Validation of nrDNA ITS as a DNA barcode for *Marshallia* (Asteraceae)

Curtis J. Hansen* and Leslie R. Goertzen

Department of Biological Sciences and Auburn University Museum of Natural History, Auburn University, Auburn, AL 36849, U.S.A.
* Correspondence: hansecj@auburn.edu

Abstract

We examine the utility of nrDNA ITS as a barcode marker in the southeastern endemic wildflower genus *Marshallia*, sequencing multiple individuals of all species. With the exception of the putative hybrid *M. mohrii*, all species had a unique ITS sequence and intraspecific variation was very low. Intragenomic polymorphism was also quantified using high-throughput sequencing reads. The ITS data support specific status for the recently described *M. legrandii* and indicate a relationship between *M. trinervia* and the federally listed *M. mohrii*.

Introduction

DNA barcoding uses standardized genetic markers for the identification and discrimination of individual species. In animals, the mitochondrial cytochrome oxidase 1 (CO1) gene has been a widely used barcode but finding a plant equivalent has proven challenging (Hollingsworth *et al.* 2011). The internal transcribed spacer (ITS) region of nuclear ribosomal DNA (nrDNA) has been proposed as a DNA barcode in numerous organisms including plants (Chase *et al.* 2007; Lahaye *et al.* 2008; Chen *et al.* 2010) and appears to show specific utility within the family Asteraceae (Gao *et al.* 2010; Hollingsworth 2011; Timme *et al.* 2007).

Recently, Schilling & Floden (2013) surveyed native and endemic genera from the tribes Helenieae and Polymnieae as part of a larger systematic ITS barcoding effort for Asteraceae in Tennessee (Schilling 2013; Schilling & Floden 2012, 2014; Schilling *et al.* 2014). Their data suggest that differences in tempo and mode of diversification among lineages of Asteraceae in the southeast are reflected in diverse levels of interspecific ITS variation. Within some genera ITS sequences enabled clear discrimination among species but for others there was insufficient or no variation. Their study included three species of *Marshallia* Schreb. (Asteraceae: Helenieae), a genus largely endemic to the southeast United States. In that study, each *Marshallia* species as well as two additional genbank accessions had unique ITS sequences. Here we extend their analysis to all species and subspecific taxa in the genus to determine if the nrDNA ITS region can distinguish the most closely related lineages. We also extend the barcode validation process to examine intraspecific and intragenomic variation for this marker by sequencing multiple individuals of each species with high-throughput methods.

Materials and Methods

Study System

Marshallia, commonly known as Barbara's buttons, consists of eight species (sensu Watson 2006) of herbaceous perennials ranging from southwestern Missouri and central Texas east to Florida and north into southwestern Pennsylvania. *Marshallia* are often recognized by a pink or white 'puffball' inflorescence formed by unusually elongate and deeply lobed disc flowers (Fig. 1). Individual species can be difficult to identify because of continuous, occasionally overlapping variation in certain morphological characters. Species are sporadically distributed within their respective ranges and often associated with unique habitats and substrates including rocky soils (both sandstone and limestone derived), open sandy pine forests and also wet habitats such as streamsides, shoals or bogs.

Three of the species are fairly common (*M. caespitosa*, *M. graminifolia* and *M. obovata*) while four are uncommon or rare (*M. grandiflora*, *M. ramosa* and *M. trinervia*), including the federally listed threatened species *M. mohrii* (U.S. Fish and Wildlife Service 1988). An especially tall,

Table 1. *Marshallia* specimen and ITS sequence information used in this study including taxon, population identifier, GenBank accession number and voucher information including the Index Herbariorum acronym for herbarium of deposition.

