Plant identification **using DNA "barcodes"**

Nic Tippery

Overview

Organismal diversity, often measured as the number of biological species present in a habitat, is a primary concern for lake managers. High-quality, pristine habitats often have an abundance of native species, whereas non-native species, particularly invasive species, can be cause for concern. In many cases, closely related native and non-native species can be challenging to tell apart, even for experts. Aquatic plants are particularly notorious for being difficult to identify, because they often have simplified morphology, because specimens may lack reproductive structures like flowers or fruits, and because many species have the potential to hybridize and produce morphologically intermediate plants. Fortunately, there are methods besides morphological features that can be employed to identify plants.

One strategy is to use "molecular" data, typically DNA sequences, to identify plant specimens. DNA consists of nucleotides (A / C / G / T), and the two strands of DNA make complementary "base pairs"(A with T and C with G) (Figure 1). The full set of nucleotide sequences for an organism comprise its genetic code (or "genome"), basically the instructions for how to build and maintain it. In a string of just 20 nucleotides there are over one trillion ways to arrange the nucleotide letters, so it's effectively impossible for two species to have entirely identical DNA sequences. (However, closely related individuals can have very similar DNA, because they inherited the same genetic code from a recent shared ancestor.) Over the past 50 years, there have been myriad efforts to document the diversity of life by sequencing DNA, sometimes one gene at a time, and sometimes an entire genome (i.e., all the DNA that exists in the organism).

"Identifying" an organism by its DNA sequence amounts to matching a sequence from the unknown organism against a database of known sequences (Figure 2). DNA sequences that have been reported in scientific publications reside on a public database called GenBank (https://www.ncbi.nlm.nih.gov/genbank/). If there are enough sequences published online to make an effective comparison, then new DNA sequences can be matched to those. However, if reference sequences are lacking, new sequences must be generated before an effective comparison can be made. Ultimately, DNA sequences are only reliable if the source material was correctly identified and if the molecular lab work was done carefully. Occasionally there are errors in the system, but these occur very rarely; a recent study found only about 50 misidentified sequences out of around 5,000 that were being evaluated (Tippery 2026).

There have been various efforts to decide on a consensus set of DNA regions that can be used across plants as a "universal barcode". When researchers publish data for the same DNA region across many species, then that region can be sequenced in unknown plants, regardless of which group of plants they belong to. Among the most commonly available and frequently used "barcodes" are several chloroplast regions, including the *matK* and *ndhF* genes. DNA from the

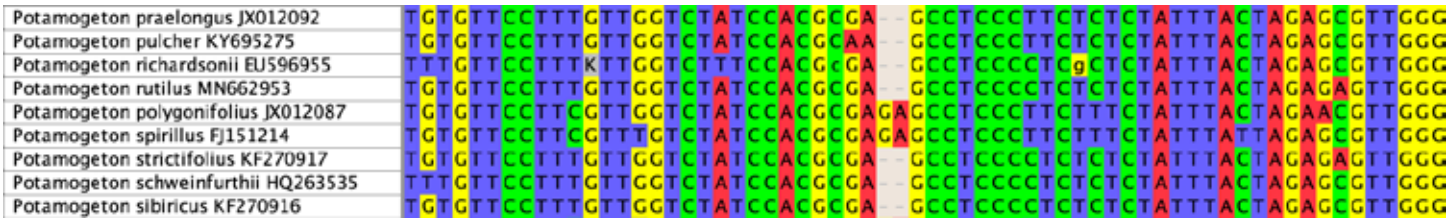

*Figure 1. Portion of an alignment of DNA sequences from the internal transcribed spacer (ITS) region for species of pondweed (*Potamogeton*). One such alignment often consists of 500–2,000 nucleotides (A / C / G / T). (This alignment shows one letter "K," which means that position could be G or T, indicating some uncertainty or variation.) At most of the positions (i.e., columns), the letters are the same, and this gives a measure of confidence that the same DNA region is being compared across all species. Positions with nucleotide differences are useful for identifying specimens and also for research endeavors like constructing phylogenetic trees. Sometimes DNA sequences have insertions or deletions when compared across species, and these appear as gaps in the alignment (i.e., the positions that have no letters). This alignment was visualized using the program Mesquite (https://www.mesquiteproject.org/).*

