Jiang *et al. Plant Methods* (2025) 21:61
https://doi.org/10.1186/s13007-025-01380-x

Plant Methods

METHODOLOGY Open Access

SSR_VibraProfiler: a Python package for accurate classification of varieties using SSRs with intra-variety specificity and inter-variety polymorphism

Chenhao Jiang[1†], Chuan Dong[1†], Zhenzhen Wu[2], Chenyi Shi[1], Qiannan Ye[2], Xiaopei Wu[1], Siyi Ma[3], Yuming Wen[4], Guoping Yu[4,5], Jiasheng Wu[1*] and Chengjun Zhang[1,2,5*]

Abstract

Background Simple sequence repeats (SSRs) are widely used as molecular markers; however, traditional development of SSR molecular markers heavily relies on experimental methods. The advancement of modern sequencing technology has provided the possibility of directly extracting SSR characteristics from sequencing data and using them for variety identification.

Results We have developed a computational framework for variety identification, treating the presence or absence of each SSR in sequencing data as a numerical characteristic while ignoring specific loci, flanking sequences, and occurrence counts. Therefore, subsequent variety identification does not rely on experimental validation but is directly performed based on the numerical characteristic matrix. Using a formula, we measure the variance of these numerical characteristics both within and among varieties, and select SSRs that exhibit intra-variety specificity and inter-variety polymorphism, forming a 0,1 matrix. We use t-SNE (t-distributed Stochastic Neighbor Embedding) to project the matrix onto a two-dimensional plane, followed by K-means clustering of the individuals. The classification performance of the matrix is preliminarily assessed by comparing the cluster labels with the true labels, providing an initial evaluation of its effectiveness in variety detection. Ultimately, we construct a recognition model based on the SSRs matrix and apply it for variety identification. The process has been encapsulated into the package SSR_VibraProfiler, which can serve as a tool for constructing an SSR variety DNA fingerprint database. We tested this package on a *Rhododendron* dataset that included 40 individuals from 8 varieties. The accuracy achieved through t-SNE dimensionality reduction and K-means clustering was 100%. Furthermore, we used the leave-one-out method to validate the accuracy of our method in predicting variety, and confirmed the reliability of our method in detecting varieties. The package is freely available at https://github.com/Olcat35412/SSR_VibraProfiler.

†Chenhao Jiang and Chuan Dong contributed equally to this work.

*Correspondence:
Jiasheng Wu
wujs@zafu.edu.cn
Chengjun Zhang
zhangcj@zafu.edu.cn

Full list of author information is available at the end of the article

Conclusion We introduced SSR_VibraProfiler, a Python package for distinguishing and predicting individual varieties without a reference genome by extracting SSR numerical characteristics from next-generation sequencing data. This tool will contribute to the development, identification, and protection of new varieties.

Keywords SSRs, Variety identification, In silico-based method, *Rhododendron*

Background

Simple sequence repeats (SSRs), also known as microsatellites, are motifs consisting of 1–6 bases that repeat multiple times and are found widely across the genomes of both prokaryotic and eukaryotic species [1–3]. Due to the high polymorphism, SSRs have been widely used in molecular marker studies at different taxonomic levels, including species [4], varieties [5, 6] and population structures [7, 8]. According to the repeat patterns of SSRs, they can be classified into four major categories: simple perfect, where the repeat sequence is continuous and uninterrupted by any base outside the motif; simple imperfect, consisting of one or more repeat units of varying lengths; compound perfect, composed of two or more distinct repeat motifs of the same length; and compound imperfect, where different-length repeats disrupt the motif continuity [9]. Regardless of the repeat patterns in SSR molecular markers, the time consumption and financial costs involved in primer synthesis and gel electrophoresis are drawbacks of conventional molecular marker development [10, 11]. With the rapid development of sequencing technology and bioinformatics, it has become possible to directly utilize the information from sequencing datasets for marker development and individual identification. However, most current in silico-based methods rely on either a single reference genome with a corresponding resequencing dataset [12] or multiple reference genomes [13]. Their primary goal is to identify SSRs with well-defined locations and distinct polymorphisms. On the one hand, they are not applicable in the absence of a reference genome. On the other hand, determining individual label using these SSRs still requires experimental validation, including primer synthesis and gel electrophoresis. In this study, we introduced a method that directly utilized the presence or absence of each SSR as its numerical characteristic from short-read data, without focusing on the specific locations or occurrence frequencies, thereby eliminating the need for a reference genome. We then highlight SSRs with intra-specific specificity and inter-specific polymorphism, facilitating variety classification and identification. We have integrated this method into a Python package named SSR_VibraProfiler. This method may applicable to species with numerous varieties, where reference genomes are lacking.

A potential target for utilizing the method is the variety identification under genus *Rhododendron*. The genus *Rhododendron*, belonging to the family *Ericaceae*, comprises over 1,000 species [14]. As of 2018, the International Rhododendron Register and Checklist lists over 28,000 cultivated varieties within the genus *Rhododendron* [15]. Due to their outstanding ornamental characteristics, such as diverse flower forms and vibrant colors, *Rhododendron* plants are highly valued in horticulture and possess significant economic value. Many ornamental hybrid varieties have been developed through long-term cultivation, leading to *Rhododendron* plants being widely grown around the world as important ornamental plants [16]. Therefore, there exists a practical necessity for the classification and prediction of *Rhododendron* varieties. In this study, we selected 8 varieties and 40 individuals from the genus *Rhododendron* as experimental subjects, and achieved good results in variety classification and prediction. These results demonstrate the applicability of our method and outline its potential areas of application. In addition, to test the dataset requirements and scope of our method, we performed down-sampling on this *Rhododendron* dataset and tested it on a publicly available rice dataset.

