high abundance of below-prediction OTUs are likely dominated by few taxa, processing a simple microbiota that may not well support the classification. The poor assignment for samples with high abundance of stochastic OTUs supported the negative effect of stochastic microbial assembly on the random forest classification.

Then, we chose the 8:2 set, under which all AUC values were higher than 0.9 in 10 replications). Results of leave-one-out validation suggested that some batches (e.g., T and W), could poorly be assigned correctly (Figure 5C). It indicates that good sample representability is a prerequisite for the performance of the machine learning classifier. Moreover, since batch W is from a remote geographical location to most other batches, the low classifying performance of may be partially attributed to biogeography.

As suggested by the $k$-fold cross-validation (Figure S10), using 15 OTUs could generate the lowest error rate of prediction. Thus, the top 15 classifier OTUs that contributed to the accuracy of classification were listed (Figure 5D). A large proportion of them were core OTUs from the wild and farmed groups. Psychrobacter spp. were highly weighted in the algorithm, and its relative abundance is the strongest factor that is positively related to the predicted probabilities for the wild individuals. The optimized classifier was kept for downstream analysis. A wild batch S ($n=15$), which was collected in the same bay as most farmed batches, was additionally tested. It was noticed that the frequency of the top 15 classifier OTUs are significantly lower in this batch than in other wild batches (all $P<0.001$, two-tailed student’s $t$-test, Figure 5E). Batch S was
enriched with only two *Psychrobacter* OTUs, while most other OTUs were extremely low or missing (Figure 5F). Intriguingly, all predicting results were correct, although the probabilities were low (0.70±0.09, Figure 5F). To further validate the superiority of the machine learning algorithm, all samples were clustered with Bray-Curtis distance based on all, core, and the 15 classifier OTUs, respectively (Figure S11). There were 23-24 (approximately 10%) wild samples, including 1-3 from batch S, were clustered with most (>95%) farmed individuals for each samples no matter the referring dataset. Surprisingly, high proportion (>90%) of these wild samples can be correctly assigned by the machine learning approach. These results indicated the robustness of the random forest classifier.

**Discussion**

Although machine learning classification based on gut microbiota has been extensively used in the prediction of host phenotypes in human [65, 66], to our knowledge, this is the first time to utilize it in host source tracking in fish. Similar strategy can be applied in discrimination of other wild and captive fishes or animals. It is well known that insufficient representativeness of a training dataset can cause overfitting and the failure of prediction on newly introduced samples [67]. To determine the divergence between wild and farmed fishes, most previous studies collected few or even a single batch samples, overlooking potential high inter-batch divergence that was observed by our study and potentially causing biases in profiling the taxonomic and functional features. It was noteworthy that even definition of core taxa can be seriously biased when samples are underrepresented as shown in the croakers.

It is not surprising that obvious divergences between the gut microbiome of wild and
farmed croakers were determined in terms of alpha- and beta-diversity. An unexpected phenomenon is that rectum bacterial diversity was lower in the wild individuals than in the farmed samples (Figure 2A). The apparent reason for this is the domination of a single genus *Psychrobacter*. Higher alpha diversity of gut microbiota can be triggered by simplified diets have been reported in fish [68-70]. In addition, formulated food is highly advance to increase digestibility and feed conversion ratio, which might select certain gut bacteria, as suggested by the functional prediction of microbiome and isolates. Notably, apart from source and diet, living habitat, season, and drug utilization may be the other factors shaping gut microbiome [15, 71, 72]. The effect of host size was excluded from our analysis due to its poor capability in explaining inter-batch divergence (Figure 2B), possibly due to all examined individuals are adults. However, other undetermined factors can also be responsible for the gut microbiota divergence between batches and sources.