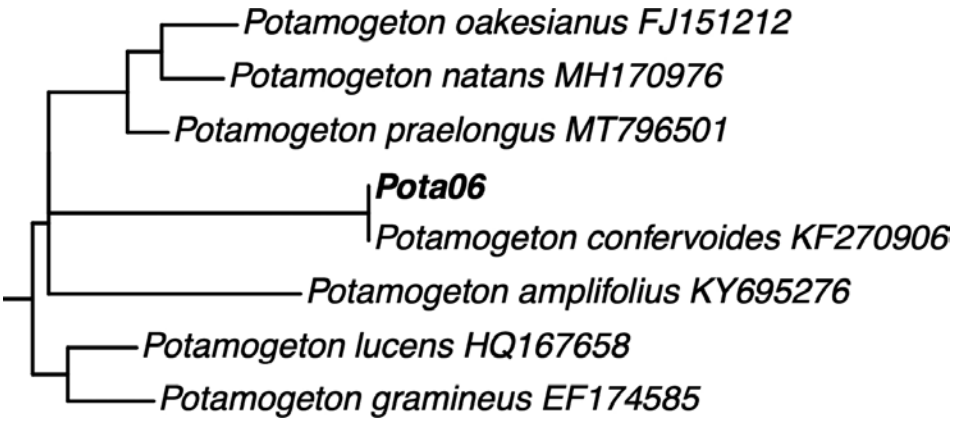

Figure 2. A portion of a phylogenetic tree, showing relationships among species of Potamogeton*. This phylogeny was constructed using DNA sequences from the ITS region. After each species name is an alphanumeric code (e.g., "FJ151212") that represents a sequence in GenBank. The boldface label "Pota06" indicates a plant sample that is being identified, and in this case Pota06 is an identical match to P.* confervoides *KF270906. The phylogenetic tree helps to depict how similar any two species are. Horizontal lines or "branches" have lengths that are proportional to the number of DNA sequence*

chloroplast has two main advantages for barcoding: firstly, there are multiple chloroplasts per cell, thus the DNA can be obtained at a higher frequency from DNA extractions, and secondly, chloroplasts are inherited as one unit, usually from the female parent, thus individual plants usually have only one kind of sequence for a given chloroplast gene. The chloroplast genome is relatively small, at about 150,000 base pairs and 130 genes in most plants (Daniell et al. 2016). DNA from the nucleus, by comparison, is far more expansive and diverse, comprising 125 million base pairs and 25,000 genes in the well-studied model plant *Arabidopsis thaliana* (Arabidopsis Genome Initiative 2000). (Plants also have DNA in their mitochondria, consisting of about 367,000 base pairs and 57 genes, but this DNA has proven to be less useful than chloroplast DNA and is rarely sequenced.)

DNA in the nucleus is far more diverse and expansive than chloroplast DNA. One popular nuclear DNA region used in molecular studies is the internal transcribed spacer (ITS), which is fairly reliable to sequence and provides a rich amount of data across species. Nuclear DNA generally exists as two copies per gene (i.e., "alleles"), and plants can have multiple copies of gene regions if they are "polyploid" (i.e., having 2×, 4×, or more copies of the normal amount of chromosomes). Having multiple gene copies to work with can be helpful for some research endeavors like phylogenetics, but the multiple copies can introduce complications when it comes to simple plant identification. However, nuclear DNA has one major advantage over chloroplast DNA for plant identification, namely that it can be used to identify hybrids. Hybrids are the result of mating between genetically dissimilar individuals, and the term "hybrid" is often used to describe combinations between two biological species. At the level of DNA, hybrids typically retain one version of a nuclear gene from each parental species. Several noteworthy invasive plants are hybrids, such as the hybrid between Eurasian watermilfoil and northern watermilfoil (*Myriophyllum spicatum* × *M. sibiricum*; Moody and Les 2002), and their hybrid origin may allow them to evade detection, adapt in new ways to compete against native species, or escape control efforts like herbicides (Ellstrand and Schierenbeck 2000). Aquatic managers are keen to know when hybrids are present in their lakes, because they might require different control approaches or additional vigilance.

