

# Dix-seq: An integrated pipeline for fast amplicon data analysis

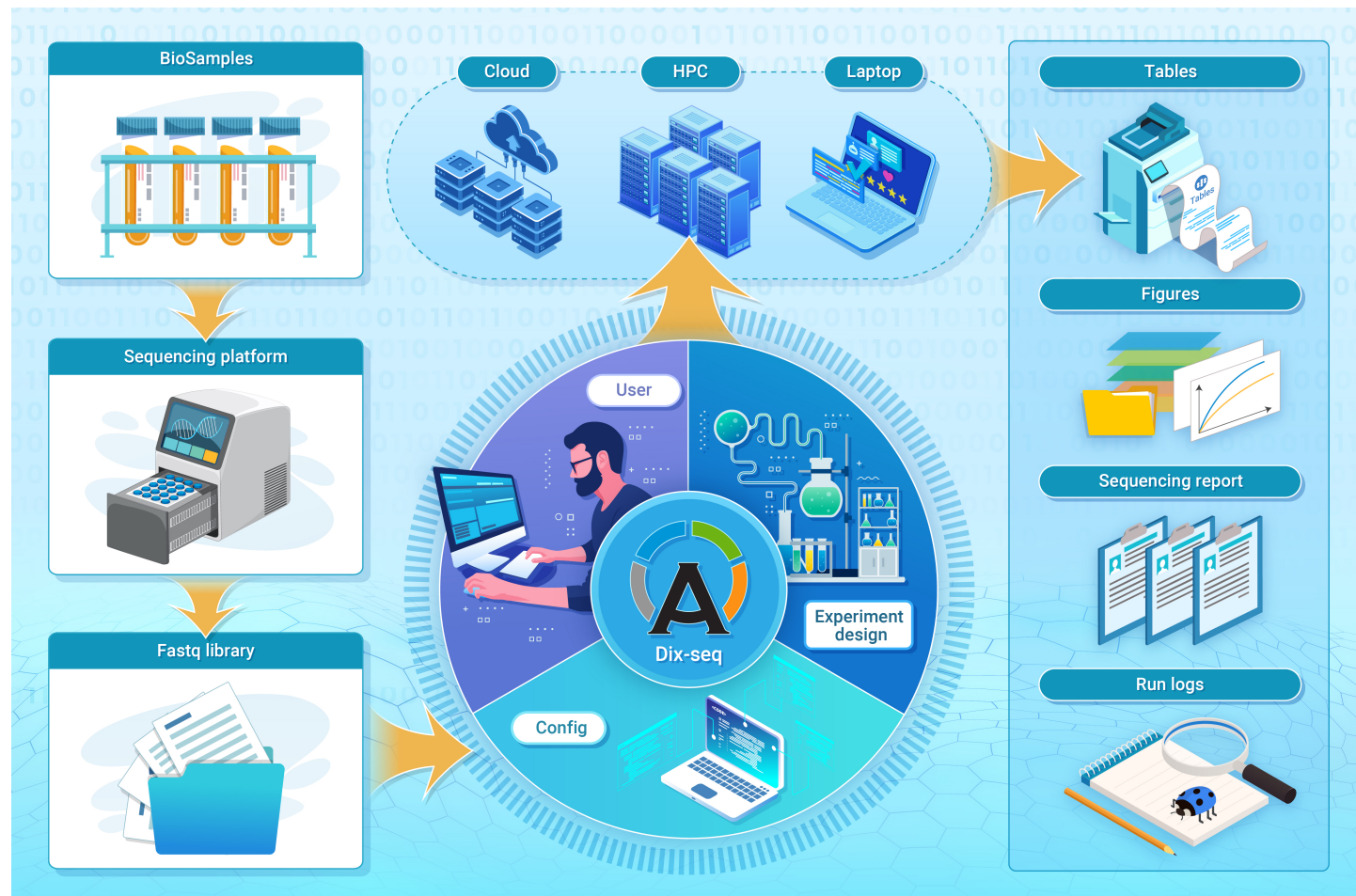
Pengsheng Dong,<sup>1,2</sup> Yubin Chen,<sup>2</sup> Yongjun Wei,<sup>3,4,\*</sup> Xiangyang Zhao,<sup>5</sup> Ting Wang,<sup>2</sup> Shan Jiang,<sup>6</sup> Jiajie Xu,<sup>2</sup> Tianqi Ren,<sup>7</sup> Ming Li,<sup>1,\*</sup> and Lei Zhang<sup>8,\*</sup>

\*Correspondence: yongjunwei@zzu.edu.cn (Y.W.); liming@henau.edu.cn (M.L.); zhanglei@logictek.cn (L.Z.)

Received: October 18, 2024; Accepted: February 21, 2025; Published Online: February 22, 2025; <https://doi.org/10.59717/j.xinn-life.2024.100120>

© 2025 The Author(s). This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>).

## GRAPHICAL ABSTRACT



## PUBLIC SUMMARY

- An integrated pipeline for optimized amplicon analysis has been developed.
- Dix-seq offers a one-step process with a single parameter sheet file to complete the entire pipeline.
- The modular design of Dix-seq supports custom analysis.
- Retrospective scripts make reproducing results or debugging errors easy.

# Dix-seq: An integrated pipeline for fast amplicon data analysis

Pengsheng Dong,<sup>1,2</sup> Yubin Chen,<sup>2</sup> Yongjun Wei,<sup>3,4,\*</sup> Xiangyang Zhao,<sup>5</sup> Ting Wang,<sup>2</sup> Shan Jiang,<sup>5</sup> Jiajie Xu,<sup>2</sup> Tianqi Ren,<sup>7</sup> Ming Li,<sup>1,\*</sup> and Lei Zhang<sup>8,\*</sup>

<sup>1</sup>College of Animal Science and Technology, Henan Agricultural University, Zhengzhou 450046, China

<sup>2</sup>State Key Laboratory for Managing Biotic and Chemical Threats to the Quality and Safety of Agro-Products, School of Marine Sciences, Ningbo University, Ningbo 315211, China

<sup>3</sup>Key Laboratory of Advanced Drug Preparation Technologies, Ministry of Education, School of Pharmaceutical Sciences, Zhengzhou University, Zhengzhou 450051, China

<sup>4</sup>Laboratory of Synthetic Biology, Zhengzhou University, Zhengzhou 450051, China

<sup>5</sup>Jiangsu Provincial Key Lab for Organic Solid Waste Utilization, Jiangsu Collaborative Innovation Center for Solid Organic Wastes, Educational Ministry Engineering Center of Resource-saving fertilizers, Nanjing Agricultural University, Nanjing 210095, China

<sup>6</sup>State Key Laboratory of Estuarine and Coastal Research, East China Normal University, Shanghai 200241, China

<sup>7</sup>Qilu Normal University, Jinan 250200, China

<sup>8</sup>DeepBiome. Co. Ltd., Shanghai 200031, China

\*Correspondence: [yongjunwei@zzu.edu.cn](mailto:yongjunwei@zzu.edu.cn) (Y.W.); [liming@henau.edu.cn](mailto:liming@henau.edu.cn) (M.L.); [zhanglei@logictect.cn](mailto:zhanglei@logictect.cn) (L.Z.)

Received: October 18, 2024; Accepted: February 21, 2025; Published Online: February 22, 2025; <https://doi.org/10.59717/j.xinn-life.2024.100120>

© 2025 The Author(s). This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>).

Citation: Dong P., Chen Y., Wei Y., et al. (2025). Dix-seq: An integrated pipeline for fast amplicon data analysis. *The Innovation Life* 3:100120.

Rapid advancements in sequencing technologies in the past decade have driven the widespread adoption of amplicon metagenome. However, current amplicon data analysis software/pipelines often require manual intervention spanning multiple steps, necessitating a clear understanding of parameters and hindering inexperienced users from automating their workflows. Here, we introduce Dix-seq, a fully containerized tool for rapid, automated, and scalable amplicon data analysis. With one single command, Dix-seq can process raw amplicon sequences down to various statistical and visualization results, generate html-based reports, and retrospective logfiles. Dix-seq utilizes a single parameter sheet file to drastically simplify its command line interface, making it much more approachable by inexperienced users while improving study reproducibility. The modular design of Dix-seq enables rapid adoption of new methods and databases into its software frame. Currently, more than 21 algorithms, software, and third-party procedures have been integrated into eight modules in Dix-seq, while more are coming down the line. This approach also allows experienced users to fine-tune the workflow, facilitating customized analysis. Benchmarks performed on datasets from real-world case studies demonstrated Dix-seq's capabilities in generating publish-ready figures integrated with statistical information and extracting biologically meaningful patterns. Furthermore, it remained highly effective at detecting variance upon simulated sequencing depth drop, the results remained robust down to a depth of 11000 and 1000 in all and certain fronts, such as phylogenetic diversity and Pearson correlation, respectively. In summary, Dix-seq is a convenient yet highly customizable tool for amplicon data analysis, making it an ideal choice for both entry-level and experienced users.