Stochastic process plays key roles in shaping the microbial assemblages in many environments [73, 74]. However, previous and our studies indicated that deterministic processes usually play important roles in the gut microbial assembly in fishes, suggesting the high niche selection stress on the community structure [15, 17]. In the present study, despite the highly variable microbiota among different batches, we found that almost all major or core OTUs were deviated from the neutral model, indicating their underlying deterministic assembly pattern. A recent study revealed that lower contribution of neutral processes in the gut microbial assembly of wild Atlantic salmon than farmed individuals [10]. Our study also discovered the higher goodness-of-fit to the neutral model in farmed individuals than wild ones, suggesting a disconnection of host-microbe interaction in farmed circumstance. More importantly, stochastic systems
are intrinsically unfavorable for machine learning classification and could generate classifier established by false signals, e.g., p-hacking [64]. Despite the overall poor goodness-of-fit of gut microbial assembly and major OTUs in both wild and farmed croakers, results still support that the high proportion of stochastic OTUs can be linked to the errors in host source tracking (Figure 5B). Successful host source tracking for the newly introduced batch S verified the robustness of the classifier building upon the non-stochastic assembly pattern of gut microbiota. Therefore, it is recommended to evaluate and exclude the effect of stochastic events when applying machine learning host source tracking based on microbiota.

Despite the good performance in discriminating wild and farmed individuals, the microbiota-based classifier may have other untested problems during the practice of wild resource assessment, i.e., to discriminate true wild fishes from those escaping from farming cages and artificial releasing of fry. The dynamics and profile of their gut microbiota have not been examined in the present study. The rapid shift (from days to few months) of gut microbiota during domesticated and diet change have been revealed in African cichlid, European seabass, grass carp, perch, etc. [75-78]. In carnivorous European seabass, mucosa-associated microbiota was found more stable than the corresponding digesta microbiota when shifting to a plant-based diet [77]. Investigations on the effect of diet shift on the dynamics of gut microbiota in different intestinal locations may provide more basis for the practice of wild resource assessment for croaker.

Lastly, although it is not the major aim of the present study, understanding the underlying mechanisms responsible for the divergence of gut microbiome between the
wild and farmed fishes can also provide key information on improving production of aquaculture [21]. Dysbiosis has been widely reported in the aquaculture fishes fed with formulated feed [79]. Our study revealed possible dysbiosis in farmed croaker since potential pathogenic bacterial taxa, such as *Vibrio* spp., *Photobacterium* spp. etc., can be the core taxa, whereas they were less frequently and abundantly detected in wild individuals. Instead, the wild samples were dominated by *Psychrobacter* spp., which was widely detected in the gut of marine fish [5]. A few strains of this genus have been tested for their probiotic applications in fish diets [53-55]. Our results indicated that many strains in this genus can form dense biofilm in abiotic surface and were thus selected by natural diet and alkaline gut environment. These provided additional guidelines for various applications, including improving stock enhancement (e.g., domestication by specialized diet before the release of fries) and diet-based gut microbiota regulation for farmed croakers.

**Conclusions**

Gut microbiome is not only closely related to the health and metabolism of the hosts, but also contains key information on physiological and ecological circumstance of the hosts. Our study revealed the divergence of the gut microbiota between wild and farmed large yellow croakers and suggested that diet is an underlying key factor for the divergence. As the first time, we verified that with less biased datasets and non-stochastic pattern, gut microbiota can be robustly applied to the tracking of host source even in carnivorous fish. Similar strategy can be applied in other fish species in need of discriminating source-unknown individuals. Furthermore, the potential gut microbiota regulating pathways through feeding control may provide improvements in the mariculture and wild stock enhancement.
List of abbreviations

AMOVA, analysis of molecular variance

AUC, area-under-curve

βNTI, beta nearest-taxon index

CAZyme, carbohydrate-active enzyme

GHs, glycoside hydrolases

GTs, glycosyltransferases

NTI, net relatedness index

NMDS, Nonmetric multidimensional scaling

OTU, operational taxonomic unit

qPCR, quantitative PCR

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The sequencing datasets generated during the current study are available in the NCBI database. The 16S rRNA gene datasets were deposited in the Sequence Read Archive under accession number PRJNA679381. The genome sequences of *Photobacterium* and *Psychrobacter* strains in this study have been deposited in the NCBI database under the accession number PRJNA678775.
Competing interests

The authors declare that they have no conflict of interest.

