Expressed sequence tag library development and characterization of polymorphic microsatellite markers for the Neotropical spiral gingers, *Costus* (Costaceae)

Kathleen M Kay, Vanessa E Apkenas, and Jennifer M Yost

Department of Ecology & Evolutionary Biology, University of California Santa Cruz, 1156 High Street, Santa Cruz, CA 95064 USA

Correspondence: Kathleen M. Kay, EE Biology, 1156 High Street, Santa Cruz, CA 95064 USA, Fax: (831) 459-5353, E-mail: kmkay@ucsc.edu

Keywords: *Costus*, EST library, EST-SSR, microsatellites

Abstract

We present an expressed sequence tag (EST) library and a set of 15 polymorphic microsatellite markers developed for the Neotropical understory rainforest herbs, *Costus scaber* and *C. pulverulentus* (Costaceae). The EST library consists of 1221 reads, assembled into 912 unigenes. We tested primers for 90 microsatellites from the EST library across 5 geographically disparate populations each of *C. pulverulentus* and *C. scaber* and 6 more distantly related species from the genus. These resources will be useful for ongoing ecological and evolutionary studies of this rapidly diversifying genus.

The spiral ginger genus *Costus* (Costaceae) has undergone a rapid and recent radiation in the Neotropical forests, and provides an excellent study system for investigating ecological and evolutionary processes underlying tropical plant diversity and floral evolution. *Costus* is thought to have dispersed from Africa approximately 1.5-7.1 Ma and diversified into more than 50 species across the Neotropics (Kay *et al.*, 2005). *Costus scaber* and *Costus pulverulentus* are closely related species that have been a focus of studies of speciation (Kay, 2006). They, and other Neotropical *Costus* species, have also been the focus of ecological studies of species interactions and mating systems (e.g., Kay, Schemske, 2003). Here we report our efforts to develop an expressed sequence tag (EST) library and polymorphic microsatellite markers, tools that will expand the types of studies feasible in this genus.

We extracted total RNA from floral bud and leaf meristem tissue from a greenhouse-grown F1 hybrid between *C. scaber* (dam) and *C. pulverulentus* (sire). Both parent plants originated from La Selva Biological Station, Herédia Province, Costa Rica (La Selva). We used the Invitrogen PureLink™ RNA Mini Kit, with the addition of ABI RNA Isolation Aid during tissue homogenization. mRNA was then isolated with the Qiagen Oligotex mRNA Mini Kit and evaluated with a NanoDrop 1000 and an agarose gel. A cDNA library enriched for full-length transcripts was constructed from 174 ng pooled mRNA using the SMART cDNA library construction kit (Clontech) with the following modifications to the protocol. PCR amplification of first strand cDNA was done with Platinum-pfx DNA polymerase and its corresponding buffer (Invitrogen). We omitted Sfi I digestion, ligated the cDNA to the pCR-Blunt II-TOPO vector with a 1:1 vector:insert ratio, and transformed TOP10 cells with the Zero Blunt TOPO PCR cloning kit (Invitrogen). A subset of colonies were checked for successful inserts with PCR (forward primer 5'-AAGCAGTGGTATCAACGCAGAGT, reverse primer 5'-AGGCGGCCGACATGTTTTTTTTTTTT). Colonies were picked and grown in LB broth overnight, and we isolated DNA for sequencing using the AccuPrep Plasmid MiniPrep DNA Extraction Kit (Bioneer). Inserts from 1221 colonies were sequenced with the 5' SMART PCR primer on the ABI 3100 machine in the UCSC MEEG Facility. We performed base calling with Phred v.0.020425.c (Green and Ewing, 2002) and trimmed low quality and vector sequence and poly-A tails using Lucy v 1.20 (Chou, Holmes, 2001). The reads were submitted to NCBI GenBank dbEST (JK216135-JK217355). We assembled sequences with CAP3 (Huang, Madan, 1999), and created a unigene file containing 171 assembled contigs and 741 singletons.

