Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Discovery of novel SSR markers from transcriptome data of Astronium fraxinifolium Schott, a threatened tree species in Brazil

Maiara Cornacini
Universidade Estadual Paulista Julio de Mesquita Filho
Ricardo Oliveira Manoel (rickom.is@gmail.com)
UNESP - Univ Estadual Paulista, Alameda das Tecomarias s/n, Botucatu https://orcid.org/0000-0002-7582-1829
Marcelo Alcantara
Universidade Estadual Paulista Julio de Mesquita Filho
Mário Moraes
Universidade Estadual Paulista Julio de Mesquita Filho
Edvaldo Silva
Universidade Estadual Paulista Julio de Mesquita Filho
Leonel Pereira Neto
Empresa Brasileira de Pesquisa Agropecuaria Recursos Geneticos e Biotecnologia
Alexandre Sebbenn
Instituto Florestal de São Paulo
Bruno Rossini
Universidade Estadual Paulista Julio de Mesquita Filho
Celso Marino
Universidade Estadual Paulista Julio de Mesquita Filho

Research article

Keywords: Anacardiaceae, conservation genetics, management, microsatellite markers, population genetics

Posted Date: August 25th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-65083/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Abstract

Background

Astronium fraxinifolium is an endangered tree species from Brazil. Due to its high importance for environmental reforestation, as well as for the use of its wood, it is necessary to implement management programs for conservation of this species. Simple sequence repeats (SSR) or microsatellite markers have been widely used in population genetic studies across diverse organisms. In this study, we reported for the first time SSR markers for *A. fraxinifolium* as well as its frequency and distribution from transcriptome data.

Results

More than 125 thousand RNA-seq sequences derived microsatellites, with predominant distribution of trinucleotides repeats. From the initial screening, we selected 20 microsatellite loci, validated and evaluated genetic indexes in two natural populations. All loci were polymorphic, ranging from four to eleven alleles per locus. The observed and expected heterozygosities ranged from 0 to 1.0 and from 0.533 to 1.0, respectively. Genetic differentiation between populations (= 0.363) showed higher diversity within than among populations.

Conclusions

The developed SSR loci from RNA-seq consists in a base for future studies of genetic diversity and population structure, mating system and gene flow in *A. fraxinifolium* populations as well as related species, aiming the conservation and management of the species.

Background

Forest fragmentation directly impacts the genetic diversity, population structure, mating system and gene flow of tree populations. Population genetic studies based on genetic markers are keys to understand the effects of anthropogenic interventions on natural populations, conservation and improvement of trees species. *Astronium fraxinifolium* Schott (Anacardiaceae) tree occur discontinuously distributed on rocky terrains of different Brazilian biomes, from Cerrado to Caatinga [1, 2]. It is an insect pollinated dioecious tree, used for restoration of degraded areas [2]. Due to the intense fragmentation of its biomes, *A. fraxinifolium* is placing as threatened of extinction and remaining populations are found as isolated trees along highways margins or in small forest fragments [3, 4]. Thus, the development of genetic markers as microsatellite loci (SSR) is urgent to be used as tool for genetic investigation of genetic diversity, population structure, mating system, and gene flow of the remaining species populations.

Genomic studies in *Astronium* genus are scarce. More recently, with the development of high throughput sequencing technologies has increased the development of molecular markers in a broad range of organisms [5, 6]. The search against these datasets has increased the chances of finding SSR without any prior enrichment. From this, the development of SSR markers from transcriptome sequences has become an effective tool for population genetic investigations [7, 8], in special for endangered species [8]. Here, we developed a set of 20 polymorphic microsatellites loci for *A. fraxinifolium* and evaluated its frequency and distribution based on Illumina sequencing from a RNA-Seq data. These loci were then validated to be reproducible in order to assess genetic diversity, mating

system and gene flow in this tree species. We also included an analysis of repeats and GO classification of these reads.

Results

Identification and classification of SSR markers

In this work, the sequence run produced 189 million of paired-end reads, of which more than 95 million successfully joined reads. The trinucleotides motifs the most abundant followed by di-, hexa-, penta, and tetranucleotide motifs (approximately 41.8, 37.5, 9.80, 7.6, and 3.4%, respectively, Table 1). From these, on average, 32% of the sequences had enough flanking sequence for designing of the primers, except for the tetra- and trinucleotides loci where this was 40.9 and 36.9%, respectively. From an initial screening, more than 125 thousand sequences (redundant sequences; available upon request) were identified with tandem repeats and then we designed twenty primer pairs for amplification and test in a population study.

Table 1
Summary of Illumina paired-end sequence data; it includes non- and perfect motifs di-, tri-, tetra-, penta- and hexa-nucleotides for *Astronium fraxinifolium*

Motif	Di-	Tri-	Tetra-	Penta-	Hexa-
Number of contigs containing microsatellites	146,182	163,062	13,274	29,540	38,160
Number of contigs with flanking sequence	37,389	60,331	5,434	8,852	13,202
Total number of reads = 189,057,492; All contigs = 95,768,009					

Insert Table 1

The SSR functional annotation was classified under the three major categories: cellular component, molecular function and biological process (Fig. 1). The GO classification related to cellular component showed that the most abundant GO are intracellular (GO:0005622) or intracellular part (GO:0044424). Considering molecular function, the most representative GOs were related to organic cyclic compound binding (GO:0097159) and heterocyclic compound binding (GO:1901363). The biological process category are mainly represented by genes involved in organic substance metabolic process (GO:0071704) and cellular metabolic process (GO:0044237). Considering the tissue analyzed in the transcriptome analysis, the GO classification also show a great number of genes related to cell communication (GO:0007154), anatomical structure development (GO:0048856) and multicellular organism development (GO:0007275) which are consistent with the initial steps of the development, showing a great potential for future studies in these regions.

