frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 24 July 2023
DOI 10.3389/fmicb.2023.1210638

OPEN ACCESS

EDITED BY
Isabel Legaz,
University of Murcia, Spain

REVIEWED BY
Chitrak Banerjee,
Wells Fargo, United States
Li Xia,
South China University of Technology, China
Piyali Basak,
Merck, United States

*CORRESPONDENCE
Shujin Li
✉ shujinli@163.com;
✉ shujinli@hebmu.edu.cn

†These authors have contributed equally to this work and share first authorship

RECEIVED 23 April 2023
ACCEPTED 06 July 2023
PUBLISHED 24 July 2023

CITATION
Fu G, Ma G, Dou S, Wang Q, Fu L, Zhang X, Lu C, Cong B and Li S (2023) Feature selection with a genetic algorithm can help improve the distinguishing power of microbiota information in monozygotic twins' identification. *Front. Microbiol.* 14:1210638. doi: 10.3389/fmicb.2023.1210638

Feature selection with a genetic algorithm can help improve the distinguishing power of microbiota information in monozygotic twins' identification

Guangping Fu[1†], Guanju Ma[1†], Shujie Dou[1], Qian Wang[1], Lihong Fu[1], Xiaojing Zhang[1], Chaolong Lu[1], Bin Cong[1,2] and Shujin Li[1*]

[1]College of Forensic Medicine, Hebei Medical University, Hebei Key Laboratory of Forensic Medicine, Research Unit of Digestive Tract Microecosystem Pharmacology and Toxicology, Chinese Academy of Medical Sciences, Shijiazhuang, China, [2]Hainan Tropical Forensic Medicine Academician Workstation, Haikou, China

Introduction: Personal identification of monozygotic twins (MZT) has been challenging in forensic genetics. Previous research has demonstrated that microbial markers have potential value due to their specificity and long-term stability. However, those studies would use the complete information of detected microbial communities, and low-value species would limit the performance of previous models.

Methods: To address this issue, we collected 80 saliva samples from 10 pairs of MZTs at four different time points and used 16s rRNA V3–V4 region sequencing to obtain microbiota information. The data formed 280 inner-individual (Self) or MZT sample pairs, divided into four groups based on the individual relationship and time interval, and then randomly divided into training and testing sets with an 8:2 ratio. We built 12 identification models based on the time interval ($\leq$ 1 year or $\geq$ 2 months), data basis (Amplicon sequence variants, ASVs or Operational taxonomic unit, OTUs), and distance parameter selection (Jaccard distance, Bray-Curist distance, or Hellinger distance) and then improved their identification power through genetic algorithm processes. The best combination of databases with distance parameters was selected as the final model for the two types of time intervals. Bayes theory was introduced to provide a numerical indicator of the evidence's effectiveness in practical cases.

Results: From the 80 saliva samples, 369 OTUs and 1130 ASVs were detected. After the feature selection process, ASV-Jaccard distance models were selected as the final models for the two types of time intervals. For short interval samples, the final model can completely distinguish MZT pairs from Self ones in both training and test sets.

Discussion: Our findings support the microbiota solution to the challenging MZT identification problem and highlight the importance of feature selection in improving model performance.

KEYWORDS

forensic microbiology, monozygotic twins, personal identification, 16s rRNA sequencing, machine learning, genetic algorithm

1. Introduction

Distinguishing monozygotic twins (MZT) has long been a challenging task in individual identification, and it has significant implications in some cases. For example, Jobling (2013) introduced several of such cases, including a rape case where the characterization was obstructed by the existence of the suspect's MZT brother and the difficulty of distinguishing them. Various techniques have been established to identify MZTs. One such technique involves the analysis of Single Nucleotide Polymorphism (SNP), Count Number Variation (CNV), and Insertion/Deletion variation (InDel) markers (Hannelius et al., 2007; Abdellaoui et al., 2015; Xu et al., 2017). Researchers have reported that there may be differences in these genomic biomarkers between MZTs. However, due to the nearly identical nuclear and mitochondrial genome information shared by MZTs, such differences are infrequent (McRae et al., 2015; Ming et al., 2019) and genome-wide sequencing is typically necessary to differentiate MZTs. Another MZT identification strategy involves the investigation of various biomarkers from the perspective of epigenetics, including DNA methylation (Stewart et al., 2015; Xu et al., 2015) and microRNA (Fang et al., 2019; Xiao et al., 2019). Applying such biomarkers, the difference between MZTs would be more significant and the identification efficacy can be promising (Xu et al., 2015). However, using such biomarkers would necessitate strict material standards which would limit their application in daily work. Specially, RNA is easily degraded by exonuclease hydrolysis in the environment and DNA methylation method requires high quality and quantity of DNA because sodium bisulfite may lead to DNA fragmentation and loss. Therefore, more suitable biomarkers for MZT identification are still required.

Advances in microbiology may provide another possible solution to the MZT identification problem. As it is known, human bodies contain two genomes: the human genome is inherited from parents, and the microbiota reside on/in the human body. The latter, known as the "second genome" of the human body, is thought to contain genetic information that is ∼50–100 times greater than the human genome (Grice and Segre, 2012). Recent advances in sequencing technology have helped the growth of forensic microbiology, an interdisciplinary field of forensic medicine and microbiology (Ventura Spagnolo et al., 2019; Speruda et al., 2022). With a sharp focus much attention has been focused on the application of microbial information, with a particular emphasis on postmortem interval estimation (Liu et al., 2022), cause of death inference (Kaszubinski et al., 2020), body fluid type identification (Yao et al., 2021), and individual identification (Watanabe et al., 2018).

Numerous studies have revealed that everyone has a unique microbiome and differs, including between MZTs (Stahringer et al., 2012; Abeles et al., 2014; Suzuki et al., 2022; Sukumar et al., 2023). For example, Stahringer et al. (2012) demonstrated that the difference in salivary microbiome 16s rRNA data between MZT individuals was not less than that between dizygotic twins (DZT) and that this difference gradually increased after they lived apart. Based on such differences, primary studies on the MZT identification model constructions have been conducted. For example, Bozza et al. (2022) collected saliva samples from individuals with different relationships (unrelated individuals, MZT, or inner-individual samples) for sequencing and then built a model that can roughly distinguish between MZT and inner-individual samples. However, such studies would include all tested species in the model construction. In theory, the sensitivity of different microbial species to environmental changes should differ, resulting in different stability within the same individual. Therefore, changes in sequencing data of relatively unstable species would partially offset support for the same identification from relatively stable species. Therefore, there is still room for improvement in this field.

In this study, 80 saliva samples were collected from 10 MZT pairs to collect microbiota information via 16s rRNA sequencing. Twelve models were developed, each considering sample collection intervals, sequence data basis, and distance selection. High-value species were selected using Genetic Algorithm (GA) processes in such models. Finally, the two models with the best distinguishing power in the training test (under two different sampling intervals) were selected. The model developed for the short interval samples could completely separate MZT from inner-individual sample pairs in training and test sets.

2. Materials and methods

2.1. Sample collection

The present study included 10 pairs of MZT (four male and six female pairs). Participants signed informed consent forms before data collection, and the research content met the medical ethics requirements of Hebei Medical University. For each participant, four freely flowing saliva samples (1 mL each) were collected at four different time points (TP 1-4), the second, third, and fourth of which were collected 12, 13, and 14 months after the first one. Each participant stated that they had not used antibiotics within 6 months before sampling and had not consumed food or water within 2 h of sampling. A total of 80 samples were collected and labeled based on the participant and the time point. For example, "S10A1" denoted the sample belonging to individual A of the 10th MZT pair and was collected at the first time point. Genomic DNA was extracted from the saliva samples using the TGuide S96 Magnetic Soil/Stool DNA Kit (TIANGEN Biotech, Beijing) as recommended by the manufacturer and stored at −80°C until needed.

