Plant Ecology and Evolution 158 (1): 43–52, 2025
https://doi.org/10.5091/plecevo.138834

PLANT ECOLOGY EVOLUTION

RESEARCH ARTICLE

High-throughput sequence-based microsatellite genotyping for the non-model Neotropical tree species *Anadenanthera colubrina* (Leguminosae)

Alejandra Lorena Goncalves[1,2], María Victoria García[1,2], Emilie Chancerel[3], Olivier Lepais[3], Myriam Heuertz[3]

1 Universidad Nacional de Misiones, Facultad de Ciencias Exactas, Químicas y Naturales, Posadas, Argentina
2 Consejo Nacional de Investigaciones Científicas y Técnicas, Instituto de Biología Subtropical (UNaM – CONICET), Argentina
3 Université de Bordeaux, INRAE, BIOGECO, Cestas, France

Corresponding author: Alejandra Lorena Goncalves (alejandragoncalves@fceqyn.unam.edu.ar)

Academic editor: Katharina Budde ♦ **Received** 10 October 2024 ♦ **Accepted** 3 December 2024 ♦ **Published** 29 January 2025

Abstract

Background and aims – *Anadenanthera colubrina* is a Neotropical native forest tree species with significant ecological importance in Seasonally Dry Tropical Forests. Developing genetic markers for this species is relevant for conservation, breeding, and evolutionary studies. Previously available genetic markers for *A. colubrina* consisted of a few microsatellites. Next-generation sequencing (NGS) strategies allow simple and cost-effective development of new SSR loci from low-coverage whole genome shotgun sequencing. The main aim was to develop microsatellite markers for sequence-based high-throughput genotyping (SSRseq) in the species and to characterize their information content against traditional capillary electrophoresis-based microsatellite data by estimating the amount of molecularly accessible size homoplasy of each locus. Additionally, the reliability of these markers for population genetic analysis was assessed by genotyping two age classes (reproductively mature trees and seedlings) in a typical location in Argentina.
Key results – Sixty primer pairs targeting microsatellites were designed, of which 25 were validated with allelic error rates < 3% and genotype missingness < 20%. A significantly higher number of alleles per locus and heterozygosity was detected for SSRseq considering sequence polymorphisms compared to analysing the same data based on sequence size (length) only. Size homoplasy, calculated as the proportion of mismatches between datasets relative to the number of alleles differing in length, averaged 97.85% over all SSR loci. High levels of population genetic diversity were detected in adults and seedlings from Paranaense forests, exceeding those reported in previous studies of *A. colubrina* using traditional SSRs. The generated datasets increase the resolution of capillary-based microsatellite genotyping, allowing for more accurate inference of eco-evolutionary processes in non-model tree species.

Keywords

genotyping, multiplex PCR, next-generation sequencing, nuclear microsatellites, size homoplasy, SSRseq

INTRODUCTION

Anadenanthera colubrina (Vell.) Brenan (Leguminosae, Caesalpinioideae) is a non-model Neotropical forest tree species inhabiting Seasonally Dry Tropical Forests (SDTF). The high plant population biodiversity that characterizes this biome exhibits a fragmented distribution across Latin America and the Caribbean (Särkinen et al. 2011; DRYFLOR 2016). Until now, given the phylogeographic and ecological significance of this species, the genetic diversity and population genetic structure of *A. colubrina* have primarily been studied using plastid DNA microsatellites (Barrandeguy et al. 2016), sequences of non-coding regions of plastid DNA (Calonga Solís et

Plant Ecology and Evolution is published by Meise Botanic Garden and the Royal Botanical Society of Belgium.

al. 2014; Zerda Moreira et al. 2024), sequences of ITS regions (de Viana et al. 2014; Mangaravite et al. 2023), and species-specific nuclear microsatellites (Barrandeguy et al. 2012; Feres et al. 2012).

Microsatellites or SSRs (Simple Sequence Repeats) remain one of the most widely used molecular marker types; their codominance and high polymorphism characterize them as robust tools for population genetics analyses of plant species (Stefanini et al. 2023; Wang et al. 2023; Scotti-Saintagne et al. 2024). Ubiquitous in nuclear and organellar genomes, these markers are distinguished by a variable number of short repetitive sequence units, typically ranging from two to four nucleotides (Ellegren 2004). At the level of repetitive units, highly variable numbers of stepwise mutations (Kimura and Ohta 1978) support the high polymorphism that can be detected (Estoup et al. 2002). High levels of polymorphism are expected when analysing genetic diversity using nuclear SSR markers, as mutation rates at these loci can vary between 10^2 and 10^6 mutations per locus per generation (with an average of 5×10^4). In this way, these markers generate high allelic polymorphism, which is key in population genetic studies of processes acting on microevolutionary time scales (Schlötterer 2000). Traditionally, the allele information for estimating genetic distances among individuals is based on fragment length, as obtained from capillary sequencing instruments. However, fragments of the same length for a given locus can feature sequence differences, which results in fragment length (size) homoplasy. Homoplasy can be due to different mutational histories due to the high mutation rate of SSR (identity by state, not by descent) or to nucleotide point mutations, which can lead to an underestimation of genetic diversity (Estoup et al. 2002; Šarhanová et al. 2018). Indeed, point mutations in SSR flanking sequences may help to resolve microsatellite alleles identical by their repeat number, providing a better inference of ancient history (Ramakrishnan and Mountain 2004; Payseur and Cutter 2006).

Sequence-based microsatellite genotyping (SSRseq) is a new high-throughput, accurate, and rapid technique by next-generation sequencing (NGS) that allows the detection of higher levels of variation compared to traditional fragment size scoring (Šarhanová et al. 2018). Darby et al. (2016) and Vartia et al. (2016) validated SSRseq as a reliable method, while Lepais et al. (2020) proposed an integrative workflow for the development of SSRseq markers and their analysis for application to non-model species. Hence, genotyping by sequencing (GBS) can provide higher levels of diversity resolution than classical SSR genotyping, facilitating the characterization of genetic diversity and population genetic structure of non-model species, and could thus offer important new insights into evolutionary, ecological, and conservation biology. Therefore, SSRseq is a promising approach to studying relevant Neotropical native forest tree species such as *A. colubrina*, which is recognized as a key species in the strongly fragmented landscapes from the SDTF (Särkinen et al. 2011; DRYFLOR 2016).