Examples from aquatic plants

Plant identification using DNA barcodes has opened up new avenues for understanding aquatic plant diversity. Several important discoveries have relied upon DNA sequencing, and additional insights are gained every year. Hybrid invasives have been identified in watermilfoil (*Myriophyllum*; Moody and Les 2002), floating heart (*Nymphoides*; Harms et al. 2021), *Phragmites* (Meyerson et al. 2010), and other terrestrial and aquatic plant genera. Some newly arrived aquatic plants in North America have been confirmed using DNA evidence, including *Potamogeton wrightii* (Tippery and Warman 2024), *Utricularia tenuicaulis* and *U.* × *neglecta* (Tippery et al. 2024b), and a distinct subspecies of hydrilla (*Hydrilla verticillata* subsp. *lithuanica*; Tippery et al. 2020; Tippery 2023). A study of DNA sequences in naiads (*Najas*) led to the identification of a previously overlooked species, *N. canadensis*, that is minimally different from the related *N. flexilis* (Les et al. 2015).

Ongoing efforts to catalog and produce reference DNA sequences for plants in North America have enabled a more thorough understanding of some groups. Studies of bladderworts (*Utricularia*; Tippery et al. 2024b), waterhyssop (*Bacopa*; Tippery et al. 2024a), marsh-pennywort (*Hydrocotyle*; Tippery et al. 2026), and others have filled gaps in the available sequence data and enabled all North American species in the respective groups to be identified with barcodes. In addition to targeting specific taxonomic groups, an alternative approach is to barcode all species in a particular geographic region, and such efforts have been fruitful in Canada (Kuzmina et al. 2017) and Wisconsin (Spalink et al. 2018). It is increasingly feasible to build a global database of DNA sequences for all plant species (Zeng et al. 2018; Brewer et al. 2019), and such efforts may be complete in the next decade.

New technologies that enable DNA-based plant identification are enticing and show great promise, but it remains important to stay grounded in traditional taxonomy and morphological plant identification techniques. Plant identifications are only meaningful if we understand them in the context of the biology and ecology of the species involved.

References

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. nature, 408(6814), pp.796-815.

Brewer, G.E., Clarkson, J.J., Maurin, O., Zuntini, A.R., Barber, V., Bellot, S., Biggs, N., Cowan, R.S., Davies, N.M., Dodsworth, S. and Edwards, S.L., 2019. Factors affecting targeted sequencing of 353 nuclear genes from herbarium specimens spanning the diversity of angiosperms. F*rontiers in plant science*, 10, p.477275.

Daniell, H., Lin, C.S., Yu, M. and Chang, W.J., 2016. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. *Genome biology*, 17(1), p.134.

Ellstrand, N.C. and Schierenbeck, K.A., 2000. Hybridization as a stimulus for the evolution of invasiveness in plants?. *Proceedings of the National Academy of Sciences*, 97(13), pp.7043-7050.

Harms, N.E., Thum, R.A., Gettys, L.A., Markovich, I.J., French, A., Simantel, L. and Richardson, R., 2021. Hybridization between native and invasive *Nymphoides* species in the United States. *Biological Invasions*, 23(10), pp.3003-3011.

Kuzmina, M.L., Braukmann, T.W., Fazekas, A.J., Graham, S.W., Dewaard, S.L., Rodrigues, A., Bennett, B.A., Dickinson, T.A., Saarela, J.M., Catling, P.M. and Newmaster, S.G., 2017. Using herbarium-derived DNAs to assemble a large-scale DNA barcode library for the vascular plants of Canada. *Applications in Plant Sciences*, 5(12), p.1700079.