Funding

This study was supported by the National Key Research and Development Program of China (No. 2018YFC1406303), the National Natural Science Foundation of China (No. 31670492 and No. 41976092), Innovation Group Project of Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai) (No. 311021006), and the Fundamental Research Funds for the Central Universities (No. 20720200114).

Author information

Jun Zhu, Hao Li and Ze-zhou Jing contributed equally to this work.

Authors’ contributions

GF and SXC conceived the study. JZ and ZZJ conducted the experiments and collected the samples. JZ and ZZJ performed the analysis of samples. HL, JZ and WZ analyzed the data. GF, JZ, HL and SXC wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Prof. Ling-feng Huang, Prof. Shao-xiong Ding and Dr. Pu-qing Song were thanked for offering partial fish samples.
References

1. Clements KD, Angert ER, Montgomery WL, Choat JH. Intestinal microbiota in fishes: what's known and what's not. Mol Ecol; 2014;23:1891-8.

2. Lee WJ, Hase K. Gut microbiota-generated metabolites in animal health and disease. Nat Chem Biol. 2014;10:416-24.

3. Ley RE, Hamady M, Lozupone C, Turnbaugh PJ, Ramey RR, Bircher JS, et al. Evolution of Mammals and Their Gut Microbes. Science. 2008;320(5883):1647-51.

4. McKenney EA, O'Connell TM, Rodrigo A, Yoder AD. Feeding strategy shapes gut metagenomic enrichment and functional specialization in captive lemurs. Gut Microbes. 2018;9(3):202-17.

5. Egerton S, Culloty S, Whooley J, Stanton C, Ross RP. The gut microbiota of marine fish. Front Microbiol. 2018;9:873.

6. Ghanbari M, Kneifel W, Domig KJ. A new view of the fish gut microbiome: advances from next-generation sequencing. Aquaculture. 2015;448:464-75.

7. Wang AR, Ran C, Ringo E, Zhou ZG. Progress in fish gastrointestinal microbiota research. Rev Aquac. 2018;10:626-40.

8. Ingerslev HC, Strube ML, Jørgensen LVG, Dalsgaard I, Boye M, Madsen L. Diet type dictates the gut microbiota and the immune response against *Yersinia ruckeri* in rainbow trout (*Oncorhynchus mykiss*). Fish Shellfish Immunol. 2014;40(2):624-33.

9. Kormas KA, Meziti A, Mente E, Frentzos A. Dietary differences are reflected on the gut prokaryotic community structure of wild and commercially reared sea bream (*Sparus aurata*). Microbiologyopen. 2014;3:718-28.

10. Llewellyn MS, McGinnity P, Dionne M, Letourneau J, Thonier F, Carvalho GR,
et al. The biogeography of the Atlantic salmon (*Salmo salar*) gut microbiome.

ISME J. 2016;10(5):1280-4.

11. Baldo L, Pretus JL, Riera JL, Musilova Z, Nyom ARB, Salzburger W. Convergence of gut microbiotas in the adaptive radiations of African cichlid fishes. ISME J. 2017;11:1975-87.

12. Miyake S, Ngugi DK, Stingl U. Diet strongly influences the gut microbiota of surgeonfishes. Mol Ecol. 2015;24(3):656-72.

13. Heys C, Cheaib B, Busetti A, Kazlauskaite R, Maier L, Sloan WT, et al. Neutral processes dominate microbial community assembly in Atlantic salmon, *Salmo salar*. Appl Environ Microbiol. 2020;86:e02283-302.

