frontiers | Frontiers in Microbiology

TYPE Technology and Code
PUBLISHED 07 March 2024
DOI 10.3389/fmicb.2024.1358257

OPEN ACCESS

EDITED BY
Chen Li,
Northeastern University, China

REVIEWED BY
Bin Hu,
Los Alamos National Laboratory (DOE), United States
Xiaowen Feng,
Dana–Farber Cancer Institute, United States

*CORRESPONDENCE
Fang Liu
✉ liufcri@163.com
Weihua Pan
✉ panweihua@caas.cn

†These authors have contributed equally to this work

RECEIVED 20 December 2023
ACCEPTED 20 February 2024
PUBLISHED 07 March 2024

CITATION
Hui X, Yang J, Sun J, Liu F and Pan W (2024) MCSS: microbial community simulator based on structure.
Front. Microbiol. 15:1358257.
doi: 10.3389/fmicb.2024.1358257

MCSS: microbial community simulator based on structure

Xingqi Hui[1,2†], Jinbao Yang[2,3†], Jinhuan Sun[4], Fang Liu[1,5*] and Weihua Pan[2*]

[1]Zhengzhou Research Base, State Key Laboratory of Cotton Biology, School of Agricultural Sciences, Zhengzhou University, Zhengzhou, China, [2]Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences (ICR, CAAS), Shenzhen, China, [3]College of Informatics, Huazhong Agricultural University, Wuhan, China, [4]Key Laboratory of Plant Molecular Physiology, CAS Center for Excellence in Molecular Plant Sciences, Institute of Botany, Chinese Academy of Sciences, Beijing, China, [5]National Key Laboratory of Cotton Bio-Breeding and Integrated Utilization, Institute of Cotton Research, Chinese Academy of Agricultural Sciences (ICR, CAAS), Anyang, China

De novo assembly plays a pivotal role in metagenomic analysis, and the incorporation of third-generation sequencing technology can significantly improve the integrity and accuracy of assembly results. Recently, with advancements in sequencing technology (Hi-Fi, ultra-long), several long-read-based bioinformatic tools have been developed. However, the validation of the performance and reliability of these tools is a crucial concern. To address this gap, we present MCSS (microbial community simulator based on structure), which has the capability to generate simulated microbial community and sequencing datasets based on the structure attributes of real microbiome communities. The evaluation results indicate that it can generate simulated communities that exhibit both diversity and similarity to actual community structures. Additionally, MCSS generates synthetic PacBio Hi-Fi and Oxford Nanopore Technologies (ONT) long reads for the species within the simulated community. This innovative tool provides a valuable resource for benchmarking and refining metagenomic analysis methods.

Code available at: https://github.com/panlab-bio/mcss

KEYWORDS

metagenome, microbiome communities, long reads, simulator, assembly

1 Introduction

Metagenomic sequencing treats microbes in the environment as a unified entity to obtain genomic sequences, which can be used to study the taxonomic composition of microbial communities and identify novel species (Handelsman et al., 1998; Handelsman, 2004; Chen and Pachter, 2005; Frioux et al., 2020; Yang et al., 2021). And the assembly of metagenomic sequencing reads into metagenome-assembled genomes (MAGs) is a crucial step in the metagenomic analysis. Assembly tools, such as hifiasm_meta (Feng et al., 2022) and metaFlye (Kolmogorov et al., 2020) enhance contiguity in assemblies using nanopore (Wang et al., 2021) and PacBio (Rhoads and Au, 2015) sequencing data compared to short-read assemblies. And they can effectively address challenges of uneven species composition and intra-species heterogeneity in complex microbial communities (Marx, 2021; Bickhart et al., 2022). The development and testing of these metagenome assembly algorithms require high-quality benchmark datasets with ground truth, but obtaining ground truth for real datasets can

be challenging, making it difficult to assess the accuracy of the algorithms (Escalona et al., 2016; Zhao et al., 2017; Alosaimi et al., 2020). Therefore, the development of simulation software that can generate synthetic metagenomic data is highly meaningful.

So far, several simulation tools have been developed. Read simulators like Pbsim3 (Ono et al., 2022) and NanoSim (Yang et al., 2017) can generate simulated third-generation sequencing reads, which provide foundational data for benchmark testing. However, they cannot simulate metagenomic data. Meta-NanoSim (Yang et al., 2023) and CAMISIM (Fritz et al., 2019) can simulate metagenomic datasets but require users to provide additional information, such as a reference metagenome list or the composition of the microbial community. This requires users to have a prerequisite level of domain expertise, so in many cases, users may not be certain about the species composition of the microbial community they want to simulate. M&Ms García-García et al. (2022) can simulate datasets based on environmental parameters, allowing users to specify the environment they want to simulate and obtain simulated metagenomes. Since M&Ms acquire species within genera through random sampling, it does not consider the structural characteristics of communities at the species level and cannot learn from the characteristics of the real sequencing samples entered by the user. Additionally, the sequencing data simulated by M&Ms. is limited to 16S rDNA and cannot generate third-generation sequencing data for whole genomes.

Therefore, we have developed MCSS, which can simulate microbial communities and generate third-generation sequencing data. MCSS generates simulated data based on community structure at the species level, preserving the structural features of real samples while expanding the diversity of the simulated community. Moreover, MCSS can simulate both the abundance of species within the community and intra-species heterogeneity, which increases the complexity of the simulated data, making it more closely resemble real samples. Finally, the generated long reads can be used directly as input for assembly tools, greatly reducing the workload for users.

2 Materials and methods

MCSS generates simulated microbial communities and sequencing data by learning the structure, abundance, and intra-species heterogeneity information from real samples of microbial communities. MCSS primarily generates simulated data through the following four steps (Figure 1): (1) determine the species composition, (2) determine the abundance of each species in the community, (3) find the reference genomes of the species in the GTDB reference database (Parks et al., 2020), and (4) call Pbsim3 to generate simulated long reads.

The core function of step 1 is to determine the species composition of the simulated microbial community based on real samples. In the community, each species can be mapped to a corresponding taxonomic rank (domain, phylum, class, order, family, genus, and species) within the GTDB database. This taxonomic rank resembles a branch, and all the species' taxonomic ranks form a tree. High-level taxonomic units may have one or more subordinate low-level taxonomic units. Consequently, we have decided to represent the taxonomic profiles of microbial communities using a multiway tree structure, which is a data structure allowing multiple branches for each node (Figure 2). In our study, we use a multiway tree to represent the structural characteristics of a microbial community and construct a multiway tree for each real sample's community. Then, we construct a multiway tree based on all species in the GTDB reference database using the same method, to serve as a reference multiway tree. For each simulation, we sample a multiway tree from real samples, and then identify the optimal subtree of the sampled multiway tree within the reference multiway tree, representing the species composition of the simulated community. We have pre-generated community multiway trees for multiple samples under various environmental conditions (Table 1), which serve as the basis for sampling and creating the sampled multiway trees. Furthermore, MCSS can produce sampled multiway trees using user-input sequencing data.

