2010

Analysis of putative DNA barcodes for identification and distinction of native and invasive plant species

Shari Laprise
Babson College

Vikki Rodgers
Babson College

This research was supported by the **Babson Faculty Research Fund**
Summer Stipend / 2009

BABSON
FACULTY RESEARCH FUND

ANALYSIS OF PUTATIVE DNA BARCODES FOR IDENTIFICATION AND DISTINCTION OF NATIVE AND INVASIVE PLANT SPECIES

SHARI LAPRISE
Babson College
Mathematics and Science Division

VIKKI RODGERS
Babson College
Mathematics and Science Division

No. WP 1014 – SS09

BFRF WORKING PAPER SERIES

This research was supported by the
Babson Faculty Research Fund

Running title: Using DNA barcodes to distinguish native and invasive plant species

DNA barcoding provides a relatively new method for identifying species using short DNA sequences. A region of the mitochondrial cytochrome c oxidase subunit I (*COI*) gene has been validated as a useful DNA barcode for most animal species, but plant identification using this method has been more challenging. In this study we tested three plastid DNA regions (*matK*, *trnH-psbA*, and *rbcL*) for their suitability as DNA barcoding regions. We chose two groups of three common tree species each to perform our experiment, with each group containing two native species and one invasive species. The first group contained three congeneric species: red maple, sugar maple, and Norway maple and the second group contained three more distantly related, but morphologically similar species: staghorn sumac, white ash, and tree of heaven. Our sequence analysis showed that *rbcL* displayed the robustness required for DNA barcodes. We suggest that a tiered-approach for DNA barcoding, involving analysis of *rbcL* followed by *matK* may be the best method. Within our two groups, species identification was successful, but further analysis is needed to assess the utility of DNA barcoding to distinguish in general invasive plant species from native plant species. Our results indicate that more than one DNA barcode may be needed to clearly and correctly identify land plants.

Keywords: DNA barcoding, angiosperms, plastid DNA, *matK*, *psbA-trnH*, *rbcL*, invasive species

Introduction

Traditional methods for identifying and classifying species have relied upon the observation of morphological traits. Characterization of closely related or similar-looking species, however, often requires a more rigorous examination. DNA sequencing can serve as a

standardized method for species identification due to the fact that more closely related species have more homologous DNA sequences; the converse is true for more distantly related species.

DNA barcodes have arisen as a promising new method for correctly identifying species using short DNA sequences (Hebert et al. 2003a). This new technology utilizes a short region of DNA to distinguish taxa, therefore increasing speed, objectivity, and efficiency of species identification (Meyer and Paulay 2005). In animals, a 648 base pair region of the mitochondrial cytochrome c oxidase subunit I (*COI*) gene has been validated as a useful DNA barcode across all phyla in the animal kingdom, except *Cnidaria* (Hebert et al. 2003a, Hebert et al. 2003b, Hebert et al. 2004). Plant identification using DNA barcodes has proven more difficult and currently there is no standardized method for identifying plant species using DNA barcodes (Chase et al. 2007).

In order to have widespread utility, a DNA barcode must be conserved, such that the same genetic sequence can be isolated and analyzed with a universal primer set across diverse species. At the same time, it must also exhibit adequate sequence diversity to facilitate identification at the species level. Unfortunately mitochondrial DNA of land plants is not a valid choice for a DNA barcode region because of its well-known slow evolutionary rate and therefore relatively low variability between different species. The internal transcribed spacer portions of nuclear ribosomal DNA (*nrITS*) region has been utilized as a DNA barcode by a number of groups (Alvarez and Wendel, 2003); however, it suffers from complications that render it unusable as a universal barcode for all land plants (Chase et al. 2005, Kress et al. 2005). *nrITS* has 3–4 times more variable sites that evolve up to 4 times more rapidly than chloroplast DNA (Van den Berg et al. 2000); therefore, this region cannot correctly identifiy species in some groups. *nrITS* also has a complicated molecular evolution pattern, such that multiple copies of

this region are present in some species (Rapini et al., 2006). Lastly, amplifying the *nrITS* region is problematic, and often requires specialized PCR conditions (Chase et al., 2007). For these reasons, *nrITS* is not suitable as a universal DNA barcode for land plants.

In the absence of nuclear genome regions suitable for a DNA barcode in plants (Chase et al., 2007), as well as problems with the plant mitochondrial genome, the only remaining DNA is the plastid (chloroplast) genome. Although plastid DNA has a low level of variability, similar to that for plant mitochondrial DNA, it nonetheless contains regions with sufficient variability, which, in combination, may be the best potential barcodes for plants (Chase et al. 2007).

Lahaye et al. (2008) analyzed the potential of eight loci to serve as components of a plant barcode in 1600 different plant specimens, comprising more than 1000 different species. Data from this study revealed that the combination of two loci, the chloroplast *matK* gene (coding region) and the non-coding *trnH-psbA* spacer region, resulted in correct species identification >90% of the time; the addition of the coding locus *rbcL* did not improve species identification by >3%, with none of the combinations yielding 100% identification (Lahaye et al. 2008). Other studies recognize the utility of the *trnH-psbA* locus for barcoding of plant species (Kress et al. 2005, Newmaster et al. 2006, Chase et al. 2007). Based on the available literature, we chose to evaluate the suitability of either the *matK,* or *rbcL* coding loci, in combination with the non-coding *psbA-trnH* spacer region, as DNA barcodes in land plants. This conclusion is supported by the recommendations of the Consortium for the Barcode of Life (CBOL Plant Working Group 2009).

In order to confirm the utility of a particular locus as a DNA barcode in plants, the following standards must be met: 1) robustness, described as the ability to design universal primer sets and PCR amplification and sequencing conditions in order to produce reliable sequence data across all plant taxa; 2) the region must be short (<1000bp) such that it can be PCR amplified from degraded or forensic samples; 3) resolution, which is defined as having

been conserved throughout >400 million years of land plant evolution such that it is present in all plant species, yet its sequence variable enough to discriminate between specimens at the species level; and 4) the region should be informative such that bioinformatics tools can be used to identify a given species (Taberlet et al., 2006; Chase et al., 2007).

