ARTICLE (*Pre-print*)

Species Delimitation Using Machine Learning Methods

Philip Lin* and Bing Li

University of Wisconsin - Madison, Madison, US

*Corresponding author: pjlin2001@gmail.com

Abstract

This study examines species delimitation within the Bromeliaceae family using machine learning algorithms to analyze single nucleotide polymorphism (SNP) variants. Traditional species identification methods often rely on morphological traits, which may not capture the full extent of genetic diversity. This project employs two machine learning approaches, XGBoost (Chen 2016) and Multi-Layer Perceptron (MLP, Rosenblatt 1962), to explore the genetic foundations of species differentiation. Our analysis includes eight species across five genera, sequenced by low-coverage genome skimming. The results demonstrate the potential of machine learning to enhance species delimitation with high accuracy and confidence, offering significant contributions to taxonomy and conservation biology.

Introduction

This study addresses the significant challenge of species delimitation within plant studies, meaning identifying species-level genetic and morphological diversity among organisms, which plays a central issue that touches on multiple scientific disciplines, including systematics, conservation biology, and evolutionary biology. Recent approaches using genomic data to address species delimitation focuses on developing statistical approaches and incorporating more large datasets. With the recent popularity in machine learning and its application in clustering, it is surprising that machine learning methods have been rarely applied to study taxonomic delimitation, especially in plants. This project aims to leverage the power of single nucleotide polymorphism (SNP) variants to investigate the genetic foundations underlying species differentiation across five genera within the Bromeliaceae family, sequenced by low-coverage genome skimming by Novogene Inc. Bromeliaceae is not only important due to its commercial crops, pineapple, but also as one of the most diverse flowering plant families, contributing greatly to the rich biodiversity of the Neotropics, an understudied global biodiversity hotspot. By focusing on the genetic level, we hope to offer a more nuanced and accurate method for species delimitation, contributing valuable insights to the fields of taxonomy and conservation.

Motivation

The primary motivation behind this study is to evaluate two machine learning methods, XGBoost and multi-layer perceptron, for identifying species boundary using plant genomic data using single nucleotide polymorphism (SNP) variants. This research is particularly relevant for taxonomists, conservation biologists, and evolutionary biologists, who require reliable methods for species identification to conduct biodiversity assessments, conservation planning, and evolutionary studies.

Background

Species is the fundamental unit in the study of biodiversity and evolutionary trends, holding remarkable importance in conservation biology and evolutionary studies. Distinguishing and classifying species, however, presents significant challenges. Historically, the identification of species largely relies on morphological traits (eg. Pedersen 2010), which may not fully reflect the extent of genetic variation within and between species. Recent breakthroughs in statistical methods of genomic studies and next generation sequencing technologies have facilitated a range of phylogenetic and

taxonomic delimitation studies using different types of molecular genomic data, such as whole-genome resequencing (e.g., Liang et al., 2019; Ma et al., 2020), single or low-copy nuclear genes or ultraconserved loci (e.g., Johnson et al., 2019; Yardeni et al., 2022, Heidin et al. 2019), extensive transcriptome sequencing (e.g., Ma et al., 2020), and large-scale SNP data (eg. Adam et al. 2014), etc. Moreover, species delimitation research is increasingly incorporating sophisticated statistical approaches, such as utilizing multivariate algorithms to interpret morphological data (Ezard et al. 2010) and advancing molecular diagnostics with multi-locus sequencing data (Yang and Rannala, 2010).

