Short communication

Molecular insight for baru *Dipteryx alata* (Fabaceae) populations based on novel SSRs

Rayleen Whaiti Lopes da Silva[1], Sarah Silva Machado[2], Karina de Cassia Faria[3] (iD), Fernanda Ancelmo de Oliveira[4], Anete Pereira de Souza[4], Ivandilson Pessoa Pinto de Menezes[5] (iD) and Joaquim Manoel da Silva[2, 3]* (iD)

ABSTRACT

Baru tree (*Dipteryx alata*) is an arboreal, fruitful plant native to the Cerrado biome with an important socioeconomic impact. Populations of this species are a good model to study anthropogenic disturbances on the biome through the genetic information. In this study, we developed seven new polymorphic microsatellite markers for *D. alata,* using an enriched genomic library. We performed loci characterization in three populations, obtaining a total of 49 alleles, with an average of 5 to 5.57 alleles per locus. A significant content of polymorphic information was obtained, as indicated by the average expected heterozygosity (*uHE*), with a total average of 0.58 to 0.65 per locus. The average value of the observed heterozygosity (*Ho*) was also high, with a total average of 0.73 to 0.85 per locus. Some of the loci are in linkage disequilibrium, such as DalatG6 with DalatB3, DalatH3 and DalatB4. The estimate of the combined loci for the probability of paternity exclusion obtained an average value of 1.00 for all loci, and the average combined probability of identity, the values were $(1.2^{10^{-5}})$ to $(4.4^{10^{-6}})$. All markers are informative and suitable for studies on genetic diversity and population structure, aiming at the conservation and management of the species.

Keywords: Baru tree, *Dipteryx alata,* genetic variability, molecular markers, microsatellites.

Understanding patterns in which genetic variability is organized in populations is of fundamental importance for the development of conservation and sustainable use strategies for species and their habitats (Hoban *et al.* 2022). *Dipteryx alata* Vog. (Fabaceae), popularly known as baru tree, is a native tree species that has multiple uses ranging from human and animal food, to its pharmaceutical properties (Soares *et al.* 2008). It is an endemic species to the Cerrado biome, and due to its wide distribution throughout the area and abundance in several of its habitats (Canuto *et al.* 2008), it has potential as a model to check the impact of human intervention on the environment,

1 Universidade Federal de Mato Grosso, Programa de Pós-Graduação em Biotecnologia e Biodiversidade, 78300-000, Tangará da Serra, MT, Brazil
2 Universidade do Estado de Mato Grosso, Programa de Pós-Graduação em Ecologia e Conservação, 78690-000, Nova Xavantina, MT, Brazil
3 Universidade do Estado de Mato Grosso, Faculdade de Ciências Agrárias, Biológicas e Sociais Aplicadas, 78690-000, Nova Xavantina, MT, Brazil
4 Universidade Estadual de Campinas, Instituto de Biologia, Departamento de Biologia Vegetal, 13083-862, Campinas, SP, Brazil
5 Instituto Federal de Educação, Ciência e Tecnologia do Ceará, Departamento de Biologia, 63560-000, Acopiara, CE, Brazil

* Corresponding author: joaquimmanoel@unemat.br

through the characterization of its diversity and genetic structure and the effective size of populations (Collevatti *et al.* 2013). For this, the microsatellite markers (SSR - Single Sequence Repeats) are considered great tools to access the genetic variability of populations due to their codominant inheritance, high degree of polymorphism, multi-allelic and high reproducibility properties (Cosson *et al.* 2014). Thus, this marker provides information on the long-term evolutionary history of species, mutation, isolation of the population, as well as the mechanisms of genetic drift, gene flow and selection (Allendorf 2017).

Despite the importance of *D. alata* for the Cerrado, little is known about the genetic variability in populations (Soares *et al.* 2008; Soares *et al.* 2012; Guimarães *et al.* 2017). Although specific SSR markers are already available (Soares *et al.* 2012; Guimarães *et al.* 2017), these studies show a low level of polymorphism in loci, limiting the type of studies that can be carried out. Thus, in order to increase the coverage of loci representative of the genome, we developed and characterized a set of new microsatellite loci for *D. alata*. These markers have the potential to be used in population studies at various scales, in order to provide subsidies for the conservation and management of the species and its area of occupation.