In the second step, the abundance of each species in the community is determined by sampling based on species abundance observed in real samples. Because not all community species abundance distributions can be fitted with appropriate models, sampling from real samples is widely applicable across various environments.

In the third step, MCSS searches the GTDB reference database for the genomes of each species in the community. If the user specifies the number of strains within each species, the tool will search for that specified quantity of genomes for each species to reflect the diversity between species.

In the final step, the user needs to specify the minimum depth of coverage or average depth of coverage, as well as the sequencing model for Pbsim3. MCSS then calls Pbsim3 to generate simulated sequencing data. By using simulated community genome list as input, Pbsim3 simulation allows the generation of synthetic metagenomic sequencing data. Pbsim3 can generate both high-accuracy Hi-Fi reads and ultra-long ONT reads, with sequencing costs higher than those associated with second-generation sequencing reads. In addition, we provide the function to exclusively generate simulated community genome data. Users can choose a sequence simulation tool that suits their research to generate sequencing reads for simulated genomes.

2.1 Determine the species composition

2.1.1 Data sources

We downloaded sequencing data for six environmental conditions (Table 1) from the MGnify (Richardson et al., 2023) and NCBI SRA (Sayers et al., 2023) database, which include the gut (PRJNA398089) (Zhang et al., 2022), marine (PRJNA329908) (Tremblay et al., 2017), oral (PRJNA362687) (Qiao et al., 2018), rhizosphere (PRJEB23682) (Maarastawi et al., 2018), skin (PRJEB26427) (Lam et al., 2018), and soil (PRJNA252425) (Větrovský et al., 2020). For each environment, we downloaded multiple samples where, each set of sequencing data represents a single sample.

2.1.2 Real and reference multiway tree

We assign GTDB taxonomic labels (Youngblut and Ley, 2021) to sequencing reads with Kraken2 (Wood et al., 2019) (GTDB_release207) and retrieve the taxonomic information (kingdom, phylum, class, order, family, genus, and species) of each species in the samples from the GTDB reference database for constructing the multiway tree of the microbial community structure (Figure 2). Each node in the multiway tree (T_real) represents a taxonomic unit. The edge lengths are determined by calculating the evolutionary distance in the GTDB reference tree, representing the evolutionary distance between two taxonomic units. Construct a multiway tree based on all

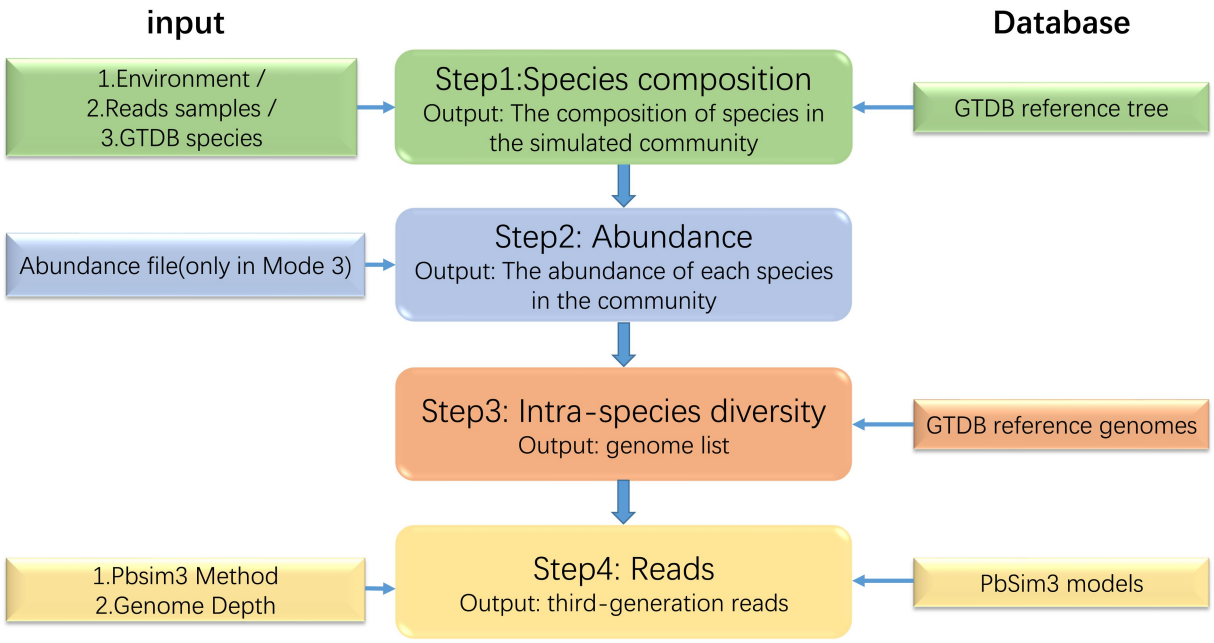

FIGURE 1
Flow diagram of the MCSS workflow.