The major difficulty of addressing species delimitation is to detect the genetic differentiation in population-level and species-level diversity. Speciation is a gradual evolutionary process with species often representing points along a continuum. This results in blurry species boundaries, particularly considering highly differentiated population structure associated with limited gene flow and geographic barriers. For instance, a plant species with a disjointed distribution that experiences limited gene flow between populations tends to exhibit different population structure in genetics. This can lead to confusion, mistaking population structure for species boundaries. The current most popular model for delimiting species boundary, multispecies coalescent model (MSCM) on Bayesian estimation (Rannala and Yang 2003; Yang and Rannala 2010), accommodates both pre-speciation mutations and incomplete lineage sorting. This approach helps to resolve discrepancies between actual species relationships and those inferred from multi-locus gene trees (Rannala and Yang 2003; Yang and Rannala 2010; Adams et al. 2014). However, MSCM often assumes speciation to be an instantaneous point event rather than a process, potentially inflating species counts due to the genetic differentiation arising from population structuring (Sukumaran and Knowles, 2017; Yang et al., 2019; Smith and Carstens, 2018).

Machine learning methods are increasingly applied to identify patterns in evolution and biology. Historically, such methods in plant research have focused on detecting phenotypic traits through image processing in model species, like Arabidopsis (Ma et al. 2014; Singh et al. 2016). However, the use of machine learning to interpret plant genomic data, particularly for species delimitation in non-horticultural species, is rare. Techniques like Principal Component Analysis (PCA) have proven successful in using morphological and genomic data to clarify species boundaries and elucidate population structures (Pedersen 2010; Yang et al. 2019; Cheng et al. 2021). Support vector machines (SVMs) have been developed to optimize likelihood scores for species-population assignments (Pei et al. 2018). Smith and Carstens (2018) have employed random forests (RF) to assess various speciation models incorporating demographic processes. Additionally, unsupervised machine learning techniques, which do not require predefined species labels, have been adopted to explore genetic structures (Derkarabetian et al. 2019).

Species delimitation using unsupervised machine learning techniques is fundamentally a clustering challenge. It involves identifying and grouping patterns present in a matrix dataset, where the goal is to cluster these patterns based on their similarity across samples. This clustering is not merely about detecting similarities but also about interpreting the biological significance of the resulting groups. Previous studies utilizing machine learning for species delimitation have generally focused on datasets comprising approximately 5 populations or species, each with a moderate sample size (approximately 80) and a moderate number of SNPs (1,000 to 10,000) (Smith and Carsten 2018; Derkarabetian et al. 2019).

This study presents the first effort to address the issue of species delimitation using two machine learning approach, XGBBoost (Chen and Guestrin 2016) and multi-layer perceptron (MLP, Haykin 1994), to train a SNP dataset with eight different plant species within the pineapple family (Bromeliaceae). To our best knowledge, while XGBoost has been utilized for deciphering tree-based diversification driver in the Cactus family (Cactaceae) (Thompson et al. 2023), its application to SNP datasets

for species delimitation remains unexplored. Similarly, our MLP approach echoes neural network strategies like convolutional neural networks, which have been used to discern species or population structure in taxonomically complex Cactaceae. However, these applications have largely focused on cryptic diversity, showing potential but also highlighting the need for more extensive research to validate their accuracy (Perez et al. 2022). This investigation marks the initial phase in tackling species delimitation using XGBBoost and MLP, aiming to assess the training accuracy of our methods and to explore the limitations and potential enhancements for future studies.

Data

Eight species across five genera within the Bromeliaceae family were sequenced by low-coverage genome skimming by Novogene Inc. We downloaded the reference genome of Ananas comosus (pineapple, NC033621, Ming et al. 2015) from GenBank. Raw sequencing reads were trimmed using fastp (Chen et al. 2018), which eliminated adapters and excluded reads with a quality score below 20 or length under 100 base pairs. To ensure our low-coverage genome skimming sufficiently captured the necessary genetic information, we implemented three distinct validation methods. BUSCO (Manni et al. 2021) was used to evaluate the capture of universal single-copy orthologs, while BLAST+ (Camacho et al. 2009) and MiniMap2 independently verified the capture of single-copy orthologs (SCOs) based on criteria from Yardeni et al. (2022), adjusting for mismatch thresholds. These checks confirmed that the genome skimming data was of high quality, with adequate and diverse genetic representation across each genus. For SNP variant calling, we followed a standard protocol that involved indexing the reference genome with BWA (Li and Durbin 2009), mapping the trimmed reads for each sample using Bowtie2 (Langmead and Salzberg 2012), and applying BCFtools (Li et al. 2011) for the variant calling process.