Insert Fig. 1

Figure 1 Functional annotation of SSRs in coding regions for *A. fraxinifolium* transcriptome, including the number of genes putatively involved in different subcellular functions in GO classification.

Genetic diversity of natural populations

The use of SSR-derived from RNA-Seq studies increases the success of amplification, including related species, due to its conservation of transcribed regions flanking sites. All microsatellite primer pairs designed were successfully

amplified and polymorphic in the studied populations, being detected between 4 to 11 alleles per locus and ranging from 0.346 to 0.857 (Table 2), which indicate useful markers for population studies.

Table 2
Results of screening in two populations of *Astronium fraxinifolium*

	Ilha Solteira (*n* = 30)						Selvíria (*n* = 30)					
Locus												
Ga01	8	0.760	0.933	0.799	-0.171	-	9	0.754	0.654	0.792	0.177*	-	0.757
Ga02	11	0.551	0.867	0.626	-0.394	-	7	0.784	0.808	0.824	0.105	-	0.513
Ga03	8	0.346	0.000	0.533	0.094	-	8	0.375	1.000	1.000	0.465*	+	0.088
Ga04	11	0.788	0.467	0.818	0.434*	+	10	0.804	0.533	0.836	0.366*	+	0.044
Ga05	8	0.656	0.367	0.718	0.494*	+	5	0.664	0.296	0.726	0.597*	+	0.785
Ga06	9	0.755	0.733	0.797	0.081*	-	8	0.712	0.769	0.754	-0.020	-	0.959
Ga07	10	0.762	0.800	0.804	0.005	-	8	0.693	0.630	0.746	0.158*	-	0.484
Ga08	8	0.840	0.933	0.871	-0.073	-	8	0.719	0.821	0.769	-0.069	-	0.583
Ga09	10	0.805	0.667	0.833	0.202*	+	11	0.857	0.733	0.885	0.174*	+	0.337
Ga10	7	0.569	0.308	0.654	0.475*	+	4	0.725	0.600	0.785	0.106	-	0.263
Ga11	8	0.724	0.800	0.777	-0.030	-	6	0.785	0.731	0.827	0.159*	-	0.058
Ga12	8	0.657	0.944	0.730	-0.175	-	6	0.743	0.800	0.799	0.137*	-	0.314
Ga13	7	0.756	0.722	0.810	0.124*	-	8	0.725	0.762	0.779	0.089	-	0.194
Ga14	9	0.734	0.722	0.794	0.069	-	8	0.746	0.824	0.804	0.133*	-	0.088
Ga15	8	0.711	0.826	0.769	0.085	-	8	0.757	0.500	0.817	0.377*	+	0.013
Ga16	9	0.580	0.600	0.623	0.038	-	6	0.761	0.724	0.797	0.093	-	0.693
Ga17	5	0.669	0.828	0.721	-0.129	-	8	0.655	0.567	0.719	0.215*	-	0.340
Ga18	8	0.690	0.655	0.743	0.125	-	11	0.815	0.733	0.849	0.138*	-	0.149
Ga19	4	0.558	0.552	0.634	0.140	-	7	0.551	0.133	0.606	0.783*	+	0.086
Ga20	8	0.710	0.759	0.760	0.010	-	9	0.781	0.733	0.821	0.109	-	0.072
Mean	8.2	0.681	0.674	0.741	-0.029	-	7.5	0.720	0.668	0.797	0.059	-	0.363

= number of alleles per locus; = polymorphism information content; = observed heterozygosity; = expected heterozygosity; = fixation index; = null alleles occurrence; = genetic differentiation between populations; $P < 0.05$

Insert Table 2

Therefore, despite the history of fragmentation of the *A. fraxinifolium* populations, the SSR loci showed a large amount of genetic variation: the observed heterozygosity () ranged from 0 to 0.944 (mean of 0.674) and expected heterozygosity () ranged from 0.533 to 0.871 (mean of 0.741) in Ilha Solteira, and in Selvíria ranged from 0.133 to

1.0 (mean of 0.668) and ranged from 0.606 to 1.0 (mean of 0.797). The fixation index () ranged from − 0.394 to 0.494 (mean of 0.090) in Ilha Solteira, and in Selvíria ranged from − 0.069 to 0.783 (mean of 0.162). Null alleles were observed in four and six loci in Ilha Solteira and Selvíria, respectively. After Bonferroni sequential correction, genotypic linkage disequilibrium (LD) was observed in four pairs of loci in Ilha Solteira and in three pairs in Selvíria (Table 3).

Table 3
Genotypic disequilibrium between pairwise microsatellite loci in adult of *Astronium fraxinifolium*

Pairwise loci	Ilha Solteira	Selvíria
Ga01xGa05	0.00005	1.00000
Ga01xGa06	0.00005	1.00000
Ga01xGa07	0.00005	1.00000
Ga01xGa08	0.00011	1.00000
Ga01xGa19	0.89137	0.00005
Ga01xGa20	0.73405	0.00005
Ga02xGa03	0.00047	0.00005
The values represent the probability of genotypic disequilibrium after 19.000 permutations of alleles among individuals. Probability after Bonferroni's corrections: P = 0.000263 (α = 0.05).		

Insert Table 3

To test the genetic similarity between populations, we used DAPC and STRUCTURE analysis. The genetic differentiation () between populations (0.363) was high great part of genetic is distribute within than among populations. In fact, the high genetic differentiation among the populations was expected given their geographical distance (50 km). The species is pollinated by bees, which have limited distances reported [1], which can explain the high genetic differentiation between both. The PCA showed a clear differentiation of both populations, with some individuals mixed between them. Furthermore, DAPC and both assignment probability tests (from adegenet package and STRUCTURE) resulted in the similar results of population structure, with two distinct populations (Fig. 2).

