GMR

Discovery and characterization of new microsatellite loci in *Dipteryx alata* Vogel (Fabaceae) using next-generation sequencing data

R.A. Guimarães[1], M.P.C. Telles[1,2], A.M. Antunes[1], K.M. Corrêa[1], C.V.G. Ribeiro[1], A.S.G. Coelho[3] and T.N. Soares[1]

[1]Laboratório de Genética & Biodiversidade, Instituto de Ciências Biológicas, Universidade Federal de Goiás, Goiânia, GO, Brasil
[2]Escola de Ciências Agrárias e Biológicas, Pontifícia Universidade Católica de Goiás, Goiânia, GO, Brasil
[3]Escola de Agronomia, Universidade Federal de Goiás, Goiânia, GO, Brasil

Corresponding authors: R.A. Guimarães / T.N. Soares
E-mail: rejanearaujog@hotmail.com / tnsoares@gmail.com

Genet. Mol. Res. 16 (2): gmr16029639
Received February 2, 2017
Accepted March 7, 2017
Published April 28, 2017
DOI http://dx.doi.org/10.4238/gmr16029639

ABSTRACT. The use of next-generation sequencing (NGS) technologies provides a great volume of genome sequence data even for non-model species. The development of microsatellite markers using these data is a relatively quick and easy process. *Dipteryx alata* Vogel (Fabaceae) is an arboreal species from the Cerrado biome and is considered an important plant genetic resource. Here, we report the development of microsatellite markers for *D. alata* using NGS data. DNA samples from four individuals were sequenced using the Illumina MiSeq platform and high-quality reads were assembled into contigs of the *D. alata* genome sequence. Microsatellite regions were identified using the IMEX webserver and primer pairs were designed using the Primer3 software. The amplification settings for each locus were optimized. Fluorescent-labeled primers were developed and used to genotype individuals derived from three natural populations of *D.*

alata. Fifty-four microsatellite regions were identified, from which 27 were elected to primer design. Among the amplified loci, 11 were polymorphic, with the number of alleles ranging from 2 to 10. The expected heterozygosity under Hardy-Weinberg Equilibrium (HWE) per locus varied from 0.191 to 0.807. Genotype and allele frequencies for all loci agreed with those expected under HWE and linkage disequilibrium was not significant for all pairs of loci. The probabilities of exclusion of paternity and of combined identity were equal to 0.993 and 5.65 x 10^{-8}, respectively. The markers developed in this study are useful to several types of population genetic studies with *D. alata* and, eventually, for closely related species.

Key words: Baru; Cerrado; Illumina MiSeq; SSR

INTRODUCTION

The species *Dipteryx alata* Vogel belongs to the Fabaceae family and is popularly known as Baru. Baru is widely used in the food and pharmaceutical industry, thus presenting good economic potential (Pineli et al., 2015). This economic value makes *D. alata* a very important genetic resource (Sano et al., 2004).

The availability of microsatellite markers for *D. alata* enables the development of population genetic studies with the species and contributes to increase our knowledge about microevolutionary processes in the Cerrado biome. To date, only a small number of microsatellite markers are available for *D. alata*, each presenting a low level of polymorphism, restricting the type of studies that can be performed (Tarazi et al., 2010; Melo et al., 2011; Soares et al., 2012).

The development of microsatellite markers has traditionally been a difficult and expensive process. Next-generation sequencing technologies (NGS) enable efficient identification of large numbers of microsatellites at a fraction of the cost and effort compared to traditional approaches (Zalapa et al., 2012). In the present study, we report the development of new microsatellite markers for *D. alata* from NGS data, increasing the set of polymorphic microsatellite markers available for the species.

MATERIAL AND METHODS

DNA was extracted from leaf tissues from four *D. alata* individuals, using the CTAB (2%) protocol described by Doyle and Doyle (1987). The four DNA samples were used in the preparation of four barcoded libraries following the Nextera Illumina protocol (Illumina, Inc.). Quantification and validation procedures for each library were performed in Bioanalyzer 2100 (Agilent Technologies) using the High Sensitivity DNA kit, and by real-time PCR using the Kappa kit. The libraries were pooled and sequenced on the Illumina MiSeq platform using the MiSeq Reagent Kit v3 (PE 2 x 300 bp).

