Appalachian
STATE UNIVERSITY®
BOONE, NORTH CAROLINA

Eleven Diverse Nuclear-Encoded Phylogenetic Markers For The Subfamily Panicoideae (Poaceae)

Authors

Matt C. Estep, Dilys M. Vela Diaz, Jinshun Zhong And Elizabeth A. Kellogg

Abstract

Premise of the study: Polyploidy is common in the grasses and low-copy nuclear loci are needed to further our understanding of phylogenetic relationships.

• Methods and Results: Genetic and genomic resources were combined to identify loci known to influence plant and inflorescence architecture. Degenerate primers were designed and tested to amplify regions of 11 nuclear-encoded loci across the panicoid grasses.

• Conclusions: The primers designed in this study amplify regions of a diverse set of genes within the panicoid grasses. Properly employed, these markers will allow the identification of allopolyploid taxa and their diploid progenitors.

Subfamily Panicoideae Link (Poaceae) is a diverse group of mainly tropical grasses that includes ~3300 species, grouped into ~200 genera (Grass Phylogeny Working Group, 2001). This subfamily of the grasses contains crop species (*Zea mays*, *Sorghum bicolor*, *Setaria italica*), biofuel stocks (*Panicum virgatum* and *Saccharum* sp.), agricultural weeds (*Setaria viridis*, *Rottboellia cochinchinesis*, and *Sorghum halapense*), and dominant species in tropical grasslands around the world (*Andropogon* sp., *Hyparrhenia* sp., and *Schizachyrium* sp.). Polyploidy and reticulate evolution are common in the grasses, requiring data from low-copy nuclear loci to resolve relationships. Although the internal transcribed spacer of the ribosomal genes (ITS) is widely used, its high copy number and concerted evolution make it inadequate for this purpose (Álvarez and Wendel, 2003).

Over the past decade, extensive genetic and genomic resources have been developed to study agriculturally important grasses. Genome sequences are available for several crop species in the panicoids, including maize (*Zea mays*), sorghum (*Sorghum bicolor*), and foxtail millet (*Setaria italica*). These genome sequences can be used as a reference for primer design using databases such as Phytozome and the National Center for Biotechnology Information (NCBI) databases. In addition to genome sequences, the genetics community has conducted extensive studies to identify genes that play functional roles in many agriculturally and evolutionarily important traits. Many of these loci influence plant and inflorescence architecture, the same traits used by botanists to define groups and species. Combining these resources allowed us to identify and design primers for 11 new single to low-copy nuclear loci for use within the panicoid grasses.

METHODS AND RESULTS

We identified genes from the literature that are known to influence grass morphology and inflorescence architecture, resulting in a large number of candidate loci. This list was reduced to ~20 gene candidates by removing genes that are part of large gene families. Published sequences (usually cDNA) from these studies were then used to query a database of translated nucleotide sequences from plant genome projects using Phytozome 8.0 (http://www.phytozome. net). This allowed the identification of homologous sequences in the three panicoid grass taxa for which genomes are available (*Setaria*, *Sorghum*, and *Zea*). Sequences were downloaded and used to query the high-throughput genomic sequences (HTGS) database on NCBI for the maize genome to aid in identifying exon/intron boundaries. All sequences were then translated and aligned using MUSCLE, as implemented in Geneious Pro 5.5.6 (BioMatters, Auckland, New Zealand). Alignments were then trimmed to exclude ambiguously aligned regions and to avoid large introns. Portions of coding sequence of ~800–1200 bp in the reference alignment were identified for primer design. The web-based software Primaclade was used to develop multiple primer pairs for each locus (Gadberry et al., 2005).

Each primer pair was then tested on a diverse group of 10 taxa (Appendix 1). PCR was carried out using primer-specific conditions in 25-µL reactions containing 100 ng (1 µL) of template DNA using GoTaq Flexi DNA polymerase kits (Promega Corporation, Madison, Wisconsin, USA). Reactions included 5× green flexi buffer (5 µL), 25 mM MgCl₂ (2 µL), 2.5 mM dNTPs (2 µL), 10 µM of each primer (1 µL), and 1 unit of *Taq* (0.2 µL). Five molar betaine (2.5 µL) and dimethyl sulfoxide (DMSO; 1.25 µL) were added to reactions when primer sequences contained a GC content above 50%. Reactions were performed at an annealing temperature equal to the melting temperature (T_m) − 5°C or using a touchdown protocol. Thermocycler conditions for standard reactions were: initial denature at 96°C for 5 min, followed by 32 cycles of denature at 96°C for 1 min, annealing at T_m − 5°C for 1 min, elongation at 72°C for 1 min, and a final elongation step at 72°C for 5 min. Thermocycler conditions for touchdown reactions were similar except three rounds of five cycles were used starting with an annealing temperature of 68°C and reducing it each round by 3°C; these 15 cycles were then followed by 25 cycles at an annealing temperature of 58°C. Twelve primer pairs produced a single band for all 10 taxa upon gel electrophoresis (Table 1). These bands were purified using a QIAGEN gel extraction kit, following the manufacturer's protocol (Valencia, California, USA). Purified products were cloned using pGEM-T Easy Vector and transformed into JM109 High-Efficiency Competent Cells, following manufacturer's protocols (Promega