## INTRODUCTION

The launch of the Human Microbiome Project (HMP) and the Earth Microbiome Project (EMP) ignited research hotspots in microbiome science.<sup>1,2</sup> The exploration of microbiome and microbial community helps us better understand the impact of microbes, especially their ecological roles on human disease, agricultural productivity, environmental protection, and One Health, fostering interdisciplinary collaboration and paving the way for innovative applications and economic growth.<sup>3</sup> Amplicon sequences segmented from marker genes, such as 16S rRNA gene, 18S rRNA gene, internal transcribed spacer (ITS) sequences, functional genes (i.e., denitrification genes *nirS* and *nirK*) contain valuable taxonomy information with fine-scale resolution.<sup>4,5</sup> Coupled with next-generation sequencing techniques, amplicon metagenome offers rapid, affordable, and host-contamination proof insights into the microbial diversity, taxonomic composition, and functional information of the microbial community, making it an ideal choice for many researchers.<sup>6</sup> Consequentially, amplicon data analysis of microbiomes has become the first/indispensable step for quickly understanding their basic characteristics, assembly mechanisms, interaction mode, and ecological functions.<sup>7</sup>

Various amplicon analysis tools and pipelines have been developed for sequence processing, taxonomy classification, quantitative ecology, visualization, and process integration to help us better understand the environmental or human microbiome.<sup>1,7</sup> However, data analysis of amplicon sequencing

is still challenging in many ways, barring beginners from entry.<sup>8</sup> Most software, methods, or algorithms usually focus on specific steps and functions of data analysis. For example, mulitqc,<sup>9</sup> Trimmomatic,<sup>10</sup> Cutadapt,<sup>11</sup> and flash were used for sequence processing;<sup>12</sup> UNOISE3,<sup>13</sup> DADA2,<sup>14</sup> and Deblur for sequence denoising;<sup>15</sup> UPARSE,<sup>16</sup> Uclust,<sup>17</sup> and SortMeRNA for calling OTU (operational taxonomic unit);<sup>18</sup> SINTAX,<sup>19</sup> RDP,<sup>20</sup> the classify-sklearn method in QIIME2,<sup>21</sup> and kraken2 as taxonomic profilers;<sup>22</sup> the R packages, vegan,<sup>23</sup> phyloseq,<sup>24</sup> MicrobiomeAnalystR,<sup>25</sup> Microeco,<sup>26</sup> EasyMicroPlot,<sup>27</sup> and ggClusterNet for downstream analysis or visualization.<sup>28</sup> Therefore, pipelines designed to integrate the whole analysis processes from raw sequencing data to visualization were popular, such as mothur,<sup>29</sup> USEARCH,<sup>17</sup> VSEARCH,<sup>30</sup> QIIME1 & 2,<sup>21,31</sup> Microbiome Helper,<sup>32</sup> and EasyAmplicon.<sup>33</sup> However, they still require users to have a clear understanding of various parameters in each step during the analysis, and familiar skills with computer and bioinformatics knowledge, making it a challenge for researchers, even those experienced in bioinformatics, let alone those who are not tech-savvy.<sup>34</sup> For instance, DADA2-based pipelines require the user to set reads truncation lengths when processing paired-end datasets, which often involves multiple attempts of trial-and-errors or additional third-party software.<sup>35</sup> Recently, software that supports web-based graphical interface had been developed to improve the user experience, such as Qiita,<sup>36</sup> MG-RAST,<sup>37</sup> Galaxy,<sup>38</sup> gcMeta,<sup>39</sup> q2Galaxy,<sup>21</sup> MicrobiomeAnalyst 2,<sup>25</sup> and Parallel-Meta Suite.<sup>40</sup> However, they were usually limited in debugging support at user level, large-scale raw sequence data analyses, customized analyses, scalability, and supports for third-party software. Public servers may suffer further performance hits by network latency or limitations of web-frame and web server resources. Additionally, due to their complex nature, some tools or pipelines may take too long to run, creating another headache for their users. Finally, the outputs obtained from those pipelines often need further polishing with additional software/tools before publication.

How the sequence data is processed could also considerably impact the final result, especially in calling microbial features (such as OTUs) and taxonomic annotations. The popular algorithms for calling microbial features had shifted from clustering chimera-checked reads based on sequence identity (such as QIIME1 and mothur) to directly extracting amplicon sequence variant (ASV) or zero-radius operational taxonomic unit (ZOTU) from reads by implementing sequence error correction using approaches of sequence variant resolution (such as DADA2,<sup>14</sup> Amplicl,<sup>41</sup> Swarm,<sup>42</sup> and UNOISE3).<sup>13</sup> Benchmarks performed in several prior studies showed that different methods of calling microbial features, including different denoisers, could generate statistically different results, which was commonly pronounced in alpha diversity metrics.<sup>43-44</sup> Different taxonomic profilers have also shown a similar trend, trading between sensitivity and accuracy.<sup>45-46</sup> The variance caused by different analysis strategies could also be related to the diversity, type of samples, and read depth.<sup>47</sup> For instance, DADA2 was often recommended as the better method to process gut microbiome data;<sup>14</sup> and Nilsen et al. have surfaced that Swarm and UNOISE performed better than DADA2 and Deblur for high-diversity environmental samples.<sup>42</sup> Therefore, choosing appropriate indicators or parameters to evaluate the applicability of different amplicon analysis pipelines, especially on their feature callers and taxonomic profilers, is crucial.

In this study, we developed Dix-seq (<https://github.com/jameslz/dix-seq>, Gitee mirror: <https://gitee.com/dpsh/dix-seq>), an open-source, cross-platform, and cloud-ready software package for rapid and comprehensive microbiome analysis with special focus on the user experience. Currently, native installation on Unix-like systems such as Linux, MacOS, and Windows Subsystem for Linux 2 (WSL 2) are supported. A pre-configured container image is also provided for rapid deployment on the cloud or local platforms (detailed in Materials and Methods). To appeal to users for standardization and repeatability of data analysis, the design of the parameter sheet in Dix-seq enables users to overview and modify all the parameters in a centralized way. Which enables the one-step pipeline through which publication-quality statistical results and figures are generated. The user can also easily share the parameter sheet file and re-run their analysis on other systems via the same file, thereby improving reproducibility. For experienced analysts who seek fine-tuning of their workflows, the modularized design of Dix-seq allows the users to integrate other third-party procedures, which provides a strong basis for custom analysis. Additionally, Dix-seq integrates various statistical and data visualization procedures, which were carefully polished to make the auto-generated results ready for publication. Finally, benchmarks against currently popular tools of amplicon analysis (such as QIIME1, QIIME2, and EasyAmplicon) were also performed using a real-world microbiome dataset comprised of highly diverse environmental samples. The performance of feature callers, taxonomic profilers, and diversity analysis in different conditions were systematically evaluated to demonstrate the applicability of Dix-seq.

## MATERIALS AND METHODS

Dix-seq integrates several different amplicon analysis algorithms and tools for microbiome data analysis, most of which require manual inputs of various parameters across multiple steps. The parameter sheet was designed to collect all the parameters in the global environment and step-by-step commands into a single text file. Dix-seq performs raw sequence data filtering, calling ZOTUs (including denoising and chimera removal and table construction), taxonomic annotation, and diversity analyses by the built-in USEARCH11 and our self-developed tool atlas-utils (<https://github.com/jameslz/atlas-utils>) in the basic analysis pipeline.<sup>17</sup> The advanced mode of Dix-seq could automatically generate the results of functional prediction differential analyses and integrate the final statistical and visual results by using the PICRUST2 algorithm,<sup>48</sup> DESeq2-kit, kwt-kit, lefse-kit, wrist-kit (<https://github.com/jameslz/dix-seq/tree/master/thirdpart>), respectively. The low-coupling modular design realized by subfunctions ensures the independence of subprograms in Dix-seq. The new module of third-party functions could also be added in Dix-seq by using the built-in Perl script (also in supplemental information). Both the general (e.g., Greengenes,<sup>49</sup> RDP<sup>20</sup> SILVA,<sup>50</sup> and UNITE<sup>51</sup>) and customized [e.g., human oral microbiome database (HOMD),<sup>52</sup> and microbial database for activated sludge (MiDAS4)<sup>53</sup>] microbiome reference database with the UDB format (those six databases could be downloaded from figshare, <https://doi.org/10.6084/m9.figshare.27978744.v2>) were supported in Dix-seq for using the SINTAX algorithm,<sup>19</sup> and the detail method of formatting annotation database could be approached in supplemental information and Gitee wiki (<https://gitee.com/dpsh/dix-seq/wikis>). To help the user take fine control of their workflow and improve error tracebacks, Dix-seq was designed to record all the invoked commands from the encapsulated programs into \*.sh files in the "run" folder. The version and other information of all the statistical analyses and visualization implemented in Dix-seq were summarized in Table S1.

Dix-seq is coded mainly in Perl, R, and Python languages and could be run in command-line (terminal) supported by Unix-like systems such as MacOS and Linux and Windows Subsystem for Linux 2 (WSL 2). The Dix-seq program could be used by (1) cloning from GitHub (<https://github.com/jameslz/dix-seq>) and Gitee (<https://gitee.com/dpsh/dix-seq>) and installing dependent software and packages listed in Table S1, (2) rapid deployment from an Apptainer container within Dix-seq (SIF file hosted at <https://doi.org/10.6084/m9.figshare.23257829>), and (3) online-workspace incorporated in Bio-OS (<https://www.volcengine.com/product/bioos>), as part of the standard operating procedure for analyzing microbiome data. Note that if the user chooses the deployment method (1), administrator privileges may be required

depending on the OS.

Two studies of microbiomes in natural and artificial environments were re-analyzed for benchmark purposes, and the performance of Dix-seq was accessed. Study I: Biogeography study on microbiome of estuarine sediment. A total of 48 surface sediment samples were collected across four zones, including two near (MG1 and MG2) and two outer (MG3 and MG4) of the Changjiang River estuary.<sup>47</sup> Study II (Wang et al.): Time-series study on the microbiome of aquaculture animals. Fine-scale succession patterns of bacterial community for shrimp larvae were analyzed by using 101 shrimp larvae and 92 rearing water samples across the whole developmental cycle, including Nauplius, Zoea I, II, and III, Mysis, and early Postlarvae stages.<sup>54</sup> Five alternate mainstream pipelines including USEARCH (v11.0.667) and VSEARCH (v2.22) based pipeline EasyAmplicon<sup>33</sup> (v1.19), both the DADA2 (v1.26.0) and Deblur (v1.1.1) workflows in QIIME2 (amplicon distribution, v2023.9) and the open-reference OTU clustering pipeline with both the 97% (S97%, employed in Wang et al.) and 99% (S99%) sequence similarity in QIIME1 (v1.9.1) were also executed<sup>14, 15, 21, 31</sup>. To maximize comparability, the SILVA database (v132) was used as the reference taxonomy database for all the pipelines, while other parameters remained as defaults. The detailed methods for comparison of those tools were shown in supplemental methods, and the analysis codes could be accessed through GitHub ([https://github.com/biodps/Performance\\_evaluation\\_of\\_Dix-seq](https://github.com/biodps/Performance_evaluation_of_Dix-seq)).