The current study tests three loci – *trnH-psbA*, *matK*, and *rbcL* – for their ability to serve as a DNA barcode in two groups of three species of plants. These species were chosen to represent native and invasive plant species in Massachusetts. Invasive species are those that are non-native and cause environmental or economic harm within the region they invade. A method for identification of invasive plant species which does not rely solely upon morphological characteristics is needed, especially in determining the initial stages of plant invasion, such as the appearance of small seedlings, seeds, pollen and even partially digested leaf fragments within insects, which are often difficult to identify. To this end, we performed experiments designed to reveal the resolution, robustness, and species discriminating power using bioinformatics tools for the three aforementioned loci. Our research is unique in that we have specifically tested the ability of putative DNA barcodes to distinguish closely-related (congeneric) species vs. more distantly related species. Secondly, this study aids in the identification and distinction of native tree species from invasive tree species, as there have been no tests to date for identifying invasive plants using DNA barcoding.

Methods

Two different groups of common New England tree species were chosen for these experiments. The first group consisted of three congeneric species: red maple (*Acer rubrum*), sugar maple (*Acer saccharum*), and the invasive species Norway maple (*Acer platanoides*) (Table 1). The second group of species was chosen due to their morphological similarity, and the fact that the species are not closely

related. This second group consisted of: white ash (*Fraxinus Americana*), staghorn sumac (*Rhus typhina*), and the invasive species tree of heaven (*Ailanthus altissima*) (Table 1).

Field Collection

In August 2008, seven leaf samples of each of the six species (forty-two samples in total) were collected from across the Boston metropolitan area. To reduce the chance of collecting genetically identical individuals, the samples were taken from trees that were at least 200 meters apart. Upon collection, the leaf samples were rinsed with distilled water and immediately immersed in liquid nitrogen and then stored at -80°C until processed.

Molecular Analysis

Individual plant tissue samples were ground in liquid nitrogen and total DNA was extracted using the Dneasy Plant Mini Kit (Qiagen, Valencia, CA). DNA barcode sequences (*matK*, *trnH-psbA*, and *rbcL*) were amplified by the polymerase chain reaction (PCR) using the Taq PCR Core Kit (Qiagen), following previously published specifications for each sequence (Kress et al. 2005, Newmaster et al. 2006, Taberlet et al. 2006, Kress and Erickson 2007, Lahaye et al. 2008). Following verification of amplification by agarose gel electrophoresis, PCR products were purified with the QIAquick Gel Extraction kit (Qiagen) and sent for bidirectional sequencing at The DNA Analysis Facility at Yale University, using both forward and reverse primers for each locus for most samples.

Sequence Analysis

Sequence data was analyzed using Sequencher 4.9, demo version. Sequence data for each locus within each species were compared and contigs assembled for analysis of overlap and intraspecific % sequence divergence. For interspecific sequence analysis, individual comprehensive sequences and smaller contigs assembled to yield comprehensive sequence data were compared for overlap and % sequence divergence.

To verify that sequence data could be identified as the correct locus from the correct species using bioinformatics tools, sequences for each locus of each species sampled were separately entered into the NCBI (National Center for Biotechnology Information) website (http://blast.ncbi.nlm.nih.gov) and a nucleotide BLAST (Basic Local Alignment Search Tool) analysis was performed.

Results

The first test of the potential DNA barcode regions was the robustness of PCR amplification. The three loci were PCR amplified from forty-two DNA samples, using previously published primer sets for each locus and empirically derived optimization conditions (Kress et al. 2005, Newmaster et al. 2006, Taberlet et al. 2006, Kress and Erickson 2007, Lahaye et al. 2008). The length of the amplicons obtained from each locus were all <1000bp in length, satisfying the second criterion for an appropriate DNA barcode (see Tables 2, 3, and 4). PCR success, defined as the amplification and recovery of sufficient DNA to be analyzed by sequencing, was determined for each species at each locus (Figure 1). PCR amplification was most successful for the *trnH-psbA* and *rbcL* loci (overall success rates among all six species of 29/41 or 70% and 30/41 or 73%, respectively), compared to the more problematic *matK* locus (overall success rate among all six species of 16/41 or 39%). This low PCR success rate for *matK* was not unanticipated, as it has been reported by previous researchers (Kress and Erickson, 2007). Interestingly, we observed widespread PCR failure for almost all samples of staghorn sumac. Due to the nature of these experiments, internal controls allowed for an objective analysis of this outcome as PCR failure. For example, those samples which failed to amplify at one locus did yield quality PCR product at another locus, establishing that the quality of DNA extraction was not a source of experimental failure. Furthermore, the successful amplification of

each locus from many different species confirms that the proper amplification conditions were used (Figure 1).

For the *trnH-psbA* locus, doublets were obtained following PCR amplification. Attempts to optimize the PCR reaction conditions to achieve a single successful amplicon were unsuccessful. Futher alterations to the PCR protocol were not attempted, due to the nature of DNA barcodes; that is, the goal of PCR robustness includes the ability to PCR amplify a particular locus using universally established conditions. Therefore, we continued with these procedures, and instead simply purified the PCR products by gel extraction prior to DNA sequence analysis.

The resolution of the three barcode loci was investigated by bidirectional DNA sequencing of purified PCR amplicons. Sequencing success, defined as achievement of > 50bp of quality sequence data, was determined for each species and locus (using both forward and reverse primers) (Tables 2, 3 and 4). Staghorn sumac was eliminated from this analysis for both *trnH-psbA* and *matK*, due to the failure of PCR amplification at these loci. For the remaining species, sequencing with the forward primer was more successful for *trnH-psbA*, while reverse primers yielded better sequence data for *matK*, as well as *rbcL*. This suggests that bidirectional sequencing is advisable for these loci. In general, the most consistent, and best quality, sequence data was obtained at the *rbcL* locus, while sequencing reactions were most problematic at the *matK* locus.

To continue the assessment of *trnH-psbA*, *matK*, and *rbcL* resolution, intraspecific and interspecific sequence divergences were analyzed for each locus. Sequence data for all samples of each species were assembled into contigs for each locus (data not shown), and percent sequence divergence was calculated by the number of mismatches and gaps within an

overlapping region between samples. Average intraspecific percent sequence divergences are shown in Table 5 and Figure 2. The most variable intraspecific sequence divergence was observed for *matK*; however , since the greatest divergence was seen in staghorn sumac with only two samples analyzed, more samples would need to be screened to validate this observation.

