WHOLE GENOME SHOTGUN SEQUENCES FOR MICROSATellite DISCOVERY AND APPLICATION IN CULTIVATED AND WILD MACADAMIA (PROTEACEAE)  

CATHERINE J. NOCK2,4, MARTIN S. ELPHINSTONE2, GARY ABLETT2, ASUKA KAWAMATA2, WAYNE HANCOCK2, CRAIG M. HARDNER3, AND GRAHAM J. KING2

2Southern Cross Plant Science, Southern Cross University, Lismore, New South Wales, Australia; and 3Queensland Alliance for Agriculture and Food Innovation, University of Queensland, St. Lucia, Queensland, Australia

• Premise of the study: Next-generation sequencing (NGS) data are widely used for single-nucleotide polymorphism discovery and genetic marker development in species with limited available genome information. We developed microsatellite primers for the Proteaceae nut crop species Macadamia integrifolia and assessed cross-species transferability in all congeners to investigate genetic identification of cultivars and gene flow.

• Methods and Results: Primers were designed from both raw and assembled Illumina NGS paired-end reads. The final 12 microsatellite markers selected were polymorphic among wild individuals of all four Macadamia species—M. integrifolia, M. tetraphylla, M. ternifolia, and M. jansenii—and in commercial macadamia cultivars including hybrids.

• Conclusions: We demonstrate the utility of raw and assembled Illumina NGS reads from total genomic DNA for the rapid development of microsatellites in Macadamia. These primers will facilitate future studies of population structure, hybridization, parentage, and cultivar identification in cultivated and wild Macadamia populations.

Key words: crop; cultivar; horticulture; Macadamia; nut; Proteaceae.

Macadamia is a recently domesticated nut crop derived from the Australian subtropical rainforest species Macadamia integrifolia Maiden & Betche and M. tetraphylla L. A. S. Johnson and their hybrids. Within the genus, all species, including M. ternifolia F. Muell. and M. jansenii C. L. Gross & P.H. Weston, are under threat of genetic erosion (Mas et al., 2008; Costello et al., 2009). Commercial cultivars were developed primarily in Hawaii and are only a few generations removed from Australian wild progenitors (Hardner et al., 2009). Macadamia are preferentially out-crossing and take four to five years to reach maturity. For breeding programs to progress effectively, there is a need to discriminate among clonally propagated industry standard cultivars and novel selections well before maturity. Although the 17 available M. integrifolia microsatellite markers with perfect repeats were tested in our laboratory (Schmidt et al., 2006), only four amplified successfully. These results are consistent with previous research on M. integrifolia (Neal, 2008), and no published study has used more than four polymorphic markers (Shapcott and Powell, 2011; Spain and Lowe, 2011). Additional microsatellite markers are needed to support conservation studies and breeding programs.

Next-generation sequencing (NGS) platforms are now routinely used for isolation of microsatellite, or simple sequence repeat (SSR), loci from plants (Egan et al., 2012). Long-read platforms are commonly used because reads of 300 to 500 bp in length may contain both the SSR motif and flanking sequence for primer design (Zalapa et al., 2012). Together, paired-end reads from short-read platforms also contain the SSR motif and flanking sequence for primer design at a lower cost per base (Silva et al., 2013). The aim of this study was to develop polymorphic microsatellite markers for Macadamia using paired-end Illumina reads with and without de novo assembly.

METHODS AND RESULTS

Fresh leaf material was collected from macadamia nut cultivars at Clunes Variety Trial M2, Clunes, New South Wales, Australia (Stephenson and Gallagher, 2000). Additional cultivars and clones of wild-collected individuals of all four Macadamia species were sourced from the Australian Macadamia Germplasm Collection at Alstonville Tropical Fruit Research Station, NSW Department of Primary Industries. Herbarium material is deposited at the Southern Cross University Medicinal Plant Herbarium (PHARM), Lismore, New South Wales, Australia (Appendix 1). Fresh leaf material was stored at −80°C (for Illumina sequencing) or after collection dried in a sealed container with 10× silica gel by fresh weight. Total DNA was extracted using a QIAGEN DNeasy Plant Kit (QIAGEN, Valencia, California, USA) according to manufacturer’s protocols. Approximately 4.5 μg of DNA extracted from one individual of M. integrifolia was submitted to the Australian Genome Research Facility, Melbourne, for sequencing. A DNA library was prepared with an Illumina TruSeq Sample Preparation Kit (version 2) following the manufacturer’s instructions (Illumina, San Diego, California, USA). Genomic DNA was sheared using a
Table 1. Characterization of 12 polymorphic microsatellite loci developed in Macadamia integrifolia.4