Materials and methods

Collection of materials and preparation of dataset

The method is based on short-read data from multiple individuals across different varieties. For our case study in *Rhododendron*, 40 individuals from 8 varieties (Table 1) were collected from the Kunming Institute of Botany, Chinese Academy of Sciences. For each individual sample, 2 g of tender leaves was collected and placed in a 10 ml cryotube, frozen rapidly using liquid nitrogen for 2 hours, and then stored in the freezer at -80 ℃. The samples were transported to Beijing Biomarker Technologies Co. Ltd. for DNA extraction, library construction, and sequencing. After genomic DNA extraction and passing quality control for the sample DNA, the genomic DNA underwent fragmentation via ultrasonication. The fragmented DNA was purified, end-repaired, polyadenylated at the 3' end, and sequencing adapters were ligated. Agarose gel electrophoresis was used for fragment size selection, followed by PCR enrichment to construct sequencing libraries. The libraries were further purified to remove adapter contamination. Then, Illumina paired-end sequencing was performed with a read length of 150 bp (PE150). In total, approximately 1286 million (M) clean reads were obtained from the 40 samples, representing 384.40 Gb of sequencing data. The average sequencing depth for each variety ranges from 13.40 to

Table 1 The sequencing information of individual of each variety

Variety Name	Accession	Number of Individuals	Sequencing depth
Rhododendron x *pulchurum* 'ZiHuDie'	W-1	10	14.84
Rhododendron 'HongBaoShi'	W-2	10	13.40
Rhododendron x hybridum Ker Gawl.	W-3	5	18.39
Rhododendron 'WanXia'	W-4	5	19.42
Rhododendron simsii	W-5	3	17.35
Rhododendron 'Fanxing'	W-6	3	19.27
Rhododendron 'Liuguang Yicai'	W-7	2	16.35
Rhododendron 'Yuzhuo'	W-8	2	16.38

19.42 (For most varieties with unknown genome sizes, we estimated a genome size of 600 MB based on published *Rhododendron* genomes, except for *Rhododendron simsii*, whose genome size has been reported as 528.6 MB [17]. And the sequencing depth is then determined by the ratio of the total base count to the genome size). The average sequencing depth per individual is 16.17. The average Q30 score (a quality assessment parameter for sequencing) reached 93.72%. These results indicate that the sequencing data quality is reliable and can be used for subsequent SSRs extraction and variety identification. We also conducted two rounds of down-sampling on this dataset using seqtk (the random state was set to 42) [18], with proportions of 50% and 75%.

To further assess the generalizability of our method, we evaluated it using a public rice dataset [19]. We filtered this dataset based on the following criteria: each individual was sequenced at > 10× depth, and each subpopulation included at least three individuals. As a result, we obtained 21 individuals from six subpopulations (Supplementary Table 1). In this dataset, subpopulations were defined below the variety level, with each individual's subpopulation determined by ecosystem type, geographic distribution and grain morphology. The raw dataset is available from the European Nucleotide Archive (ENA) (https://www.ebi.ac.uk/ena/browser/view/PRJEB36631).

Data processing and numerical characterization of SSRs

After obtaining the sequencing data for each individual, we use Minia (version 3.2.5) [20] to assemble it to the contigs, and use MISA (Microsatellite Identification Tool, version 2.1) [21] to detect all the SSRs within each individual. For convenience and to streamline the approach, only the simple perfect SSRs (as mentioned in the background) from the MISA results are retained for further evaluation.

Regarding an SSR, we use its presence or absence in an individual as a numerical characteristic. In more details, if an SSR appears in one assembled contigs of an individual, its characteristic value is 1, otherwise it is 0. Base on the above procedures, we could obtain an initial feature matrix after compiling all SSRs information from all individuals (Fig. 1a and Supplementary Fig. 1).

Selection of SSRs with intra-variety specificity and inter-variety polymorphism

Based on the initial SSRs feature matrix, we further filter SSRs exhibiting intra-variety specificity and inter-variety polymorphism through the following principles and steps (Fig. 1b). For a given SSR, among all the individuals within a variety, we expect its numerical characteristics to be the same, while showing variation beyond the variety. Consequently, we employ formulas (1) and (2) to determine the standard deviation of these numerical characteristics for each SSR, first within the same variety (S_{in}) and then across individuals beyond the variety (S_{out}). In formula (1), the n_{in} represents the number of all individuals within a variety, and $\overline{x}_{in}$ refers to the average of all numerical characteristics within the variety. In formula (2), n_{out} represents the number of all individuals out of the variety, and $\overline{x}_{out}$ refers to the average of all numerical characteristics outside the variety. In both formula (1) and formula (2), x_i represents the numerical characteristic value of the i_{th} individual. When the standard deviation of an SSR within a variety (S_{in}) is smaller than S_{out}, we consider that this SSR to exhibit intra-variety specificity and inter-variety polymorphism for this variety. However, we expect the final screened SSRs to exhibit this property across multiple varieties. Therefore, we designed formula (3).