Species	Pop.	GenBank	Voucher	Herbarium
Marshallia Schreb.				
M. caespitosa Nutt. ex. DC.	–	AF229262	*Panero 7426*, Travis Co., TX	TEX
var. *caespitosa*	M26	KP715512	*Watson 12-01*, Pottawatamie Co., OK	AUA
var. *signata* Beadle & F.E. Boynton	M28	KP715513	*Watson 12-03*, Parker Co., TX	AUA
M. graminifolia (Walt.) Small	M1	KP715498	*Hansen 4951*, Covington Co., AL	AUA
	M39	KP715501	*Hansen 5813*, Jackson Co., MS	AUA
	M5	KP715499	Seed from NC Bot. Garden, no voucher	–
M. grandiflora Beadle & F.E. Boynton	M17	KP715506	Plant nursery, no voucher	–
	M30	KP715507	*Hansen 5645*, Upshur Co., WV	AUA
	3303	KF607075	*Floden s.n.*, garden grown, TN	TENN
M. legrandii Weakley	M31	KP715514	*Hansen 5646*, Granville, Co., NC	AUA
M. mohrii Beadle & F.E. Boynton	M21	KP715510	*Hansen 5056*, Bibb Co., AL	AUA
	M24	KP715511	*Hansen 5564*, Cherokee Co., AL	AUA
M. obovata (Walt.) Beadle & F.E. Boynton	M3	KP715503	*Hansen 4956*, Macon Co., AL	AUA
	M32	KP715505	*Hansen 5647*, Durham Co., NC	AUA
	M4	KP715504	Seed from NC Bot. Garden, no voucher	–
	3306	KF607076	*Rothberger s.n.*, Polk Co., TN	TENN
	–	AF229261	*Baldwin s.n.*, Durham Co., NC	UC
M. ramosa Beadle & F.E. Boynton	M19	KP715508	*Hansen 5054*, Ben Hill Co., GA	AUA
	M38	KP715509	*Hansen 5795*, Washington Co., FL	AUA
M. trinervia (Walt.) Trel.	M2	KP715500	*Hansen 4954*, Lee Co., AL	AUA
	M33	KP715502	*Hansen 5785*, Pearl River Co., MS	AUA
	3067	KF607077	*Horn 2006-7*, Lawrence Co., TN	TENN

Figure 1. *Marshallia obovata* capitulum, Bullock County, Alabama. Photo: C.J. Hansen.

extremely rare, unbranched species with morphological affinities to *M. grandiflora* and *M. obovata* var. *obovata*, from unique mafic barren habitats of North Carolina and southern Virginia was recently named a new species (*M. legrandii*, Weakley & Poindexter 2012) and is included here. Most taxa are diploid [2n=18] but two polyploid entities have been described (*M. caespitosa* var. *caespitosa* and *M. mohrii*, Watson & Estes 1990). See Table 1 for complete list of taxonomic authorities.

Taxonomic Sampling

A total of 17 individuals were sampled to represent all major lineages of *Marshallia* based on previous taxonomic studies. Where multiple individuals of a species were included they were sampled from diverse localities across the range of that particular species (Table 1).

DNA Sequencing and Assembly

DNA was extracted from fresh leaf material of greenhouse-maintained, wild-collected plants using either the modified CTAB protocol of Doyle & Doyle (1987) or an E.Z.N.A.® Plant DNA Kit (Omega Bio-tek, Inc., Norcross, GA). Samples were submitted to the Genomic Services Lab, HudsonAlpha Institute for Biotechnology (Huntsville, AL) where paired-end libraries were prepared and sequencing (100 bp) was performed on an Illumina Hi-Seq 2000 platform. De novo assembly of each set of reads was performed using Ray (Boisvert *et al.* 2010) with a kmer length of 31. Large contigs containing the 35S rDNA were identified by BLAST (Zhang *et al.* 2000) and the ITS/5.8S region

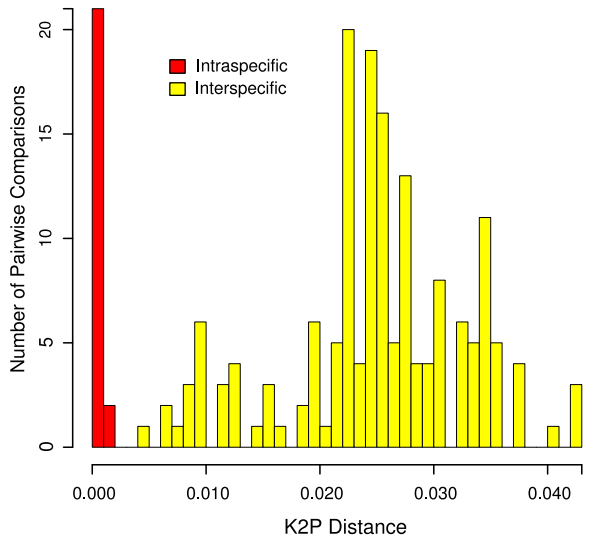

Figure 2. Pairwise K2P distances between all *Marshallia* ITS sequences excepting *M. mohrii*. Intraspecific comparisons shown in red, interspecific in yellow.

was extracted for each accession. Reads for each accession were filtered for quality (min Q = 33, minlength = 100) and then re-mapped at an average depth > 400× onto the corresponding ITS reference with Bowtie2 (Langmead & Salzberg 2012). The resulting map was examined visually in Tablet (Milne *et al.* 2013) and a consensus sequence was generated with Samtools (Li *et al.* 2009). In two cases a complete ITS region did not assemble in Ray and a reference sequence from another species was used in the first round of mapping. A consensus sequence was then obtained from Samtools and reads mapped again onto this result.

