Applications in Plant Sciences

Applications in Plant Sciences 2013 1(7): 1200002

Primer Note

Single-copy nuclear gene primers for *Streptanthus* and other Brassicaceae from genomic scans, published data, and ESTs[1]

N. Ivalú Cacho[2,3] and Sharon Y. Strauss[2]

[2]Department of Evolution and Ecology, University of California, Davis, One Shields Avenue, Davis, California 95616-5294 USA

- *Premise of the study:* We report 11 primer sets for nine single-copy nuclear genes in *Streptanthus* and other Thelypodieae (Brassicaceae) and their utility at tribal-level and species-level phylogenetics in this poorly resolved group.
- *Methods and Results:* We selected regions based on a cross-referenced matrix of previous studies and public *Brassica* expressed sequence tags. To design primers, we used alignments of low-depth-coverage Illumina sequencing of genomic DNA for two species of *Brassica* mapped onto *Arabidopsis thaliana*. We report several primer combinations for five regions that consistently amplified a single band and yielded high-quality sequences for at least 70% of the species assayed, and for four additional regions whose utility might be clade specific.
- *Conclusions:* Our primers will be useful in improving resolution at shallow depths across the Thelypodieae, and likely in other Brassicaceae.

Key words: Brassicaceae; Illumina; rapid diversification; single-copy nuclear gene; species-level phylogeny; Thelypodieae.

Despite the great importance of members of the Brassicaceae in agriculture and the extensive genomic resources available for *Arabidopsis thaliana* (L.) Heynh., our knowledge of phylogenetic relationships within the family is still murky in many clades (Franzke et al., 2011). Among the hurdles to elucidating phylogenetic relationships within this family are extensive gene duplication and polyploidization, and past and recent hybridization (Franzke et al., 2011). Main lineages have been identified using a variety of regions (e.g., ITS, *nad4*, *ndhF*, *phyA*, *Adh*, *chs*, *matK*, *rbcL*, and *trnLF*). However, relationships at shallower levels (e.g., within some tribes or at the species level) are often characterized by poor resolution. New genomic tools have much to contribute to our understanding of evolution in Brassicaceae, but to date, technological, analytical, and logistical limitations have slowed down the wide-scale applicability of genomic approaches for phylogenetics (Egan et al., 2012). Thus, phylogenetic studies at the species level or of rapidly diversified groups still rely widely on single marker primer development.

We developed a strategy to identify and design primers for single-copy nuclear genes (SCNGs) focusing on *Streptanthus* Nutt. and other members of the tribe Thelypodieae whose phylogenetic relationships and circumscription have remained a challenge (Warwick et al., 2010). Our strategy combines Illumina reads from genomic scans (low-depth-coverage sequencing of total genomic DNA) and public expressed sequence tags (ESTs) from *Brassica* L., a close relative to the Thelypodieae, with results from previous studies that identified putative SCNGs at wider taxonomic scales using algorithmic methods. We report several primer combinations that might be of utility for informing relationships within and across groups in the Thelypodieae.

METHODS AND RESULTS

Our approach to identify SCNGs is outlined in Fig. 1. We cross-referenced the results of three previous studies that used algorithms to identify putative SCNGs with published ESTs as follows: for the APVO loci (file 1471-2148-10-61-S1.xls from Duarte et al., 2010), we kept only those loci reported to have introns and be SCNGs in *A. thaliana*; for the COSII set (file available at: http://solgenomics.net/documents/markers/cosii.xls), we included all that were single-copy in *A. thaliana*; for the PPR genes (file NPH_2739_sm_TableS1.xls from Yuan et al., 2009), we kept those unique in rice and *Arabidopsis*; and we included all ESTs between *B. napus* and *A. thaliana* (Ilut and Doyle, 2012) after removing duplicates. Our final matrix contained 10,817 loci (APVO, 5381; COSII, 2869; PPR, 90; ESTs, 2477). The vast majority of loci (5596; 69.86%) were represented by a single source, 25% (2025) were represented by two sources, 5% (385) by three, and only 0.05% (4) were present in all four sources. We selected loci for primer design at random and verified that the following four criteria were met for each locus (if not, we picked another locus): (1) it was identified as SCNG by multiple sources in the matrix above; (2) it was represented by a single gene model in the *A. thaliana* genome (Tair10); (3) it contained an estimated length range between 600–1200 bp to allow for assembly from a single pass of Sanger sequencing; and (4) it possessed 40–60% intron content to maximize potential phylogenetic utility at species-level relationships (Rodríguez et al., 2009). In addition, we chose loci to span all five *A. thaliana* linkage groups.