The core of our dataset is in Variant Call Format (VCF), a specialized format designed to store gene sequence variations with high precision. This study specifically zooms in on the SNP variants, represented in the GT:PL format. Here, GT denotes the genotype, revealing the genetic constitution at a specific loci, while PL indicates the normalized Phred-scaled likelihoods, offering a quantitative measure of the genotype's accuracy by estimating the probability of each genotype possibility. This dual-layered data structure allows for a nuanced exploration of genetic variations, serving as a critical tool for understanding the genetic diversity and evolutionary relationships among species. Focusing on whole-genome sequencing data, our study encompasses 8 species, delving into the genetic intricacies of single-copy nuclear genes distributed across 53 different species, enabling a robust analysis of genetic diversity.

Prior to the analytical phase, the datasets undergo a rigorous preprocessing stage. This essential step ensures that all data adheres to strict quality and consistency standards, paving the way for accurate and reliable species delimitation analysis. The final processed dataset is encapsulated in a dataframe with dimensions of shape (210859, 33), signifying the integration of 210859 SNP records across 33 distinct features. This structured approach not only facilitates efficient data manipulation and analysis but also enhances the clarity and interpretability of the results.

Methods

Our approach to addressing the problem of species delimitation through genetic data involves treating it as a classification problem using machine learning algorithms. Once the data is transformed into a matrix form, we will employ techniques such as XGBoost and multi-layer perceptrons (MLPs) to potentially enhance the predictive accuracy beyond that achievable with traditional clustering models. Our baseline for comparison will be a random forest model, configured with 800 estimators and trained on the same dataset.

The effectiveness of our models will be rigorously evaluated using the F1 score. F1 score will be used to assess the balance between precision (the proportion of positive identifications that are correct)

and recall (the proportion of actual positives that were correctly identified) for each species. This is particularly important in the context of imbalanced classes, which is common in species data where some species may be more frequent than others.

$$\text{Precision} = \frac{TP}{TP + FP} \qquad \text{Recall} = \frac{TP}{TP + FN} \qquad \text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where:

- TP = True Positives: the number of correct predictions that an instance is positive,
- TN = True Negatives: the number of correct predictions that an instance is negative,
- FP = False Positives: the number of incorrect predictions that an instance is positive,
- FN = False Negatives: the number of incorrect predictions that an instance is negative.

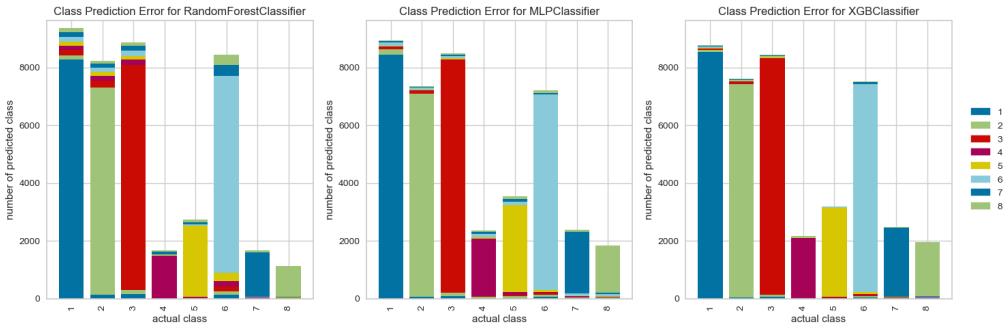
Model validation will employ the hold-out method, reserving 20% of the data for testing the models' ability to accurately delimit species not encountered during the training phase. The assignment of labels for training will be meticulously conducted based on manually verified genetic data. To ensure accuracy and consistency, we use the GT:PL format (Genotype: Probability List) which integrates expert knowledge into the genetic data interpretation process. Each species label is derived from annotations made by seasoned biologists or from previously validated datasets known for their reliability.

