BMC Bioinformatics

RESEARCH Open Access

A new alignment-free method for classification of fungi

Lily He[1], Mochao Huang[1†], Gulinisha Yiming[1†], Yi Zhu[1†], Ruowei Liu[1], Jinghan Chen[1] and Stephen S. T. Yau[2,3*]

† Mochao Huang, Gulinisha Yiming and Yi Zhu have contributed equally to this work.

*Correspondence: yau@uic.edu

[1] School of Science, Beijing University of Civil Engineering and Architecture, Beijing 102616, People's Republic of China
[2] Beijing Institute of Mathematical Sciences and Application, Beijing 100084, People's Republic of China
[3] Department of Mathematical Science, Tsinghua University, Beijing 100084, People's Republic of China

Abstract

As eukaryotic organisms, fungi play a pivotal role within ecosystems and exert profound influences on agriculture, the pharmaceutical industry, and human health. The classification of fungi in databases has emerged as a crucial and complex issue in the field of biology. In this study, by leveraging the local distribution of k-mer in nucleotide sequences, we introduce a novel alignment-free method, denoted as k-mer SNV, to address this challenge. On a large fungi dataset including 120,140 sequences, our innovative approach has achieved remarkable success in predicting the taxonomic labels of fungi across six hierarchical taxonomic levels: phylum (99.52%), class (98.17%), order (97.20%), family (96.11%), genus (94.14%), and species (93.32%). The approach is also evaluated on the common Taxxi benchmark dataset. Based on these results, it has been convincingly demonstrated that the k-mer SNV method exhibits outstanding performance in processing large-scale fungal sequence data.

Keywords: K-mer SNV, Fungi, Alignment-free, Classification

Introduction

Approximately 144,000 species of organisms have been documented, with fungi representing one of the most widely distributed groups on Earth and exhibiting substantial environmental and medical significance [1, 2]. Since the 1990 s, the issue of fungal classification has emerged as a critical area of focus [3, 4]. The classification of fungi can be categorized into three main types: classical, culture-based, and modern. The classical approach achieves the purpose of classifying fungi by identifying specific morphological areas [5], but requires specialized knowledge. The culture-based method identifies fungal classes by examining colonies grown in culture, but it is not suitable for fungi that cannot grow or produce reproductive structures in culture, or for those that are difficult to reproduce naturally. These methods are time-consuming and labor-intensive, leading to their declining use.

Modern methods have shifted towards DNA-based technology due to the rapid development of biotechnology. The accurate classification at each taxonomic level is crucial for future ecological and physiological studies [6]. With the advancement of molecular biology, more studies are being conducted on the classification and analysis of fungi

Journal : **BMCOne 12859**	Dispatch : **22-5-2025**	Pages : **15**
Article No : **6152**	☐ LE	☐ TYPESET
MS Code :	☑ CP	☑ DISK

based on genome sequences. DNA barcoding is widely used in species classification, and the internal transcribed spacer (ITS) region is often used for fungal identification due to its simplicity and effectiveness [7, 8].

Some researches utilize targeted sequencing analysis followed by the BLAST method [32] to identify fungi. Although these methods often provide a new means of sequencing [10–12], they require a high level of expertise, making it difficult for teams without relevant experimental foundations. There are also many approaches based on phylogenetic trees to determine fungal types [13, 15, 16]. However, phylogenetic-based methods struggle with processing big data and may yield inconsistent results due to variations in the evolutionary models employed [17]. Phenotype-based approaches [18–20]are relatively time-consuming and costly.

Machine learning-based methods have become popular in recent years [21]. Combining correlations between nucleotides with machine learning methods, Yau proposed an 18-dimensional Natural Vector approach for fungal classification [22]. However, this approach only utilizes the distribution of single nucleotide ignoring that of k-mers. Two other prominent techniques in the field of fungal classification are the Hitac method, which is based on DNA barcodes, and the Kraken2 method, which is purely based on k-mer. Hitac is a hierarchical taxonomic classifier for fungal ITS sequences [27]. On the other hand, the Kraken2 use k-mers for mapping sequences to a database for classification [26, 28]. Given the vast number of fungal taxa across multiple hierarchical levels, ranging from kingdom to species, and the uneven distribution of taxa within each class, many methods are only capable of identifying fungi at specific taxonomic levels [30, 31].Even when these methods can be applied to other levels, their accuracy is often compromised.

To overcome these challenges, we propose a new alignment-free method: K-mer Subsequence Natural Vector (K-mer SNV). The K-mer SNV method divides fungal ITS region sequences into segments, then utilizes the frequency, average positions, and variance of positions of K-mers to represent each segment. By using the distribution of K-mers, the method better adapts to the diversity of fungal sequences. In particular, this technique can classify fungi from phylum to species with high accuracy.

Materials and methods

Dataset

The dataset used in this study was downloaded on January 22, 2024 from the Bold Systems (https://portal.boldsystems.org/result?query=Fungi[tax]). Regarding data processing, for each taxonomic level, samples with fewer than 20 occurrences and species without clear classification were removed. This decision was made because, for taxonomic categories with too few samples, the model may struggle to learn meaningful patterns. Concurrently, the dataset containing fungal ITS region data was retained. Finally, a total of 120,140 barcode entries were included for this research. This dataset contains six categories: phylum, class, order, family, genus, and species, with their numbers being 118,918, 115,241, 113,683, 105,513, 92,141, and 38,646, respectively. As the taxonomic rank decreases from phylum to species, the entry counts for each level decreasing. (see Fig. 1)

Journal : BMCOne 12859 | Dispatch : 22-5-2025 | Pages : 15
Article No : 6152 | ☐ LE | ☐ TYPESET
MS Code : | ☑ CP | ☑ DISK