We selected 15 loci for primer design. We designed primers based on alignments of genomic scans generated from low-coverage Illumina sequencing of total genomic DNA (Illumina GAIIx [Illumina Inc., San Diego, California, USA],

[1]Manuscript received 13 November 2012; revision accepted 5 February 2013.

We appreciate Luca Comai's generosity in facilitating the *Brassica* BAM files and providing laboratory space. Support for this study comes from the National Science Foundation (DEB 0919559 to S.Y.S.), Plant Genome Program award DBI 0733857 "Functional genomics of plant polyploids" (L. Comai), and a Consejo Nacional de Ciencia y Tecnología (CONACyT) fellowship to N.I.C. (EPSCI no. 187083).

[3]Author for correspondence: ivalu.cacho@gmail.com

doi:10.3732/apps.1200002

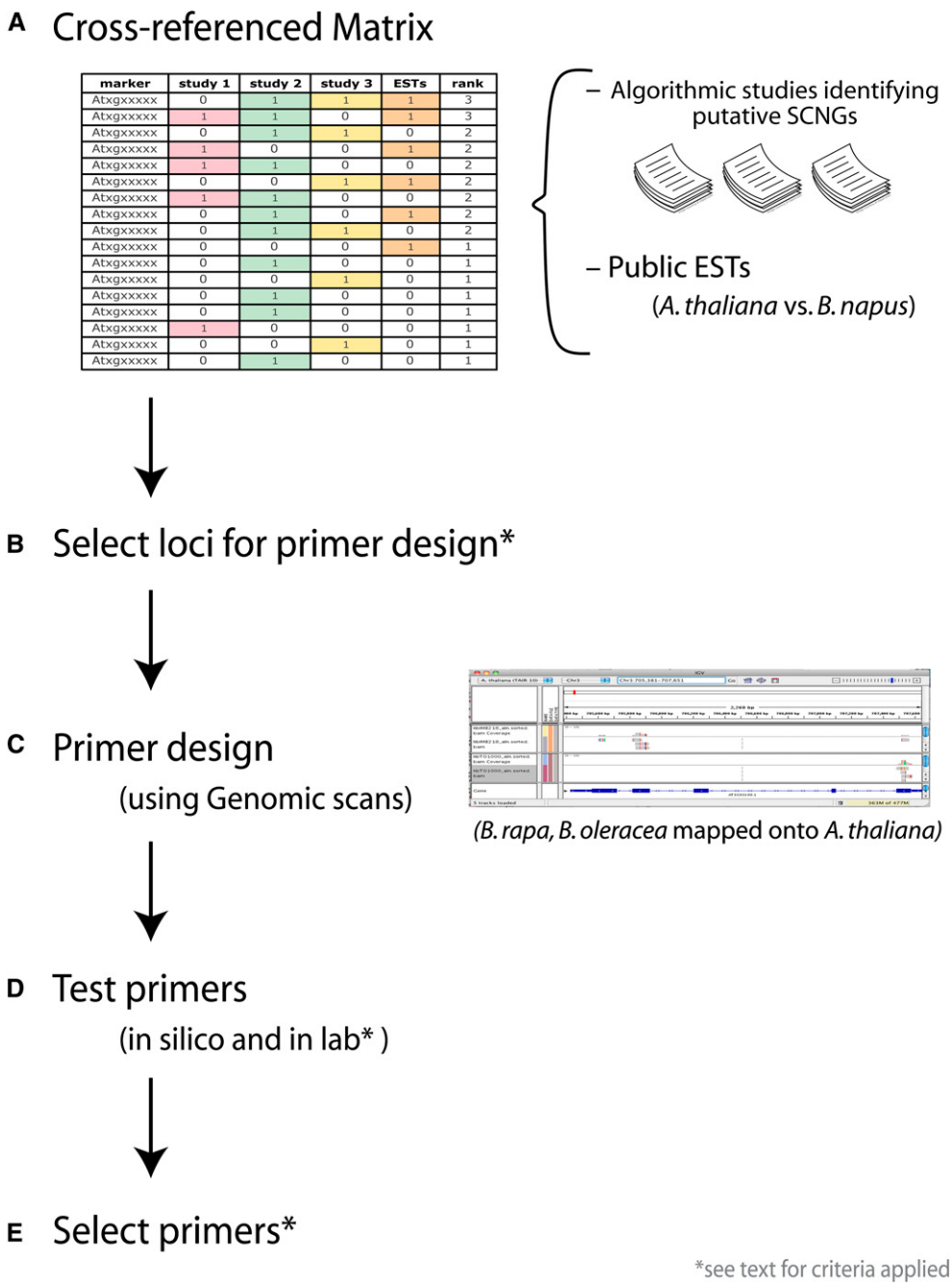

Fig. 1. Approach used to identify SCNGs in this study. (A) Genes that were identified as putative SCNGs across different algorithmic studies were cross-referenced with public ESTs. (B) Loci for primer design were selected according to criteria outlined in the text. (C) Primer design was based on alignments of reads from shallow-depth Illumina sequencing of genomic DNA of *Brassica rapa* and *B. oleracea* (2× and 9× coverage, respectively) mapped onto the *Arabidopsis thaliana* genome. (D) Primers were tested in silico and in the laboratory before final selection for sequencing of products (E).

80 bp reads) of *B. rapa* L. (2× coverage) and *B. oleracea* L. (9× coverage; L. Comai, unpublished data) mapped onto *A. thaliana* using the Burrows-Wheeler Alignment tool (BWA) (Li and Durbin, 2009) and visualized in IGV version 1.5 (Thorvaldsdottir et al., 2013). We located the selected regions based on their locus ID and followed standard primer design guidelines, aiming for primers with a length of 22–25 bp, 40–60% GC content, melting temperature (T_m) = 55–62°C, the presence of a 3′ GC clamp, and without repeats, runs, or secondary structures such as hairpins, dimers, and cross-dimers. Prior to testing in the laboratory, we tested primer performance in silico using Amplify 3× version 3.1.4 (http://engels.genetics.wisc.edu/amplify/). We designed 250 primer combinations for the 15 selected regions, and chose 52 to test in the laboratory.

