GMR
The Original by FUNPEC-RP

Large number of repetitive elements in the draft genome assembly of *Dipteryx alata* (Fabaceae)

A.M. Antunes [1], R. Nunes [1], E. Novaes [2], A.S.G. Coelho [3], T.N. Soares [1] and M.P.C. Telles [1,4]

[1] Laboratório de Genética & Biodiversidade, Instituto de Ciências Biológicas, Universidade Federal de Goiás, Goiânia, Goiás, Brasil
[2] Setor de Genética e Melhoramento de Plantas, Departamento de Biologia, Universidade Federal de Lavras, Lavras, Minas Gerais, Brasil
[3] Laboratório de Genética e Genômica de Plantas, Escola de Agronomia, Universidade Federal de Goiás, Goiânia, Goiás, Brasil
[4] Escola de Ciências Agrárias e Biológicas, Pontifícia Universidade Católica de Goiás, Goiânia, Goiás, Brasil

Corresponding author: A.M. Antunes
E-mail: adrianaantunesbio@gmail.com

Genet. Mol. Res. 19 (2): gmr18463
Received August 26, 2019
Accepted January 11, 2020
Published May 28, 2020
DOI http://dx.doi.org/10.4238/gmr18463

ABSTRACT. *Dipteryx alata* (Fabaceae), locally known as Baru, is a non-model, native tree species endemic to the Brazilian Savanna (Cerrado), with economic potential due to its use as food, medicine, animal forage, lumber, and in recovery of degraded areas and landscaping. Although *D. alata* is recognized as an important Brazilian resource, currently there is no genomic information for this species. We generated 22 Gb raw reads from the genomes of *D. alata* trees using the Illumina MiSeq platform. These were assembled in 275,707 nuclear genomic sequences (N50 = 1598 bp) with a total of 355 Mb, which corresponds to 44% of the whole genome. We detected 21,981 microsatellite regions, of which 49.3% were dinucleotides and 42.7% trinucleotides. We found 421,701 transposable elements (TEs) in 39.29% of the sequences. Long terminal repeat retrotransposons were the most abundant TEs. This is one of the first genomic scale studies for a native Cerrado species. The results can be used for the development of molecular markers for studies on evolution, population genetics and conservation of *D. alata*.

Key words: Baru; Conservation; Genomic Resources; Microsatellites; Transposable Elements

INTRODUCTION

The Cerrado is a savannah-type biome that occurs in Brazil and has high species richness and endemism. As a biodiversity hotspot, the Cerrado contains many species that are threatened due to the fragmentation of habitat and alteration of the natural landscape with agricultural expansion. The neotropical and native species of the Cerrado include *Dipteryx alata*, a species belonging to family Fabaceae (formerly Leguminosae) popularly known as Baru. It is a large tree with an average mature height of 15 meters and a maximum 25 meters. The species is allogamous and pollination is mainly performed by bees (Oliveira and Sigrist, 2008).

Baru has a wide geographical distribution in the Brazilian Cerrado, being found in the states of São Paulo, Minas Gerais, Mato Grosso, Mato Grosso do Sul, Tocantins and Goiás, as well as in Paraguay and Bolivia (Ratter, 2000). In Brazil *D. alata* is found mainly in seasonal savannah habitats and growing in eutrophic and drained soils, and in plantations, where the species provides an alternative income for agroextractivists. It is planted commercially in some locations in Brazil and can produce on average 850 kg of seeds and 19 tons of fruit pulp per hectare (Ribeiro et al., 2000; Arakaki et al., 2009). This tree is of economic importance mainly because of the pulp (mesocarp) and seeds (seeds) of its fruits, which are edible and used to manufacture ice creams, liqueurs, toasted nuts, and other products by small and medium-sized industries. The seeds have a pleasant taste and are rich in protein, essential minerals, fiber and lipids (Ferreira et al., 2018). In addition, *D. alata* has value in medicine, as a forage plant, in the recovery of degraded areas, in landscaping and as wood (Sano et al., 2004; Soares et al., 2015).