In formula (3) (Supplementary Fig. 2), the parameter *threshold* refers to the proportion of variety satisfying $S_{in} < S_{out}$ among all varieties, while N_v refers to the number of varieties. $S_{in,j}$ and $S_{out,j}$ refer to the S_{in} and S_{out} of the j_{th} variety. The function I is a conditional judgement, and when the condition is met, the value of I is 1; when the condition is not met, the value of I is 0. For the SSRs that satisfied formula (3), we deemed such SSRs exhibit intra-variety specificity and inter-variety polymorphism among the varieties. Accordingly, after applying these formulas to all SSRs, we generate a new 0, 1 matrix of the retained SSRs to characterize all individuals from all varieties.

$$S_{in} = \sqrt{\frac{1}{n_{in}-1}\sum_{i=1}^{n_{in}}\left(x_i - \overline{x}_{in}\right)^2} \quad (1)$$

$$S_{out} = \sqrt{\frac{1}{n_{out}-1}\sum_{i=1}^{n_{out}}\left(x_i - \overline{x}_{out}\right)^2} \quad (2)$$

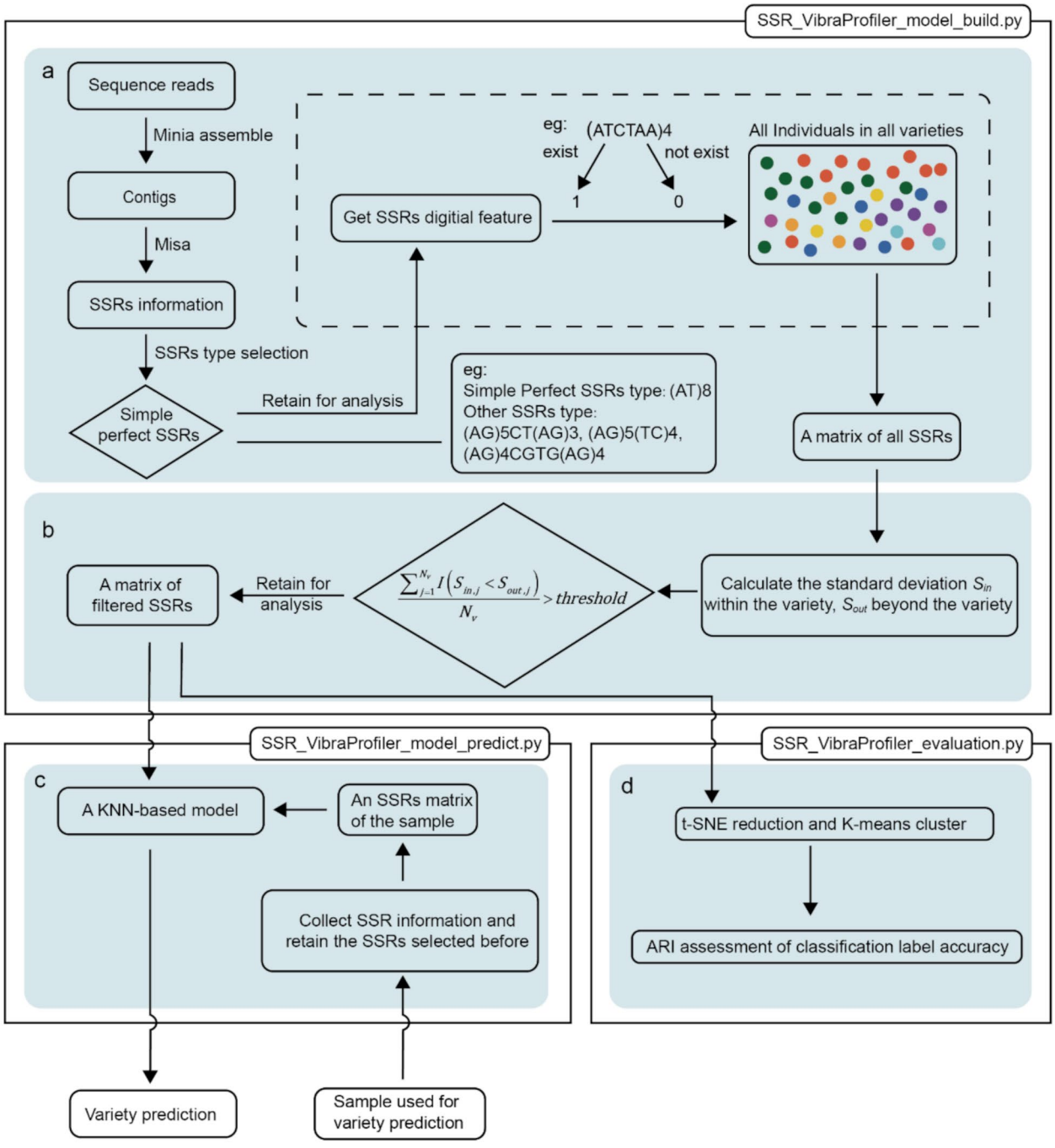

Fig. 1 The complete process of the method and SSR_VibraProfiler. **a**. SSRs information collection process. **b**. SSRs selection process. **c**. Variety prediction process of the unknown individual. **d**. Classification evaluation of SSRs. The outer black box outlines show the content integrated by each of the three scripts in the SSR_VibraProfiler

$$\frac{\sum_{j=1}^{N_v} I\left(S_{in,j} < S_{out,j}\right)}{N} > threshold \tag{3}$$

Construction of the KNN-based identification model

The 0, 1 matrix of retained SSRs is used to build a model based on KNN (K-Nearest Neighbors) classification model (Fig. 1c), where K refers to several nearest neighboring points. The model determines the type of the predicted point based on the types of K nearest neighbors to identify the variety of a query sample (an individual whose variety needs to be predicted). However, due to the possibility of unequal numbers of individuals within each variety in our case study (the situation also exists in

reality) when constructing the model, we do not specify a *K* value, and the model does not display the query's variety as well. On the contrary, it prints the distance rankings between the individuals within model and the query. The distances between points are calculated using Euclidean Distance, which is equivalent to the square root of the number of different SSRs among the individuals.