Between one and five primer pairs for each of 15 selected regions were tested for single band amplification in a set of taxa spanning several genera in the Thelypodieae. Here, we report statistics on primer combinations that consistently yielded a single band and whose product generated a clean sequence in at least 70% of taxa tested (five loci), as well as sequences for a few primer sets that could be of utility with additional optimization or in a different subset of taxa (Tables 1 and 2).

***Laboratory*—**Genomic DNA was extracted from tissue dried in silica gel using either the cetyltrimethylammonium bromide (CTAB) method (Doyle and Doyle, 1987) or the DNeasy Plant Mini Kit (QIAGEN, Valencia, California,

Table 1. Primer regions that amplify a single band and yield clean sequences (first five) and others that might be useful on a clade-by-clade basis.

Locus ID[a]	Primer set	Primer sequences (5′–3′)	Length (bp)	Cycle[b]	Max no. of bands
AT1G56590	702F/1535R	F: AARGAYAATTTCATCATTGTCTATGAG R: GCATCCATYTCTTCAACAAGGTCG	900	PCR55	1
AT1G61620	F2/R2	F: GCAAAGACACTCGAAGAACA R: AGGTTTGTCACACACAAGACTT	1500	PCR60	1
AT2G40600	F2/R2	F: TCAAATGAGACGACAATGGTT R: CACCTCCTTTGCTTTGTTTAC	1300	TDPCR56	1
AT4G34700	F1/R1	F: GAAGTTCAATGTCAACCAAGATG R: ACCATGAGCAATCAGTTTGT	710	ANNEXT63	1
AT5G25630	F2/R2	F: RAAGAAGGTGGAGGAGGCGT R: ACATCTGCCTTCACSTTACACTC	450	TDPCR56	1
AT5G27620[c]	F1/R1	F: GAAGTTCAATGTCAACCAAGATG R: ACCATGAGCAATCAGTTTGT	1500	TDPCR54	2
AT3G03100[c]	F3/R3	F: CCTACCAGATGGGAACCTCT R: GTACCGAGTCCAGTTCTTTTG	1500	TDPCR54	2
AT3G03100[c]	F4/R3	F: CATAACATAGGAGCGACACTT R: GTACCGAGTCCAGTTCTTTTG	950	TDPCR54	1
AT1G50020	F3/R1	F: GTGTGGCTCTCCTGTATCGTTT R: CGCCGAGAGTTAAGACTATCAAT	950	TDPCR54	1
AT5G26680	5F/9R	F: AAGAGACAGGAACTGGCTAAACG R: TGCAAAGTGCAGCACATTGC	600	PCR60	1
AT5G26680	12F/13R	F: ACTGCTCTAAAGCTTATTCGCCAG R: AAAGTTTCGAGCTTCATTATATGG	450	PCR60	1

