Data in Brief 66 (2026) 112855

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Draft genome assemblies and SSR-Seq dataset for *Simarouba amara* and *S. versicolor*, native species to Latin America

Marla A. Almeida-Silva[a,b,*], Leonardo C.J. Corvalán[a,c], Ramilla S. Braga-Ferreira[a,g], Cíntia P. Targueta[a], Edivani V. Franceschinelli[d], Carlos M. Silva-Neto[e,f], Thannya N. Soares[a], Rhewter Nunes[a,c], Mariana P.C. Telles[a,h]

[a] *Genetics & Biodiversity Laboratory (LGBio), Federal University of Goiás, Goiânia, GO, Brazil*
[b] *State University of Piauí,* Campus *Prof. Ariston Dias Lima, São Raimundo Nonato, PI, Brazil*
[c] *Bioinformatics and Biodiversity Laboratory (LBB), State University of Goiás, Academic Institute of Health and Biological Sciences, West Campus, University Unit of Iporá, Iporá, GO, Brazil*
[d] *Laboratory of Plant Reproductive Biology, Federal University of Goiás, Goiânia, GO, Brazil*
[e] *Federal Institute of Goiás-Innovation Hub, Goiânia, GO, Brazil*
[f] *State University of Goiás, University Unit of Ipameri, GO, Brazil*
[g] *Federal University of Rondonópolis, Institute of Exact and Natural Sciences, Rondonópolis, MT, Brazil*
[h] *School of Medical and Life Sciences, Pontifical Catholic University of Goiás, Goiânia, GO, Brazil*

ARTICLE INFO

Article history:
Received 17 March 2026
Revised 21 April 2026
Accepted 11 May 2026
Available online 15 May 2026

Dataset link: Simarouba amara isolate SamPIRGO, whole genome shotgun sequencing project (Original data)
Dataset link: Simarouba versicolor isolate Sve-GOIGO, whole genome shotgun sequencing project (Original data)

ABSTRACT

Simarouba amara (known in Brazilian Portuguese as "marupá") and *S. versicolor* (known in Brazilian Portuguese as "pé-de-perdiz") are Neotropical species belonging to the family Simaroubaceae. These species have historically been used in folk medicine to treat conditions such as malaria, cancer, helminthiasis, viral infections, gastritis, ulcers, diarrhea, and diabetes. Recent advances in high-throughput sequencing (HTS) technologies have improved the acquisition of genomic datasets for economically wild species. This genomics data enables the development of microsatellite markers (SSR), which are valuable tools in genetic analysis, mainly in species with absence of genomic resources, as *S. amara* and *S. versicolor*. In this study, we generated high-quality draft assemblies and developed SSR-Seq primers

* Corresponding author.
E-mail address: marla.arianne@srn.uespi.br (M.A. Almeida-Silva).

https://doi.org/10.1016/j.dib.2026.112855

Keywords:
Assembly
Genotyping-by-sequencing
Markers microsatellites
Marupá
Pé-de-perdiz
Simaroubaceae

from these assemblies, for both species. We sequenced a total of 20,273,467 and 16,800,708 reads from *S. amara* and *S. versicolor*, respectively, with an estimated genome sizes of 372.16 Mb and 249.78 Mb. The genome assemblies by SPAdes resulted in 23,601 and 23,722 total contigs and an N50 value of 28,440 bp and 22,312 bp. Using the QDD pipeline, we identified 11,348 and 12,084 microsatellite regions that are putative for primers design. Using the openPrimeR tool, this dataset was filtered and 87 and 77 sets of SSR-Seq primers survived. Using physicochemical properties, 55 and 56 SSR-seq primer pairs for *S. amara* and *S. versicolor* were organized into five and four multiplex sets. The SSR-Seq dataset developed in this study enables the acquisition of genetic information and performs genetic and evolutionary analyses in these wild populations.