Nucleotide frequencies for all mapped reads at each position were extracted from Samtools mpileup output with a script (courtesy of Dr. Eric Archer, NOAA Fisheries) written for the R statistical environment (R Core Team 2013). For each sequence, polymorphic sites were represented by IUPAC ambiguity codes if variant nucleotides occurred at a frequency equal to or greater than 5% of total coverage at that position. The 17 newly obtained ITS sequences, along with five existing *Marshallia* ITS sequences from GenBank (Table 1), were visually aligned using Seaview (Galtier *et al.* 1996). Phylogenetic analyses were conducted in PAUP* v4.10b (Swofford 2003). Parsimony was chosen as an optimality criterion and a heuristic search with TBR branch swapping and 10 000 random sequence addition replicates was performed. Trees were visualized in FigTree (Rambaut 2010).

Results

Marshallia ITS sequences ranged in length from 640–645 bp, with 45–46% GC content. Individual accessions differed from zero up to 26 positions (4%) across this range, with an average of 12.6 differences (2%) among all pairwise comparisons. Within species, individuals differed by at most one nucleotide and had on average 0.1 differences for all within-species comparisons. There was an average of 14.6 nucleotide differences (2.3%) for pairwise comparisons between different species, excluding *M. mohrii* which had ITS sequences identical to *M. trinervia*.

Barcode gap analysis (Meyer & Paulay 2005; Lahaye *et al.* 2008) based on Kimura 2-parameter corrected pairwise distances revealed a small but distinct break delineating intraspecific and interspecific comparisons (Fig. 2). The relative level of within and among-species sequence variation can also be seen in the parsimony tree shown in Fig. 4. Observed pairwise distances are found in Table 2.

Intragenomic polymorphism was observed for all accessions, ranging from three sites (0.5%) in *Marshallia legrandii* (M31) to 27 sites (4.2%) in *M. caespitosa* var. *caespitosa* (M26). On average 9.6 positions or 1.5% of the ITS were polymorphic across all individuals. Haplotype variation was concentrated in the ITS1 and ITS2 regions with some positions polymorphic in more than one individual and species (Fig. 3).

Discussion

Among the 22 *Marshallia* ITS sequences examined in this study there were sufficient nucleotide differences to clearly differentiate most species. In contrast, intraspecific variation was low, tending towards near identity even for individuals separated by significant geographic distance. For example, samples of *M. graminifolia* ranging from North Carolina to Louisiana and *M. obovata* from North

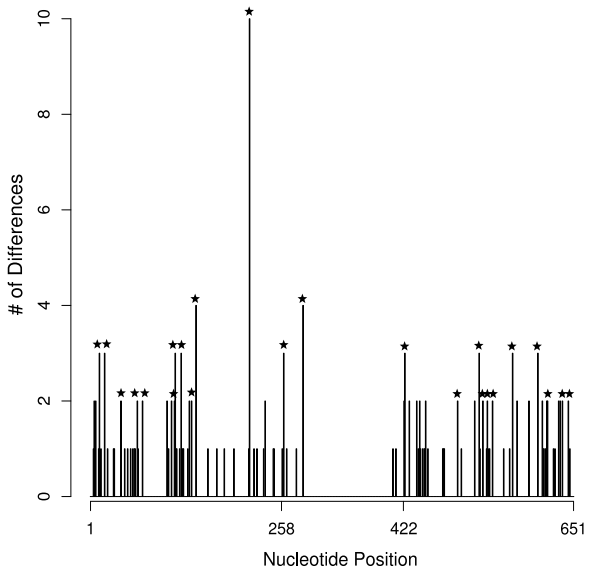

Figure 3. Number of observed nucleotide differences per position of the aligned ITS matrix for all *Marshallia*. Positions 258 and 422 define the 5.8S region. Stars indicate sites where multiple species vary in sequence.