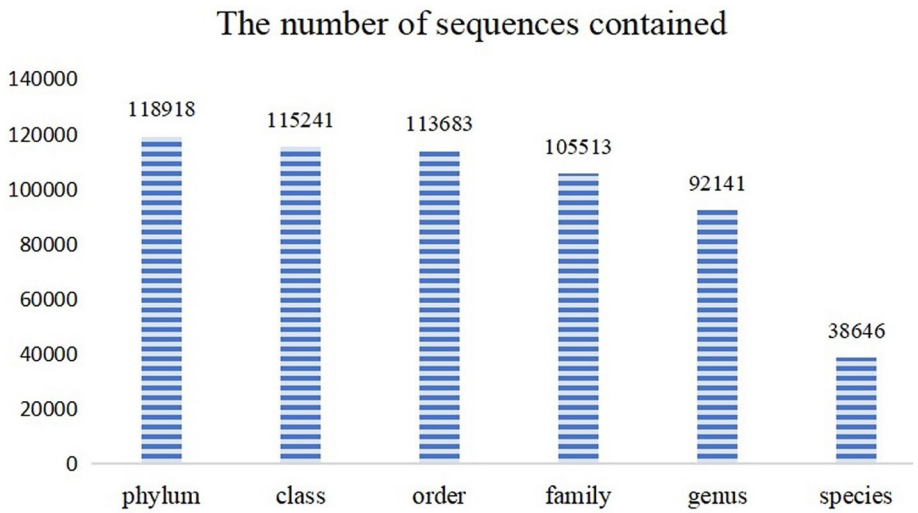

Fig. 1 Fungal ITS regional data distribution map

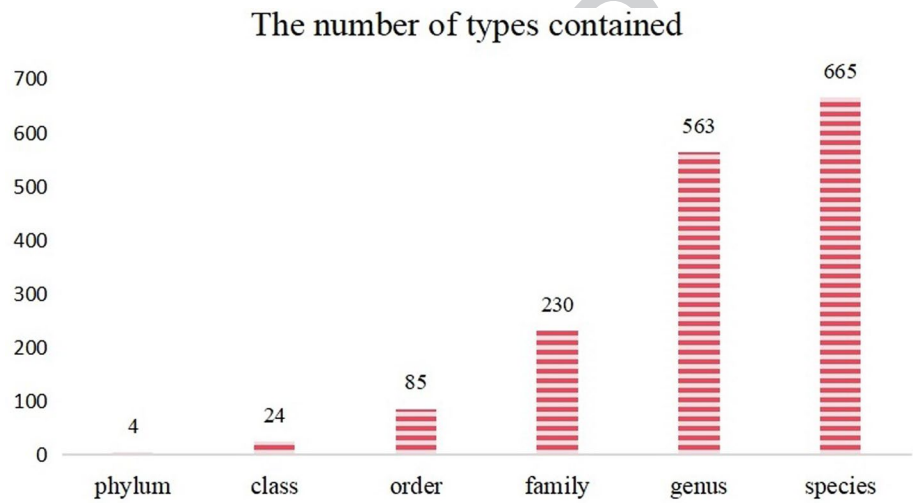

Fig. 2 Distribution map of fungal ITS regional data types

In addition, the distribution across taxonomic levels is as follows: there are 4 phyla, 24 classes, 85 orders, 230 families, 563 genera, and 665 species. (see Fig. 2)

Subsequence

We take as a example one ITS DNA sequence S as our input data. Let $S = S_1, S_2, S_3, ..., S_N, S_i \in \{A, C, G, T\}$, we first divide the sequence into L segments. The method proposed in [23] is used to make the number of nucleotides for all segments basically equal. This method divides the sequence into L non-overlapping subsequences by using the formula (1):

$$M = [\frac{N}{L}], J = N - L * M. (0 \leq J < L) \quad (1)$$

Journal : BMCOne 12859 | Dispatch : 22-5-2025 | Pages : 15
Article No : 6152 | ☐ LE | ☐ TYPESET
MS Code : | ☑ CP | ☑ DISK

where M is the quotient and J is the remainder when dividing N by L. Therefore, for the first J segments $(S_1, S_2, \cdots, S_J)$, each segment consists of $M+1$ nucleotides. At the same time, for the remaining $L-J$ segments $(S_{J+1}, S_{J+2}, \cdots, S_L)$, each segment consists of M nucleotides. L is a preset integer ($L \ll N$) which can be adjusted based on the dataset.

K-mer subsequence natural vector (K-mer SNV)

Before introducing our proposed method, it is worth reviewing the 18-dimensional Natural Vector (18-NV) for classifying fungi based on their ITS region DNA barcodes in [23]. The 18-NV method captures sequence information through the following features: the count of each nucleotide (4 features), the mean position of each nucleotide (4 features), the normalized variance of the position for each nucleotide (4 features), and the covariance between each pair of nucleotides (6 features). This method has demonstrated the potential of utilizing sequence distribution and nucleotide correlations for effective classification. Building on this concept, we introduce the K-mer Subsequence Natural Vector (K-mer SNV) method, which further enhances feature extraction efficiency and adaptability to sequence diversity.

After cutting the sequence into L segments, we calculate the K-mer values for each subsequence. We begin by introducing the concept of a K-mer in the DNA sequence. A K-mer refers to a sequence of K nucleotides. For instance: there are 4 possibilities for 1-mer: A, C, G, T ($K=1$), 16 possibilities for 2-mer: AA, AC, AT, AG, CA, CC, CT, CG, GA, GC, GG,GT, TA, TC, TG, TT ($K=2$).Continuing in this manner, when $K=k$, there will be 4^k possibilities of combining.Next, we calculate the K-mer Subsequence Natural Vector (K-mer SNV), which is ultimately used as the feature vector for the sequence. Before calculating the feature, we define an indicative function:

$$w_\alpha(s_i) = \begin{cases} 1, & s_i = \alpha \\ 0, & s_i \neq \alpha \end{cases} \quad i = 1, 2, \cdots, N. \tag{2}$$

Where, $\alpha \in \{x \mid x:$ all combinations of k-mer $\}$, for example, when K=2, $\alpha \in \{AA, AC, AT, AG, CA, CC, CT, CG, GA, GC, GG, GT, TA, TC, TG, TT\}$.

Subsequently, for each α in each subsequence, three statistics are used for feature extraction:

1. Let $n_\alpha = \sum_{i=1}^{n} \omega_\alpha(s_i)$ describe the number of α.
2. Let $\mu_\alpha = \sum_{i=1}^{n} \frac{\omega_\alpha(s_i) * i}{n_\alpha}$ be the mean position of α.
3. Let $D_\alpha = \sum_{i=1}^{n} \frac{(i-\mu_\alpha)^2 \omega_\alpha(s_i)}{n_\alpha n}$ be the normalized second central moment of position of α.

Consequently, when the sequence is separated into L subsequences: $S_1, S_2, \cdots, S_L$, we ultimately obtain an $L * 3 * 4^k$ dimensional numeric vector, (see (3)), and use it as a feature vector. This vector called the K-mer Subsequence Natural Vector(K-mer SNV).