Classification evaluation of retained SSRs based on t-SNE dimensionality reduction and K-means clustering

After obtaining the SSR matrix, we aim to evaluate its classification capability by assessing whether it could effectively cluster individuals of the same variety into a single group. The process is as follows:

1. The retained SSR matrix is reduced to two dimensions for data point (the point represents individual) visualization using the t-SNE algorithm [22].
2. Subsequently, the data points are clustered into different groups using the K-means algorithm.
3. The accuracy of classification can be assessed by comparing the actual labels with the assigned cluster labels using the ARI (Adjusted Rand Index) parameter, which measures the similarity between the two label sets (Fig. 1d).

The t-SNE is a non-linear dimensionality reduction technique, whose main concept is to map similar data points in high-dimensional space to low-dimensional space while preserving the relative distances between data points as much as possible. K-means is a clustering algorithm used to divide data points into *K* clusters. ARI value ranges between −1 and 1, with values closer to 1 indicating a closer resemblance to the true partitioning results. ARI is commonly used to evaluate the performance of clustering algorithms.

Validation of model identification effects based on leave-one-out method

We utilize the leave-one-out (LOO) method to assess the predictive performance of the model on unknown individuals. Specifically, one individual is taken out as our query, and the remaining individuals are used to construct a KNN-based identification model, which is then used to predict the query. This process can be repeated multiple times based on the number of individuals. Due to the imbalance in the variety composition, we opted to display the ranked distances between the query and other individuals within the model, rather than displaying its specific variety.

Result

The architecture of SSR_VibraProfiler

We organized and packaged our method into a package named SSR_VibraProfiler. The package includes four Python scripts:

1) SSR_VibraProfiler_model_build.py. This script can automatically complete the following tasks (Fig. 1a, b):

1. Assembling sequencing data into segments using Minia;
2. Extracting SSRs information with MISA;
3. Summarizing and filtering SSRs data features;
4. Identifying SSRs with classification ability and generating a KNN-based model.

This process is relatively time-consuming; therefore, we have implemented parameters that enable the script to execute step-by-step. Additionally, we have preserved the output files at each stage, such as the segment file assembled by Minia and the SSRs information file generated by MISA.

2) SSR_VibraProfiler_evaluation.py. This script is utilized to evaluate the classification performance of the SSRs retained by SSR_VibraProfiler_model_build.py. It employs the t-SNE, K-means algorithms, and ARI parameter mentioned before. Due to the presence of a random state in both t-SNE and K-means, we set a range to iterate over them. It will output the top ARI results and generate a classification effect diagram corresponding to the highest ARI value (Fig. 1c).

3) SSR_VibraProfiler_model_predict.py. This script uses the KNN-based model to accept a new individual's SSRs information file (MISA output file) and ultimately returns the ranking of distances between this query individual and all other individuals in the model (Fig. 1d).

4) SSR_VibraProfiler_cross_validation.py. This script takes an index file as input and automatically executes a leave-one-out cross-validation process. In this process, it directly invokes the script "SSR_VibraProfiler_model_build.py" for model construction and "SSR_VibraProfiler_model_predict.py" for making predictions on the pick out individual, ultimately outputs a file containing the predict results for all individuals.

Figure 1 illustrates the content corresponding to each of the first three scripts (the fourth script invokes the model construction and prediction scripts). They have been packaged and submitted to the Conda (https://anaconda.org/oldcat931/ssr_vibraprofiler) and GitHub (https://github.com/Olcat35412/SSR_VibraProfiler) platforms, we also have provided a corresponding user manual that thoroughly introduces the installation and usage methods, encompassing the introductions of all available parameters for SSR_VibraProfiler. This manual can be obtained from GitHub platform or Supplementary

Table 2 Retained SSRs number and ARI evaluation results under different *thresholds*

threshold	0	0.125	0.25	0.375	0.5	0.625	0.75	0.875
Number of retained SSRs	8062	8054	8014	7785	7092	5729	3693	94
Highest ARI	0.858	0.844	0.85	1	0.85	0.85	0.944	0.84