The main aim of this study was to develop new SSRseq loci and generate sequence-based microsatellite data for *A. colubrina*, for which the only prior genomic data consists of a few dozen flanking regions of SSR, and some sequenced fragments of the nuclear, chloroplast, and mitochondrial genome. We characterized the advantages of the SSRseq method against traditional capillary electrophoresis-based microsatellite genotyping by considering nucleotide polymorphisms and we estimated the amount of molecularly accessible size homoplasy of each locus as support for the use of sequencing over assessing length polymorphism for genotyping. The methodological relevance of these markers was also tested in a biological framework by comparing reproductively mature trees and seedlings from the same population in a typical location of *A. colubrina* (Paranaense forest, Argentina), which is especially relevant for fragmented landscapes.

MATERIAL AND METHODS

Plant material and DNA extraction

Young leaves from 107 individuals of *A. colubrina* (adults and seedlings) were collected from four different Argentinean ecoregions: Paranaense forest (n = 95), Yungas (n = 6), Humid Chaco (n = 2), and the Delta and Islands of the Paraná River (n = 4). In the southern region of the Paranaense forests (Santa Ana, Misiones; -27.43372198, -55.579419), where *A. colubrina* characterizes forest patches within grassland landscapes, two life stages were sampled: 31 reproductively mature trees and 64 seedlings resulting from the germination of seeds collected from the fruits of four mother trees. The seedlings thus represent different sample sizes of four half-sib families and may contain full sibs since a previous study of an *A. colubrina* population in the same region suggested a selfing rate of 51–56% estimated from the inbreeding coefficient $s = 2F_{IS}/(1 + F_{IS})$ (Hartl and Clark 2007; Goncalves et al. 2019). Trees were georeferenced by GPS (Geographic Position System) using a Garmin eTrex® 20× receiver (precision of ± 3 m) and identified by an individual code. Leaves were dried with silica gel in the field and stored at room temperature.

Total genomic DNA was extracted for each individual using the modified cetyl-trimethylammonium bromide (CTAB) method (Doyle 1991). We added 15 mg of dry leaves and two steel beads (4 mm in diameter) into a 2 mL Eppendorf tube. Liquid nitrogen was used to break the sample into a powder using a Geno Grinder 2010 m (Spex Sample Prep, New Jersey) tissue homogenizer. Then, we added 400 µL of CTAB buffer and put the tubes into a 65°C water bath for lysis. The subsequent steps followed Doyle (1991). The DNA quality was assessed using a Nanodrop 2000 spectrophotometer (Thermo Fisher

Scientific, Waltham, USA) showing that the genomic DNA concentration was at least 80 ng/µL.

SSRseq development and genotyping

SSRseq markers were developed from low-coverage shotgun sequencing of a single library, prepared using the Qiaseq FX DNA library kit (Qiagen, Hilden, Germany). This library was generated from four pooled samples, each representing a different ecoregion (YS113, F134, IT185, SB5). Sequencing was conducted using Illumina MiSeq v.3 (Illumina, San Diego, USA) 2 × 300 bp paired-end sequencing, generating 4,497,218 read pairs. Overlapping forward and reverse reads were merged using BBMerge v.38.87 (Bushnell et al. 2017) with a minimum quality of 25, an overlap greater than 100 bp without any authorized mismatch, resulting in 3,239,270 merged reads. The QDD pipeline v.3.1.2 (Meglécz et al. 2014) was used to detect sequences containing microsatellites (SSR mining), identify good quality sequences, and design candidate primer pairs flanking the identified SSR. From 991,696 sequences longer than 200 bp, a total of 54,149 sequences containing microsatellites were extracted by QDD pipe 1, from which 28,197 singleton sequences were retained for QDD pipe 2. Then, QDD pipe 3 designed primers in singleton sequences with the following filter parameters: amplicon size between 100 and 180 bp, primer size between 21 and 26 bp, melting temperature (Tm) of primers between 60 and 75°C, a maximum difference of Tm between primers of 10°C, and GC content between 40 and 60%. Several criteria were used to select 60 markers among the 10,372 candidates with designed primers: the selection of one primer set per sequence, the exclusion of motifs consisting exclusively of CG or AT, the retention of only perfect or compound microsatellites, ensuring minimal distance between primers and microsatellites, retaining loci with at least 10 repeats, and excluding loci whose primers contain multiple identical repeated bases. The 60 selected SSR primer pairs were tagged at the 5'-end with universal Illumina adapter overhang sequences: TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG for forward primers, and GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG for reverse primers. These primers were ordered with standard desalted purification at Integrated DNA Technologies (Coralville, USA) and tested in simplex PCR. SSR markers that successfully amplified in simplex PCR were amplified in a multiplexed PCR following Lepais et al. (2020), but using a single-step multiplexed PCR before the indexing PCR, and sequenced on an Illumina iSeq 100 2 × 150 bp. The genotyping of 107 samples was blind-repeated to assess the genotyping error rate based on genotype reproducibility using a pipeline (Lepais et al. 2020) integrating FDSTools v.1.2.0 (Hoogenboom et al. 2016). We kept as validated loci those showing less than 20% of missing genotypes across samples and less than 3% of allelic genotyping error estimated based on repeated genotypes.