[a] Locus ID from The Arabidopsis Information Resource database (http://www.arabidopsis.org/).

[b] PCR conditions are as follows: PCR55, PCR60: 94°C, 2:00; (94°C, 0:30; 55°C or 60°C, 1:10; 72°C, 2:00) 35×; final extension 72°C, 7:00. ANNEXT63: 94°C, 2:00; (94°C, 0:30; 63°C, 4:00) 30–34×; final extension 72°C, 7:00. TDPCR54: 94°C, 1:00; (94°C, 0:30; 58°C, 1:10; 72°C, 1:30) 1×; (94°C, 0:30; 56°C, 1:10; 72°C, 1:30) 1×; (94°C, 0:30; 54°C, 1:10; 72°C, 1:30) 32×; final extension 72°C, 7:00. TDPCR56: 94°C, 1:00; (94°C, 0:30; 58°C, 1:10; 72°C, 1:30) 1×; (94°C, 0:30; 56°C, 1:10; 72°C, 1:30) 33×; final extension 72°C, 7:00.

[c] Some clade variation observed; additional optimization might be required in some cases.

USA). PCR reactions consisted of 5 µL of 5× Green GoTaq Reaction Buffer (M791A; Promega Corporation, Madison Wisconsin, USA), 0.5 µL dNTP mix (10 mM each), 0.5 µL of each primer (10 µM), and 0.2 µL (5 units/µL) of GoTaq (M3001; Promega Corporation) in a total volume of 25 µL. Cycling conditions are presented in Table 2. Bidirectional sequencing was performed at Beckman Coulter Genomics (Danvers, Massachusetts, USA). When more than one band amplified, we isolated bands, reamplified, and sequenced directly. If cloning was necessary, PCR products were gel-purified (QIAquick Gel Extraction Kit, QIAGEN), ligated into pGEM T-Vector (Promega Corporation), cloned into *E. coli* DHB-5α–competent cells (Invitrogen, Carlsbad, California, USA), reamplified (eight colonies per PCR product), and sequenced.

Table 2. Summary of parsimony-informative characters for those regions for which we obtained sequence data (due to financial limitations we could only sequence a reduced number of amplicons). For those taxa where cloning (see Appendix 1) was necessary, the allele that yielded the shortest tree was selected.

Region	# Tips	# Chars	CTE	# Var	% Var	nPIC	PIC	% PIC
AT2G40600	10	1569	1178	391	24.9	229	162	10.3
AT5G25630	10	296	259	37	12.5	23	14	4.7
AT4G34700	10	722	509	213	29.5	180	33	4.6
AT1G56590	11	891	646	245	27.5	199	46	5.2
AT1G61620	9	1530	1331	199	13.0	146	53	3.5
ITS[a]	11	681	587	94	13.8	69	25	3.7

Notes: # Tips = number of tips or accessions; # Chars = total number of characters; CTE = number of constant characters; # Var = number of variable characters; % Var = percentage of total characters that are variable; nPIC = number of nonparsimony informative characters; PIC = number of characters that are parsimony informative; % PIC = percentage of total characters that are parsimony informative.

[a] ITS is included as a reference.

Sequences were assembled and edited in Sequencher version 4.7 (Gene Codes Corporation, Ann Arbor, Michigan, USA). Potential PCR recombinants, assessed by manual examination of the sequences, were excluded. Alignment was performed manually in MacClade version 4.08 (Maddison and Maddison, 2002), and proportion of informative characters calculated in PAUP* version 4.0b10 (Swofford, 2002).