## RESULTS AND DISCUSSION

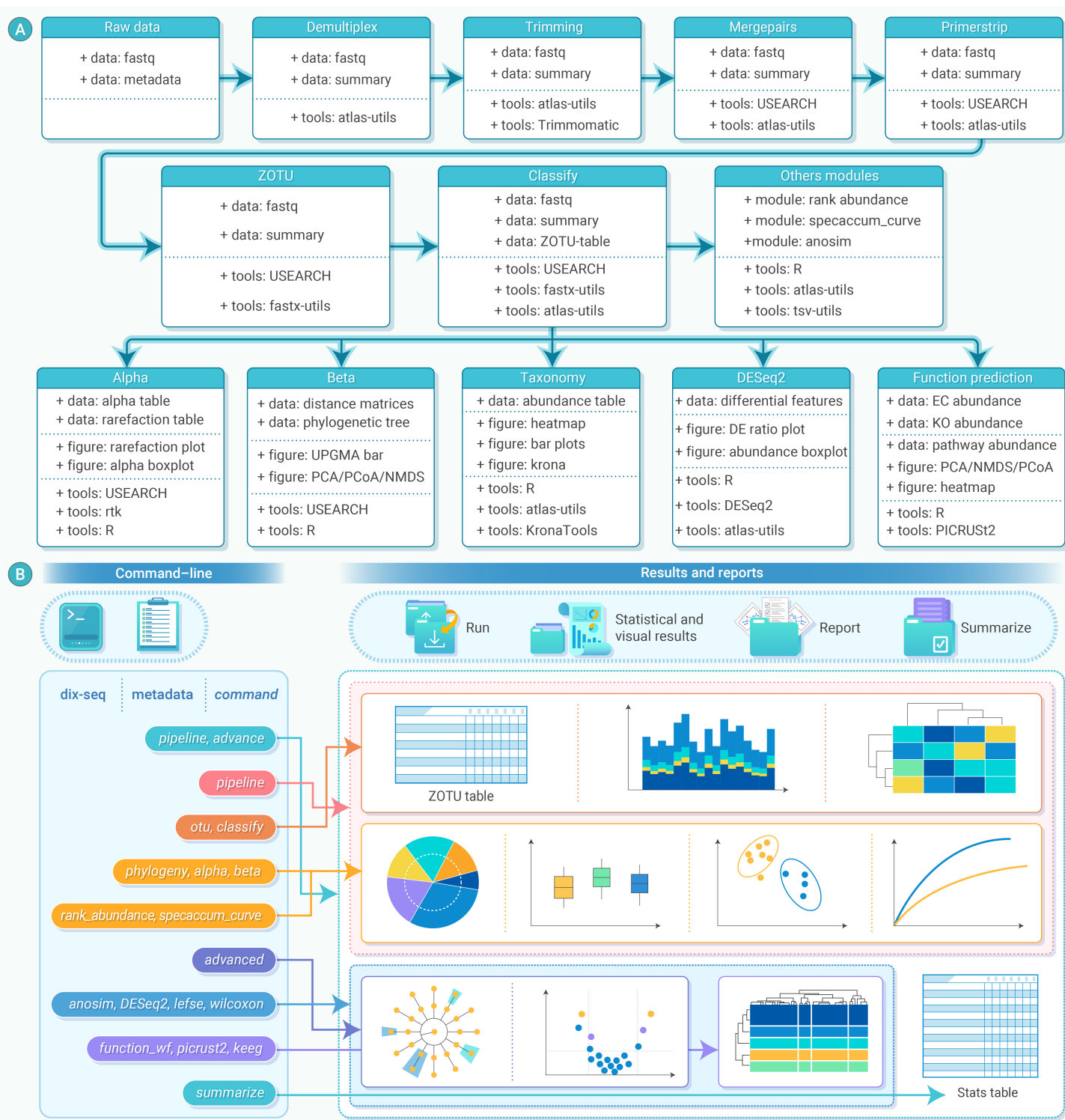
### Overview of Dix-seq

Currently, several suite tools for amplicon data analysis are available, and each of them has unique strengths.<sup>7,21</sup> Dix-seq packs multiple modules in its software frame, making it capable of raw data demultiplex, chimeric sequence removal, sequence classification, alpha and beta diversity, functional prediction, and differential analysis (Figure 1). It has a user-friendly command-line interface and generates publication-ready figures in vector (PDF) and bitmap (PNG) formats, making it a welcoming addition for researchers. Dix-seq is also very flexible and highly customizable, making the integration of diverse third-party procedures and pipeline optimizations possible (Table S1). To facilitate rapid deployment on different operating systems, Dix-seq (version 1.0.0) and all its dependencies were packaged into an Apptainer image (<https://doi.org/10.6084/m9.figshare.23257829>), enabling the user to reproduce runs on high-performance computing (HPC) clusters, local workstations, and laptops. Moreover, Dix-seq could also be deployed in online workspaces as part of established public platforms (such as Bio-OS, <https://www.volcengine.com/product/bioos>) or as independent cloud instances.

### Ease of use

The command line of Dix-seq was designed with special emphasis on its ease of use. It only contains the following three parts: "dix-seq", the parameter sheet (a text file), and a sub-command. The parameter sheet file stores all the global and local parameters of each analysis step, while the sub-command dictates what module/pipelines to run. This approach drastically reduces complexity and enables the user to tune all the parameters easily and all at once (details in supplemental methods). For example, different options of taxonomic databases and classifiers (e.g., SINTAX or RDP Naive Bayesian Classifier) could be selected in the parameter sheet to flexibly modulate the taxonomic annotation strategies.<sup>19,20</sup> Dix-seq provides a total of 20 sub-commands, including two global workflows ("pipeline" and "advance") and 18 command interfaces. The global workflows of "pipeline" and "advance" in Dix-seq perform the basic data analysis and advanced data analysis, respectively. The basic data analysis includes results and visualization of community diversity, composition, and function prediction, while the advanced data analysis includes the algorithms of similarity analysis (ANOSIM), DESeq2,<sup>55</sup> and Linear discriminant analysis (LDA).<sup>56</sup> Furthermore, by using the combined sub-command, "dix-seq metadata pipeline, advanced", these two workflows could be chained together into one step, enabling the fully automated analysis (Table S2) to transform raw sequence data to publication-ready statistical and visualization results without any user intervention. Meanwhile, the 18 command interfaces pack individual steps of the pipelines, such as reads preparation, ZOTU tables construction, taxo-





**Figure 1. Dix-seq: An integrated pipeline for fast amplicon data analysis** (A) The steps and software used in the Dix-seq pipeline. (B) A schematic overview demonstrating one-step or step-by-step mode to finish the whole pipeline, and modular design for custom analysis using Dix-seq.

nomic composition, diversity analysis, miscellaneous analysis, variance analysis, and functional prediction (Figure 1 & Table S2). A built-in example was also packaged alongside all distributions of Dix-seq, which includes the raw paired-end sequence data, the parameter sheet file (metadata.txt), and sample group information (mapping\_file.txt) for users to try out.

### Modularity and reproducibility

Dix-seq also took a highly modular approach that functionally resembles QIIME2, which ensures the independence of subprograms in Dix-seq.<sup>21</sup> This allows for rapid future development, testing, and adoption of additional third-

party modules while smoothing the collaboration between developers by using a model script for adding custom functions to Dix-seq (details in supplemental information). Analysis by Dix-seq is also highly reproducible, due to the centralized settings in the parameter sheet file which guarantees pipeline reproducibility. Dix-seq also records and summarizes individual steps into ready-to-execute separate bash scripts and log files, which enables the user to reproduce individual steps during the analysis and debug errors (Figure 1). This differs from most automated pipelines and web-based software,<sup>25, 33, 36–37</sup> which does not support provenance tracking natively. Hence, to reproduce the analysis, the user must remember to note down



every step they run and capture the standard output. While QIIME2 possesses the most comprehensive provenance tracking mechanisms to date, it trails behind in terms of flexibility. In QIIME2, intermediate files generated between steps were stored in temp folders by default, and external commands invoked could not be re-run again, making debugging at user level almost impossible.<sup>21</sup> In contrast, Dix-seq preserves all the intermediate results and logs all the external commands, making debugging less frustrating to users. We hope these efforts, especially through sharing the parameter sheet file, would further pave the way for promoting reproducible studies in this field.

### Application to analysis and visualization of samples in natural environments

Case Study I featured a total of 48 surface water samples collected across four sediment zones, including two near (MG1 and MG2) and two outer (MG3 and MG4) of the Changjiang River estuary (Figure S1).<sup>57</sup> The basic and advanced pipelines of Dix-seq were performed, yielding the results of sequence processing, functional prediction, general ecological and statistical analysis, and visualization within 6,180.26 seconds on a workstation with Intel E5-2640 v3 CPU × 2 and 64 GB DDR4 memory. This suggested that Dix-seq is fairly fast and efficient. In case the user has limited access to computational resources, Dix-seq can be rapidly deployed on the cloud via Apptainer. The cost of running Dix-seq on public clouds is also minimal. At the time of the writing, a cloud instance with better specs (dedicated high-performance CPU, 32vcpus, 64GB memory, 500GB Nvme SSD, 9TB bandwidth limit) on Vultr (vultr.com) could be rented for just \$0.952 per hour, hence putting the estimated total cost of the analysis in below \$2.