14. Roeselers G, Mittge EK, Stephens WZ, Parichy DM, Cavanaugh CM, Guillemin K, et al. Evidence for a core gut microbiota in the zebrafish. ISME J. 2011;5(10):1595-608.

15. Yan Q, Li J, Yu Y, Wang J, He Z, Van Nostrand JD, et al. Environmental filtering decreases with fish development for the assembly of gut microbiota. Environ Microbiol. 2016;18(12):4739-54.

16. Burns AR, Stephens WZ, Stagaman K, Wong S, Rawls JF, Guillemin K, et al. Contribution of neutral processes to the assembly of gut microbial communities in the zebrafish over host development. ISME J. 2016;10(3):655-64.

17. Razak SA, Scribner KT. Ecological and ontogenetic components of larval lake sturgeon gut microbiota assembly, successional dynamics, and ecological evaluation of neutral community processes. Appl Environ Microbiol. 2020;86(10):e02662-19.

18. Stagaman K, Burns AR, Guillemin K, Bohannan BJ. The role of adaptive immunity as an ecological filter on the gut microbiota in zebrafish. ISME J.
19. Kim D-H, Kim D-y. Microbial diversity in the intestine of olive flounder (*Paralichthys olivaceus*). Aquaculture. 2013;414:103-8.

20. Holben WE, Williams P, Saarinen M, Särkilähti L, Apajalahti JH. Phylogenetic analysis of intestinal microflora indicates a novel *Mycoplasma* phylotype in farmed and wild salmon. Microb Ecol. 2002;44:175-85.

21. Limborg MT, Alberdi A, Kodama M, Roggenbuck M, Kristiansen K, Gilbert MTP. Applied hologenomics: feasibility and potential in aquaculture. Trends Biotechnol. 2018;36:252-64.

22. Ramírez C, Romero J. The microbiome of *Seriola lalandi* of wild and aquaculture origin reveals differences in composition and potential function. Front Microbiol. 2017;8:1844.

23. Ramírez C, Coronado J, Silva A, Romero J. *Cetobacterium* is a major component of the microbiome of giant Amazonian fish (*Arapaima gigas*) in Ecuador. Animals. 2018;8:189.

24. Johansen LH, Jensen I, Mikkelsen H, Bjørn PA, Jansen P, Bergh Ø. Disease interaction and pathogens exchange between wild and farmed fish populations with special reference to Norway. Aquaculture 2011;315:167-86.

25. Einum S, Fleming I. Genetic divergence and interactions in the wild among native, farmed and hybrid Atlantic salmon. J Fish Biol. 1997;50:634-51.

26. Brigante R, Lem A. Price interaction between aquaculture and fishery. Working paper XIII EAFE Conference, Salerno. 2001.

27. Bjørndal T, Guillen J. Market integration between wild and farmed seabream and seabass in Spain. Appl Econ. 2017;4567-68.

28. Karlsson J, Gattami A, Oechtering TJ, Skoglund M. Iterative source-channel
coding approach to Witsenhausen's counterexample. In Proceedings of the 2011 American Control Conference. IEEE; 2011;5348-53.

29. Chen S, Su Y, Hong W. Aquaculture of the large yellow croaker. Aquaculture in China: Success stories and modern trends. John Wiley & Sons Oxford; 2018: 297-308.

30. Guo Y, Zhao W, Gao H, Wang S, Yu P, Yu H, et al. China fishery statistical yearbook. China: Fisheries Agency of China Agriculture Ministry. 2017.

31. Liu M, De Mitcheson YS. Profile of a fishery collapse: why mariculture failed to save the large yellow croaker. Fish and Fisheries. 2008;9:219-42.

32. Wang L, Shi X, Su Y, Meng Z, Lin H. Loss of genetic diversity in the cultured stocks of the large yellow croaker, Larimichthys crocea, revealed by microsatellites. Int J Mol Sci. 2012;13(5):5584-97.