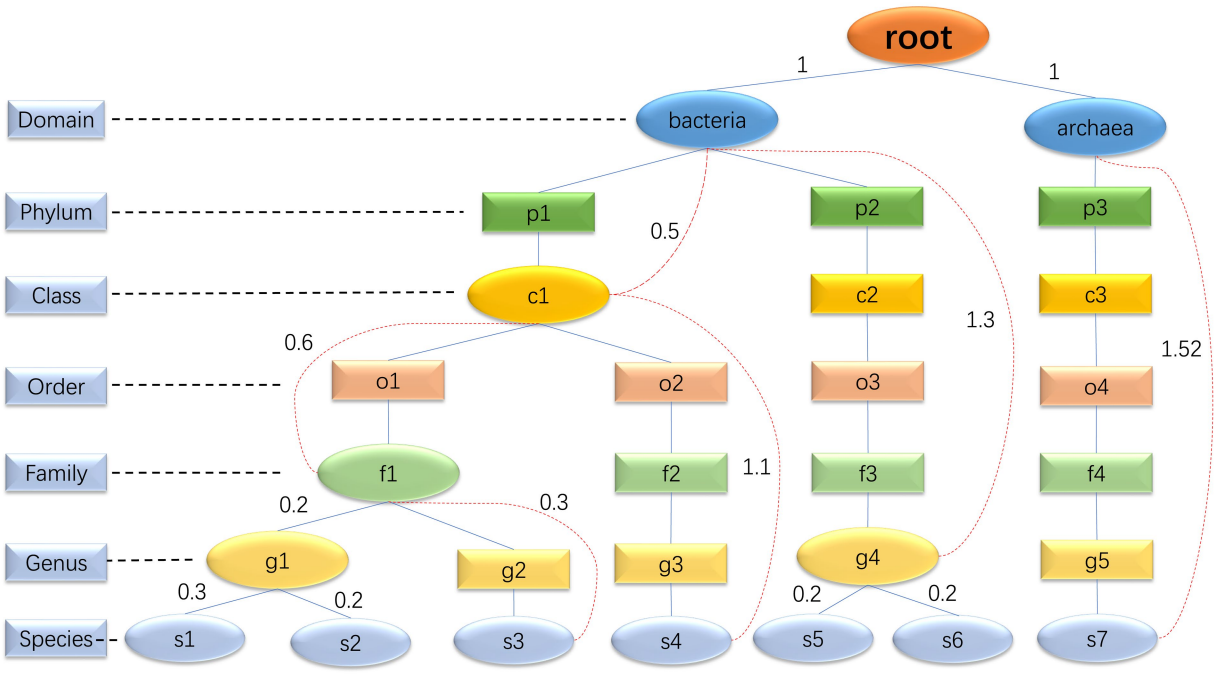

FIGURE 2
Multiway trees: the diagram is an illustration of a microbial community multiway tree, where oval nodes represent entity nodes, and rectangular nodes represent non-entity nodes. The evolutionary distance between entity nodes can be calculated, but it is not possible to calculate the evolutionary distance between entity nodes and non-entity nodes, or between non-entity nodes themselves. For example, the genus g2 only has one species, s3. Although g2 can be mapped to a specific taxonomic unit in the GTDB database, the information in the GTDB database about this genus corresponds to s3 rather than the genus itself. Consequently, it is not possible to calculate the evolutionary distance from other nodes to g2. Hence, we consider g2 as a non-entity node. All domains and species are entity nodes. For other taxonomic hierarchy levels, a node is considered an entity node if it has two or more directly adjacent child nodes; otherwise, it is considered a non-entity node. The red dashed lines are used to connect a parent entity node and its non-directly adjacent child node to record the evolutionary distance between these two nodes. The evolutionary distance between two directly adjacent entity nodes is directly represented by the edge length.

TABLE 1 The sources and quantities of actual samples from different environments.

Environment	Project	Sample count
Gut	PRJNA398089	104
Soil	PRJNA252425	118
Oral	PRJNA362687	111
Skin	PRJEB26427	102
Marine	PRJNA329908	123
Rhizosphere	PRJEB23682	120

species in the GTDB reference database using the same method, to serve as a reference multiway tree.

2.1.3 Simulated multiway tree

We randomly sample the multiway trees of environmental samples to obtain the sampled multiway tree. In the reference multiway tree, use Algorithm 1 to identify the subtree that closely resembles the sampled multiway tree. The essence of Algorithm 1 lies in finding a subtree in the reference multiway tree such that the difference in evolutionary distance between this subtree and the sampled multiway tree is minimized, which can be addressed using recursion. The species within this subtree constitute the simulated microbial community. To make MCSS more practical, users can specify two search modes to find subtrees, which are the accurate mode and the prolific mode. In the accurate mode, the search in reference multiway tree is based solely on sampled multiway tree, while in the prolific mode, adjustments are made using the mean and standard deviation of evolutionary distances for species in the sample to expand the simulated data. The process of calculating evolutionary distance is described by the following formula (formulas (1), (2), (3), (4), and (5)):

$$p = B(1, 0.5) \tag{1}$$

$$\sigma = std(dreal) \tag{2}$$

$$u = mean(dreal) \tag{3}$$

$$dnref = \{ y \in dref \mid u - \sigma \le y \le u + \sigma \} \tag{4}$$

$$dsam_g = \begin{cases} dsam_{gs} \sim dreal_g, s \in sam_g; & acc\ mode \\ dsam_{gs} \sim \begin{cases} dreal_g, s \in sam_g, p = 1 \\ dnref_g, s \in sam_g, p = 0 \end{cases}; & prol\ mode \end{cases} \tag{5}$$

where $B(1, 0.5)$ represents a Bernoulli distribution with parameters 1 and 0.5; $dref$, $dsam$, and $dreal$ respectively represent the evolutionary distances of species within the GTDB database, sampled data, and real samples; $dnref$ is the result of filtering $dref$ based on $dreal$; $dreal_g$ and $dnref_g$ represent the evolutionary distances of species within the genus g; sam_g represents the set of species within genus g in the sampled data.

When the user provides FastQ files, Kraken2 is used to assign GTDB taxonomic labels to the reads in the FastQ files, and then real multiway trees are constructed.

ALGORITHM 1 Get_SubTree (T_sample, T_ref): Recursively search for the closest subtree.

```
Input: sampled multiway tree T_sample, reference multiway tree T_ref
Output: simulated multiway tree T_sim
If T_sample has two layers:
    if prolific mode:
        adjust T_sample
    calculate evolutionary distance difference between species in T_sample and T_ref
    return dis, node_list
Else:
    min_dis = INF
    choice_node = []
    for child_node_T_sample in T_sample.child_nodes
        for child_node_Tref in in T_ref.child_nodes
            dis_tree, node_list = get_subTree(T_sample_child, T_ref_child)
            if min_dis > dis_tree and child_node_Tref not chosen
                min_dis = dis_tree
                choice_node.
                append(node_list)
    return min_dis, choice_node
```

2.2 Determine the abundance of each species in the community

For a specific environment, we analyze and record statistics on the abundance of species in each sample based on the Kraken2 results. Then, based on the number of species in the simulated multiway tree, we sample and normalize to obtain the abundance of microbial community species. Since the normalization process can affect the abundances of sampled species, we ensure that the sum of the sampled results approximates 100%, mitigating the impact of normalization on abundances.

When the input file consists of FastQ reads, we fit the species abundance distribution using a log-normal distribution and then sample to obtain the community species abundance.

Users can either specify both the species in a community and their respective abundances or select a pre-learned environment to generate species abundances based on that environment.

2.3 Find the reference genomes of the species

We downloaded multiple genomes for each species based on the correspondence between species and accession numbers in the GTDB database. These genomes are used to represent different strains within the same species. For each species in the previously determined simulated community, we randomly select the user-specified number of strain genomes. We select multiple reference genomes for a species because real samples often exhibit genetic variation within the same species. To capture these intra-species differences, we consider all