Journal : BMCOne 12859 | Dispatch : 22-5-2025 | Pages : 15
Article No : 6152 | ☐ LE | ☐ TYPESET
MS Code : | ☑ CP | ☑ DISK

$$\begin{bmatrix} n^{S_1}_{\underbrace{AA\cdots A}_{k}}, n^{S_1}_{\underbrace{AC\cdots A}_{k}}, \cdots, n^{S_1}_{\underbrace{TT\cdots T}_{k}}, \\ \mu^{S_1}_{\underbrace{AA\cdots A}_{k}}, \mu^{S_1}_{\underbrace{AC\cdots A}_{k}}, \cdots, \mu^{S_1}_{\underbrace{TT\cdots T}_{k}}, \\ D^{S_1}_{\underbrace{AA\cdots A}_{k}}, D^{S_1}_{\underbrace{AC\cdots A}_{k}}, \cdots, D^{S_1}_{\underbrace{TT\cdots T}_{k}}, \\ n^{S_2}_{\underbrace{AA\cdots A}_{k}}, n^{S_2}_{\underbrace{AC\cdots A}_{k}}, \cdots, n^{S_2}_{\underbrace{TT\cdots T}_{k}}, \\ \mu^{S_2}_{\underbrace{AA\cdots A}_{k}}, \mu^{S_2}_{\underbrace{AC\cdots A}_{k}}, \cdots, \mu^{S_2}_{\underbrace{TT\cdots T}_{k}}, \\ D^{S_2}_{\underbrace{AA\cdots A}_{k}}, D^{S_2}_{\underbrace{AC\cdots A}_{k}}, \cdots, D^{S_2}_{\underbrace{TT\cdots T}_{k}}, \\ \cdots\cdots \\ n^{S_L}_{\underbrace{AA\cdots A}_{k}}, n^{S_L}_{\underbrace{AC\cdots A}_{k}}, \cdots, n^{S_L}_{\underbrace{TT\cdots T}_{k}}, \\ \mu^{S_L}_{\underbrace{AA\cdots A}_{k}}, \mu^{S_L}_{\underbrace{AC\cdots A}_{k}}, \cdots, \mu^{S_L}_{\underbrace{TT\cdots T}_{k}}, \\ D^{S_L}_{\underbrace{AA\cdots A}_{k}}, D^{S_L}_{\underbrace{AC\cdots A}_{k}}, \cdots, D^{S_L}_{\underbrace{TT\cdots T}_{k}} \end{bmatrix}_{L*3*4^k} \quad (3)$$

(3)

In Supplementary File 1, we provide a detailed demonstration of the calculation process.

At this point, we successfully convert a DNA sequence into a numeric feature vector, known as the K-mer Subsequence Natural Vector (K-mer SNV). This vector will serve as our input for the machine learning model. In the next step, we will utilize machine learning algorithms to classify fungi, as illustrated in the flowchart (see Fig. 3)

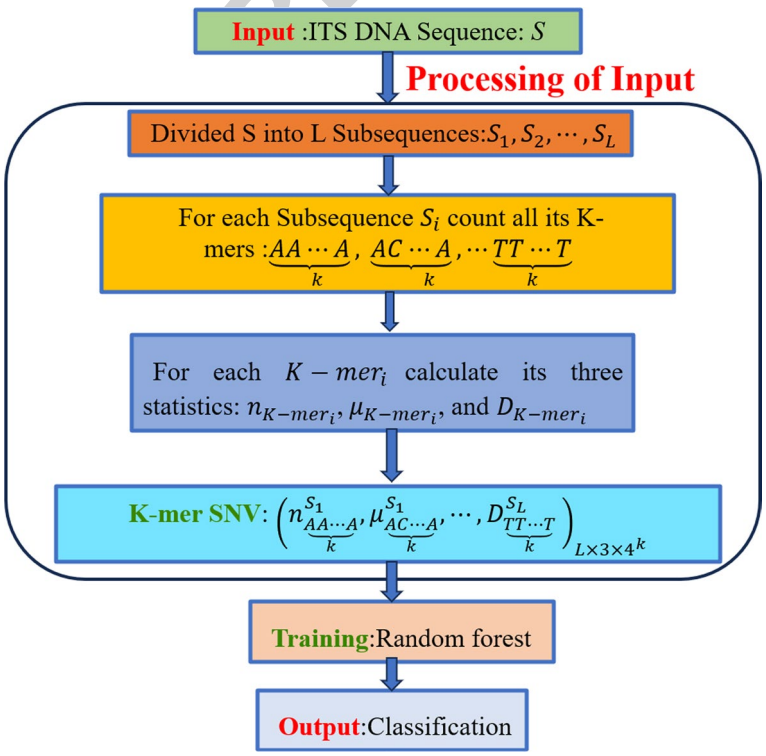

Fig. 3 Flowchart of the K-mer SNV method

Supervised classification

Due to the complexity and data characteristics of the fungal datasets, we have chosen the random forest algorithm as the classifier for our study. Random forest [24] is a widely-used supervised learning method, which employs an ensemble learning-based approach that performs the classification task by constructing multiple decision trees and synthesizing their predictions.

To ensure the robustness of our model and to prevent data leakage, we have carefully de-duplicated the dataset to ensure that identical sequences do not appear in both the training and test sets. We allocate 80% of the data to the training set and 20% to the testing set respectively, ensuring that there is no overlap between the training and testing datasets. Next, we train the model, where only the training set participates in the training phase, and the test set is an independent set that is used to test the results of the model. We employ 5-fold cross-validation for both modeling and testing, which further enhances the reliability of our evaluation. The parameters of the random forest are optimized through a grid search, ensuring that our model benefits from the most effective hyperparameter configuration. The output of our method is a classification result that categorizes the fungal sequences into various taxonomic levels, specifically: phylum, class, order, family, genus, and species.

In this study, we employed several evaluation metrics to assess the performance of our Random Forest model, beginning with accuracy. Accuracy [14] is defined as the ratio of correctly predicted samples to the total number of samples. Specifically, it is calculated by dividing the number of correctly classified samples by the total number of samples in the dataset. The formula for accuracy is:

$$\text{Accuracy} = \frac{\text{Number of Correctly Classified Samples}}{\text{Total Number of Samples}}. \tag{4}$$

Correctly classified samples are those for which the model's predictions match the actual labels, while the total number of samples refers to all the samples used for evaluation. To compute accuracy, we compare the model's predicted outputs with the true labels, count the number of samples that were classified correctly, and divide this count by the total number of samples.