33. Zhang QY, Hong WS, Yang SY, Liu M. Discussion on the division of geographic populations for the large yellow croaker (Larimichthys crocea). Modern Fisheries Information. 2011;2:3-8(in Chinese).

34. Eraslan G, Avsec Ž, Gagneur J, Theis FJ. Deep learning: new computational modelling techniques for genomics. Nat Rev Genet. 2019;20:389-403.

35. Roguet A, Eren AM, Newton RJ, McLellan SL. Fecal source identification using random forest. Microbiome. 2018;6:185.

36. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Appl Environ Microbiol. 2013;79(17):5112-20.

37. MacConaill LE, Burns RT, Nag A, Coleman HA, Slevin MK, Giorda K, et al. Unique, dual-indexed sequencing adapters with UMIs effectively eliminate...
index cross-talk and significantly improve sensitivity of massively parallel sequencing. BMC Genomics. 2018;19:30.

38. Muyzer G, Hottentrager S, Teske A, Wawer C. Denaturing gradient gel electrophoresis of PCR-amplified 16S rDNA. A new molecular approach to analyze the genetic diversity of mixed microbial communities. Microbiol Ecol Management. 1996;3:1-23.

39. Bolger A, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-20.

40. Edgar RC. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. BioRxiv. 2016:081257.

41. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, et al. A communal catalogue reveals Earth’s multiscale microbial diversity. Nature. 2017;551(7681):457-63.

42. Lozupone C, Lladser ME, Knights D, Stombaugh J, Knight R. UniFrac: an effective distance metric for microbial community comparison. ISME J. 2011;5(2):169-72.

43. Yoon SH, Ha SM, Kwon S, Lim J, Kim Y, Seo H, et al. Introducing EzBioCloud: A taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. Int J Sys Evol Microbiol. 2017;67(5):1613-7.

44. Mukuta Y, Harada T. Probabilistic partial canonical correlation analysis. In: Proceedings of The 31st international conference on machine learning. PMLR; 2014; 1449-57.

45. Kolde R. Pheatmap: pretty heatmaps. R package version. 2012;1(2).

46. Oksanen J, Blanchet F, Friendly M, Kindt R, Legendre P, McGlinn D, et al. Package 'vegan': community ecology package, version. 2018.
47. Sloan WT, Lunn M, Woodcock S, Head IM, Nee S, Curtis TP. Quantifying the roles of immigration and chance in shaping prokaryote community structure. Environ Microbiol. 2006;8(4):732-40.

48. Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, etc. Picante: R tools for integrating phylogenies and ecology. Bioinformatics 2010;26:1463-4.

49. Douglas GM, Maffei VJ, Zaneveld JR, Yurgel SN, Brown JR, Taylor CM, et al. PICRUSt2 for prediction of metagenome functions. Nat Biotechnol. 2020;38:685-8.

50. Fernandes AD, Reid JN, Macklaim JM, McMurrough TA, Edgell DR, Gloor GB. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. Microbiome. 2014;2:1-13.

51. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19:455-77.

52. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119.

53. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 2018;9(1):5114.

54. Zhang H, Yohe T, Huang L, Entwistle S, Wu P, Yang Z, et al. dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. 2018;46(W1):W95-101.
55. Petersen TN, Brunak S, Von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 2011;8:785-6.

56. Lin H, Chen W, Ding H. AcalPred: a sequence-based tool for discriminating between acidic and alkaline enzymes. PloS One. 2013;8:e75726.

57. Shaffer M, Borton MA, McGivern BB, Zayed AA, La Rosa SL, Solden LM, et al. DRAM for distilling microbial metabolism to automate the curation of microbiome function. Nucleic Acids Res. 2020;48:8883-900.

58. O'Toole GA. Microtiter dish biofilm formation assay. J Vis Exp. 2001;47:10-1.

59. Breiman L. Random forests. Mach Learn. 2001;45:5-32.

60. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;12:77.