Part of the visual outputs were shown (Figures 2-3 & S1-S5). The curves of rank-abundance and rarefaction both saturated along the increasing number of species and read depth (Figure S1), indicating sufficient sequencing depths. The alpha diversity of microbial species (ZOTUs) and function (KOs), including equitability/evenness, Richness, Chao1, Shannon, and reciprocal Simpson indices, were calculated and visualized (Figures 2A-B & S2), which indicated the disparities of microbial community diversity between zones. The contour maps of Shannon indices for species (Figure S2A) and function (Figure S2B) further revealed their spatial partitioning across the four zones. Statistical analysis of the difference in microbial diversity indices (Figure S2C) across four zones was accessed by Kruskal-Wallis test, which was significant between multiple zones. Principal coordinate analysis (PCoA) based on Bray-Curtis dissimilarity was then applied to visualize the variation in the microbial composition of ZOTUs (Figure 2C) and KOs (Figure 2D). The results showed the vast difference in community and function composition between zones, with the difference in community composition far exceeding function composition. Furthermore, Procrustes analysis was used to compare the correlation between the composition of microbial species and function (Figure S3), from which a significant correlation ( $P = 0.001$ ) was obtained. The relative abundance of phyla / proteobacterial classes microbiota in 48 samples could be clustered by the four zones (Figure 2E), indicating a more significant difference between zones than within zones. The Linear discriminant analysis (LDA) effect size (LEfSe) showed the discriminatory taxa of four zones (Figure S4A), and Dix-seq also calculated the LDA scores (Figure S4B). In addition, the differences of some selected taxa from LEfSe were tested by Kruskal-Wallis test in Dix-seq (Figure S4C). Pair-wise comparisons of four zones at ZOTU level were performed and volcano plots were created to visualize the results (Figures 3A & S5). The results from the differential abundance analysis conducted above revealed that compositional difference was driven by a wide range of microbes between zones. Finally, the top 12 pathway-level predicted functions (Figure 3B) in 48 samples were calculated and visualized using Dix-seq, and a clear enrichment in metabolic-related pathways was observed. Overall, the results revealed clear spatial partitioning of microbial community across zones, emphasizing Dix-seq's capabilities in extracting meaningful biological patterns even on complex environmental samples. Additionally, the related statistical tables and the plot codes were also exported in the same folder as those plots and the "run" folder, respectively. This saves the user valuable time should they wish to polish the auto-generated figures or conduct customized analysis. Moreover, Dix-seq also supports outputting a full report in HTML format using the "rmarkdown"

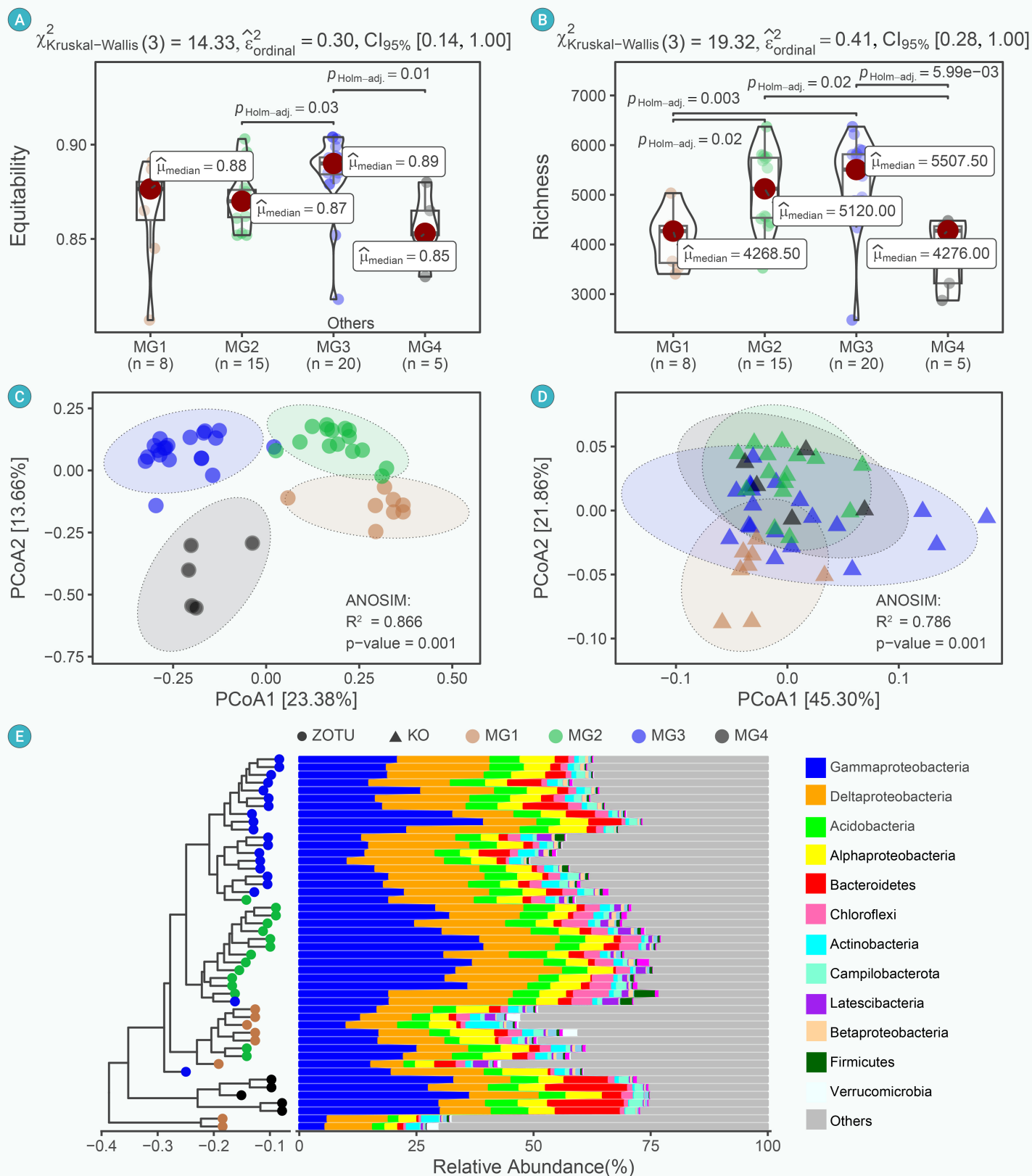
subprogram in "utils" module from Dix-seq. The report can then be opened in a web browser and viewed on practically any modern desktop or laptop. This would help anyone to quickly overview and assess the data and facilitate smoother collaborations among researchers since no technical knowledge of the software is required.

### Performance evaluation using samples from anthropically engineered ecosystems

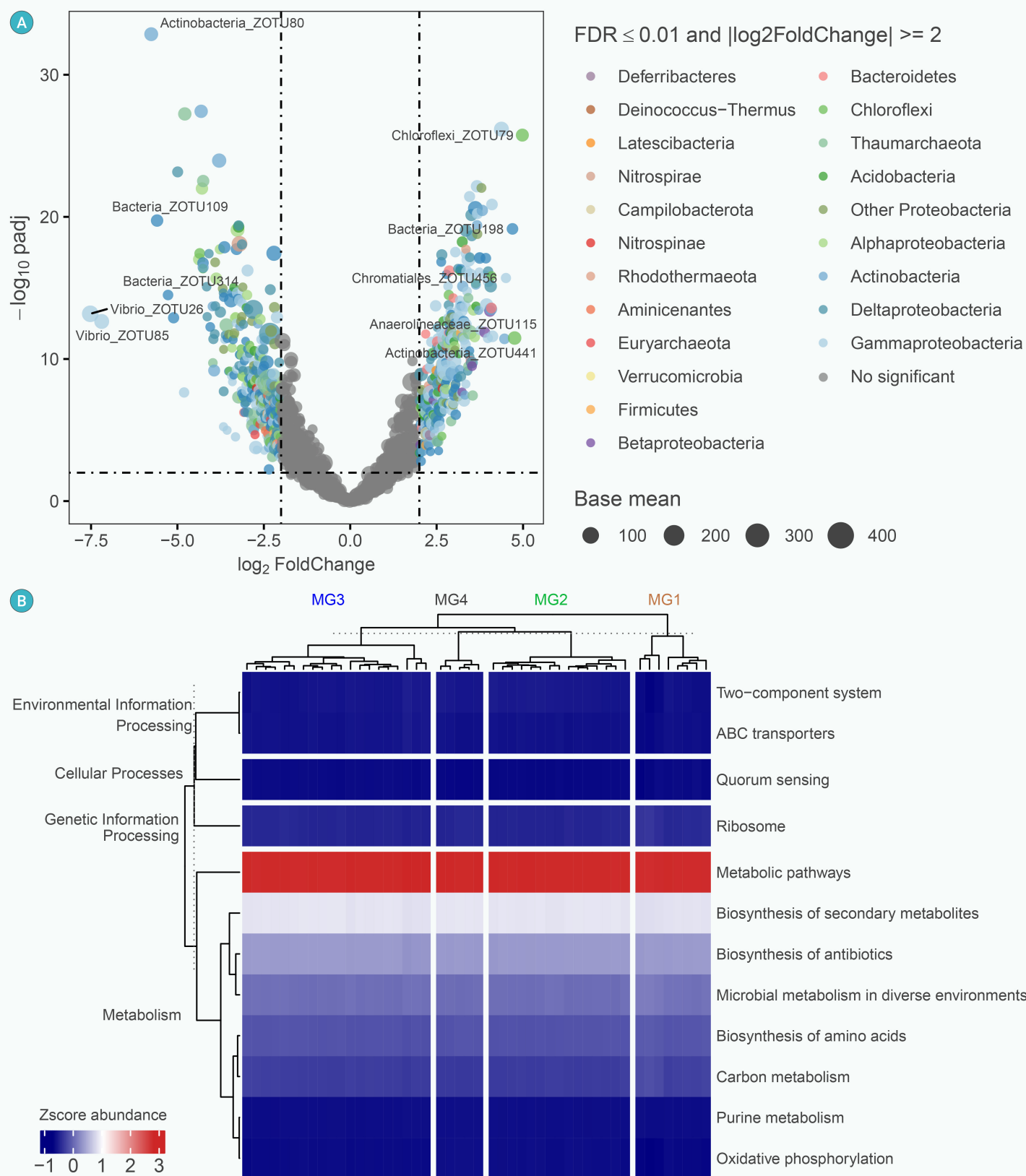
The performance of Dix-seq against currently adopted tools for analyzing complex microbial communities was also benchmarked by re-analyzing the dataset from a study featuring larval shrimp and wearing water samples along larval development stages with a fine temporal scale (Wang et al., Study II).<sup>54</sup> Specifically, the set of pipelines included Dix-seq, EasyAmplicon, QIIME2 with two different denoising methods (DADA2 and Deblur), and QIIME1 open-reference OTU picking strategies with 97% (S97%, also used in Wang et al.) and 99% (S97%) sequence similarity. Default parameters were used for all six pipelines (details in supplemental methods). The performance was evaluated based on six criteria: (1) elapsed time in each step of pipelines, (2) "mappability" of representative sequences for raw data, (3) recovery of the representative sequences in Study II against those in tested tools, (4) probabilities of feature abundance distribution, (5) percentage of unclassified feature (relative abundance), and (6) difference in taxonomy classifications between Study II and tested tools. Among the pipelines listed above, Dix-seq performed the best (ranked 1st) within all these aspects, except for the probability distribution of features and percentage of unclassified features (Figures 4A-F). Dix-seq classified more features at higher taxa ranks (from phylum to order) than all its counterparts, while at family and genus level it still achieves comparable performance similar to those in QIIME2 (Figure 4E). The high percentage of unclassified features at lower ranks is an inevitable limitation of annotations by generic taxonomic databases, especially in uncommon or rare habitats with abundant unknown microbes, such as the mariculture environment in Study II. The taxonomic profiler in Dix-seq defaults to a strict confidence threshold (details in supplemental methods). While no definitive confidence threshold settings were reported largely due to the species coverage and diversity in the taxonomic database, higher values generally reduce false positives at the cost of increasing the ratio of unclassified taxa in the lower levels (e.g., genus).<sup>19,58</sup> Therefore, a higher threshold was generally recommended when using common taxonomic databases for taxonomic annotation.<sup>58</sup> Furthermore, the higher average abundance of features was found in Dix-seq with a more even distribution pattern than other pipelines, except for QIIME2 with Deblur (Figure 4D). The criteria of mappability and number of extracted ASVs in Deblur were both the lowest than those in other pipelines and showed significant differences with the previous report in Study II (Richness, in Figure 4G), although QIIME2 with Deblur performed better than Dix-seq in criteria of unclassified feature ratio and distribution of features. When denoising sequences and calling ASVs, Deblur only supported merged and globally trimmed reads as input and filtered away most reads compared to other pipelines.<sup>15</sup> Consequently, the number of extracted ASVs was drastically reduced and the average abundance of ASV was greatly inflated. This trend was observed in other studies, too.<sup>42,43</sup> In summary, those evaluations demonstrated that Dix-seq performed better than other pipelines for complex mariculture samples due to the high read recovery and effective taxonomic profiler supported by USEARCH.