Quality control of sequencing data was performed using FastQC (Andrews, 2010) and Trimmomatic (Bolger et al., 2014). Contigs were assembled using MaSuRCA (Zimin et al., 2013). Contigs were submitted to the IMEX webserver (Imperfect Microsatellite Extractor Webserver) (Mudunuri and Nagarajaram, 2007) to identify perfect and imperfect

(<10%) microsatellite regions in the *D. alata* genome. Critical size parameters considered in the search for microsatellites regions were: for dinucleotides, a minimum of 20 tandem repeats; for tetranucleotides, pentanucleotides, and hexanucleotides, a minimum of 8 tandem repeats. Mononucleotide and trinucleotide microsatellites were removed from the analysis. Mononucleotide microsatellite loci present high genotyping errors and trinucleotides are more abundant in coding regions of the genome (Tóth et al., 2000).

Primer pairs were designed using the Primer3 software (Rozen and Skaletsky, 2000) with the following parameters: amplicon size between 100 and 400 bp, primer length between 20 and 24 bp, melting temperature between 57° and 63°C, and CG content between 40 and 80%.

Designed primer pairs were tested for amplification in PCRs with different annealing temperatures using *D. alata* DNA samples. The reactions were assembled to a final volume of 10 µL using 7.5 ng DNA and 0.22 µM primers (forward + reverse), 0.23 µM dNTPs, 3.25 mg BSA, 1X buffer (10 mM Tris-HCl, pH 8.3, 50 mM KCl, 1.5 mM $MgCl_2$), 0.75 U Taq DNA polymerase. The amplification process followed the steps of initial denaturation at 94°C for 5 min; 30 cycles of 94°C for 1 min, 58°-62°C (depending on primer) for 1 min, and 72°C for 1 min; and a final extension of 72°C for 45 min.

Forward sequences of selected primer pairs were labeled with one of the four fluorescent dyes (VIC, NED, 6-FAM, or PET). Lengths of the amplification products were determined using the internal marker GeneScan 600 LIZ (Applied Biosystems), in the ABI PRISM® 3500 DNA Genetic Analyzer (Applied Biosystems). Microsatellite loci that presented a good pattern in capillary electrophoresis were arranged in multiplex panels for routine analysis.

The levels of polymorphism of the developed markers were evaluated in 72 individuals, derived from three natural populations of *D. alata*. The sampled natural populations are from three geographically distant municipalities in Brazil: Pirenópolis, GO (15°59'82.9"S, 49°02'06.9"W), Sonora, MS (17°51'18.1"S, 54°42'21.1"W), and Alvorada, TO (12°26'92.8"S, 49°06'88.0"W).

After PCR amplification and capillary electrophoresis, genotypes were called using GeneMapper (Applied Biosystems). Genetic diversity of each locus in each population was evaluated using the following parameters: average number of alleles per locus and per population (N_A), observed heterozygosity (H_O), expected heterozygosity (H_E) under Hardy-Weinberg equilibrium (HWE). Intrapopulation fixation index (F_{IS}) was also estimated for each locus in each population. The significance of genotypic linkage disequilibrium estimates for all pairs of loci was also evaluated. These analyses were performed in the FSTAT 2.9.3 program (Goudet, 2002). The power of individual discrimination for each locus and for the total set of loci was evaluated by estimating the probability of genetic identity (I) and the probability of exclusion of paternity (Q), with the aid of the IDENTITY 4.0 program (Wagner and Sefc, 1999).

RESULTS

The assembled contigs comprised 27 Mb of the genomic sequence of *D. alata*. Using the IMEx webserver, 46 imperfect and 8 perfect microsatellite regions were identified. The most frequent identified class was imperfect dinucleotide microsatellites (56.52%), followed by microsatellites with tetranucleotide (28.26%), pentanucleotide (4.34%), and hexanucleotide (10.86%) repeat motifs. Twenty-seven microsatellite regions with favorable characteristics for designing primers were obtained. Among the 27 designed pairs of primers, 12 presented a good amplification pattern, although only 11 were polymorphic (Table 1). The overall probability of exclusion and probability of identity was 0.993 and 5.65 x 10^{-8}, respectively. The mean number of alleles per locus was 4.64, ranging from 2 to 10 alleles. The average estimates of H_O and H_E under HWE were 0.348 and 0.515, respectively. The deviations from the expected genotype proportions under HWE given the allele frequencies were not significant for any loci. Using the Bonferroni criteria, no linkage disequilibrium were detected for any pair of loci in the three populations.