61. Kuhn M. Caret: classification and regression training. R package version 6.0-86. 2020.

62. Xu J, Chen J, Tian F, Xu Z. Summer diet composition and feeding ecology of large yellow croaker (Larimichthys crocea) in Guanjing Yang. Journal of Fishery Sciences of China. 2012;19 (in Chinese).

63. Wu C, Zhang D, Kan M, Lv Z, Zhu A, Su Y, etc. The draft genome of the large yellow croaker reveals well-developed innate immunity. Nat Commun. 2014, 5:5227.

64. Stegen JC, Lin X, Fredrickson JK, Chen X, Kennedy DW, Murray CJ, et al. Quantifying community assembly processes and identifying features that impose them. ISME J. 2013;7:2069-79.

65. Subramanian S, Huq S, Yatsunenko T, Haque R, Mahfuz M, Alam MA, et al. Persistent gut microbiota immaturity in malnourished Bangladeshi children.
66. Kong F, Hua Y, Zeng B, Ning R, Li Y, Zhao J. Gut microbiota signatures of longevity. Curr Biol. 2016;26:R832-3.

67. Schat E, van de Schoot R, Kouw WM, Veen D, Mendrik AM. The data representativeness criterion: predicting the performance of supervised classification based on data set similarity. PLoS One. 2020;15(8):e0237009.

68. Schat E, van de Schoot R, Kouw WM, Veen D, Mendrik AM. The data representativeness criterion: predicting the performance of supervised classification based on data set similarity. PLoS One. 2020;15(8):e0237009.

69. Bolnick DI, Snowberg LK, Hirsch PE, Lauber CL, Org E, Parks B, et al. Individual diet has sex-dependent effects on vertebrate gut microbiota. Nat Commun. 2014;5(1):4500.

70. Li P, Wu G. Roles of dietary glycine, proline, and hydroxyproline in collagen synthesis and animal growth. Amino Acids. 2018;50:29-38.

71. Dhanasiri AK, Brunvold L, Brinchmann MF, Korsnes K, Bergh Ø, Kiron V. Changes in the intestinal microbiota of wild Atlantic cod Gadus morhua L. upon captive rearing. Microb Ecol. 2011;61:20-30.

72. Dulski T, Kozłowski K, Ciesielski S. Habitat and seasonality shape the structure of tench (Tinca tinca L.) gut microbiome. Sci Rep. 2020;10:4460.

73. Roguet A, Laigle GS, Therial C, Bressy A, Soulignac F, Catherine A, et al. Neutral community model explains the bacterial community assembly in freshwater lakes. FEMS Microbiol Ecol. 2015;91(11):fiv125.

74. Chen W, Ren K, Isabwe A, Chen H, Liu M, Yang J. Stochastic processes shape microeukaryotic community assembly in a subtropical river across wet and dry seasons. Microbiome. 2019;7(1):138.
75. Baldo L, Riera JL, Tooming-Klunderud A, Albà MM, Salzburger W. Gut microbiota dynamics during dietary shift in eastern African cichlid fishes. PloS One. 2015;10:e0127462.

76. Hao YT, Wu SG, Xiong F, Tran NT, Jakovlić I, Zou H, et al. Succession and fermentation products of grass carp (Ctenopharyngodon idellus) hindgut microbiota in response to an extreme dietary shift. Front Microbiol. 2017;8:1585.

77. Serra CR, Oliva-Teles A, Enes P, Tavares F. Gut microbiota dynamics in carnivorous European seabass (Dicentrarchus labrax) fed plant-based diets. Sci Rep. 2021;11:447.

78. Zha Y, Eiler A, Johansson F, Svanbäck R. Effects of predation stress and food ration on perch gut microbiota. Microbiome. 2018;6:28.

79. Infante-Villamil S, Huerlimann R, Jerry DR. Microbiome diversity and dysbiosis in aquaculture. Rev Aquacult. 2020;1753-5131.
**Figure Legends**

Figure 1 Bacterial alpha diversity (A), loading (B), and high-rank taxa (C) in the rectum of the wild and farmed croakers. Wilcoxon test was applied to the comparisons.