TABLE 1. Primer sequences and PCR conditions for 11 nuclear-encoded loci.

Gene name	Abbreviation	Primer sequences (5′–3′)	T_a (°C)	GC rich
*Aberrant Panicle Organization*1	*apo*1	F: TCTACTCGTCGCCGTTCCT	58	Yes
		R: AGGAACCTCCTCATCGGC		
*Dwarf*8	*d*8	F: ATGAAGCGCGAGTACCAAGAC	TD 68–55	Yes
		R: GAAGTCGACGCGGATGGT		
		F2seq: CWTGCTGTCCGAGSTCAAC	58	No
		R2seq: AGTTCTCCTGCTGCACGG		
Erect Panicle	*ep*2–*ex*7	F: GAGCAGACTCCAAGTTCTGGT	58	No
		R: GGATGATTGCCTGCTCAAAT		
	*ep*2–*ex*8	F: TCAATCGAGGAGGTCACAATC	58	No
		R: TGAGGCAGTGGAAGCAGATA		
*Floral Organ Regulator*1	*for*1	F: ACMTGCASGCGCTGCTGAG	58	Yes
		R: GTTGTGGGASAGGTCCAGGTA		
*Liguleless*1	*lg*1	F: ATCTSCTGGATGAGTTCGAYGA	58	No
		R: AAGTCGAGATCRAACATMGCCTG		
Monoculm	*moc*1	F: CTGGCGTACAACCAGATCG	58	Yes
		R: GTCTGCCAGCCGAGGAAG		
*Ramosa*1	*ra*1	F: CGCCRCAGRTAAGGTCGTC	52	No
		R: GCCCAGTCTAAGCTGAAGATCC		
*Ramosa*1 *Enhancer Locus*2	*rel*2	F: ATGTCKTCKCTKAGCAGGGA	58	Yes
		R: RAGCATGATATTCCGRGCT		
*Ramosa*2	*ra*2	F: ATGGCRTCSYCGTCGAGCACC	57	No
		R: CATGCTGCTGTCTCCYCCTTCC		
*Retarded Palea*1	*rep*1	F1: CACTTCTTCYCCRGCCAYGSCC	TD 57–52	No
		F2: CAAYTCYGARACNCTGGAGGC		
		R: GTCSARGAAATCRYCRYCKTC		
*Vanishing Tassel*2	*vt*2	F: TACATCGAGCTGGTGTGCTC	58	No
		R: AKATGTCGTAGGCRTCGTCC		

Note: T_a = annealing temperature.