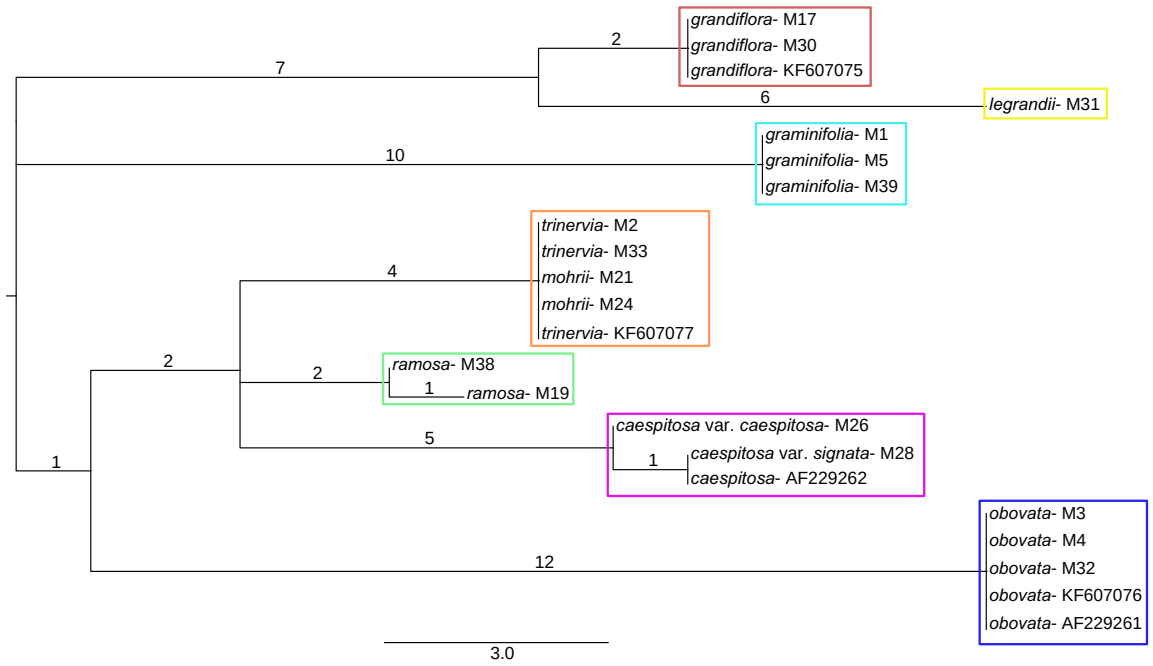

Figure 4. One of 39 equally parsimonious trees of length 53 from analysis of *Marshallia* ITS sequences. The tree is midpoint rooted and branch lengths are indicated.

Carolina to Alabama were identical in their respective nuclear ITS sequence. Sampling within *M. graminifolia* and *M. caespitosa* included representatives of named sub-specific taxa that were indistinguishable using ITS sequence data.

The high-throughput, genome-skimming approach (Straub *et al.* 2012) used here provided essentially complete resolution of intragenomic polymorphism in nrDNA repeats. The ITS haplotype phase was readily apparent from the consistent pattern of base frequencies at polymorphic sites as well as from the occurrence of multiple polymorphisms on individual or paired reads. Owing to the high sequence coverage, the relative abundance of haplotypes could be established with precision. These data suggest that for most individuals, 2–3 major haplotypes possibly representing 2–3 separate rDNA loci account for almost all of the observed polymorphism. There were minor single nucleotide variants of each but in general the level of homogenization was high.

Despite the polymorphism, haplotypes coalesced at the species level suggesting a sufficient amount of inter-locus concerted evolution over time or a common ancestor within each species. For example the significant intragenomic polymorphism identified in *M. caespitosa* var. *caespitosa*, sample M26 (27 polymorphic sites), can be traced to an unusually divergent paralog that is possibly a result of the genome duplication in this polyploid subspecies. For all its variation however, this haplotype does not differ at any position that distinguishes *M. caespitosa* overall from other species and it groups with other *M. caespitosa* sequences when it is included independently in phylogenetic analyses (data not shown).

Overall, the results provide additional justification for the use of ITS sequences as DNA barcode markers in *Marshallia*. The ITS sequences are readily amplified from diverse plant tissues and possess sufficient variation to differentiate most species in this genus. Individual plants will likely be polymorphic at a handful of sites but this variation does not involve indels and does not obscure differences among species. However, the ITS marker alone is insufficient to distinguish between lower ranked intraspecific taxa in this group.