Size homoplasy, genetic diversity, and heterozygosity

Size homoplasy was calculated for 25 high-quality validated loci as the number of alleles differing in sequence minus the number of alleles differing by length and divided by the number of alleles differing by their length. Data analyses were conducted per locus on the entire set of individuals (n = 107) and based on two different datasets, for which alleles were coded according to the amplicon length, and the sequence identity. The genetic variability of each dataset was characterized by locus by the number of alleles (N_A), the effective number of alleles (N_E), the allelic richness (R), the observed heterozygosity (H_O), and the expected heterozygosity (H_E). Rarefied allelic richness for a random subsample of gene copies (k = 166) was calculated based on the minimum sample size per locus (n = 83). The genetic diversity estimates were computed in SPAGeDi v.1.5a (Hardy and Vekemans 2002). Significant differences for N_A, N_E, R, H_O, and H_E between the amplicon length and the sequence identity datasets were tested using the paired Wilcoxon signed-rank test in R v.4.4.1 (R Core Team 2024).

Biologically informed statistical analyses were performed on population samples to determine the reliability of the SSRseq data for subsequent population genetic analyses. Genotyping errors and null alleles were assessed using Micro-Checker v.2.2.3 (Van Oosterhout et al. 2004). Hardy-Weinberg equilibrium and linkage disequilibrium were tested using the MCMC algorithm with the following parameters: 1000 iterations per batch, 100 batches, and 1000 dememorization steps, using Genepop v.4.7.5 (Raymond and Rousset 1995). The genetic diversity parameters (N_A, N_E, R, H_O, H_E) and the inbreeding coefficient F_{IS} were estimated and compared between the two life stages of the forest from the Paranaense region. To standardize comparisons, the rarefied allelic richness (R) was calculated using the minimum sample size (n = 26). Significant differences between the life stages were also tested as described above.

RESULTS

Marker development and validation

Forty-six developed loci out of 60 were successfully amplified in the simplex PCR test and 25 were validated as good-quality loci after amplification of the 46 loci in a single multiplex and sequencing (Table 1; Supplementary material 1). These 25 SSR loci produced high-quality genotypes for 107 *A. colubrina* samples (4.45% of missing genotypes and 0.8% of genotyping error).

Table 1. Characterization of the 25 validated SSRseq loci developed to *Anadenanthera colubrina* and genetic diversity indices per locus for the two different datasets based on allele length, and SSRseq (sequence-identity).

Locus		Primer sequences (5' – 3')	Missing rate	Genotyping error rate	SSR-length						SSR-seq					
					N	N_A	N_E	*R*	H_O	H_E	*N*	N_A	N_E	*R*	H_O	H_E
SSRseq.A06	F:	CGATCCACATGGTTCGCTGGGTCAT	7.48%	1.20%	99	7	4.107	6.93	0.455	0.760	99	28	9.599	24.67	0.505	0.900
	R:	GGCATTTCAACATATAGGACCACCA														
SSRseq.A14	F:	TGCACTTAATGTGGAAGGGATTGCA	0.93%	0.00%	106	4	2.103	3.68	0.509	0.527	106	7	2.969	6.36	0.698	0.666
	R:	TTGCCAACAAAGATTTCCTGGAGTT														
SSRseq.A17	F:	GTGGTTGAACGCCGCCCATTCTCAA	0.00%	0.00%	107	6	1.874	5.86	0.477	0.469	107	8	1.992	7.85	0.523	0.500
	R:	TCTCGAGATAGATGATTTGTCCAGGA														
SSRseq.A19	F:	GAAACTTGAAAGCAATTAGGCGGGT	14.95%	1.27%	91	15	8.407	14.33	0.857	0.886	91	18	11.183	17.32	0.901	0.916
	R:	CGGAGAGCCTTCTGTGCCTTGAGCA														
SSRseq.A20	F:	TCCCGACTAACCCTGACTTGCCACT	13.08%	1.23%	93	11	4.955	10.85	0.505	0.802	93	18	5.715	16.72	0.527	0.829
	R:	ACGGATCCACGTTCTGCAATGTATG														
SSRseq.A21	F:	ACTGCAAAGATAATGCCAACATGTC	0.93%	0.00%	106	6	1.787	5.65	0.358	0.443	106	11	2.412	9.67	0.491	0.588
	R:	TGCCTAAGTGGTCAGGTCTCATTCA														
SSRseq.A27	F:	ACCTCACATTTAACACCACAAGCCC	0.00%	0.00%	107	5	1.995	4.35	0.467	0.501	107	11	2.452	9.54	0.617	0.595
	R:	TGGCTAGTGAGGAAGACGAAGACGA														
SSRseq.A29	F:	AGCTCAGCTCTTTCTTTCATACGCA	0.93%	0.00%	106	9	2.826	8.47	0.604	0.649	106	12	3.664	10.83	0.651	0.730
	R:	CCGAGTTTGTGTTGCACCCAGCTCA														
SSRseq.A30	F:	AATCATCTAACACGCAGCCTCACTT	1.87%	0.55%	105	6	2.069	5.8	0.571	0.519	105	14	5.087	12.86	0.829	0.807
	R:	AGGCCGGATACATGGTTTGCTGACC														
SSRseq.A33	F:	AGCTCTGCTTCAATGGCGGAACTGA	0.93%	0.56%	106	10	5.139	9.57	0.840	0.809	106	11	5.219	10.25	0.84	0.812
	R:	CGGTGTTGGTTACTGGCAACCCACC														
SSRseq.A34	F:	GCCGCCTACTATACCAAGCCATGCA	2.80%	1.69%	104	13	3.613	12.22	0.721	0.727	104	27	10.594	24.40	0.846	0.910
	R:	ACTGCGTCTAACCCATATGTGATGGT														
SSRseq.A37	F:	GCCAATATTTAAGACCGGCGTGACCA	0.93%	0.00%	106	9	2.823	7.93	0.613	0.649	106	24	6.658	20.95	0.858	0.854
	R:	ACCTAATCTGGAGACGACCGTCCGA														
SSRseq.A38	F:	GGTTAACGACCCAAGAGCAATAAGA	14.02%	1.95%	92	7	2.851	6.78	0.522	0.653	92	9	2.881	8.69	0.522	0.656
	R:	GGGATTGAGTGGTGAAGTGTAGAAA														
SSRseq.A39	F:	TTCCTCTCCTTCTCCGCCACTTGCC	0.00%	0.00%	107	6	3.140	5.66	0.701	0.685	107	20	6.947	18.14	0.869	0.860
	R:	ACGCGCCGTTTCATCTGGTTGATGC														

Table 1 (continued). Characterization of the 25 validated SSRseq loci developed to *Anadenanthera colubrina* and genetic diversity indices per locus for the two different datasets based on allele length, and SSRseq (sequence-identity).