2.2. Library preparation and sequencing

A two-round-PCR workflow was used for the library preparation for the 16s rRNA V3–V4 sequencing that involves four steps: **(i) First round PCR:** The following primers were used to amplify the target regions of 16S rDNA: 338F: 5′-ACTCCTACGGGAGGCAGCA-3′ and 806R: 5′-GGACTACHVGGGTWTCTAAT-3′. PCR was performed in 10 μL reactions, with 0.3 μL forward primers (10 μM), 0.3 μL reverse ones (10 μM), 50 ng ± 20% template DNA, 0.2 μL KOD FX Neo, 5 μL KOD FX Neo Buffer, and 2 μL dNTP (2 mM of each

type). The PCR program consisted of 95°C for 5 min, followed by 25 cycles of 95°C for 30 s, 50°C for 30 s, and 72°C for 40 s, with a final extension at 72°C for 7 min. **(ii) Second round PCR:** The first round PCR product was used as a template in the second round PCR to add index labels to each sample. PCR was performed in 20 μL reactions containing 5 μL of the first PCR round product, 2.5 μL MPPI-a (2 μM), 2.5 μL MPPI-b (2 μM), and 10 μL 2 × Q5 HF MM. The PCR program consisted of 98°C for 30 s, followed by 10 cycles of 98°C for 10 s, 65°C for 30 s, and 72°C for 30 s, with a final extension at 72°C for 5 min. **(iii) Quantifying and sample mixing:** The integrity of the second round PCR product for each sample was tested using 1.8% agarose gel electrophoresis (120 V, 40 min), and then quantitative analysis was performed based on electrophoresis findings using ImageJ software. The samples were mixed by equal mass based on the quantification results. **(iv) Purifying and recovering:** The mixed samples were purified using OMEGA DNA purification columns. The fragments with target lengths were then cut and recovered in a second round of 1.8% agarose gel electrophoresis (120 V, 40 min).

Then, sequencing was performed on the Illumina Novaseq 6000 sequencing platform using a double-end sequencing strategy (PE250, Biomarker Technologies Co, Ltd., Beijing, China) according to standard protocols.

2.3. Preliminary analysis of microbial diversity

Several steps were taken on the original sequence data to obtain high-quality reads: (i) FLASH (Magoč and Salzberg, 2011) (version 1.2.11) was used to splice the original sequencing data and obtain Raw Tags, with a minimum overlap length of 10 bp and a maximum mismatch ratio set as 0.2; (ii) Trimmomatic (Bolger et al., 2014) (version 0.33) was used to filter the spliced sequence and obtain the Clean Tags: Raw Tags were checked sequentially using a 50 bp window, once the average quality value within the window was <20, the subsequent bases would be truncated. Truncated Tags with <75% length would be filtered; (iii) UCHIME (Edgar et al., 2011) (version 8.1) was used to obtain Clean Reads by removing chimeras. The threshold for confirming a chimera was set at 80% similarity between the query sequence and parent sequences selected from the reference database.

Two strategies were employed to achieve two types of microbial information (i.e., the "feature" to be selected) based on these filtered high-quality sequences: (i) The UPARSE algorithm was used on USEARCH software version 10.0 (Edgar, 2013) to cluster the sequences at a level of 97% similarity to obtain Operational Taxonomic Unit (OTU); and (ii) the dada2 workflow (Callahan et al., 2016) on QIIME2 platform (Bolyen et al., 2019) (2020.6) was used to perform noise reduction and to obtain Amplicon Sequence Variant (ASV, i.e., OTU with a level of 100% similarity). The OTUs/ASVs with <0.005% of total reads were filtered out (Bokulich et al., 2013). The SILVA database Release 138.1 (Quast et al., 2012) was used as a reference for species annotation based on ASV information, using a naive Bayesian classifier combined with alignment. RDP Classifier version 2.2 (Wang et al., 2007) was used as the annotating tool. The abundance information of various species at different taxonomic levels was then summarized using QIIME2.

The Chao 1, ACE, Shannon-wiener, and Simpson indices were calculated using Mothur v.1.30 (Schloss et al., 2009) for alpha diversity analysis for each sample. In addition, principal coordinate analysis (PCoA) was performed using the distance matrix of beta diversity parameters (such as Bray-Curtis distance).

2.4. Feature selection to promote the distinguishing ability of microbial model in MZT identification

2.4.1. Division of data sets

There would be eight samples for each MZT pair, resulting in $C_8^2 = 28$ sample pairs. All the 280 pairs (28 sample pairs per MZT pair × 10 MZT pairs) were divided into four groups based on two dimensions: (i) the relationship between the individuals from which the two samples were collected (samples from the same individual were labeled as "Self," and samples from different individuals within an MZT pair were labeled as "MZT"); and (ii) the sample collection interval ($\leqslant$ 2 months was labeled as "short", or $\geqslant$ 12 months was labeled as "long"). The sample pairs in the four groups were randomly divided into training or test sets with an 8:2 ratio, where the test set is used only for the validation of the final models. The sample amounts in each group are listed in Table 1.

2.4.2. Feature type and beta distance parameter choice

As stated in Section 2.3, two types of features (i.e., microbial sequencing information) would be used as the data basis for model construction: OTUs or ASVs. Three types of beta diversity distances were calculated based on these two types of features:

(i) Jaccard distance (JD), which only considers whether a specific feature was detected in two samples. JD between two samples, A and B, is denoted as $JD(A, B)$ and calculated as shown in Equation (1), if the symbol "$|A \cap B|$" denoted the number of features detected in both samples A and B, and "$|A \cup B|$" represented the total number of features detected in the two samples.

$$JD(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

(ii) Bray-Curtis Distance (BC), which represents the difference in absolute abundance information between two samples of the

TABLE 1 The division of data sets.

Relationship	Interval	Training set	Test set
Self	Short	48	12
	Long	48	12
MZT	Short	80	20
	Long	48	12

tested features, calculated as Equation (2), if symbols "a_i" and "b_i" denote the absolute abundance of the *i*th feature in samples A and B, respectively.

$$BC(A,B) = \frac{\sum_i |a_i - b_i|}{\sum_i (a_i + b_i)} \quad (2)$$

(iii) Hellinger distance (HD), which calculates the difference in relative abundance information between two samples of the tested features, calculated as Equation (3), if "$\sum_j a_j$" and "$\sum_j b_j$" denote the total abundance of all features considered in samples A and B, respectively.

$$HD(A,B) = \frac{1}{\sqrt{2}} \sqrt{\sum_i \left(\sqrt{\frac{a_i}{\sum_j a_j}} - \sqrt{\frac{b_i}{\sum_j b_j}} \right)^2} \quad (3)$$

In summary, six possible models could be constructed based on data basis and distance choices for distinguishing the Self and MZT pairs under each sampling interval. Therefore, a total of 12 models will be constructed. The distinguishing ability of the 12 choices would be measured based on all features detected. The area under the receiver operating characteristic curve (AUC) would be used in such measurements. The AUC parameter serves as a commonly used metric for evaluating binary classifiers. For each of the 12 models, the following steps can be taken to calculate the AUC: Firstly, calculation results for all sample pairs are sorted from smallest to largest and assigned rank numbers, during which the average rank would be assigned for pairs with identical results. Secondly, if the number of MZT pairs and Self pairs are indicated as n_M and n_S, respectively, and the sum of ranks in MZT pairs is represented with symbol "R_M," the AUC can be obtained by using the following equation:

$$AUC = \frac{R_M - \frac{n_M(n_M+1)}{2}}{n_M \times n_S} \quad (4)$$

2.4.3. Feature selection based on genetic algorithm

To construct more efficient discrimination models, GAs were performed 12 times on MATLAB (R2021a update 4, MathWorks, United States) to screen the high-value OTUs or ASVs in the identification models. Since their introduction in the 1970s, GAs have been widely used in feature selection. The algorithm uses mathematical methods and computer simulation operations to transform the process of solving complex problems into processes similar to the chromosome genes crossover and mutation in biological evolution. Then it searches for the optimal solution by simulating natural evolution processes.

The operational process of GA involves several fundamental concepts that are presented in bold and italicized words: (i) An ***individual*** presents a potential solution to the problem. Each ***individual*** is a digital string of the same length as the number of detected features in all samples. Each position on the string is called a ***gene***, one-to-one corresponding to the detected features. Each ***gene*** has two possible states called ***alleles***, 1 if the corresponding feature is selected or 0 if not. (ii) A ***population*** is a group of ***individuals***. Besides the initial ***population*** [**P(0)**], ***populations*** would be generated iteratively and denoted as **P(*x*)**, where "***x***" denote the iteration generation and ***x*** = 10,000 is set as the first termination condition. (iii) The value of an ***individual*** in MZT identification would be measured with the "fitness function." In the present study, the corresponding distance parameter would be calculated for each pair in the training set with the corresponding interval based on the features being selected according to a specific ***individual*** (i.e., the corresponding ***allele*** being 1); then AUC would be calculated with Equation (4) and used as the fitness function. The second termination condition would be AUC = 1, indicating that the two groups can be completely separated.