Les, D.H., Peredo, E.L., King, U.M., Benoit, L.K., Tippery, N.P., Ball, C.J. and Shannon, R.K., 2015. Through thick and thin: cryptic sympatric speciation in the submersed genus *Najas* (Hydrocharitaceae). *Molecular phylogenetics and evolution*, 82, pp.15-30.

Meyerson, L.A., Viola, D.V. and Brown, R.N., 2010. Hybridization of invasive *Phragmites australis* with a native subspecies in North America. *Biological Invasions*, 12(1), pp.103-111.

Moody, M.L. and Les, D.H., 2002. Evidence of hybridity in invasive watermilfoil (*Myriophyllum*) populations. *Proceedings of the National Academy of Science*s, 99(23), pp.14867-14871.

Soltis, D.E., Soltis, P.S., Bennett, M.D. and Leitch, I.J., 2003. Evolution of genome size in the angiosperms. *American Journal of Botany*, 90(11), pp.1596-1603.

Spalink, D., Kriebel, R., Li, P., Pace, M.C., Drew, B.T., Zaborsky, J.G., Rose, J., Drummond, C.P., Feist, M.A., Alverson, W.S. and Waller, D.M., 2018. Spatial phylogenetics reveals evolutionary constraints on the assembly of a large regional flora. *American Journal of Botany*, 105(11), pp.1938-1950.

Tippery, N.P., 2023. Subspecies classification of *Hydrilla verticillata* (Hydrocharitaceae). *Journal of the Botanical Research Institute of Texas*, 17(2), pp.401-412.

Tippery, N.P., 2026. Using molecular phylogenetic analyses to assess the timing and biogeography of aquatic plant origins. *Aquatic Botany*, p.103928.

Tippery, N.P. and M.J. Warman. 2024. *Potamogeton wrightii* (Potamogetonaceae) naturalized in Ohio, U.S.A. *Journal of the Botanical Research Institute of Texas* 18: 225-232.

Tippery, N.P., Bugbee, G.J., and S.E. Stebbins. 2020. Evidence for a genetically distinct strain of introduced *Hydrilla verticillata* (Hydrocharitaceae) in North America. *Journal of Aquatic Plant Management* 58: 1–6.

Tippery, N.P., Gonzalez-Socoloske, D., Leliaert, F., Thompson, T.A., Scatigna, A.V., and V.C. Souza. 2024a. Systematics and biogeography of *Bacopa* (Plantaginaceae). *Plant Systematics and Evolution* 310: 7.

Tippery, N.P., Moreland, J., Wild, T., Bugbee, G.J., Stebbins, S.E., Mickley, J.G., Franck, A.R., Wefferling, K.M., and Warman, M.J., 2024b. Toward a comprehensive phylogeny of North American bladderworts (*Utricularia*, Lentibulariaceae). *Aquatic Botany*, 195, p.103794.

Tippery, N.P., Abbott, K.M., Leliaert, F., Wefferling, K.M., Franck, A.R., and Forbes, H., 2026. Phylogenetic relationships and molecular identification of *Hydrocotyle* species in North America (Araliaceae) *Plant Systematics and Evolution*, 312(1), p.7.

Zeng, C.X., Hollingsworth, P.M., Yang, J., He, Z.S., Zhang, Z.R., Li, D.Z. and Yang, J.B., 2018. Genome skimming herbarium specimens for DNA barcoding and phylogenomics. *Plant Method*s, 14(1), p.43.

Nic Tippery is a professor in the Department of Biology at the University of Wisconsin-Whitewater. He conducts DNA sequencing to identify plants, predominantly aquatic plants, and he has published the results of several focused efforts to document the molecular side of plant diversity.

Please take a moment to ensure NALMS has your correct email and mailing address. Log into the member-only area of www.nalms.org to view the information we currently have on file.

Send any corrections to membershipservices@nalms.org