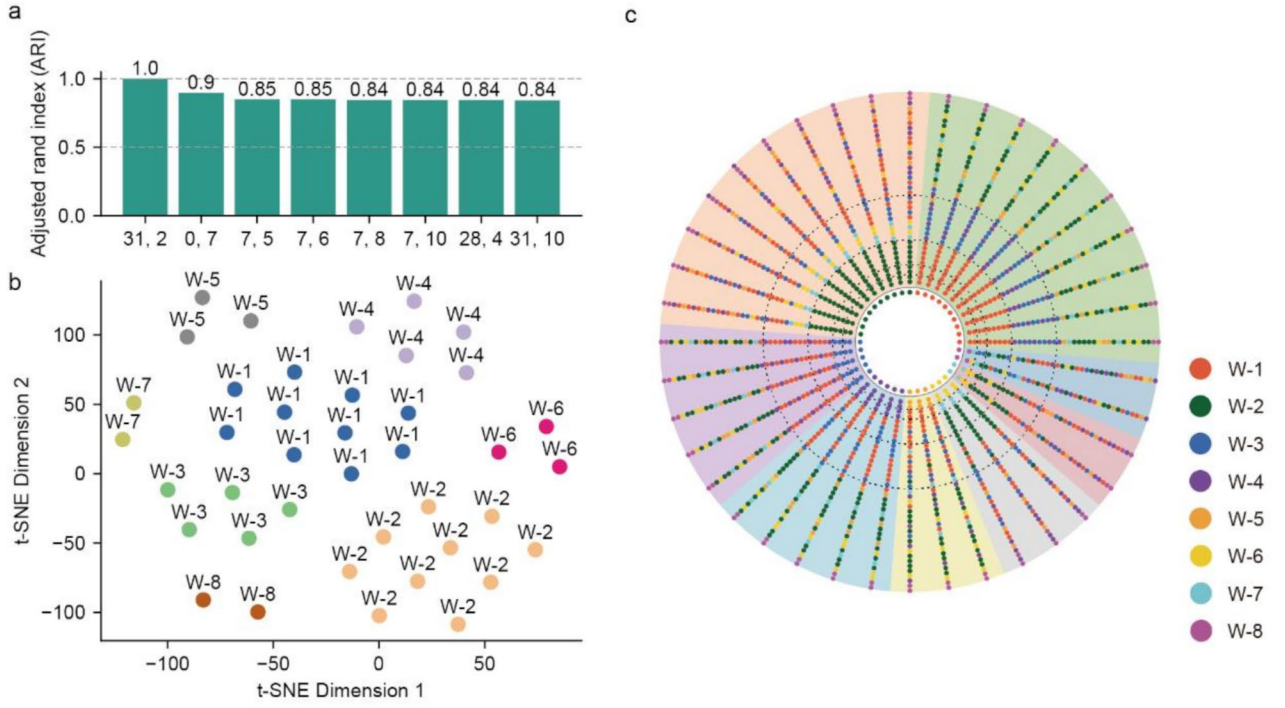

Fig. 2 The evaluation of the differentiation effect of screened SSRs and the results of leave-one-out cross-validation. **a**. ARI evaluation result. There are two random states in the process of dimensionality reduction and clustering; we take them from 0 to 50 and 0 to10, respectively. On the x-axis, "a, b" represent random state of t-SNE and random state of clustering, respectively. This figure displays the top 8 best results. **b**. The best clustering result achieved by the SSR matrix after dimensionality reduction using t-SNE and clustering using k-means (corresponding to the highest ARI value of 1 in figure a). The same color indicates that they are clustered as one variety, and the labels near the points represent the true labels. **c**. Result of LOO cross-validation. The innermost point on each axis represents the individual used for validation. The distance between query and sample within the model is sorted according to Euclidean Distance. The four black dashed lines correspond to the 2 closest, 4 closest, 9 closest, and 18 closest points to the query, respectively

materials of this article. Furthermore, to further enhance users' convenience and lower the barrier to using the software, we have also provided a docker image, which can be available at https://hub.docker.com/repository/docker/oldcat931/ssr_vibraprofiler freely.

SSR_VibraProfiler can accurately classify and predict varieties within the *Rhododendron*

A total of 8796 SSRs were identified using the SSR_VibraProfiler_model_build.py script from the SSR_VibraProfiler package. The presence or absence of these SSRs form an initial 0, 1 matrix (Supplementary Table 2, Sheet1). Given that we have 8 varieties, we considered all possible "*thresholds*" from 0 to 0.875 and used the SSR_VibraProfiler_evaluation.py script to evaluate the classification performance of SSRs under these different *threshold*. The corresponding numbers of retained SSRs and the highest ARI results for each "*threshold*" have been summarized in Table 2 (with the complete information in Supplementary Table 2, Sheet 2 to 9). When the "*threshold*" set to 0.375, the selected SSRs show intra-variety specificity and inter-variety polymorphism in at least 4 varieties ($8 \times 0.375 = 3$). At this "*threshold*", the highest ARI reaches 1 (Table 2; Fig. 2a). The corresponding dimensionality reduction and clustering results completely match the true label classification. This process was executed on two platforms, with the runtime displayed in Table 3. This indicates that these SSRs achieve 100% accuracy in classifying the varieties of individuals (Fig. 2b).

Based on this "*threshold*", we used the script SSR_VibraProfiler_cross_validation.py to perform the leave-one-out cross-validation process, and then aggregated the corresponding results (Fig. 2c, Supplementary Table 3). Notably, in the final result of leave-one-out, the nearest multiple individuals to the queries are always those of the same variety. For example, variety W-1 has ten individuals. When we remove any given individual from this variety and build a model to predict its variety, we find

Table 3 Runtime of SSR_VibraProfiler on the *Rhoddendron* dataset

Platform	Intel(*R*) Core (TM) i9-14900 K 32 Cores 128GB Memory (deployed on the workstation)	Intel(*R*) Xeon(*R*) Gold 5318Y CPU 96 cores 2 TB Memory (deployed on the server)
Model Build Command	SSR_VibraProfiler_model_build.py -i index.txt -s 1 -e 3 -pp 0.375 -index index.txt -o output_dir/ -miniap 4 -miniac 6 -misap 24	SSR_VibraProfiler_model_build.py -i index.txt -s 1 -e 3 -pp 0.375 -index index.txt -o output_dir/ -miniap 8 -miniac 5 -misap 40
Cores Used	24	40
Minia Assembly time	9h8m55s	7h27m11s
MISA Running time	15m27s	32m45s
Model Building time	8s	30s
Cross Validation time	5m26s	24m2s
Dimensionality Evaluation time	27s	6m1s