Insert Fig. 2

Figure 2 PCA, DAPC and assignment tests for the two populations of *Astronium fraxinifolium* analyzed. **a** PCA showing the dsitribution of genotypes. **b** DAPC clearly showing the differences between popualtions. **a, b** Colors reflect each population: Ilha Solteira (red) and Selviria (blue). **c** assignment test from adegenet package. **d** STRUCTURE results from analysis at optimal K = 2. **c, d** Each column and colors reflect the genetic assignment of individuals: in **c** Ilha Solteira (brown) and Selviria (blue); **d** Ilha Solteira (green) and Selviria (red).

Discussion

In *A. fraxinifolium* transcriptome data, a predominance of -tri, followed by dinucleotides motifs, with more than 79% of all identified contigs, which could not affect the protein structure, with non-perturbation of the reading frame [9, 10]. When analyzing all repeats number, we identified that repeats number greater than 10 corresponds with less than 7.8% of all SSRs (Table 4). The SSR frequency decreased with an increase in motif length, as reported for

Magnolia wufengensis [11]. The frequency of motifs from AG/CT corresponds to more than 24.3%, being the most abundant motif in this species, followed by TCT/AGA repeats with less than 6%. (Fig. 3). These frequencies of AG/CT repeats are higher than found for other species such bamboo (17.11%) [12], but less than for Magnolia (37.8%) [11]. High frequencies of AG repeats are also reported for other plant species being suggested that could be related to mutation mechanism of generation of SSRs or selective pressure to particular sequences [9, 10, 11, 13, 14].

Table 4
Frequency distribution of SSRs identified in the *A. fraxinifolium* transcriptome.

Repeat number	Di-	Tri-	Tetra-	Penta-	Hexa-	Total
4	N/A	N/A	N/A	6685	11008	17693
5	N/A	N/A	3003	1740	1644	6387
6	N/A	33155	2139	365	466	36125
7	13340	15352	217	56	55	29020
8	6167	7539	52	3	20	13781
9	5489	2413	10	2	4	7918
10	3511	1013	4	0	2	4530
> 10	8882	859	9	1	3	9754
Total	37389	60331	5434	8852	13202	125208

Insert Table 4

Insert Fig. 3

Figure 3 Frequency distribution of the most representative SSR motifs types in the *A. fraxinifolium* transcriptome.

The microsatellite markers developed were efficient in the genetic differentiation among populations sampled. The average levels of heterozygosity observed and expected were above that reported for populations of *Astronium graveolens* [15] and in other tropical species, as in populations of *Cedrela fissilis* (Meliaceae) [16], *Campomanesia xanthocarpa* (Myrtaceae) [17], *Myracrodruon urundeuva* [18] and *Eugenia uniflora* L. (Myrtaceae) [19], which confirms the existence of high genetic variability in the populations studied here. Therefore, these genetic markers are reliable to be used in population genetics studies, as such in the investigation of the pollen and seeds dispersal patterns aid to understand the actual distribution of natural populations, with impacts in the evolutionary history of a species. Previous studies with other tree species show a great range of pollen dispersal, such in *Hymenaea stignocarpa*, showed long-distance pollen dispersal reaching values of more than 8 km between the populations analyzed [20] and even more in *Ceiba pentandra*, reaching 18 km [21, 22]. However, these long distances are mainly related to the dispersion by bats, which have a large feeding area. For *A. fraxinifolium*, which are pollinated by bees, distances of almost 6 km were reported of the insects feeding behavior [23, 24]. Such results indicate that further investigations of pollen/seed dispersion are necessary for the species. To date, few studies were conducted based in natural populations of *A. fraxinifolium*, focused on silvicultural traits [25]. Recently in the genus, *A. graveolens* SSR loci were described, but not tested in other *Astronium* species [15]. Given this, the microsatellite markers in this work developed may be useful in genetic studies such as diversity and genetic structure, gene flow and mating

system, providing information for conservation, breeding and reforestation plans of the species. In addition, our study provides a database with more than 125 thousand of expressed SSR sequences in the genome that will serve as a basis for studies of consequences of forest fragmentation in tropical forest of Brazil, thus contributing to the development of adequate strategies for the conservation of *A. fraxinifolium* and related species from Anacardiaceae family.

Conclusions

This study reports the first SSR markers for *A. fraxinifolium*. The frequency and distribution of SSR motif types showed great diversity, with a predominance of trinucleotides motifs, as reported for other plant species. Functional annotation of SSRs can help future breeding programs in the selection of genes related to important characteristics of the development. Also, the use of transcriptome derived SSR can increase the rate of amplification in related species, due the conservative flanking sequence of these loci. At population level, these SSR markers showed enough levels of polymorphism in both populations analyzed. Therefore, the obtained results suggest that these markers can be used as tools for ecological population genetic studies, such as genetic diversity, spatial genetic structure, mating system and gene flow, besides improving the development of genetic conservation strategies and management of fragmented populations and related species.