Locus		Primer sequences (5' – 3')	Missing rate	Genotyping error rate	SSR-length						SSR-seq					
					N	N_A	N_E	*R*	H_O	H_E	*N*	N_A	N_E	*R*	H_O	H_E
SSRseq.A40	F:	TGCAGAGGTATTTGAAATTAGGGCT	1.87%	0.00%	105	12	4.377	11.62	0.790	0.775	105	21	6.417	19.62	0.867	0.848
	R:	ATGATCAGTGGACCCATTGACCTGA														
SSRseq.A43	F:	AGGAATCATTGCACACCCAAAGATGA	22.43%	2.67%	83	11	4.097	10.93	0.217	0.760	83	12	4.108	11.93	0.217	0.761
	R:	GGCCGTCAATCGCTAGTGGCAGAAG														
SSRseq.A44	F:	GCTAGGCCACTCCACAACATTGCAGG	3.74%	0.55%	103	13	5.945	12.29	0.864	0.836	103	30	8.354	26.00	0.903	0.885
	R:	TCGAGGAGATTAGGTGGTGACTTGT														
SSRseq.A45	F:	TGCTTCCACGACGTTATTCTCTAGCA	8.41%	2.75%	98	11	4.103	10.64	0.653	0.760	98	13	4.315	12.34	0.653	0.772
	R:	CCGAGATGCAGGCTATCTGTTCAAC														
SSRseq.A47	F:	TTTCCGTCTCTGTCTTCCTGCTATA	2.80%	0.57%	104	10	4.27	9.36	0.798	0.770	104	15	6.516	13.86	0.846	0.851
	R:	TGCCTTCCTCCATGCTGTTATCTGC														
SSRseq.A48	F:	CGCGAACTTCACTTTGGCGTAGGTG	2.80%	1.12%	104	10	5.478	9.66	0.865	0.821	104	13	5.854	12.23	0.865	0.833
	R:	GCGAGCTGTTGCAATGCCGGAATTG														
SSRseq.A51	F:	CCCTTTGCAGTTTATGGTCCCAGCA	1.87%	0.00%	105	4	2.265	3.90	0.600	0.561	105	8	2.409	7.03	0.61	0.588
	R:	GGACTTATGGGATTGGGCCGAGAG														
SSRseq.A54	F:	AAAGCTCTCGCCGTTCAAACCTGCC	0.93%	0.00%	106	6	1.629	5.87	0.396	0.388	106	6	1.629	5.87	0.396	0.388
	R:	TGACGATTAGGAGGGCGAGCTCTGA														
SSRseq.A55	F:	GGGAACAGAAGCGGGAATCTTGAAG	2.80%	2.75%	104	8	1.735	7.40	0.433	0.426	104	9	1.738	8.10	0.442	0.427
	R:	TGCATCAGCCTGCCACTTGCATGAT														
SSRseq.A59	F:	ACATGAAGCAGCTGATTGAGGAAAGT	0.93%	0.00%	106	12	4.586	11.14	0.717	0.786	106	34	8.496	28.60	0.84	0.886
	R:	CACAATCCTGCCTTGTGGGTCCAACA														
SSRseq.A60	F:	TGAACAGGAACTTGTTGGCGGAGGG	3.74%	1.10%	103	8	1.709	6.99	0.369	0.417	103	12	3.722	10.99	0.66	0.735
	R:	CGGCCTCTTTGTCCACCTTCCCAGT														
Mean			**4.45%**	**0.80%**	**102**	**9**	**3.515**	**8.32**	**0.596**	**0.655**	**102**	**16**	**5.237**	**14.19**	**0.679**	**0.744**
SE			0.012	0.002	1.237	0.606	0.334	0.585	0.036	0.031	1.237	1.553	0.561	1.323	0.038	0.031

N: Number of genotyped samples; N_A: mean number of alleles per locus; N_E: effective number of alleles; *R*: allelic richness estimated after rarefaction to 83 samples; H_O: observed heterozygosity; H_E: expected heterozygosity; SE: Standard error.

Table 2. Genetic diversity indices per life stage for the dataset based on 22 SSRseq of *Anadenanthera colubrina* from a forest site of the Paranaense region.

Life stages		N_A	N_E	R	H_O	H_E	F_{IS}
Adults (N = 31)	**Mean**	10.955	6.226	10.488	0.751	0.773	0.023
	SE	0.913	0.831	0.851	0.031	0.029	0.026
Seedlings (N = 64)	**Mean**	9.136	4.116	7.651	0.696	0.679	-0.018
	SE	0.836	0.454	0.652	0.043	0.037	0.021
Total	**Mean**	**12.500**	**4.919**	**9.175**	**0.714**	**0.720**	**0.009**
	SE	1.095	0.605	0.743	0.036	0.034	0.018

N: Number of genotyped samples; N_A: mean number of alleles per locus; N_E: effective number of alleles; R: allelic richness estimated after rarefaction to 26 samples; H_O: observed heterozygosity; H_E: expected heterozygosity; F_{IS}: inbreeding coefficient; SE: Standard error.

Genetic diversity, heterozygosity, and advantage of sequence to detect size homoplasy

The microsatellite loci were highly polymorphic in both sequences and length for most of them (Table 1). A high number of alleles per locus was detected in both datasets. Significantly more alleles were observed per locus based on sequence identity (16) than based on allele length (9) and expected heterozygosity was also higher in the sequence-based dataset (H_E = 0.744 vs H_E = 0.655) (Table 1). Size homoplasy reached a mean of 97.85% across loci, or, in other words, for each allele observed based on length, next to two (1.98) alleles were observed based on sequence. The number of alleles, allelic richness, and both observed and expected heterozygosity showed significant differences between allele calling methods ($p = 4.645 \times 10^{-7}$ for N_A, $p = 2.751 \times 10^{-5}$ for N_E, $p = 2.019 \times 10^{-7}$ for R, $p = 4.768 \times 10^{-5}$ for H_O, and $p = 8.108 \times 10^{-6}$ for H_E) (Table 1).