Despite the importance of *D. alata* to local communities in central Brazil, few genetic and genomic investigations of the species have been performed. Currently, only 32 sequences of this species are deposited in the NCBI GenBank database; these are mainly microsatellites and complete or partial chloroplast genes. Therefore, large-scale DNA sequencing strategies based on next-generation sequencing platforms and genome annotation can generate new insights into the biology of *D. alata* (Metzker, 2010). Sequencing and characterizing the genomes of all Earth's eukaryotic biodiversity is currently a scientific challenge. This challenge includes studies of neotropical species that do not have commercial appeal at a global scale (Cheng et al., 2018), such as *D. alata*. Information on the genome of *D. alata* may be used as a key element in formulating appropriate conservation and use. In addition, it may support the development of domestication and breeding programs in the future. Therefore, we sequenced and assembled a draft version of the genome of *D. alata* and subsequently performed annotation of repetitive elements in the genomic sequences.

MATERIAL AND METHODS

DNA extraction and Illumina sequencing

In this work, leaf samples were collected from four mature *D. alata* trees from different regions of the Brazilian Cerrado: Natividade-TO, Camapuã-MS, Aquidauana-

MS and Cáceres-MT. In each region, leaf samples were collected from a specimen of *D. alata* found in the wild. DNA was extracted using the CTAB protocol. The preparation steps of the libraries were carried out following the protocol of the Illumina Nextera™ kit. Libraries were pooled and sequenced with the MiSeq Illumina platform using a MiSeq sequencing kit v3 with 600 cycles (PE - 2 x 300bp). Sequencing was performed in two runs.

Quality analysis and assembly of nuclear genomic sequences

The quality of the reads was evaluated using FastQC software (Andrews, 2010), and low-quality reads, as well as sequences of adapters, were removed using Trimmomatic software (Bolger et al., 2014). Assembly was performed with dipSPAdes v.1.0 software (Safonova et al., 2015). Contaminants sequences, organellar sequences and sequences shorter than 500 bp were excluded from the assembly. The assembly quality was evaluated using the script "assemblathon_stats.pl". The sequencing coverage was estimated using the genomeCoverageBed tool from the BEDTools suite (Quinlan and Hall, 2010).

Annotation of repetitive elements and primer design

Assembled nuclear sequences were investigated to identify microsatellite regions and design primers using QDD v. 3.1.2 software (Meglécz et al., 2009). The minimum number of repetitions for each motif size used to search the microsatellites was 10 for dinucleotide, 6 for tri- and tetranucleotide and 5 for penta- and hexanucleotide microsatellites. The primer design for microsatellite regions was performed using the following parameters: an amplicon size of 150-400 bp, a GC content of 30-60%, a Tm of 56-62°C, and primer lengths of 22-25 bp.

TEs were detected and annotated in the sequences of the nuclear genome of *D. alata* using RepeatMasker and RepeatModeler software for similarity-based and *de novo* annotation, respectively (Tarailo-Graovac and Chen, 2009). Gypsy and Copia LTR retrotransposons were selected for primer design, aiming to aid the future development of retrotransposon-based insertion polymorphism (RBIP) markers for *D. alata*. Primer design was performed with the web version of Primer3 software, using the following parameters: a GC content of 30-60%, a Tm of 52-62°C, and primer lengths of 20-25 bp (Rozen and Skaletsky, 2000).

RESULTS

A total of 22 Gb raw data was generated that corresponded to 61.62 million paired-end reads. After quality control, 60.55 million reads (14.6 Gb) were retained for genome assembly. The reads were assembled into 275,707 scaffolds with an N50 equal to 1,598 bp. The size of the largest and smallest scaffold and the total size of the scaffolds were 87,189, 500 and 363,921,722 bp, respectively (Figure 1). Among the assembled scaffolds, 137,850 were greater than 1k nucleotides in length, and 140

scaffolds were greater than 10k nucleotides in length. The final assembled sequence comprised 355 Mb of the nuclear genome of *D. alata*, which corresponded to 44% of the flow cytometry-determined genome size, which was estimated to be 807 Mb (unpublished data). The coverage of the assembly was estimated to be 12.78X.