| Locus  | Primer sequences (5′-3′) | Repeat motif | Fluorescent label | Allele size range (bp) | T_a (°C) | GenBank accession no. |
|--------|-------------------------|--------------|-------------------|------------------------|----------|----------------------|
| Mac001 | F: GTGACTGGTGACCAAAACCA (AT)_11 | VIC | 412–420 | 60 | KF130888 |
| Mac002 | R: GCACCTGGTACCCCGCCTTT | (CT)_6 | NED | 283–297 | 60 | KF130889 |
| Mac003 | F: TGGACCTTTGGAGGATTGACT | (AT)_10 | FAM | 258–276 | 60 | KF130890 |
| Mac004 | R: TCCACGTTTACCTTGCTAGC | (AT)_10 | NED | 224–240 | 60 | KF130891 |
| Mac005 | F: CATCACTGAUTGATGAGAACTAA | (AAG)_10 | FAM | 331–343 | 60 | KF130892 |
| Mac006 | R: ATTCAACCAACTCTGTGATTT | (AG)_11 | PET | 322–360 | 55 | KF130893 |
| Mac007 | F: AGGCCCTGGATTGCTTCGAT | (CT)_11 | NED | 368–390 | 60 | KF130894 |
| Mac008 | R: GCAAAATACCAAGGACCCCTGGCC | (AT)_10 | FAM | 388–398 | 60 | KF130895 |
| Mac009 | F: AGACCTGGCTCTGTCTCTTGA | (AAG)_13 | VIC | 241–244 | 60 | KF130896 |
| Mac010 | R: TGCCTATGATCCTTTACGAG | (AG)_11 | PET | 259–297 | 55 | KF130897 |
| Mac011 | F: GGCAATCGATAGCATACAGAAT | (CT)_6 | FAM | 175–199 | 60 | KF130898 |
| Mac012 | R: TCTGATGATCCTCTTTACGAG | (AC)_10 | VIC | 309–321 | 60 | KF130899 |

Note: T_a = annealing temperature used for all Macadamia species and cultivars.

*Values based on 22 samples representing Macadamia cultivars located at Chunes Varietal Trial M2, New South Wales, Australia.

Covaris S2 sonication device (Covaris, Woburn, Massachusetts, USA). DNA fragments were end-repaired, A-tailed, and ligated to adapters. Size and concentration of DNA fragments were assessed using a DNA 100 chip on a Bioanalyzer 2100 instrument (Agilent Technologies, Santa Clara, California, USA).

Average insert size of the library was 424 bp. Approximately 4 pmol of the library was paired-end sequenced (100 x 2 cycles) on an Illumina Hi-Seq 2000 instrument.

Paired-end reads were imported into CLC Genomics Workbench (version 4.9; CLC Bio, Aarhus, Denmark) and trimmed to remove low-quality base calls (qadro > 20; P < 0.01) and adapter sequences. For the purpose of primer design, reads containing SSR motifs were identified as follows. Raw sequence reads: the search function was used to identify di- and trinucleotide SSR motifs with a minimum of eight repeats in raw sequence reads. SSR regions were identified at the 3′-end of a read. Primers were then designed in the flanking regions (i.e., 5′-end of read containing SSR) and in the matching paired-end read. De novo contigs: trimmed reads were assembled de novo with the following parameters: similarity index = 0.8; length fraction = 0.8; insertion/deletion cost = 3; mismatch cost = 2. Contigs were screened for SSR regions using the search function described above. To develop and optimize a suite of SSR markers for cultivar identification and gene flow studies, primers were designed for 48 loci, 24 for each method using a batch function in Primer3 version 2 (Rozen and Skaletsky, 2000) specifying a primer melting temperature (T_m) range 58–70°C, maximum T_m difference 5°C, and primer GC content 40–60%. To minimize the cost of primer synthesis during the testing phase, one primer from each pair was 5′ modified with an engineered sequence (5′-CCCCCGGGGGCC-3′) to enable the attachment of a third primer that was fluorescently labeled using a two-step PCR protocol (Pacey-Miller and Henry, 2003). Primer pairs were tested for amplification success and polymorphism among 12 DNA samples including eight Mac. integrifolia cultivars and one individual from each Macadamia species. Of the 48 primer pairs tested, six did not amplify and seven produced multiple bands. Of the remaining 35 loci, none were monomorphic, with two or more alleles detected among the 12 test individuals. Primer sequences for these loci are available on request from the author.