The SSRseq adult trees dataset from the Paranaense forest showed no evidence of scoring errors due to stuttering or large allele dropout across all 25 loci in the genotyping error analysis. However, null alleles were detected at three loci (*SSRseq.A06*, *SSRseq.A20*, *SSRseq.A43*), potentially causing an excess of homozygosity. Consequently, these loci were excluded from subsequent population analyses. No statistically significant deviations from Hardy-Weinberg equilibrium were detected in none of the two life stages, despite expectations of higher relatedness due to family structure in the seedlings. No evidence of linkage disequilibrium was observed between pairs of loci.

High population genetic diversity was detected in adults and seedlings from the Paranaense forest (based on 22 loci). The mean and effective numbers of alleles per locus (N_A, N_E), observed and expected heterozygosity (H_O, H_E), and the inbreeding coefficient (F_{IS}) did not show significant differences between life stages. However, rarefied allelic richness (R) values were significantly higher in adults than in seedlings (p = 0.023) (Table 2).

DISCUSSION

Development of SSRseq markers

High-throughput sequencing allowed the de novo development of an NGS-based multiplex marker panel for *Anadenanthera colubrina*. The effectiveness of SSRseq strengthens the advantage of using modern NGS platforms for qualitatively and quantitatively increasing data in large-scale population genetic studies.

The quality of DNA extraction is a crucial starting point for genetic studies. For Leguminosae forest tree species, the CTAB method is highly recommended because it is an inexpensive and rapid protocol that continues to be widely used in several species of the family (Riahi et al. 2010; Silva et al. 2023; Caycho et al. 2023; da Rocha et al. 2024).

Microsatellites have been used as highly polymorphic markers in previous population genetic studies of *Anadenanthera colubrina*, although the number of nuclear and plastid SSRs was limited (Barrandeguy et al. 2014; Goncalves et al. 2019; Feres et al. 2021). The de novo development of SSRs using traditional methods constitutes a time-consuming and costly endeavour (Edwards et al. 1996), particularly for non-model species such as *A. colubrina*, of which genomic resources are scarce. Conversely, the newer SSRseq technology is an efficient method for obtaining high-quality and informative species-specific nuclear microsatellite markers in non-model species. Additionally, multiplexing samples in a single sequencing run using individual-specific barcoding reduces the cost and time required for marker development and large-scale genotyping (Guo et al. 2020).

Sequence information reduces size homoplasy at microsatellite loci

The 25 new SSRseq markers provide an additional advantage over traditional fragment-length genotyping.

Here, the NGS-based method allowed us to detect a high percentage of size homoplasy and to resolve almost double the number of alleles than expected based on fragment length accessible by capillary electrophoresis-based SSR. The generated datasets increase the resolution for more accurate eco-evolutionary inference in *A. colubrina* populations. The percentage of increase in the detected number of alleles due to sequence analysis resulted as high as 97.85% in *A. colubrina*, while in other tree species, it was 36% in chestnut (*Castanea sativa*; *C. crenata*; Laurent et al. 2020), and 55% on oaks (*Quercus faginea* and *Q. canariensis*; Lepais et al. 2022).

Beyond detecting size homoplasy, the sequencing approach expands the scope of SSRs due to the use of compound marker systems integrating linked polymorphisms with different mutational dynamics, such as a microsatellite and its flanking sequences, allowing improvements in the estimation of population structure and inferences of demographic history (Payseur and Cutter 2006). This is especially relevant in the case of complex population dynamics (Lepais et al. 2022). Hence, the development and analysis of SSRseq provide an effective means to enhance our understanding of the processes affecting genetic diversity patterns, offering a valuable tool for informing conservation and management decisions.

Genetic diversity for seedling and adult trees

High levels of population genetic diversity were detected in adults and seedlings from a Paranaense forest. The mean number of alleles per locus (N_A = 12.50) and the allelic richness after rarefaction (R = 10.49 in adults) were particularly high. These estimates exceed those reported in previous studies of *A. colubrina* at various spatial scales, which employed traditional nuclear microsatellites. For example, a survey of two populations located in northeastern São Paulo state, Brazil, detected a mean N_A = 7 (Feres 2013), while an analysis of four populations from the Yungas and Paranaense regions of Argentina found N_A = 9.72 and R = 6.75 (Barrandeguy et al. 2014). More recently, the genetic diversity of 79 populations representing four *A. colubrina* varieties from Brazil and northeastern Bolivia was characterized using 12 SSR loci, revealing a range of N_A = 3.36–5.59 (Mangaravite et al. 2023). Additionally, in other tree species from Seasonally Dry Tropical Forests, such as *Tibouchina papyrus* (R = 1.50–2.10, n = 66; Collevatti et al. 2012), and *Myroxylon peruiferum* (R = 2.73–3.01, n = 37; Silvestre et al. 2018), the estimated allelic richness was lower than that reported here.

Forest tree species generally exhibit a high genetic diversity, primarily distributed within populations, a pattern often attributed to life history traits such as longevity and outcrossing mating systems (Hamrick and Godt 1996; Petit and Hampe 2006). In our sample, adults had a higher allelic richness than seedlings (standardized for the same sampling size), which was expected considering that seedlings represented four half-sib families, i.e. 50% of their alleles corresponding to the maternal copies came from four mothers only, whereas adults represented a sample of naturally regenerated trees without a priori family structure. The overall high genetic diversity detected in this study may result from synergistic factors, including elevated multilocus outcrossing rates (Goncalves et al., manuscript in preparation), which suggest substantial pollen-mediated gene flow, and high mutation rates at SSR loci, further amplified by the polymorphisms identified within the SSRs and their flanking sequences.