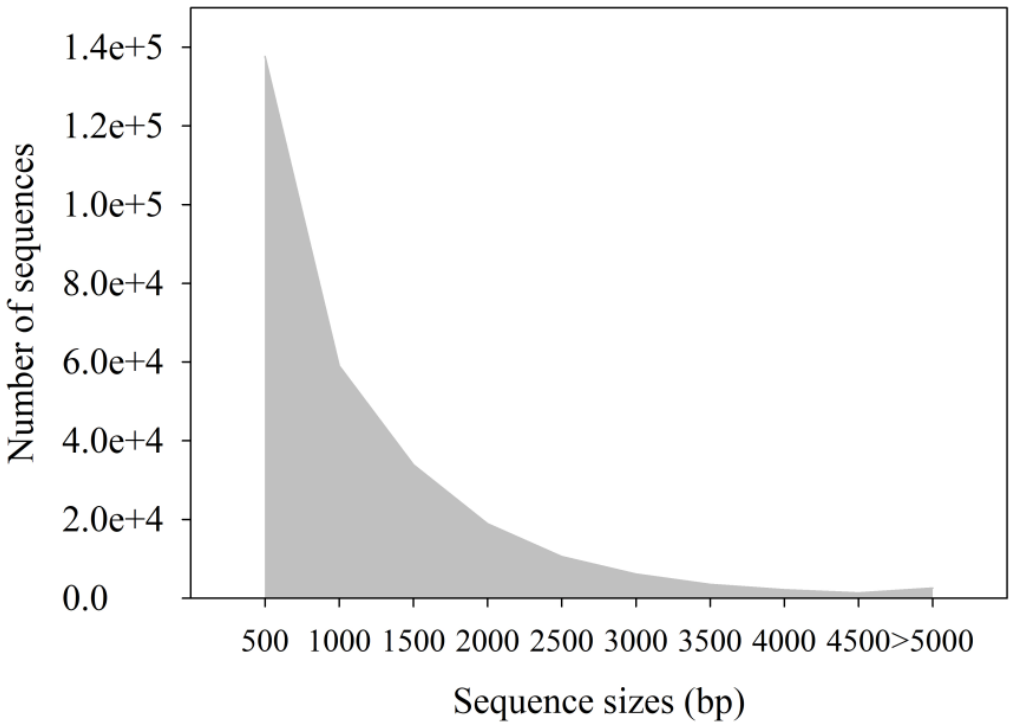

Figure 1. Distribution of scaffold sizes of the nuclear genome" assembly of *Dipteryx alata* obtained from the Illumina sequencing data.

A total of 21,981 microsatellite regions were identified, 49.3% (10,832) of which were composed of dinucleotides; 42.7% (9,389), of trinucleotides; 4% (887), of tetranucleotides; 2.7% (589), of pentanucleotides; and 1.3% (284), of hexanucleotides. The most common repeat motif was "AT" (Table 1). Considering all of the identified microsatellite regions and all sizes of motifs, the most frequent motifs were the dinucleotides AT (24%), AG (8.6%) and CT (8.3%); the trinucleotides AAT (10.9%), TTA (10.5%), and AAG (5.5%); the tetranucleotides TATT (0.75%), AAAT (0.73%) and CTTT (0.45%); the pentanucleotides ATTTT (0.4%), AAAAT (0.3%) and CTTTT (0.3%); and the hexanucleotides AAAAAT (0.10%), TATTTT (0.08%) and AAAAAC (0.06%). The microsatellite motifs were tandemly repeated between 5 and 29 times (Figure 2). The identified microsatellites totaled 516,684 bp in length, which corresponded to 0.14% of the sequences of *D. alata*. We designed 31,480 pairs of

primers for 8,332 microsatellite regions in *D. alata*, 120 of which were selected for future marker development (Supplementary 1).

Table 1. Frequency of microsatellite repeat motifs annotated in genomic sequences of *Dipteryx alata*.