For that purpose, we construct a microsatellite-enriched genomic library according to the protocol adapted from Billotte *et al.* (1999). An individual of *D. alata* was chosen randomly at the Mário Viana Municipal Park, a Conservation Unit, located in the municipality of Nova Xavantina, state of Mato Grosso, Brazil, and total genomic DNA was extracted from young leaves using the commercial kit Plant DNeasy® (Qiagen). A sample of 5 µg DNA was digested with the restriction enzyme *AfaI* (Invitrogen). The digested fragments were ligated to specific adapters Rsa21 and Rsa25 using T4 DNA ligase and an enriched library was obtained by hybridization of probes with Biotin-IIIIII (CT)8 and Biotin-IIIIII (GT)8 and magnetic beads coated with streptavidin (MagneSphere Magnetic Separation Products; Promega Corporation, Madison, Wisconsin, EUA). The captured fragments were amplified by PCR, with the primer adapter Rsa21 (10µM), cloned into a pGEM-T Easy Vector (Promega, Madison, WI, USA), and then inserted into competent *Escherichia coli* cells 10H10b by electroporation. We selected positive clones using white/blue screening, and then we checked the presence and size of cloned inserts through PCR and a 2% agarose gel, then the best 48 clones were sequenced on an ABI 3500xL Genetic Analyzer (Applied Biosystems, Foster City, California, USA) using primers T7 and SP6, and BigDye Terminator version 3.1 Cycle Sequencing Kit (Perkin Elmer – Applied Biosystems).

Microsatellite regions were identified using SSRIT (The Simple Sequence Repeat Identification Tool) (Temnykh *et al.* 2001). The primer pairs were designed using the software primer3Plus (Untergasser *et al.* 2007), with the following parameters: maximum primer size of 25 bp, annealing temperature of the primer (Tm) varying from 52 °C to 65 °C, minimum and maximum GC percentage of 40% and 60%, respectively, and amplification product range between 100 and 700 bp.

The microsatellite markers were characterized using 90 individuals of *D. alata* sampled in three populations in the Cerrado-Amazon transition region (Fig. S1), 30 individuals were collected from each population: 1) "BASC" population - natural population, with a history of anthropization (pasture area), located in the municipality of Nova Xavantina - MT (14° 67' 78.5" S 52° 54' 36.6" W); 2) "BAP" population - it is a conserved population, at the Bacaba Municipal Park (14° 70' 82.7" S 52° 35' 31.9" W); 3) "BAVC" population - is a young population, located in a restoration area carried out by manual direct seeding, located in the municipality of Barra do Garças - MT (14°87' 64.0" S 52°11' 81.4" W) (Fig. S1).

The total genomic DNA was extracted from young leaves using the cetyltrimethylammonium bromide (CTAB) method (Doyle 1991). PCR cocktail (20 µL) contained, 2.0 µL 10x PCR *Buffer*, 0.60 µL $MgCl_2$ (50 mM), 1.6 µL DNTPs (0.2 mM each), 0.40 µL each *primer* (0.20 µM), 1 U Taq DNA Polymerase, 4.0 µL DNA (5ng/µL) and 10.8 µL milliQ H_2O. To optimize the annealing temperature of the primer pairs, the amplification of fragments was performed in a Biocycler® thermocycler, following the steps described by (Soares *et al.* 2012). The loci were subjected to the multiplex system, in which the forward primer was labeled with one of the following fluorescences (FAM, NED, VIC and ATTO565) and analyzed in an ABI 3500 automated sequencer (Applied Biosystems, Foster City, CA, USA).

For genotyping the loci, alleles were scored against the GeneScan-600 (LIZ) internal Size Standard Kit (Applied Biosystems, Foster City, CA, USA) using the Geneious 8.1.6 software (Kearse *et al.* 2012). Descriptive statistics were run in the GenAlex 6.5 software (Peakall and Smouse 2012), in which the estimates of number of alleles per locus (N_A), observed heterozygosity (*Ho*), expected heterozygosity (uH_E) and fixation index (*f*) and the probability of identity (*PI*) and probability of paternity exclusion (*PE*) were evaluated. The FSTAT 2.9.3.2 software (Goudet 2001) was used to analyze the linkage disequilibrium between the loci.

Out of the 48 clones sequenced, 29% sequences contained microsatellite regions. The microsatellites found in these sequences were classified as perfect microsatellites, with dinucleotide-type repeat motifs, this motif is present in 100% microsatellites found in the library.

From the primer pairs designed, a set of 10 loci were selected for synthesis and validation (Tab. 1). Of these, seven loci were polymorphic and three were monomorphic in the evaluated populations. Similar parameters in the percentage of sequences obtained with genomic library construction was also observed in the protocol of Billotte *et al.* (1999), which varied from 20 to 90% for tropical plant species. However, based on genetic diversity studies for *D. alata* (Soares *et al.* 2008; Soares

et al. 2012; Guimarães *et al.* 2017), this is the first to develop SSR markers using the microsatellite-enriched genomic library technique.