These expert annotations are critical as they serve as the gold standard for our analyses. The robustness of this gold standard is essential because it directly impacts the reliability of our model's performance evaluation. By grounding our label assignment in well-established genetic evidence and expert validation, we ensure that our model validation reflects the models' effectiveness in a realistic and operational context where precise species identification is paramount. This comprehensive approach guarantees that the machine learning models are not only theoretically sound but also practically effective in real-world biodiversity research settings.

Results

The application of machine learning techniques to species delimitation within the Bromeliaceae family was undertaken using Random Forest, Multi-Layer Perceptron (MLP), and XGBoost algorithms. These methods were chosen based on their theoretical capability to manage high-dimensional data, like the SNP variants used in this study.

Class Prediction Error



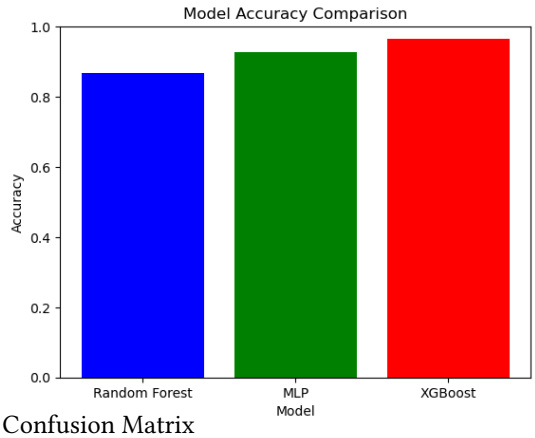
The class prediction error was visualized through histograms for each classifier (Random Forest,

MLP, XGBoost). The histograms display the count of misclassified instances across the eight actual plant species. It appears that the Random Forest exhibits the highest level of misclassification among the three models. This inference is drawn from the variety of colors across the bars corresponding to each actual class, indicating that predictions were more frequently assigned to incorrect classes. The presence of multiple colors within the bars of the Random Forest plot suggests a less consistent prediction for several classes compared to the other models.

The MLP shows a moderate degree of misclassification, with fewer instances of incorrect class predictions than the Random Forest, as indicated by the lesser variety of colors within the bars. This suggests that while the MLP may not be as precise as the XGBoost model, it still maintains a reasonable level of accuracy across most classes.

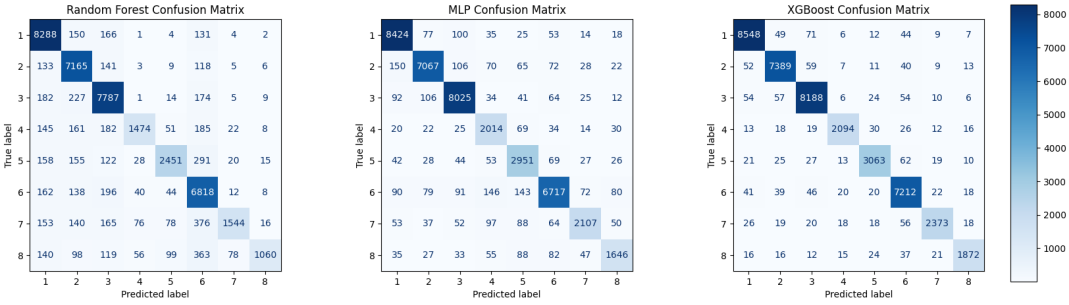
The XGBoost model, meanwhile, seems to demonstrate the best performance in terms of class prediction accuracy. The bars in the XGBoost plot predominantly match the color that represents the correct class, with minimal presence of other colors. This uniformity suggests that the XGBoost model has the most consistent and accurate class predictions, making it the superior model among the three in handling this particular dataset.