The GA operation would be carried out in the following steps for each of the 12 model construction processes:

(i) Initialization: Randomly generate 2,000 ***individuals*** to form **P(0)**. Calculate AUC for each ***individual***; if the second termination condition is fulfilled, output the optimal solution directly; otherwise, enter the evolution process.

(ii) Evolution: Three steps would be taken to generate a new ***population***, **P(*x*+1)**, from the existing one, **P(*x*)**:

(a) Selection: ***Individuals*** with the lowest 10% AUC in **P(*x*)** would be eliminated. According to their AUCs, 2,000 "fathers" and 2,000 "mothers" would be selected from the remaining 90% **P(*x*)** ***individuals*** according to their AUC:

$$Pr(individual\ i\ is\ selected) = \frac{AUC_i}{\sum_{j \in the\ top\ 90\%\ individuals} AUC_j} \quad (5)$$

(b) Crossover: ***Individuals*** in **P(*x*+1)** would inherit ***alleles*** from their "parents," ***alleles*** from both of which have the same probability of being inherited for each gene.

(c) Mutation: A part of the ***alleles*** in the next generation would be reversed between 0 and 1. In the present study, the mutation rate, i.e., the probability of such reversion occurring, was 1%.

(iii) Evaluation and iteration: Calculate AUC for each ***individual*** of **P(*x*+1)**. If any of the two termination conditions are fulfilled, output the optimal solution and terminate the process; otherwise, repeat the evolution process in step (ii).

After the 12 GA processes, select the two feature-distance combinations with the best AUCs for each type of interval to build the MZT identification model.

2.5. Introducing Bayes theory in the final models

After feature selection, the specific distance between a sample pair can be used to infer the relationship between the sourcing individuals. However, in some cases, it may be necessary to answer the question, "How confident are we in judging the relationship between these two individuals." The likelihood ratio calculation could be helpful in answering the question.

2.5.1. Basic theories

When only one individual of the MZT pairs is available as the suspect for MZT identification, there are more than two exclusive assumptions (Bozza et al., 2022), such as H_1: the query saliva trace originates from the suspect; or H_2: the query saliva trace originates from the suspect's MZT sibling. The present study aimed to estimate the probability of H_1 being true, considering the distance obtained by sequencing and calculating, i.e., $Pr(H_1|D = d)$ where "d" stands for the specific distance value and "$D = d$" denote the event that d is calculated between the two samples. According to the Bayes Rules, such a probability can be calculated by Equation (6):

$$Pr(H_1|D = d) = \frac{Pr(H_1) \times Pr(D = d|H_1)}{Pr(D = d)} \tag{6}$$

Where $Pr(H_1)$ denote the probability of H_1 being true without considering the microbial evidence, and $Pr(D = d)$ denote the probability of two individuals exhibiting corresponding distance. $Pr(D = d)$ is usually difficult to calculate accurately because there may be infinite hypotheses, resulting in a high difficulty in accurately calculating $Pr(H_1|D = d)$. However, calculating the ratio of it to the probability of another hypothesis being true would be much easier, such as:

$$\frac{Pr(H_1|D = d)}{Pr(H_2|D = d)} = \frac{Pr(H_1)}{Pr(H_2)} \times \frac{Pr(D = d|H_1)}{Pr(D = d|H_2)} \tag{7}$$

Where the ratio $Pr(H_1)/Pr(H_2)$ is known as the ratio of prior probabilities, and if the probability of any hypothesis being true is assumed to be equal, such a ratio should equal 1. Thus, if we define a parameter likelihood ratio (LR or BF in some studies; Bozza et al., 2022) as Equation (8), it can represent the degree to which H_1 is more likely to be true than H_2. When considering a determined value, the probability ratio should equal the ratio between the probability density in the Self group and that in the MZT group, labeled with $\hat{f}$:

$$LR = \frac{Pr(D = d|H_1)}{Pr(D = d|H_2)} = \frac{\hat{f}(d|Self, short/long)}{\hat{f}(d|MZT, short/long)} \tag{8}$$

2.5.2. Model construction based on LR results

To obtain the two probabilities in Equation (8), the distribution of the specific distance between Self or MZT sample pairs must be estimated while considering the time interval. Gaussian Kernel density estimation (KDE) was used as the estimation method. Assume that n pairs of specific samples are tested, and the specific distance between each pair is labeled as X_1, X_2, ..., and X_n, then the probability density can be estimated by Equation (9):

$$\hat{f}(d) = \frac{1}{nh\sqrt{2\pi}} \sum_{i=1}^{n} \left[e^{-\frac{(d-X_i)^2}{2h^2}} \right] \tag{9}$$

And the bandwidth h was set as Equation (10), where δ denotes the standard deviation of all X_i.

$$h = 1.06\delta n^{-\frac{1}{5}} \tag{10}$$

Thus, the LR of each pair could be calculated by Equation (11)

$$LR = \frac{n_M h_M \sum_{i=1}^{n_S} \left[e^{-\frac{(d-S_i)^2}{2h_S^2}} \right]}{n_S h_S \sum_{j=1}^{n_M} \left[e^{-\frac{(d-M_j)^2}{2h_M^2}} \right]} \tag{11}$$

Where n_M or n_S denotes the number of MZT or Self pairs within a specific time interval, and h_M or h_S represents the bandwidth of corresponding groups. S_i denotes the specific distance between the *ith* Self pair and M_j the distance between the *jth* MZT pair.

After calculating LR for each of the 224 sample pairs, diagnostic tests were performed between MZT and Self pairs at each time interval. Two types of thresholds would be used in the final models: (i) T_{max}: the LR thresholds that may provide the best Youden index [YI, which can be calculated with Equation (12), if *Sen* and *Spe* denote sensitivity and specificity, respectively] in the training sets; and (ii) LR = 1.

$$\left.\begin{aligned} Sen &= \frac{number\ of\ Self\ pairs\ being\ confirmed\ as\ Self\ pairs}{n_S} \\ Spe &= \frac{number\ of\ MZT\ pairs\ being\ confirmed\ as\ MZT\ pairs}{n_M} \end{aligned}\right\} \Longrightarrow YI = Sen + Spe - 1 \tag{12}$$

2.6. Validation of the model

Based on the two feature-distance combinations selected, LRs would be calculated for each of the 56 sample pairs in the test set, and diagnostic tests would be performed. The validation parameter of the final models would be YI under the two types of threshold sets.

3. Results

3.1. Sequencing results

Supplementary Table 1 contains basic information about the 10 MZT pairs: age, gender, and population. After sequencing, 6,134,347 pairs of reads were obtained, with 6,107,044 clean reads remaining after splicing and filtering, as described in section 2.3. At least 48,389 clean reads were obtained for each of the 80 samples, with the mean number of clean reads per sample being 76,338. After clustering at 97% similarity and filtering with a threshold of 0.005% in total reads, 369 types of OTUs were detected across all 80 samples. The number of OTUs obtained per sample ranged from 141 to 332, averaging 230. As shown in Figure 1A, OTUs obtained in samples collected at the first time point for each participant would differ from the other three. Those differed significantly between inner-individual samples could be "burdens" on the identification model, implying the importance of feature selection. A total of 1130 ASVs were obtained after noise reduction using DADA2 and filtering with a 0.005% threshold in the 80

