Identification of weed seeds in soil samples intercepted at the New Zealand border

T.K. James1,4, P.D. Champion2, C.A. Dowsett1,4, M.R. McNeill1,4 and G.J. Houliston3,4

1AgResearch, Ruakura Research Centre, Private Bag 3123, Hamilton 3240, New Zealand
2National Institute of Water & Atmospheric Research, PO Box 11-115, Hamilton 3251, New Zealand
3Landcare Research, PO Box 69040, Lincoln 7640, New Zealand
Corresponding author: trevor.james@agresearch.co.nz

Abstract Identification of weed seeds from overseas countries can be problematic, particularly when diagnostic tools are lacking or incomplete. A well trained seed analyst will usually be able to identify seed to generic level but not always to the species level. Resources for identification of mature plants are usually more complete. Using a seed germination method for intercepted soil samples achieved two goals; it provided an easier route to identification and a measure of viable seed. The drawback of this method is the need to grow the plants through to flowering. DNA identification is an emerging method for more rapid identification, but it relies on availability of matching sequences in an existing database and validation of source plant identification with reliable voucher specimens. The limitations and advantages of both techniques are discussed, and ways in which timely and accurate identification can be provided for biosecurity practitioners are suggested.

Keywords weed seeds, identification, biosecurity, taxonomy, DNA.

INTRODUCTION

A cornerstone to the plant biosecurity of New Zealand’s borders to pest incursions is the accurate and timely identification of unknown species for developing appropriate management response. For plants, this can be difficult as they are often intercepted in their dormant form, i.e. as seeds. Interception of non-indigenous seeds may occur in many ways, e.g. as contaminants of imported seed or fresh produce (Rahman et al. 2005), on footwear (McNeill et al. 2011) or in soil attached to containers, machinery or vehicles (M.R. McNeill, AgResearch, unpublished data). Determining the identity of the seed arriving at the border at the species level provides plant biosecurity authorities with information on potential risks as well as an opportunity to develop mitigation strategies for protecting primary industry and the natural estate. For similar reasons, the importation of seeds for sowing is controlled by the Ministry for Primary Industries through its Import Health Standards (MPI 2011; MPI 2013), which contain a schedule of 291 regulated (quarantine) weed seeds. No seed lot will be released for sowing if seed of any of the quarantine weed species are found to be present or if it contains any unidentified seed (MPI 2013). Presently all seed identification is carried out taxonomically using a variety of techniques.
Biosecurity of resources including voucher specimens, taxonomic keys and image resources (e.g. Musil 1963; Martin & Barkley 2000; Anonymous 2004; James et al. 2013). This system works well for species that have distinctive seeds, e.g. *Tribulus terrestris* and *Emex australis*, but it is difficult where seeds are very similar to related species already present in New Zealand, e.g. *Amaranthus blitoides* and *Digitaria abyssinica*.

In other situations, contaminant weed seeds may be difficult to find and identify. When seeds contaminate soil, their small size and colour can make them difficult to find. Although there are several methods, such as flotation, washing or dry sieving, to separate seed from soil, all are time consuming or require specialised equipment. A simple way to circumvent this is to germinate the seeds while still in the soil sample. This also bypasses the need for complex viability testing but may not find dormant seed or those not favoured by the germination conditions. With this method the plants are grown on until sufficiently mature for taxonomic identification. For some common species identification may be possible at an early growth stage, but for many species the plants need to either be grown to flowering or tested using molecular methods. These methods are usually easier and more reliable than seed identification and in most cases lead directly to a species whereas seed identification often stops at the generic level. The downside of growing plants to flowering, assuming that the seed can be germinated, is the length of time required for the plants to grow, premature plant mortality that prevents identification, and the need for containment facilities to maintain plants in quarantine.

Molecular methods have the advantage of being suitable for both seed and vegetative material. However, they require a specialised laboratory and rely on either a verified reference sample in one of the DNA databases or a well characterised collection for comparison, usually either cultivated material or herbarium voucher specimens.

This paper reports on the identification of 11 unknown and suspected ‘new to New Zealand’ plant species whose seeds were present in intercepted soil samples taken from sea freight, using both classical plant diagnostics and DNA analysis. The challenges of developing an effective identification method for intercepted seed samples using the currently available technology are discussed.