The low inbreeding coefficients detected contrast with those reported in a previous study on fine-scale population genetic structure in *A. colubrina*, where both adults and saplings exhibited high F_{IS} coefficients (Goncalves et al. 2019). These discrepancies may arise from biological differences in the mating system, from differences in sampling design and/or differences in the nature of molecular marker variation. In the present study, distances between mother trees were kept above 50 m to minimize the likelihood of sampling closely related individuals, consistent with previous findings on a mixed mating system and relatively short-range seed dispersal in *A. colubrina* populations (Goncalves et al. 2019; Feres et al. 2021). Non-significant F_{IS} values in combination with high polymorphism in our study suggest high outcrossing rates and gene flow in Paranaense populations. The approximate selfing rate, s, estimated from the inbreeding coefficient using $s = 2F_{IS}/(1 + F_{IS})$ (Hartl and Clark 2007), suggests 4.5% selfing in this forest site. Therefore, the influence of the mating system and pollen dispersal on the genetic diversity of natural populations of *A. colubrina* in remnant forests can now be studied using SSRseq. This type of molecular marker may help foster research into non-model species, such as a recent study in other Neotropical forest tree species (Corvalán et al. 2023). This approach emphasizes the importance of understanding evolutionary and ecological processes and their impacts on ecosystem responses, thereby promoting the sustainability of South American fragmented native forests.

Recent advances in molecular and computational techniques, combined with transdisciplinary research that integrates ecology, evolution, and genetics, are crucial for understanding and conserving processes that support plant genetic diversity in a changing world. Overall, the development of SSRseq is a powerful tool for genetic analysis and can be used to identify genetic variation and diversity within and among populations, which can be useful for sustainable management and conservation policy.

CONCLUSIONS AND PERSPECTIVES

The primers designed for SSR high-throughput genotyping-by-sequencing are promising genetic tools

useful in many population genetics applications such as genetic characterization of entire populations with less sequencing effort. The developed SSRseq markers may be particularly advantageous for efficient analysis of genetic diversity, providing renewed opportunities to explore ecological and evolutionary processes that shape population genetic structure in non-model species such as *A. colubrina*.

The comparison between allele length and SSR sequence identity revealed that SSRseq are more informative markers than traditional SSRs due to their sequence-based nature, allowing for greater variability in repeat numbers and flanking sequences. The SSRseq development involves clear criteria and allows, after proper laboratory testing, multiplexing, and high-throughput genotyping, simultaneously enabling efficient and cost-effective analysis of multiple markers. Therefore, this study builds the basis for new approaches using the information provided by NGS technologies that can also be used for developing molecular markers for genotyping on a large scale mainly in population genetic and genomic studies.

DATA AVAILABILITY

Raw data from the shotgun whole genome sequencing are available in the Sequence Read Archive (SRA) under BioProject PRJNA1033700 with SRA number SRR26587918 from the National Center for Biotechnology Information Repository. Genotype data for every individual and microsatellite loci are available on Zenodo: https://doi.org/10.5281/zenodo.11106586.

ACKNOWLEDGEMENTS

Technical developments and sequencing were performed at the PGTB (https://doi.org/10.15454/1.5572396583599417E12) with the help of Z. Compagnie and E. Guichoux. The authors thank C. Lalanne for their technical assistance. Also, ALG wishes to thank "Consejo Nacional de Investigaciones Científicas y Técnicas" (CONICET) for providing a postdoctoral fellowship for a short research stay at INRAE. This research has benefited from the support of a grant from "Investissement d'Avenir" grants of the French National Research Agency (CEBA:ANR-10-LABX-25–01) to MH and ALG. Also, this research was partially supported by a multiannual research project (PIP N° 112–2015001-00860CO) from "Consejo Nacional de Investigaciones Científicas y Técnicas" (CONICET) to MVG.

REFERENCES

Barrandeguy ME, Prinz K, García MV, Finkeldey R (2012) Development of microsatellite markers for *Anadenanthera colubrina* (Fabaceae), a native tree from South America. American Journal of Botany 99(9): e372–e374. https://doi.org/10.3732/ajb.1200078

Barrandeguy ME, García MV, Prinz K, Rivera Pomar R, Finkeldey R (2014) Genetic structure of disjunct Argentinean populations of the subtropical tree *Anadenanthera colubrina* var. *cebil* (Fabaceae). Plant Systematics and Evolution 300: 1693–1705. https://doi.org/10.1007/s00606-014-0995-y

Barrandeguy ME, Prado DE, Goncalves AL, García MV (2016) Demografía histórica de *Anadenanthera colubrina* var. *cebil* (Leguminosae) en Argentina. Boletín de la Sociedad Argentina de Botánica 51(4): 689–703.

Bushnell B, Rood J, Singer E (2017) BBMerge – Accurate paired shotgun read merging via overlap. PLoS ONE 12(10): e0185056. https://doi.org/10.1371/journal.pone.0185056

Calonga Solís V, Barrandeguy ME, García MV (2014) Divergencia histórica en *Anadenanthera colubrina* var. *cebil* (Leguminosae) analizando una región intrónica del ADN cloroplástico. Boletín de la Sociedad Argentina de Botánica 49(4): 547–557.