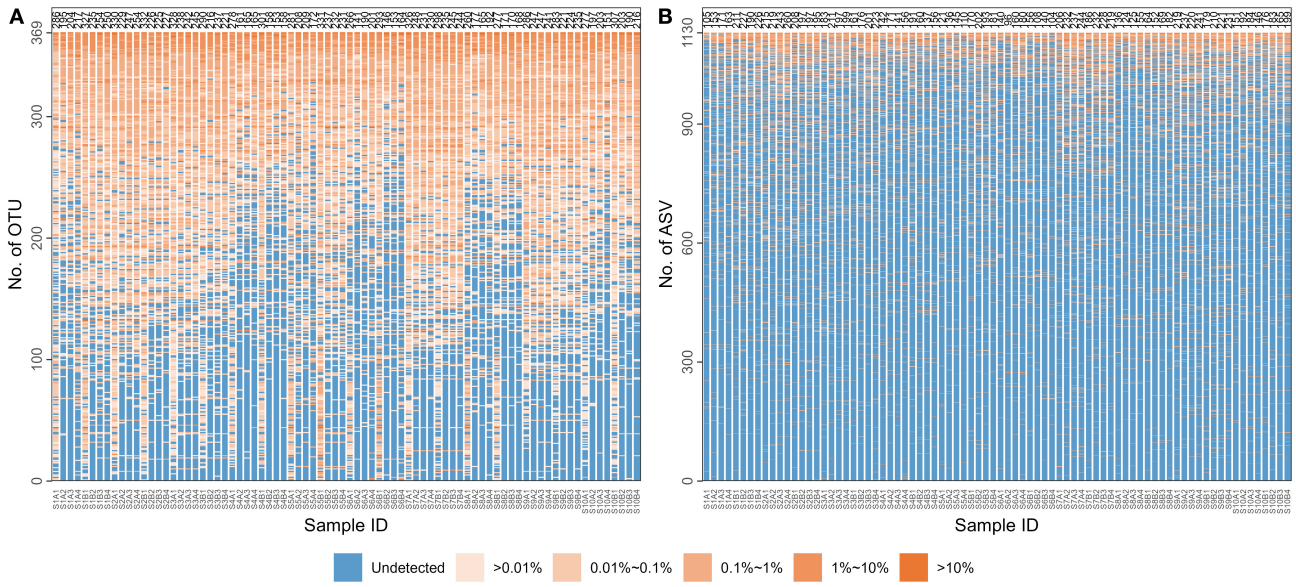

FIGURE 1
The distribution of microbial species characteristics in each sample. The microbiological information of 80 samples (each column for a sample) is represented at two levels: **(A)** Operational taxonomic unit (OTU); **(B)** Amplicon sequence variants (ASV). The numbers of OTUs or ASVs detected in each sample are listed at the top of each column. A color gradient represents the percentages for each detected OTU/ASV in the corresponding sample.

samples. Figure 1B indicates that the number of ASVs obtained per sample ranged from 96 to 260, with an average of 181.

Taxonomic annotation was performed using the SILVA 138 database, as mentioned in Section 2.3. There were 13 phyla detected in all 80 samples, with the top five (Firmicutes, Proteobacteria, Bacteroidota, Actinobacteriota, and Fusobacteriota) accounting for 97%. At the genus level, 131 genera were detected, with Streptococcus and Neisseria accounting for about 21 and 13% of the bacteria in saliva, respectively. The top five genera account for more than half of all species. Figure 2 depicts the taxonomic composition of all samples at the phylum and genus levels, and Supplementary Figure 1 indicates the compositions in each sample. Differences in taxonomic compositions can be found between both Self sample pairs and MZT sample pairs, indicating that microbiota information may have the potential to distinguish MZT. However, feature selection would be required in the model construction.

Basic information of OTUs and ASVs applied in the model construction is presented in Supplementary Table 2, including taxonomic annotation, Reads in each sample and the status of selection in the final model.

3.2. Preliminary analysis

3.2.1. Alpha diversity analysis

Based on ASV information, four alpha diversity indices were calculated for each sample. After calculation, it was found that both ACE and Chao 1 indices of any sample equaled the number of ASV detected from the corresponding sample. For each individual, the four samples collected at different time points can form six comparisons; and for each type of comparison, paired *t*-test or Wilcoxon sign rank-sum test was performed to assess the difference between the distributions of the alpha indices between the corresponding two time points, according to whether the differences between all samples fit the normal distribution. As shown in Figure 3 and Table 2, if the time interval was <2 months within the same individuals, no significant difference could be found regardless of which index was calculated. More significant differences could be found between samples collected at the first time point and the other three if Shannon and Simpson's indices were more calculated than the other two.

3.2.2. PCoA based on beta diversity analysis

The three types of beta diversity parameters mentioned in Section 2.4.2 were used to perform principal coordinate analysis as shown in Figure 4, indicating that the results were similar to each other. The samples collected at the first time point were significantly different from other samples, i.e., the difference between the collection times point would overcome the difference between MZT pairs.

3.3. Feature selection with GA processes

As mentioned in Section 2.4.1, the 80 samples could form 280 Self or MZT sample pairs and be divided into four groups based on the relationship of the sourcing individuals and the time interval. Each group was randomly divided into training and test sets in an 8:2 ratio. A total of 12 GA processes were performed based on different feature-distance combinations in the training set, with the best AUC of each population generation listed in Supplementary Figure 2. Except for two GAs that stopped due to satisfying AUC = 1, the generational-best AUC of the remaining 10 GAs fluctuated within a relatively flat range after 1,000-1,500 iterations. The final output result can be considered

10.3389/fmicb.2023.1210638

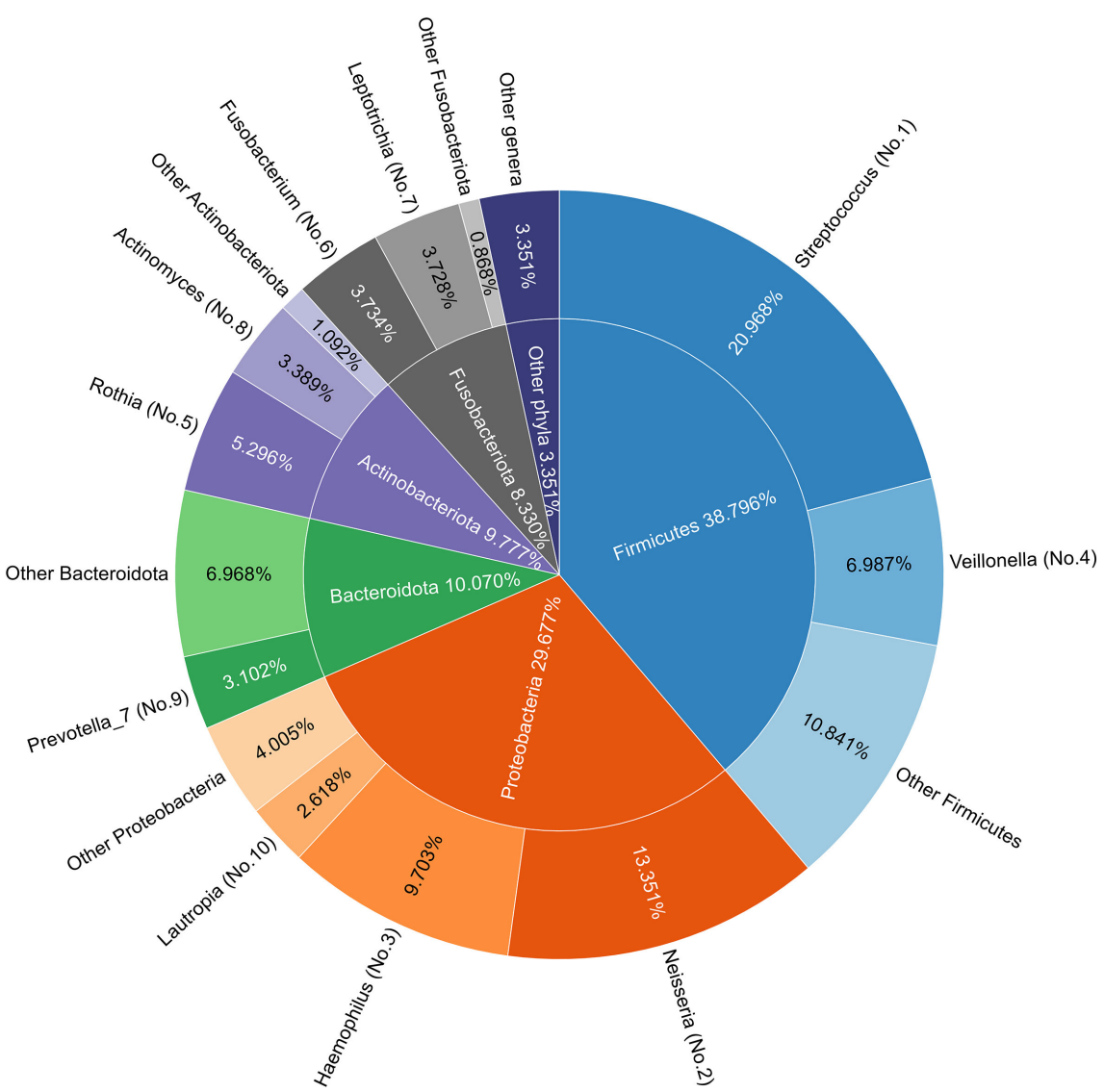

FIGURE 2
Species composition in all samples. **(Inner)** Pie chart of phylum composition of all samples: the top five phyla are arranged clockwise, with names and proportions marked in corresponding positions. **(Outer)** A doughnut chart of genus composition in all samples, with the top 10 genera listed. All genera belonging to the same phylum are grouped and labeled with a gradient of color for the corresponding phylum in the inner pie chart. The names and proportions of each genus are marked in the corresponding positions, where serial numbers in parentheses indicate the proportion rankings. Streptococcus and Neisseria are the most common genera in saliva and the top five genera accounting for more than half of all species.

close to the optimal solutions. Table 3 indicates the best AUCs for each GA, each of which is better than the corresponding AUC if all species detected were considered. Supplementary Figure 3 depicts the distance distributions based on the corresponding selected species and the KDE results. If the time interval is short and JD is used, the two groups could be completely separated, regardless of the choice in OTUs or ASVs. However, the KDE results indicate that if the final chosen combination of ASVs is used, the gap between the two curves would be larger, implying a better distinguishing ability. Meanwhile, when the time interval is long, the ASV-JD combination can also provide the best AUC. Therefore, ASV-JD combinations are selected for further model construction. As mentioned in Section 3.1, the status of whether each ASV is applied in the final model is annotated in Supplementary Table 2B.