**MATERIALS AND METHODS**

**Germination of seeds and taxonomic identification of plants**

Soil samples were taken as part of a study looking at the risk organisms associated with soil arriving on sea containers and general freight into New Zealand’s seaports. Samples were collected at the Ports of Auckland container terminal and shipped under quarantine conditions to the Ruakura PC2 plant facility. Soil samples (5-500 g) were pre-soaked with water for 72 h prior to being laid out in a 5 mm thick layer on a paper towel over a wet medium (potting mix) for small samples or on a layer of towels in a small tray for the larger samples. The soil was kept moist under natural light in the PC2 glasshouse (15°C–35°C) and observed daily for germinated seedlings. Germinated seedlings were transplanted into sterile potting mix 1-2 days after emergence and grown on for identification to the lowest possible taxonomic level. When the plants had flowered, voucher specimens were collected and transferred to the Allan Herbarium’s PC2 specimen reception. Aids to identification were a combination of web-based keys (e.g. [http://www.efloras.org/index.aspx](http://www.efloras.org/index.aspx)) and published literature relevant to the suspected origin of the soil samples (e.g. Whistler 1995).

**Molecular identification of plants**

Molecular identification of plant material was performed on the 11 plant species by Ecogene, Landcare Research, where morphological identification was uncertain. DNA was extracted from leaf tissue using a Qiagen DNeasy Plant Mini kit (Qiagen, USA) following the manufacturer’s instructions. DNA samples were PCR amplified and sequenced for the chloroplast and ribosomal nuclear regions, following the protocols in the papers cited for each gene region (see below). The primers chosen were known to cross-amplify in diverse plant families and focussed on the following six regions: (1) a combination of ITS1 and ITS2 that flank the 5.8s nuclear ribosomal DNA region (White et al. 1990); (2) chloroplast trnL-trnF spacer region (Sang et al. 1997); (3) chloroplast *rbcL* gene – when possible the complete gene (about 1200 base pairs (bp)) using
primers developed by Steve Wagstaff (Landcare Research) or alternatively approximately 600 bp using internal primers from Levin et al. (2003) and Kress & Erickson (2007); (4) chloroplast ndhF gene (Olmstead & Sweere 1994); (5) intron region of a chloroplast tRNA gene (trnL, Taberlet et al. 1991); and (6) the external transcribed spacer (ETS) of the 18S–26S nuclear ribosomal DNA from Wright et al. (2001).

For each sample, multiple gene regions were included; these were selected from the list above as being most likely to be diagnostic based on the morphological identification. To do this, the utility of each of the regions for other congeneric species was compared, as an indicator of how variable that region is within the genus. For some individuals this would indicate that either other commonly used regions should be added or that regions particularly informative for the genus should be used. In this study the latter option was not pursued, but rather the question of how far towards a diagnosis could be achieved using widely amplifiable gene regions was examined.

Edited DNA sequences (ends trimmed and base reads checked against chromatograms) were compared against sequences from GenBank (www.ncbi.nlm.nih.gov/Genbank/) using a BLAST search (Altschul et al. 1997). Original comparisons were made in 2011 to NCBI Genbank, and these were repeated in 2014, to determine if more reference sequence had been deposited in this time.

RESULTS

Plant ID No. 618833 Ludwigia octovalvis (Jacq.) Raven

A wetland forb of the Onagraceae family, this species is native to tropical America but is now pantropic in distribution (Whistler 1995). It closely resembles its more widespread relative L. hyssopifolia (G.Don) Exell, which has previously been reported as an incursion species with cocopeat (James et al. 2011). However, L. octovalvis has larger flowers, 25-35 mm diameter cf. L. hyssopifolia 6-8 mm diameter. There was no exact GenBank match for this species in 2011, with a BLAST search finding the closest match for L. hyssopifolia for one nuclear and two chloroplast regions (Table 1).

In 2014 an ITS sequence for L. octovalvis from Taiwan was lodged in GenBank, which was a better match than the previous nuclear data (99% similarity versus a 97% similarity for L. hyssopifolia).

Plant ID No. 618834 Hyptis pectinata (L.) Poit.

A forb of the Lamiaceae family, this species is also native to tropical America and now widespread through the tropics. It is a serious weed of pastures due to grazing avoidance (Whistler 1995). This plant grows to 3 m tall and had to be grown to its full height before it flowered and was able to be identified. The best GenBank matches for this species in 2011 were two genera of Lamiaceae (Hyptis sp. and Hanceola sp.). Plectranthus barbartus was another species with a close match (Table 1). All data generated were for chloroplast regions.

In 2014 there were several new GenBank entries for ITS with a better match than the previous references, the best of which was an entry for H. pectinata which had a match percentage of 98%, versus 93% and 92% respectively for the previous search in 2011.