The identical ITS sequences obtained from *M. trinervia* and *M. mohrii* are an interesting result in the context of the latter species' putative hybrid origin and uncertainty regarding its parental species. Watson *et al.* (1991) suggested an allopolyploid origin for *M. mohrii* and hypothesized that the parents were *M. trinervia* and *M. grandiflora* or possibly an unknown diploid species. The ITS data presented here are consistent with this scenario, given subsequent biased gene conversion in the direction of one parent, *M. trinervia*, as is commonly described in even early hybrid lineages (Álvarez & Wendel 2003).

The ITS data also provide clear support for the recognition of the recently described *M. legrandii* from North Carolina and Virginia (Weakley & Poindexter 2012). The single ITS sequence we obtained from this critically rare species clustered with those from *M. grandiflora* which it resembles closely. Eight nucleotide differences separated *M. legrandii* from any *M. grandiflora* ITS sequence and although this is slightly lower than the average interspecific pairwise distance among *Marshallia* reported here, it is greater than the pairwise distance values noted for several other pairs of easily distinguished species.

Table 2. Observed pairwise distances between *Marshallia* ITS sequences included in this study.

Individual	M1	M5	M39	M2	M33	KF77	M3	M4	M32	KF76	AF61	M17	M30	KF75	M19	M38	M21	M24	M26	M28	AF62	M31
graminifolia-M1	-																					
graminifolia-M5	0	-																				
graminifolia-M39	0	0	-																			
trinervia-M2	16	17	15	-																		
trinervia-M33	16	17	15	0	-																	
trinervia-KF607077	16	17	15	0	0	-																
obovata-M3	21	22	20	16	16	16	-															
obovata-M4	22	23	21	15	15	15	0	-														
obovata-M32	19	20	18	14	14	14	0	0	-													
obovata-KF607076	22	23	21	16	16	16	0	0	0	-												
obovata-AF229261	22	23	21	16	16	16	0	0	0	0	-											
grandiflora-M17	15	16	14	15	15	15	21	21	20	21	21	-										
grandiflora-M30	15	17	14	14	14	14	21	21	19	21	21	0	-									
grandiflora-KF607075	16	17	15	15	15	15	21	21	19	21	21	0	0	-								
ramosa-M19	14	15	13	7	7	7	17	16	15	17	17	14	12	14	-							
ramosa-M38	14	15	13	6	6	6	16	15	14	16	16	14	12	14	1	-						
mohrii-M21	10	11	9	0	0	0	8	8	6	8	8	10	9	10	3	2	-					
mohrii-M24	16	17	15	0	0	0	16	15	14	16	16	15	14	15	7	6	0	-				
caesp. caespitosa-M26	11	12	10	5	5	5	14	13	12	14	14	12	11	12	4	3	1	5	-			
caesp. signata-M28	13	14	12	6	6	6	16	15	14	16	16	14	13	14	5	4	2	6	0	-		
caespitosa-AF229262	18	19	17	10	10	10	20	19	18	20	20	18	17	18	9	8	6	10	1	0	-	
legrandii-M31	17	17	17	19	19	19	25	26	23	26	26	8	8	8	18	18	14	19	15	18	22	-

Acknowledgments

We greatly appreciate the assistance of Lindsey Garner, Andrew Brane, Nathan Hall, Steve Biggerstaff, Eldon and Susan Larsen, Alan Weakley, and Linda Watson with field work. We also thank Anthony Melton and Pamela Brannock for technical assistance and discussion. We are grateful for funding from the Alabama Wildflower Society and the U.S. Fish and Wildlife Service (Award F12AP00728 to LRG) for this research. We gratefully acknowledge Ed Schilling and Alan Weakley for very thorough peer review of this paper and Derick Poindexter for additional helpful comments. This is contribution no. 706 of the Auburn University Museum of Natural History.

References

Álvarez I, Wendel J (2003) Ribosomal ITS sequences and plant phylogenetic inference. *Molecular Phylogenetics and Evolution*, **29**, 417–434.

Boisvert S, Laviolette F, Corbeil J (2010) Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *Journal of Computational Biology*, **17**, 1519–1533.

Chase M, Cowan R, Hollingsworth P, *et al.* (2007) A proposal for a standardised protocol to barcode all land plants. *Taxon*, **56**, 295–299.

Chen S, Yao H, Han J, *et al.* (2010) Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. *PLoS ONE*, **5**, e8613.

Doyle J, Doyle J (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin*, **19**, 11–15.

Galtier N, Gouy M, Gautier C (1996) SEAVIEW and PHYLO_WIN, two graphic tools for sequence alignment and molecular phylogeny. *Bioinformatics*, **12**, 543–548.