3.4. Construction and validation of the final models

In the present study, the training set comprised 224 sample pairs. JD was calculated for each sample pair based on one of the two ASV combinations, as selected in Section 3.3, according to the time interval between the samples. Further, probability density curves were obtained for the four groups by processing the calculation results using KDE. These groups were segregated based on the relationship and time interval between the samples. Then, LR was computed for each of the 224 sample pairs and the corresponding results are shown in Table 4 (the first 4 rows) and Figures 5A, B. In the training set, the minimum LR of the Short-Self group was 0.8152, which was slightly higher than the maximum LR

10.3389/fmicb.2023.1210638

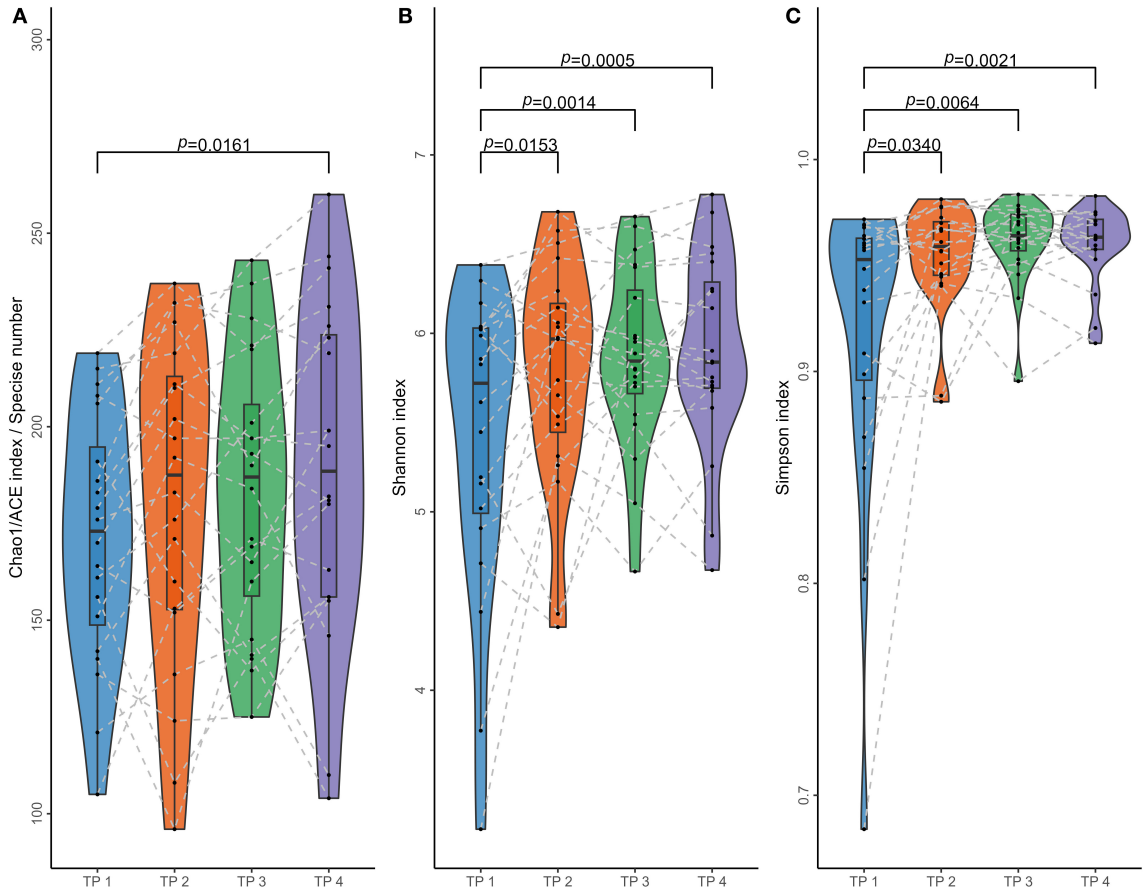

FIGURE 3
Paired comparisons of microbial alpha diversity in samples collected at different times. Four parameters were calculated to access microbial alpha diversity within a single sample and compare them in different time groups. **(A)** Chao 1 and ACE indices, measuring the species richness of a sample; **(B)** the Shannon indices and **(C)** Simpson indices, considering community evenness. These alpha diversity parameters were calculated using ASV sequence data of 80 samples from 20 individuals at four-time points (TP 1–4), with samples from TP 2/3/4 collected at least 1 year after TP 1. Dots present the parameters of each sample, and dotted lines link samples of the same individual. Paired *t*-test or Wilcoxon sign rank-sum test was performed for each of the six comparisons between the four groups, and $P < 0.05$ are listed in the figure.

of Short-MZT group (0.3638). Therefore, if the confirming Self pair threshold is set between the two LR values (e.g., 0.4), the two groups in the training set can be completely separated, i.e., YI = 1. If LR = 1 is set as the threshold, one in 48 Short-Self pairs will be identified as an MZT pair, while no MZT pair will be mistakenly identified, i.e., YI = 0.9792. Meanwhile, in the training set, the minimum LR of the Long-Self group was 0.9553, and 3 Long-MZT pairs had LR higher than that. The best YI was achieved when the threshold was set between 1.2214 and 1.2548 when the Long-Self pair with the minimum LR would be the only pair incorrectly identified, and YI = 0.9792. If LR = 1 is set as the threshold, three Long-MZT pairs and one Long-Self pair would be mistakenly identified, resulting in YI = 0.9167.

Furthermore, JD was calculated similarly for each sample pair in the test set, and then LR was calculated based on the above mentioned probability density curves. The two types of threshold sets were validated in the test set, as shown in Table 4 (the last four rows) and Figures 5C, D (the two lower sub-figures). Comparing to the training set, the model under short intervals could provide a similar YI in the corresponding test set. In contrast, two types of YI would drop sharply for identification between Long-Self and Long-MZT groups, from ~0.9 in training set to 0.5 (with both sensitivity and specificity being 0.75) in test set.

4. Discussion

In the present study, MZT identification models were built using microbiota 16s rRNA V3–V4 region information from 80 saliva samples collected from 10 pairs of MZT individuals while considering the time interval of sample collections. GAs were used to select high-value species-distance combinations for such models. As a result, JD calculated from 516 selected ASVs was used to construct a model that could completely separate MZT sample pairs collected in a short time interval (⩽2 months, i.e., Short-MZT pairs) from Short-Self pairs in both the training and test sets. The distinguishing power of such a model was improved when compared to all 1,130 detected types of ASVs with the same distance parameter. A similar improvement in the GA process could be found for sample pairs with long time intervals (⩾1 year), and

TABLE 2 Comparison between α diversity between inner-individual samples.

Parameter	Time points	P1*	P2**	Note***
ACE/Chao 1	TP 1 vs. TP 2	0.1412	0.1332	T
	TP 1 vs. TP 3	0.8144	0.0675	T
	TP 1 vs. TP 4	0.4232	0.0161	T
	TP 2 vs. TP 3	0.7515	0.6714	T
	TP 2 vs. TP 4	0.6478	0.2619	T
	TP 3 vs. TP 4	0.0833	0.3042	T
Shannon	TP 1 vs. TP 2	0.0001	0.0153	R
	TP 1 vs. TP 3	0.0011	0.0014	R
	TP 1 vs. TP 4	0.0003	0.0005	R
	TP 2 vs. TP 3	0.0023	0.4221	R
	TP 2 vs. TP 4	0.0880	0.2026	T
	TP 3 vs. TP 4	0.8462	0.7192	T
Simpson	TP 1 vs. TP 2	0.2056	0.0340	T
	TP 1 vs. TP 3	0.1974	0.0064	T
	TP 1 vs. TP 4	0.3124	0.0021	T
	TP 2 vs. TP 3	0.0119	0.9854	R
	TP 2 vs. TP 4	0.2653	0.2760	T
	TP 3 vs. TP 4	0.5051	0.6319	T

*P-value of Shapiro-wilk tests on the differences of specific parameters in samples from the same individual.
**P-value of paired t-test (if corresponding $P_1 > 0.05$) or Wilcoxon sign rank-sum test (if not) of each paired comparison, in which the significant differences ($P_2 < 0.05$) are colored red and underlined.
***T, paired t-test; R, Wilcoxon sign rank-sum test.

