# The development of a DNA barcode system for species identification of *Conyza* spp. (fleabane)

Karen Alpen[1], David Gopurenko[1,2], Hanwen Wu[1,2], Brendan J. Lepschi[3] and Leslie A. Weston[1]
[1] Graham Centre for Agricultural Innovation, Locked Bag 588, Wagga Wagga, NSW 2678, Australia
[2] NSW Department of Primary Industries, Private Mail Bag, Wagga Wagga, NSW 2650, Australia
[3] Australian National Herbarium, Centre for Australian National Biodiversity Research, CSIRO Plant Industry, GPO Box 1600,Canberra, ACT 2601, Australia
(dalpen@bigpond.com.au)

**Summary** The genus *Conyza* Less. includes numerous species of invasive annual weeds that are a threat to the cropping regions of Australia. *Conyza* species are successful ruderal invaders, tolerating a wide range of climates, soils and habitats. There are eight recorded species of *Conyza* in Australia, the most prevalent and invasive being *C. bonariensis* (L.) Cronquist (flax leaf fleabane), *C. sumatrensis* (Retz.) E.Walker (tall fleabane) and *C. canadensis* (L.) Cronquist (Canadian fleabane). These species exhibit differences in susceptibility to commonly applied post emergent herbicides, and herbicide resistance has been confirmed in flax leaf fleabane in New South Wales and other eastern states of Australia. Herbicide application is most effective at an early stage of plant growth before flowering but identification of *Conyza* species using morphological characters at these early stages is often only achievable to the genus level. This study explored the utility of DNA barcoding as a method for accurate and rapid species identification of *Conyza* species in Australia to facilitate management of these weeds. It assessed the ability of one nuclear (ITS) and three chloroplast gene regions (*rbcL*, *matK*, and *trnL-trnF*) to discriminate between *Conyza* species from populations collected across Australia, and also from herbarium voucher specimens for each species. The results showed that a combination of ITS and *rbcL* DNA barcode regions generally provided a suitable platform for potential identification of *Conyza* at the species level.

**Keywords** Flax leaf fleabane, Canadian fleabane, tall fleabane, *Conyza bonariensis, C. canadensis, C. sumatrensis, C. parva, C. bilbaoana, C. aegyptiaca, C. leucantha, C. primulifolia*, DNA sequence analysis, DNA barcoding, ITS, *rbcL, trnL-trnF, matK*.

## INTRODUCTION

DNA barcoding is a method used for genetically identifying taxonomically described species based on sequence content at a standard comparable genomic region (Hebert and Gregory 2005). A single gene region (mitochondrial CO1 gene) has been established as the standard DNA barcode for species identification in animals but the discovery of a similar single gene DNA barcode for plants has proved more challenging. Multiple gene regions are almost always required. The chloroplast gene regions *rbcL* and *matK* where recommended by the CBoL Plant Working Group (2009) as a standard two-locus DNA barcode for identification of plants, but for some genera these gene regions are not powerful species discriminators and additional/alternative gene regions must be identified and used. The aim of this study was to explore several gene regions for their potential use as DNA barcodes for identification of *Conyza* species.

## MATERIALS AND METHODS

DNA sequences were sourced from *Conyza* specimens held in the Australian National Herbarium, Canberra and the Queensland Herbarium, Brisbane; and previously identified field collected/greenhouse specimens. Sequences were also sourced from publicly available gene sequence databases such as GenBank and Barcode of Life Data systems (BOLD). Where possible, specimens used for sequence analysis were obtained from a wide geographical area in order to capture the depth of genetic variation within each species at the particular DNA region under scrutiny. Eighty nine specimens were obtained, of these, 76 generated suitable sequences for further analysis (*C. bonariensis* n = 39; *C. sumatrensis* n = 21; *C. canadensis* n = 6; *C. parva* n = 3; *C. bilbaoana* J.Remy n = 2; *C. primulifolia* (Lam.) Cuatrec. & Lourteig n = 1; *C. aegyptiaca* (L.) Aiton n = 2; *C. leucantha* (D.Don) Ludlow n = 2).

**DNA extraction, PCR and sequencing** Specimen tissue samples (<0.4 $cm^2$) were digested overnight at 55°C in 240 µL of 1% DX digest enzyme/DXT (v/v) digest solution (Qiagen). DNA was extracted using a Corbett Research X-tractor Gene™ (CAS-1820) robot with recommended Qiagen QIAxtractor DNA plasticware and associated DX solid tissue DNA extraction buffers.

PCR was used to amplify chloroplast *(rbcL, matK* and *trnL-trnF*) and nuclear (ITS) gene regions using the forward and reverse primers shown in Table 1.