Model Accuracy



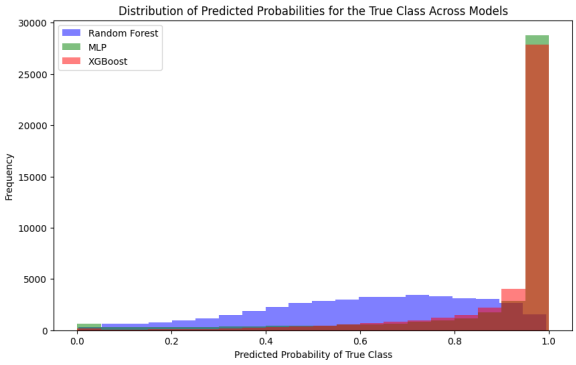
When comparing model accuracies, the XGBoost model outperforms both the Random Forest and MLP models, with a near-perfect accuracy score. This high level of accuracy indicates that the gradient boosting technique employed by XGBoost is highly effective for this dataset, possibly due to its capacity for handling the complexities and interactions within the SNP data.

Confusion Matrix



The confusion matrices for each model provide a detailed breakdown of the predictions. In all three models, some pairs of species (for example, species 1 and 2) are more commonly confused with one another than with other species, which may be due to genetic similarities between these species. The XGBoost model demonstrates a more distinct diagonal in its confusion matrix, which represents a higher rate of correct predictions, while the Random Forest and MLP show more spread across non-diagonal cells, indicating more frequent misclassifications.

Predicted Probability Distribution



The distribution of predicted probabilities for the true class illustrates the confidence of each model in its predictions. The XGBoost model again shows a strong performance, with a significant majority of predictions having high confidence (probability close to 1). This suggests that XGBoost not only makes accurate predictions but also makes these predictions with a high degree of certainty. Meanwhile, Random Forest and MLP exhibit broader distributions of probabilities, indicating less consistent confidence across their predictions.

Discussion

The results of this study provide compelling evidence for the utility of machine learning algorithms in species delimitation using SNP data. The high accuracy and confidence of the XGBoost model demonstrates its potential as a powerful tool for taxonomists and conservation biologists, enabling precise identification of species boundaries that may not be discernible through traditional methods.

However, the variation in model performance also highlights the complexity of species delimitation. It suggests that certain species pairs with close genetic relationships or similar evolutionary histories may require more refined data or advanced modeling techniques to differentiate accurately.

The variability in error rates across species implies that while machine learning can significantly enhance species delimitation, there remains a need for domain expertise to interpret the biological significance of the predictions. Inconsistencies between models for certain species suggest that integrating multiple modeling approaches or refining feature selection could improve performance, especially for species with high misclassification rates.

The superior performance of XGBoost could be attributed to its ensemble learning approach that effectively captures complex patterns in the data by combining multiple weak learners into a strong one. The presence of highly confident predictions aligns with the theoretical efficiency of the algorithm in handling various types of data distributions and interactions among features.

The Random Forest model, while generally robust and less prone to overfitting, may have been challenged by the high dimensionality and complexity of SNP data. Its misclassifications across species suggest a need for parameter tuning or potentially incorporating feature selection methods to reduce noise and focus on more informative genetic markers.

The MLP, representing neural network-based approaches, showed varying degrees of confidence in its predictions, possibly due to its sensitivity to the scale of input data and the choice of architecture. The observed errors suggest that further optimization of network hyperparameters or the inclusion of regularization techniques might enhance its predictive capacity.

These findings highlight the importance of comprehensive model evaluation, not only in terms of overall accuracy but also in considering the reliability of predictions for each class. It also points to the significance of probability scores in understanding model confidence and making informed decisions in species conservation efforts.

Conclusion and Future Work

The study presented here effectively demonstrates the use of machine learning techniques, particularly XGBoost, in the challenging field of species delimitation based on SNP data. The XGBoost model, in particular, has shown remarkable accuracy and prediction confidence, establishing its potential as a robust tool for taxonomic classification. Despite its success, the research also unveiled several challenges, such as misclassifications when dealing with species that have close genetic relationships. These findings highlight the necessity for more refined data and sophisticated modeling techniques to accurately delineate species.