Methods

Sample collection and DNA extraction for validation analysis

For the validation step, a total of 60 *A. fraxinifolium* samples from two natural populations were collected: 30 individuals from along the SP-595 highway in the municipality of Ilha Solteira, in the State of São Paulo (20° 21' 36.46" S, 51 °01 '15.52" W), Brazil, characteristic of the Semideciduous Seasonal Forest. The second population is composed of 30 adult trees from along the BR-158 highway in the municipality of Selvíria, in the state of Mato Grosso do Sul (20 °12 '02.30" S, 51° 14' 56.81" W), Brazil, characteristic of the Cerrado biome. Collection of leaf tissues was authorized by the Institute for Biodiversity Conservation (ICMBio), linked to the Ministry of the Environment (MMA) under the number 73998-1. The identification of trees in field was performed by PhD. José Cambuim, São Paulo State University (UNESP), based on the testimony plants of the areas that were sampled and are deposited at the Herbarium of Ilha Solteira of the São Paulo State University according to the following vouchers: HISA 1566, HISA 1570, HISA 4213, HISA 4214, HISA 4215, HISA 4216, HISA 4217, HISA 4218, HISA 4219, HISA 4220, HISA 4221, HISA 4222, HISA 4223, HISA 4224, HISA 6041, HISA 10299, HISA 10300, HISA 10301, HISA 10302, HISA 10303, HISA 10304, HISA 10305, HISA 10308, HISA 10531.

SSR loci identification and characterization

Species-specific microsatellite primers were generated from cDNA genomic library of the species [26] (data not published) and submitted to new generation sequencing (HiSeq Sequencing 2500 System, Illumina). Briefly, the total RNA was extracted from fresh embryonic axis of seeds collected from Alto Paraíso, Cavalcante, Colinas do Sul and Niquelândia cities of Goiás state, and Montes Claros, Mirabela and Lontra municipalities of Minas Gerais state, using NucleoSpin RNA Plant Kit (Macherey-Nagel, Düren, Germany. Seed collection was authorized by the Institute for Biodiversity Conservation (ICMBio) under the number 41166-1. The library was constructed using the TrueSeq RNA Library Prep Kit V2 (Illumina) and sequenced in a 2 × 100 bp paired-end run. For development of useful microsatellite markers for population genetics, we considered a subset of four samples from this run and the software SSR_pipeline [27] was used for search of SSRs by the following parameters: di-, tri-, tetra-, penta- and

hexanucleotides by the minimum of seven, six, five, four and four repeats respectively, with 40 bp flanking sequence. This software used as input the raw data reads, processed in a module of quality sort, followed by joining reads and completed the search for the SSRs. Primer design was conducted in BatchPrimer3 v1.0 [28]. Functional annotation of SSR-containing coding sequences were analyzed in Blast2GO software, using EnsemblePlants database from UniProt [29].

For the population analysis, genomic DNA was isolated from fresh leaves using the cetyltrimethylammonium bromide (CTAB) protocol [30], quantified in NanoDrop ND-1000 Spectrophotometer (NanoDrop Products, DE, USA) and its integrity was verified in 1% agarose gels running with TBE (1X), at constant voltage of 5V/cm. We selected 20 SSR loci for population validation. Polymerase chain reactions were performed in a Mastercyler thermocycler (Eppendorf, Hamburg, Germany) with an addition of a M13 tail for fluorescent labeling [31] in a reaction mixture containing 5.0 μL GoTaq Colorless Master Mix, 0.3 μL of forward primer (2 pmol), 0.3 μL of reverse primer (8 pmol) 0.3 μL of fluorescent primer with M13 (8 pmol; 6-FAM, VIC, PET or NED, Applied Biosystems), 0.5 μL of BSA (Bovine Serum Albumin), 1.0 μL genomic DNA (approximately 50 ng) and nuclease-free water for 10 ul of final reaction. The PCR program was as follows: 2 min at 96 °C, 35 cycles of 30 s at 96 °C, a primer-specific annealing temperature (Table 5) for 1 min 30 s, 72 °C for 1 min 30 s, followed by 12 cycles of 96 °C for 30 s, 53 °C for 1 min 30 s, 72 °C for 1 min 30 s, with final extension at 72 °C for 20 min. The PCR products were genotyped in ABI3130xl Genetic Analyzer (Applied Biosystems) with GeneScan 500 LIZ (Applied Biosystems). Genotypes were assigned in GeneMapper v.5.0 software (Applied Biosystems).

Table 5
Characteristics of the 20 microsatellite markers isolated for *Astronium fraxinifolium*

Locus	Primer sequences 5'- 3'	Repeat motif	Size range (bp)	Tm (°C)	Genbank accession
Ga01	F[a]: CATTCCACATTTGCCCTTG R: GCTTTCCTGTTCCCTAAATCC	$(GA)_{13}$	151–170	54	MN129762
Ga02	F[a]: TCTCCCCTCTACCTGCACTC R: ATCACCTTCTCCCACGAAGA	$(CT)_{12}$	92–118	54	MN129763
Ga03	F[a]: CAAGAACGAAAGAAGATCAACC R: ATGCTGATGGATCGATTGTG	$(TA)_{24}$	84–100	54	MN129764
Ga04	F[a]: AAACAAACATGATGCCAGGA R: CACTTCAATCCAGGGAAAAGA	$(AG)_{20}$	120–142	54	MN129765
Ga05	F[a]: CATCGATGATGAACCCAGAA R: CTCTACGGAAGCGGAGTCAC	$(TC)_{12}$	230–260	54	MN129766
Ga06	F: CGGCGCAAATAGTTAATAGGA R[a]: TGATCTACAGCCCATCCTGA	$(AG)_{13}$	104–138	54	MN129767
Ga07	F: GATGGTAGAGGCTGGTGTTTG R[a]: CCAAGCTCATTTCCGTCATC	$(AG)_{10}$	123–139	54	MN129768
Ga08	F[a]: GAAAGAGAAAGAGATCGGAACG R: CATCACCACCCACCAAACTC	$(GA)_{37}$	105–121	54	MN129769
Ga09	F[a]: GCCCAGATTCTCTCAATCTCC R: CGCGTTCTTTGTAGCAGACA	$(CT)_{24}$	73–123	54	MN129770
Ga10	F[a]: TCGCACACCTCATCTACAGC R: AACGACAAGTCGCTTTGACA	$(AG)_{14}$	97–111	54	MN129771
Ga11	F[a]: GACCTTCAAAGCCAACCAAC R: TTAGCTTGAGTGGCGATGTG	$(GGT)_{11}$	93–120	54	MN129772