PCR products were visualised for quality and approximate size using a UV transilluminator (Bio-Rad Molecular Imager® Gel Doc™ XR System) after electrophoresis of 2.5 µL of PCR products and reference size markers through a 1.5% agarose gel immersed in 1% TAE buffer at 190V for seven minutes. Quality PCR products were picked and outsourced to the Australian Genome Research Facility (AGRF) in Brisbane, Australia for purification and bi-directional sequencing.

**Sequence editing, alignment and analysis** Forward and reverse chromatograms at each gene were checked for quality, assembled by specimen ID and aligned to a comparative reference sequence (derived from GenBank), using Lasergene® SeqMan Pro™Seqman software. The chromatograms were first checked for quality of read and as a means to detect and confirm sequence polymorphisms. Sequences identified at AGRF as containing low sequence signal strength (relative to background fluorescent noise) or less than acceptable pre-determined levels of signal homogeneity at >50% of the nucleotides in the sequence read, were discarded. Low quality nucleotide sites observed within sequences which passed these initial criteria were manually scored as unknown sites. Sequences were trimmed at both ends of the alignment to remove primer sequences and poor quality read. Quality checked sequences were further aligned using the CLUSTALW algorithm with default parameters as implemented in Bioedit 7.0. Four methods were used to assess the ability of each gene region to discriminate among the species; i) neighbour-joining (NJ) tree, ii) presence of a DNA barcode gap, iii) BLAST of sequences against GenBank, and iv) diagnostic character method.

**Table 1.** Primers used in PCR.

| Gene region | Primer (direction) | Sequence (sequence direction 5'–3') |
|---|---|---|
| rbcL | rbcLF(F)1 | ATGTCACCACAAACAGAGACTAAAGC |
| rbcL | rbcLajf634(R)2 | GAAACGGTCTCTCCAACGCAT |
| matK | 3FX_KIM(F)3 | CGTACAGTACTTTTGTGTTTACGNG |
| matK | 1RX_KIM(R)3 | ACCCAGTCCATCTGGAAATCTTGGTNC |
| trnL-trnF | Ucp-e(F)4 | GGTTCAAGTCCCTCTATCCC |
| trnL-trnF | Ucp-f(R)4 | ATTTGAACTGGTGACACGAG |
| ITS | ITS5a(F)5 | TATCATTTAGAGGAAGGAG |
| ITS | ITS4(R)5 | GCATATCAATAAGCGGAGGA |

[1] (Kress and Erickson 2007); [2] (Fazekas *et al*. 2008); [3] (Dunning and Savolainen 2010); [4] (Taberlet *et al*. 1991); [5] (Baldwin 1992).

RESULTS

DNA sequences were recovered from all four gene regions, *rbcL, matK, trnL-trnF* and ITS, with high success rates across species and specimens within species. All four gene regions had an amplification success rate > 85% among specimens while sequencing success rates ranged from 93% to 100% of PCR products.

**NJ tree analysis** Tree based analysis using NJ was used to determine the extent of species monophyly at each gene region. A species was considered to be discriminated if all its specimen sequences formed a single, well supported (>70% bootstrap support) monophyletic clade distinct from all other species. The three chloroplast gene regions, *rbcL*, *matK*, and *trnL-trnF*, did not generate NJ trees with monophyletic clusters for each species at this level of support and hence were unable to discriminate among all the *Conyza* species. The nuclear ITS region exhibited the greatest species discrimination and was able to separate five of the eight species.

**DNA barcode gap analysis** A DNA barcode gap is observed among species at a gene when the minimum interspecific distance is higher than the maximum intraspecific distance (Hebert *et al*. 2004). The greater the overlap between intraspecific and interspecific genetic distances, the less effective a gene region will be as a DNA barcode for species discrimination (Meyer and Paulay 2005). Each of the three chloroplast gene regions, *rbcL*, *matK*, and *trnL-trnF,* had considerable overlap between intraspecific and interspecific distances indicating absence of a DNA barcode gap. The ITS region also displayed a region of overlap but it was marginal compared to those of the chloroplast regions. Additional analysis of the ITS region identified absence of a DNA barcode gap at five of the 28 possible pairwise species comparisons.

**Sequences queried against GenBank** Sample sequences at each gene region were queried against pre-existing sequence accessions at the GenBank database using the BLAST algorithm. The absence of some *Conyza* species in the GenBank database at particular gene regions prevented some sequences from being further scrutinised. The chloroplast regions had less ability than the ITS region to discriminate at the species level with both positive and unique identification rates of 0%, 85% and 78% for the *rbcL, matK and trnL-trnF* regions

respectively. The ITS region was able to positively identify 100% of the sequences queried against pre-existing specimen accessions**.** The complete failure of the *rbcL* gene region to correctly identify all species can be attributed to two GenBank records which have potentially been misidentified. These two GenBank sequences were also anomalies in the NJ tree analysis.

