Microsatellite Primer Development for Post Oak, Quercus stellata (Fagaceae)

Authors: Chatwin, Warren B., Carpenter, Kyrie K., Jimenez, Felix R., Elzinga, Dave B., Johnson, Leigh A., et al.

MICROSATELLITE PRIMER DEVELOPMENT FOR POST OAK, QUERCUS STELLATA (FAGACEAE)[1]

WARREN B. CHATWIN[2], KYRIE K. CARPENTER[2], FELIX R. JIMENEZ[2], DAVE B. ELZINGA[2], LEIGH A. JOHNSON[3], AND PETER J. MAUGHAN[2,4]

[2]Department of Plant and Wildlife Sciences, Brigham Young University, 5144 Life Sciences Building, Provo, Utah 84602 USA; and [3]Department of Biology, Brigham Young University, 4058 Life Sciences Building, Provo, Utah 84602 USA

- *Premise of the study:* The American Cross Timbers forest ecosystem runs from southeastern Kansas to Central Texas and is primarily composed of post oak (*Quercus stellata*). This old-growth forest currently occupies only about 2% of its ancestral range. To facilitate genetic research on this species, we developed microsatellite primers specific to post oak from reduced genomic libraries.
- *Methods and Results:* Two *Q. stellata* individuals, sampled from the northern and southern range of the post oak forest, were subject to genomic reduction and 454 pyrosequencing. Bioinformatic analysis identified putative microsatellites from which 12 polymorphic primer sets were screened on three populations. The number of alleles observed ranged from five to 20 across all populations, while observed and expected heterozygosity values ranged from 0.05 to 0.833 and 0.236 to 0.893, respectively, within individual populations.
- *Conclusions:* We report the development of microsatellite markers, specific to post oak, to aid the study of genetic diversity and population structure of extant forest remnants.

Key words: 454 pyrosequencing; Cross Timbers; old-growth forest; post oak; *Quercus stellata*; simple sequence repeat.

The Cross Timbers forests of central North America contain some of the largest tracts of old-growth forests remaining in the continental United States and are predominantly populated with two species of oak: post oak (*Quercus stellata* Wangenh.) and blackjack oak (*Q. marilandica* Münchh.; Therrell and Stahle, 1998). Originally, these forests covered nearly 79,062 km^2 (Kuchler, 1964), but remnant old-growth stands are now estimated to cover only about 1200 km^2 (Bayard, 2003; Peppers, 2004) ranging from eastern Oklahoma to North Texas. The remnant stands in eastern Kansas have never been spatially quantified. The remaining stands of old-growth forest are highly fragmented but relatively undisturbed because they grow in areas that are very rocky and difficult to navigate, farm, or develop. Post oak is a slow-growing tree, typically 9–12 m tall with trunk diameter of 38–46 cm in the Cross Timbers eco-region (Stransky, 1990). Because post oak is such a slow-growing species, fire is necessary to maintain post oak as the dominant canopy species. In the absence of fire or under active fire-suppression regimes, the forest canopy composition changes from a fire-tolerant post oak canopy to faster-growing fire-intolerant species such as softwoods and mesophytic plants (Abrams, 1992).

Little genetic or ecological research has been performed on this characteristic old-growth forest species. A genetic examination of the extant forest remnants is needed to quantify remaining genetic diversity in post oak and to identify populations of highest conservation priority. Here we use 454 pyrosequencing to discover microsatellites from a pooled library of post oak samples using reduced representational libraries based on restriction site conservation (Maughan et al., 2009), with each sample labeled independently using a unique multiplex identifier barcode. The incorporation of the barcodes into specific DNA sequence fragments allows for the unambiguous assignment of sequences to specific DNA samples in the sequence pool, which can then be used to identify putatively polymorphic microsatellite loci among samples from the sequence data itself, effectively eliminating the need for polymorphism screening and genotyping validation (an expensive and often laborious process).

METHODS AND RESULTS

DNA was extracted from bud and catkin tissue from each of three populations located across the Cross Timbers range, including a northern (Kansas; $n = 29$), a central (Oklahoma; $n = 30$), and a southern (Texas; $n = 28$) population (Appendix 1). Voucher specimens are archived at the Stanley L. Welsh herbarium (BRY) at Brigham Young University, Provo, Utah.