Accuracy is a commonly used and straightforward metric to evaluate classification models, as it provides a general measure of the model's performance. In our study, accuracy effectively reflects the overall performance of the model in classifying ITS DNA sequences. However, we recognize that accuracy alone may not fully capture the model's performance, especially in cases where the dataset is imbalanced. For this reason, we also employ additional evaluation metrics to provide a more comprehensive understanding of the model's effectiveness.

In addition to accuracy, we calculate the F1-score, which is the harmonic mean of precision and recall. The F1-score [14] is calculated using the following formula:

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{5}$$

The F1-score provides a balanced measure between precision and recall, particularly useful when there is an uneven class distribution. We also use the Area Under the Curve

Journal : BMCOne 12859 | Dispatch : 22-5-2025 | Pages : 15
Article No : 6152 | ☐ LE | ☐ TYPESET
MS Code : | ☑ CP | ☑ DISK

(AUC), which measures the model's ability to distinguish between different classes. AUC is derived from the Receiver Operating Characteristic (ROC) curve and provides a summary of the model's classification performance across different thresholds.

Lastly, we assess recall (also known as sensitivity), which calculates the proportion of actual positives correctly identified by the model. The recall [14] is calculated as:

$$\text{Recall (Sensitivity)} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}. \tag{6}$$

Recall is particularly important when we need to minimize false negatives in classification tasks, ensuring that as many relevant samples as possible are correctly identified.

Together, these metrics-accuracy, F1-score, AUC, and recall-offer a thorough evaluation of our model's performance, addressing various aspects of classification accuracy, precision, and reliability.

The source code has been made publicly available on GitHub. You can access it through the following link https://github.com/xxxx.

Results

Selection of parameters *K* and *L*

In our study, K and L are parameters that need to be determined based on the dataset of the experiment. To determine the optimal values for parameters K and L, we compute e a series of accuracies for different K and L. For the Phylum dataset, using the random forest method with 20 trees, we adjusted the values of K and L, calculating the prediction accuracy for the test set and selecting the parameter values corresponding to the highest accuracy, as shown in Fig. 4. We observed that accuracy initially increased and then decreased, with the model achieving its peak accuracy of 99.35%. when $K = 4$ and $L = 3$. Consequently, we settled on $K = 4$ and $L = 3$ for our model. These parameters, once optimized in the training set, were then applied to the test set to evaluate the model's performance.This approach was consistently applied to other datasets, which also demonstrated peak performance with$K$ and L set to 3 and 4, respectively.

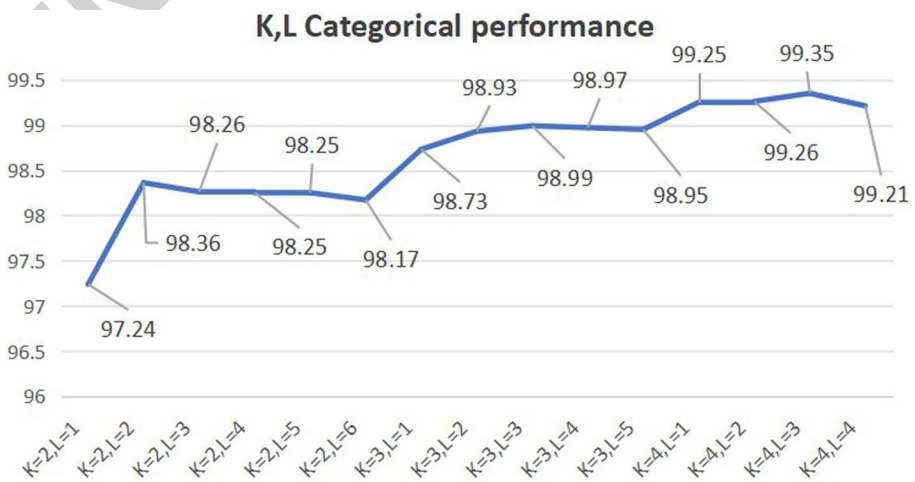

Fig. 4 On our dataset, the prediction accuracy of Phylum dataset with changes in *K* and *L*

Journal : BMCOne 12859 | Dispatch : 22-5-2025 | Pages : 15
Article No : 6152 | ☐ LE | ☐ TYPESET
MS Code : | ☑ CP | ☑ DISK

Furthermore, Fig. 5 illustrates the changes in the AUC values on the species dataset with variations in K and L. It can be seen from the graph that the AUC value generally shows an upward trend with the increase of K and L, which further supports the rationale behind our selection of the K and L parameters. In particular, the AUC value reached its highest point when K=4,L=3, which is consistent with our previous findings on the phylum dataset and further substantiates our parameter choices. These results indicate that by adjusting the K and L parameters, we can effectively enhance the model's classification performance.

Regarding the choice of L and K, due to computational resource limitations, the parameter range we selected is within what can be handled by the current hardware. Larger values of L or K would significantly increase the computational burden, especially when processing large amounts of data on a standard laptop, leading to higher time and memory costs. Through experimentation, we found that the performance reached a plateau near K=4 and L=3. Increasing the parameters further did not result in significant improvements. With more powerful computational resources, increasing the parameters might be beneficial, but under the current conditions, we believe the chosen settings already provide satisfactory results.

Computations were performed on a personal computer with an Intel Core i5-7200U CPU @ 2.50 GHz and 8 GB RAM.