To determine interspecific differences, sequence data from each locus were compared between all six species. Contig analysis for each locus for each species was used to choose representative sequences from each species for comparison (data not shown). In some cases, individual sequences were used; in others, small contigs were assembled in order to cover as much of the DNA region as possible for a more accurate comparison. Matrices displaying the interspecific percent sequence divergences are shown in Tables 6, 7, and 8. For the data that was available for *matK*, much higher interspecific vs. intraspecific sequence divergences were observed than for *rbcL* and *trnH-psbA*. The most complete data were obtained for the *rbcL* locus (Table 8). The greatest sequence differences were seen in non-congeneric (more distantly related) species (e.g., within the staghorn sumac, white ash, and tree of heaven group, as well as between these species and all maple species). However, in the comparison of the congeneric species, we observed very low interspecific vs. intraspecific differences, with the notable 100% sequence similarity between sugar and Norway maple at this locus.

The most intriguing result was obtained with the interspecific analysis of *trnH-psbA*. When a contig was assembled for this locus including all samples for which sequence data were available, it was noted that few sequences overlapped (Figure 3); therefore, a direct comparison between many of these species was not possible (Table 6). The lengths of sequence overlap between some samples were calculated by the Sequencher software as larger than what would be evident by the contig (Table 6). Other published reports have indicated great alignment

difficulties with *trnH-psbA* due to complex evolutionary events that resulted in high rates of insertions and deletions, as well as the presence of a pseudogene between the *trnH* and *psbA* regions. (Chase et al., 2007; Lahaye et al., 2008; Kress and Erickson, 2007). Further analyses of the sequences at this locus are necessary before a conclusion can be drawn regarding the accuracy of the interspecific sequence divergence values reported herein.

The last analysis performed was a test of the ability for each locus to correctly identify the plant species sequenced, using bioinformatics tools. In all BLAST analyses, the sequence submitted was identified as the correct locus. For sugar maple and red maple, sequence for all three loci correctly identified the species. For Norway maple, half of the samples were incorrectly identified as either red maple or sugar maple. There was no clear pattern as to which loci correctly identified this grouping of species better (Appendix 1). White ash and staghorn sumac were not correctly identified within the NCBI website, but the majority of samples for tree of heaven correctly identified this species using all three loci (Appendix 1).

Discussion

The power of a single genetic region to serve as DNA barcode for species within a kingdom has been confirmed for the mitochondrial COI gene in animals (Hebert et al. 2003a, Hebert et al. 2003b, Hebert et al. 2004). Recently, the Consortium for the Barcode of Life (CBOL) validated the choice of either *trnH-psbA*, *matK*, or *rbcL* as a suitable DNA barcode for land plants (CBOL Plant Working Group 2009). This conclusion was based on published reports of the resolution, robustness, and discriminating ability of these loci. While recognizing that neither of these regions fulfills all criteria for a suitable DNA barcode, the CBOL espoused the robustness of *rbcL* (Fazekas et al., 2008; Newmaster et al., 2006), and the resolution of *matK* and

trnH-psbA (Fazekas et al., 2008 – in CBOL paper; Lahaye et al., 2008), but acknowledged the lack of primer universality for *matK* leading to PCR amplification failure (Kress and Erickson, 2007; Fazekas et al., 2008), and the ambiguous sequences obtained from *trnH-psbA* (Chase et al., 2007; Lahaye et al., 2008; Kress and Erickson, 2007). The final recommendation of this group was the use of a two locus barcode of *matK* and *rbcL* (CBOL Plant Working Group, 2009).

In the studies reported herein, we sought to verify the use of the *trnH-psbA*, *matK*, and *rbcL* loci as DNA barcodes for four native and two invasive common tree species. As ecological niches are often similar, many invasive species directly compete with closely related or similar-looking native species (Villa and Weiner 2004), often reducing the growth and abundance of native species (Drake et al. 1989, Williamson 1996). The spread of invasive plant species can also result in reduced native species diversity and a decline in forest ecosystem stability and productivity (Chapin et al. 2000, Cock et al. 2003). Therefore, a strategy for rapid identification of key problematic invasive species is necessary. The BLAST analysis recognized all three of the loci as equally adequate for identification of the six species used here, including correct discrimination of the two invasive species from the four native species; however, further analysis is needed to test the utility of DNA barcoding in generally distinguishing invasive plants from native plants.

In the design of this study, we were able to assess intraspecific sequence divergence for the three loci tested. This is something that has not been previously reported within the literature. Due to our large replicate sample size (seven for each of our six species) we were able to test the utility of DNA barcodes across different individuals from the same species. Our results show that all three loci had less than 3% intraspecific sequence divergence for all six species. *trnH-psbA* provided the lowest intraspecific sequence divergence, with averages below

0.5%, whereas *matK* provided the largest intraspecific divergences, therefore suggesting it may not be an ideal candidate for a universal DNA barcode.

In our analysis, we found that *rbcL* displayed the robustness required for DNA barcodes, as of the three loci, *rbcL* was the only one for which we obtained sequence data for all six species. *rbcL* exhibited low intraspecific sequence divergence, and more variable interspecific sequence divergence, with the exception of 100% sequence similarity between congeneric sugar maple and Norway maple. *rbcL* has been reported to discriminate between 85% of congeneric species (Newmaster et al., 2006); therefore, this result was not unexpected. In contrast, the resolution and discriminating power of *matK* could not be accurately assessed in this study given the widespread problems with lack of primer universality. Likewise, the alignment difficulties with *trnH-psbA* made analysis of resolution at this locus impossible.

Given the rigorous standards by which DNA barcodes are measured, we agree with previous reports that *rbcL* has the potential to serve as a barcode for our chosen tree species. However, with its limitations of resolution, we also agree that it should be combined with a second locus, for a tiered approach. A multigene tiered approach to DNA barcoding has been suggested recently as a viable approach for plant identification (Newmaster et al. 2006). This strategy employs one universal locus from a more conserved coding region to serve as a first tier for establishing identity at the family or genus level. The second tiered analysis typically focuses on a more variable non-coding region, which would be specific to the species level. Despite the lack of robustness of *matK*, we conclude that this locus is more suitable than *trnH-psbA*, due to the general inability to align the latter sequences. With the design of more universal *matK* primers, the complement of *rbcL* and *matK* together could have wide utility as desirable plant DNA barcodes. However, even within our small sample size (six species), we found one species

(staghorn sumac) that was especially problematic with all of these loci. This therefore suggests that a single, universal plant barcode may not be achievable.