Using SSRIT (Temnykh *et al.*, 2001) to find di-, tri-, tetra-, penta-, and hexa-nucleotide repeat motifs with a minimum of 5, 4, 3, 3, and 3 subunits, respectively, we identified 112 microsatellites in our 912 unigenes, including 21 di-, 60 tri-, 23 tetra-, 6 penta-, and 2 hexa-nucleotide repeats. We designed primers for 90 of these loci using Primer3 (Rozen, Skaletsky, 2000). We first screened all primer pairs for successful amplification using a single individual from 5 populations each of *C. scaber* and *C. pulverulentus* and from six other *Costus* species spanning the phylogeny of the genus. The populations of *C. scaber* and *C. pulverulentus* encompassed their combined geographic ranges from Mexico to Bolivia. The additional six species included Neotropical *C. malortieanus*, *C. lima*, *C. spiralis*, *C. laevis*, and *C. ricus*, and Paleotropical *C. tappenbeckianus*. Loci that amplified consistently and exhibited more than one allele across all species tested were then evaluated for polymorphism in a minimum of 20 individuals from the La Selva populations of *C. pulverulentus* and *C. scaber*, using DNA from leaf tissue that we collected and silica dried in the field. Polymorphism levels within the additional six species remain to be tested. All genomic DNA for this screening was extracted from specimens growing in the UCSC greenhouses using Qiagen DNEasy Plant Mini Kit. Except for *C. tappenbeckianus*, plants were originally collected in the field or acquired from the collections of the University of Utrecht in the Netherlands, and voucher information for all populations can be found in Kay *et al.* (2005). *Costus tappenbeckianus* DNA was sampled from a clonal division of W.J. Kress 94–3697 (US).

We screened loci with a nested PCR method with labeled 5' M13-FAM and 5' M13-HEX primers (Schuelke, 2000). Reactions consisted of 12.5 µl Promega GoTaq Hotstart Colorless Mastermix, 0.65 µl of 10 pmol/µl 5' M13-tailed forward primer, 2.5 µl of 10 pmol/µl reverse primer, 2.5 µl of 10 pmol/µl 5' M13 HEX or FAM labeled primer, 1 µl of DNA (concentration varied from 10-200 ng/µl) and 5.9 µl water for a final volume of 25 µl. All reactions were run with the following conditions: 94 °C for 5 min; 30 cycles of 94 °C for 30 s, touchdown annealing starting at either 60, 62, or 64 °C for 45 s and decreasing by 0.5 °C each cycle, 72 °C for 45 s; followed by 8 cycles of 94 °C for 30 s, 53 °C for 45 s, 72 °C for 45 s; and a final extension at 72 °C for 10 min. Products were verified on 0.8% agarose gels using 1x TBE or 1x SB buffer with Biotium GelRed™ Nucleic Acid Gel Stain.

Amplicons were sized at the UC Berkeley DNA Sequencing Facility, and alleles were scored using Applied Biosystems Peak Scanner v1.0 software. We evaluated loci for allelic diversity and Hardy-Weinberg equilibrium (HWE) with HW-QUICKCHECK (Kalinowski, 2006), tested for linkage disequilibrium within each species with Genepop 4.1 (Raymond, Rousset, 1995; Rousset, 2008), and tested for null alleles with MICRO-CHECKER using both the Brookfield and Chakraborty estimators and a 99% confidence interval (Van Oosterhout *et al.*, 2004). In addition, representatives of successful loci were sequenced to reconfirm their identity.

Seventy-four microsatellite loci out of ninety amplified consistently well across populations and species with a single PCR product. Fifteen of these showed products that were bigger than expected, indicating a possible intron. Forty-four loci amplified well but did not exhibit polymorphism. The remaining 15 loci amplified consistently and showed polymorphism in the La Selva populations (Table 1). No loci deviated significantly from HWE (Bonferroni-corrected $P < 0.05/15$; Table 1), and no significant linkage disequilibrium was found between these polymorphic loci following sequential Bonferroni correction. We also did not find any evidence for null alleles in these populations. These 15 loci successfully amplified in the six other *Costus* species, with the following rare exceptions: cdi4G6 in *C. tappenbeckianus*, ncdi8A10 in *C. malortieanus* and *C. ricus*, and nctet3E6 in *C. laevis*.