A total of 49 alleles were identified at the seven polymorphic loci, with the number of alleles per locus ranging from two at the locus (DalatB3) to twelve at the locus (DalatB4) (Tab. 2). The total average of alleles in the three populations ranged from 5 to 5.57 alleles per locus (Tab. 2). The observed heterozygosity ranged from 0.00 to 1.0 per loci depending on the sampled population, and the expected heterozygosity ranged from 0.34 to 0.74 per loci, with total averages ranging from 0.73 to 0.85 for H_O and from 0.58 to 0.65 for uH_E (Tab. 2). The observed heterozygosity (H_O) was greater than the expected heterozygosity (uH_E) for most loci, except for loci (DalatB3 and DalatB4). As a consequence, the fixation index (*f*) was negative for most loci, except for loci (DalatB3 and DalatB4), with total averages ranging from -0.15 to -0.28 (Tab. 2).

Table 1. Characteristics of the ten microsatellite loci developed for *D. alata*.

Loci	Primer Sequences (5′–3′)	Repeat Motif	Allelic range	Ta (°C)	Dye	Multiplex	GenBank Accession Number
DalatA1	F:GCCTTATCCGCAAACTAACG R: AAGCCATTTGTTGCAACTGA	$(GA)_{10}$	406	60 °C	FAM	A	MZ229345
DalatB3	F:TGCAACTTGAAACACGTCTT R: ACAAGTCGCAAAGGTTTCTT	$(GT)_8$	528-530	58 °C	VIC	A	MZ229346
DalatB4	F:CGTTTGCACACACTTAAAAC R: GGGCTGAATTTGAAAGCACT	$(AC)_7$	554-586	60 °C	NED	A	MZ229347
DalatB5	F:TAGACAGAGACCCAGAGACC R: TCTGGTGAGCTTTTATGTGT	$(GT)_{10}$	362-382	60 °C	ATTO565	A	MZ229348
DalatC3	F:ACCTATATGCTGCTAGCTCA R: GTCATGACCCTAAGCTTTAGC	$(GT)_7$	390-424	60 °C	VIC	B	MZ229349
DalatD6	F:TGGGGAGTTTTGAATTTGGTG R: ACAAGACAGCTGCATAGAAAG	$(TG)_7$	501	60 °C	NED	B	MZ229350
DalatF3	F:AGAAAGCAACATGAGTCTGTG R: ACATATCTGCCATTACGCCT	$(CT)_{14}$	558	60 °C	ATTO565	B	MZ229351
DalatF6	F:ACCACCTGTTACACTTGCAA R: TCCACTTCCATCACTTCTCT	$(AC)_{12}$	288-310	58 °C	FAM	C	MZ229352
DalatG6	F:TTGGGTGATTTTGATGGTTGG R: GTTCTACCCTCTGACACTCG	$(TG)_5$	480-506	60 °C	VIC	C	MZ229353
DalatH3	F:TGGGTGATTTTGATGGTTGG R: CACCTCTACCTCCCCTAGAG	$(TG)_5$	512-538	60 °C	NED	C	MZ229354

Ta(°C) = Annealing temperature set after optimization tests.

Table 2. Genetic parameters of microsatellite loci determined in three populations of *D. alata*.

Loci	Santa Célia Farm (BASC)							Bacaba Park (BAP)							Vera Cruz Farm (BAVC)						
	N	N_A	H_O	uH_E	*f*	*PI*	*PE*	*N*	N_A	H_O	uH_E	*f*	*PI*	*PE*	*N*	N_A	H_O	uH_E	*f*	*PI*	*PE*
DalatF6	29	5	0.93	0.72	-0.31	0.13	0.65	30	6	0.97	0.74	-0.33	0.12	0.67	30	8	0.97	0.71	-0.38	0.13	0.68
DalatC3	30	7	1.00	0.72	-0.41	0.12	0.69	29	4	1.00	0.67	-0.52	0.17	0.59	30	7	0.97	0.73	-0.34	0.12	0.69
DalatG6	30	4	1.00	0.66	-0.54	0.18	0.57	30	3	1.00	0.57	-0.79	0.29	0.39	30	3	1.00	0.63	-0.60	0.21	0.49
DalatB3	26	2	0.00	0.40	1.00	0.45	0.25	26	2	0.00	0.14	1.00	0.75	0.12	26	2	0.00	0.46	1.00	0.40	0.27
DalatH3	29	5	1.00	0.63	-0.60	0.21	0.51	29	5	1.00	0.66	-0.54	0.18	0.56	30	5	1.00	0.67	-0.52	0.17	0.59
DalatB4	30	8	1.00	0.71	-0.43	0.12	0.70	30	9	0.83	0.66	-0.28	0.15	0.66	26	5	0.31	0.34	0.09	0.45	0.33
DalatB5	30	8	1.00	0.70	-0.45	0.14	0.64	30	6	0.93	0.63	-0.51	0.20	0.54	29	8	0.90	0.71	-0.28	0.13	0.68
Mean	29.14	5.57	0.85	0.65	-0.25	$4.4^{10^{-6}}$*	1.00*	29.14	5.00	0.82	0.58	-0.28	$2.4^{10^{-5}}$*	1.00*	28.71	5.43	0.73	0.61	-0.15	$1.2^{10^{-5}}$*	1.00*
SE	0.55	0.84	0.14	0.04	0.21	-	-	0.55	0.87	0.14	0.07	0.22	-	-	0.71	0.90	0.15	0.06	0.21	-	-