510 ASVs were selected to construct a model that would make correct decisions for 95 out of 96 pairs of the training set. However, the accuracy of such a model would drop sharply in the test set, where 6 of 24 pairs were misjudged, indicating the probability of over-fitting.

It is imperative that applied biomarkers exhibit individual specificity (i.e., distinction between different individual) to distinguish between MZT pairs effectively. Microbiota information is a more promising avenue for this purpose than the traditional genomic biomarkers, such as short tandem repeats (STRs) or single nucleotide polymorphisms (SNPs), which theoretically should be identical between MZTs. The first phase of the Human Microbiome Project (HMP) demonstrated that each collected sample contained a unique set of microorganisms (Human Microbiome Project Consortium, 2012), and multiple studies have highlighted the occurrence of similar differences between MZTs (Bowyer et al., 2022; Sukumar et al., 2023). The composition of an individual's microbiota undergoes a succession process throughout their lifetime (Martino et al., 2022), which can be influenced by several factors such as dietary habits, antibiotic treatments (Bokulich et al., 2016), diseases, or stochastic factors (Turnbaugh et al., 2010). While these factors shape the individual specificity of microbiota communities, they also continuously affect the stability of the communities. And another important requirement of the MZT distinguishing biomarkers is inner-individual stability, which ensures that the difference between the corresponding biomarkers from inter-individual samples is highly likely to exceed the difference between inner-individual samples. The sensitivity of different microorganisms to the above-mentioned factors should differ; the more sensitive species would significantly affect the inner-individual stability and thus "drag down" the application value of the MZT identification models. This could partially explain the improved performance of the model after the feature selection process, as represented by AUC (as illustrated in Table 3).

Based on ASV data, four types of alpha diversity parameters were calculated for each of the 80 samples. ACE and Chao 1 estimators measure species richness among these parameters, while Shannon and Simpson's indices consider community evenness. Analysis of Figure 3 reveals no significant differences in inner-individual samples collected at TP 2–4. Meanwhile, significant differences were identified between these samples and those collected at TP 1 when Shannon and Simpson's indices were used. This suggests that microbiota composition can be stable over a relatively short period of time. In contrast, calculating the ACE or Chao 1 estimator only revealed a significant difference between samples with the longest time intervals (14 months between TP 1 and TP 4). This indicates that the inner-individual stability of microbial species richness is superior to the stability considering community composition. This finding may help to explain why subsequent model constructions based on JD (which also considers only species richness) performed better under the same conditions than models based on the other two types of distances (which consider the community composition).

The calculated ACE and Chao 1 indices for each sample equaled the detected ASV of the corresponding sample. The reason would be that after being confirmed as ASVs, sequences were filtered with a 0.005% threshold, making the minimum reads of the final used ASVs $> 48{,}389 \times 0.005\% = 2.41945$. Consequently, parameters "F1" and "F2" used in the calculation of ACE and Chao 1 indices were set to zero for each sample, and the two indices equaled the observed ASV number, as shown in Equation (13).

$$\begin{cases} ACE & = S_{abund} + \frac{S_{rare}}{1-\frac{F1}{C_{ACE}}} + \frac{F1}{C_{ACE}}\gamma^2_{ACE} = S_{abund} + S_{rare} = S_{obs} \\ Chao\ 1 & = S_{obs} + \frac{F1(F1-1)}{2(F2+1)} = S_{obs} \end{cases} \quad (13)$$

Where S_{obs}, S_{abund}, and S_{rare} denote the total number of observed ASVs, the number of ASVs with large reads, and the number of ASVs with small reads; $F1$ and $F2$ represent the number of ASVs with reads 1 and 2, respectively. C_{ACE} and γ^2_{ACE} can be calculated with several additional steps, but their values do not affect the final result when $F1 = 0$.

Based on the data presented in Supplementary Figure 3, it appears that distances exhibited a unimodal distribution across all 24 groups, accounting for three types of distances, two types of time intervals, two types of databases, and two types of sourcing individuals' relationships. Therefore, the conversion of LRs based on the aforementioned distributions results in a rough negative correlation with distance results. Consequently, there was no significant change in the numerical ranking of calculation results between samples, indicating that adding LR did not improve the distinguishing capability of the specific model in this study. Nevertheless, LR can be an effective tool for evaluating evidence

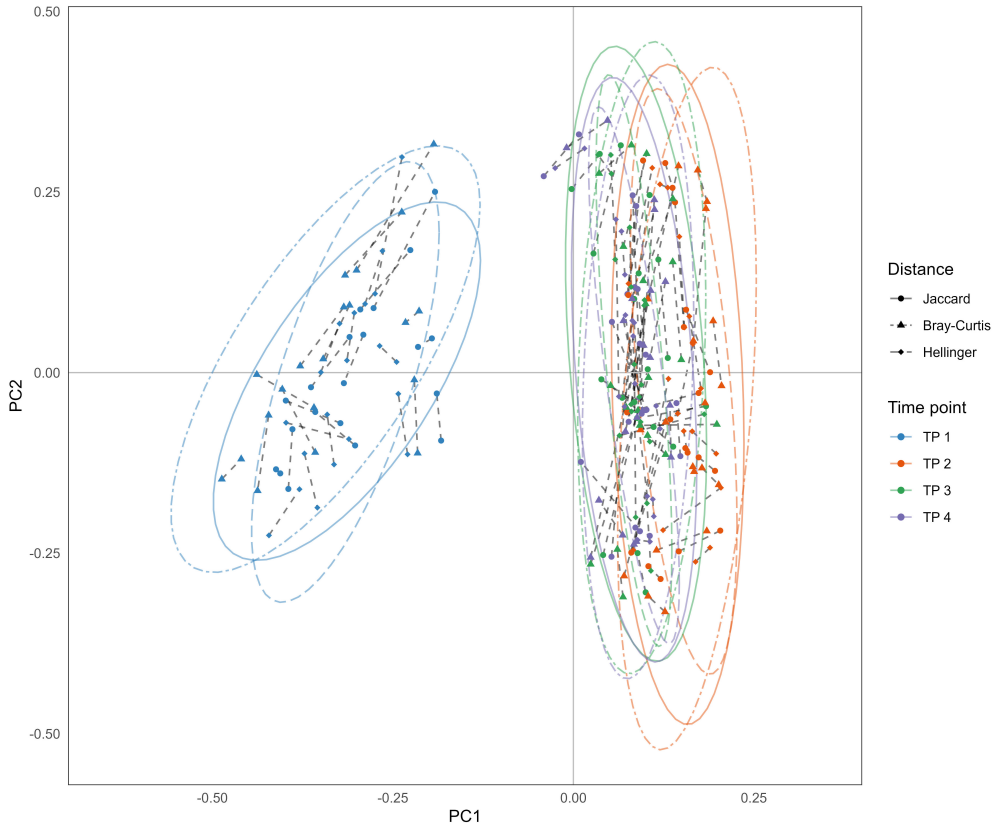

FIGURE 4
Results of principal coordinate analysis (PCoA). PCoA was performed using three types of beta density parameters: (i) Dots: Jaccard distance; (ii) Triangles: Bray-Curitis distance; (iii) Rhombus: Hellinger distance. Each point represents a sample, which is colored according to the time point of sample collection. Dotted lines connect samples collected from an MZT pair at the same time.

TABLE 3 Best AUC achieved by 12 GA processes.