The 11 polymorphic microsatellite markers were arranged in three multiplex panels (Table 2) for capillary electrophoresis for the population analysis. The mean N_A per population ranged from 2.9 to 3.8. Estimates of expected heterozygosity under HWE conditions for each population ranged from 0.363 to 0.469. A significant value (0.301) for the F_{IS} was found only for the population of Alvorada, TO (Table 3).

Table 1. Genetic parameters estimates for the 11 microsatellite markers, developed for *Dipteryx alata*.

Marker	GenBank accession No.	Primer sequence (5'→3')	Motif	Ta (°C)	N_A	H_E	H_O	Q	I
Dal11	KX427119	F: CATTTGCCCCTTCTTGTCTTT R: AATGCCGATAATTTGTGTTGTTC	$(TC)_{22}$	62	2	0.407	0.186	0.332	0.220
Dal12	KX427121	F: TGCTGCGTTCATTTTATAGTTTT R: CTTTCTTCTTTGGGAGTTTGCT	$(TC)_{22}$	58	4	0.445	0.127	0.201	0.376
Dal14	KX427125	F: CATCTCACCAAAGCCATACAGA R: TTGTTGCTTCCCGTTTTCTC	$(AG)_{21}$	58	5	0.473	0.429	0.279	0.295
Dall5	KX427115	F: CAACTCCATCAACCAATACACC R: AAATGAACCCCTCCAACACTT	$(CAGGCA)_9$	58	6	0.463	0.361	0.267	0.321
Dal16	KX427122	F: AATTGCGAGGCACAAAAACT R: TGAATGATAATGGGGGCAAA	$(TC)_{22}$	58	4	0.721	0.322	0.560	0.086
Dal18	KX427124	F: CGATAAGCATTACTATTTCCCTTT R: GTGATTGACATCTAACCTCCTCT	$(AG)_{22}$	58	6	0.609	0.592	0.376	0.196
Dal20	KX427118	F: CACGACTAGGAACCCCTTATTTT R: TCTTTGTCATCCTTTCCCTTTG	$(TC)_{27}$	58	5	0.191	0.174	0.141	0.563
Dal21	KX427120	F: TCTAGCCTCAACACACTGCTTC R: CAAGAAAGATGATATGGGAAAAGG	$(TC)_{21}$	62	4	0.530	0.250	0.339	0.227
Dal23	KX427116	F: TAGACAAAGTGCTTGGGGAAA R: TTCTTGATTTTTGGATCTCTATCG	$(AG)_{23}$	58	10	0.807	0.772	0.672	0.047
Dal24	KX427123	F: ATTTTGAGGAAGTCTCTTTGGT R: TGGAGTTCATATCCTTATCTTTG	$(AC)_{22}$	58	2	0.396	0.358	0.230	0.344
Dal25	KX427117	F: CCAGGTGGTGGGATGAGATA R: ATGATGGACACACGAAAGTGAA	$(ATTGAG)_5$	58	3	0.595	0.261	0.364	0.199

Ta: Annealing temperature; N_A: number of alleles; H_E: expected heterozygosity under HWE; H_O: observed heterozygosity, Q: probability of exclusion paternity; I: probability of genetic identity.

Table 2. Panels of microsatellite markers for *Dipteryx alata*, arranged in multiplex for genotyping in capillary electrophoresis.

Multiplex	Markers	Dye	Allele size range (bp)
1	*Dal15*	6FAM	190-214
	Dal23	NED	134-172
	Dal25	VIC	152-164
	Dal20	PET	184-208
2	*Dal11*	NED	162-170
	Dal12	6FAM	188-194
	Dal16	VIC	174-190
	Dal21	6FAM	158-162
3	*Dal14*	6FAM	164-172
	Dal18	VIC	186-198
	Dal24	VIC	124-128

Table 3. Genetic characterization of 11 microsatellite loci in three natural populations of *Dipteryx alata* (n = 3 x 24).