Caycho E, La Torre R, Orjeda G (2023) Assembly, annotation and analysis of the chloroplast genome of the Algarrobo tree *Neltuma pallida* (subfamily: Caesalpinioideae). BMC Plant Biology 23(1): 570. https://doi.org/10.1186/s12870-023-04581-5

Collevatti RG, Terribile LC, Lima-Ribeiro MS, Nabout JC, de Oliveira G, Rangel TF, Rabelo SG, Diniz-Filho JAF (2012) A coupled phylogeographical and species distribution modelling approach recovers the demographical history of a Neotropical seasonally dry forest tree species. Molecular Ecology 21: 5845–5863. https://doi.org/10.1111/mec.12071

Corvalán LC, Carvalho LR, Melo-Ximenes AA, Targueta CP, Braga-Ferreira RS, Nunes R, Telles MP (2023) Data of SSRs primers for high-throughput genotyping-by-sequencing (SSR-Seq) based on the partial genome assembly of *Eugenia klotzschiana* (Myrtaceae). Data in Brief 19: 108917. https://doi.org/10.1016/j.dib.2023.108917

Darby BJ, Erickson SF, Hervey SD, Ellis-Felege SN (2016) Digital fragment analysis of short tandem repeats by high-throughput amplicon sequencing. Ecology and Evolution 6(13): 4502–4512. https://doi.org/10.1002/ece3.2221

da Rocha VD, Dal'Sasso, TC, Williams CCV, Simon MF, Bueno ML, de Oliveira LO (2024) From forest to savanna and back to forest: Evolutionary history of the genus *Dimorphandra* (Fabaceae). Journal of Plant Research 137: 377–393. https://doi.org/10.1007/s10265-024-01523-6

de Viana ML, Giamminola E, Russo R, Ciaccio M (2014) Morphology and genetics of *Anadenanthera colubrina* var. *cebil* (Fabaceae) tree from Salta (Northwestern Argentina). Revista de Biología Tropical 62(2): 757–767. https://doi.org/10.15517/rbt.v62i2.10404

Doyle J (1991) DNA protocols for plants. In: Hewitt GM, Johnston AWB, Young JPW (Eds) Molecular Techniques in Taxonomy NATO ASI Series 57. Springer, Berlin, 283–293. https://doi.org/10.1007/978-3-642-83962-7_18

DRYFLOR (2016) Plant diversity patterns in neotropical dry forests and their conservation implications. Science 353(6306): 1383–1387. https://doi.org/10.1126/science.aaf5080

Edwards KJ, Barker JHA, Daly A, Jones C, Karp A (1996) Microsatellite libraries enriched for several microsatellite sequences in plants. BioTechniques 20(5): 758–760. https://doi.org/10.2144/96205bm04

Ellegren H (2004) Microsatellites: simple sequences with complex evolution. Nature Reviews Genetics 5: 435–445. https://doi.org/10.1038/nrg1348

Estoup A, Jarne P, Cornuet JM (2002) Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. Molecular Ecology 11: 1591–1604. https://doi.org/10.1046/j.1365-294X.2002.01576.x

Feres JM (2013) Diversidade genética, fluxo génico e sistema de cruzamento de *Anadenanthera colubrina* (Vell.) Brenan e *Anadenanthera peregrina* (L.) Speg: duas espécies que ocorrem em alta densidade no interior do Estado de São Paulo. PhD Thesis, University of São Paulo, Brazil. https://doi.org/10.11606/T.17.2014.tde-21052014-105106

Feres JM, Monteiro M, Zucchi MI, Pinheiro JB, Mestriner MA, Alzate-Marin AL (2012) Development of microsatellite markers for *Anadenanthera colubrina* (Leguminosae), a neotropical tree species. American Journal of Botany 99(4): e154–e156. https://doi.org/10.3732/ajb.1100446

Feres JM, G Nazareno A, Borges LM, Corbo Guidugli M, Bonifacio-Anacleto F, Alzate-Marin AL (2021) Depicting the mating system and patterns of contemporary pollen flow in trees of the genus *Anadenanthera* (Fabaceae) PeerJ 9: e10579 https://doi.org/10.7717/peerj.10579

Goncalves AL, García MV, Heuertz M, González-Martínez SC (2019) Demographic history and spatial genetic structure in a remnant population of the subtropical tree *Anadenanthera colubrina* var. *cebil* (Griseb.) Altschul (Fabaceae). Annals of Forest Science 76(1): 18. https://doi.org/10.1007/s13595-019-0797-z

Guo L, Yang Q, Yang JW, Zhang N, Liu BS, Zhu KC, Guo H-Y, Jiang S-G, Zhang DC (2020) MultiplexSSR: a pipeline for developing multiplex SSR-PCR assays from resequencing data. Ecology and Evolution 10(6): 3055–3067. https://doi.org/10.1002/ece3.6121

Hamrick JL, Godt MJW (1996) Effects of life history traits on genetic diversity in plant species. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 351(1345): 1291–1298. https://doi.org/10.1098/rstb.1996.0112

Hardy OJ, Vekemans X (2002) SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Molecular Ecology Notes 2(4): 618–620. https://doi.org/10.1046/j.1471-8286.2002.00305.x

Hartl DL, Clark AG (2007) Principles of Population Genetics. Fourth edition. Sinauer Associates, Sunderland, 1–653.

Hoogenboom J, De Knijff P, Laros JFJ, De Leeuw RH, Van der Gaag KJ, Sijen T (2016) FDSTools: a software package for analysis of massively parallel sequencing data with the ability to recognise and correct STR stutter and other PCR or sequencing noise. Forensic Science International: Genetics 27: 27–40. https://doi.org/10.1016/j.fsigen.2016.11.007

Kimura M, Ohta T (1978) Stepwise mutation model and distribution of allelic frequencies in a finite population. PNAS 75(6): 2868–2872. https://doi.org/10.1073/pnas.75.6.2868

Laurent B, Larue C, Chancerel E, Guichoux E, Petit RJ, Barreneche T, Robin C, Lepais O (2020) Microhaplotype genotyping-by-sequencing of 98 highly polymorphic markers in three chestnut tree species. Conservation Genetics Resources 12: 567–580. https://doi.org/10.1007/s12686-020-01157-5

Lepais O, Chancerel E, Boury C, Salin F, Manicki A, Taillebois L, Dutech C, Aissi A, Bacles CF, Daverat F, Launey S (2020) Fast sequence-based microsatellite genotyping development workflow. PeerJ 8: e9085. https://doi.org/10.7717/peerj.9085