Classification performance

In order to demonstrate the advantages of our method in the classification of fungi, we initially assessed its predictive performance across each taxonomic level using aggregated data from each level and compared it with the 18-NV method [23]. As illustrated in Table 1, the 18-NV method analyzes fungi species by leveraging correlations between nucleotides, relying on the convex hull method. In contrast, our method has yielded promising results across various datasets. Although the number

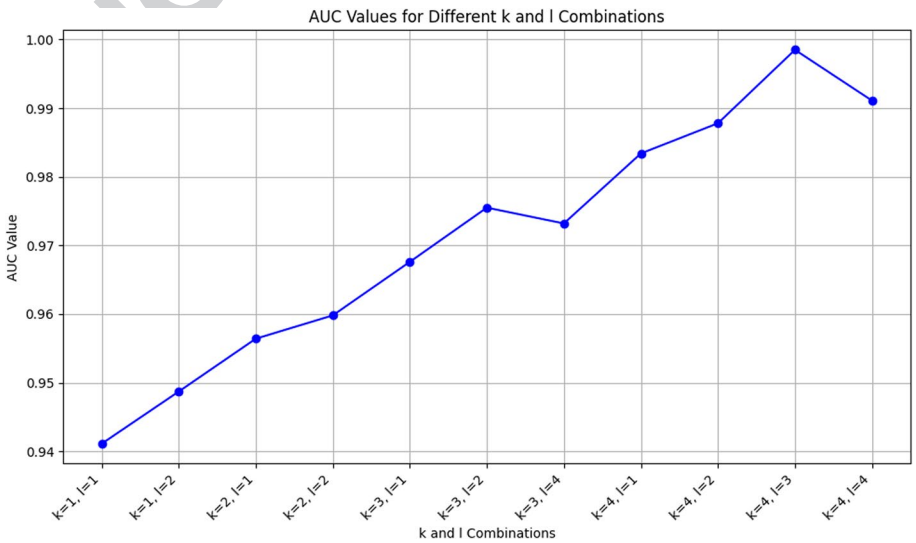

Fig. 5 On our dataset, the AUC at the Species dataset with changes in K and L

Journal : BMCOne 12859 | Dispatch : 22-5-2025 | Pages : 15
Article No : 6152 | ☐ LE | ☐ TYPESET
MS Code : | ☑ CP | ☑ DISK

Table 1 Classification prediction results of fungal at each category

Dataset	Sequence number	Type number	18-NV (%)	Kraken2 (%)	Hitac (%)	BTOP (%)	TOP (%)	K-mer SNV (%)
Phylum	118,918	4	96.17	67.01	99.75	84.15	98.11	99.52
Class	115,241	24	91.41	45.44	97.45	83.40	96.75	98.17
Order	113,683	85	87.10	29.12	96.40	82.46	94.18	97.20
Family	105,513	230	82.99	15.05	92.27	80.02	87.90	96.11
Genus	92,141	563	81.44	13.82	82.89	75.75	81.90	94.14
Species	38,646	665	84.68	6.94	56.42	0.00	11.07	93.32

of categories increases and becomes uneven from Phylum to Species, leading to a decrease in prediction accuracy, our method still maintains an accuracy above 93%.

In our study, when using the method of only building fungi libraries, running according to the program of kraken2 build always fails to complete, and it seems that the required fungi sequence cannot be found on the NCBI server. We took the approach of building the entire library, but it didn't always work out due to network issues. Finally, we download the database from the https://benlangmead.github.io/aws-indexes/k2 website. We have selected Standard plus Refeq protozoa & fungi which contains the latest kraken2 database with fungi released on 2024-12-28. The database is k2_pluspf_20241228.tar.gz, 70GB in compressed package, 92GB after decompression. At the same time, we compared it with the BTOP [32], which is a BLAST-based method, and TOP, which is a usearch-based techique. [33] The outcomes of our experiments are presented in Table 1.We compared our method with these existing ones at various taxonomic levels. The performance of Hitac showed a significant improvement over the previous 18-NV method. However, as shown in Table 1, our proposed method demonstrated better advantages at all levels.

Additionally, the time required to complete this work at each level is less than 30 min. These calculations were conducted on a personal computer equipped with an Intel Core i5-7200U CPU @ 2.50 GHz and 8 GB of RAM, showcasing the efficiency and feasibility of our method even on standard hardware.

From the Fig. 6, The 18-NV method shows a high performance with AUC values starting at 0.99 for the Phylum level and slightly decreasing to 0.96 at the Species level. On the other hand, the K-mer SNV method demonstrates perfect or near-perfect performance across all taxonomic levels, with AUC values ranging from 0.99 to 1.00. This indicates that while both methods perform well, the K-mer SNV method is more consistent and robust, especially as the classification becomes more refined. The slight decrease in AUC for the 18-NV method from Phylum to Species suggests that it may be slightly less effective at handling the increased complexity and uneven distribution of categories at finer taxonomic levels. In contrast, the K-mer SNV method maintains high accuracy, indicating its superior capability in classifying fungi across different taxonomic levels. This analysis highlights the advantages of the K-mer SNV method in fungal classification, particularly its ability to maintain high performance even as the classification task becomes more challenging.

Journal : BMCOne 12859 | Dispatch : 22-5-2025 | Pages : 15
Article No : 6152 | ☐ LE | ☐ TYPESET
MS Code : | ☑ CP | ☑ DISK

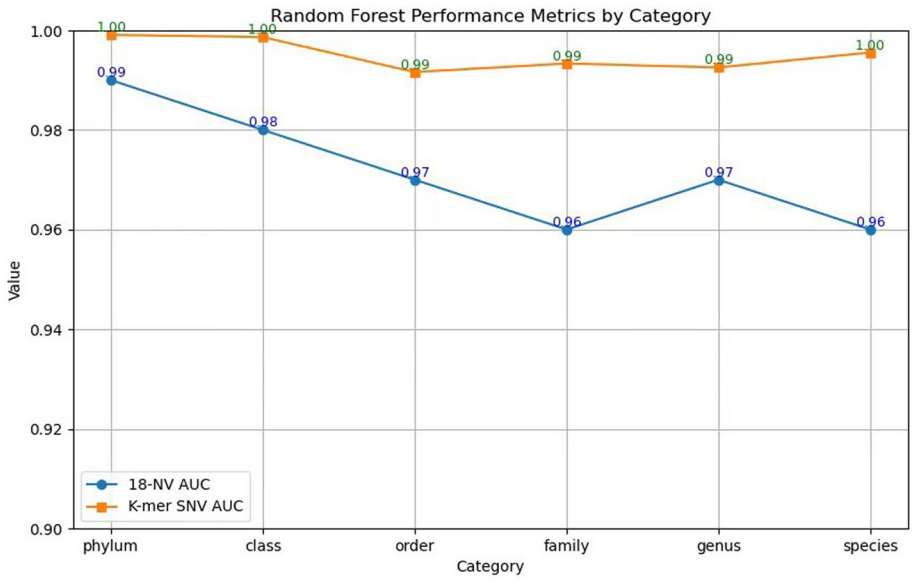

Fig. 6 On our dataset, comparison of the AUC between the 18-NV method and our K-mer SNV method

Table 2 Classification prediction results of fungal phylum by K-mer SNV with Random Fforest (RF) and Logistic Regression (Lg) methods with $K = 4$ and $L = 3$.