### Dix-seq is robust and enhances explained variance in complex community

One major challenge of amplicon data analysis is the low consistency between results generated from different pipelines, especially between OTU-based and ASV-based ones. Even different parameter settings of the same tool could yield highly inconsistent results.<sup>44,58</sup> In general, calling features is a key step in those pipelines that significantly affected the analysis results.<sup>7</sup> However, it is challenging to find the perfect balance between increasing the sensitivity of denoisers and error-correcting models and minimizing the impacts of sequencing errors and chimeras formed during amplification.<sup>12</sup> This challenge is further compounded by the growing size of sequence data, especially in building denoising models based on sub-sampled reads from the entire dataset.<sup>14,41,59</sup> Therefore, the methods in calling feature (ASVs /



**Figure 2. Analysis and visualization results of Study I (environmental microbiome) reveal differentiation of microbial species (ZOTUs) and function (KOs) across estuary zones by using one-step workflows "pipeline, advanced" in dix-seq (A-B) Violin boxplots showcasing the differences and probability distribution of microbial alpha diversity indices including equitability/evenness (A), and richness (B) across four zones. Statistical differences were calculated using Kruskal-Wallis tests, and the adjusted  $p$  values were calculated using the "holm" method. The bold line and red point within boxes represent medians. The whiskers in the top and bottom of the box represent 1.5 times the IQR, respectively. The upper and lower whiskers extend to data no more than  $1.5 \times$  the interquartile range from the upper and lower edge of the box, respectively. (C and D) Principal Coordinate Analysis (PCoA) based on Bray-Curtis dissimilarity visualizing the variation in the ZOTU (C) and KO (D) compositions of 48 samples across four marine zones. The variations of microbial and functional composition were corroborated by similarity analysis (ANOSIM) with significant distinction across four geographic zones. (E) UPGMA phylogenetic tree based on weighted UniFrac distance combined with species distribution at phyla / proteobacterial classes level. The numbers of samples within different coastal zones are as follows: in MG1 (n=8), MG2 (n=15), MG3 (n=20), and MG4 (n = 5). MG1-4 were located in the Changjiang River plume and the outer margin of this estuary, and all the sample sites were represented in Figure S1.**



**Figure 3. Examples of publication-quality visualization results of Study I by using step-by-step mode with "DESeq2, picrust2, keeg" subcommand combination in dix-seq (A)** Comparison of the difference ZOTUs between MG2 and MG3 groups. The diameter of the circle is proportional to the mean abundance of a given bacterial ZOTU. The ZOTUs with  $|\log_2(\text{FoldChange})| > 2$  and false discovery rate (FDR)  $\leq 0.01$  were selected as different ZOTUs and plotted with different colors based on their taxonomy. (B) The relative abundance of the top 12 pathway-level predicted functions across four marine zones. The abundance of KEGG level 2 category was normalized by calling third-party "PICRUST2" and scaled with Z-score in dix-seq.

OTUs) by using clustering and denoising approaches were not perfect. Hence, it is necessary to try multiple pipelines to cross-validate the results.

Major inconsistencies were observed in taxonomic composition and alpha diversity, which were similar to other metabarcoding studies evaluating the