Time interval	Species-level	Distance	AUC_ALL*	Best AUC**	N***
Short	OTUs	Jaccard	0.6730	1	113
		Bray-Curtis	0.7299	0.8977	176
		Hellinger	0.7669	0.9560	134
	ASVs	Jaccard	0.8964	1	516
		Bray-Curtis	0.8448	0.9820	514
		Hellinger	0.8828	0.9880	514
Long	OTUs	Jaccard	0.5855	0.9123	111
		Bray-Curtis	0.6263	0.8320	141
		Hellinger	0.6385	0.8533	137
	ASVs	Jaccard	0.6795	0.9985	510
		Bray-Curtis	0.6102	0.9201	517
		Hellinger	0.6302	0.9536	494

*AUC in the training set if all detected features were involved in the model.
**Best AUC achieved after GA processes, the largest of which under each time interval are labeled red and underlined.
***Number of features selected in the final model.

credibility with respect to specific distance calculations. It enables a numerical representation of our confidence level in making a particular decision when faced with definite distance outcomes in real-world situations.

Limitations can be found in the present study. For example, unrelated sample pairs from different MZT pairs were not considered during the model construction process in this study. Fresh saliva samples were used, from which the host's DNA

TABLE 4 Model evaluation results.

Data set	Time interval	Threshold	Sen	Spe	YI*	Misjudged pairs
Training set	Short	$T_{max} = 0.4$**	1	1	1	None
		1	0.9792	1	0.9792	1 Self
	Long	$T_{max} = 1.25$	0.9792	1	0.9792	1 Self
		1	0.9792	0.9375	0.9167	1 Self and 3 MZT
Test set	Short	$T_{max} = 0.4$	1	1	1	None
		1	0.9167	1	0.9167	1 Self
	Long	$T_{max} = 1.25$	0.75	0.75	0.5	3 Self and 3 MZT
		1	0.75	0.75	0.5	3 Self and 3 MZT

*Sen, Sensitivity; Spe, Specificity; YI, Youden index; The calculation methods of these three parameters are presented in Equation (12).
** T_{max}: the LR thresholds could provide the best YI in the training sets.

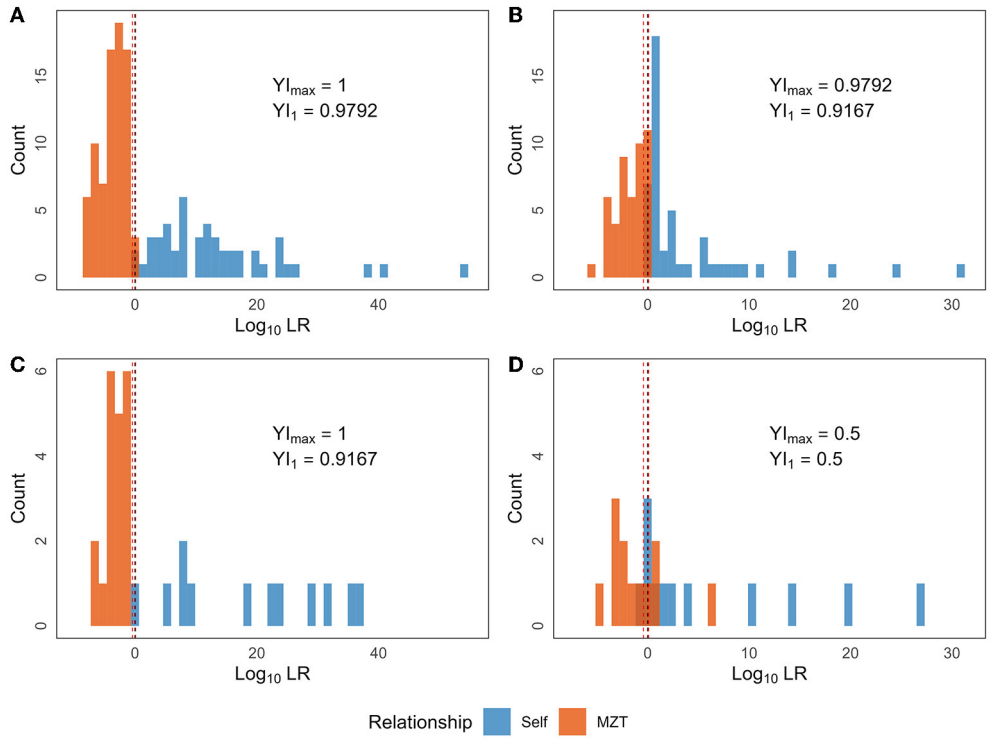

FIGURE 5
The LR distribution based on the final selected combination of ASVs. LR was calculated for each of the 280 sample pairs, which can be divided into eight groups based on the following three dimensions: (i) time interval (Long or Short); (ii) sourcing individual relationship (Self or MZT pair); (iii) training set or test set. The four sub-figures contain the LR distribution of Self (blue) or MZT (red) pairs in **(A)** Short+training set; **(B)** Long+training set; **(C)** Short+test set; **(D)** Long+test set. The red and black dotted lines denote T_{max} and LR = 1, respectively.

should be easily obtained. Therefore, individual identification issues other than MZT cases can be solved using traditional host-genome-level biomarkers like STRs or SNPs. If it is difficult to obtain the host's DNA, the sample is likely collected from another part of the body or is not fresh. The reference significance of results from fresh saliva samples to such cases would be reduced. This also highlights a problem that hinders using microorganisms in individual identification. The composition of microbial communities is greatly affected by the source body sites (Ward et al., 2018) and the time of isolation (Liu et al., 2022).

Another challenge to applying microbial 16s rRNA data in the forensic field is the lack of a standardized method of sequencing nomenclature, i.e., specific generic names for detected sequences, similar to the RefSNP (rs) system of SNP markers (Sherry et al., 2001). Because of this, the ASVs or OTUs in all samples were numbered sequentially, e.g., ASV1, and ASV2, which is obviously of a limited reference value. Such a problem would be covered if the model used all of the sequencing information, but it would be widened when feature selection is used. Before establishing the standardized method, researchers must construct their own microbial MZT identification models based on their sequencing

data, instead of applying existing models, such as the ones we constructed in the present study.

Moreover, the present study used a relatively small sample size, which limited the consideration of various factors that could affect the model. For example, recent studies have demonstrated that co-habitation may be the primary factor in bacterial sharing between MZT pairs due to person-to-person oral microbiome transmission between co-living family members (Valles-Colomer et al., 2023). Therefore, there would be an increase in microbiota similarity between MZT pairs with increased co-habitation time, and a decrease after they live apart (Stahringer et al., 2012; Valles-Colomer et al., 2023). The number of co-living and separated pairs among the 10 MZT pairs investigated in the present study were four and six, respectively, which are insufficient for statistical analysis when considering time intervals and dividing data sets. Similar insufficiency would occur when considering gender, age, diseases, and antibiotics applications. Therefore, large-scale research and experimentation considering co-living factors may provide a more accurate assessment of the potential of microbiota information in resolving the MZT identification problem.

In summary, the investigation of the microbial solution to the MZT identification problem is still in its early stages. And thus, rather than recommending a preconceived model, the primary objective of the current research is to underscore the potential utility of feature selection in facilitating the process of model construction when identifying MZT pairs with microbial information.

5. Conclusions

The present study used 80 saliva samples from 10 pairs of MZT to develop models for identifying MZT based on their microbiota information. The feature selection method was used to build models that effectively distinguish MZT from Self pairs. When the sampling interval was <2 months, the final model could completely separate the two types of sample pairs. This highlights the potential value of microbiota information as a biomarker in MZT identification. Moreover, the feature selection process demonstrated its ability to refine models and filter out irrelevant data in this field, resulting in more accurate and reliable models.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: NCBI—PRJNA983501.

Ethics statement

The studies involving human participants were reviewed and approved by Medical Ethics Committee of Hebei Medical University. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

Author contributions

SL, BC, and GF contributed to the conception and design of the study. GF, SD, QW, and XZ collected saliva samples. GF and SD performed 16s rRNA sequencing. GM conducted data analysis, carried out genetic algorithm in model construction, and wrote the first draft of the manuscript. LF and CL verified the model. BC and SL supervised the entire study and provided funding support. All authors have read, commented on, and approved the submitted version.

Funding

This work was supported by the National Natural Science Foundation of China (82072118), Natural Science Foundation of Hebei Province (H2021206451), and S&T Program of Hebei (225A5602D).