Tissue was ground in liquid nitrogen and DNA extracted using a modified mini-salts protocol as reported by Todd and Vodkin (1996) supplemented with 1% (w/v) polyvinylpyrrolidone-10 (Sigma-Aldrich, St. Louis, Missouri, USA). Microsatellite markers were identified from reduced representational libraries, also known as genomic reduction libraries, of two individuals collected from the northern and southern range of post oak. The reduced representational

libraries were generated following Maughan et al. (2009) and were 454 pyro-sequenced at the DNA Sequencing Center at Brigham Young University (DNASC), Provo, Utah. A total of 305,717 reads were identified (approximately 150,000 reads from each individual) representing 184 Mb with an average read length of 601 bp (N50 = 623 bp). All reads were assembled into 18,270 contigs using the default parameters of the Roche Newbler assembler (version 2.3; 454 Life Sciences, a Roche Company, Branford, Connecticut, USA). A total of 1191 perfect microsatellite repeats were identified from the contigs using the computer program MISA (Thiel et al., 2003), including 809 di-, 291 tri-, and 91 tetranucleotide repeat motifs, with minimum repeat units of 8, 6, and 5, respectively. The most common motifs were AG/CT, AAT/ATT, and AAAT/ATTT.

Putative microsatellite markers were then further filtered using custom Perl scripts (Stajich et al., 2002) to identify 48 microsatellite loci that varied in repeat length between the consensus reads of each post oak individual. Flanking primers for all 48 loci were designed using Primer3 version 2.0 (Rozen and Skaletsky, 2000) with default parameters except for: product size = 150–250 bp, maximum melting temperature (T_m) difference = 1°C, and maximum poly X = 3. All 48 primer pairs were screened for strong amplification and polymorphism on an initial panel of four post oak individuals. The PCR reactions used a HotStarTaq Plus Master Mix Kit (QIAGEN, Germantown, Maryland, USA) in 10-μL reactions with 10 ng of genomic DNA and 0.5 μM of each primer according to the manufacturer's guidelines with the following amplification parameters: 95°C for 5 min, followed by 35 cycles of 94°C for 45 s, 55°C for 30 s, 72°C for 60 s, followed by a final extension cycle of 72°C for 10 min. The amplicons were visualized in a UV transilluminator using 3% MetaPhor Agarose (Lonza Ltd., Allendale, New Jersey, USA) gels electrophoresed for 6 h and stained with ethidium bromide. Twenty-four primer pairs failed to amplify or produced complex (multiple bands) banding patterns.

Twelve of the polymorphic microsatellites were then fluorescently labeled with VIC, FAM, or NED (Life Technologies, Grand Island, New York, USA; Table 1) and amplified across all individuals in each of the three populations ($n = 87$) using the multiplex PCR Type-it Microsatellite PCR Kit (QIAGEN) according to the manufacturer's instructions in 10-μL reactions with 30 ng of genomic DNA and 0.2 μM of all primers with the following amplification profile: 95°C for 5 min, followed by 30 cycles of 95°C for 30 s, 55°C for 90 s, 72°C for 30 s, followed by a final extension cycle of 60°C for 30 min. Four microliters of the amplified products were processed for fragment analysis on an ABI 3730xl using GeneScan 500 ROX (Applied Biosystems, Carlsbad, California, USA) as the size standard. Fragment analysis and scoring was performed using the microsatellite plugin in Geneious 5.6.7 (Drummond et al., 2011), and the data analyzed for genic differentiation (F_{ST}), deviation from Hardy–Weinberg

equilibrium (HWE), observed heterozygosity (H_o), expected heterozygosity (H_e), and total number of alleles using the computer program Arlequin version 3.5 (Excoffier and Lischer, 2010). Linkage disequilibrium was tested for each pair of loci across and within populations using the log likelihood ratio statistic and default Markov chain parameters in GENEPOP (Raymond and Rousset, 1995). BLASTX searches of the RefSeq_Protein data with the flanking microsatellite DNA sequences produced only a single significant hit (E-value = 2e-55) for QS00766 to GenBank accession number XP_007012674.1, an RNA polymerase-associated protein CTR9 isoform 5 from *Theobroma cacao* L. The primer sequences, repeat motif type, and expected product size for the microsatellites not selected for fluorescent analysis are provided in Appendix S1.

The 12 microsatellite loci produced a total of 168 alleles across all populations, with an average number of alleles per locus of 14. The total number of alleles varied from a low of five (QS02499 locus) to a high of 20 (QS01386 locus). Within individual populations, the average number of alleles ranged from a low of eight (Oklahoma) to a high of 9.167 (Texas). The mean H_o and H_e were 0.417 and 0.752, respectively, across all populations, while in individual populations the H_o and H_e values ranged from 0.05 to 0.833 and 0.236 to 0.893, respectively (Table 2). F_{ST} among populations was 0.066. Significant departure from HWE was identified for 11 out of the 12 microsatellite markers when measured across all individuals, without respect to population (Table 2). When tested within populations, four to five markers were in HWE in each population, suggesting possible subpopulation structure due to geographic isolation and genetic drift (we note that the sample sizes are too small for definitive conclusions). Of the 66 possible pairwise comparisons, significant linkage disequilibrium ($P < 0.01$) was identified across populations between two sets of loci, specifically between QS01386 and QS08928, and between QS00984 and QS02499. Within populations, significant linkage disequilibrium was identified between a single set of loci for both the Texas (loci QS00984 and QS02499) and Oklahoma (loci QS01386 and QS08928) populations, while three pairwise comparisons evidenced linkage disequilibrium (loci QS00766 and QS01904, QS00984 and QS00562, and QS00984 and QS03297) in the Kansas population.