Phylum	Number	Length range	Precision		Recall		F1-score		Accuracy	
			RF	Lg	RF	Lg	RF	Lg	RF	Lg
Ascomycota	72,385	(1021, 1066)	1	0.99	1	1	1	0.99	0.996	0.994
Basidiomycota	40,204	(1011, 1055)	0.99	0.99	0.99	0.99	0.99	0.99	0.994	0.991
Glomeromycota	3,499	(1031, 1067)	1	0.99	0.98	0.99	0.99	0.99	0.999	0.991
Zygomycota	2,830	(1015, 1056)	1	0.98	0.96	0.97	0.98	0.98	0.996	0.983

Classification of phylum

For the Phylum, there are 118,918 barcodes. This dataset is divided into four categories: Ascomycota, Basidiomycota, Glomeromycota, and Zygomycota. For more information about the number and length range of datasets, see Table 2. As can be seen from the Table 2, there is little difference in the length of each fungal phyla, but there is a significant difference in the amount of Ascomycota and Basidiomycota compared to Glomeromycota, and Zygomycota. Although the amount of data varies widely, the results are good. For each class, the values of the four quantities Precision, Recall, F1 score, and Accuracy are close to 1. After calculating the k-mer SNV, we used the Random Forest (RF) and Logistic Regression (Lg) classifiers for verification, and from the results, it can be seen that the Random Forest and Logistic Regression classifiers have little impact on the results, indicating that the K-mer SNV encoding effect plays a major role in data prediction. Meanwhile, this proves that our method works even when the amount of data is uneven, and that our method is robust.

Journal : BMCOne 12859 | Dispatch : 22-5-2025 | Pages : 15
Article No : 6152 | ☐ LE | ☐ TYPESET
MS Code : | ☑ CP | ☑ DISK

Classification of class

Unlike Phylum, Class has 115,241 barcodes, and it contains 24 classes, and the number of classes is even more uneven, the largest number of class includes 34,307 barcodes, however, the class contains only 21 barcodes, the difference between the two is more than 1,000 times. Still, the results of our algorithm at the level of Class have been shown to be very effective, see Table 3. Except for Agaricostilbomycetes, which has a relatively small number, it is only 0.8 on the Precision indicator, and everything else is close to 1.

The remaining prediction results are included in the Supplementary Files 2 (Order, Family, Genus, and Species).

Discussion

Our method was evaluated on the Taxxi (sp_rdp_its.90) dataset and compared with other existing methods. Note that this dataset is the most difficult to analyze among five fungi ITS datasets in the TAXXI benchmark.

Using the hierarchical precision, recall and F1-score defined in [27], our method achieved an F1-score of 89, demonstrating good performance, yet there is still room for improvement. Figure 7 presents a comprehensive comparison between k-mer SNV and other state-of-the-art methods. Our method's F1-score of 89 indicates a satisfactory balance between precision and recall. Although this score is slightly lower than that of the best-performing methods, it still showcases the potential and

Table 3 Classification prediction results of fungal class by K-mer SNV method with $K = 4$ and $L = 3$

Class	Number	Length range	Precision	Accuracy
Agaricomycetes	34,307	(632,664)	0.98	0.976
Agaricostilbomycetes	61	(611,673)	0.8	1
Archaeorhizomycetes	69	(622,669)	1	1
Arthoniomycetes	561	(632,665)	1	1
Atractiellomycetes	30	(611,674)	1	1
Dacrymycetes	28	(622,670)	1	1
Dothideomycetes	12,633	(632,666)	0.98	0.981
Eurotiomycetes	9,479	(611,675)	0.98	0.979
Exobasidiomycetes	286	(622,671)	1	1
Glomeromycetes	3,420	(632,667)	1	0.994
Lecanoromycetes	12,148	(611,676)	0.97	0.972
Leotiomycetes	4,513	(622,672)	0.97	0.976
Microbotryomycetes	615	(632,668)	0.98	0.982
Pezizomycetes	3,197	(611,677)	0.99	0.992
Pneumocystidomycetes	59	(622,673)	1	1
Pucciniomycetes	2,155	(632,669)	1	1
Saccharomycetes	4,125	(611,678)	0.99	0.999
Sordariomycetes	22,618	(622,674)	0.99	0.990
Taphrinomycetes	29	(632,670)	1	1
Tremellomycetes	1,852	(611,679)	0.98	0.980
Trichomycetes	21	(622,675)	1	1
Ustilaginomycetes	264	(632,671)	1	1
Wallemiomycetes	65	(611,680)	1	1
Zygomycetes	2,706	(622,676)	1	0.992

Journal : BMCOne 12859 | Dispatch : 22-5-2025 | Pages : 15
Article No : 6152 | ☐ LE | ☐ TYPESET
MS Code : | ☑ CP | ☑ DISK