Specifications Table

Subject	Biological Sciences
Specific subject area	Genomics, Plant Science.
Type of data	Raw sequencing reads obtained by high-throughput sequencing (FASTQ), Genome assembly files (fasta), and primers (Table).
Data collection	Fresh leaves from four individuals of *Simarouba amara* and one individual of *S. versicolor* were collected in Pirenópolis-GO and Goiânia-GO, Brazil. DNA was extracted using the CTAB 2% protocol, and library preparation was performed using the Illumina DNA Prep Kit. Samples were sequenced on the Illumina MiSeq platform using the V3 600 cycles and V2 300 cycles kits, for *S. amara* and *S. versicolor*, respectively. Genome assemblies were performed using the SPADES tool and microsatellite marker extraction was performed using the QDD program. SSR-Seq were designed using the Primer3 tool, and primer multiplexing was performed using the openPrimeR tool.
Data source location	*Simarouba amara*: Pirenópolis, Goiás, Brazil (−15.80294, −48.84820)/ Collection data: 18 September 2019. *Simarouba versicolor:* Goiânia, Goiás, Brazil (−15.929415, −50.154083) Collection data: 2 March 2022.
Data accessibility	**Repository name:** Sequence Read Archive (SRA) Raw reads data link **Simarouba amara -** https://www.ncbi.nlm.nih.gov/sra/?term=SRR34954384 **Simarouba versicolor -** https://www.ncbi.nlm.nih.gov/sra/?term=SRR34954690 **Repository name:** Mendeley data Supplementary data link https://data.mendeley.com/datasets/rnsd2m5tc5/2 Repository name: NCBI Nucleotide Draft genome data link Simarouba amara accession link: https://www.ncbi.nlm.nih.gov/nuccore/JBSQFA000000000.1/ Simarouba versicolor accession link: https://www.ncbi.nlm.nih.gov/nuccore/JBSQFB000000000

1. Value of the Data

- The genomic data presented here represents the first data sets for species belonging to the genus *Simarouba*, and these data demonstrate the retrieval of genomic information. Genomic researchers working with this botanical group can benefit from this dataset and enhance their research with more complete and up-to-date data. While draft assemblies may be in-

complete or fragmented, they still provide critical insights and a reference framework that researchers can build upon.

- The Simple Sequence Repeat (SSR) primers derived from high-throughput sequencing (SSR-Seq) datasets have proven to be useful, including the assessment of evolutionary history, delineation of population structures, evaluation of the conservation status of the subject species, germplasm characterization, and incorporation into breeding programs.
- This study presents a dataset of multiplex primers, 55 primers in 5 multiplexes for *S. amara* and 56 primers in 4 multiplexes for *S. versicolor*, providing a valuable resource for studies population structure and phylogeography in these Neotropical tree species. Research using primers in multiplex increases throughput, reduces costs, and facilitates large-scale population-level sampling.
- The *in-silico* validation, in conjunction with the successful implementation of multiplexing and transferability assessments for the SSR-Seq markers delineated in this study, serves to enhance their applicability. For genomic studies of non-cultivated plants, these data are valuable because they allow new studies on native flora.
- A total of 27 SSR-Seq primer pairs designed for *S. amara* successfully amplified in *S. versicolor*, reduces costs without compromising data quality. Given the limited genetic markers for native species, cross-species amplification is a valuable tool for genetic studies.

2. Background

The motivation for this research stems from the potential and significance of generating molecular data sets for the neotropical species *S. amara* and *S. versicolor*. These species have gained renown for their application in the treatment of ailments such as malaria and cancer, particularly within the context of traditional communities. Furthermore, this study acknowledges the significance of native and endemic species as crucial sources for the study of biodiversity. The paucity of genomic and molecular resources has limited the understanding of the evolutionary history and genetic diversity of these species. Moreover, the present article contributes to the field by offering the initial genomic data and SSR-Seq primers derived from high-throughput sequencing technologies for the genus *Simarouba*. This provides a foundational framework for the identification of genetic variations and the study of genomic synteny. Furthermore, they enable comparative analyses between species and facilitate the acquisition of genetic information, which is critical for the conservation and management of wild populations.

3. Data Description

A total of 20,273,467 and 16,800,708 paired-end sequencing reads were generated from *S. amara* and *S. versicolor*, respectively, which were used for genome assembly. The assembled genome sizes are 286,173,048 bp and 241,977,620 bp for *S. amara* and *S. versicolor*, respectively, with GC content of 32.25% and 32.41%. The total number of contigs was 23,601 and 23,722, and the largest contig was 304,853 bp and 176,144 bp (Table 1) for *S. amara* and *S. versicolor*, respectively. The estimated genome sizes were 372,156,664 bp (best kmer 121) and 249,780,341 bp (best kmer 37) for *S. amara* and *S. versicolor*, respectively (Table 1). The discrepancy between the assembled and estimated genome sizes of *S. amara* (approximately 23%) can be attributed to repetitive genomic regions (e.g. transposable elements) and heterozygous regions were excluded during assembly with SPAdes. Besides, there are inherent limitations of short-read sequencing technologies. Data on genome size estimation by flow cytometry for the genus *Simarouba* are not available yet. Within the family Simaroubaceae, there is a record of genome size estimation by flow cytometry in *Ailanthus altissima* (1020 Mbp) [1]. The BUSCO (Benchmarking Universal Single-Copy Orthologs) completeness metrics showed 94.08% and 91.06% of complete single-copy orthologous genes considering a given taxonomic group, and the sequencing depth, with values of 36.25x and 17.97x, for *S. amara* and *S. versicolor*, respectively (see Fig. 1 and Table 1).