**Diagnostic character based analysis** An analysis of the ITS sequences present among the surveyed *Conyza* species identified several polymorphic nucleotide positions that were unique in character state (A, C, G or T) for particular *Conyza* species and therefore potentially useful as diagnostic nucleotide characters for distinguishing species. Four of the species, *C. aegyptiaca, C. leucantha, C. parva* and *C. primulifolia* have multiple nucleotide sites which can uniquely identify each of these species from the eight *Conyza* species under review. *Conyza aegyptiaca* has at least thirteen unique nucleotide sites that are fixed for a character observed at all specimens within the species but absent at all other surveyed *Conyza*; *C. leucantha* possesses seven sites, while *C. parva* and *C. primulifolia* both possess two unique sites. *Conyza canadensis* has four nucleotide sites that can be used to distinguish this species from *C. sumatrensis, C. bonariensis* and *C. bilbaoana*. *Conyza sumatrensis* has one site that allows for separation from *C. bonariensis* and *C. bilbaoana*. In summary, diagnostic nucleotide sites exist in the ITS region that can separate all the *Conyza* species reviewed in this study except for *C. bonariensis* and *C. bilbaoana.* Regarding *C. bonariensis* and *C. bilbaoana,* two nucleotide sites (positions 143 and 551 (Table 2)) were of further interest. The two nucleotide sites distinguished *C. bilbaoana* from all Australian sourced *C. bonariensis* samples, but not the GenBank accessions of *C. bonariensis* sourced from Hawaii and Taiwan. These two overseas *C. bonariensis* specimens differed from the Australian sourced *C. bonariensis* at these two sites, and shared the same nucleotide type as that of *C. bilbaoana.*

**Table 2.** Polymorphic nucleotide sites of interest in the ITS alignment at *C. bonariensis* in comparison to *C. bilbaoana*.

| | | Nucleotide position | |
|---|---|---|---|
| Species | Sample source | 143 | 551 |
| *C. bonariensis* | 37 × Australian specimens plus 1 × overseas specimen | C | A |
| *C. bonariensis* | 2 × GenBank accessions* | A | G |
| *C. bilbaoana* | 2 × Australian specimens | A | G |

* Specimens from Hawaii and Taiwan.

A review of the *rbcL* gene region for diagnostic nucleotide sites identified one potential unique diagnostic site that may differentiate *C. bonariensis* from all other species including *C. bilbaoana* but the results were confounded by the two potentially misidentified GenBank specimens discussed previously in the section titled 'Sequences queried against GenBank'.

## DISCUSSION

All four methods of analysis demonstrated that the chloroplast gene regions (*rbcL*, *matK* and *trnL-trnF*) individually were poor species discriminators of *Conyza*. The nuclear ITS region was however superior in all cases.

NJ trees are regularly used in DNA barcoding for species discrimination (Van Velzen *et al*. 2012). In the case of *Conyza*, NJ trees based on individual and concatenated gene regions were unable to differentiate among all eight species. NJ analysis of the ITS region provided greater resolution than at the chloroplast regions and supported the monophyly of five of the eight morpho-species. The genetic distance relationships among *C. bonariensis*, *C. bilbaoana* and *C. sumatrensis* were ambiguous and poorly resolved at all gene regions using NJ and bootstrap analysis. *Conyza bonariensis* was paraphyletic with respect to both *C. bilbaoana* and *C. sumatrensis*; the latter two species were each well supported as reciprocally monophyletic sister species but they only differed from one another by a few nucleotides and were nested within the clade containing all *C. bonariensis*. There was however, no sharing of identical DNA barcode sequences among these three species. The close genetic relationship among the three species indicates they recently diverged from a shared common ancestor; it is therefore likely the paraphyly of *C. bonariensis* relative to *C. sumatrensis* and *C. bilbaoana* is due to incomplete sorting of ancestral polymorphisms among the species (Fazekas *et al*. 2009). The absence of well resolved species monophyly also indicates there has been insufficient time for novel and unique mutations to accumulate within the species at the surveyed genes (Van Velzen *et al*. 2012). Our results indicate genetic distance based methodologies, such as NJ cannot accurately identify all species within *Conyza* at the genes surveyed here.

Diagnostic nucleotide character methods have been found to outperform genetic distance based methods (such as NJ trees) when dealing with recently diverged species as these methods look for unique nucleotide site characters in the DNA sequence that are fixed within a species and different to that observed at other species within a genus (Lowenstein *et al*. 2009).