Motif	Absolute frequency	Relative frequency (%)	Motif	Absolute frequency	Relative frequency (%)
AT	5290	24.06	AATAT	22	0.10
AAT	2396	10.90	AATTT	21	0.09
ATT	2312	10.51	AAATT	20	0.09
AG	1893	8.61	AAAAC	19	0.08
CT	1831	8.32	ATTT	18	0.08
AAG	1203	5.47	GTTT	18	0.08
CTT	1151	5.23	TATTTT	18	0.08
GT	927	4.21	ATCT	17	0,07
AC	885	4.02	GATA	16	0.07
AAC	565	2.57	ATTTG	16	0.07
GTT	543	2.47	GTTTT	15	0.06
ATC	233	1.06	ATTTT	15	0.06
ATG	190	0.86	AAAAAC	15	0.06
TATT	165	0.75	TCTTTT	14	0.06
AAAT	161	0.73	AAAC	13	0.05
CTTT	103	0.46	AAAAGA	13	0.05
AAAG	99	0.45	CTCTT	9	0.04
ACC	89	0.40	CTCCTT	8	0.03
ATTTT	89	0.40	CCTT	7	0.03
ACAT	88	0.40	CAGAA	7	0.03
GGT	87	0.39	CG	6	0.02
CCT	86	0.39	AAGG	6	0.02
CTG	86	0.39	AGTG	6	0.02
AGC	82	0.37	AAATC	6	0.02
AGG	81	0.36	GAGAA	6	0.02
AAAAT	81	0.36	TCTCTT	6	0.02
ATGT	76	0.34	AATC	5	0.02
CTTTT	73	0.33	ATTC	5	0.02
AAAAG	64	0.29	AATTG	5	0.02
AATT	46	0.20	AATCC	5	0.02
CGG	41	0.18	GGAAG	5	0.02
CTC	41	0.18	TTGTTT	5	0.02
CCG	38	0.17	TTTAAT	5	0.02
GTC	38	0.17	AAATTT	5	0.02
GAG	37	0.16	ATATTT	5	0.02
ACG	34	0.15	AACT	4	0.01
CTA	30	0.13	CGTG	4	0.01
AGT	26	0.11	TGGTT	4	0.01
ATATT	26	0.11	Others	274	1.28
AAAAAT	23	0.10	TOTAL	21981	100

In total, 39.29% of the *D. alata* genome contained TEs, which corresponded to 143,004,122 bp and 421,701 TEs. The long terminal repeat retrotransposons Gypsy/DIRS1 (28.6%) and Ty1/Copy (12.4%) were the most abundant TEs in the sequences, followed by the retrotransposon LINE/L1 (2.5%), DNA transposon CMC-ENSPM (1.7%) and retrotransposon LINE RTE-BovB (1.6%) (Table 2). A total of 45.8% of the identified TEs did not fit this particular class. We designed 100 primer pairs to amplify the TEs in *D. alata* (Supplementary 2).

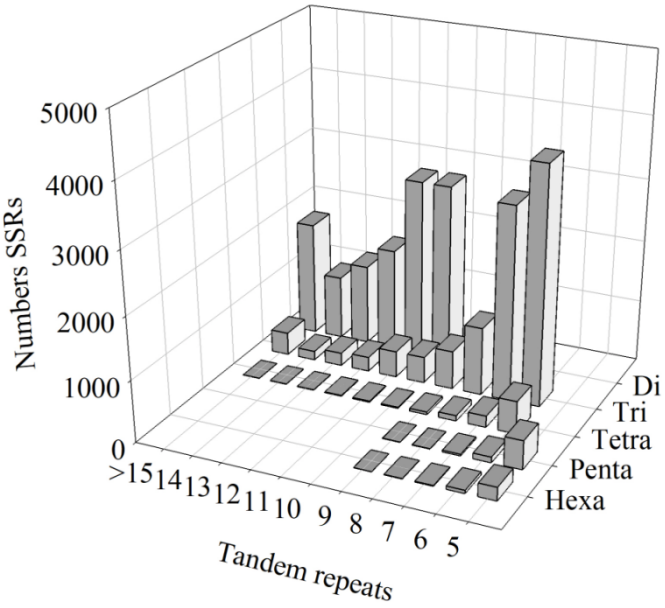

Figure 2. Distribution of microsatellites annotated in *Dipteryx alata* for repeat motif size and number of tandem repeats.

Table 2. Classes and frequencies of TEs identified in sequences of *Dipteryx alata* using RepeatModeler and RepeatMasker.