Table 1
Statistics of *de novo* genome assemblies of *S. amara* and *S. versicolor* using SPAdes software.

Parameter	Value	
	S. amara	*S. versicolor*
Number of contigs	32,007	471,109
Number of contigs ≥ 1000 bp	23,601	23,722
Total length	286,173,048 bp	241,977,620 bp
Estimated genome size	372,156,664 bp	249,780,341 bp
Genome completeness	94.08%	91.06%
Genome depth	36,25	17,97
Largest contig (bp)	304,853 bp	176,144 bp
Shortest contig (bp)	1000 bp	1000 bp
N50	28,440	22,312
L50	2823	3207
GC%	32.25	32.41

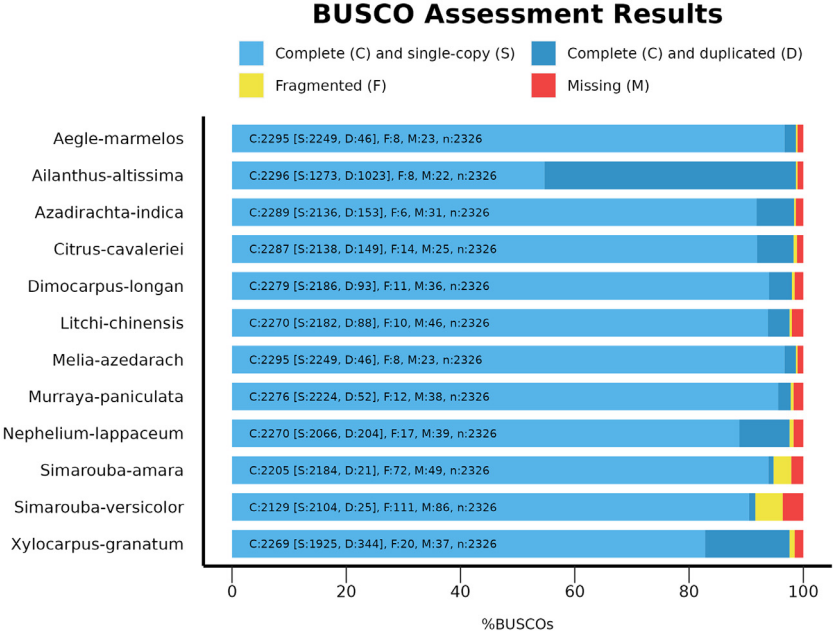

Fig. 1. Comparative data from BUSCO (Benchmarking Universal Single-Copy Orthologs) for quantitative assessment of genome assembly and gene set in the species belonging to order Sapindales, including *S. amara* and *S. versicolor*. Database used: eudicots (eudicots_odb10).

From these *de novo* genome assemblies and using the QDD version 3 software [2]. We initially identified 13,226 (*S. amara*) and 13,011 (*S. versicolor*) regions with microsatellites and utilized these fragments to develop SSR-Seq molecular marker primers for both species. As a first result, we developed 11,348 and 12,084 primers for regions with microsatellite sequences for *S. amara* and *S. versicolor*, respectively. After that, the *in silico* tests were applied using openPrimeR, an R-based tool [3]. The results after these filters were a final dataset of 87 and 77 SSR-Seq primers pairs for *S. amara* and *S. versicolor*, respectively (Table S1 and Table S2). (See the filters and physical and chemical characteristics in section Experimental design, Materials, and Methods, subsection Development of SSR-Seq primers, multiplexing and cross-species in silico amplification). The set of SSR-seq primers generated by this study has the potential to be used in future analyses of genetic diversity and understanding of population dynamics, helping to guide strate-

gies for conservation and management of native species, as well as for the characterization of germplasm and breeding programs.

The SSR-Seq primers designed for *S. amara* (Table S1) and *S. versicolor* (Table S2) and the performance has been evaluated in a multiplex setting. According to the physicochemical properties previously determined and using the openPrimeR tool, we were able to multiplex 55 SSR-seq primers pairs for *S. amara*, organized in five multiplex sets (Table 2). In *S. versicolor*, we were able to multiplex 56 pairs of primers, organized in four different multiplexes (Table 3). Without compromising the quality and biological interpretation of the data, multiplex PCR has the potential to save time and effort in the laboratory [4,5].

The set of 87 SSR-Seq primers pairs developed for *S. amara* was tested *in silico* for cross-species amplification in the *S. versicolor* draft genome. This step was performed with the aim of providing alternatives in the use of the primers sets presented in this research. Having the same set of primers that can be used in the study of both species, *S. amara* and *S. versicolor,* will reduce the cost of obtaining these molecular markers without compromising the quality of the results. A total of 27 SSR-Seq primers pairs designed for *S. amara* showed cross-species amplification in *S. versicolor* (Table 4). Due to the scarcity of marker genetics in native species, the cross-species amplification of microsatellite markers has proven to be an excellent tool and alternative in genetic studies in these plants [6].