Looking ahead, the potential for enhancing species delimitation using machine learning is vast. Exploring additional machine learning models could provide complementary strengths to those observed with XGBoost and MLP. The integration of deep learning and ensemble methods might offer new insights and potentially superior performance in managing complex genetic data. Moreover, augmenting the data used in training these models by increasing sample sizes and diversifying the range of SNP datasets could significantly improve both training and predictive capabilities. Such expansion, including the integration of broader genomic regions and multi-omics data, may deepen our understanding of species differentiation.

Advanced feature selection methods represent another promising avenue for future research. These methods could help pinpoint the key genetic markers that are crucial for distinguishing species, thereby not only enhancing the accuracy of models but also aiding in the understanding of genetic differentiation mechanisms. Additionally, developing a hybrid approach that combines the computational power of machine learning with the detailed biological insights offered by traditional phylogenetic and coalescent-based methods could lead to more accurate and robust species delimitation tools.

Furthermore, applying these methods to other plant families and even non-plant organisms could help validate the versatility and effectiveness of machine learning approaches in species delimitation across various biological systems. This could pave the way for their application in conservation biology, where they could assist in the accurate identification of endangered species and the planning of effective conservation strategies based on detailed understandings of genetic diversity and species boundaries. Through these efforts, we can enhance our ability to understand and protect biological diversity using the latest advances in computational techniques.

References

[1] Adam DL, Matthew KF, Vladimir NM, Remco RB (2014) Species Delimitation using Genome-Wide SNP Data, Systematic Biology 63(4): 534–542.

[2] Camacho, C., Coulouris, G., Avagyan, V. *et al.* (2009) BLAST+: architecture and applications. **BMC Bioinformatics 10**, 421 (2009).

[3] Chen S, Zhou Y, Chen Y, Gu J (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor, *Bioinformatics* 34(17): i884–i890.

[4] Cheng S, Zeng W, Wang J, Liu L, Liang H, Kou Y, Wang H, Fan D, Zhang Z. (2021) Species Delimitation of Asteropyrum (Ranunculaceae) Based on Morphological, Molecular, and Ecological Variation. *Front Plant Sci.* 12:681-864. doi: 10.3389/fpls.2021.681864. PMID: 34567021; PMCID: PMC8461316.

[5] Chen T and Guestrin C. (2016). XGBoost. Proceedings of the 22nd ACM SIGKDD 834 International Conference on Knowledge Discovery and Data Mining. 835 https://doi.org/10.1145/2939672.2939785

[6] Derkarabetian S, Castillo S, Koo PK, Ovchinnikov S, Hedin M. (2019) A demonstration of unsupervised machine learning in species delimitation. *Mol Phylogenet Evol.* 139:106562. doi: 10.1016/j.ympev.2019.106562. Epub 2019 Jul 16. PMID: 31323334; PMCID: PMC6880864.

[7] Ezard THG, Pearson PN, Purvis A (2010) Algorithmic approaches to aid species' delimitation in multidimensional morphospace. *BMC Evol. Biol.* 10, 175.

[8] Ma C, Xin M, Feldmann KA, Wang X (2014) Machine learning–based differential network analysis: A Study of stress-responsive transcriptomes in Arabidopsis. *The Plant Cell* 26(2): 520–537.

[9] Johnson MG, Pokorny L, Dodsworth S et al. (2019) A Universal probe set for targeted sequencing of 353 nuclear genes from any flowering plant designed using k-medoids clustering *Systematic Biology* 68(4): 594–606, https://doi.org/10.1093/sysbio/syy086

[10] Haykin S. (1994). Neural networks: a comprehensive foundation. Prentice Hall PTR.

[11] Hedin M, Derkarabetian S, Alfaro A, Ramírez MJ, Bond JE. (2019) Phylogenomic analysis and revised classification of atypoid mygalomorph spiders (Araneae, Mygalomorphae), with notes on arachnid ultraconserved element loci. *PeerJ.* 7:e6864. doi: 10.7717/peerj.6864. PMID: 31110925; PMCID: PMC6501763.