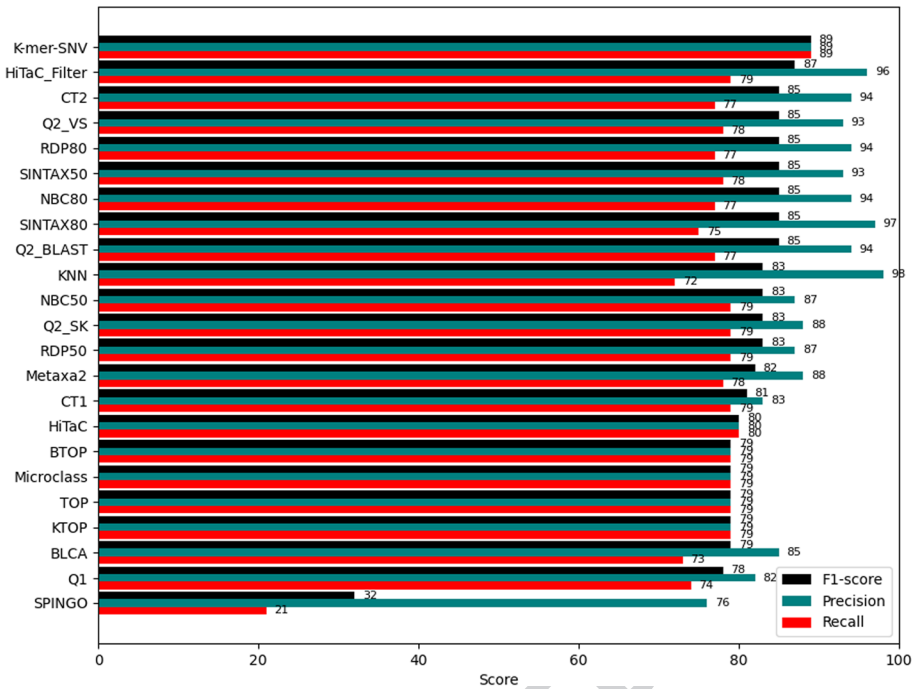

Fig. 7 Comparison of methods on the Taxxi (sp_rdp_its.90) dataset

competitiveness of our approach. One of the strengths of our method is the flexibility in parameter adjustment. By tuning the k and l parameters, we can fine-tune the model's performance. Despite the current F1-score already indicating strong performance, further optimization of these parameters could lead to even higher performance.

While the results are promising, our method, like any other, has its limitations. The selection of K and L parameters is crucial and may require manual adjustment based on the specific characteristics of different datasets. Future work could focus on developing automated parameter selection strategies, potentially leveraging machine learning techniques to predict optimal parameter values based on dataset attributes.

In addition to this, our methods can also be applied to the phytogenetic analysis of fungi. After calculating the corresponding K-mer SNV, the following Fig. 8 of the UPGMA [25] phylogenetic tree is obtained using MEGA software [29]. Through the phylogenetic tree, it can be seen that our method can show the evolutionary relationship of fungi very well.

In summary, we propose a new computational method that can be effectively applied to the identification of fungi. Traditional sequence alignment methods are widely used, but they are time-consuming and require a high-performance computer to process large data sets. However, our approach is effective in overcoming these problems. Compared to 18-NV, although our method achieves good classification results on all layers, we still need to optimize our method for some layers, such as genus and species. Our approach also shows great advantages in some sub-categories.

Journal : BMCOne 12859 | Dispatch : 22-5-2025 | Pages : 15
Article No : 6152 | ☐ LE | ☐ TYPESET
MS Code : | ☑ CP | ☑ DISK

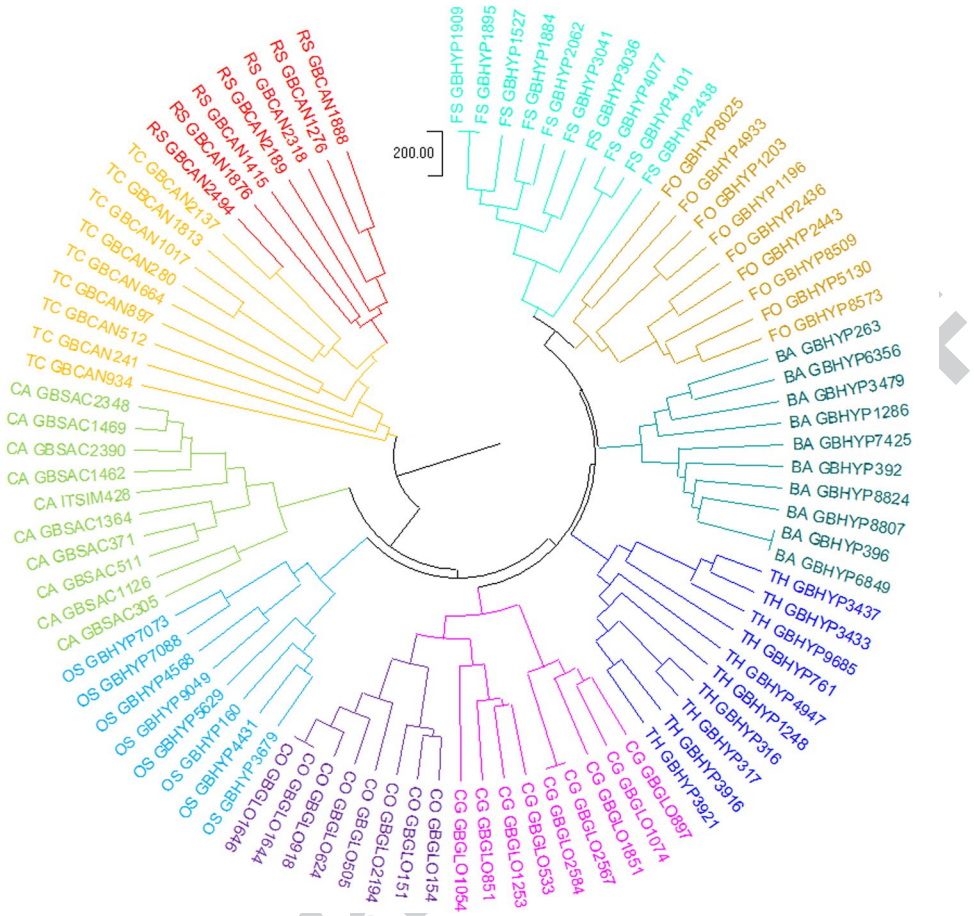

Fig. 8 UPGMA phylogenetic tree of ITS sequences from 90 fungal Species in our dataset, constructed using the K-mer SNV method with K=4,L=3

Conclusions

Moreover, in the selection of parameters K and L, our method needs to do more comprehensive experiments to explore the problem of parameter selection. At the same time, we need to see if our approach can be applied to other species classifications. These problems should be investigated in further study.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-025-06152-x.

Supplementary file 1.

Supplementary file 2.

Acknowledgements

This study is supported by the National Natural Science Foundation of China (Z23146), and Beijing Municipal Education Commission (Z22053 and Z22055) for providing excellent research environment while part of this research was done. This work is supported in part by funds from National Natural Science Foundation of China (NSFC) grant (12171275) Tsinghua University Education Foundation fund (042202008).

Author contributions

Conceptualization, L.H., S.T. Yau; methodology, M.H., L.H., G.Y., R.L., J.C.; software, M.H.; validation, M.H., G.Y; formal analysis, G.Y.; investigation, R.L.; resources, J.C.; data curation, L.H.; writing—original draft preparation, G.Y.; writing—review and editing, L.H., Y.Z.; visualization, G.Y.; supervision, M.H.; project administration, L.H.; funding acquisition, L.H., S.S.-T.Y.; All authors have read and agreed to the published version of the manuscript.