Corporation). Eight positive clones for each PCR product were sequenced in both directions using universal primers (M13, Sp6, or T7) on an ABI3730 DNA sequencer (Applied Biosystems, Carlsbad, California, USA). Chromatograms were trimmed of vector and resulting sequences were translated and aligned with the reference sequences using MUSCLE, as implemented in Geneious Pro 5.5.6 (BioMatters). The resulting sequences were compared using a maximum likelihood tree, and an exemplar was chosen for each distinct clade to represent a species or paralogues within a species. Regions not suitable for phylogenetic analysis (introns, ambiguously aligned regions, and primer annealing sites) were annotated on a consensus sequence and excluded during character analysis (Fig. 1). Known protein domains were also identified using Pfam databases and annotated on the consensus sequence (Punta et al., 2012). Statistics describing each alignment were generated using parsimony in PAUP* 4.0 (Swofford, 2002). An uncorrected distance matrix was calculated to identify the percent variation between different sequences (Table 2).
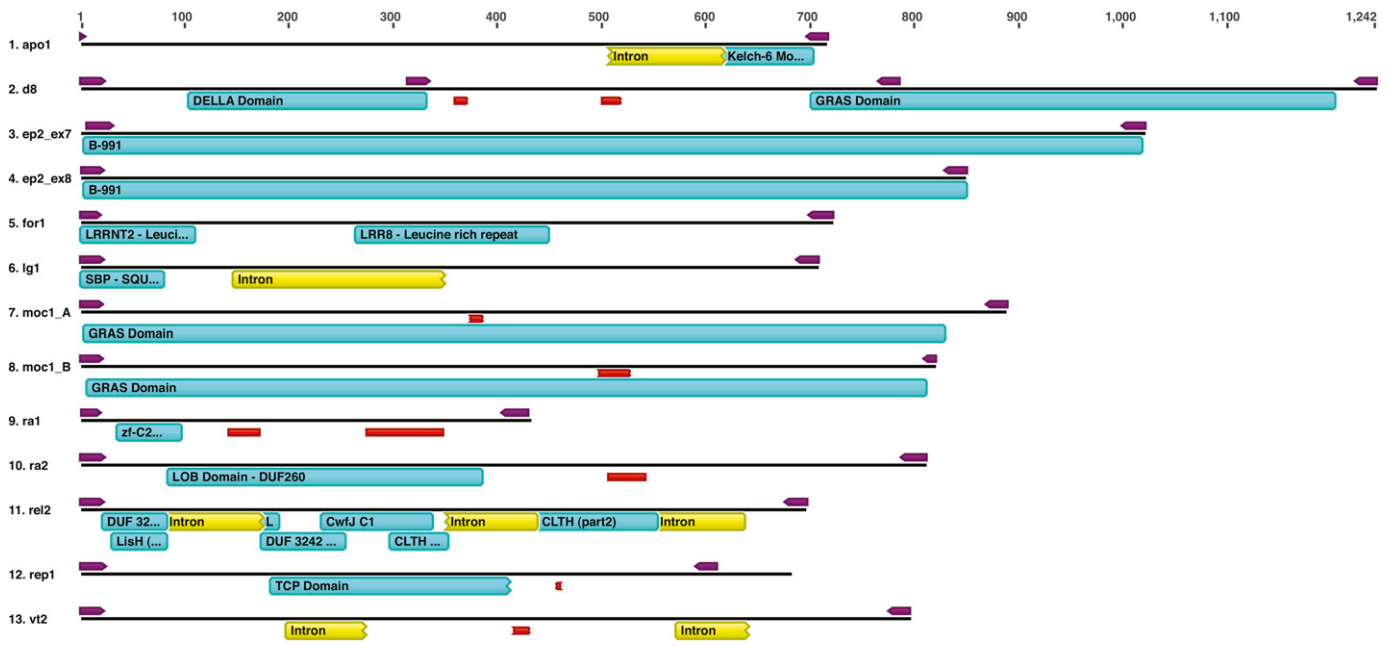

Fig. 1. Annotation of consensus sequences generated from 13 alignments. Sequence length (black line), primer binding locations (purple arrows), introns (yellow), ambiguously aligned regions (red), and Pfam domains (turquoise) are indicated.

TABLE 2. Characteristics of 13 sequence alignments generated from 10 diverse taxa and five reference sequences for 11 nuclear-encoded genes.

| | Alignment | | | Characters | | | | Pairwise variation | |
| | | | | | | Parsimony variable | | | |
Locus	Length (bp)	Intron	Ambiguous region	Total excluded	Constant	Uninformative	Informative	Min	Max
*apo*1	860	1	0	208	512	78	62	0.016	0.115
*d*8	1273	0	2	39	918	190	126	0.031	0.101
*ep*2_ex7	1054	0	0	0	746	199	109	0.018	0.158
*ep*2_ex8	869	0	0	0	640	144	85	0.007	0.128
*for*1	721	0	0	0	551	91	79	0.036	0.117
*lg*1	880	1	0	322	431	51	76	0.031	0.126
*moc*1 A	933	0	1	31	714	125	63	0.036	0.100
*moc*1 B	894	0	1	48	606	143	97	0.038	0.137
*ra*1	505	0	2	129	234	90	52	0.030	0.281
*ra*2	849	0	1	39	643	86	81	0.014	0.110
*rel*2	790	3	0	355	388	27	20	0.017	0.055
*rep*1	864	0	1	48	527	145	144	0.055	0.209
*vt*2	1101	2	1	450	510	60	81	0.034	0.103