Tm = melting temperature; [a]M13-tailed primer fluorescent labeled, following Schuelke (2000)

Locus	Primer sequences 5'- 3'	Repeat motif	Size range (bp)	Tm (°C)	Genbank accession
Ga12	F[a]: TGACCATCCCAAACACATTCT R: TGCGGTTAGATAGGGAATGC	$(ACC)_{09}$	92–115	54	MN129773
Ga13	F[a]: TGCGGTTAGATAGGGAATGC R: TGACCATCCCAAACACATTCT	$(GGT)_{11}$	91–112	54	MN129774
Ga14	F[a]: TGCGGTTAGATAGGGAATGC R: TGACCATCCCAAACACATTCT	$(GGT)_{09}$	92–113	54	MN129775
Ga15	F[a]: TCAGGACACTCTCCACGGTAG R: TACACGTTTTCACCGCCAAC	$(ATC)_{10}$	92–113	54	MN129776
Ga16	F[a]: CCACCACACCTCAGGAACTC R: GAGGGTGGAGTGGTCAGTGT	$(CCT)_{10}$	115–133	54	MN129777
Ga17	F[a]: GCGAAAGTGATGCTGTATTGG R: CATCGACACCAACCTCATGT	$(AAG)_{11}$	108–120	54	MN129778
Ga18	F[a]: TTCTAACACAGCAACCAAACG R: TGGTGGGTTTTCAGTTGTGA	$(TCA)_{12}$	108–126	54	MN129779
Ga19	F[a]: GGCACGAAAAAGAAGAGGAG R: CATGGCAGTGCAAGAATGTT	$(GAG)_{11}$	107–123	54	MN129780
Ga20	F[a]: TTCTAACACAGCAACCAAACG R: TGGTGGGTTTTCAGTTGTGA	$(TCA)_{12}$	111–129	54	MN129781

Tm = melting temperature; [a]M13-tailed primer fluorescent labeled, following Schuelke (2000)

Insert Table 5

Statistical population analysis

The number of alleles per locus (), observed () and expected () heterozygosity, polymorphism information content () was estimated using CERVUS 3.03 software [32]. The fixation index () and genotypic linkage disequilibrium (LD) were estimated for each population using FSTAT software [33]. To test if the values and to the LD were significantly different from zero, we used Monte Carlo permutations and a Bonferroni correction (95%, $\alpha = 0.05$). Micro-Checker v.2.2.3 [34] was used to detect the possibility of occurrence of null alleles () and estimated the genetic differentiation () with base in Hedrick's statistics [35]. The adegenet package [36] in R environment was used to conduct the

principal component analysis (PCA), discriminant analysis of principal components (DAPC) and the assignment probability of each individual. Additionally, a bayesian analysis was performed in STRUCTURE [37] assuming admixture model, correlated allele frequencies, testing each K (1–4) in 10 independent runs with 100,000 'burn-in' period and 1,000,000 generations. The search of optimal K was inferred using Evanno method [38] in Structure Harvester [39].

Abbreviations

GO: Gene Ontology; BSA:Bovine Serum Albumin; CTAB:Cetyltrimethylammonium bromide; DAPC:Discriminant analysis of principal components; F:Fixation index; GST:Genetic differentiation; H_e:Expected heterozygosity; H_o:Observed heterozygosity; LD:linkage disequilibrium; PCA:Principal component analysis; PIC:Polymorphism information content; SSR:Simple sequence repeats

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

All data generated or analysed during this study are included in this published article.

Competing Interests

The authors declare that they have no competing interests.

Funding

The Post-Doctoral fellowship (grant number 2018/00898-0; ROM) and genetic analyzes for the validation SSR loci were paid by the Sao Paulo Research Foundation. The funding agency played role in the design of the study, collection, analysis, interpretation of data and in writing the manuscript. E. A. Amaral da Silva was financed by CNPq (Process number 309718/2018-0).

Author Contributions

MRC, MAA and ROM carried out the field and laboratory steps of the research. EAAS and LGPN developed the cDNA genomic library. BCR and ROM performed analysis and wrote the manuscript. MLTM, BCR and CLM designed the experiment. BCR and CLM supervised the research and with MLTM and AMS reviewed the manuscript. All of the authors read and approved the final manuscript.

Acknowledgments

The authors are thankful for the technical support in field by José Cambuim, Alexandre Marques da Silva, Alonso Angelo da Silva and for Sao Paulo Research Foundation (grant number 2018/00898-0) for supporting this research through fellowships.