Our results provide a direct test of DNA barcoding in land plants. We were able to specifically test the ability of putative DNA barcodes to distinguish closely-related (congeneric) species vs. more distantly related species and to measure intraspecific sequence divergence by using replicated samples. From our research it appears that a single DNA barcode, at least of the three tested here, is not satisfactory to clearly and correctly identify land plant species. Therefore we suggest using a multigene tiered approach using *rbcL*, followed by *matK* for the most accurate analysis.

Acknowledgements

This research was supported by a grant from the Babson Faculty Research Fund, granted to Shari Laprise. We would like to thank Joanne Pratt, Sadie Aznavoorian-Cheshire, and Jodi Schaefer for equipment and support. We would also like to thank Babson undergraduate students Jeffrey Holoubek and Mark Paolucci for laboratory assistance.

References

Alvarez, I. and J.F. Wendel, *Ribosomal ITS sequences and plant phylogenetic inference*. Mol Phylogenet Evol , 2003. **29**(3): p. 417-434.

Chapin, F.S.I., et al., *Consequences of changing biodiversity.* Nature, 2000. **405**: p. 234-242.

Chase, M.W., et al., *Land plants and DNA barcodes: short-term and long-term goals.* Phil. Trans. R. Soc. Lond. B Biol. Sci., 2005. **360**: p. 1889-1895.

Chase, M.W., et al., *A proposal for a standardized protocol to barcode all landplants.* Taxon 2007. **56(**2): p. 295-299.

Cock, M.J., et al., *Biosecurity and forests: an introduction-with particular emphasis on forest pests*. FAO Forest Health and Biosecurity Working Paper FBS/2E. 2003.

CBOL Plant Working Group. 2009. *A DNA barcode for land plants*. PNAS 106: 12794-12797.

Drake, J.A., et al., *Biological invasions: a global perspective*. 1989: John Wiley & Sons.

Fazekas, A.J., et al., *Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well*. PLoS ONE, **3**: e2802.

Hebert, P.D.N., et al., *Biological identifications through DNA barcodes.* Proc. R. Soc. Lond. B, 2003a. **270**: p. 313-321.

Hebert, P.D.N., et al., *Barcoding animal life:cytochrome c oxidase subunit 1 divergences among closely related species.* Proc. R. Soc. Lond. B, 2003b. **(Suppl.) 270**: p. S96-S99.

Hebert, P.D.N., et al., *Identification of Birds through DNA Barcodes.* PLoS Biol, 2004. **2**(10): p. e312.

Hebert, P.D.N., et al., *Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly Astraptes fulgerator.* Proc. Nat. Acad. Sci. USA, 2004. **101**(41): p. 14812-14817.

Kress, W.J. and D.L. Erickson, *A two-locus global DNA barcode for land plants: The coding rbcL gene complements the non-coding trnH-psbA spacer region.* PLoS ONE, 2007. **2**(6): p. e508.

Kress, W.J., et al., *Use of DNA barcodes to identify flowering plants.* Proc. Nat. Acad. Sci. USA, 2005. **102**: p. 8369-8374.

Lahaye, R., et al., *DNA barcoding the floras of biodiversity hotspots.* Proc. Nat. Acad. Sci. USA, 2008. **108**(8): p. 2923-2928.

Matz, M.V. and R. Nielsen, *A likelihood ratio test for species membership based on DNA sequence data.* Phil. Trans. R. Soc. Lond. B. Biol. Sci., 2005. **360**: p. 1969-1974.

Meyer, C.P. and G. Paulay, *DNA barcoding: Error rates based on comprehensive sampling.* PLoS Biol., 2005. **3**(12): p. e422.

Newmaster, S.G., et al., *DNA barcoding in land plants: evaluation of rbcL in a multigene tiered approach.* Can. J. Bot., 2006. **84**: p. 335-341.

Pimentel, D., et al., *Environmental and economic costs of non-indigenous species in the United Stated.* BioScience, 2000. 50(1): 53-65.

Rapini, A., et al., *Phylogenetics of the New World Asclepiadeae (Apocynaceae)*. Taxon, **55**: p. 119-124.

Taberlet, P., et al., *Power and limitations of the chloroplast trnL (UAA) intron for plant DNA barcoding.* Nucleic Acids Res., 2006. **35**(3): p. e14.

Tautz, D., et al., *A plea for DNA taxonomy.* Trends Ecol. Evol., 2003. **18**(2): p. 70-74.

Van den Berg, C., et al., *A phylogenetic analysis of Laeliinae (Orchidaceae) based on sequence data from nuclear internal transcribed spacers (ITS) of ribosomal DNA*. Lindleyana, 2000. **15**: p. 96-114.

Vila, M. and J. Weiner, *Are invasive plant species better competitors than native plant species? - evidence from pair-wise experiments.* Oikos, 2004. **105**: p. 229-238.

Williamson, M., *Biological invasions*. 1996: Chapman and Hall.

Figures and Tables

Table 1: List of plant species used in this study, including common and scientific nomenclature.

Common species name	Scientific nomenclature	Invasive vs. Native	Classification
Staghorn sumac	*Rhus typhina*	Native	Non-congeneric, but morphologically similar
White ash	*Fraxinus Americana*	Native	
Tree of heaven	*Ailanthus altissima*	Invasive	
Red maple	*Acer rubrum*	Native	Congeneric, and morphologically similar
Sugar maple	*Acer saccharum*	Native	
Norway maple	*Acer platanoides*	Invasive	

Figure 1: Percent PCR success for each species at each DNA barcode locus. Seven samples from each species were tested, with the exception of white ash, for which only six samples were available.

