[BIOINFO 2025]

Novel Random forest-based identification of a single chloroplast DNA barcode gene for Lemnoideae Classification

Halim Park[1], Yang Jae Kang[1,2,3,*]

[1] *Division of Bio & Medical Bigdata Department (BK4 Program), Gyeongsang National University, Republic of Korea*

[2] *Division of Life Science Department, Gyeongsang National University, Republic of Korea*

[3] *Research Institute of Molecular Alchemy, Gyeongsang National University, Republic of Korea*

**Corresponding author: kangyangjae@gnu.ac.kr*

Species classification serves as a foundational framework in biological research, crucial for biodiversity understanding and ecological studies. While traditional morphological methods and universal DNA barcoding approaches are useful, they often face limitations, including subjective interpretations, environmental plasticity, and insufficient resolution for distinguishing closely related species. The necessity for combining multiple barcodes to enhance discriminatory power adds considerable time and effort. These limitations are particularly problematic for taxa with simplified morphology, such as duckweed (Lemnoideae), where standard barcoding techniques have shown difficulty in resolving species boundaries. To overcome these challenges and identify a highly effective and practical single marker for species identification, this study proposes a novel data-driven taxonomic framework leveraging machine learning. Focusing on the chloroplast genome due to its advantageous characteristics for plant barcoding, we analyzed Single Nucleotide Polymorphisms (SNPs) located within gene regions, excluding potentially ambiguous intergenic spacers. Utilizing the Random Forest algorithm for its robust feature selection and classification capacities, we identified SNPs that are powerful for duckweed species classification and revealed the single most effective gene region. Our analysis successfully identified the *ndhH* gene as the optimal single barcode for duckweed. This framework provides a powerful and efficient methodological alternative to current barcoding techniques. It is applicable to other groups facing similar taxonomic difficulties and is expected to contribute to a better understanding of biodiversity.