Plant ID No. 618836 Chamaesyce hyssopifolia (L.) Small

A common synonym for this low growing forb of the Euphorbiaceae family is Euphorbia hyssopifolia. This native from tropical and temperate America is a quick growing, frost tender summer annual that quickly colonises waste and amenity areas. BLAST searches in 2011 found this plant to have a 100% trnL-F to Chamaesyce lasiocarpa (Table 1), whereas rbcL matched at 96% to C. lasiocarpa. Chamaesyce lasiocarpa and C. hyssopifolia are taxonomically quite different. The fruit of C. lasiocarpa are hairy whereas the plant in the present study had hairless fruit; the leaves of C. lasiocarpa are serrated and hairy whereas those on the plant being investigated had few serrations and were hairless; C. lasiocarpa grows to 2 m tall whereas this plant was prostrate. Thus, despite the molecular analysis, the taxonomic description led to the conclusion that the plant in the present study was C. hyssopifolia.

Plant ID No. 618839 Pycreus polystachyos (Rottb.) P.Beauv.

In its early stages of flower head development, this plant was identified as Cyperus iria, but...
Table 1  Taxonomic and DNA identification of 11 unknown plant species intercepted in soil samples at New Zealand’s border.

<table>
<thead>
<tr>
<th>Plant ID No</th>
<th>Origin</th>
<th>Sea freight type</th>
<th>Taxonomic identification</th>
<th>Best molecular match from a BLAST search in NCBI GenBank</th>
<th>GenBank match (%)¹</th>
</tr>
</thead>
<tbody>
<tr>
<td>618833</td>
<td>Tonga</td>
<td>Sea container</td>
<td>Ludwigia octovalvis (Jacq.) Raven</td>
<td>Ludwigia hyssopifolia</td>
<td>99% (trnL)</td>
</tr>
<tr>
<td>618834</td>
<td>Tonga</td>
<td>Sea container</td>
<td>Hyptis pectinata (L.) Poit.</td>
<td>Hyptis sp. or Hanceola sp.</td>
<td>99% (trnL-trnF)</td>
</tr>
<tr>
<td>618836</td>
<td>Tonga</td>
<td>Sea container</td>
<td>Chamaesyce hyssopifolia (L.) Small</td>
<td>Chamaesyce lasiocarpa</td>
<td>100% (trnL-trnF)</td>
</tr>
<tr>
<td>618839</td>
<td>Tonga</td>
<td>Sea container</td>
<td>Cyperus iria L.</td>
<td>Pycreus nuerensis</td>
<td>100% (rbcL)</td>
</tr>
<tr>
<td>618840</td>
<td>American Samoa</td>
<td>Sea container</td>
<td>Eriochloa procera (Retz.) C. E. Hubb.</td>
<td>Urochloa trichopus</td>
<td>99% (ITS)</td>
</tr>
<tr>
<td>618841</td>
<td>Tonga</td>
<td>Sea container</td>
<td>Fimbristylis complanata (Retz.) Link</td>
<td>Fimbristylis cymosa</td>
<td>99% (ITS)</td>
</tr>
<tr>
<td>618842</td>
<td>Australia</td>
<td>Used machinery</td>
<td>Tribolium sp.</td>
<td>Tribolium obliterum</td>
<td>100% (ndhF)</td>
</tr>
<tr>
<td>618843</td>
<td>Wallis and Futuna Islands</td>
<td>Flat rack</td>
<td>Chamaesyce prostrata (Aiton) Small</td>
<td>Chamaesyce prostrata</td>
<td>100% (trnL-trnF)</td>
</tr>
<tr>
<td>618846</td>
<td>Tonga</td>
<td>Sea container</td>
<td>Macroptilium atropurpureum</td>
<td>Macroptilium atropurpureum</td>
<td>100% (trnL-trnF)</td>
</tr>
<tr>
<td>618847</td>
<td>Tonga</td>
<td>Sea container</td>
<td>Chamaesyce hypericifolia (L.) Millsp</td>
<td>Chamaesyce hypericifolia</td>
<td>100% (ITS)</td>
</tr>
<tr>
<td>618848</td>
<td>Vanuatu</td>
<td>Sea container</td>
<td>Cyperus sp.</td>
<td>Cyperus iria or Pycreus sanguinolentus</td>
<td>Both 100% (rbcL part sequence)</td>
</tr>
</tbody>
</table>