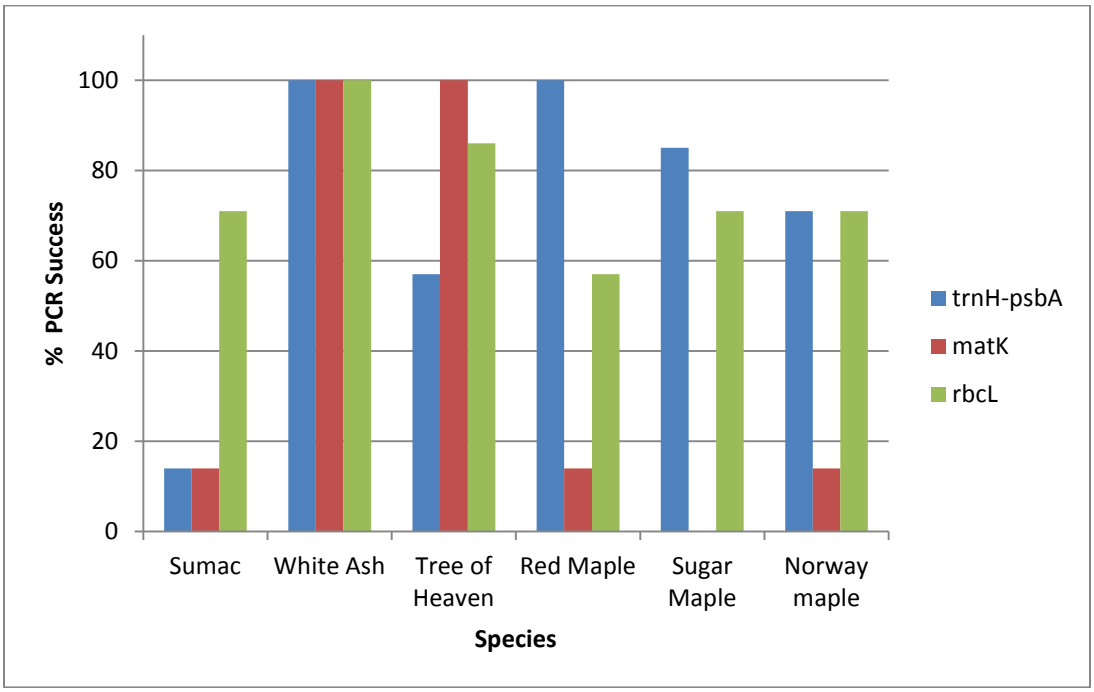


Table 2: Sequencing success at the *trnH-psbA* locus. (SE = standard error of the mean; ND = no data available; n/a = not applicable)

		Forward primer sequencing					Reverse primer sequencing				
Species	Amplicon length range (bp)	Average seq. length (bp)	Range (bp)	SE	# seq success/samples sequenced	% seq. success	Average seq. length (bp)	Range (bp)	SE	# seq success/samples sequenced	% seq. success
Staghorn sumac	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND
White ash	600	368	120-450	53.8	6/6	100	360	n/a	n/a	1/1	100
Tree of heaven	600-700	153	40-360	103	3/4	75	0	n/a	n/a	0/3	0
Red maple	450-600	278	180-420	29	7/7	100	255	90-420	117	2/2	100
Sugar maple	450-550	352	120-500	63	5/6	83	420	420-420	0	4/4	100
Norway maple	450-600	163	22	120-190	3/5	60	135	90-180	45	2/4	50
		Totals:			**24/28**	**86**	**Totals:**			**9/14**	**64**

Table 3: Sequencing success at the *matK* locus (SE = standard error of the mean; NS = sample not sequenced with this primer; n/a = not applicable)

		Forward primer sequencing					Reverse primer sequencing				
Species	Amplicon length range (bp)	Average seq. length (bp)	Range (bp)	SE	# seq success/samples sequenced	% seq. success	Average seq. length (bp)	Range (bp)	SE	# seq success/samples sequenced	% seq. success
Staghorn sumac	700	0	n/a	n/a	0/1	0	NS	NS	NS	NS	NS
White ash	900	810	780-840	30	2/6	33	420	120-660	141	4/6	66
Tree of heaven	900-950	540	n/a	n/a	1/6	17	227	50-570	76	6/7	86
Red maple	900	0	n/a	n/a	0/1	0	90	n/a	n/a	1/1	100
Sugar maple	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS
Norway maple	700	660	n/a	n/a	1/1	100	600	n/a	n/a	1/1	100
				Totals:	**4/15**	**27**			**Totals:**	**12/15**	**80**

Table 4: Sequencing success at the *rbcL* locus. (SE = standard error of the mean; n/a = not applicable)

		Forward primer sequencing					Reverse primer sequencing				
Species	Amplicon length range (bp)	Average seq. length (bp)	Range (bp)	SE	# seq success/samples sequenced	% seq. success	Average seq. length (bp)	Range (bp)	SE	# seq success/samples sequenced	% seq. success
Staghorn sumac	600-800	330	120-540	n/a	2/6	33	600	n/a	n/a	1/2	50
White ash	600-800	560	360-600	40	6/6	100	600	570-600	5	6/6	100
Tree of heaven	700	400	170-640	80	6/6	100	600	n/a	n/a	1/1	100
Red maple	700-800	410	300-615	70	4/4	100	540	480-600	60	2/2	100
Sugar maple	700-800	250	180-360	60	3/5	60	420	100-640	130	4/4	100
Norway maple	600-800	550	420-630	70	3/4	75	570	480-640	40	5/5	100
		Totals:			**21/31**	**68**		**Totals:**		**19/20**	**95**

Figure 2: Average intraspecific % sequence divergence for the three loci tested. Compare to values listed in Table 5, and note that for some species, the intraspecific % sequence divergence was 0, while no data was available for others (e.g., staghorn sumac).

