Materials and methods

Preprocessing of microbial relationships

The MEFE algorithm measures the relationships among microbes, and forming by integrating their phylogenetic hierarchies, taxonomy similarity, and functional annotations from the reference sequences. The multidimensional relationships serve as the biological foundation for subsequent analyses.

1. Phylogeny similarity

The sequence similarity of microbes is determined by a sequence alignment similarity threshold. After global alignment of whole-genome or marker gene sequences (e.g., 16S rRNA) in the reference database, a specific similarity threshold (e.g. 97%) is set to define *"high phylogeny similarity"*. The biological basis of this threshold originates from the conserved characteristics of key genes in microbial evolution (e.g., the conventional standard for bacterial species demarcation).

2. Taxonomy similarity

Based on the taxonomic annotation system of the reference database, when two sequences have the same annotation at a specific taxonomic level (e.g. genus or species), they are defined as having *"taxonomy consistency"*. The reliability of this determination depends on the completeness of the database's taxonomic annotations. High-quality annotations can significantly reduce the proportion of unclassified sequences, thereby enhancing the accuracy of comparative analyses.

3. Functional homology

Functional homology is evaluated by the consistency of normalized functional features and quantified by Hierarchical Meta-Storms (HMS) [1], with 100% similarity defined as "*identical functions*". Functional features (e.g., KEGG Orthology or metabolic pathways) of metagenomic data can be directly derived from reference genomes (e.g., RefSeq [2]); and those of amplicon sequencing data (e.g., 16S rRNA gene) were inferred by PICRUSt2 [3] or Taxa4Fun [4] from full-length targeted genes.

Finally, for each reference sequence in the database, the set of sequences that have *"high phylogeny similarity"*, *"taxonomy consistency"*, and *"functions identity"* with that sequence is considered as the close neighbors, and the sequence similarities were set as the elastic weights."

16S rRNA data analysis pipeline

In this study, the reference framework of the MEFE algorithm is constructed based on the Greengenes2 database [5]. First, 331,269 16S rRNA backbone sequences were extracted from this database, and a maximum likelihood phylogenetic tree was constructed through MAFFT multiple alignment to ensure compatibility with the beta-diversity analysis module of the Parallel-Meta Suite (PMS) [6] software. The determination of phylogeny neighbor was carried out by VSEARCH [7] for full-length sequence global alignment, with a 97% similarity threshold set according to the conventional standard for bacterial species demarcation.

Taxonomy similarity analysis was based on the annotation system of Greengenes2, which has a genus-level annotation completeness of 90% and a species-level annotation completeness of 60%. In the experiment, only sequences annotated at least to the genus level were retained, and consistency in genus-level annotation was used as the criterion for high taxonomy similarity.

Functional homology was predicted by PICRUSt2 for KEGG Orthology functional profiles, and the HMS algorithm was used to calculate functional similarity. When the KO functional profile similarity between two sequences reached 100%, they were defined as having identical functions. Finally, sequences with high phylogeny similarity (⩾97% sequence similarity), high taxonomy similarity (genus-level consistency), and identical function (HMS similarity = 100%) were integrated to construct the approximate sequence group for each reference sequence.

Microbiome profiling

The microbiome datasets used in this study and their corresponding information are listed in **Table 1**. After extracting 16S rRNA gene sequences from the raw sequence data, the data was preprocessed using the PMS, which involved merging paired-end sequences and denoising the raw sequences. VSEARCH was then used to select OTUs based on sequence similarity of 0.99 from the Greengenes2 reference database. The relative abundance of microorganisms was normalized and corrected based on the 16S rRNA gene copy numbers.

Elastic biomarker selection

Initially, each microbial member in the processed abundance matrix is treated as an index, and its approximate neighbors are identified from the preprocessed reference database (the approximate neighbors must be present in the abundance matrix). Considering that microorganisms may be misclassified due to sequencing errors, algorithmic mistakes, and other factors, while also recognizing the influence of approximate neighbors on the microorganism, we adjust the abundance of the microorganism in a flexible manner.

$$\omega_{ij} = S_{ij} - S_i^{max} \tag{1}$$

$$Nabd_i = \frac{e^{\omega_{ij}}}{\sum_{k=1}^{n} e^{\omega_{ijk}}} \tag{2}$$

We quantify the similarity relationship between microorganisms based on sequence similarity. Since the sequence similarity between approximate neighbors is greater than 97%, the approximate neighbors that exist in the abundance matrix are first smoothed using the softmax function. As shown in *Eq1*, S_{ij} represents the sequence similarity between microorganism *i* and its *j*-th approximate neighbor, and it is the highest similarity value between microorganism *i* and all of its neighboring sequences. The difference between the two gives the weight of the *j*-th approximate neighbor of microorganism *i*, and the weighted average of relative abundance is calculated using *Eq2*.

$$Abd_i^* = \alpha \times Abd_i + (1 - \alpha) \times Nabd_i \quad (3)$$

The processed abundance is then obtained as the final abundance through the weighted summation method in *Eq3*. Here, α serves as the weight of the microorganism itself, and *1-*α is the weight for each of its approximate neighbors. The final weighted average is computed, representing the final abundance value for microorganism *i*.

Then, we filtered out features that do not meet the criterion of "relative abundance > 0.001% and present in ⩾10% of samples" to reduce the interference of sequencing noise and data sparsity.

To highlight the superiority of this algorithm rather than the advantages of various complex models, we used only the Wilcoxon rank-sum test to select microbes with significant differences between groups (*p*-value < 0.01) as the final biomarkers. Finally, the obtained biomarkers were used as features, and the prediction results were obtained using the random forest model and leave-one-out cross-validation.

References

1 Zhang Y, Jing G, Chen Y, Li J, Su X. Hierarchical Meta-Storms enables comprehensive and rapid comparison of microbiome functional profiles on a large scale using hierarchical dissimilarity metrics and parallel computing. Bioinformatics advances, 2021, 1(1): vbab003

2 Zhang W, Fan X, Shi H, Li J, Zhang M, Zhao J, Su X J M S. Comprehensive assessment of 16S rRNA gene amplicon sequencing for microbiome profiling across multiple habitats. 2023, 11(3): e00563-00523

3 Douglas G M, Maffei V J, Zaneveld J R, Yurgel S N, Brown J R, Taylor C M, Huttenhower C, Langille M G I. PICRUSt2 for prediction of metagenome functions. Nature Biotechnology, 2020, 38(6): 685-688

4 Aßhauer K P, Wemheuer B, Daniel R, Meinicke P J B. Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. 2015, 31(17): 2882-2884

5 McDonald D, Jiang Y Y, Balaban M, Cantrell K, Zhu Q Y, Gonzalez A, Morton J T, Nicolaou G, Parks D H, Karst S M, Albertsen M, Hugenholtz P, DeSantis T, Song S J, Bartko A, Havulinna A S, Jousilahti P, Cheng S, Inouye M, Niiranen T, Jain M, Salomaa V, Lahti L, Mirarab S, Knight R. Greengenes2 unifies microbial data in a single reference tree. Nature Biotechnology, 2024, 42(5): 715-+

6 Chen Y Z, Li J, Zhang Y F, Zhang M Q, Sun Z, Jing G C, Huang S, Su X Q. Parallel-Meta Suite: Interactive and rapid microbiome data analysis on multiple platforms. Imeta, 2022, 1(1): 11

7 Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. Peerj, 2016, 4: 22