Loci	Pirenópolis-GO				Sonora-MS				Alvorada-TO			
	N_A	H_E	H_O	F_{IS}	N_A	H_E	H_O	F_{IS}	N_A	H_E	H_O	F_{IS}
Dal15	3	0.228	0.167	0.273	6	0.665	0.542	0.189	3	0.414	0.375	0.096
Dal23	4	0.683	0.739	-0.084	8	0.810	0.938	-0.163	3	0.646	0.667	-0.033
Dal25	2	0.496	0.250	0.502	3	0.318	0.208	0.350	3	0.528	0.333	0.375
Dal20	3	0.228	0.250	-0.100	1	0.000	0.000	0.000	5	0.327	0.273	0.168
Dal11	2	0.413	0.222	0.469	2	0.480	0.067	0.865*	2	0.268	0.300	-0.125
Dal12	2	0.043	0.043	0.000	4	0.264	0.292	-0.107	2	0.042	0.042	0.000
Dal16	3	0.412	0.300	0.276	3	0.513	0.455	0.116	4	0.619	0.176	0.721*
Dal21	4	0.241	0.087	0.644	3	0.616	0.500	0.192	3	0.528	0.143	0.735*
Dal14	3	0.377	0.375	0.005	5	0.673	0.864	-0.291	3	0.082	0.083	-0.011
Dal18	5	0.382	0.391	-0.026	5	0.649	0.792	-0.226	4	0.665	0.583	0.125
Dal24	2	0.488	0.625	-0.287	2	0.169	0.182	-0.077	2	0.438	0.238	0.462
Average	2.9	0.363	0.314	0.138	3.8	0.469	0.440	0.063	3.0	0.414	0.292	0.301*

N_A: number of alleles; H_E: expected heterozygosity under HWE; H_O: observed heterozygosity; F_{IS}: intrapopulation fixation index. *Statistically different from 0.000.

DISCUSSION

Several studies have demonstrated success in the development of microsatellite markers for native Cerrado species (Croft and Schaal, 2012; Soares et al., 2012; Bernardes et al., 2014). However, these studies use methodologies that limit the quantity and quality of microsatellite markers obtained. In the present study, using NGS data, 54 microsatellites were detected in a relatively quick and easy way. From the 54 pre-selected microsatellites, further selection was made based on favorable characteristics for design and amplification, resulting in 11 polymorphic microsatellite markers. This short set of markers showed a diversity of repeat motifs, dinucleotides being the most frequent, but also tetra-, penta-, and hexanucleotides. The combined probability of exclusion and of identity showed that the set of developed microsatellite loci has a high power of exclusion of false paternities (Evett and Weir, 1998) and is suitable for individual discrimination (Paetkau et al., 1995).

The polymorphism found in this study for *D. alata* was similar to the observed by Tarazi et al. (2010), Melo et al. (2011), and Soares et al. (2012), who found an average N_A per locus of 4.7, 2.7 and 4.3, respectively. The genetic diversity estimates observed in this study is also similar to other studies with *D. alata* populations. The increase in the number of available markers will allow the use of a wider set of informative markers in more powerful genetic studies. Of the 27 pairs of primers designed in the present study, 11 were polymorphic in *D. alata*. It is suggested that the set of 27 primers be tested for transferability to other Cerrado species of the Fabaceae family, further reducing the costs in obtaining this kind of markers.

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

Research has been continuously supported by several grants and fellowships to the research network GENPAC (Geographical Genetics and Regional Planning for Natural Resources in Brazilian Cerrado) from CNPq/FAPEG (projects #563839/2010-4 and #201110267000125), by the CERGEN (Núcleo de Excelência em Recursos Genéticos Vegetais do Cerrado - PRONEX/FAPEG/CNPq 07-2012) and our study has also been developed in the context of the National Institutes for Science and Technology (INCT) in Ecology, Evolution and Biodiversity Conservation, supported by MCTIC/CNPq (process #465610/2014-5) and FAPEG, besides support of PPGS CAPES/FAPEG (Public Call #08/2014). R.A. Guimarães, A.M. Antunes, and K.M. Corrêa were supported by fellowships from CAPES. M.P.C. Telles and T.N. Soares are supported by productivity grants from CNPq.