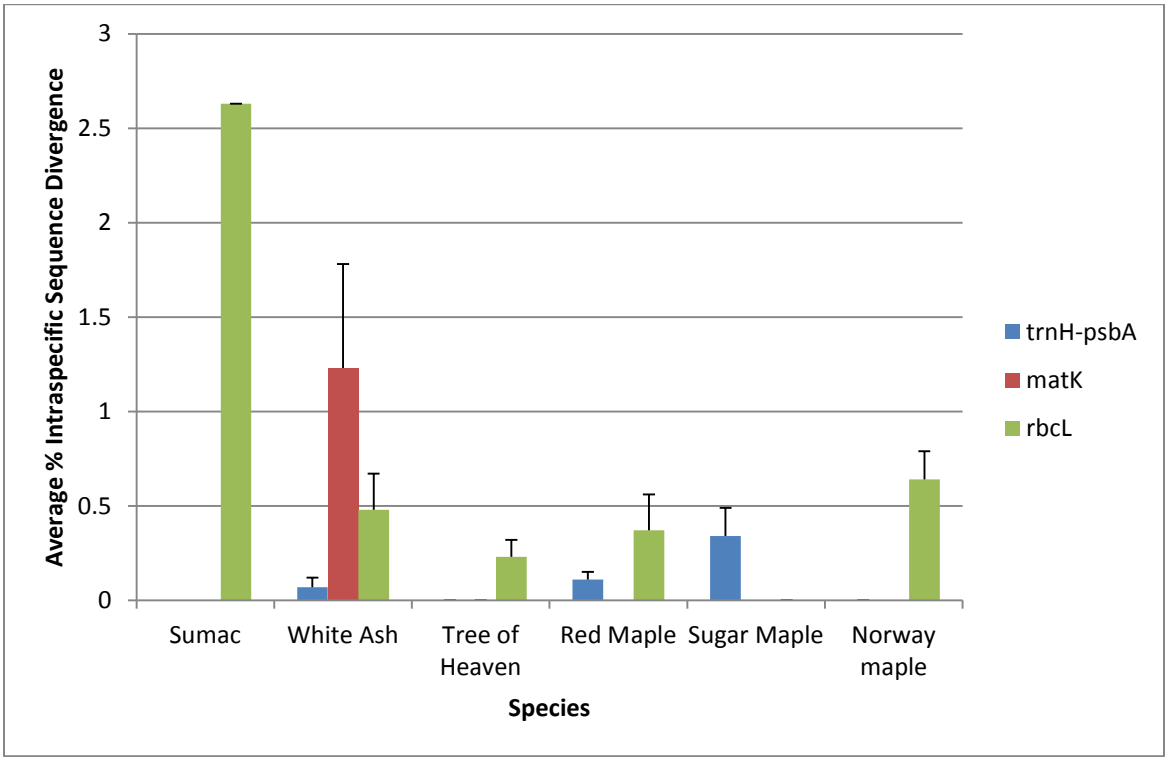


Table 5: Intraspecific sequence divergence for each barcode locus among six species. (n total = number of samples for which DNA was available for sequence analysis; range = low and high % sequence divergence values; SE = standard error of the mean for % sequence divergence; ND = no data available; n/a = not applicable)

	trnH-psbA			matK			rbcL		
Species	n	% seq. div (range)	SE	n	% seq. div (range)	SE	n	% seq. div (range)	SE
Staghorn sumac	0	ND	ND	0	ND	ND	2	2.63 (n/a)	n/a
White ash	6	0.07 (0-0.33)	0.05	4	1.23 (0-2.74)	0.55	6	0.48 (0-1.7)	0.19
Tree of heaven	2	0 (n/a)	n/a	6	0 (0-0)	0	6	0.23 (0-1.43)	0.09
Red maple	7	0.11 (0-0.56)	0.04	1	n/a	n/a	4	0.37 (0-0.85)	0.19
Sugar maple	5	0.34 (0-0.84)	0.15	0	ND	ND	5	0 (0-0)	0
Norway maple	3	0 (0-0)	0	1	n/a	n/a	5	0.64 (0-1.25)	0.15

Table 6: Matrix showing interspecific sequence divergence at the *trnH-psbA* locus. Full-length sequences or contigs for each species were compared. The overlap between the sequences compared and the % sequence divergence are shown for each species pair. Sequence divergences between species pairs without overlapping sequences could not be determined. Asterisks indicate data that appears to contradict the *trnH-psba* contig shown in Figure 3. *trnH-psbA* sequence data was not available for staghorn sumac(ND=no data).

	White ash		**Tree of heaven**		**Red maple**		**Sugar maple**		**Norway maple**	
	Overlap (bp)	% seq. div.	Overlap (bp)	% seq. div.	Overlap (bp)	% seq. div.	Overlap (bp)	% seq. div.	Overlap (bp)	% seq. div.
Staghorn sumac	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND
White ash			127*	30.71	173	17.35	190*	21.58	0	
Tree of heaven					0		0		0	
Red maple							429	9.56	370*	9.19
Sugar maple									0	0

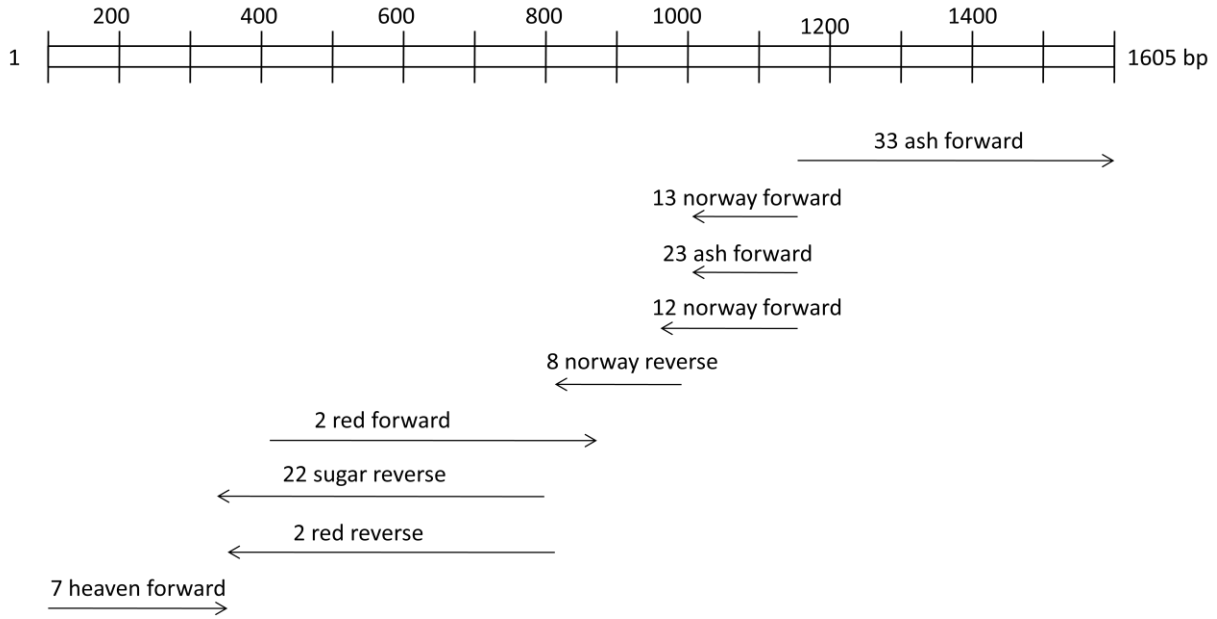
Table 7: Matrix showing interspecific sequence divergence at the *matK* locus. Full-length sequences or contigs for each species were compared. The overlap between the sequences compared and the % sequence divergence are shown for each species pair. *matK* sequence data was available only for white ash, tree of heaven, Norway maple, and red maple (ND=no data).