ESTs and the simple sequence repeats (SSRs) identified within them have become popular resources for microsatellite marker development due to their low cost, wide accessibility in online databases, transferability between taxa, and decreased error with null alleles (Ellis, Burke, 2007). However, they may have lower rates of polymorphism due to their location within coding DNA. In contrast to our expectations, there were no consistent patterns governing the successful development of reliable polymorphic SSRs in terms of the type of repeat or its location: di- and tetra-nucleotide repeats were just as likely to be polymorphic as tri-nucleotide repeats, and SSRs in the predicted open reading frame (ORF) were just as likely to amplify consistently and be polymorphic as those outside the predicted ORF. In contrast to other reports in the literature that SSRs are most common in the 5' UTR (reviewed in Bouck, Vision, 2007), we found 65 SSRs in the predicted ORF, 33 in the predicted 3' UTR, and only 7 in the predicted 5' UTR (the remaining 7 were found in unigenes without a predicted ORF). The number of alleles per locus within a population for our successful microsatellites ranged from 1 to 11 (median 2), which is relatively low for microsatellites. However, this relatively conservative rate of evolution facilitates their wide transferability across the genus, as there appears to be little divergence in the priming sites. With rare exception, all loci amplified across the genus, including the African species *C. tappenbeckianus*, which falls well outside the Neotropical radiation of *Costus*. These results suggest that these loci will infrequently exhibit null alleles and will be useful for future ecological and evolutionary studies throughout the Neotropical *Costus*.

References

Bouck A, Vision T (2007) The molecular ecologist's guide to expressed sequence tags. *Molecular Ecology* **16**, 907-924.

Chou H-H, Holmes MH (2001) DNA sequence quality trimming and vector removal. *Bioinformatics* **17**, 1093-1104.

Ellis JR, Burke JM (2007) EST-SSRs as a resource for population genetic analyses. *Heredity* **99**, 125-132.

Huang X, Madan A (1999) CAP3: A DNA sequence assembly program. *Genome Research* **9**, 868-877.

Kalinowski ST (2006) HW-QUICKCHECK: an easy-to-use computer program for checking genotypes for agreement with Hardy-Weinberg expectations. *Molecular Ecology Notes* **6**, 974-979.

Kay KM (2006) Reproductive isolation between two closely related hummingbird-pollinated Neotropical gingers. *Evolution* **60**, 538-552.

Kay KM, Reeves PA, Olmstead RG, Schemske DW (2005) Rapid speciation and the evolution of hummingbird pollination in Neotropical *Costus* subgenus *Costus* (Costaceae): Evidence from nrDNA ITS and ETS sequences. *American Journal of Botany* **92**, 1899-1910.

Kay KM, Schemske DW (2003) Pollinator assemblages and visitation rates for 11 species of Neotropical *Costus* (Costaceae). *Biotropica* **35**, 198-207.

Raymond M, Rousset F (1995) GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *Journal of Heredity* **86**, 248-249.

Rousset F (2008) Genepop'007: a complete reimplementation of the Genepop software for Windows and Linux. *Molecular Ecology Resources* **8**, 103-106.

Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods in Molecular Biology* **132**, 365-386.

Schuelke M (2000) An economic method for the fluorescent labeling of PCR fragments. *Nature Biotechnology* **18**, 233-234.

Temnykh S, DeClerck G, Lukashova A*, et al.* (2001) Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. *Genome Research* **11**, 1441-1452.

Van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P (2004) MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. *Molecular Ecology Notes* **4**, 535-538.

Acknowledgements

We thank A. Louthan for assistance with lab work, Y. Surget-Groba for assistance with microsatellite scoring, and C. Machado for advice on cDNA library construction. This work was partially supported by NSF DEB-0947138 to KMK.

Data Accessibility:

DNA sequences: Genbank accessions JK216135-JK217355

Table 1. Characterization of 15 microsatellite loci [individuals screened (*n*), alleles observed at each locus (*k*), observed (H_O) and expected (H_E) heterozygosity, Hardy-Weinberg equilibrium probability (HWE)].