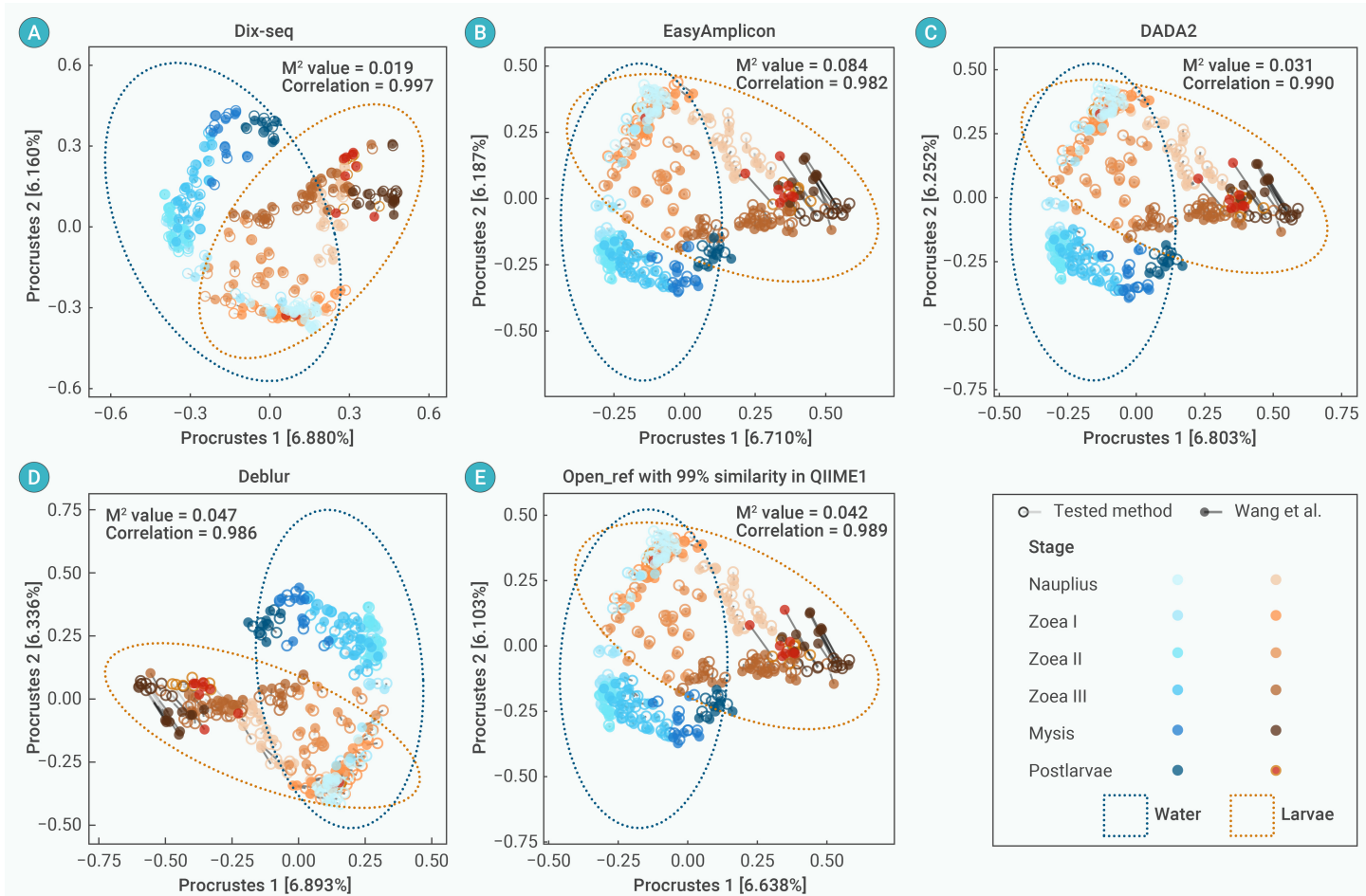


**Figure 4. Comparison of runtime, taxonomic annotations, and alpha diversity indexes on the data set of Study II (Wang et al.) using alternative protocols** (A) Comparison of the elapsed time of recommended protocols for microbiome analysis with six methods. The protocols were further separated into six steps: 1. read preparation, 2. calling ASVs / ZOTUs, 3. feature table construction, 4. Annotation, 5. beta diversity, 6. alpha diversity. The detail statistical method of running time was in supplemental information. (B and C) Comparison of sequence availability by using alternative protocols. The mappability was calculated by the percentage of raw reads retained in the unfiltered feature table for each sample (B). The representative sequences previously reported in Study II were aligned against those in other protocols to evaluate the mapping recovery (C). The bold line and the whiskers of the boxplots represent the median and 1.5 times the IQR, respectively. (D) Ridge plot revealing the probability distribution of ZOTUs/ASVs abundance in the feature tables for six protocols. The vertical lines represent the median (center line), upper and lower quartiles, respectively. (E and F) Comparison of taxonomic annotation by using alternative protocols. The mean abundance of ZOTUs/ASVs in percentages with unclassified taxonomic units at different ranks was calculated for six taxonomic profilers (E). The differences of taxonomic units at five taxa ranks were compared between the previous report of Study II and taxonomic results in other protocols (F). (G) Violin plots showcasing the differences and probability distribution of microbial alpha diversity between the results of Study II (Wang et al.) and the other five protocols. The asterisks indicate statistical significance between Wang et al. and the other five protocols (\*,  $P < 0.05$ , and \*\*,  $P < 0.01$ , Kruskal-Wallis test). The bold points within violin plots represent medians. The top and bottom of vertical lines represent upper and lower quartiles, respectively. (C, F, G) Comparison of consistency between the previous report (Study II) and the results of the other five protocols with representative sequences, taxonomy, and alpha diversity, respectively. EasyAmp is short for EasyAmplicon.

different methods of clustering or denoising sequences.<sup>42,44,58</sup>

It is interesting to note that taxonomic information annotated by Dix-seq was much closer to the original results reported in Wang et al. than other pipelines (Figure 4F). A trend that consistently occurred in the changes in the percentage of unclassified features, taxonomy assignment differences, and Evenness diversity, where no significant difference was found between the

values calculated by Dix-seq and Wang et al., while the results from other pipelines obviously differed from Wang et al. (Figure 4G). The Procrustes analysis showed that community composition in Study II was less vulnerable to the difference imposed by the tested pipelines (Figure 5), and all the patterns from the tested pipeline were highly consistent with the previous report of Wang et al. This may be caused by the strong effects of environ-



**Figure 5. Procrustes analysis showed the correlation between the composition previously reported in Study II (Wang et al.) and the results of alternative protocols** Procrustes analysis based on Bray-Curtis dissimilarity matrices calculated with the ZOTUs / ASVs table. Samples (dots) are colored by shrimp developmental stages and habitats of microbial communities. Correlation and  $M^2$  represented Procrustes correlation and Sum of Squares between two matrices. The closer the  $M^2$  value gets to 0, the higher the consistency between the results of Wang et al. and the tested protocols. The significance of Procrustes statistic was estimated with 999 permutations, and all  $P$  values < 0.01.

mental factors.<sup>46</sup> Still, when compared to other pipelines, community compositions were the most similar between those calculated from Dix-seq and Wang et al. (Figure 5). These analyses suggested that Dix-seq could better capture community characteristics from the dataset in Study II than other tested pipelines. Meanwhile, Dix-seq packed robust denoising algorithms that isolated sequencing noise, resulting in a reduction of false-positive assignments compared to the QIIME1 pipeline used in Wang et al.

Dix-seq also supports the calculation of essential indicators that reflect the community composition, which was used to extract biologically meaningful patterns in Study II (Figure 6). Statistical analysis was performed to evaluate the variance caused by different pipelines on the interpretation of driving factors (such as microbe habitats and shrimp developmental stages) of community composition (Figure 6). Considering the interference among environmental factors on community composition, different types of factors were analyzed separately with the same statistical methods.<sup>60,61</sup> The results showed that regardless of the Bray-Curtis or weighted UniFrac distance, the analysis by using Dix-seq pipelines explained a greater proportion of variance (Figure 6A). Permutational multivariate analysis of variance (PERMANOVA) of the beta diversity distance matrices suggested that both methods were able to clearly differentiate samples by microbe habitats and shrimp developmental stages ( $P = 0.001$ ), with the Dix-seq pipeline generating more statistical difference than its counterparts, except for QIIME1 pipeline with 99% similarity (Figure 6B). The distinction by developmental stages was less pronounced. Nevertheless, the Dix-seq pipeline was able to explain greater variance explanation of habitat-stage interaction in microbial communities ( $F_{\text{weighted UniFrac}} = 17.31$ ,  $F_{\text{Bray-Curtis}} = 23.01$ ,  $P < 0.01$ ). The relative distance based on beta diversity distance matrices from any sample with the

Nauplius stage to other samples belonging to the Nauplius stage versus that of other developmental stages (ratio of the Nauplius stage vs. others) was also calculated. The results showed that this distance was significantly shorter with the Dix-seq pipeline than with other pipelines, suggesting that Dix-seq was more effective at distinguishing subgroups with shared properties (Figure 6C).

#### Dix-seq remains effective at low sequencing depths.

The efficacy of the Dix-seq pipelines in analyzing amplicon data with varying sequencing depths was assessed by re-analyzing sub-sampled sequences along a gradient of decreasing read depths from the data set of Study II. The performance was evaluated based on eight criteria: (1) mappability of representative sequences for raw data, (2) recovery of the representative sequences from Study II against those from Dix-seq, (3) relative abundance of unclassified features, (4) difference in taxonomy assignments between Study II and tested tools, (5-6) Evenness and phylogenetic diversity, (7) the correlation between the original feature table and each of the subsampled tables, and (8) PERMANOVA pseudo- $F$  statistics. Near the sequencing depths of 11,000 (unit: read pairs, same below), mappability and recovery remained above 90% and 95%, respectively (Figures 7A-B). Similar trends were observed in the changes in the percentage of unclassified features, taxonomy assignment differences, and evenness diversity. Those indices tend to stabilize with the value of the original results (before subsample) when sequencing depths increased to 11,000 (Figures 7C-E). While the change of phylogenetic diversity steadily declined all the way down to the sequencing depth of 15,000 (Figure 7F). A Pearson's  $r$  of  $0.982 \pm 0.0032$  (mean and SD, same below) was retained even at the sequencing depth of



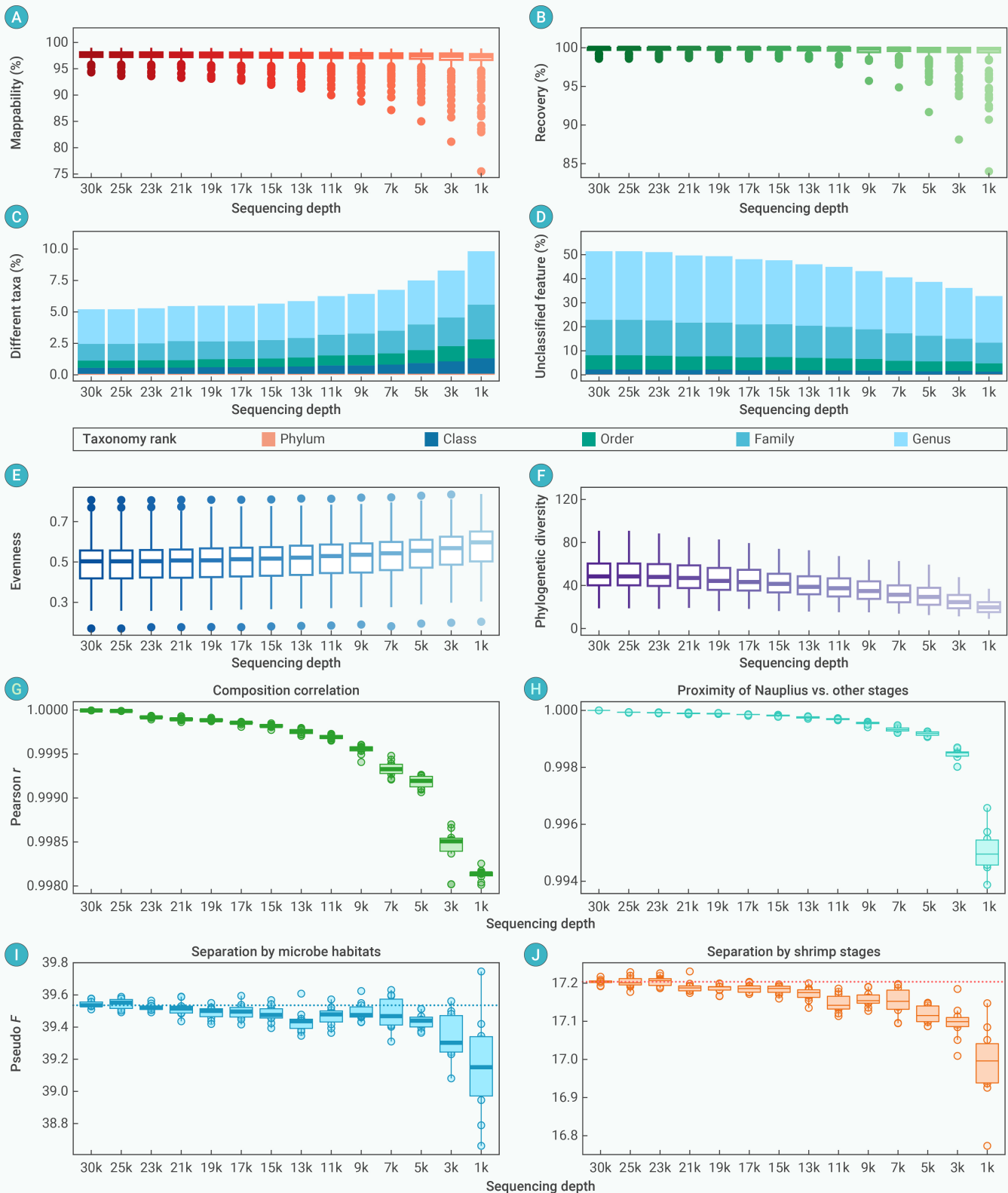
**Figure 6. Comparison of beta diversity analysis results on the data set of Study II using alternative protocols** (A) Proportions of variance explained by the first three axes of PCoA in different habitats of microbial communities (left panel: only in larvae, middle: only in water, and right: the total microbial community). (B) PERMANOVA pseudo-*F* statistics for the separation between samples from microbial habitats in larvae and water (left panel), by different larvae stages (middle), and their interaction (right). (C) Mean ratios of the distances from any sample with Nauplius stage to other samples belonging to Nauplius stage versus to that of other developmental stages. The Mean ratios were also calculated in different habitats (left panel: only in larvae, middle: only in water, and right: the total microbial community). The bold line and the whiskers represent the median and 1.5 times the IQR, respectively. The closer the ratio gets to 1, the lower resolution falls between samples of Nauplius and other stages for the tested protocols. EasyAmp is short for EasyAmplicon.