Classe	Order	Superfamily	Number of elements	Percentage of elements (%)	Occupied sequences (bp)	Percentage of sequences in the genome (%)
Retroelements		Caulimovirus	5,477	1.29	2,617,339	0.71
	LTR elements	*Copia*	52,388	12.42	22,676,307	6.23
		Gypsy	120,642	28.60	50,992,893	14.01
		Unclassified	657	0.15	166,344	0.04
	LINE	LINE/L1	10,887	2.58	3,719,137	1.02
		LINE/L2	10	0.002	415	0.0001
		RTE-BovB	6,966	1.65	2,120,408	0.58
	SINE		808	0.19	101,495	0.02
DNA transposons	TIR	CMC-EnSpm	7,451	1.76	2,738,611	0.75
		Crypton	261	0.06	91,752	0.02
		hAT-Ac	2,778	0.65	758,007	0.20
		hAT-Charlie	72	0.01	8,773	0.002
		hAT-Tag1	4,295	1.01	1,382,478	0.37
		hAT-Tip100	1,016	0.24	302,076	0.08
		MuLE-MuDR	6,133	1.45	1,942,663	0.53
		PIF-Harbinger	2,678	0.63	628,614	0.17
		TcMar-Fot1	112	0.02	32,231	0.008
		TcMar-Stowaway	257	0.06	44,481	0.01
		Unclassified	163	0.03	11,485	0.003
Rolling-circles		Helitron	5,257	1.24	2,178,518	0.59
Unclassified			193,393	45.86	50,490,095	13.87
Total			421,701	100	143,004,122	39.29

DISCUSSION

A large number of repetitive elements was annotated in the draft genome assembly of *D. alata.* Microsatellite regions and TEs have been identified on a large scale. The results generated genomic information hitherto unavailable for this species from the Brazilian Cerrado (Antunes, 2016). Repetitive elements are essential components that influence the size, operation, organization and evolution of genomes (Pisupati et al., 2018). Usually, there is a positive correlation between genome size and TE content. For example, among the species of the Leguminosae family, *Glycine max* has a 1335 Mb genome and 58.7% TEs (Schmutz et al., 2010), *Phaseolus vulgaris* has a 773 Mb genome and 45.4% TEs (Schmutz et al., 2014) and *Medicago truncatula* has a 479 Mb genome and 30.5% TEs (Young et al., 2011). The TEs contribute significantly to the increase in the genome size of the Leguminosae species, in *D. alata* they correspond to 39.29% of the genome. Different compositions of TEs in genomes reflect the different evolutionary histories of species. TEs are usually silenced during plant development, and activation of transcription, as well as transposition, is mainly induced by biotic and abiotic stress, which may be regarded as an adaptive response of the genome. The mechanisms of TE transposition contribute to the emergence of new genes and regulatory networks in the genome. (Sahebi et al., 2018).

TEs are scattered throughout the genome, while microsatellites are predominant in noncoding regions. Microsatellite regions are also abundant in the genome of other legumes, such as *Cicer arietinum* (81,845 microsatellites) (Varshney et al., 2013). Microsatellite regions play an important role in the evolution of the genome because they show high mutation rates. Microsatellites generally show high levels of intra- and interpopulation polymorphism (Bagshaw, 2017). Additionally, repetitive DNA plays a role in chromosomal stability and is the main constituent of the centromeric and telomeric regions of eukaryotic chromosomes (Klein and O'Neill, 2018). Due to the various roles played by repetitive DNA sequences, it was important to identify and characterize these regions in the *D. alata* genome.

This study allowed the identification of a large number of microsatellite regions and TEs in *D. alata*, which are potentially useful in the development of molecular markers. Microsatellite markers are valuable tools in several types of studies in plants such as genetic diversity studies, population structure analyses, kinship studies, forensic investigations, evolutionary studies, and phylogenetic studies, among others (Deng et al., 2016; Bagshaw, 2017). Markers based on TEs have important applications in the assessment of genetic diversity and construction of phylogenies (Kumar and Hirochika, 2001; Karakülah and Pavlopoulou, 2018).

D. alata is a species that still has few microsatellite markers available today. Collevatti et al. (2013) evaluated the effects of demographic history on the genetic diversity and population structure of *D. alata* using data from 8 microsatellite markers, which were used to genotype individuals from 25 populations. Soares et al. (2015) studied the spatial distribution of genetic variability in 23 populations of *D. alata* using 8 nuclear microsatellites. The low genetic diversity within populations makes it difficult to develop microsatellite markers for *D. alata* using conventional methods based on the use of artificial probes, fragment cloning and Sanger sequencing (Collevatti et al., 2013). High throughput sequencing, followed by assembly of genomic sequences, identification of microsatellite regions, and primer design will facilitate the development of molecular markers for *D.*