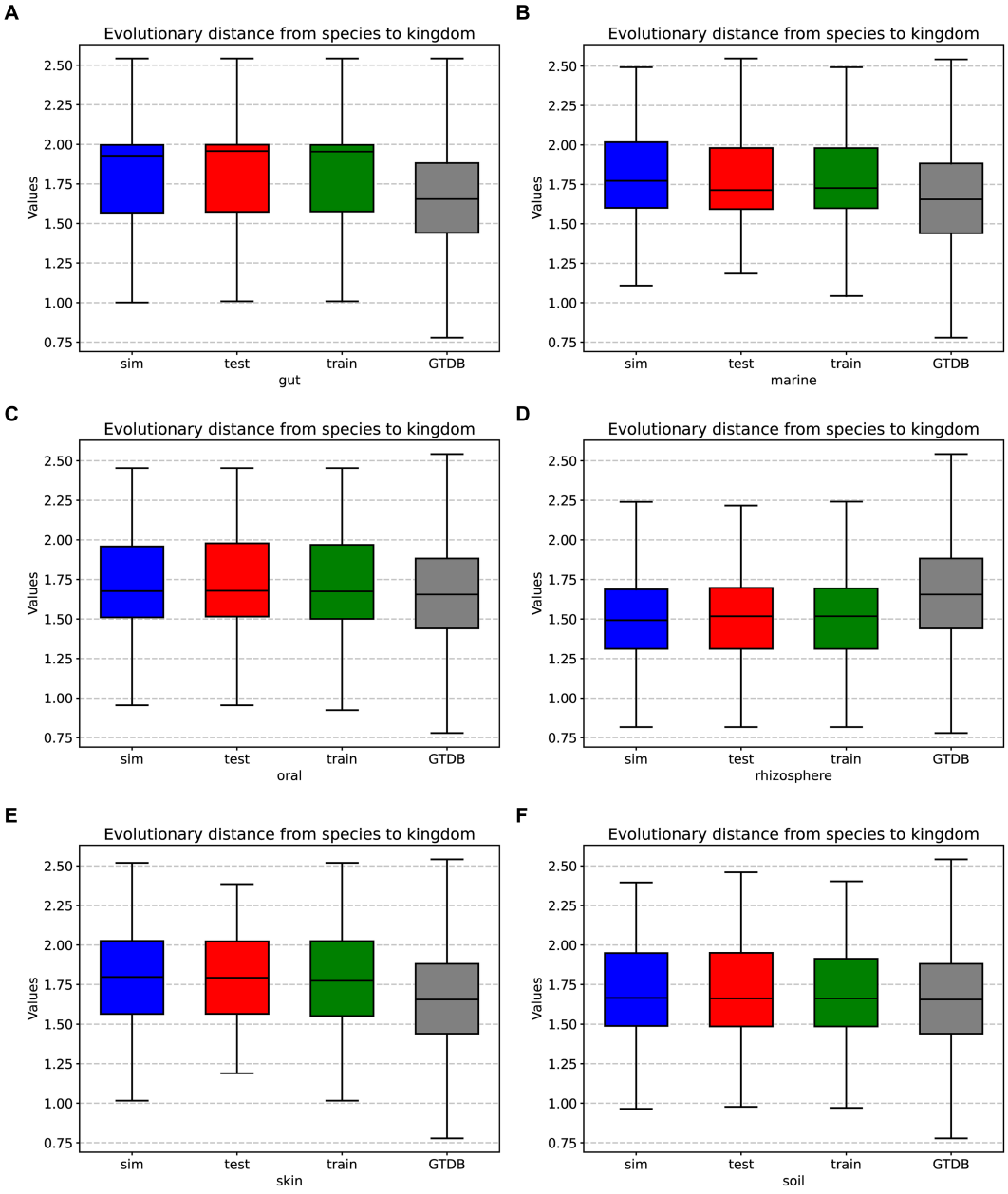

FIGURE 3
Box plots displaying the evolutionary distances from species to kingdom in the training set, test set, simulated data from different environments, and the GTDB reference database: **(A)** gut: 30 test samples and 74 training samples, **(B)** marine: 30 test samples and 93 training samples, **(C)** oral: 30 test samples and 81 training samples, **(D)** rhizosphere: 30 test samples and 90 training samples, **(E)** skin: 30 test samples and 72 training samples, **(F)** soil: 30 test samples and 88 training samples.

distinct genomes classified under the same species when selecting species genomes. This approach ensures that the simulated genome dataset incorporates internal variations within species, facilitating an effective evaluation of the performance of metagenomic tools in handling highly similar genomes.

2.4 Generate simulated long reads

Since Pbsim3 (using the qshmm model by default and other error models can also be chosen) generates PacBio continuous long reads (CLR), we employ SAMtools (Danecek et al., 2021) to convert the SAM format data produced by Pbsim3 into BAM format data. Subsequently, we utilize CCS (Wenger et al., 2019) to generate PacBio high-fidelity (Hi-Fi) reads.

3 Results

To ascertain the extent to which the simulated tree accurately represents the structural characteristics of actual microbial communities, we randomly choose 30 samples for the test set and use

the remaining ones for the training set and executed the subsequent validation procedure.

3.1 Consistency in the structural characteristics

In prolific mode, we generated 30 simulated samples using the features learned from the training dataset. We analyzed the evolutionary distances from species to kingdom in the training set, test set, simulated data, and the GTDB reference database. The species in the environment are a subset of the species in the GTDB database, reflecting the community characteristics of that environment. The community simulation process involves searching for species in the overall GTDB database that match the environmental characteristics. Figures 3, 4 show that species evolutionary distance distributions in different environments have distinct features, and they are highly consistent between the simulated and test data in each environment. The results indicate that MCSS has captured the feature of species evolutionary distances within the community from the environment.