1,000 (Figure 7G). The Nauplius-versus-other relative distance retained above a Pearson's  $r$  of  $0.995 \pm 0.0013$  when sampling depth dropped to 1000 (Figure 7H). The PERMANOVA pseudo-*F* statistics steadily declined with the reduction at sequencing depth, where results close to the original were observed down to the depth of 11,000 (Figures 7I-J). These findings suggested that the results generated by the Dix-seq remained highly consistent with varying read depth of samples. It should be noted that sufficient sequencing depths are still required to better capture the taxonomy and phylogenetic diversity, especially when the samples originate from highly diverse environmental sources.<sup>42,46</sup>

## CONCLUSION

The rapid adoption of amplicon metagenome sequencing necessitated the development of computational tools to harness its full potential. Dix-seq, a fast and versatile tool, offers a one-stop solution for comprehensive ampli-

con data analysis. Dix-seq was designed with reduced command line complexity, which was achieved by consolidating all parameters into a single configuration file. It features a fully automated pipeline, which turns raw sequence data into integrated statistical and visualization results in one step without human intervention during the whole process, ideal for biologists with limited microbiome data analysis experience. For experienced analysts, Dix-seq offers debug modes, flexibility, and customizability due to its modular design. It integrates third-party procedures for functional prediction and variance analysis, expanding capabilities and enabling workflow automation. For complex microbial communities, Dix-seq could better capture the variance even at low sequencing depths. Future plans include the deployment of a web-based GUI for ease of use in Supercomputing Network (<https://www.scnet.cn>) and Bio-OS volcengine (<https://www.volcengine.com/>) and the incorporation of more algorithms to support data from third-generation sequencing platforms.



**Figure 7. Impact of sequencing depth on the result of the Dix-seq software** (A-B) The relationship between sequence availability of Dix-seq and the subsampled sequencing depth. (C and D) The relationship between taxonomic annotation and the read depth when using the Dix-seq software. The methods of calculating mappability (A), mapping recovery (B), the different percentages between tested pipelines and Study II of taxonomic units (C), and the mean percentage of ZOTUs with unclassified taxonomic units (D) were similar in Figure 4. (E and F) The relationship between alpha diversity including evenness (E), and phylogenetic diversity (F) and the read depth with using Dix-seq software. To ensure visual resolution of phylogenetic diversity (F), the outliers of the box were not visible in the range of the y-axis. (G) Pearson correlation coefficient ( $r$ ) between the composition of the original ZOTU table and the subsampled tables. (H) Pearson's  $r$  between the ratios of the distances from any sample in Nauplius stage to other samples belonging to Nauplius stage versus to that of other developmental stages. (I and J) PERMANOVA pseudo- $F$  statistics on Separation by microbe habitats (I) and shrimp stages (J). The orange dashed lines indicate the statistics calculated in the previous report of Study II. The bold line and the whiskers in boxplots represent the median and 1.5 times the IQR, respectively. The results presented in panels E to J were calculated from 10 replicates of random subsampling at each depth. The original ZOTU table generated from 7 million paired-end reads of Study II data set was randomly subsampled to each of the sampling depths indicated at the x-axis, which is equivalent to the number of paired-end reads in the raw sequencing data (see supplementary methods).