¹Gene regions sequenced: rbcL (ribulose-bisphosphate carboxylase gene), ITS (Internal transcribed spacer of the 18S ribosomal gene), trnL-trnF (trn spacer regions of the chloroplast, with capital letters denoting positions on the chloroplast), ndhF (NADH dehydrogenase F gene).
doubt was cast on that identification when the individual seed heads did not elongate as for C. iria but remained compact at maturity. The GenBank search in 2011 matched this plant to Pycreus nuerensis or a close match to P. mundtii (Table 1). However, the plant being investigated did not match the taxonomic description for P. nuerensis, “Distinguished from the other dark brown or black spikeletted species by the broader flat leaves and long linear spikelets” (Anonymous 2014) and a broad-oblong brown nut (Thistleton-Dyer 1902). The present plant had light brown or tan spikelets that were not elongated and the nut trigonal and elongated. Further, P. nuerensis is a threatened species that appears to have a distribution limited to tropical Africa. Pycreusmundtii also originates from tropical Africa but has also spread to the Mediterranean countries. However, this species is strongly rhizomatous while the present plant had fibrous roots. It is also notable that, using BLAST, this specimen did not match to C. iria for rbcL, despite there being a reference for this species in GenBank.

No new GenBank entries since 2014 improved identification of this species, other than further confirming it is in the subfamily Cyperaceae.

Based on the genera identification from the GenBank search, this species was taxonomically identified as P. polystachyos.

Plant ID No. 618840 Erichloa prosera (Retz.) C. E. Hubb.
A tufted grass species of the Poaceae family, this plant is native to the African/Asian tropics. Identified from the taxonomic description by Whistler (1995), there was no GenBank match in 2011; the closest reference was to Urochloa trichopus (Table 1). This species, although superficially similar to Erichloa prosera, appears to be limited in its distribution to tropical Africa and its taxonomy is not well documented (Anonymous 2014b).

Identification of this specimen was not improved by later GenBank entries, rather, these entries further reinforced that this group is poorly resolved and likely to be taxonomically difficult.

Plant ID No. 618841 Fimbristylis cymosa R. Br. var. cymosa
This member of the Cyperaceae family was initially identified from the Flora of Pakistan (www.eFloras.org) as Fimbristyis complanata. However, there was no 100% sequence match from GenBank in 2011 with the closest match being a member of the same genus F. cymosa (Table 1), and the reference sequence for F. complanata in GenBank having low similarity to this specimen (91%). The plant in this investigation closely fitted the descriptors for F. complanata, e.g. plant height, leaf width, leafy bracts and inflorescence structure, but the nut colour of the present specimen was dark brown, not white. Descriptions for F. cymosa varied between floras (North America versus Pakistan, www.eFloras.org) with the present specimen matching the Flora of North America. Further taxonomic referencing carried out at the Allan Herbarium confirmed the GenBank identification and added that it was the variety cymosa.

New GenBank entries accessed in 2014 for rbcL confirmed this is Fimbristyis, but offered no further resolution.

Plant ID No. 618842 Tribolium oblitterum (Hemsl.) Renvoize
This tufted, South African grass (Poaceae) was initially only identified to genus level. However, there was 100% match with Tribolium oblitterum in GenBank in 2011 (Table 1). The present plant was then compared to the taxonomic description (Jessop et al. 2006) and found to be a good fit.

Plant ID No. 618843 Chamaesyce prostrata (Aiton) Small
Similar in many respects to C. hyssopifolia above, this member of the Euphorbiaceae is characterised by its smaller leaves and leafy cymes. The GenBank match in 2011 for this DNA sequence agreed with the morphological determination.

Plant ID No. 618846 Macroptilium atropurpureum (DC.) Urb.
This vine of the Fabaceae family has dark maroon flowers and is a native to tropical America (Whistler 1995). GenBank also had a match for this species in 2011, specifically a voucher specimen from Canada. A further specimen in GenBank had a lower match percentage, and was collected from Argentina.
Plant ID No. 618847 *Chamaesyce hypericifolia* (L.) Millsp.
The third member of the Euphorbiaceae identified here, this species is an upright growing plant with larger leaves than either of the two previous species. GenBank matches in 2011 for both nuclear and chloroplast sequence agreed with this determination.

Plant ID No. 618848 *Cyperus iria*
This sedge is widespread throughout the tropics, reaching into sub-tropical and temperate areas in some regions. It is a common weed of rice paddies and is widely referred to as rice sedge. GenBank BLAST searches in 2011 for two different parts of *rbcL* found matches with sequences for *Cyperus iria* and *Pycreus sanguinolentus*, indicating that this is likely a species where limited reference data were available.

New data in 2013, lodged after the initial GenBank search had been performed, has a 100% match for *rbcL* and *trnL-K* for *C. microiria* from South Korea, which may represent another putative identification for this individual. Another sequence for *rbcL* for *C. microiria* was lodged in this time, but has a lower match percentage (97%). However, that sequence was generated from a metagenomics study and therefore is likely to not have optimal sequence quality. The shape of the flower head indicates that it is not *P. sanguinolentus* but rather one of the *Cyperus* sp. with *C. iria* having the best taxonomic fit.