Gao T, Yao H, Song J, Zhu Y, Liu C, Chen S (2010) Evaluating the feasibility of using candidate DNA barcodes in discriminating species of the large Asteraceae family. *BMC Evolutionary Biology*, **10**, 324.

Hollingsworth PM (2011) Refining the DNA barcode for land plants. *Proceedings of the National Academy of Sciences*, **108**, 19451–19452.

Hollingsworth PM, Graham SW, Little DP (2011) Choosing and using a plant DNA barcode. *PLoS ONE*, **6**, e19254.

Lahaye R, van der Bank M, Bogarin D, *et al.* (2008) DNA barcoding the floras of biodiversity hotspots. *Proceedings of the National Academy of Sciences*, **105**, 2923–2928.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**, 357–359.

Li H, Handsaker B, Wysoker A, *et al.* (2009) The sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Meyer C, Paulay G (2005) DNA barcoding: error rates based on comprehensive sampling. *PLoS Biol*, **3**, e422.

Milne I, Stephen G, Bayer M, *et al.* (2013) Using Tablet for visual exploration of second-generation sequencing data. *Bioinformatics*, **14**, 193–202.

R Core Team (2013) R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria.

Rambaut A (2010) FigTree. Ver.1.3.1. Institute of Evolutionary Biology, University of Edinburgh.

Schilling EE (2013) Barcoding the Asteraceae of Tennessee, tribes Cardueae and Vernonieae. *Phytoneuron*, **2013-19**, 1–7.

Schilling EE, Floden A (2012) Barcoding the Asteraceae of Tennessee, tribes Gnaphalieae and Inuleae. *Phytoneuron*, **2012-99**, 1–6.

Schilling EE, Floden A (2013) Barcoding the Asteraceae of Tennessee, tribes Helenieae and Polymnieae. *Phytoneuron*, **2013-81**, 1–6.

Schilling EE, Floden A (2014) Barcoding the Asteraceae of Tennessee, tribe Senecioneae. *Phytoneuron*, **2014-34**, 1–5.

Schilling EE, Mattson N, Floden A (2014) Barcoding the Asteraceae of Tennessee, tribe Coreopsideae. *Phytoneuron*, **2014-101**, 1–6.

Straub S, Parks M, Weitemier K, Fishbein M, Cronn R, Liston A (2012) Navigating the tip of the genomic iceberg: next generation sequencing for plant systematics. *American Journal of Botany*, **99**, 349–364.

Swofford DL (2003) PAUP*. Phylogenetic analysis using parsimony (*and other methods). Ver 4.0b10. Sinauer Associates, Sunderland Massachusetts.

Timme RE, Simpson BB, Linder CR (2007) High-resolution phylogeny for *Helianthus* (Asteraceae) using the 18S–26S ribosomal DNA external transcribed spacer. *American Journal of Botany*, **94**, 1837–1852.

US Fish and Wildlife Service (1988) Endangered and threatened wildlife and plants; determination of *Marshallia mohrii* (Mohr's Barbara's buttons) to be a threatened species. *Federal Register*, **53**, 34698–34701.

Watson LE (2006) *Marshallia*. In *Flora of North America North of Mexico* (edited by Flora of North America Editorial Committee), vol. 21, pp. 456–458. Oxford University Press, New York and Oxford.

Watson LE, Elisens WJ, Estes JR (1991) Electrophoretic and cytogenetic evidence for allopolyploid origin of *Marshallia mohrii* (Asteraceae). *American Journal of Botany*, **78**, 408–416.

Watson LE, Estes JR (1990) Biosystematic and phenetic analysis of *Marshallia* (Asteraceae). *Systematic Botany*, **15**, 403–414.

Weakley AS, Poindexter DB (2012) A new species of *Marshallia* (Asteraceae, Helenieae, Marshalliinae) from mafic woodlands and barrens of North Carolina and Virginia. *Phytoneuron*, **2012-105**, 1–17.

Zhang Z, Schwartz S, Wagner L, Miller W (2000) A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology*, **7**, 203–214.

Citation: Hansen CJ, Goertzen LR (2014) Validation of nrDNA ITS as a DNA barcode for *Marshallia* (Asteraceae). *Paysonia* **3**, 5–10.

Keywords: Asteraceae, *Marshallia*, DNA barcoding, internal transcribed spacer, ribosomal DNA.

Peer Review: Edward Schilling, University of Tennessee and Alan Weakley, University of North Carolina.