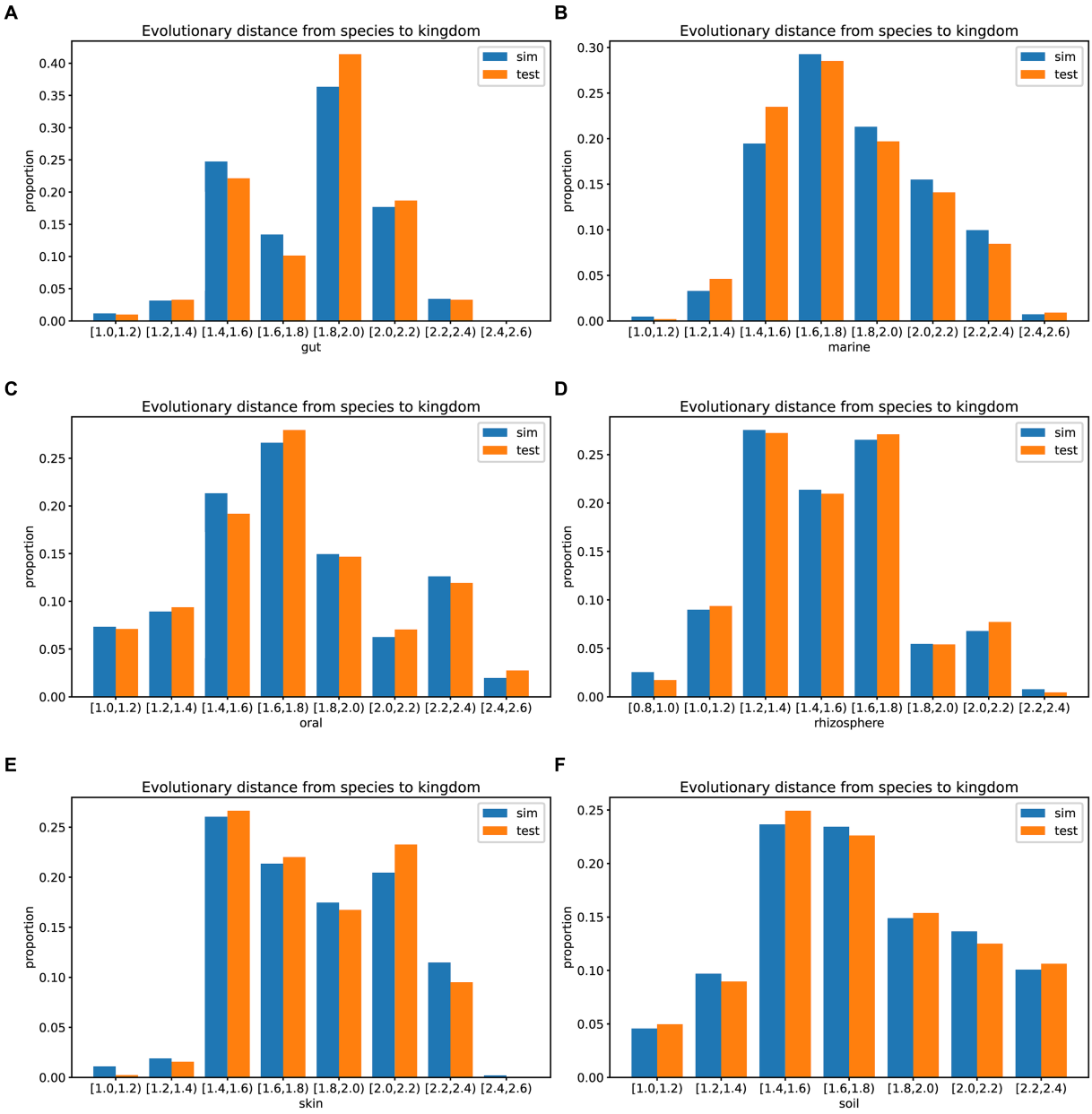

FIGURE 4
Histograms showing the evolutionary distances from species to kingdom in the training set, test set, simulated data from different environments, and the GTDB reference database: **(A)** gut: 30 test samples and 74 training samples, **(B)** marine: 30 test samples and 93 training samples, **(C)** oral: 30 test samples and 81 training samples, **(D)** rhizosphere: 30 test samples and 90 training samples, **(E)** skin: 30 test samples and 72 training samples, **(F)** soil: 30 test samples and 88 training samples.

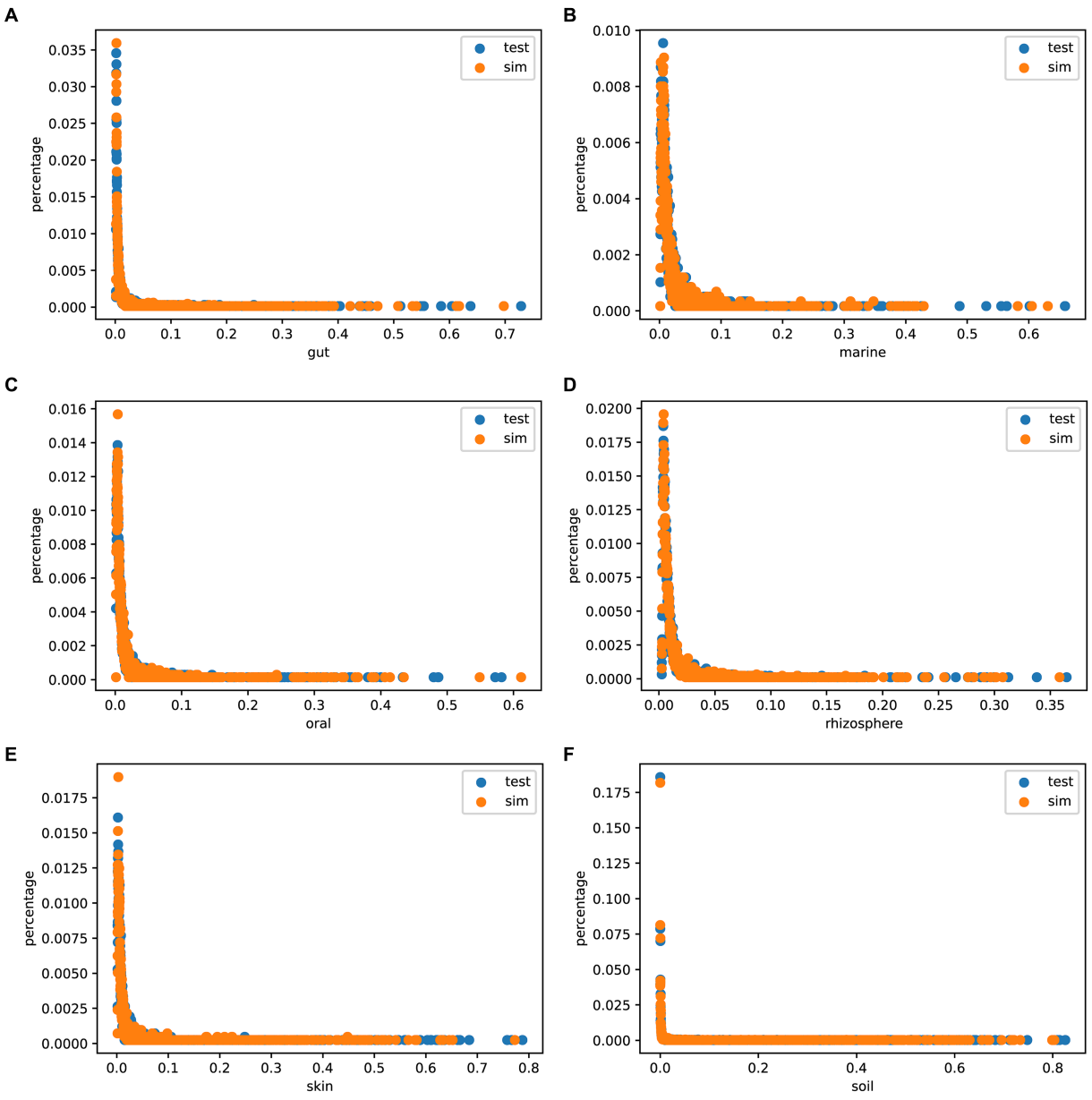

FIGURE 5
Species abundance scatter plot. The x-axis ($0 < x < 1$) represents species abundance, and the y-axis ($0 < y < 1$) represents the proportion of that abundance appearing in the samples: **(A)** gut: 104 samples, **(B)** marine: 123 samples, **(C)** oral: 111 samples, **(D)** rhizosphere: 120 samples, **(E)** skin: 102 samples, and **(F)** soil: 118 samples.

3.2 Species abundance in different environments

The Species abundance is a crucial metric of microbial communities. To assess the authenticity of the species abundance generated by MCSS, we compared the abundance distribution between the simulated data and the test data, and plotted scatter diagrams. Figure 5 shows that the sampled species abundances closely match the species abundances in the real samples. In environments like marine, oral, and rhizosphere, outliers are noticeable, and sampling from real samples can capture this feature, while obtaining species abundance from a distribution function fails to capture these characteristics. These results indicate that MCSS has the capability to generate relatively realistic species abundance data.