References

1. Leite EJ. State-of-knowledge on *Astronium fraxinifolium* Schott (Anacardiaceae) for genetic conservation in Brazil. Perspectives in Plant Ecology, Evolution and Systematics. 2002; 5: 63–77.
2. Carvalho PER. Espécies arbóreas brasileiras, Brasília: Embrapa Informação tecnológica; Colombo: Embrapa florestas. 2010; 231–240.
3. Aguiar AV, et al. Genetic variation in *Astronium fraxinifolium* populations in consortium. Crop Breeding and Applied Biotechnology. 2003; 3(2).
4. Ibama. Portaria Ibama nº 37-N, de 03 de abril de 1992 – Available from: http://www.mma.gov.br/estruturas/179/_arquivos/179_05122008034139.pdf (1992).
5. González-castellano I, et al. Isolation and characterization of 21 polymorphic microsatellite loci for the rockpool shrimp *Palaemon elegans* using Illumina MiSeq sequencing. Sci Rep. 2018;8:17197.
6. Souza DCL, et al. Development of microsatellite markers for *Myracrodruon urundeuva* (FF & MF Allemão), a highly endangered species from tropical forest based on next-generation sequencing. Molecular biology reports. 2018;45(1):71–5.
7. Zheng XF, et al. Development of microsatellite markers by transcriptome sequencing in two species of Amorphophallus (Araceae). BMC Genom. 2013; 14.
8. Zhou T, et al. Transcriptome sequencing and development of genic SSR markers of an endangered Chinese endemic genus *Dipteronia Oliver* (Aceraceae). Molecules. 2016; 21(3).
9. Li YC, Korol AB, Fahima T, Nevo E. Microsatellites within genes: structure, function, and evolution. Mol Biol Evol. 2004;21:991–1007.
10. Mrázek J, Guo X, Shah A. Simple sequence repeats in prokaryotic genomes. P Natl Acad Sci USA. 2007;104:8472–7.
11. Wang L, et al. Development and validation of EST-SSR markers of *Magnolia wufengensis* using de novo transcriptome sequencing. Trees. 2019;33:1213.
12. Cai K, et al. Development and characterization of EST-SSR markers from RNA-Seq data in *Phyllostachys violascens.* Frontiers in. Plant Sci. 2019;10:50.
13. Wang P, et al. Characterization and Development of EST-SSR Markers from a Cold-Stressed Transcriptome of Centipedegrass by Illumina Paired-End Sequencing. Plant Molecular Biology Reporter. 2017;35:215.
14. Cho YG, et al. Diversity of microsatellites derived from genomic libraries and GenBank sequences in rice (*Oryza sativa* L.). Theor Appl Genet. 2000;100:713–22.
15. Sanchez-Gomez KF, et al. Isolation and characterization of microsatellite loci in *Astronium graveolens* (Anacardiaceae) and cross amplification in related species. Mol Biol Rep. 2020;47:4003–7.
16. Gandara FB, et al. Development and characterization of microsatellite loci for *Cedrela fissilis* Vell (Meliaceae), an endangered tropical tree species. Silvae Genetica. 2014;63:240–3.

17. Góes BD, et al. Development and characterization of microsatellite loci for *Campomanesia xanthocarpa* (Myrtaceae) and cross amplification in related species. Acta Scientiarum. 2019;41(1):43454.
18. Souza DCL, et al. Development of microsatellite markers for *Myracrodruon urundeuva* (F.F. & M.F. Allemão), a highly endangered species from tropical forest based on next-generation sequencing. Molecular Biology Reports. 2017; 45(1): 71–75.
19. Sarzi DS, et al. Discovery and characterization of SSR markers in *Eugenia uniflora* L. (Myrtaceae) using low coverage genome sequencing. Anais da Academia Brasileira de Ciências. 2019; 91(1).
20. Moraes MA, et al. Long-distance pollen and seed dispersal and inbreeding depression in *Hymenaea stigonocarpa* (Fabaceae: Caesalpinioideae) in the Brazilian savannah. Ecology Evolution. 2018;8:7800–16.
21. Dick CW, et al. Spatial scales of pollen and seed-mediated gene flow in tropical rain forest trees. Tropical Plant Biology. 2008;1:20–33.
22. Gribel R, Lemes MR. Mating system and pollen flow of *Ceiba pentandra* (Bombacaceae) in Central Amazon Assessment of levels and dynamics of intra-specific genetic diversity of tropical trees. Second Annual Report to the European Commission (1997).
23. Abou-shaara HF. The foraging behaviour of honey bees, *Apis mellifera*: a review. Vet Med. 2014;59(1):1–10.
24. Hagler JR, et al. Foraging range of honey bees, *Apis mellifera*, in alfalfa seed production fields. Journal of Insect Science. 2011;11(144):1–12.
25. Aguiar AV, et al. Genetic variation in *Astronium fraxinifolium* populations in consortium. Crop Breeding Applied Biotechnology. 2003;3(2):95–106.
26. Pereira Neto LG. *Astronium fraxinifolium* Schott seed longevyti: physiological, biochemical and molecular studies. Thesis in Agronomy. Faculty of Agronomic Sciences, Paulista State University, Botucatu (2016).
27. Miller MP, et al. SSR_pipelin: A bioinformatic infrastructure for identifying microsatellites from paired-end Illumina high-throughput DNA sequencing data. J Hered. 2013;104:881–5.
28. You FM, et al. BatchPrimer3: a high throughput web application for PCR and sequencing primer design. BMC Bioinformatics. 2008;9(1):253.
29. Götz S, et al. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 2008;36:3420–35.
30. Doyle JJ, Doyle JL. Isolation of plant DNA from fresh tissue. Focus. 1990;12:13–5.
31. Schuelke M. An economic method for the fluorescent labeling of PCR fragments. Nat Biotechnol. 2000;18:233–4.
32. Kalinowski ST, Taper ML, Marshall TC. Revising how the computer program CERVUS accommodates genotyping error increases success in paternity assignment. Mol Ecol. 2007;16:1099–106.
33. Goudet J. FSTAT version 2.9.3.2, a program to estimate and test gene diversities and fixation index. Lausanne: Institute of Ecology; 2002.
34. Van Oosterhout C, et al. MICRO-CHECKER: Software for identifying and correcting genotyping errors in microsatellite data. Mol Ecol Notes. 2004;4:535–38.
35. Hedrick PW. A standardized genetic differentiation measure. *Evolution; International Journal of Organic Evolution.* 2005; 59(8): 1633–1638.
36. Jombart T, Almed I. Adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics. 2011;27(21):3070–1.

37. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155(2):945–59.
38. Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Molecular Ecology. 2005;14(8):2611–35.
39. Earl DA, Vonholdt BM. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conserv Genet Resour. 2012;4(2):359–61.