CONCLUSIONS

We report the development of 12 highly polymorphic microsatellite markers for post oak. Our preliminary analysis of three post oak populations suggests that genetic diversity within populations remains high and that only low levels of among-population

TABLE 1. Characteristics of 12 polymorphic microsatellite primers developed for *Quercus stellata*.

Locus	Primer sequences (5′–3′)	Repeat motif	Product size (bp)	Multiplex set	GenBank accession no.
QS01386	F: 6FAM-CCCTCACTCTTATGATTGTCCA R: ATGCTGCTTTTTGGGGAGTA	$(CT)_6$	189	1	KM027297
QS00984	F: VIC-TTTCATTTCAAGAAACAACAAGTGA R: CAATCTCATCGTCCAAGCCT	$(AAT)_6$	152	1	KM027296
QS03297	F: NED-ACACAAAGAGCCATTCGCTT R: GAGGCATACCTACGGGACAA	$(CA)_7$	220	1	KM027301
QS00314	F: 6FAM-TCAAAACGCAACGTTTCAAG R: TTCGGGTTTTCTTTGTGGTC	$(GAA)_6$	174	2	KM027293
QS10751	F: VIC-CCATACAATCAAAAACACCGA R: TTTTTCCCCTTTGGAATTGA	$(AATA)_5$	148	2	KM027304
QS02499	F: NED-TCCTTGTAAGTCCAAGCAAACA R: TGCTGAAGCAACCAATGAAG	$(CTT)_5$	216	2	KM027300
QS08928	F: 6FAM-GTAAGGAAGTTGGGGGTTCC R: TGCCTTTAGCTCTATATTTCACCC	$(ATA)_5$	171	3	KM027303
QS03797	F: VIC-AGTTTTGGTTGATTGTGACGG R: GCAAGGTCTGTCATTGGGTT	$(TCT)_5$	147	3	KM027302
QS00766	F: 6FAM-CCATTGGCATGGATGTAATTT R: CAATAGCAACAGCAATGTGAGA	$(TC)_9$	165	4	KM027295
QS01904	F: VIC-TCAGTCAAAAACCCACCTCC R: GGGTTTTCTTCAGTTTGCTTGT	$(TC)_{10}$	141	4	KM027299
QS01729	F: VIC-TCCTTCACAGAAACTCAGCAAA R: GAGTCCTGAGTAGGAACTGATGTG	$(AAG)_8$	125	5	KM027298
QS00562	F: NED-ACCCCCACCTAATCCCAAC R: TGCAAACACACAGAGACACTTTT	$(GA)_7$	225	5	KM027294

TABLE 2. Number of alleles detected at each microsatellite marker. Bold values indicate a significant deviation ($P < 0.001$) from Hardy–Weinberg equilibrium.

Locus	Kansas ($n = 29$)			Texas ($n = 28$)			Oklahoma ($n = 30$)			Total ($n = 87$)		
	H_o	H_e	A	H_o	H_e	A	H_o	H_e	A	H_o	H_e	A
QS01386	**0.240**	**0.893**	13	**0.440**	**0.856**	13	**0.520**	**0.859**	13	**0.400**	**0.887**	20
QS00984	**0.538**	**0.892**	11	0.769	0.795	9	0.379	0.554	5	**0.556**	**0.754**	13
QS03297	**0.111**	**0.786**	7	**0.385**	**0.821**	12	**0.260**	**0.798**	7	**0.250**	**0.803**	16
QS00314	0.438	0.476	6	0.833	0.743	7	0.538	0.551	7	0.600	0.648	11
QS10751	**0.333**	**0.87**	10	**0.278**	**0.862**	9	**0.363**	**0.890**	13	**0.327**	**0.904**	18
QS02499	**0.150**	**0.506**	4	**0.050**	**0.583**	4	**0.074**	**0.609**	4	**0.090**	**0.586**	5
QS08928	0.556	0.778	7	0.600	0.778	10	**0.310**	**0.633**	8	**0.472**	**0.738**	13
QS03797	**0.346**	**0.760**	10	0.833	0.820	9	**0.276**	**0.716**	7	**0.468**	**0.789**	15
QS00766	0.607	0.701	6	**0.536**	**0.785**	8	**0.300**	**0.597**	6	**0.477**	**0.768**	9
QS01904	0.577	0.797	9	0.720	0.862	12	0.815	0.823	10	**0.705**	**0.846**	15
QS01729	**0.545**	**0.856**	12	**0.222**	**0.440**	5	**0.476**	**0.775**	10	**0.400**	**0.793**	18
QS00562	0.333	0.566	4	0.350	0.637	12	0.125	0.236	6	**0.265**	**0.509**	15
Mean	0.398	0.734	8.25	0.501	0.748	9.167	0.37	0.67	8	0.417	0.752	14