alata. Guimarães et al. (2017) developed 11 polymorphic microsatellite markers for *D. alata* using next generation sequencing data. The markers developed by Guimaraes et al. (2017) were efficient for evaluating the *D. alata* mating system and comparing pollen dispersal patterns between in situ and *ex situ* conditions (Guimarães et al., 2019). We designed primers to amplify 120 microsatellite regions potentially useful for the development of molecular markers. In addition, we identify and design primers for TEs we identified and designed, also aiming the development of molecular markers for *D. alata*. TE-based markers provide a new system for genetic analyses, contributing to the already powerful set of genetic tools available to study plant biology (Kumar and Hirochika, 2001; Karakülah and Pavlopoulou, 2018). Therefore, the results for *D. alata* obtained in this study expand the possibilities for genetic research on this species.

ACKNOWLEDGMENTS

This work was supported by several grants and fellowships to the research network "Geographic Genetics and Regional Planning for natural resources in Brazilian Cerrado" (GENPAC) from CNPq/MCT/CAPES/FAPEG (projects no. 564717/2010-0, 563727/2010-1 and 563624/2010-8), CNPq Universal (475182/2009-0) and by "Núcleo de Excelência em Genética e Conservação de Espécies do Cerrado" - GECER (PRONEX/FAPEG/CNPq CP 07-2009; 07/2012). M.P.C.T., T.N.S. and E.N. have been continuously supported by productivity fellowships from "Conselho Nacional de Desenvolvimento Científico e Tecnológico" (CNPq) and A.M.A has been supported by doctorate and postdoctoral fellowships from “Coordenação de Aperfeiçoamento de Pessoal de Nível Superior” (Capes), which we gratefully acknowledge. Also, this work is in the context of the GT - Genetics and Evolutionary Genomics, linked to the research line "planning in conservation and sustainable use of biodiversity" of the Instituto Nacional de Ciência e Tecnologia – Ecologia, Evolução e Conservação da Biodiversidade (INCT_EECBio).

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Andrews S (2010). FastQC: a quality control tool for high throughput sequence data.

Antunes AM (2017). Tamanho, montagem de novo e anotação do genoma de *Dipteryx alata* (Leguminosae). Doctoral thesis. Goias Federal University, Goiás. Available at [http://repositorio.bc.ufg.br/tede/handle/tede/7297]

Arakaki AH, Scheidt GN, Portella AC, Arruda EJD, et al. (2009). O baru (*Dipteryx alata* Vog.) como alternativa de sustentabilidade em área defragmento florestal do Cerrado, no Mato Grosso do Sul. *Interações*. 10(1): 31-39.

Bagshaw ATM (2017). Functional mechanisms of microsatellite DNA in eukaryotic genomes. *Genome Biol. Evol.* 9: 2428-2443.

Bolger AM, Lohse M and Usadel B (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*. 30: 2114-2120.

Cheng S, Melkonian M, Smith SA, Brockington S, et al. (2018). 10KP: A phylodiverse genome sequencing plan. *GigaScience*. 7: giy013.

Collevatti RG, Telles MPC, Nabout JC, Chaves LJ, et al. (2013). Demographic history and the low genetic diversity in *Dipteryx alata* (Fabaceae) from Brazilian Neotropical savannas. *Heredity*. 111: 97-105.

Deng P, Wang M, Feng K, Cui L, et al. (2016). Genome-wide characterization of microsatellites in Triticeae species: Abundance, distribution and evolution. *Sci. Rep.* 6: 32224.

Fan F, Cui B, Zhang T, Ding G, et al. (2014). LTR-retrotransposon activation, IRAP marker development and its potential in genetic diversity assessment of masson pine (*Pinus massoniana*). T*ree Genet. Genomes*. 10: 213-222.

Ferreira CM, Gabriel, GH, Nepomuceno L, Cruz VS, et al. (2018). Caracterização botânica e cadeia produtiva da espécie *Dipteryx alata* Vogel. *Enciclopedia Biosfera.* 15(28): 201-217.

Foulongne-Oriol M, Murat C, Castanera R, Ramírez L, et al. (2013). Genome-wide survey of repetitive DNA elements in the button mushroom *Agaricus bisporus*. *Fungal Genet. Biol.* 55: 6-21.

Guimarães RA, Telles MPC, Antunes AM, Correa KM, et al. (2017). Discovery and characterization of new microsatellite loci in *Dipteryx alata* Vogel (Fabaceae) using next-generation sequencing data. *Genet. Mol. Re*s. 16: gmr16029639.