3.3 The consistency and diversity of the species composition

We examined the species composition of real samples and simulated samples under user-input sample patterns. Selecting 30 samples from each environment, we generate simulated data for each sample, and then compare the species composition between

10.3389/fmicb.2024.1358257

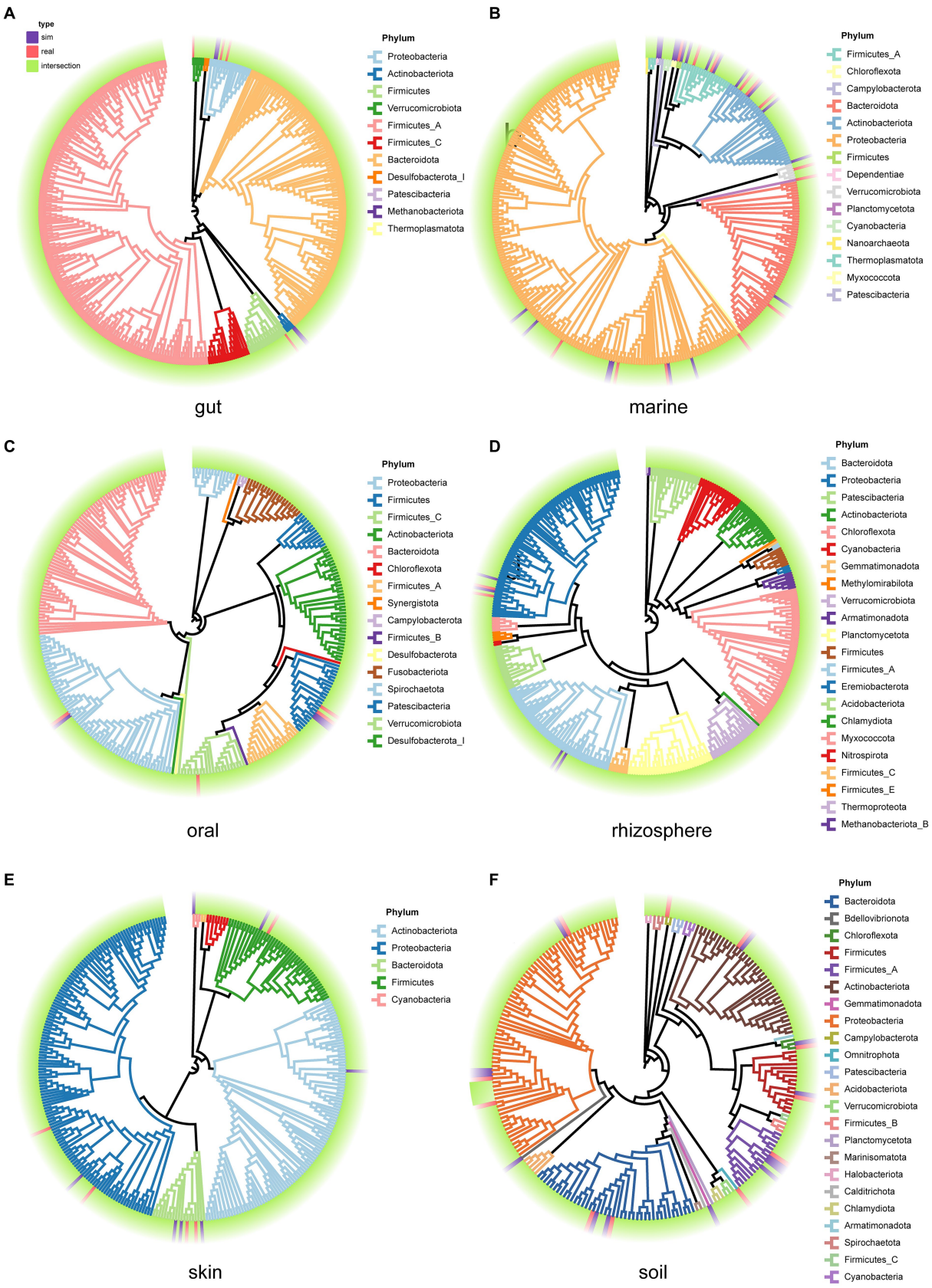

FIGURE 6
Phylogenetic trees for the species in both the real samples and the simulated samples from each environment. Purple blocks represent species that appear only in the simulated samples, red blocks represent species that appear only in the real samples, and green blocks represent species that are present in both real and simulated samples. The branch colors represent the phylum categories: **(A)** gut, **(B)** marine, **(C)** oral, **(D)** rhizosphere, **(E)** skin, and **(F)** soil.

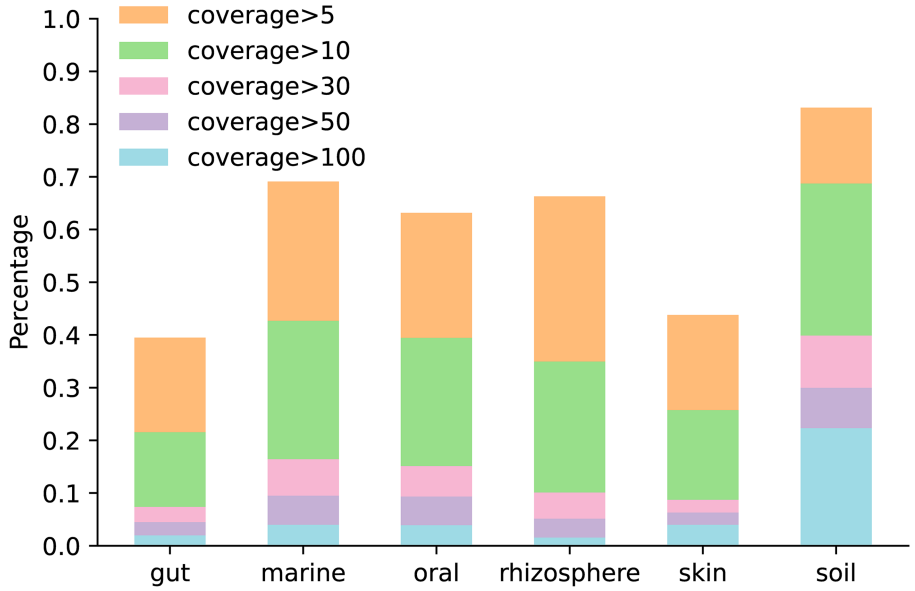

FIGURE 7
Proportion of genome coverage at various levels in different environments: the bar chart in orange and beneath represents the number of genomes with coverage greater than 5. The bar chart in green and beneath represents the number of genomes with coverage greater than 10. The bar chart in pink and beneath represents the number of genomes with coverage greater than 30. The bar chart in purple and beneath represents the number of genomes with coverage greater than 50. The bar chart in blue represents the number of genomes with coverage greater than 100.

the 30 real samples and the simulated samples. This process is used to evaluate the performance of MCSS in generating simulated data based on user-input samples. Compare the phylogenetic trees (Xie et al., 2023) for the species in both the real and simulated samples from each environment (Figure 6). In each environment, a high degree of overlap between the species in the simulated samples and the species in the real samples is evident. Meanwhile, there is a slight difference in the species composition between the simulated communities and the real datasets, suggesting MCSS can capture real community species composition characteristics while introducing diversity.

3.4 The assembly results of both real and simulated data

We generated simulated data based on SRR15275210 (Kim et al., 2022), assembled both real and simulated data separately, and analyzed the results. Despite the reduced read count in the simulated data (which can be increased by adjusting coverage), the outcomes of high-quality contigs do not differ significantly compared to real data. This is particularly evident in the number of contigs exceeding 1 M (see Table 2).