Note: A = number of alleles; H_e = expected heterozygosity; H_o = observed heterozygosity; n = number of individuals sampled.

differentiation (F_{ST}) exist. These observations suggest that the remnant populations we sampled have not had sufficient time to show the effects of severe population reduction (inbreeding and genetic drift), perhaps due to the long lifespan (300+ years) of post oak. The markers reported here should provide the molecular tools needed for larger investigations of population structure and genetic diversity in this key species of the Cross Timbers forest of North America.

LITERATURE CITED

ABRAMS, M. D. 1992. Fire and the development of oak forests. *Bioscience* 42: 346–353.

BAYARD, A. R. 2003. Quantifying spatial distribution of ancient oaks with predictive modeling. Master's thesis, University of Arkansas, Fayetteville, Arkansas, USA.

DRUMMOND, A. J., B. ASHTON, S. BUXTON, M. CHEUNG, A. COOPER, C. DURAN, M. FIELD, ET AL. 2011. Geneious v5.4 created by Biomatters. Website http://www.geneious.com [accessed 8 September 2014].

EXCOFFIER, L., AND H. E. L. LISCHER. 2010. Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources* 10: 564–567.

KUCHLER, A. W. 1964. Potential natural vegetation of the coterminous United States. American Geographical Society Special Publication 36. American Geographical Society, New York, New York, USA.

MAUGHAN, P. J., S. M. YOURSTONE, E. N. JELLEN, AND J. A. UDALL. 2009. SNP discovery via genomic reduction, barcoding and 454-pyrosequencing in amaranth. *Plant Genome* 2: 260–270.

PEPPERS, K. C. 2004. Old-growth forests in the western Cross Timbers of Texas. PhD Thesis, University of Arkansas, Fayetteville, Arkansas, USA.

RAYMOND, M., AND F. ROUSSET. 1995. GENEPOP (version 1.2): Population genetics software for exact tests and ecumenicism. *Journal of Heredity* 86: 248–249.

ROZEN, S., AND H. SKALETSKY. 2000. Primer3 on the WWW for general users and for biologist programmers. *In* S. Misener and S. A. Krawetz [eds.], Methods in molecular biology, vol. 132: Bioinformatics methods and protocols, 365–386. Humana Press, Totowa, New Jersey, USA.

STAJICH, J. E., D. BLOCK, K. BOULEZ, S. E. BRENNER, S. A. CHERVITZ, C. DAGDIGIAN, G. FUELLEN, ET AL. 2002. The Bioperl toolkit: Perl modules for the life sciences. *Genome Research* 12: 1611–1618.

STRANSKY, J. J. 1990. *Quercus stellata* (Wangenh.), Post Oak. *In* R. M. Burns and B. H. Honkala [eds.], Silvics of North America, Vol. 2, Hardwoods. U.S. Department of Agriculture, Forest Service, Agriculture Handbook 654, 738–744. U.S. Department of Agriculture, Forest Service, Washington, D.C., USA.

THERRELL, M. D., AND D. W. STAHLE. 1998. A predictive model to locate ancient forests in the Cross Timbers of Osage County, Oklahoma. *Journal of Biogeography* 25: 847–854.

THIEL, T., W. MICHALEK, R. VARSHNEY, AND A. GRANER. 2003. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theoretical and Applied Genetics* 106: 411–422.

TODD, J. J., AND L. O. VODKIN. 1996. Duplications that suppress and deletions that restore expression from a chalcone synthase multigene family. *Plant Cell* 8: 687–699.

APPENDIX 1. Geographic origin of the 87 *Quercus stellata* accessions studied. Plants sampled within a location were separated by a minimum of 10 m.

Population	Voucher no.[a]	Location	n	Latitude	Longitude
KS1	623718	Independence, KS	29	37°15′25.06″N	95°46′37.71″W
OK1	623720	Eufala Lake, OK	30	35°16′33.57″N	95°32′27.48″W
TX1	623719	Bowie, TX	28	33°28′2.06″N	97°52′34.84″W

Note: n = number of individuals sampled at each location.

[a] Voucher samples for each population are maintained in the Stanley L. Welsh herbarium (BRY) at Brigham Young University, Provo, Utah.