N= number of individuals analyzed, N_A = number of alleles observed, H_O = observed heterozygosity, uH_E = expected heterozygosity, *f* = fixation index, *PI* = probability of genetic identity, *PE* = probability of paternity exclusion and means of combined analysis (*)

The variation in the number of alleles per locus is common in microsatellite regions due to the high levels of polymorphism found in these markers (Ellegren 2004), which is directly related to the high mutation rates occurring in these regions (Gao *et al.* 2013). The $H_O > uH_E$ relationship for most loci suggests an excess of heterozygotes in relation to that expected by the Hardy-Weinberg equilibrium. This excess of heterozygotes may be occurring due to the reproductive characteristics of the species or even due to a selection in favor of heterozygotes for *D. alata* (Gomes and Moura 2010), causing a lack of inbreeding for most loci (*f* values close to or less than zero). This result reveals the mixed mating system with a predominance of outcrossing found for *D. alata* (Tambarussi *et al.* 2017).

The probability of identity (*PI*) of each locus was estimated, which ranged from 0.12 (DalatB4, DalatC3 and DalatF6) to 0.75 (DalatB3), with average estimates of the combined analysis ranging from $(1.2^{10^{-5}})$ to $(4.4^{10^{-6}})$. The probability of paternity exclusion (*PE*), based on the seven pairs of microsatellite loci, ranged from 0.12 for locus (DalatB3) to 0.70 for locus (DalatB4), in the combined analysis, where no parent is known, the average value was 1.00 for all loci (Tab. 2). The set of microsatellite markers showed a low combined *PI*, indicating the probability of finding, by chance, two individuals from a sample with the same genotype in each set of markers is minimal (Val *et al.* 2020). The *PE* through the combined analysis for the developed SSR loci was high, providing high reliability to correctly exclude an individual from paternity (Rocha *et al.* 2018). The seven SSR loci were sufficient to distinguish between ninety individuals, indicating that the set of microsatellite markers is efficient in discriminating individuals in populations (Val *et al.* 2020).

According to the linkage disequilibrium (LD) test, most loci segregate independently, with no significant association between them, except only for pairs of loci (DalatG6 X DalatB3, DalatG6 X DalatH3, DalatG6 X DalatB4 and DalatB4 X DalatB5), for which the LD was significant even after Bonferroni correction ($\alpha = 0.0023809524$) (Tab. S1). This indicates that some of them were more correlated in the total sample population than would be expected with random crossover. Different biological factors can increase the LD, for example, small population size, low recombination rate, natural or artificial selection, population mixture, among others (Flint-Garcia *et al.* 2003). In our study, these observed LDs suggest they undergo selection, corroborating the revealed excess of heterozygotes (*f* < 0). However, further studies are required to consider such markers in LD as "blocked".

The developed microsatellite markers have an expressive power of discrimination and gene diversity. Therefore, these new loci added to the number of SSR markers available for *D. alata* will contribute to a greater random sampling of the genome and, consequently, estimates of genetic variability with greater precision, more effectively contributing to the conservation of this species.

Supplementary material

The following online material is available for this article:
Figure S1 - Distribution of the sampled populations of *D. alata* in the municipalities of Nova Xavantina and Barra do Garças, state of Mato Grosso.
Table S1 - Linkage disequilibrium between pairs of microsatellite loci in *D. alata*.

Acknowledgements

References

Allendorf FW. 2017. Genetics and the conservation of natural populations: Allozymes to genomes. Molecular Ecology 26: 420–430. doi: 10.1111/mec.13948

Billotte N, Lagoda PJL, Risterucci AM, Baurens FC. 1999. Microsatellite-enriched libraries: applied methodology for the development of SSR markers in tropical crops. Fruits 54: 277–288.