				Costus scaber					***Costus pulverulentus***				
Locus/ GenBank Accession	**Primer sequence (5' – 3')**	**Repeat motif in clone**	**Size (bp)†**	*n*	*k*	H_O	H_E	HWE	*n*	*k*	H_O	H_E	HWE
nctri170/ JK217202	F: TGGAGGGAATAGAGGTCGTG R: GCCGTGATCCATCCATTATT	(TCG)7	237-249	20	3	0.45	0.50	0.36	23	3	0.30	0.33	0.52
ncdi8A10/ JK216680	F: GGGGTTTCTTCTCCGAGTCT R: AGGATAACACACACGCCTCC	(TC)17	180-210	20	11	0.75	0.85	0.15	20	11	1.00	0.89	0.08
chex12B9/ JK216985	F: TGACAGCAGAGAGCGTATCG R: CTACCTCCGAATGTTTCCCA	(TTGCTG)4	189-207	21	4	0.24	0.34	0.07	23	2	0.30	0.41	0.20
ctri13A12/ JK217056	F: TTGGGAACCAGAGGAAAATG R: ACGAACAGGTTCAATCCGTC	(GGC)7	253-274	20	4	0.50	0.57	0.29	22	6	0.59	0.71	0.13
ctri2D9/ JK216253	F: GGAGAGCGAGCAGAGAACAC R: ATTGAACAGGGCGTCGATAG	(TCT)8	152-170	21	5	0.38	0.41	0.46	23	4	0.52	0.54	0.50
ctri4A11/ JK216382	F: AGACGAAGACGACGATGTCC R: GCTGAGGTATTCAGATCGCC	(GAC)5	230-233	21	2	0.24	0.47	0.03	22	1	-	-	-
nctri1C9/ JK216161	F: GAGACCCCTGTTGTTGTCGT R: GTTCTCCATCACCACCATCA	(TGT)5	151-154	20	2	0.05	0.05	0.50	23	2	0.09	0.23	0.02
nctri113/ JK216242	F: GCTCCTGTGGTTGCTTCTTC R: CTGCAACATGGAATCCAACA	(CAT)4	135-138	20	2	0.10	0.10	0.97	20	1	-	-	-
ctri3B1/ JK216305	F: CCCGTCATTTCTGCTGTGTA R: GACAACAGGGCCTCTTTGAA	(TGA)4	246-255	23	2	0.09	0.09	0.98	20	1	-	-	-

Table 1. Characterization of 15 microsatellite loci [individuals screened (*n*), alleles observed at each locus (*k*), observed (H_O) and expected (H_E) heterozygosity, Hardy-Weinberg equilibrium probability (HWE)].

				Costus scaber					***Costus pulverulentus***				
Locus/ GenBank Accession	**Primer sequence (5' – 3')**	**Repeat motif in clone**	**Size (bp)†**	*n*	*k*	H_O	H_E	HWE	*n*	*k*	H_O	H_E	HWE
nctet3E6/ JK216336	F: CAGTTGGAGGAAGAATCCGA R: CGGCACACCCCTTTTTAAT	(TGTA)3	144- 148	21	2	0.05	0.05	0.50	23	2	0.13	0.20	0.22
cdi10E12/ JK216871	F: CACGAGCACCATGAGAAGAA R: TCTTCACAAGCCACAAGCAG	(AG)6	156- 168	20	2	0.05	0.05	0.50	20	2	0.10	0.10	0.97
cdi4G6/ JK216441	F: TAGCCCGAGTCAAGCAGATT R: GTTTCGCCCGTGATACAACT	(AT)6	233- 243	20	6	0.50	0.69	0.04	20	4	0.60	0.59	0.57
ctri3D11/ JK216330	F: CTCGAGACTTCTCCTCGTCG R: AATATGTCACGGTTACCGCC	(TCC)5	270- 276	21	2	0.38	0.48	0.29	20	3	0.15	0.14	0.92
ctet53/ JK216440	F: CAAGAACGCCGTCAAGTACC R: ACTGATCTGTCGTTTGCACG	(TGTT)3	172- 184	25	2	0.32	0.49	0.08	26	3	0.46	0.50	0.45
ctet5C2/ JK216473	F: TCCGATGCGTGTAGTTTCTG R: ATGCACAAGAAGAGGCCTGA	(GAAA)3	256- 259	20	1	-	-	-	20	2	0.05	0.05	0.50

†Size given includes 18 bp M13-tail