A total of 13 primer pairs was generated for amplifying (12 pairs) and sequencing (1 pair) regions of 11 distinct loci found in the panicoid grasses (Table 1). These loci include *Aberrant panicle organization*1 (*apo*1), *Dwarf*8 (*d*8), two exons of *Erect panicle*2 (*ep*2), *Floral organ regulator*1 (*for*1), *liguleless*1 (*lg*1), *Monoculm*1 (*moc*1), *Ramosa*1 (*ra*1), *Ramosa*1 *Enhancer locus*2 (*rel*2), *Ramosa*2 (*ra*2), *Retarded palea*1 (*rep*1), and *Vanishing Tassel*2 (*vt*2) (Moreno et al., 1997; Peng et al., 1999; Jang et al., 2003; Li et al., 2003; Ikeda et al., 2005; Vollbrecht et al., 2005; Bortiri et al., 2006; Yuan et al., 2009; Gallavotti et al., 2010; Zhu et al., 2010; Phillips et al., 2011). Sequences generated with these primers produced 13 alignments with a combined alignable length of almost 10 000 bp. Two of these alignments (*moc*1 A & B) are generated with a single pair of primers and represent a duplication that occurred prior to the divergence of grasses. The *rep*1 primers occasionally amplify multiple loci as a single band, so a third primer (F2) was designed for nested PCR using the same thermocycler conditions for greater specificity. The number of parsimony informative characters ranged from 20 to 144, and the amount of pairwise variation ranged from a minimum of 0.007–0.055 to a maximum of 0.055–0.281. *d*8 and *rep*1 provide the most informative characters, with pairwise variation from 0.031–0.209 between sequences.

CONCLUSIONS

Combining genetic and genomic resources within the grasses is a powerful approach for designing phylogenetic markers. The 11 loci targeted for primer design in this study belong to a diverse set of genes that are known to influence plant and inflorescence architecture. These loci are distributed on nine of the 10 maize chromosomes, suggesting they are unlinked and providing independent records of duplications and gene loss within the tribe. Cloning PCR products and sequencing multiple clones allows the identification of paralogues, helping to identify allopolyploid events and reducing conflict between individual phylogenetic trees. These markers offer a valuable alternative to the nuclear-encoded ITS locus.

LITERATURE CITED

ÁLVAREZ, I., AND J. F. WENDEL. 2003. Ribosomal ITS sequences and plant phylogenetic inference. *Molecular Phylogenetics and Evolution* 29: 417–434.

BORTIRI, E., G. CHUCK, E. VOLLBRECHT, T. ROCHEFORD, R. MARTIENSSEN, AND S. HAKE. 2006. *ramosa2* encodes a LATERAL ORGAN BOUNDARY domain protein that determines the fate of stem cells in branch meristems of maize. *Plant Cell* 18: 574–585.

GADBERRY, M., S. MALCOMBER, A. DOUST, AND E. KELLOGG. 2005. Primaclade—A flexible tool to find conserved PCR primers across multiple species. *Bioinformatics (Oxford, England)* 21: 1263–1264.

GALLAVOTTI, A., J. A. LONG, S. STANFIELD, X. YANG, D. JACKSON, E. VOLLBRECHT, AND R. J. SCHMIDT. 2010. The control of axillary meristem fate in the maize *ramosa* pathway. *Development* 137: 2849–2856.

GRASS PHYLOGENY WORKING GROUP. 2001. Phylogeny and subfamilial classification of the Poaceae. *Annals of the Missouri Botanical Garden* 88: 373–457.

IKEDA, K., N. NAGASAWA, AND Y. NAGATO. 2005. *ABERRANT PANICLE ORGANIZATION 1* temporally regulates meristem identity in rice. *Developmental Biology* 282: 349–360.

JANG, S., B. LEE, C. KIM, S.-J. KIM, J. YIM, J.-J. HAN, S. LEE, ET AL. 2003. The *OsFOR1* gene encodes a polygalacturonase-inhibiting protein (PGIP) that regulates floral organ number in rice. *Plant Molecular Biology* 53: 357–372.

LI, X., Q. QIAN, Z. FU, Y. WANG, G. XIONG, D. ZENG, X. WANG, ET AL. 2003. Control of tillering in rice. *Nature* 422: 618–621.

MORENO, M., L. HARPER, R. KRUEGER, S. DELLAPORTA, AND M. FREELING. 1997. *liguleless1* encodes a nuclear-localized protein required for induction of ligules and auricles during maize leaf organogenesis. *Genes & Development* 11: 616–628.

PENG, J., D. E. RICHARDS, N. M. HARTLEY, G. P. MURPHY, K. M. DEVOS, J. E. FLINTHAM, J. BEALES, ET AL. 1999. 'Green revolution' genes encode mutant gibberellin response modulators. *Nature* 400: 256–261.

PHILLIPS, K. A., A. L. SKIRPAN, X. LIU, A. CHRISTENSEN, T. L. SLEWINSKI, C. HUDSON, S. BARAZESH, ET AL. 2011. *vanishing tassel2* encodes a grass-specific tryptophan aminotransferase required for vegetative and reproductive development in maize. *Plant Cell* 23: 550–566.