Lepais O, Aissi A, Véla E, Beghami Y (2022) Joint analysis of microsatellites and flanking sequences enlightens complex demographic history of interspecific gene flow and vicariance in rear-edge oak populations. Heredity 129: 169–182. https://doi.org/10.1038/s41437-022-00550-0

Mangaravite E, Silveira TC, Vinson CC, Bueno ML, Silva RS, Carniello MA, Veldman JW, Gonçalves MG, Oliveira LO (2023) Unlocking the secret diversity of *Anadenanthera*: insights from molecular genetics of four evolving species. Botanical Journal of the Linnean Society 37: 1–16. https://doi.org/10.1093/botlinnean/boad037

Meglécz E, Pech N, Gilles A, Dubut V, Hingamp P, Trilles A, Grenier R, Martin JF (2014) QDD version 3.1: a user-friendly computer program for microsatellite selection and primer design revisited: experimental validation of variables determining genotyping success rate. Molecular Ecology Resources 14: 1302–1313. https://doi.org/10.1111/1755-0998.12271

Payseur BA, Cutter AD (2006) Integrating patterns of polymorphism at SNPs and STRs. Trends in Genetics 22(8): 424–429. https://doi.org/10.1016/j.tig.2006.06.009

Petit RJ, Hampe A (2006) Some evolutionary consequences of being a tree. Annual Review of Ecology, Evolution, and Systematics 37: 187–214. https://doi.org/10.1146/annurev.ecolsys.37.091305.110215

Ramakrishnan U, Mountain JL (2004) Precision and accuracy of divergence time estimates from STR and SNPSTR variation. Molecular Biology and Evolution 21(10): 1960–1971. https://doi.org/10.1093/molbev/msh212

Raymond M, Rousset F (1995) GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. Journal of Heredity 86: 248–249. https://doi.org/10.1093/oxfordjournals.jhered.a111573

R Core Team (2024) R: a language and environment for statistical computing. Version 4.4.1. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/ [accessed 17.12.2024]

Riahi M, Zarre S, Maassoumi AA, Attar F, Kazempour Osaloo S (2010) An inexpensive and rapid method for extracting papilionoid genomic DNA from herbarium specimens. Genetics and Molecular Research 9(3): 1334–1342. https://doi.org/10.4238/vol9-3gmr839

Šarhanová P, Pfanzelt S, Brandt R, Himmelbach A, Blattner FR (2018) SSR-seq: genotyping of microsatellites using next-generation sequencing reveals higher level of polymorphism as compared to traditional fragment size scoring. Ecology and Evolution 8(22): 10817–10833. https://doi.org/10.1002/ece3.4533

Särkinen T, Iganci JR, Linares-Palomino R, Simon MF, Prado DE (2011) Forgotten forests - issues and prospects in biome mapping using Seasonally Dry Tropical Forests as a case study. BMC Ecology 11(1): 27. https://doi.org/10.1186/1472-6785-11-27

Schlötterer C (2000) Evolutionary dynamics of microsatellite DNA. Chromosoma 109(6): 365–371. https://doi.org/10.1007/s004120000089

Scotti-Saintagne C, de Sousa Rodrigues A, Roig A, Fady B (2024) A comprehensive strategy for the conservation of forest tree genetic diversity: an example with the protected *Pinus nigra* subsp. *salzmannii* (Dunal) Franco in France. Conservation Genetics 25(2): 469–480. https://doi.org/10.1007/s10592-023-01581-8

Silva, RWLD, Machado SS, Faria KDC, Oliveira FAD, Souza APD, Menezes IPPD, Silva J MD (2023) Molecular insight for baru *Dipteryx alata* (Fabaceae) populations based on novel SSRs. Acta Botanica Brasilica 37: e20220168. https://doi.org/10.1590/1677-941X-ABB-2022-0168

Silvestre EDA, Schwarcz KD, Grando C, de Campos JB, Sujii PS, Tambarussi EV, Macrini CMT, Pinheiro JB, Brancalion PHS, Zucchi MI (2018) Mating system and effective population size of the overexploited Neotropical tree (*Myroxylon peruiferum* Lf) and their impact on seedling production. Journal of Heredity 109(3): 264–271. https://doi.org/10.1093/jhered/esx096

Stefanini C, Csilléry K, Ulaszewski B, Burczyk J, Schaepman ME, Schuman MC (2023) A novel synthesis of two decades of microsatellite studies on European beech reveals decreasing genetic diversity from glacial refugia. Tree Genetics & Genomes 19: 3. https://doi.org/10.1007/s11295-022-01577-4

Van Oosterhout C, Hutchinson WF, Wills DP, Shipley P (2004) MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. Molecular Ecology Notes 4(3): 535–538. https://doi.org/10.1111/j.1471-8286.2004.00684.x

Vartia S, Villanueva-Cañas JL, Finarelli J, Farrell ED, Collins PC, Hughes GM, Carlsson JE, Gauthier DT, McGinnity P, Cross TF, FitzGerald RD (2016) A novel method of microsatellite genotyping-by-sequencing using individual combinatorial barcoding. Royal Society Open Science 3(1): 150565. https://doi.org/10.1098/rsos.150565

Wang X, Wang L, Sun Y, Chen J, Liu Q, Dong S (2023) Genetic diversity and conservation of Siberian apricot (*Prunus sibirica* L.) based on microsatellite markers. Scientific Reports 13(1): 11245. https://doi.org/10.1038/s41598-023-37993-2

Zerda Moreira A, García MV, Barrandeguy ME (2024) Unravelling the evolutionary history and promoting conservation genetics of *Anadenanthera colubrina* var. *cebil* (Leguminosae), a paradigmatic species in Seasonally Dry Tropical Forests. Botanical Journal of the Linnean Society 205(2): 177–189. https://doi.org/10.1093/botlinnean/boad078

SUPPLEMENTARY MATERIAL

Supplementary material 1

Characterization of the 60 SSRseq markers generated from low-coverage shotgun sequencing of a single *Anadenanthera colubrina* library.

https://doi.org/10.5091/plecevo.138834.suppl1