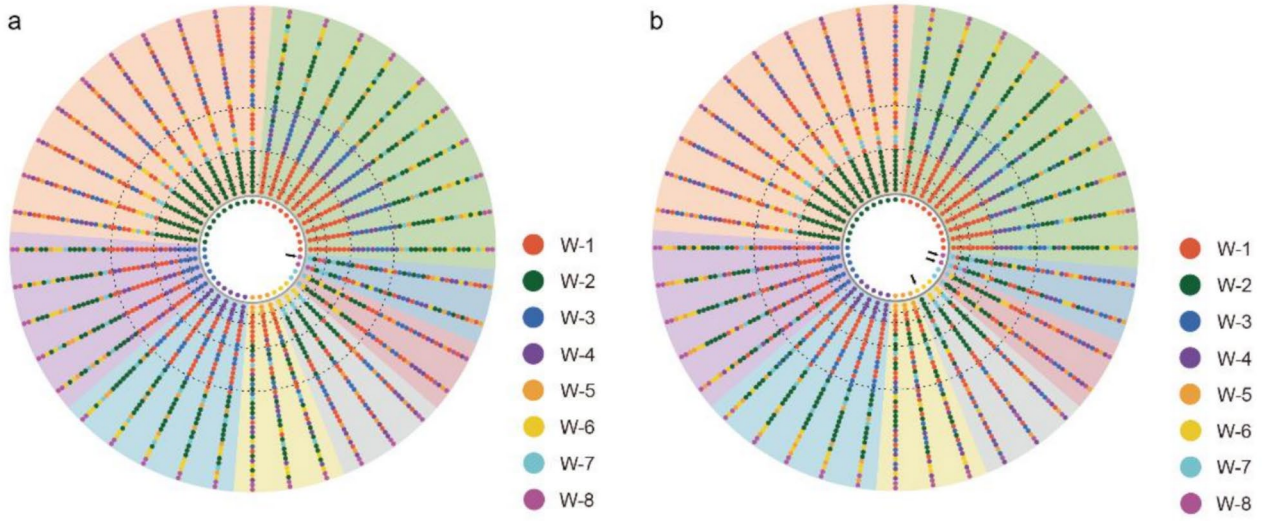

Fig. 3 Cross-validation results after down-sampled the *Rhododendron* dataset. **a**. Cross-validation result when down-sampling rate is 75%. **b**. Cross-validation result when down-sampling rate is 50%. The black arrows point to the individuals with incorrect cross-validation results

that the nine individuals closest to it in the model prediction all come from the same variety. This consistent pattern is also observed in other varieties. Despite there is existing risk of overfitting, our analysis demonstrates that our model possesses the capability to accurately predict the variety of individuals. The workflow of this case study has been provided as an operational example in the software manual.

Reducing sequencing depth can affect the variety identification capability of SSR_VibraProfiler

The performance of our method may be influenced by the sequencing depth of the samples. Therefore, we performed two rounds of down-sampling on the sequencing data, with down-sampling rates of 75% and 50%, corresponding approximately to 12x and 8x sequencing depths, respectively. After constructing models with the corresponding down-sampled data, we re-evaluated the classification performance of the models using dimensionality reduction-based clustering and cross-validation. The "*threshold*" was set to 0.375 in this process.

The ARI results based on t-SNE dimensionality reduction and k-means cluster still reach 1 when the down-sampling rates are 75% and 50% (Supplementary Fig. 3). However, the cross-validation results show varying degrees of decline. In detail, when the down-sampling rate is 75%, one individual of the W-8 variety is misclassified. Consequently, the model's prediction accuracy is 97.5% (Fig. 3a). When the down-sampling rate is 50%, both two individuals of the W-8 and an individual of W-6 are misclassified, resulting in a model prediction accuracy of 92.5% (Fig. 3b). This indicates that sequencing depth affects the performance of variety identification. However, accuracy only dropped slightly to 97.5% and then 92.5%. Since we consider the presence or absence of SSRs rather than their frequencies, down-sampling has limited impact on this feature, which may explain the slight decrease in accuracy.

SSR_VibraProfiler is not suitable for subpopulation identification in rice

To evaluate the performance of our method on other species and simulate cases with high intra-variety polymorphism, we used the previously published rice dataset [19]. The "*threshold*" of software was set to 0.375 in this process. After obtaining the SSRs matrix, performing t-SNE dimensionality reduction and K-means clustering, the best ARI reach 0.87 (corresponding to one misclassified individual, Supplementary Fig. 4a). However, during cross-validation, the predicted labels for 5 individuals were incorrect, resulting in an accuracy of only 76% (Supplementary Fig. 4b, Supplementary Table 4). This may be due to the long breeding history of rice, which has led to high internal SSR polymorphism that our method fails to capture effectively. These findings indicate potential limitations of our method in handling varieties characterized by high internal polymorphism.

Discussion

Advancement and convenience of the SSR_VibraProfiler

Some in silico-based tools that do not utilize resequencing datasets typically identify SSRs directly from sequences, provide flanking sequences, and filter experimentally verifiable markers [23]. In recent years, several new in silico SSR markers development methods have been proposed, relying either on a single reference genome and resequencing dataset [12] or on multiple assembled genomes [13, 24, 25]. The objective of these studies is to identify SSR markers with well-defined loci and potential polymorphism, which, upon experimental validation, can be used for individual classification. While these studies hold considerable practical significance, they are not applicable in the absence of a reference genome. Another widely used markers based on NGS (Next-Generation Sequencing) data are SNPs (Single Nucleotide Polymorphisms). Similar to SSRs, SNP-based approaches also depend on the availability of a reference genome. Besides, the process of obtaining SNPs is time-consuming especially for large genomes, which poses challenges for their practical application.

In contrast, our approach can run effectively in the absence of a reference genome. It identifies SSRs that exhibit intra-variety specificity and inter-variety polymorphism, forming a 0, 1 matrix. Subsequently, we utilize the matrix to construct a model for identifying the variety of unknown individuals. By disregarding specific loci, flanking sequences, and occurrence counts, this method sacrifices a portion of SSR information in exchange for faster computational efficiency and advantage of not relying on a reference genome. Our case study on *Rhododendron* reveals that after utilizing the t-SNE reduction technique to preprocess the SSRs matrix, the K-means clustering algorithm successfully classified the varieties (Fig. 2b). This result validates the effectiveness of the selected SSRs as a basis for variety classification. Furthermore, our LOO method result demonstrates that individuals from the same variety consistently cluster closely around the queries, forming a stable nearest-neighbor relationship. These results demonstrate the potential of our method to accurately distinguish varieties. In addition, applying our method to analyze the *Rhododendron* dataset takes only a few hours (Table 3), which further demonstrates its convenience. These advantages make our method applicable to a wider range of organisms, even in cases where genomic resources are limited.