TABLE 2 The assembly results of both real and simulated data.

Result	Real	Sim
Reads	15,240,116,452	6,707,226,299
Species	160	160
Contigs with a size >500 K	244	162
Contigs with a size >1 M	128	118

3.5 Coverage of genomes in different environments

The coverage of the genome is a critical metric, which influences the quality of the assembly. To analyze the genome coverage in simulated data generated by MCSS, we generated five simulated datasets for each environment under default parameters. Figure 7 illustrates the proportion of genome coverage at various levels in different environments.

3.6 Genome divergence between strains of species in each environment

To quantify the genetic variation between species, we used mash (Ondov et al., 2016) to obtain the genome divergence of strains in simulated data across different environments. Figure 8 displays variations in the genomic differences among strains of species across different environments.

4 Discussion and conclusions

MCSS is a convenient and versatile metagenomic community simulation software that can generate diverse simulated data while

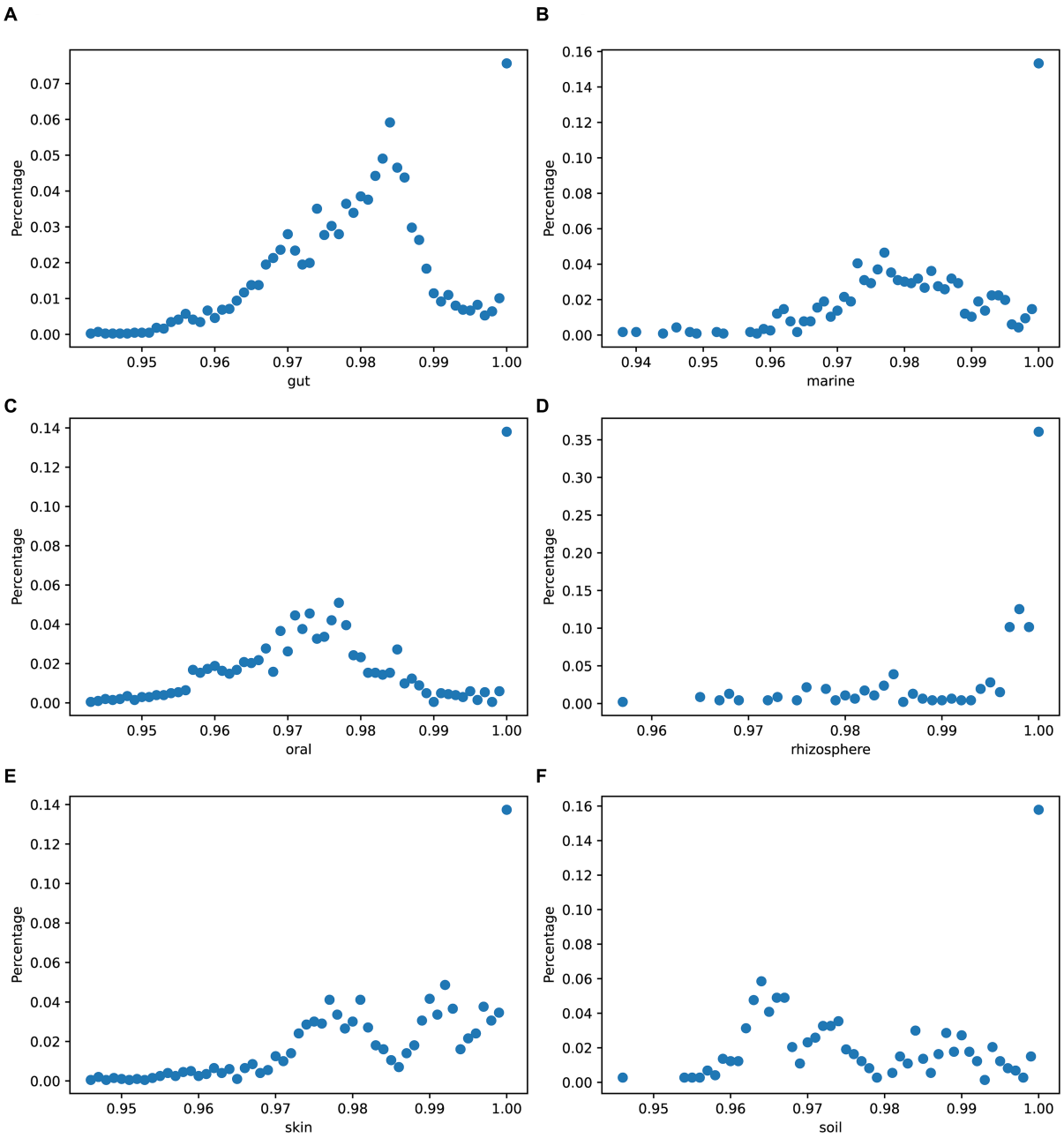

FIGURE 8
Genome divergence of strains in simulated data. The *x*-axis values represent the numbers obtained by subtracting mash results from 1: **(A)** gut, **(B)** marine, **(C)** oral, **(D)** rhizosphere, **(E)** skin, and **(F)** soil.

ensuring community similarity. MCSS can generate a simulated microbiome based on environmental parameters, learn from user-input sequencing data features, and allow users to specify the microbiome composition. Furthermore, it can simulate the species composition, species abundance, and intra-species heterogeneity of microbiomes, making the simulated communities closely resemble real metagenomic communities. In addition, the generated third-generation sequencing data increases its utility. The mentioned features allow it to cater to various cases of datasets to meet the evaluation needs of metagenomic assembly and analysis tools to help relevant researchers improve their software or algorithms.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

Author contributions

XH: Data curation, Methodology, Software, Writing – original draft, Writing – review & editing. JY: Methodology, Writing – original draft, Writing – review & editing. JS: Data curation, Writing – review

& editing. FL: Methodology, Supervision, Writing – review & editing. WP: Funding acquisition, Methodology, Project administration, Supervision, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work has been supported by the Innovation Program of Chinese Academy of Agricultural Sciences, Shenzhen Science and Technology Program (grant no. RCBS20210609103819020) and the National Natural Science Foundation of China (grant no. 32100501).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Alosaimi, S., Bandiang, A., van Biljon, N., Awany, D., Thami, P. K., Tchamga, M. S. S., et al. (2020). A broad survey of DNA sequence data simulation tools. *Brief. Funct. Genomics* 19, 49–59. doi: 10.1093/bfgp/elz033

Bickhart, D. M., Kolmogorov, M., Tseng, E., Portik, D. M., Korobeynikov, A., Tolstoganov, I., et al. (2022). Generating lineage-resolved, complete metagenome-assembled genomes from complex microbial communities. *Nat. Biotechnol.* 40, 711–719. doi: 10.1038/s41587-021-01130-z

Chen, K., and Pachter, L. (2005). Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLoS Comput. Biol.* 1:e24. doi: 10.1371/journal.pcbi.0010024

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., et al. (2021). Twelve years of SAMtools and BCFtools. *GigaScience* 10:giab008. doi: 10.1093/gigascience/giab008