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2023.1210638/full#supplementary-material

References

Abdellaoui, A., Ehli, E. A., Hottenga, J.-J., Weber, Z., Mbarek, H., Willemsen, G., et al. (2015). CNV concordance in 1,097 MZ twin pairs. *Twin Res. Hum. Genet.* 18, 1–12. doi: 10.1017/thg.2014.86

Abeles, S. R., Robles-Sikisaka, R., Ly, M., Lum, A. G., Salzman, J., Boehm, T. K., et al. (2014). Human oral viruses are personal, persistent and gender-consistent. *ISME J.* 8, 1753–1767. doi: 10.1038/ismej.2014.31

Bokulich, N. A., Chung, J., Battaglia, T., Henderson, N., Jay, M., Li, H., et al. (2016). Antibiotics, birth mode, and diet shape microbiome maturation during early life. *Sci. Transl. Med.* 8, 343ra82. doi: 10.1126/scitranslmed.aad7121

Bokulich, N. A., Subramanian, S., Faith, J. J., Gevers, D., Gordon, J. I., Knight, R., et al. (2013). Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat. Methods* 10, 57–59. doi: 10.1038/nmeth.2276

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

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* 37, 852–857 doi: 10.7287/peerj.preprints.27295v1

Bowyer, R. C., Twohig-Bennett, C., Coombes, E., Wells, P. M., Spector, T. D., Jones, A. P., et al. (2022). Microbiota composition is moderately associated with greenspace composition in a UK cohort of twins. *Sci. Tot. Environ.* 813, 152321. doi: 10.1016/j.scitotenv.2021.152321

Bozza, S., Scherz, V., Greub, G., and Taroni, F. (2022). A probabilistic approach to evaluate salivary microbiome in forensic science when the Defense says: 'it is my twin brother'. *Forens. Sci. Int.* 57, 102638. doi: 10.1016/j.fsigen.2021.102638

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., and Holmes, S. P. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. *Nat. Methods* 13, 581–583. doi: 10.1038/nmeth.3869

Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat. Methods* 10, 996–998. doi: 10.1038/nmeth.2604

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., and Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 27, 2194–2200. doi: 10.1093/bioinformatics/btr381

Fang, C., Zhao, J., Liu, X., Zhang, J., Cao, Y., Yang, Y., et al. (2019). MicroRNA profile analysis for discrimination of monozygotic twins using massively parallel sequencing and real-time PCR. *Forens. Sci. Int.* 38, 23–31. doi: 10.1016/j.fsigen.2018.09.011

Grice, E. A., and Segre, J. A. (2012). The human microbiome: our second genome. *Annu. Rev. Genomics Hum. Genet.* 13, 151–170. doi: 10.1146/annurev-genom-090711-163814

Hannelius, U., Gherman, L., Mäkelä, V.-V., Lindstedt, A., Zucchelli, M., Lagerberg, C., et al. (2007). Large-scale zygosity testing using single nucleotide polymorphisms. *Twin Res. Hum. Genet.* 10, 604–625. doi: 10.1375/twin.10.4.604

Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214. doi: 10.1038/nature11234

Jobling, M. A. (2013). Double trouble. *Invest. Genet.* 4:12. doi: 10.1186/2041-2223-4-12

Kaszubinski, S. F., Pechal, J. L., Smiles, K., Schmidt, C. J., Jordan, H. R., Meek, M. H., et al. (2020). Dysbiosis in the dead: human postmortem microbiome beta-dispersion as an indicator of manner and cause of death. *Front. Microbiol.* 11, 555347. doi: 10.3389/fmicb.2020.555347

Liu, R., Wang, Q., Zhang, K., Wu, H., Wang, G., Cai, W., et al. (2022). Analysis of postmortem intestinal microbiota successional patterns with application in postmortem interval estimation. *Microb. Ecol.* 84, 1087–1102. doi: 10.1007/s00248-021-01923-4

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

Martino, C., Dilmore, A. H., Burcham, Z. M., Metcalf, J. L., Jeste, D., and Knight, R. (2022). Microbiota succession throughout life from the cradle to the grave. *Nat. Rev. Microbiol.* 20, 707–720. doi: 10.1038/s41579-022-00768-z

McRae, A. F., Visscher, P. M., Montgomery, G. W., and Martin, N. G. (2015). Large autosomal copy-number differences within unselected monozygotic twin pairs are rare. *Twin Res. Hum. Genet.* 18, 13–18. doi: 10.1017/thg.2014.85

Ming, T., Wang, M., Zheng, M., Zhou, Y., Hou, Y., and Wang, Z. (2019). Exploring of rare differences in mtGenomes between MZ twins using massively parallel sequencing. *Forens. Sci. Int.* 7, 70–72. doi: 10.1016/j.fsigss.2019.09.028

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, 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

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., 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

Sherry, S. T., Ward, M. H., Kholodv, M., Baker, J., Phan, L., Smigielski, E. M., et al. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29, 308–311. doi: 10.1093/nar/29.1.308

Speruda, M., Piecuch, A., Borzecka, J., Kadej, M., and Ogórek, R. (2022). Microbial traces and their role in forensic science. *J. Appl. Microbiol.* 132, 2547–2557. doi: 10.1111/jam.15426

Stahringer, S. S., Clemente, J. C., Corley, R. P., Hewitt, J., Knights, D., Walters, W. A., et al. (2012). Nurture trumps nature in a longitudinal survey of salivary bacterial communities in twins from early adolescence to early adulthood. *Genome Res.* 22, 2146–2152. doi: 10.1101/gr.140608.112

Stewart, L., Evans, N., Bexon, K. J., van der Meer, D. J., and Williams, G. A. (2015). Differentiating between monozygotic twins through DNA methylation-specific high-resolution melt curve analysis. *Analyt. Biochem.* 476, 36–39. doi: 10.1016/j.ab.2015.02.001

Sukumar, S., Wang, F., Simpson, C. A., Willet, C. E., Chew, T., Hughes, T. E., et al. (2023). Development of the oral resistome during the first decade of life. *Nat. Commun.* 14, 1291. doi: 10.1038/s41467-023-36781-w

Suzuki, T. A., Fitzstevens, J. L., Schmidt, V. T., Enav, H., Huus, K. E., Mbong Ngwese, M., et al. (2022). Codiversification of gut microbiota with humans. *Science* 377, 1328–1332. doi: 10.1126/science.abm7759

Turnbaugh, P. J., Quince, C., Faith, J. J., McHardy, A. C., Yatsunenko, T., Niazi, F., et al. (2010). Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins. *Proc. Natl. Acad. Sci.* 107, 7503–7508. doi: 10.1073/pnas.1002355107

Valles-Colomer, M., Blanco-Míguez, A., Manghi, P., Asnicar, F., Dubois, L., Golzato, D., et al. (2023). The person-to-person transmission landscape of the gut and oral microbiomes. *Nature* 614, 125–135. doi: 10.1038/s41586-022-05620-1

Ventura Spagnolo, E., Stassi, C., Mondello, C., Zerbo, S., Milone, L., and Argo, A. (2019). Forensic microbiology applications: a systematic review. *Leg. Med.* 36, 73–80. doi: 10.1016/j.legalmed.2018.11.002

Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. (2007). Navie 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

Ward, T. L., Dominguez-Bello, M. G., Heisel, T., Al-Ghalith, G., Knights, D., and Gale, C. A. (2018). Development of the human mycobiome over the first month of life and across body sites. *mSystems* 3, e00140-17. doi: 10.1128/mSystems.00140-17

Watanabe, H., Nakamura, I., Mizutani, S., Kurokawa, Y., Mori, H., Kurokawa, K., et al. (2018). Minor taxa in human skin microbiome contribute to the personal identification. *PLoS ONE* 13, e0199947. doi: 10.1371/journal.pone.0199947

Xiao, C., Pan, C., Liu, E., He, H., Liu, C., Huang, Y., et al. (2019). Differences of microRNA expression profiles between monozygotic twins' blood samples. *Forens. Sci. Int.* 41, 152–158. doi: 10.1016/j.fsigen.2019.05.003

Xu, J., Fu, G., Yan, L., Craig, J. M., Zhang, X., Fu, L., et al. (2015). LINE-1 DNA methylation: a potential forensic marker for discriminating monozygotic twins. *Forens. Sci. Int.* 19, 136–145. doi: 10.1016/j.fsigen.2015.07.014

Xu, Y., Li, T., Pu, T., Cao, R., Long, F., Chen, S., et al. (2017). Copy number variants and exome sequencing analysis in six pairs of chinese monozygotic twins discordant for congenital heart disease. *Twin Res. Hum. Genet.* 20, 521–532. doi: 10.1017/thg.2017.57

Yao, T., Han, X., Guan, T., Zhai, C., Liu, C., Liu, C., et al. (2021). Exploration of the microbiome community for saliva, skin, and a mixture of both from a population living in Guangdong. *Int. J. Legal Med.* 135, 53–62. doi: 10.1007/s00414-020-02329-6