Figures

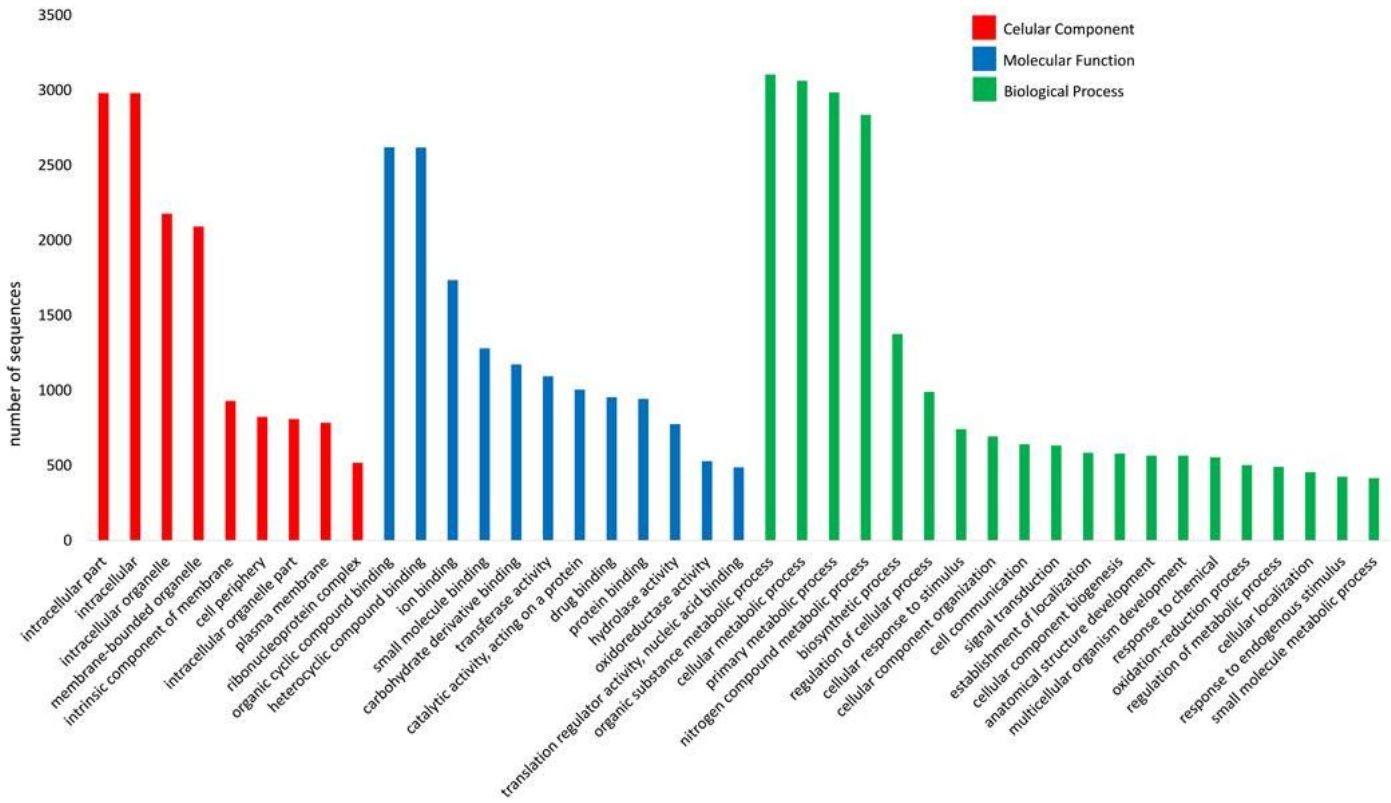

Figure 1

Functional annotation of SSRs in coding regions for A. fraxinifolium transcriptome, including the number of genes putatively involved in different subcellular functions in GO classification.