## REFERENCES

- Turnbaugh P., Ley R., Hamady M., et al. (2007). The Human Microbiome Project. *Nature* **449**:804–810 DOI:10.1038/nature06244
- Gilbert J.A., Jansson J.K. and Knight R. (2014). The Earth Microbiome Project: Successes and aspirations. *BMC Biol.* **12**:69 DOI:10.1186/s12915-014-0069-1
- Blaser M., Bork P., Fraser C. et al. (2013). The microbiome explored: recent insights and future challenges. *Nat. Rev. Microbiol.* **11**:213–217 DOI:10.1038/nrmicro2973
- Ruppert K.M., Kline R.J. and Rahman M.S. (2019). Past, present, and future perspectives of environmental DNA (eDNA). *metabarcoding: A systematic review in methods, monitoring, and applications of global eDNA. Glob. Ecol. Conserv.* **17**:e00547 DOI:10.1016/j.gecco.2019.e00547
- Zhou L. (2023). Microbial taxonomy with DNA sequence data as nomenclatural type: How far should we go. *The Innovation Life* **1**:100017 DOI:10.59717/j.xinn-life.2023.100017
- Jin J., Liu X. and Shiroguchi K. (2024). Long journey of 16S rRNA - amplicon sequencing toward cell - based functional bacterial microbiota characterization. *IMetaOmics* **1**:imo2.9 DOI:10.1002/imo2.9
- Zhang Z., Jiao N. and Zhang Y. (2024). Global ocean microbiome catalogue: A gateway to unlock marine microbial resources and ecosystem functions. *The Innovation Life* **2**:100099 DOI:10.59717/j.xinn-life.2024.100099
- Li C., Gillings M., Zhang C., et al. (2024). Ecology and risks of the global plastisphere as a newly expanding microbial habitat. *The Innovation* **5**:100543 DOI:10.1016/j.xinn.2023.100543
- Ewels P., Magnusson M., Lundin S., et al. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**:3047–3048 DOI:10.1093/bioinformatics/btw354
- Bolger A.M., Lohse M. and Usadel B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**:2114–2120 DOI:10.1093/bioinformatics/btu170
- Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**:10 DOI:10.14806/ej.17.1.200
- Magoč T. and Salzberg S.L. (2011). FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**:2957–2963 DOI:10.1093/bioinformatics/btr507
- Edgar R. (2016). UNOISE2: Improved error-correction for Illumina 16S and ITS amplicon sequencing. *BioRxiv*081257. DOI:10.1101/081257
- Callahan B.J., McMurdie P.J., Rosen M.J., et al. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**:581–583 DOI:10.1038/nmeth.3869
- Amir A., McDonald D., Navas-Molina J., et al. (2017). Deblur rapidly resolves single - nucleotide community sequence patterns. *MSystems* **2**:e00191–16 DOI:10.1128/mSystems.00191-16
- Edgar R. (2013). UPARSE: Highly accurate OTU sequences from microbial amplicon reads. *Nat. Methods* **10**(10):996–998 DOI:10.1038/nmeth.2604
- Edgar R. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**:2460–2461 DOI:10.1093/bioinformatics/btq461
- Kopylova E., Noé L. and Touzet H. (2012). SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**:3211–3217 DOI:10.1093/bioinformatics/bts611
- Edgar R. (2016). SINTAX: A simple non-Bayesian taxonomy classifier for 16S and ITS sequences. *BioRxiv*074161. DOI:10.1101/074161
- Wang Q., Garrity G.M., Tiedje J.M., et al. (2007). Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **73**:5261–5267 DOI:10.1128/AEM.00062-07
- Bolyen E., Rideout J.R., Dillon M.R., et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* **37**:852–857 DOI:10.1038/s41587-019-0209-9
- Wood D.E., Lu J. and Langmead B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**:257 DOI:10.1186/s13059-019-1891-0
- Dixon P. (2003). VEGAN, a package of R functions for community ecology. *J. Veg. Sci.* **14**:927–930 DOI:10.1111/j.1654-1103.2003.tb02228.x
- McMurdie P.J. and Holmes S. (2013). phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS ONE* **8**:e61217 DOI:10.1371/journal.pone.0061217
- Lu Y., Zhou G. and Ewald J., et al. (2023). MicrobiomeAnalyst 2.0: Comprehensive statistical, functional and integrative analysis of microbiome data. *Nucleic Acids Res.* **51**:W310–W318. DOI:10.1093/nar/gkad407
- Liu C., Cui Y., Li X. and Yao M. (2021). Microeco: An R package for data mining in microbial community ecology. *FEMS Microbiol. Ecol.* **97**:fiaa255 DOI:10.1093/femsec/fiaa255
- Liu B., Huang L., Liu Z., et al. (2022). EasyMicroPlot: An efficient and convenient R package in microbiome downstream analysis and visualization for clinical study. *Front. Genet.* **12**:803627 DOI:10.3389/fgene.2021.803627
- Wen T., Xie P., Yang S., et al. (2022). ggClusterNet: An R package for microbiome network analysis and modularity - based multiple network layouts. *IMeta* **1**:imt2.32 DOI:10.1002/imt2.32
- Schloss P.D., Westcott S.L., Ryabin T., et al. (2009). Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Appl. Environ. Microbiol.* **75**:7537–7541 DOI:10.1128/AEM.01541-09
- Rognes T., Flouri T., Nichols B., et al. (2016). VSEARCH: A versatile open source tool for metagenomics. *PeerJ.* **4**:e2584 DOI:10.7717/peerj.2584
- Caporaso J.G., Kuczynski J., Stombaugh J., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7**:335–6 DOI:10.1038/nmeth.f303
- Comeau A.M., Douglas G.M. and Langille, M.G.I. (2017). Microbiome Helper: A custom and streamlined workflow for microbiome research. *MSystems* **2**:e00127–16 DOI:10.1128/mSystems.00127-16
- Liu Y., Chen L., Ma T., et al. (2023). EasyAmplicon: An easy - to - use, open - source, reproducible, and community - based pipeline for amplicon data analysis in microbiome research. *IMeta* **2**:imt2.83 DOI:10.1002/imt2.83
- Weissbecker C., Schnabel B. and Heintz-Buschart A. (2020). Dadasnake, a Snakemake implementation of DADA2 to process amplicon sequencing data for microbial ecology. *Gigascience* **9**:giaa135 DOI:10.1093/gigascience/giaa135
- Weinstein M.M., Prem A., Jin M., et al. (2019). FIGARO: An efficient and objective tool for optimizing microbiome rRNA gene trimming parameters. *BioRxiv*:610394. DOI:10.1101/610394
- Gonzalez A., Navas-Molina J.A., Kosciolk T., et al. (2018). Qiita: Rapid, web-enabled microbiome meta-analysis. *Nat. Methods* **15**:796–798 DOI:10.1038/s41592-018-0141-9
- Keegan K.P., Glass E.M. and Meyer, F. (2016). MG-RAST, a metagenomics service for analysis of microbial community structure and function. *Methods Mol Biol.* **1399**:207–233 DOI:10.1007/978-1-4939-3369-3\_13
- Abueg L.A.L., Afgan E., Allart O., et al. (2024). The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update. *Nucleic Acids Res.* **52**(W1):W83–W94 DOI:10.1093/nar/gkae410
- Shi W., Qi H., Sun Q., et al. (2019). gcMeta: A Global Catalogue of Metagenomics platform to support the archiving, standardization and analysis of microbiome data. *Nucleic Acids Res.* **47**:D637–D648 DOI:10.1093/nar/gky1008
- Chen Y., Li J., Zhang Y., et al. (2022). Parallel - Meta Suite: Interactive and rapid microbiome data analysis on multiple platforms. *IMeta* **1**:imt2.1 DOI:10.1002/imt2.1
- Peng X. and Dorman K.S. (2021). AmpliCl: A high-resolution model-based approach for denoising Illumina amplicon data. *Bioinformatics* **36**:5151–5158 DOI:10.1093/bioinformatics/btaa648
- Nilsen T., Snipen L.-G., Angell, I.L., et al. (2024). Swarm and UNOISE outperform DADA2 and Deblur for denoising high-diversity marine seafloor samples. *ISME Communications* **4**:ycae071 DOI:10.1093/ismeco/ycae071
- Nearing J.T., Douglas G.M., Comeau A.M., et al. (2018). Denoising the Denoisers: An independent evaluation of microbiome sequence error-correction approaches. *PeerJ* **6**:e5364 DOI:10.7717/peerj.5364
- Chiarrello M., McCauley M., Villéger S., et al. (2022). Ranking the biases: The choice of OTUs vs. ASVs in 16S rRNA amplicon data analysis has stronger effects on diversity measures than rarefaction and OTU identity threshold. *PLOS ONE* **17**:e0264443. DOI:10.1371/journal.pone.0264443
- Bokulich N.A., Kaehler B.D., Rideout J.R., et al. (2018). Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. *Microbiome* **6**:90 DOI:10.1186/s40168-018-0470-z
- Edgar R.C. (2018). Accuracy of taxonomy prediction for 16S rRNA and fungal ITS sequences. *PeerJ* **6**:e4652 DOI:10.7717/peerj.4652
- Zhu Q., Huang S., Gonzalez A., et al. (2022). Phylogeny-aware analysis of metagenome community ecology based on matched reference genomes while bypassing taxonomy. *MSystems* **7**:e00167–22 DOI:10.1128/msystems.00167-22
- Douglas G.M., Maffei V.J., Zaneveld J.R., et al. (2020). PICRUST2 for prediction of metagenome functions. *Nat. Biotechnol.* **38**:685–688 DOI:10.1038/s41587-020-0548-6
- DeSantis T. Z., Hugenholtz P., Larsen N., et al. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* **72**:5069–5072 DOI:10.1128/AEM.03006-05
- Köljal U., Larsson K. and Abarenkov K. (2005). UNITE: A database providing web - based methods for the molecular identification of ectomycorrhizal fungi. *New Phytol.* **166**:1063–1068 DOI:10.1111/j.1469-8137.2005.01376.x
- Quast C., Priesse E. and Yilmaz P., et al. (2012). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**:D590–D596 DOI:10.1093/nar/gks1219
- Escapa F.I., Huang Y., Chen T., et al. (2020). Construction of habitat-specific training sets to achieve species-level assignment in 16S rRNA gene datasets. *Microbiome* **8**:65 DOI:10.1186/s40168-020-00841-w
- Dueholm M. K. D., Nierychlo M., Andersen K. S., et al. (2022). MiDAS 4: A global catalogue of full-length 16S rRNA gene sequences and taxonomy for studies of bacterial communities in wastewater treatment plants. *Nat. Commun.* **13**:1908 DOI:10.1038/s41467-022-29438-7
- Wang Y., Wang K., Huang L., et al. (2020). Fine-scale succession patterns and assembly mechanisms of bacterial community of *Litopenaeus vannamei* larvae across the developmental cycle. *Microbiome* **8**:106 DOI:10.1186/s40168-020-00879-w
- Love M.I., Huber W. and Anders S. (2014). Moderated estimation of fold change and

dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**:550 DOI:10.1186/s13059-014-0550-8

56. Khleborodova A., Gamboa-Tuz S.D., Ramos M., et al. (2024). Lefser: Implementation of metagenomic biomarker discovery tool, LEfSe, in R. *Bioinformatics* **40**. DOI: 10.1093/bioinformatics/btae707
57. Wei Y., Jiang S., Tian L., et al. (2022). Benthic microbial biogeography along the continental shelf shaped by substrates from the Changjiang River plume. *Acta Oceanol. Sin.* **41**:118–131 DOI:10.1007/s13131-021-1861-8
58. Nguyen N.-P., Warnow T., Pop M., et al. (2016). A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity. *Npj Biofilms Microbiomes* **2**:16004 DOI:10.1038/npjbiofilms.2016.4
59. Johnson J.S., Spakowicz D.J., Hong B.-Y., et al. (2019). Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat. Commun.* **10**:5029 DOI:10.1038/s41467-019-13036-1
60. Dong P., Guo H., Huang L., et al. (2023). Glucose addition improves the culture performance of Pacific white shrimp by regulating the assembly of Rhodobacteraceae taxa in gut bacterial community. *Aquaculture* **567**:739254 DOI:10.1016/j.aquaculture.2023.739254
61. Dong P., Guo H., Wang Y., et al. (2021). Gastrointestinal microbiota imbalance is triggered by the enrichment of *Vibrio* in subadult *Litopenaeus vannamei* with acute hepatopancreatic necrosis disease. *Aquaculture* **533**:736199 DOI:10.1016/j.aquaculture.2020.736199

## FUNDING AND ACKNOWLEDGMENTS

This study was financially supported by the National Natural Science Foundation of China (32303039 and 32111530179), State Key Laboratory for Managing Biotic and Chemical Threats to the Quality, Safety of Agro-products (2021DG700024-KF202425), Top-Notch Talents Program of Henan Agricultural University (30501466), Special Fund for Henan Agriculture Research System (HARS-22-13-G1 and HARS-22-16-T), and Henan Provincial Department of Science and Technology Research Project (242102111013). We gratefully thank the users (Yanting Wang, Chengjia Wu, Xiaoshuang Liu, Meina Zhang, Huidong Niu, Jiale Hong, Tong zhou, Hongjing Luo, HZYCSLKT, GGHTX, Clement Tsui, Genshen Wang, Lulu Zhu) in Biostack network forum for testing Dix-seq and editing its documentation.

## AUTHOR CONTRIBUTIONS

All authors contributed to the manuscript and approved the final version.

## DECLARATION OF INTERESTS

Lei Zhang is an employee of DeepBiome. Co. Ltd.

## ETHICAL STATEMENT AND PATIENT CONSENT

Not applicable.

## DATA AND CODE AVAILABILITY

The code of Dix-seq could be approached from GitHub (<https://github.com/jameslz/dix-seq>) and Gitee (<https://gitee.com/dpsh/dix-seq>). Dix-seq is also supported for rapid deployment in Unix-like systems such as MacOS and Linux and Windows Subsystem for Linux 2 (WSL 2) by using an Apptainer container (SIF file hosted at <https://doi.org/10.6084/m9.figshare.23257829>). The sequencing data of case study I and II used for performance evaluation are available in Sequence Read Archive of NCBI (<https://www.ncbi.nlm.nih.gov/sra>) with the accession number of PRJNA525225, and figshare (<https://doi.org/10.6084/m9.figshare.24906642>), respectively. All the codes for visualization and statistics of case study could be found on GitHub ([https://github.com/biodps/Performance\\_evaluation\\_of\\_Dix-seq](https://github.com/biodps/Performance_evaluation_of_Dix-seq)).

## SUPPLEMENTAL INFORMATION

It can be found online at <https://doi.org/10.59717/j.xinn-life.2024.100120>

## LEAD CONTACT WEBSITE

Pengsheng Dong:

<https://dwkjxy.henau.edu.cn/a/shiziduiwu/shuichan/2023/0328/4874.html>

Yongjun Wei:

[https://www7.zzu.edu.cn/synbio/info/1067/1061.htm?\\_refluxos=a10](https://www7.zzu.edu.cn/synbio/info/1067/1061.htm?_refluxos=a10)