Escalona, M., Rocha, S., and Posada, D. (2016). A comparison of tools for the simulation of genomic next-generation sequencing data. *Nat. Rev. Genet.* 17, 459–469. doi: 10.1038/nrg.2016.57

Feng, X., Cheng, H., Portik, D., and Li, H. (2022). Metagenome assembly of high-fidelity long reads with hifiasm-meta. *Nat. Methods* 19, 671–674. doi: 10.1038/s41592-022-01478-3

Frioux, C., Singh, D., Korcsmaros, T., and Hildebrand, F. (2020). From bag-of-genes to bag-of-genomes: metabolic modelling of communities in the era of metagenome-assembled genomes. *Comput. Struct. Biotechnol. J.* 18, 1722–1734. doi: 10.1016/j.csbj.2020.06.028

Fritz, A., Hofmann, P., Majda, S., Dahms, E., Dröge, J., Fiedler, J., et al. (2019). CAMISIM: simulating metagenomes and microbial communities. *Microbiome* 7:17. doi: 10.1186/s40168-019-0633-6

García-García, N., Tamames, J., and Puente-Sánchez, F. (2022). M & Ms: a versatile software for building microbial mock communities. *Bioinformatics* 38, 2057–2059. doi: 10.1093/bioinformatics/btab882

Handelsman, J. (2004). Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.* 68, 669–685. doi: 10.1128/MMBR.68.4.669-685.2004

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., and Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem. Biol.* 5, R245–R249. doi: 10.1016/S1074-5521(98)90108-9

Kim, C. Y., Ma, J., and Lee, I. (2022). Hi-Fi metagenomic sequencing enables assembly of accurate and complete genomes from human gut microbiota. *Nat. Commun.* 13:6367. doi: 10.1038/s41467-022-34149-0

Kolmogorov, M., Bickhart, D. M., Behsaz, B., Gurevich, A., Rayko, M., Shin, S. B., et al. (2020). metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat. Methods* 17, 1103–1110. doi: 10.1038/s41592-020-00971-x

Lam, T. H., Verzotto, D., Brahma, P., Ng, A. H. Q., Hu, P., Schnell, D., et al. (2018). Understanding the microbial basis of body odor in pre-pubescent children and teenagers. *Microbiome* 6:213. doi: 10.1186/s40168-018-0588-z

Maarastawi, S. A., Frindte, K., Linnartz, M., and Knief, C. (2018). Crop rotation and straw application impact microbial communities in Italian and Philippine soils and the rhizosphere of *Zea mays*. *Front. Microbiol.* 9:1295. doi: 10.3389/fmicb.2018.01295

Marx, V. (2021). Long road to long-read assembly. *Nat. Methods* 18, 125–129. doi: 10.1038/s41592-021-01057-y

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., et al. (2016). Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* 17:132. doi: 10.1186/s13059-016-0997-x

Ono, Y., Hamada, M., and Asai, K. (2022). PBSIM3: a simulator for all types of PacBio and ONT long reads. *NAR Genom. Bioinform.* 4:lqac092. doi: 10.1093/nargab/lqac092

Parks, D. H., Chuvochina, M., Chaumeil, P.-A., Rinke, C., Mussig, A. J., and Hugenholtz, P. (2020). A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat. Biotechnol.* 38, 1079–1086. doi: 10.1038/s41587-020-0501-8

Qiao, Y., Wu, M., Feng, Y., Zhou, Z., Chen, L., and Chen, F. (2018). Alterations of oral microbiota distinguish children with autism spectrum disorders from healthy controls. *Sci. Rep.* 8:1597. doi: 10.1038/s41598-018-19982-y

Rhoads, A., and Au, K. F. (2015). PacBio sequencing and its applications. *Genom. Proteom. Bioinform.* 13, 278–289. doi: 10.1016/j.gpb.2015.08.002

Richardson, L., Allen, B., Baldi, G., Beracochea, M., Bileschi, M. L., Burdett, T., et al. (2023). MGnify: the microbiome sequence data analysis resource in 2023. *Nucleic Acids Res.* 51, D753–D759. doi: 10.1093/nar/gkac1080

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., et al. (2023). Database resources of the National Center for Biotechnology Information in 2023. *Nucleic Acids Res.* 51, D29–D38. doi: 10.1093/nar/gkac1032

Tremblay, J., Yergeau, E., Fortin, N., Cobanli, S., Elias, M., King, T. L., et al. (2017). Chemical dispersants enhance the activity of oil-and gas condensate-degrading marine bacteria. *ISME J.* 11, 2793–2808. doi: 10.1038/ismej.2017.129

Větrovský, T., Morais, D., Kohout, P., Lepinay, C., Algora, C., Awokunle Hollá, S., et al. (2020). Global fungi, a global database of fungal occurrences from high-throughput-sequencing metabarcoding studies. *Sci. Data* 7:228. doi: 10.1038/s41597-020-0567-7

Wang, Y., Zhao, Y., Bollas, A., Wang, Y., and Au, K. F. (2021). Nanopore sequencing technology, bioinformatics and applications. *Nat. Biotechnol.* 39, 1348–1365. doi: 10.1038/s41587-021-01108-x

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P.-C., Hall, R. J., Concepcion, G. T., et al. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* 37, 1155–1162. doi: 10.1038/s41587-019-0217-9

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

Xie, J., Chen, Y., Cai, G., Cai, R., Hu, Z., and Wang, H. (2023). Tree Visualization By One Table (tvBOT): a web application for visualizing, modifying and annotating phylogenetic trees. *Nucleic Acids Res.* 51, W587–W592. doi: 10.1093/nar/gkad359

Yang, C., Chowdhury, D., Zhang, Z., Cheung, W. K., Lu, A., Bian, Z., et al. (2021). A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. *Comput. Struct. Biotechnol. J.* 19, 6301–6314. doi: 10.1016/j.csbj.2021.11.028

Yang, C., Chu, J., Warren, R. L., and Birol, I. (2017). NanoSim: nanopore sequence read simulator based on statistical characterization. *GigaScience* 6:gix010. doi: 10.1093/gigascience/gix010

Yang, C., Lo, T., Nip, K. M., Hafezqorani, S., Warren, R. L., and Birol, I. (2023). Characterization and simulation of metagenomic nanopore sequencing data with Meta-NanoSim. *GigaScience* 12:giad013. doi: 10.1093/gigascience/giad013

Youngblut, N. D., and Ley, R. E. (2021). Struo2: efficient metagenome profiling database construction for ever-expanding microbial genome datasets. *PeerJ* 9:e12198. doi: 10.7717/peerj.12198

Zhang, Y., Bhosle, A., Bae, S., McIver, L. J., Pishchany, G., Accorsi, E. K., et al. (2022). Discovery of bioactive microbial gene products in inflammatory bowel disease. *Nature* 606, 754–760. doi: 10.1038/s41586-022-04648-7

Zhao, M., Liu, D., and Qu, H. (2017). Systematic review of next-generation sequencing simulators: computational tools, features and perspectives. *Brief. Funct. Genomics* 16, 121–128. doi: 10.1093/bfgp/elw012