	White ash		**Tree of heaven**		**Red maple**		**Sugar maple**		**Norway maple**	
	Overlap (bp)	% seq. div.	Overlap (bp)	% seq. div.	Overlap (bp)	% seq. div.	Overlap (bp)	% seq. div.	Overlap (bp)	% seq. div.
Staghorn sumac	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND
White ash			746	19.58	93	25.81	ND	ND	749	16.83
Tree of heaven					89	15.74	ND	ND	744	12.1
Red maple							ND	ND	93	2.16
Sugar maple									ND	ND

Table 8: Matrix showing interspecific sequence divergence at the *rbcL* locus. Full-length sequences or contigs for each species were compared. The overlap between the sequences compared and the % sequence divergence are shown for each species pair.

	White ash		**Tree of heaven**		**Red maple**		**Sugar maple**		**Norway maple**	
	Overlap (bp)	% seq. div.	Overlap (bp)	% seq. div.	Overlap (bp)	% seq. div.	Overlap (bp)	% seq. div.	Overlap (bp)	% seq. div.
Staghorn sumac	599	8.02	596	4.7	600	4.5	600	4.67	600	4.67
White ash			644	8.33	647	8.81	599	8.69	613	8.81
Tree of Heaven					659	4.71	596	4.03	610	4.1
Red maple							600	0.84	614	0.82
Sugar maple									625	0

Figure 3: *trnH-psbA* contig. Representative sequences from five species were arranged into one contig of 1605bp to analyze interspecific sequence divergence. The lack of overlap between many individual samples at this locus does not allow for direct comparison of these sequences. Also note that sequences for samples 12 and 13 Norway maple, and sample 23 white ash are oriented in reverse direction. (arrows indicate the sequence data available for each sample; numbers above the arrows represent assigned sample identifiers, followed by common species name; forward or reverse indicates the primer used for sequencing).



Appendix 1: Identification of known species using nucleotide BLAST analysis. Forward/Reverse refers to the primer alignment and actual species names that matched with Blast identified species names are highlighted. All Blast loci were identified as the correct loci used. The E value provides the probability due to chance, that there is another species with a similarity greater than the identified species and percent similarity of our nucleotide sequence to the nucleotide sequence of the BLAST identified species is provided in parentheses.

Forward/ Reverse	Actual species	BLAST identified species	Actual and BLAST identified locus	E value (% Similarity)
Forward	*Acer platanoides* (Norway)	*Acer platanoides* (Norway)	rbcL	5e-19 (98%)
Reverse	*Acer platanoides* (Norway)	*Acer saccharum* (Sugar)	rbcL	0.0 (100%)
Forward	*Acer platanoides* (Norway)	*Acer platanoides* (Norway)	matK	0.0 (100%)
Reverse	*Acer platanoides* (Norway)	*Acer platanoides* (Norway)	matK	0.0 (100%)
Forward	*Acer platanoides* (Norway)	*Acer truncastum / campestre*	rbcL	0.0 (100%)
Reverse	*Acer platanoides* (Norway)	*Acer platanoides* (Norway)	rbcL	0.0 (100%)
Forward	*Acer platanoides* (Norway)	*Acer platanoides* (Norway)	trnh	1e-64 (96%)
Reverse	*Acer platanoides* (Norway)	*Acer platanoides* (Norway)	trnh	1e-13 (91%)
Forward	*Acer platanoides* (Norway)	*Acer saccharum* (Sugar)	rbcL	1e-118 (100%)
Reverse	*Acer platanoides* (Norway)	*Acer saccharum* (Sugar)	rbcL	4e-151 (100%)
Reverse	*Acer platanoides* (Norway)	*Acer platanoides* (Norway)	rbcL	4e-151 (100%)
Forward	*Acer platanoides* (Norway)	*Acer rubrum* (Red)	rbcL	0.0 (100%)
Reverse	*Acer platanoides* (Norway)	*Acer rubrum* (Red)	rbcL	0.0 (100%)
Forward	*Acer rubrum* (Red)	*Acer rubrum* (Red)	trnh	0.0 (99%)
Reverse	*Acer rubrum* (Red)	*Acer rubrum* (Red)	trnh	0.0 (99%)
Forward	*Acer rubrum* (Red)	*Acer rubrum* (Red)	rbcl	5e-167 (100%)
Reverse	*Acer rubrum* (Red)	*Acer rubrum* (Red)	matK	0.0 (100%)
Forward	*Acer rubrum* (Red)	*Acer rubrum* (Red)	trnh	3e-131 (100%)
Forward	*Acer rubrum* (Red)	*Acer rubrum* (Red)	rbcl	0.0 (100%)
Reverse	*Acer rubrum* (Red)	*Acer rubrum* (Red)	rbcl	0.0 (100%)
Forward	*Acer rubrum* (Red)	*Acer rubrum* (Red)	trnh	1e-95 (100%)
Forward	*Acer rubrum* (Red)	*Acer rubrum* (Red)	trnh	0.0 (100%)