	Journal : **BMCOne 12859**	Dispatch : **22-5-2025**	Pages : **15**
	Article No : **6152**	☐ LE	☐ TYPESET
	MS Code :	☑ CP	☑ DISK

Availability of data and materials
The datasets analysed during the current study are available in the [Bold Systems] repository, [https://portal.boldsystems.org/result?query=Fungi[tax]].

Declarations

Competing of interest
The authors declare that they have no Conflict of interest.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Received: 19 November 2024 Accepted: 30 April 2025

References

1. Moore D, Alexopoulos CJ, et al. "Fungus". Encyclopedia Britannica, 2024-07-21, https://www.britannica.com/science/fungus
2. Lücking R, Aime MC, Robbertse B, et al. Fungal taxonomy and sequence-based nomenclature. Nat Microbiol. 2021;6:540–8.
3. Gautam AK, Verma RK, et al. Current insight into traditional and modern methods in fungal diversity estimates. J Fungi. 2022;8(3):226.
4. Lin P, Kook M, Yi TH, et al. Current fungal taxonomy and developments in the identification system. Curr Microbiol. 2023;80:375.
5. Hoefnagels MH. Biodiversity of fungi: inventory and monitoring methods. BioScience. 2005;55(3):282–3.
6. Tedersoo L, Smith ME. Ectomycorrhizal fungal lineages: detection of four new groups and notes on consistent recognition of ectomycorrhizal taxa in high-throughput sequencing studies. Ecol Stud. 2017;30:125–42.
7. Schoch CL, Seifert KA, et al. Fungal barcoding consortium; fungal barcoding consortium author list. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal dna barcode marker for fungi. Proceedings of the national academy of sciences of the United States of America. 2012;109(16):6241-6.
8. Irinyi L, Serena C, Garcia-Hermoso DD, et al. International society of human and animal mycology (ISHAM)-ITS reference DNA barcoding database-the quality controlled standard tool for routine identification of human and animal pathogenic fungi. Med Mycol. 2015;53(4):313–37.
9. Altschul SF, Gish W, Miller W. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.
10. Langsiri N, Worasilchai N, Irinyi L, et al. Targeted sequencing analysis pipeline for species identification of human pathogenic fungi using long-read nanopore sequencing. IMA Fungus. 2023;14(1):18.
11. Jain M, Olsen HE, Paten B, et al. The Oxford nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol. 2016;17(1):239.
12. Lücking R, Aime MC, Robbertse B, et al. Unambiguous identification of fungi: where do we stand and how accurate and precise is fungal DNA barcoding. IMA Fungus. 2020;11:14.
13. Yuanning L, Jacob SLL, Ying Changwald C, et al. A genome-scale phylogeny of the kingdom Fungi. Curr Biol. 2021;31(8):1653–65.
14. Rainio O, Teuho J, Klén R. Evaluation metrics and statistical tests for machine learning. Sci Rep. 2025;14(1):1–14. https://doi.org/10.1038/s41598-024-56706-x.
15. Mehta N, Jadhav R, Baghela A, Molecular taxonomy A, multigene phylogeny of filamentous fungi. In: Gupta VK, Tuohy M, editors. Laboratory protocols in fungal biology. Fungal Biology. Cham: Springer; 2022.
16. Tedersoo L, Snchez-Ramrez S, Kljalg U, et al. High-level classification of the fungi and a tool for evolutionary ecological analyses. Fungal divers. 2018;90:135–59.
17. Wang Y, Dong W, Liang Y, Lin W, Chen J, Henry R, Chen F. PhyloForge: unifying micro- and macroevolution with comprehensive genomic signals. First published: 26 November 2024.
18. Hewitt SK, Foster DS, Dyer PS, et al. Phenotypic heterogeneity in fungi: importance and methodology. Fungal Biol Rev. 2016;30(4):176–84.
19. Smith TJ, Donoghue PCJ. Evolution of fungal phenotypic disparity. Nat Ecol Evol. 2022;6:1489–500.
20. Persoh D. Plant-associated fungal communities in the light of meta'omics. Fungal Divers. 2015;75:1–25.
21. Liu Z, Li Y. Fungi classification in various growth stages using shortwave infrared (SWIR) spectroscopy and machine learning. J Fungi. 2022;8:978.
22. Zhao X, Tian K, Yau SST. A new efficient method for analyzing fungi species using correlations between nucleotides. BMC Evol Biol. 2018;18:1–13.
23. Zhao B, He RL, Yau SS-T. A new distribution vector and its application in genome clustering. Mol Phylogenet Evol. 2011;59:438–43.
24. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
25. Loewenstein Y, Portugaly E, Fromer M, et al. Efficient algorithms for accurate hierarchical clustering of huge datasets: tackling the entire protein space. Bioinformatics. 2008;24(13):i41–9.
26. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019;20(257):1–13. https://doi.org/10.1186/s13059-019-1891-0.

	Journal : **BMCOne 12859**	Dispatch : **22-5-2025**	Pages : **15**
	Article No : **6152**	☐ LE	☐ TYPESET
	MS Code :	☑ CP	☑ DISK

27. Miranda FM, Azevedo VC, Ramos RJ, Renard BY, Piro VC. Hitac: a hierarchical taxonomic classifier for fungal ITS sequences compatible with QIIME2. BMC Bioinform. 2024;25(228):1–13. https://doi.org/10.1186/s12859-024-05839-x.
28. Lu J, Salzberg SL. Ultrafast and accurate 16S rRNA microbial community analysis using Kraken 2. Microbiome. 2020;8:124.
29. Tamura K, Stecher G, Kumar S. MEGA11: molecular evolutionary genetics analysis version 11. Mol Biol Evol. 2021;38(7):3022–7.
30. Nilsson RH, Abarenkov K, Ryberg M, et al. The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. Nucleic Acids Res. 2019;47(D1):D259–64.
31. Lücking R, Dentinger BM, Lumbsch HT. Fungal taxonomy and sequence-based nomenclature. Nat Microbiol. 2020;5(5):540–8.
32. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.
33. Edgar RC. Accuracy of taxonomy prediction for 16s RRNA and fungal its sequences. PeerJ. 2018;6:e4652.

Publisher's Note

	Journal : **BMCOne 12859**	Dispatch : **22-5-2025**	Pages : **15**
	Article No : **6152**	☐ LE	☐ TYPESET
	MS Code :	☑ CP	☑ DISK