[12] Pedersen H (2010) Species delimitation and recognition in the Brachycorythis helferi complex (Orchidaceae) resolved by multivariate morphometric analysis, Botanical Journal of the Linnean Society: 162(1): 64–76, https://doi.org/10.1111/j.1095-8339.2009.01015.x

[13] Langmead B and Salzberg SL. (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 4;9(4):357-9. doi: 10.1038/nmeth.1923. PMID: 22388286; PMCID: PMC3322381.

[14] Thompson JB, Hernández-Hernández T, Keeling G, Priest NK (2023) Identifying the multiple drivers of Cactus diversification. *bioRxiv* doi: https://doi.org/10.1101/2023.04.24.538150

[15] Li H. (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27(21) 2987-93.

[16] Li H and Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform, *Bioinformatics* 25(14): 1754–1760, https://doi.org/10.1093/bioinformatics/btp324

[17] Liang, Z., Duan, S., Sheng, J. et al. (2019) Whole-genome resequencing of 472 Vitis accessions for grapevine diversity and demographic history analyses. *Nat Commun* 10: 1190 https://doi.org/10.1038/s41467-019-09135-8

[18] Ma ZY, Wen J, Tian JP, Gui LL, Liu XQ. (2020) Testing morphological trait evolution and assessing species delimitations in the grape genus using a phylogenomic framework. *Mol Phylogenet Evol.* 2020 148:106809. doi: 10.1016/j.ympev.2020.106809. PMID: 32224125.

[19] Manni, M., Berkeley, M. R., Seppey, M., & Zdobnov, E. M. (2021). BUSCO: Assessing genomic data quality and beyond. *Current Protocols*, 1, e323. doi: 10.1002/cpz1.323

[20] Ming, R., VanBuren, R., Wai, C. et al. (2015) The pineapple genome and the evolution of CAM photosynthesis. *Nat Genet* 47:1435–1442.

[21] Pei J, Chu C, Li X, Lu B, Wu Y (2018) CLADES: A classification-based machine learning method for species delimitation from population genetic data. *Mol Ecol Resour.* 18: 1144–1156. https://doi.org/10.1111/1755-0998.12887

[22] Perez, M. F., Bonatelli, I. A. S., Romeiro-Brito, M., Franco, F. F., Taylor, N. P., Zappi, D. C., & Moraes, E. M. (2022). Coalescent-based species delimitation meets deep learning: Insights from a highly fragmented cactus system. *Molecular Ecology Resources.* 22: 1016–1028. https://doi.org/10.1111/1755-0998.13534

[23] Rannala B and Yang Z (2003) Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* 164(4):1645-1656

[24] Sukumaran J and Knowles LL (2017) Multispecies coalescent delimits structure, not species. *Proc Natl Acad Sci USA.* 2017 114(7):1607-1612.

[25] Yang, Z., & Rannala, B. (2010). Bayesian species delimitation using multilocus sequence data. Proceedings of the National Academy of Sciences of the United States of America, 107(20), 9264–9269. https://doi.org/10.1073/pnas.0913022107

[26] Yang L, Kong H, Huang JP and Kang M (2019) Different species or genetically divergent populations? Integrative species delimitation of the Primulina hochiensis complex from isolated karst habitats. *Molecular Phylogenetics and Evolution.* 132: 219-231. https://doi.org/10.1016/j.ympev.2018.12.011

[27] Yardeni G, Viruel J, Paris M, Hess J, Groot Crego C, de La Harpe M, Rivera N, Barfuss MHJ, Till W, Guzmán-Jacob V, Krömer T, Lexer C, Paun O, Leroy T. 2022. Taxon-specific or universal? Using target capture to study the evolutionary history of rapid radiations. *Molecular Ecology Resources* 22: 927-945.

[28] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. arXiv preprint arXiv:1603.02754

[29] Rosenblatt, F. (1962). Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Washington, DC: Spartan Books.