Figure 2 Beta-diversity of the gut microbiota of croakers. (A) NMDS carried out on weighted Unifrac distance shows a separation between the farmed and wild groups. Only inter-batch pairs without significant differences are marked (AMOVA test, \( P>0.01 \)). (B) The explanation of variance in gut microbiome by batch, source, sampling season, and body weight (ANOVA test, \(^*\) for \( P<0.05 \)). (C) Heatmap shows the main OTU relative abundance in the farmed and wild individuals. Only 24 OTUs with the mean abundance of >0.2% and occurrence of >20% are shown. Rows are clustered according to Pearson correlation, and OTUs are stratified into two clusters. The columns (samples) are clustered according to Euclidean distance. The total relative abundance of individuals is shown in the bar plot. (D) Relative abundance of OTUs belonging to Clusters I and II in the farmed (\( n=76 \)) individuals and wild (\( n=26 \)) ones that are wrongly clustered into the farmed group.

Figure 3 The functional prediction of gut microbiome and representative isolates belonging to *Photobacterium* and *Psycrobacter*. (A) Significantly differentiated MetaCyc pathways in farmed or wild individuals and only the top 10 pathways with the largest differences are shown. (B) Heatmap shows the relative abundance of MetaCyc pathways (top 20 pathways based on average relative abundance) in farmed and wild samples. (C) The pH of rectum contents in the farmed (\( n=15 \)) and wild (\( n=27 \)) individuals. (D) The distribution of CAZyme families and genes. (E) The ratio of predicted acidic and alkaline GHs and GTs in the genome. (F) Biofilm formation capability of *Photobacterium* and *Psycrobacter* isolates. Welch's t test (Benjamini-Hochberg corrected \( P \) values, \(^*\) \( P<0.05 \), \(^{**}\) \( P<0.01 \), \(^{***}\) \( P<0.001 \)) was used in (A) and (B). Wilcoxon test was used in (C), (D), (E), and (F).

Figure 4 Distribution of the core OTUs and the goodness-of-fit to neutral model for the gut microbiota. Core taxa were determined in each of four batches (A) and all individuals from each source (B). The OTUs presented in >70% samples were defined as core taxa. The boxplot shows the relative abundance (left axis) and diamonds are
represented detected frequency (right axis) in framed and wild samples (B). The fit to the neutral model for the farmed (C) and wild samples (D). The dashed curve (in C and D) represents 99% confidence intervals of the model prediction. $R^2$ indicates the goodness-of-fit to the neutral model. Pie charts show the proportion of within, above, and below the prediction of high-abundance OTUs (relative abundance $>10^{-3}$ as separated by the vertical dash line). (E) Violin plot of NTI and $\beta$NTI. Horizontal dashed lines (NTI or $\beta$NTI values at -2 and 2) indicate thresholds for determining the assembly pattern [64].

Figure 5 Performance of the random forest classifier. (A) The AUC and predicted accuracy of the farmed and wild groups under different dataset stratifications with 10 replications. (B) Effect of the abundances of OTUs within and below prediction (see Figure 4 C&D for the definition) on host source tracking. The training set: test set is 5:5 and the bootstrapping number is 100. The dashed line indicates the upper boundary of 90% confidence interval. Fisher’s exact test is applied to compare the samples above and within confidence interval. (C) The predicted accuracy of random forest classifier based on leave-one-out validation for each batch (splitting training set: test set=8:2, bootstrapping $n=20$). (D) The top 15 most important OTUs identified by the random forest classifier. (E) The detected frequency of the top 15 most important OTUs in the four main batches, two-tailed Student’s $t$-test, FDR-corrected, * $P<0.05$, ** $P<0.01$, *** $P<0.001$. (F) The predicted probabilities of samples from batch S based on the optimized classifier established using the top 15 most important OTUs (top panel). Heatmap shows the relative abundance of these OTUs in batch S (bottom panel).
Bacterial alpha diversity (A), loading (B), and high-rank taxa (C) in the rectum of the wild and farmed croakers. Wilcoxon test was applied to the comparisons.
Figure 2