Guimarães RA, Corrêa KM., Chaves LJ, Naves RV, et al. (2019). Mating system and pollen dispersal in *Dipteryx alata* Vogel (Leguminosae): comparing in situ and ex situ conditions. *Tree Genet. Genomes*. 15: 28.

Karakülah G and Pavlopoulou A (2018). In silico phylogenetic analysis of hAT transposable elements in plants. *Genes*. 9: E284.

Klein SJ and O'Neill RJ (2018). Transposable elements: genome innovation, chromosome diversity, and centromere conflict. *Chromosome Res*. 26: 5-23.

Kumar A and Hirochika H (2001). Application of retrotransposons as genetic tools in plant biology. *Trends Plant Sci*. 6: 127-134.

Lewin H, Robinson G, Kress W, Baker W, et al. (2018). Earth BioGenome Project: Sequencing life for the future of life. *Proc. Natl. Acad. Sci.* 11: 4325-4333.

Meglécz E, Costedoat C, Dubut V, Gilles A, et al. (2009). QDD: A user-friendly program to select microsatellite markers and design primers from large sequencing projects. *Bioinformatics*. 26: 403-404.

Mardis ER (2008). Next-Generation DNA Sequencing Methods. *Annu Rev. Genomics Hum. Genet.* 9: 387-402.

Metzker ML (2010). Sequencing technologies the next generation. *Nat. Rev. Genet*. 11: 31-46.

Oliveira MIB and Sigrist MR (2008). Fenologia reprodutiva, polinizacão e reprodução de Dipteryx alata Vog. (Leguminosae-Papilionoidae) em Mato Grosso do Sul. *Rev. Brasil Bot.* 31: 195-207.

Pisupati R, Vergara D and Kane NC (2018). Diversity and evolution of the repetitive genomic content in *Cannabis sativa*. *BMC Genomics*. 19: 156.

Quinlan A and Hall I (2010) BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*. 26: 841-842.

Ratter J (2000) Estudo preliminar da distribuição das espécies lenhosas da fitofisionomia Cerrado sentido restrito nos Estados compreendidos pero Bioma Cerrado. *Bol. do Herbário Esechias Paulo Heringer*. 5: 5-43.

Ribeiro JF, Sano S, Brito MAD and Fonseca CELD (2000) Baru (*Dipteryx alata* Vog.). Jaboticabal: Funep, 41 p. (Serie Frutas Nativas, 10).

Rozen S and Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.* 132: 365-86.

Safonova Y, Bankevich A, Pevzner PA (2015) DIPSPADES: Assembler for highly polymorphic diploid genomes. *J Comput Biol.* 22: 528-45.

Sahebi M, Hanafi MM, Van Wijnen AJ, Rice D, et al. (2018) Contribution of transposable elements in the plant's genome. *Gene.* 665: 155-166.

Sano S, Ribeiro J and Brito M (2004) Baru: biologia e uso. *Embrapa Cerrados*. 116: 51.

Schmutz J, Cannon SB, Schlueter J, Ma J, et al. (2010) Genome sequence of the palaeopolyploid soybean. *Nature.* 463: 178-183.

Schmutz J, McClean PE, Mamidi S, Wu GA, et al. (2014) A reference genome for common bean and genome-wide analysis of dual domestications. *Nat. Genet.* 46: 707-713.

Soares TN, Diniz-Filho JAF, Nabout JC, Telles MPC, et al. (2015). Patterns of genetic variability in central and peripheral populations of *Dipteryx alata* (Fabaceae) in the Brazilian Cerrado. *Plant Syst. Evol.* 301: 1315.

Tarailo-Graovac M and Chen N (2009) Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics*. 5: 4.10.1-4.10.14.

Varshney RK, Song C, Saxena RK, Azam S, et al. (2013) Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. *Nat. Biotechnol.* 31: 240-246.

Young N, Debellé F and Oldroyd G (2011) The Medicago Genome Provides Insight into the Evolution of Rhizobial Symbioses. *Nature.* 480: 520-524.

Yu JN, Won C, Jun J, Lim YW, et al. (2011) Fast and cost-effective mining of microsatellite markers using NGS technology: An example of a Korean water deer *Hydropotes inermis argyropus*. *PLoS One.* 6: e26933.