Diagnostic sites at the ITS region were identified that could separate all species except *C. bonariensis* and *C. bilbaoana*. The separation of *C. bonariensis* from *C. bilbaoana* using the ITS region was thwarted by polymorphisms present in the population distribution of *C. bonariensis* at two nucleotide sites (Table 2). At both sites, all 38 *C. bonariensis* samples (primarily Australian samples), carried a fixed nucleotide site which differed from *C. bilbaoana*. However, two *C. bonariensis* GenBank sequences sampled from outlying Pacific localities carried two different nucleotides to the Australian *C. bonariensis* and these nucleotides were shared with *C. bilbaoana*. Confirmation of the taxonomic identity of these two overseas specimens is required.

This failure of the ITS region to separate *C. bonariensis* from *C. bilbaoana* necessitated the review of the chloroplast regions for further diagnostic sites. One site in *rbcL* proved successful in the task of discriminating *C. bonariensis* from all other species under review. However, this is subject to further clarification of the potentially incorrect identification of GenBank specimens as previously discussed.

These results highlight the importance of wide geographical sampling in DNA barcoding studies. In the absence of the two overseas *C. bonariensis* ITS samples, it may have been erroneously concluded that the ITS region alone was able to separate all the eight *Conyza* species existing in Australia. Once specimens from a wider geographical area were included in this study intraspecific variation at the ITS region increased, eliminating several of the potentially diagnostic nucleotide sites available for species discrimination. Additional samples of *C. bonariensis,* and *C. bilbaoana* from the species' natural distribution and other exotic locations outside of Australia are required to ensure intraspecific variation has been adequately surveyed, and to further assist in clarification of the genetic relationship between the two species.

Subject to the results of future additional sampling and clarification of the two potentially misidentified GenBank samples, a combination of the ITS and *rbcL* gene regions is proposed as a suitable multi-locus DNA barcode for genetic identification of species in the *Conyza* genus. This recommendation is based on the presence of key diagnostic nucleotide sites within these regions that individually or in combination are able to identify each of the eight *Conyza* species found in Australia.

## ACKNOWLEDGMENTS

We thank the Graham Centre for Agricultural Innovation who funded this project via a Research Initiative Grant, Wagga Wagga Agricultural Institute for providing access to the DNA barcoding research laboratory, and Tony Bean of the Queensland Herbarium for providing material of *C. aegyptiaca* and *C. leucantha*.

## REFERENCES

Baldwin, B.G. (1992). Phylogenetic utility of the internal transcribed spacers of nuclear ribosomal DNA in plants: An example from the compositae. *Molecular Phylogenetics and Evolution* 1, 3-16.

CBOL Plant Working Group. (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences* 106, 12794-7.

Dunning, L.T. and Savolainen, V. (2010). Broad-scale amplification of *matK* for DNA barcoding plants, a technical note. *Botanical Journal of the Linnean Society* 164, 1-9.

Fazekas, A.J., Burgess, K.S., Kesanakurti, P.R., Graham, S.W., Newmaster, S.G., Husband, B.C., Percy, D.M., Hajibabaei, M. and Barrett, S.C. (2008). Multiple Multilocus DNA Barcodes from the Plastid Genome Discriminate Plant Species Equally Well. *Plos One* 3, e2802.

Fazekas, A.J., Kesanakurti, P.R., Burgess, K.S., Percy, D.M., Graham, S.W., Barrett, S.C.H., Newmaster, S.G., Hajibabaei, M. and Husband, B.C. (2009). Are plant species inherently harder to discriminate than animal species using DNA barcoding markers? *Molecular Ecology Resources* 9, 130-9.

Hebert, P.D. and Gregory, T.R. (2005). The Promise of DNA Barcoding for Taxonomy. *Systematic Biology* 54, 852-9.

Hebert, P.D., Stoeckle, M.Y., Zemlak, T.S. and Francis, C.M. (2004). Identification of Birds through DNA Barcodes. *PLoS biology* 2, e312.

Kress, W.J. and Erickson, D.L. (2007). A Two-Locus Global DNA Barcode for Land Plants: the Coding *rbcL* Gene Complements the Non-Coding *trnH-psbA* Spacer Region. *Plos One* 2, e508.

Lowenstein, J.H., Amato, G. and Kolokotronis, S.-O. (2009). The Real maccoyii: Identifying Tuna Sushi with DNA Barcodes–Contrasting Characteristic Attributes and Genetic Distances. *Plos One* 4, e7866.

Meyer, C.P. and Paulay, G. (2005). DNA Barcoding: Error Rates Based on Comprehensive Sampling. *PLoS biology* 3, e422.

Taberlet, P., Gielly, L., Pautou, G. and Bouvet, J. (1991). Universal primers for amplification of three non-coding regions of chloroplast DNA. *Plant molecular biology* 17, 1105-9.

Van Velzen, R., Weitschek, E., Felici, G. and Bakker, F.T. (2012). DNA Barcoding of Recently Diverged Species: Relative Performance of Matching Methods. *Plos One* 7, e30490.