REFERENCES

Andrews S (2010). FastQC a Quality Control Tool for High Throughput Sequence Data. Available at [http://www.bioinformatics.babraham.ac.uk/projects/fastqc/].

Bernardes V, Dos Anjos DE, Gondim SG, Murakami DM, et al. (2014). Isolation and characterization of microsatellite loci in *Byrsonima cydoniifolia* (Malpighiaceae) and cross-amplification in *B. crassifolia*. *Appl. Plant Sci.* 2: 1400016. http://dx.doi.org/10.3732/apps.1400016

Bolger AM, Lohse M and Usadel B (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30: 2114-2120. http://dx.doi.org/10.1093/bioinformatics/btu170

Croft GK and Schaal BA (2012). Development of microsatellite markers in *Byrsonima crassifolia* (Malpighiaceae). *Am. J. Bot.* 99: e111-e113. http://dx.doi.org/10.3732/ajb.1100457

Doyle JJ and Doyle JJ (1987). Isolation of plant DNA from fresh tissue. *Focus* 12: 13-15.

Evett IW and Weir BS (1998). Interpreting DNA evidence: Statistical for forensic scientists. Sinauer Associates, Inc. Publishers, Sunderland.

Goudet J (2002). Fstat (Version 2.9.3.2.): A computer program to calculate F-statistics. *J. Hered.* 86: 485-486. http://dx.doi.org/10.1093/oxfordjournals.jhered.a111627

Melo DB, Diniz-Filho JAF, Oliveira G, Santana LL, et al. (2011). Optimizing sampling efforts for *ex situ* conservation of genetic variability of *Dipteryx alata* Vogel. *BMC Proc.* 5: 18. http://dx.doi.org/10.1186/1753-6561-5-S7-P18

Mudunuri SB and Nagarajaram HA (2007). IMEx: Imperfect Microsatellite Extractor. *Bioinformatics* 23: 1181-1187. http://dx.doi.org/10.1093/bioinformatics/btm097

Paetkau D, Calvert W, Stirling I and Strobeck C (1995). Microsatellite analysis of population structure in Canadian polar bears. *Mol. Ecol.* 4: 347-354. http://dx.doi.org/10.1111/j.1365-294X.1995.tb00227.x

Pineli LLO, Carvalho MV, Aguiar LA, Oliveira GT, et al. (2015). Use of baru (Brazilian almond) waste from physical extraction of oil to produce flour and cookies. *LWT -. Food Sci. Technol. (Campinas)* 60: 50-55.

Rozen S and Skaletsky HJ (2000). Primer3: Bioinformatics Methods and Protocols. In: Methods in Molecular Biology (Krawetz S and Misener S, eds.). Humana Press, New Jersey, 365-386. Available at [http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi].

Sano SM, Ribeiro JF and Brito MA (2004). Baru: Biologia e Uso. Embrapa Cerrados, Brasília, DF, Brazil.

Soares TN, Melo DB, Resende LV, Vianello RP, et al. (2012). Development of microsatellite markers for the neotropical tree species *Dipteryx alata* (Fabaceae). *Am. J. Bot.* 99: e72-e73. http://dx.doi.org/10.3732/ajb.1100377

Tarazi R, Moreno MA, Gandara FB, Ferraz EM, et al. (2010). High levels of genetic differentiation and selfing in the Brazilian cerrado fruit tree *Dipteryx alata* Vog. (Fabaceae). *Genet. Mol. Biol.* 33: 78-85. http://dx.doi.org/10.1590/S1415-47572010005000007

Tóth G, Gáspári Z and Jurka J (2000). Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res.* 10: 967-981. http://dx.doi.org/10.1101/gr.10.7.967

Wagner HW and Sefc KM (1999). IDENTITY 1.0. Centre for Applied Genetics, University of Agricultural Sciences, Vienna. Available at [http://www.uni-graz.at/~sefck].

Zalapa JE, Cuevas H, Zhu H, Steffan S, et al. (2012). Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. *Am. J. Bot.* 99: 193-208. http://dx.doi.org/10.3732/ajb.1100394

Zimin AV, Marçais G, Puiu D, Roberts M, et al. (2013). The MaSuRCA genome assembler. *Bioinformatics* 29: 2669-2677. http://dx.doi.org/10.1093/bioinformatics/btt476