Dataset requirements and application scope of SSR_VibraProfiler

SSR_VibraProfiler imposes requirements on dataset composition and sequencing depth. Firstly, our method necessitates a minimum of three individuals per variety for model construction. This is because the model construction involves calculating the sample standard deviation within variety (representing the polymorphism of an SSR within a variety), which cannot be computed for a variety with only one individual. To ensure robust model construction and accurate classification results, we recommend including at least three individuals per variety. Additionally, the number of individuals per variety should be approximately equal during the initial stage of model construction. Secondly, although our method exhibits a compatibility with low sequencing depth, ensuring an appropriate sequencing depth remains essential. In *Rhododendron* case study, when we performed 75% and 50% down-sampling, the model accuracy obtained through cross-validation began to decrease (Fig. 3). However, accuracy dropped slightly because our approach encodes the presence or absence of each SSR as a numerical character, which is minimally affected by down-sampling, reducing the risk of being entirely missed at lower sequencing depths. Even if some SSRs' numerical characters change after down-sampling, our method leverages the collective effect of all selected SSRs for variety classification, minimizing the impact of losing a few SSRs on overall prediction accuracy. Nonetheless, a proper sequencing depth remains necessary. We recommend an approximate depth of 10× to enhance the model's robustness.

SSR_VibraProfiler may not apply to varieties with high intra-variety polymorphism. Due to the relatively high mutation rate of SSRs (ranging from 10^{-2} to 10^{-6} per locus per generation [26]), some existing varieties, particularly those with a long breeding history or extensive geographical distribution, may have accumulated substantial SSR variations. As a result, our method may not be able to extract reliable classification information from these varieties. The relatively poor performance

of SSR_VibraProfiler on the rice subpopulations dataset demonstrates this risk. Therefore, we exercise caution for applying our method to varieties with high internal polymorphism.

Conclusion

In this study, we introduced a method for distinguishing and predicting individual varieties without requiring a reference genome. Our approach directly extracts SSR numerical characteristics from next-generation sequencing DNA data, focusing on SSRs with the property of intra-variety specificity and inter-variety polymorphism. We have packaged this method into a tool called SSR_VibraProfiler, enabling the construction of an SSR-based DNA fingerprint database for variety identification. When applied to *Rhododendron* varieties, our method yielded excellent performance results. Additionally, our work will contribute to the development, identification, and protection of new varieties.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13007-025-01380-x.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Supplementary Material 4

Supplementary Material 5

Supplementary Material 6

Supplementary Material 7

Supplementary Material 8

Supplementary Material 9

Supplementary Material 10

Supplementary Material 11

Supplementary Material 12

Supplementary Material 13

Acknowledgements

We sincerely appreciate the insightful discussions with the members of the Zhang lab. We thank Hangbo Zhu for setting up the computing platform during the revision process. Additionally, we thank Xin Geng for assisting us with grammar and spell checking.

Author contributions

CJZ, CD and JSW designed and coordinated this work. ZZW, XPW, QNY, GPY, YMW provided and collected the samples. CHJ conducted the mainly computational analysis. CHJ, CD, CJZ, ZZW wrote the draft. CHJ packaged SSR_VibraProfiler. CHJ and CYS prepared the figures. SYM tested SSR_VibraProfiler.

Funding

This work was supported by the Zhejiang Provincial Cooperative Forestry Science and Technology Project (2025SY07); the Research and Development Foundation of Zhejiang A&F University to C.D (2023LFR103) and the Research and Development Foundation of Zhejiang A&F University to CZ (2023LFR022). The research was also supported by the Overseas Expertise Introduction Project for Discipline Innovation (111 Project D18008).

Data availability

The sequencing data supporting this study are publicly available in the NCBI BioProject database under accession number PRJNA1209532.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]National Key Laboratory for Development and Utilization of Forest Food Resources, Zhejiang A & F University, Hangzhou, Zhejiang 311300, China
[2]Germplasm Bank of Wild Species & Yunnan Key Laboratory of Crop Wild Relatives Omics, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, Yunnan 650201, China
[3]Yunnan University, Kunming, Yunnan 650091, China
[4]Haiyan Senzhi Biotechnology Co., Ltd, Jiaxing, Zhejiang 314300, China
[5]Haiyan Engineering & Technology Center, Zhejiang Institute of Advanced Technology, Jiaxing, Zhejiang 314300, China

Received: 20 November 2024 / Accepted: 25 April 2025
Published online: 16 May 2025