Reverse	*Acer rubrum* (Red)	*Acer rubrum* (Red)	matK	7e-44 (91%)
Forward	*Acer rubrum* (Red)	*Acer rubrum* (Red)	rbcl	3e-131 (100%)
Forward	*Acer rubrum* (Red)	*Acer rubrum* (Red)	trnh	1e-95 (100%)
Forward	*Acer rubrum* (Red)	*Acer rubrum* (Red)	trnh	5e-60 (100%)
Forward	*Acer saccharum* (Sugar)	*Acer saccharum* (Sugar)	trnh	1e-57 (100%)
Forward	*Acer saccharum* (Sugar)	*Acer saccharum* (Sugar)	rbcL	1e-118 (100%)
Reverse	*Acer saccharum* (Sugar)	*Acer saccharum* (Sugar)	rbcL	0.0 (100%)
Forward	*Acer saccharum* (Sugar)	*Acer saccharum* (Sugar)	trnh	1e-95 (100%)
Reverse	*Acer saccharum* (Sugar)	*Acer saccharum* (Sugar)	rbcL	0.0 (100%)
Forward	*Acer saccharum* (Sugar)	*Acer saccharum* (Sugar)	trnh	1e-118 (100%)
Reverse	*Acer saccharum* (Sugar)	*Acer saccharum* (Sugar)	trnh	4e-151 (100%)
Forward	*Acer saccharum* (Sugar)	*Acer saccharum* (Sugar)	trnh	.004 (100%)
Forward	*Acer saccharum* (Sugar)	*Acer saccharum* (Sugar)	trnh	0.0 (100%)
Reverse	*Acer saccharum* (Sugar)	*Acer saccharum* (Sugar)	trnh	4e-151 (100%)
Forward	*Acer saccharum* (Sugar)	*Acer saccharum* (Sugar)	rbcL	1e-118 (100%)
Reverse	*Acer saccharum* (Sugar)	*Acer saccharum* (Sugar)	rbcL	6e-165 (98%)
Forward	*Acer saccharum* (Sugar)	*Acer saccharum* (Sugar)	trnh	4e-151 (100%)
Reverse	*Acer saccharum* (Sugar)	*Acer saccharum* (Sugar)	trnh	1e-118 (100%)
Reverse	*Acer saccharum* (Sugar)	*Acer saccharum* (Sugar)	rbcL	3e-11 (100%)
Forward	*Acer saccharum* (Sugar)	*Acer saccharum* (Sugar)	trnh	0.0 (100%)
Reverse	*Acer saccharum* (Sugar)	*Acer saccharum* (Sugar)	trnh	0.0 (100%)
Forward	*Acer saccharum* (Sugar)	*Acer saccharum* (Sugar)	rbcL	1e-95 (100%)
Forward	*Ailanthus altissima* (Heaven)	*Ailanthus altissima* (Heaven)	matK	1e-118 (100%)
Reverse	*Ailanthus altissima* (Heaven)	*Ailanthus altissima* (Heaven)	matK	0.0 (100%)
Reverse	*Ailanthus altissima* (Heaven)	*Ailanthus altissima* (Heaven)	matK	9e-22 (100%)
Forward	*Ailanthus altissima* (Heaven)	*Ailanthus altissima* (Heaven)	rbcL	0.0 (100%)
Reverse	*Ailanthus altissima* (Heaven)	*Ailanthus altissima* (Heaven)	matK	3e-86 (100%)

Forward	*Ailanthus altissima* (Heaven)	*Ailanthus altissima* (Heaven)	rbcL	7e-54 (100%)
Forward	*Ailanthus altissima* (Heaven)	*Ailanthus altissima* (Heaven)	rbcL	3e-86 (100%)
Forward	*Ailanthus altissima* (Heaven)	*Ailanthus altissima* (Heaven)	matK	0.0 (100%)
Reverse	*Ailanthus altissima* (Heaven)	*Acer rubrum*	matK	4e-151 (100%)
Reverse	*Ailanthus altissima* (Heaven)	*Ailanthus altissima* (Heaven)	matK	1e-95 (100%)
Forward	*Ailanthus altissima* (Heaven)	*Ailanthus altissima* (Heaven)	rbcL	0.0 (100%)
Reverse	*Ailanthus altissima* (Heaven)	*Ailanthus altissima* (Heaven)	matK	1e-95 (100%)
Forward	*Ailanthus altissima* (Heaven)	*Ailanthus altissima* (Heaven)	rbcL	5e-167 (100%)
Reverse	*Fraxinus americana* (Ash)	*Fraxinus excelsior*	matK	3e-86 (100%)
Forward	*Fraxinus americana* (Ash)	*Fraxinus excelsior / augustifolia*	rbcL	0.0 (99%)
Reverse	*Fraxinus americana* (Ash)	*Fraxinus excelsior*	rbcL	0.0 (99%)
Forward	*Fraxinus americana* (Ash)	*Olea europaea*	trnh	9e-141 (97%)
Forward	*Fraxinus americana* (Ash)	*Fraxinus excelsior / augustifolia*	rbcL	0.0 (99%)
Reverse	*Fraxinus americana* (Ash)	*Fraxinus excelsior / augustifolia*	rbcL	0.0 (99%)
Forward	*Fraxinus americana* (Ash)	*Fraxinus excelsior / augustifolia*	rbcL	0.0 (99%)
Reverse	*Fraxinus americana* (Ash)	*Fraxinus excelsior / augustifolia*	rbcL	4e-178 (98%)
Forward	*Fraxinus americana* (Ash)	*Fraxinus excelsior / augustifolia*	rbcL	0.0 (99%)
Reverse	*Fraxinus americana* (Ash)	*Fraxinus excelsior / augustifolia*	rbcL	0.0 (99%)
Forward	*Fraxinus americana* (Ash)	*Fraxinus excelsior / augustifolia*	matK	0.0 (99%)
Reverse	*Fraxinus americana* (Ash)	*Fraxinus excelsior*	matK	0.0 (99%)
Forward	*Fraxinus americana* (Ash)	*Fraxinus excelsior*	matK	0.0 (98%)
Reverse	*Fraxinus americana* (Ash)	*Osmanthus suavis*	matK	0.0 (98%)
Forward	*Fraxinus americana* (Ash)	*Fraxinus excelsior / chinensis*	rbcL	0.0 (94%)
Reverse	*Fraxinus americana* (Ash)	*Fraxinus excelsior / ornus*	rbcL	0.0 (99%)
Forward	*Fraxinus americana* (Ash)	*Fraxinus excelsior*	trnh	2e-65 (96%)
Reverse	*Fraxinus americana* (Ash)	*Fraxinus excelsior*	trnh	2e-149 (100%)
Forward	*Rhus typhina* (Sumac)	*Rhus queinzii / transvaalensis*	rbcL	0.0 (99%)

Reverse	*Rhus typhina* (Sumac)	*Rhus queinzii / lucida*	rbcL	0.0 (99%)