PUNTA, M., P. C. COGGILL, R. Y. EBERHARDT, J. MISTRY, J. TATE, C. BOURSNELL, N. PANG, ET AL. 2012. The Pfam protein families database. *Nucleic Acids Research* 40(D1): D290–D301.

SWOFFORD, D. 2002. PAUP 4.0: Phylogenetic analysis using parsimony and other methods. Sinauer, Sunderland, Massachusetts, USA.

VOLLBRECHT, E., P. S. SPRINGER, L. GOH, E. S. BUCKLER IV, AND R. MARTIENSSEN. 2005. Architecture of floral branch systems in maize and related grasses. *Nature* 436: 1119–1126.

YUAN, Z., S. GAO, D.-W. XUE, D. LUO, L.-T. LI, S.-Y. DING, X. YAO, ET AL. 2009. *RETARDED PALEA1* controls palea development and floral zygomorphy in rice. *Plant Physiology* 149: 235–244.

ZHU, K., D. TANG, C. YAN, Z. CHI, H. YU, J. CHEN, J. LIANG, ET AL. 2010. *Erect panicle2* encodes a novel protein that regulates panicle erectness in *Indica* rice. *Genetics* 184: 343–350.

APPENDIX 1. List of taxa sampled with voucher information and GenBank accession numbers.

Taxa sampled	Voucher (Herbarium)	GenBank accession no.												
		apo1	*d8*	*ep2_ex7*	*ep2_ex8*	*for1*	*lg1*	*moc1_A*	*moc1_B*	*ra1*	*ra2*	*rel2*	*rep1*	*vt2*
Andropogon hallii Hack.	*Kellogg PI477973* (A)	JQ951748	JQ951713	JQ951705	JQ951694	JQ951684	JQ951738	JQ951670	JQ951662	JQ951653	JQ951644	—	JQ951634	JQ951731
Andropterum stolzii (Pilg.) C. E. Hubb.	SM3091 (MO) 09°13′S, 034°56′E	JQ951745	JQ951707	JQ951698	JQ951690	JQ951681	JQ951734	JQ951668	JQ951663	JQ951651	JQ951643	JQ951718	JQ951631	JQ951725
Chrysopogon gryllus (L.) Trin.	*Kellogg PI250984* (A)	JQ951750	JQ951709	JQ951699	JQ951687	JQ951680	JQ951739	JQ951674	JQ951667	JQ951655	JQ951640	JQ951717	JQ951636	—
Coelorachis lepidura Staph.	SM3060 (MO) 06°28′S, 038°49′W	JQ951743	JQ951708	JQ951697	JQ951692	JQ951677	JQ951740	JQ951671	JQ951661	JQ951649	—	—	JQ951635	JQ951726
Cymbopogon flexuosus (Nees ex Steud.) Stapf	*Kellogg PI209700* (A)	JQ951744	JQ951710	JQ951700	JQ951689	JQ951679	JQ951733	JQ951669	JQ951666	JQ951650	JQ951641	JQ951722	JQ951632	JQ951730
Dichanthium annulatum (Forssk.) Stapf	*Kellogg PI240155* (A)	JQ951749	JQ951711	JQ951701	JQ951688	JQ951682	JQ951735	JQ951672	JQ951658	JQ951656	JQ951642	JQ951719	JQ951630	JQ951728
Loudetia Hochst. ex Steud.	SM3084 (MO) 08°11′S, 034°54′E	JQ951742	JQ951706	JQ951696	JQ951686	JQ951676	JQ951732	—	JQ951659	JQ951648	JQ951639	JQ951720	JQ951629	JQ951724
Phacelurus digitatus (Sibth. & Sm.) Griseb.	*Kellogg PI206746* (A)	JQ951747	JQ951714	JQ951702	JQ951691	JQ951678	JQ951736	—	JQ951665	JQ951652	JQ951646	JQ951716	JQ951637	JQ951727
Schizachyrium hirtiflorum Nees	*Kellogg PI216107* (A)	JQ951746	JQ951715	JQ951704	JQ951695	JQ951685	JQ951737	JQ951673	JQ951660	JQ951654	JQ951645	JQ951721	JQ951633	—
Sorghum bicolor (L.) Moench	*Kellogg PI255738* (A)	JQ951751	JQ951712	JQ951703	JQ951693	JQ951683	JQ951741	JQ951675	JQ951664	JQ951657	JQ951647	JQ951723	JQ951638	JQ951729

Note: — = the cloning reaction failed and sequences were not obtained; A = Arnold Arboretum Herbarium; MO = Missouri Botanical Garden Herbarium.