References

1. Ellegren H. Microsatellites: simple sequences with complex evolution. Nat Rev Genet. 2004. https://doi.org/10.1038/nrg1348.
2. Mrázek J, Guo X, Shah A. Simple sequence repeats in prokaryotic genomes. Proc Natl Acad Sci U S A. 2007. https://doi.org/10.1073/pnas.0702412104.
3. Bagshaw AT. Functional mechanisms of microsatellite DNA in eukaryotic genomes. Genome Biol Evol. 2017. https://doi.org/10.1093/gbe/evx164.
4. Cheng T, Lin P, Zhou D, Wang H, Shi S, Shen J, Meng J, Ye X, Zheng K, Hu Xing, Zhuang Y. Development and characterization of novel EST-SSR markers for *Gentiana straminea* Maxim., a traditional Tibetan herb in China and cross-amplification in related species. Plant Genet Resour. 2024. https://doi.org/10.1017/S1479262124000224.
5. Singh N, Choudhury DR, Tiwari G, Singh AK, Kumar S, Srinivasan K, Tyagi RK, Sharma AD, Singh NK, Singh R. Genetic diversity trend in Indian rice varieties: an analysis using SSR markers. BMC Genet. 2016. https://doi.org/10.1186/s12863-016-0437-7.
6. Zhang J, Yang J, Lv Y, Zhang X, Xia C, Zhao H, Wen C. Genetic diversity analysis and variety identification using SSR and SNP markers in melon. BMC Plant Biol. 2023. https://doi.org/10.1186/s12870-023-04056-7.
7. Li S, Liu S, Pei S, Ning M, Tang S. Genetic diversity and population structure of *Camellia Huana* (Theaceae), a limestone species with narrow geographic range, based on Chloroplast DNA sequence and microsatellite markers. Plant Divers. 2020. https://doi.org/10.1016/j.pld.2020.06.003.
8. Yu Y, Wang H, Yu Z, Schinnerl J, Tang R, Geng Y, Chen G. Genetic diversity and structure of the endemic and endangered species *Aristolochia Delavayi* growing along the Jinsha river. Plant Divers. 2021. https://doi.org/10.1016/j.pld.2020.12.007.
9. Al-Samarai FR, Al-Kazaz AA. Molecular markers: an introduction and applications. Eur J Mol Biotechnol. 2015. https://doi.org/10.13187/ejmb.2015.9.118.
10. Hayden MJ, Nguyen TM, Waterman A, Chalmers KJ. Multiplex-ready PCR: a new method for multiplexed SSR and SNP genotyping. BMC Genomics. 2008. https://doi.org/10.1186/1471-2164-9-80.
11. SENAN S, KIZHAKAYIL D, SHEEJA TE. Methods for development of microsatellite markers: an overview. Not Sci Biol. 2014. https://doi.org/10.15835/nsb619199.
12. Perry A, Eddelbuettel D, Rosenthal G, Blackmon H, Polly. An R package for genotyping microsatellites and detecting highly polymorphic DNA markers

from short-read data. Mol Ecol Resour. 2024. https://doi.org/10.1111/1755-0998.13933.
13. Turudić A, Liber Z, Grdiša M, Jakše J, Varga F, Poljak I, et al. Dig-up primers: A pipeline for identification of polymorphic microsatellites loci within assemblies of related species. Int J Mol Sci. 2024. https://doi.org/10.3390/ijms25063169.
14. Zhang X, Liu X, Liu D, Cao Y, Li Z, Ma Y, Ma H. Genetic diversity and structure of *Rhododendron meddianum*, a plant species with extremely small populations. Plant Divers. 2021. https://doi.org/10.1016/j.pld.2021.05.005.
15. Rawat S, Jugran AK, Sharma H. Recent advancements in the physiological, genetic, and genomic research on *Rhododendrons* for trait improvement. 3 Biotech. 2024. https://doi.org/10.1007/s13205-024-04006-6.
16. Liang J, Chen Y, Tang X, Lu Y, Yu J, Wang Z, Zhang Z, Ji H, Li Y, Wu P, Liu Y, Wang L, Huang C, He B, Lin W, Guo L. Comprehensive Evaluation of Appreciation of *Rhododendron* Based on Analytic Hierarchy Process. Plants (Basel). 2024. https://doi.org/10.3390/plants13040558
17. Yang FS, Nie S, Liu H, Shi TL, Tian XC, Zhou SS, et al. Chromosome-level genome assembly of a parent species of widely cultivated azaleas. Nat Commun. 2020. https://doi.org/10.1038/s41467-020-18771-4.
18. Toolkit for processing sequences in FASTA/Q formats. https://github.com/lh3/seqtk. Accessed 15 Feb 2025.
19. Higgins J, Santos B, Khanh TD, Trung KH, Duong TD, Doai NTP, et al. Resequencing of 672 native rice accessions to explore genetic diversity and trait associations in Vietnam. Rice. 2021. https://doi.org/10.1186/s12284-021-00481-0.
20. Chikhi R, Rizk G. Space-efficient and exact de Bruijn graph representation based on a bloom filter. Algorithms Mol Biol. 2013. https://doi.org/10.1186/1748-7188-8-22.
21. Thiel T, Michalek W, Varshney R, Graner A. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L). Theor Appl Genet. 2003. https://doi.org/10.1007/s00122-002-1031-0.
22. Van der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9(11).
23. Mokhtar MM, Alsamman AM, El Allali A. MegaSSR: a web server for large scale microsatellite identification, classification, and marker development. Front Plant Sci. 2023. https://doi.org/10.3389/fpls.2023.1219055.
24. Xia EH, Yao QY, Zhang HB, Jiang JJ, Zhang LP, Gao LZ. CandiSSR: an efficient pipeline used for identifying candidate polymorphic SSRs based on multiple assembled sequences. Front Plant Sci. 2016. https://doi.org/10.3389/fpls.2015.01171.
25. Gou X, Shi H, Yu S, Wang Z, Li C, Liu S, et al. SSRMMD: a rapid and accurate algorithm for mining SSR feature loci and candidate polymorphic SSRs based on assembled sequences. Front Genet. 2020. https://doi.org/10.3389/fgene.2020.00706.
26. Li YC, Korol AB, Fahima T, Beiles A, Nevo E. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. Mol Ecol. 2002. https://doi.org/10.1046/j.1365-294X.2002.01643.x.

Publisher's note