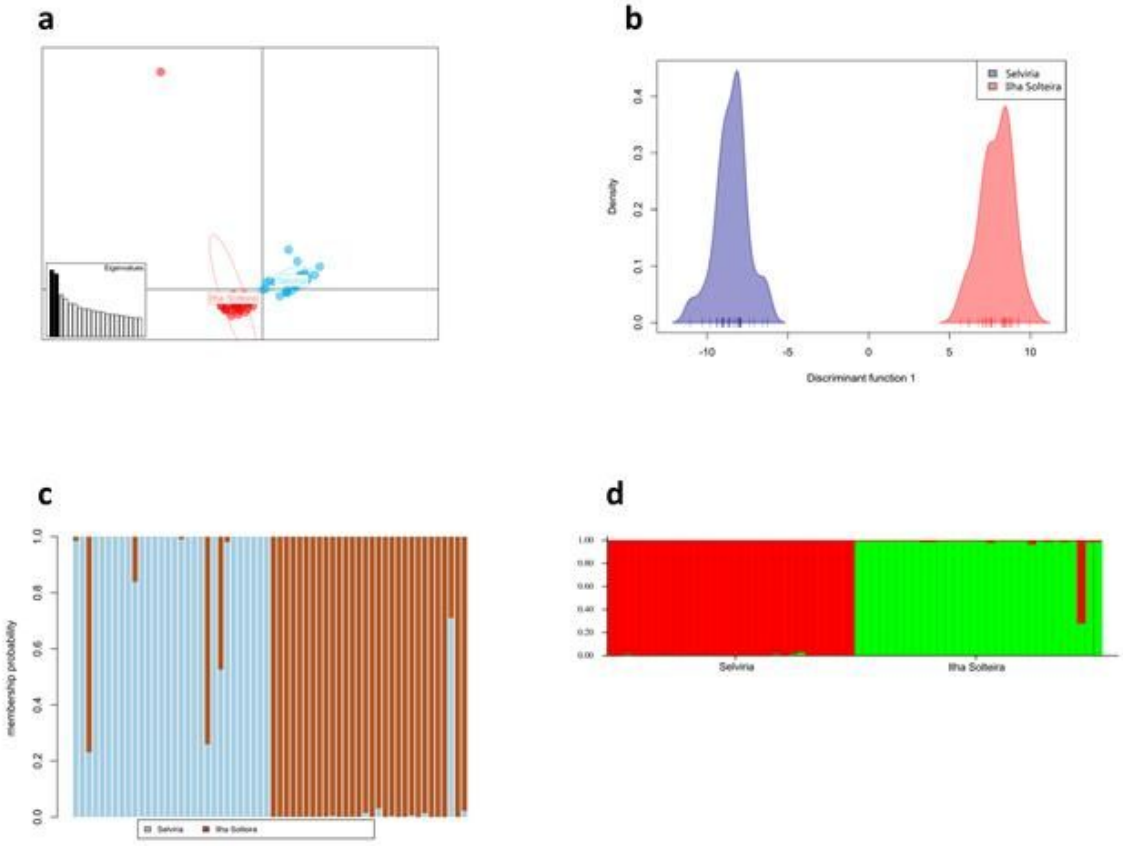

Figure 2

PCA, DAPC and assignment tests for the two populations of Astronium fraxinifolium analyzed. a PCA showing the dsitribution of genotypes. b DAPC clearly showing the differences between popualtions. a, b Colors reflect each population: Ilha Solteira (red) and Selviria (blue). c assignment test from adegenet package. d STRUCTURE results from analysis at optimal K=2. c, d Each column and colors reflect the genetic assignment of individuals: in c Ilha Solteira (brown) and Selviria (blue); d Ilha Solteira (green) and Selviria (red).

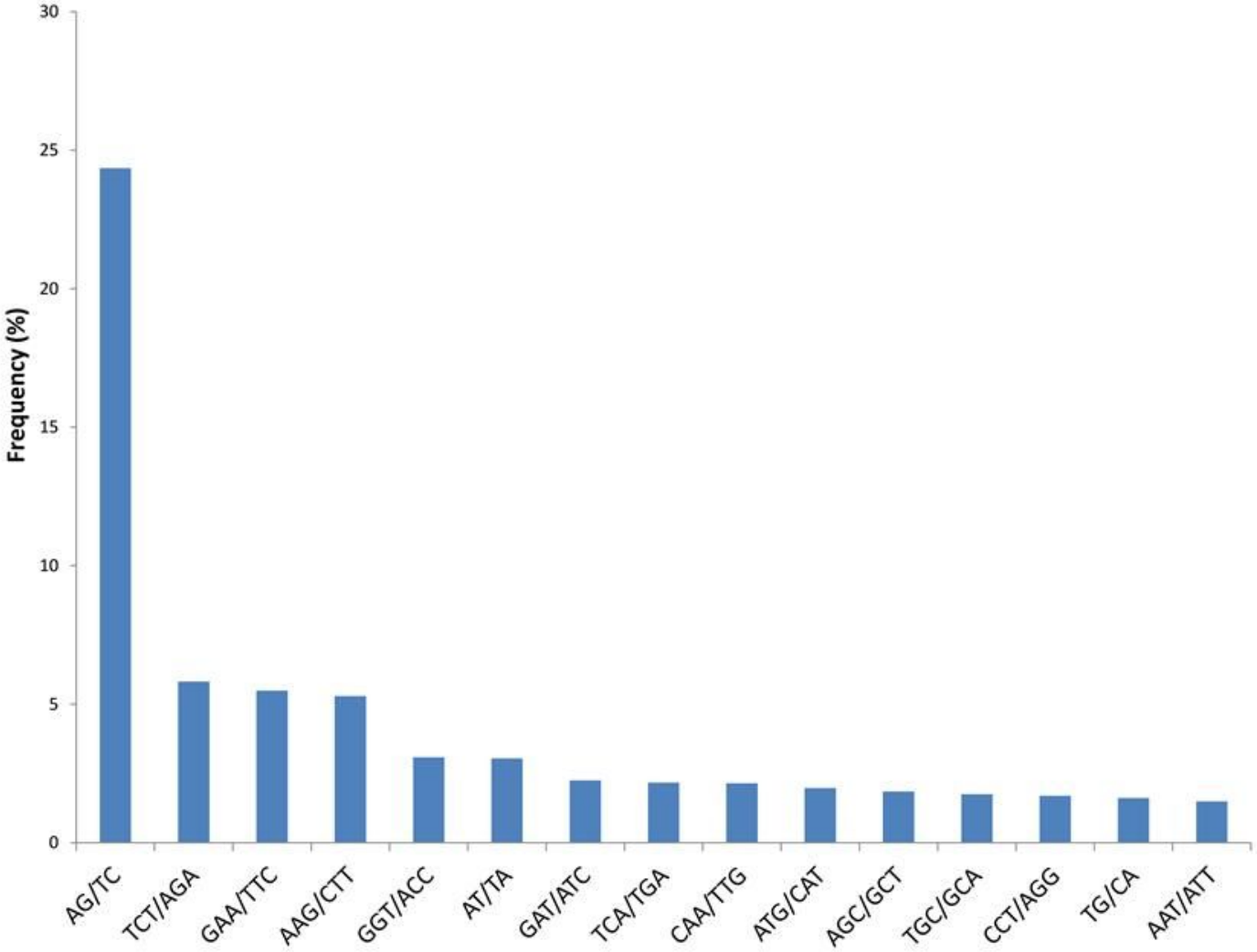

Figure 3

Frequency distribution of the most representative SSR motifs types in the A. fraxinifolium transcriptome.