We have corroborated the utility of the SCNGs reported here by using a subset to estimate phylogenies of the "Streptanthoid" complex and its allies, a group that has been subject to several substantial taxonomic revisions and whose phylogenetic relationships have remained poorly understood. While these results are beyond the scope of this paper and will be reported separately (Cacho et al., in prep.), given the level of phylogenetically informative variation that we observe (Table 2; Appendix 1) we have confidence that the SCNGs we contribute here will be useful to infer species-level phylogenies in several clades of the Thelypodieae and potentially of the Brassicaceae as a whole. These improved phylogenies could be an important stepping stone to facilitate comparative evolutionary studies in these clades until technological advances allow straightforward implementation of new sequencing technologies for low-cost phylogenetic studies.

CONCLUSIONS

Using a strategy that combines results from previous algorithmic studies identifying putative SCNGs with genomic resources from published ESTs and Illumina genomic scans, we have identified and designed primers for several SCNGs that are of phylogenetic utility. Our primers yield sequences that are informative for phylogenies at and above the species level in most species of Thelypodieae and Sisymbreae we tested, including when possible two or more species of *Streptanthus*, *Streptanthella* Rydb., *Caulanthus* S. Watson, *Guillenia* Greene, *Stanleya* Nutt., *Sisymbrium* L., *Thelypodium* Endl., and *Thysanocarpus* Hook. Given that we designed primers based on *Arabidopsis* and *Brassica* sequences, they are also likely to be useful for understanding

relationships among members of Camelineae, and potentially across Brassicaceae as a whole.

LITERATURE CITED

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

Duarte, J. M., P. K. Wall, P. P. Edger, L. L. Landherr, H. Ma, J. C. Pires, J. Leebens-Mack, and C. W. dePamphilis. 2010. Identification of shared single copy nuclear genes in *Arabidopsis, Populus, Vitis* and *Oryza* and their phylogenetic utility across various taxonomic levels. *BMC Evolutionary Biology* 10: 61.

Egan, J. S., J. Schlueter, and D. M. Spooner. 2012. Applications of next-generation sequencing in plant biology. *American Journal of Botany* 99: 175–185.

Franzke, A., M. A. Lysak, I. A. Al-Shehbaz, M. A. Koch, and K. Mummenhoff. 2011. Cabbage family affairs: The evolutionary history of Brassicaceae. *Trends in Plant Science* 16: 108–116.

Ilut, D. C., and J. J. Doyle. 2012. Selecting nuclear sequences for fine detail molecular phylogenetic studies in plants: A computational approach and sequence repository. *Systematic Botany* 37: 7–14.

Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* 25: 1754–1760.

Maddison, D., and W. Maddison. 2002. MacClade 4.05: Analysis of phylogeny and character evolution. Sinauer, Sunderland, Massachusetts, USA.

Rodriguez, F., F. Wu, C. Ané, S. D. Tanksley, and D. M. Spooner. 2009. Do potatoes and tomatoes have a single evolutionary history, and what proportion of the genome supports this history? *BMC Evolutionary Biology* 9: 191.

Swofford, D. 2002. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Sinauer, Sunderland, Massachusetts, USA.

Thorvaldsdottir, H., J. T. Robinson, and J. P. Mesirov. 2013. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Briefings in Bioinformatics* 14:178–192.

Warwick, S. I., K. Mummenhoff, C. A. Sauder, M. A. Koch, and I. A. Al-Shehbaz. 2010. Closing the gaps: Phylogenetic relationships in the Brassicaceae based on DNA sequence data of nuclear ribosomal ITS region. *Plant Systematics and Evolution* 285: 209–232.

Yuan, Y. W., C. Liu, H. E. Marx, and R. G. Olmstead. 2009. The pentatricopeptide repeat (PPR) gene family, a tremendous resource for plant phylogenetic studies. *New Phytologist* 182: 272–283.

Appendix 1. GenBank accession and population information for the specimens used in this study. All vouchers are deposited at the University of California, Davis herbarium (DAV).