Beta-diversity of the gut microbiota of croakers. (A) NMDS carried out on weighted Unifrac distance shows a separation between the farmed and wild groups. Only inter-batch pairs without significant differences are marked (AMOVA test, P>0.01). (B) The explanation of variance in gut microbiome by batch, source, sampling season, and body weight (ANOVA test, * for P<0.05). (C) Heatmap shows the main OTU relative abundance in the farmed and wild individuals. Only 24 OTUs with the mean abundance of >0.2% and occurrence of >20% are shown. Rows are clustered according to Pearson correlation, and OTUs are stratified into two clusters. The columns (samples) are clustered according to Euclidean distance. The total relative abundance of individuals is shown in the bar plot. (D) Relative abundance of OTUs belonging to Clusters I and II in the farmed (n=76) individuals and wild (n=26) ones that are wrongly clustered into the farmed group.
The functional prediction of gut microbiome and representative isolates belonging to Photobacterium and Psycrobacter. (A) Significantly differentiated MetaCyc pathways in farmed or wild individuals and only the top 10 pathways with the largest differences are shown. (B) Heatmap shows the relative abundance of MetaCyc pathways (top 20 pathways based on average relative abundance) in farmed and wild samples. (C) The pH of rectum contents in the farmed (n=15) and wild (n=27) individuals. (D) The distribution of CAZyme families and genes. (E) The ratio of predicted acidic and alkaline GHs and GTs in the genome. (F) Biofilm formation capability of Photobacterium and Psycrobacter isolates. Welch's t test (Benjamini-Hochberg corrected P values, * P<0.05, ** P<0.01, *** P<0.001) was used in (A) and (B). Wilcoxon test was used in (C), (D), (E), and (F).
Figure 4

Distribution of the core OTUs and the goodness-of-fit to neutral model for the gut microbiota. Core taxa were determined in each of four batches (A) and all individuals from each source (B). The OTUs presented in >70% samples were defined as core taxa. The boxplot shows the relative abundance (left axis) and diamonds are represented detected frequency (right axis) in framed and wild samples (B). The fit to the neutral model for the farmed (C) and wild samples (D). The dashed curve (in C and D) represents 99% confidence intervals of the model prediction. R^2 indicates the goodness-of-fit to the neutral model. Pie charts show the proportion of within, above, and below the prediction of high-abundance OTUs (relative abundance>10^-3 as separated by the vertical dash line). (E) Violin plot of NTI and βNTI. Horizontal dashed lines (NTI or βNTI values at -2 and 2) indicate thresholds for determining the assembly pattern [64].
Figure 5

Performance of the random forest classifier. (A) The AUC and predicted accuracy of the farmed and wild groups under different dataset stratifications with 10 replications. (B) Effect of the abundances of OTUs within and below prediction (see Figure 4 C&D for the definition) on host source tracking. The training set: test set is 5:5 and the bootstrapping number is 100. The dashed line indicates the upper boundary of 90% confidence interval. Fisher’s exact test is applied to compare the samples above and within confidence interval. (C) The predicted accuracy of random forest classifier based on leave-one-out validation for each batch (splitting training set: test set=8:2, bootstrapping n=20). (D) The top 15 most important OTUs identified by the random forest classifier. (E) The detected frequency of the top 15 most important OTUs in the four main batches, two-tailed Student’s t-test, FDR-corrected, * P<0.05, ** P<0.01, *** P<0.001. (F) The predicted probabilities of samples from batch S based on the optimized classifier established using the top 15 most important OTUs (top panel). Heatmap shows the relative abundance of these OTUs in batch S (bottom panel).

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

- Supplementaryresubmisson.docx