Twelve microsatellite loci were selected for further development on the basis of single band amplification, level of polymorphism, and size compatibility for pooled multilocus capillary electrophoresis. The 5′ end of one of each primer pair was fluorescently labeled (Table 1) and the following single-step PCR protocol was used: in 20-μL reaction volumes containing approximately 20 ng DNA template, 0.5 U Platinum Taq (Life Technologies, Carlsbad, California, USA), 2 μL Platinum Taq PCR buffer, 0.1 mM dNTPs, 2 mM MgCl₂, 0.2 μM of each primer, and sterile water to 20 μL. Thermal cycling was conducted in a GeneAmp PCR System 9700 (Life Technologies) with the following conditions: initial denaturation at 94°C for 2 min; followed by 35 cycles of 94°C for 10 s, annealing temperature (T_a) (Table 1) for 10 s, extension at 70°C for 1 min; followed by final extension at 70°C for 5 min. Genotypes were generated using an ABI PRISM 3730 Genetic Analyzer (Applied Biosystems, Foster City, California, USA). Allele size was scored in reference to ABI PRISM GS FLX internal size standards using the program Geneious version 6.1.6 (Biomatters Ltd., Auckland, New Zealand). We assessed variability and genotype consistency of the 12 loci in 22 macadamia cultivars (two to four replicate trees of each) including pure Mac. integrifolia and hybrids. The loci were also tested for cross-amplification in wild-collected individuals of Mac. integrifolia (n = 6), M. tetraphylla (n = 7), M. ternifolia (n = 2), and M. jansenii (n = 2).

After trimming, there were 245,099,904 reads, with an average length of 91.57 bp. We identified 2.29 million reads containing di- and trinucleotide SSR

Table 2. Genetic properties of 12 microsatellite loci in Macadamia integrifolia and hybrid industry cultivars, and M. tetraphylla.

| Locus  | Macadamia cultivars, Chunes Varietal Trial M2 (n = 22) | M. tetraphylla, northern NSW (n = 7) |
|--------|--------------------------------------------------------|-----------------------------------|
|        | A | H_o | H_e | A | H_o | H_e |
| Mac001 | 5 | 0.545 | 0.676 | 5 | 0.714 | 0.704 |
| Mac002 | 5 | 0.591 | 0.596 | 3 | 0.167 | 0.653 |
| Mac003 | 7 | 0.667 | 0.683 | 5 | 0.857 | 0.714 |
| Mac004 | 7 | 0.364 | 0.762 | 6 | 0.429 | 0.796 |
| Mac005 | 9 | 0.591 | 0.654 | 2 | 0.429 | 0.337 |
| Mac006 | 6 | 0.864 | 0.776 | 9 | 0.857 | 0.847 |
| Mac007 | 6 | 0.682 | 0.653 | 4 | 0.857 | 0.684 |
| Mac008 | 5 | 0.364 | 0.351 | 3 | 0.429 | 0.357 |
| Mac009 | 2 | 0.091 | 0.165 | 2 | 0.143 | 0.133 |
| Mac010 | 8 | 0.909 | 0.702 | 5 | 0.714 | 0.724 |
| Mac011 | 7 | 0.864 | 0.800 | 9 | 0.714 | 0.837 |
| Mac012 | 6 | 0.318 | 0.674 | 6 | 0.571 | 0.796 |
| Mean   | 5.917 | 0.571 | 0.626 | 4.917 | 0.573 | 0.632 |

Note: A = number of alleles; H_o = expected heterozygosity; H_e = observed heterozygosity; n = number of individuals sampled.
The microsatellite markers developed here enable discrimination among macadamia industry cultivars and will be used to select parental genotypes in breeding programs. Cross-amplification and polymorphism of the markers in all Macadamia species will facilitate studies of population structure, gene flow, and hybridization. In this work, we demonstrate the effectiveness of Illumina NGS paired-end sequence reads for rapid and cost-effective microsatellite development with and without prior assembly of reads.