Taxon	Collection and locality information	GenBank accession no.[a,b]
Caulanthus inflatus S. Watson	NIC-S-066, Ballinger Canyon, CA, USA	AT4G34700: NIC-S-066, KC517428 (b1); AT2G40600: NIC-S-066, KC517426; AT1G56590: NIC-S-066, KC517439; AT1G61620: NIC-S-066, KC517461 (b1a1); AT5G25630: NIC-S-066, KC517408; ITS: NIC-S-066, KC517450
Streptanthus breweri A. Gray	KBR-020A, Knoxville-Berryessa Rd. at Hwy. 128, CA, USA	AT4G34700: KBR-020A, KC517429; AT2G40600: KBR-020A, KC517419; AT1G56590: KBR-020A, KC517440; AT1G61620: KBR-020A, KC517462; AT5G25630: KBR-020A, KC517409; ITS: KBR-020A, KC517451
Streptanthus polygaloides A. Gray	WAR-045, Washington Rd., CA, USA; SKY-082, Skyway Ave., CA, USA	AT4G34700: SKY-082, KC517437; AT2G40600: WAR-045, KC517422; AT1G56590: WAR-045, KC517441; AT1G61620: WAR-045, KC517463; AT5G25630: WAR-045, KC517410; ITS: WAR-045, KC517452
Streptanthella longirostris (S. Watson) Rydb.	NIC-S-020, Coyote Canyon, CA, USA	AT4G34700: NIC-S-020, KC517430; AT2G40600: NIC-S-020, KC517427; AT1G56590: NIC-S-020, KC517442; AT1G61620: NIC-S-020, KC517468; AT5G25630: NIC-S-020, KC517411; ITS: NIC-S-020, KC517453
Sisymbrium irio L.	NIC-S-022, Coyote Canyon, CA, USA; LLSP-004, Lleida, Spain	AT4G34700: LLSP-004, KC517436; AT2G40600: NIC-S-022, KC517424; AT1G56590: LLSP-004, KC517449; AT1G61620: na; AT5G25630: NIC-S-022, KC517412; ITS: NIC-S-022, KC517454
Caulanthus coulteri S. Watson	NIC-S-001, Caliente-Bodfish Rd., CA, USA	AT4G34700: NIC-S-001, KC517431 (b2.a1); AT2G40600: NIC-S-001, KC517425; AT1G56590: NIC-S-001, KC517443; AT1G61620: NIC-S-001, KC517464; AT5G25630: NIC-S-001, KC517413; ITS: NIC-S-001, KC517455
Stanleya pinnata (Pursh) Britton	NIC-S-054, trail off Hwy. 10, CA, USA	AT4G34700: NIC-S-054, KC517432; AT2G40600: na; AT1G56590: NIC-S-054, KC517445; AT1G61620: NIC-S-054, KC517469; AT5G25630: KC517415; ITS: NIC-S-054, KC517457
Streptanthus diversifolius S. Watson	NIC-S-085, Table Mountain, CA, USA	AT4G34700: NIC-S-085, KC517433 (b2.a2); AT2G40600: NIC-S-085, KC517421; AT1G56590: NIC-S-085, KC517446; AT1G61620: NIC-S-085, KC517465; AT5G25630: NIC-S-085, KC517416; ITS: NIC-S-085, KC517458
Streptanthus drepanoides Kruckeb. & J. L. Morrison	LSAD-027A, Lime Saddle, CA, USA	AT4G34700: LSAD-027A, KC517434; AT2G40600: LSAD-027A, KC517420; AT1G56590: LSAD-027A, KC517447; AT1G61620: LSAD-027A, KC517466; AT5G25630: LSAD-027A na; ITS: LSAD-027A, KC517459

Appendix 1. Continued.

Taxon	Collection and locality information	GenBank accession no.[a,b]
Caulanthus hallii Payson	NIC-S-015, trail off Hwy. 8, CA, USA; NIC-S-285, trail off Hwy. 78, CA, USA	AT4G34700: NIC-S-015, KC517435; AT2G40600: NIC-S-015, KC517418; AT1G56590: NIC-S-285, KC517448; AT1G61620: NIC-S-015, KC517467; AT5G25630: NIC-S-015, KC517417; ITS: NIC-S-015, KC517460
Caulanthus cooperi (S. Watson) Payson	NIC-S-055, trail off Hwy. 10, CA, USA; NIC-S-007, trail off Hwy. 78, CA, USA	AT4G34700: NIC-S-007, KC517438; AT2G40600: NIC-S-055, KC517423; AT1G56590: NIC-S-055, KC517444; AT1G61620: na; AT5G25630: NIC-S-055, KC517414; ITS: NIC-S-055, KC517456

Notes: NIC = N. Ivalú Cacho; SYS = Sharon Y. Strauss.

[a] The few cases in which two bands were amplified are noted as is the allele the sequence corresponds to.

[b] Individuals grown from seed are given codes according to their populations.