Canuto DSO, Silva AM, Moraes MA, Silva CLSP, Moraes MLT, Sá ME. 2008. Genetic variability of natural populations of *Dipteryx alata* Vog. by nutrients seeds content. Revista do Instituto Florestal 20: 89-97.

Collevatti RG, Telles MPC, Nabout JC, Chaves LJ, Soares TN. 2013. Demographic history and the low genetic diversity in *Dipteryx alata* (Fabaceae) from Brazilian Neotropical savannas. Heredity 111: 97–105. doi: 10.1038/hdy.2013.23

Cosson P, Decroocq V, Revers F. 2014. Development and characterization of 96 microsatellite markers suitable for QTL mapping and accession control in an *Arabidopsis core* collection. Plant Methods 10: 2. doi: 10.1186/1746-4811-10-2

Doyle J. 1991. DNA Protocols for Plants. In: Hewitt GM, Johnston AWB, Young JPW (eds.). Molecular Techniques in Taxonomy. Berlin, Springer. p. 283–293.

Ellegren H. 2004. Microsatellites: Simple sequences with complex evolution. Nature Reviews Genetics 5: 435–445. doi: 10.1038/nrg1348

Flint-Garcia SA, Thornsberry JM, Buckler ES. 2003. Structure of linkage disequilibrium in plants. Annual Review of Plant Biology 54: 357–374. doi: 10.1146/annurev.arplant.54.031902.134907

Gao C, Ren X, Mason AS *et al.* 2013. Revisiting an important component of plant genomes: microsatellites. Functional Plant Biology 40: 645–661. doi: 10.1071/FP12325

Gomes CC, Moura T. 2010. Estrutura genética em populações de plantas do Cerrado. Revista Agrotecnologia 1: 33–51. doi: 10.12971/2179-5959.v01n01a03

Goudet J. 2001. FSTAT, a program to estimate and test gene diversity and fixation indices (version 2.9.3). http://www2.unil.ch/popgen/softwares/fstat.htm.

Guimarães RA, Telles MPC, Antunes AM *et al.* 2017. Discovery and characterization of new microsatellite loci in *Dipteryx alata* Vogel (Fabaceae) using next-generation sequencing data. Genetics and Molecular Research 16: gmr16029639. doi: 10.4238/gmr16029639

Hoban S, Archer FI, Bertola LD *et al.* 2022. Global genetic diversity status and trends: towards a suite of Essential Biodiversity Variables (EBVs) for genetic composition. Biological Reviews 97: 1511–1538. doi: 10.1111/brv.12852

Kearse M, Moir R, Wilson A *et al.* 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28: 1647–1649. doi: 10.1093/bioinformatics/bts199

Peakall R, Smouse PE. 2012. GenAlex 6.5: genetic analysis in Excel. Population genetic software for teaching and research – an update. Bioinformatics 28: 2537–2539. doi: 10.1093/bioinformatics/bts460

Rocha LL, Martinez AM, Delgado JV, Filho MAG, Ribeiro MN. 2018. Painel SRT para teste de paternidade em caprinos. Medicina Veterinária (UFRPE) 12: 52–61. doi: 10.26605/medvet-v12n1-2160

Soares TN, Chaves LJ, Telles MPC, Diniz-Filho JAF, Resende LV. 2008. Distribuição espacial da variabilidade genética intrapopulacional de *Dipteryx alata*. Pesquisa Agropecuária Brasileira 43: 1151–1158. doi: 10.1590/S0100-204X2008000900008

Soares TN, Melo DB, Resende LV *et al.* 2012. Development of microsatellite markers for the neotropical tree species *Dipteryx alata* (Fabaceae). American Journal of Botany 99: 72–73. doi: 10.3732/ajb.1100377

Tambarussi EV, Sebben AM, Alves-Pereira A *et al.* 2017. *Dipteryx alata* Vogel (Fabaceae), a neotropical tree with high levels of selfing: implications for conservation and breeding programs. Annals of Forest Research 60: 243–261. doi: 10.15287/afr.2017.842

Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S. 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. doi: 10.1101/gr.184001

Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JAM. 2007. Primer3Plus, an enhanced web interface to Primer3. Nucleic Acids Research 35: 71–74. doi: 10.1093/nar/gkm306

Val ADB, Breves SS, Cançado GMA, Ferreira JL, Pasqual M. 2020. Use of molecular markers SSR and SCAR for identification of olive accessions. Bioscience Journal 36: 1137–1145. doi: 10.14393/BJ-v36n4a2020-47959