CONCLUSIONS

The microsatellite markers developed here enable discrimination among macadamia industry cultivars and will be used to select parental genotypes in breeding programs. Cross-amplification and polymorphism of the markers in all Macadamia species will facilitate studies of population structure, gene flow, and hybridization. In this work, we demonstrate the effectiveness of Illumina NGS paired-end sequence reads for rapid and cost-effective microsatellite development with and without prior assembly of reads.

Fig. 1. Principal coordinate cluster plot based on genetic distance among multilocus genotypes for Macadamia integrifolia (white), M. tetraphylla (red), M. ternifolia (blue), M. jansenii (yellow), and macadamia cultivars (black). First and second coordinates explain 35.19% and 11.36% of the variation, respectively.

LITERATURE CITED

Costello, G., M. Gregory, and P. Donatii. 2009. Southern Macadamia species recovery plan. Report to the Department of Environment, Water, Heritage and the Arts, Canberra, by Horticulture Australia Limited, Sydney, Australia.

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

Hardner, C. M., C. Peace, A. J. Lowe, J. Neal, P. Pisani, M. Powell, A. Schmidt, et al. 2009. Genetic resources and domestication of Macadamia. Horticultural Reviews 35: 1–125.

 Mast, A. R., C. L. Willis, E. H. Jones, K. M. Downs, and P. H. Weston. 2008. A smaller Macadamia from a more vagile tribe: Inference of phylogenetic relationships, divergence times, and diaspore evolution in Macadamia and relatives (tribe Macadamieae; Proteaceae). American Journal of Botany 95: 843–870.

Neal, J. M. 2008. The impact of habitat fragmentation on wild Macadamia integrifolia Maiden and Betche (Proteaceae) population viability. Ph.D. dissertation, University of New England, Armidale, Australia.

Pacey-Miller, T., and R. J. Henry. 2003. Single-nucleotide polymorphism detection in plants using a single-stranded pyrosequencing protocol with a universal biotinylated primer. Analytical Biochemistry 317: 166–170.

Peakall, R., and P. E. Smouse. 2006. GenAlEx 6: Genetic analysis in Excel. Population genetic software for teaching and research. Molecular Ecology Notes 6: 288–295.

Peakall, R., and P. E. Smouse. 2012. GenAlEx 6.5: Genetic analysis in Excel. Population genetic software for teaching and research—An update. Bioinformatics (Oxford, England) 28: 2537–2539.

Rozen, S., and H. Skaltsky. 2000. Primer3 on the WWW for general users and for biologist programmers. In S. Misener and S. A. Krawetz
### APPENDIX 1. Voucher information for Macadamia species used in this study.

| Species                  | Voucher specimen accession no. | Collection locality                     | Geographic coordinates |
|--------------------------|--------------------------------|----------------------------------------|------------------------|
| *M. jansenii*            | PHARM-13-0809                  | Bulburin National Park, Queensland, Australia | 24°37.584'S, 151°33.291'E |
| *M. tetraphylla*         | PHARM-13-0810                  | Mullumbimby, northern New South Wales, Australia | 28°32.835'S, 153°25.455'E |
| *M. ternifolia*          | PHARM-13-0811                  | Draper, Queensland, Australia           | 27°21.268'S, 152°54.965'E |
| *M. integrifolia*        | PHARM-13-0812                  | Villeneuve, Queensland, Australia       | 26°58.384'S, 152°38.899'E |
| *M. integrifolia*, cultivar 741 | PHARM-13-0813                  | Clunes Varietal Trial M2, New South Wales, Australia | 28°43.844'S, 153°23.699'E |

*Vouchers deposited at Southern Cross University, Medicinal Plant Herbarium (PHARM), Lismore, New South Wales, Australia.*