**DISCUSSION**
There is considerable diagnostic taxonomic literature available with many countries having very good descriptions of their flora, and most are either in English or have an English language version. If the approximate source of the unknown plant is known, then the identification process can quickly access a relevant flora. However, the descriptions for the flora of some regions remain poor or non-existent. This was the situation with most of the 11 seeds in the present study, which were suspected of originating from the South Pacific Islands, the shipping ports for the containers. For this region there are some helpful botanical texts, but these are usually limited to the common or readily identified species. For the 11 species grown from seed in this study, the dicotyledonous flowering plants could be quickly identified, but identification of the grass and sedge species was more difficult due to less variation in their leaf shape and general form. This is reflected in the initial results for taxonomic identification where there were seven unknown species correctly identified to species level (six dicots and one monocot) compared with three only identified to genus level (all monocots and including an incorrectly identified species). The remaining species, a sedge (*Pycreus polystachyos*), was incorrectly identified at genus level.

The sequencing of DNA data is a new and accurate tool for plant identification. However, it relies on DNA sequences for the species in question being present in a reference library (GenBank) and also that the identification of the reference samples has been done correctly. In the present study, DNA sequencing produced only four complete identifications, two for dicots and one each for a grass and a sedge. It is important to note however that this method corrected the misidentification of a sedge (*Fimbristylis cymosa*) and resolved the identification of a grass (*Tribolium oblitterum*) that had been only resolved to the genus level through taxonomic identification. This method also offered strong leads with another sedge (*Cyperus iria*) that subsequently enabled taxonomic identification. In a further four instances the sequence matching was correct at the genus level and could have been of assistance in narrowing the taxonomic search if this was required.

Comparisons of the molecular results using data from 2011 and 2014 show an improvement in the resolution of the specimens. Given the volume of sequence data being generated and the almost exponential growth of new accessions in GenBank, this should only improve over time. Another advantage of molecular techniques is that it is not necessary to have a specialist who is able to recognise key diagnostic features for the taxonomic group in question. However, it is imperative that this work is carried out by someone able to interpret the data correctly and recognise where further work may be required. GenBank matches must be critically assessed, both in regard to the validity of the voucher material used (often available through the
linked publications in the database), and for the ability of different gene regions to be diagnostic in different groups. Chloroplast data are often useful in that they are readily generated and their single copy and high copy number nature makes amplification and sequencing technically simple. However, within plant genera chloroplast haplotypes are often either shared among species or have geographic structure (Jakob & Blattner 2006), so some further understanding of the group is required. Nuclear, including ribosomal regions, are often more diagnostic, but can be technically problematic due to multiple copies resulting from polyploidy or duplications (Alvarez & Wendel 2003). Additional issues can arise where non-specialists want to interpret BLAST match percentages as probabilities of identification, whereas they are simply match percentages to the closest sequence in the database. Experience in plant phylogenetics to interpret these data is therefore essential.

While DNA diagnostics have limitations, it is also clear that morphological identifications often don’t result in an absolute identification of a specimen. This can often be for technical reasons (incomplete characters available), but can also be due to error or misinterpretation of keys, even by experts. It can be difficult to detect error in morphological identifications, where further expertise is difficult to access or voucher specimens are geographically isolated.

A practical strategy to improve timely diagnosis of plant species would be to use both methods in combination and utilise the data from one method to inform the other. Initial morphological or molecular identification can be used to narrow the field. For both methods, the use of vouchered material can greatly improve the accuracy of the identification process, by providing comparative characters or the material to generate reference sequences from.

For biosecurity authorities, the combination of both techniques would provide a valuable tool for the rapid identification of invasive plant species that may be a potential threat to the destination country. For ‘plugging the gaps’ in the database, a proactive approach could be to identify the knowledge gaps and formulate a strategy to address them. For example, identification of key plant pests in the country of origin of New Zealand’s major trade partners could be used to develop a reference database that would permit rapid identification, and thereby appropriate control strategies. Such an approach has been applied to identify the potential for weevils and spiders to establish in New Zealand (Phillips & Vink 2013) and the development of a pheromone database of a range of insects including pests (El-Sayed 2014).

ACKNOWLEDGEMENTS
The authors thank staff from the Ministry for Primary Industry and Ports of Auckland for their support in this study and staff, W. Sykes and K. Ford, at the Allan Herbarium for confirming identifications. This research was funded by the Ministry of Business, Innovation and Employment via AgResearch through the Better Border Biosecurity research collaboration (www.b3nz.org).

REFERENCES