4. Experimental Design, Materials, and Methods

4.1. DNA extraction, sequencing, and assembly

Fresh leaves from four individuals of *S. amara* were collected in the Serra of Pirineus, city of Pirenópolis, and from one individual from *S. versicolor* in the city of Goiânia, both in the state of Goiás, Brazil (geographical coordinates: −48.84820, −15.80294 and −50.15408289, −15.92941545, respectively). Total DNA was isolated according to the CTAB protocol 2% [7]. DNA quality and concentration was determined with agarose 1% and fluorimetry with Qubit. Library preparation was performed using the Illumina DNA Prep kit with final library concentrations of 15 pM for *S. amara* and 12 pM for *S. versicolor* and sequencing was performed on the Illumina MiSeq platform in paired end using the V3 600 cycles and V2 300 cycles kits for sequencing *S. amara* and *S. versicolor*, respectively.

The raw reads were subjected to a quality control processing step using the Trimommatic software [8] with specific parameters tailored for each species. For *S. amara* the parameters used were SLIDINGWINDOW: 4:20, CROP:289, HEADCROP:15, ILLUMINACLIP: NexteraPE-PE.fa:2:15:10, LEADING:10, TRAILING:10, MINLEN:100. For *S. versicolor* the parameters applied were SLIDINGWINDOW: 4:15, CROP:150, HEADCROP:15, ILLUMINACLIP: NexteraPE-PE.fa:2:15:10, LEADING:10, TRAILING:10, MINLEN:100. Subsequently, the nuclear genomes were assembled utilizing the SPAdes genome assembler v3.13.1 tool [9]. The parameter MINLEN:100 was selected to discard short reads in excess that may compromise assembly accuracy, even in *S. versicolor* which was sequenced with shorter reads. The SLIDINGWINDOW thresholds for *S. versicolor* required a less stringent threshold (4:15) to retain sufficient data. These parameters prioritize base quality over sequencing depth and represent a compromise: although they reduce sequencing depth and may contribute to a more fragmented assembly, they were necessary to ensure overall data quality. The metrics of the initial assemblies, such as N50 and L50 values, scaffold size, assembled genome size, were accessed using the assemblathon_stats.pl script (available in https://raw.githubusercontent.com/lexnederbragt/sequencetools/master/assemblathon_stats.pl).

After analyzing the initial assemblies, contigs with sizes smaller than 1000 bp were sequences considered as unassembled and excluded from the final draft genome assembly. The search for complete genes recovered in the assemblies (genome completeness) of *S. amara, S. versicolor*, and other species from the same botanical group used for comparison was performed with BUSCO v. 5.6.1 software [10] (Fig. 1). For this analysis we used the database from eudicots

Table 2
Set of SSR-Seq primers multiplexed designed for *S. amara*. The difference in melting temperature (ΔTm) of the set primers within each multiplex is: A – 1.76 °C; B - 1.17 °C; C – 1.1 °C; D - 1.83 °C; E – 1.45 °C.

Multiplex	Loco	Contig	Forward	Reverse	Tm	Motif	N_rep	PCR product size
A	Sam_ssr1	190	cactggcatttctgttgaaa	ccatttgcagaacaggaaat	49.13	AG	15	187
A	Sam_ssr2	2156	agaacaaataacaagcagcg	tgaagagacctcgtatctga	49.34	AG	18	174
A	Sam_ssr3	2205	ggtctcaggatatggcaatt	gtaccttcagtaatgcggat	50.24	AG	19	170
A	Sam_ssr4	22,592	tctttctcttcttcggttgt	gagatgggtctgattcgaaa	49.02	AC	15	174
A	Sam_ssr5	2281	ctctcctctgcttcatcatc	aagtttagatcatgctggca	49.8	AG	16	176
A	Sam_ssr6	26	ttaccagcaaggatttagcc	ccctctctgtctgaattctg	50.78	AG	17	188
A	Sam_ssr7	2856	cctcggttatgtagtgtgag	aatacaatggtgatgtggct	49.83	AG	16	173
A	Sam_ssr8	3311	gagtcaatggaggtggattt	ctcatctttgactcctcctg	50.47	AG	17	181
A	Sam_ssr9	4912	ttggctttggcttatcagat	ccctttatggtcggctattt	49.83	AG	16	187
A	Sam_ssr10	565	aggttatagggacgaagaca	cagaagatgggtccaagaaa	50.4	AG	17	187
B	Sam_ssr11	1958	tctgtgaagttgtcaaagga	atgtgtgtctctctatgtgc	49.25	AG	15	172
B	Sam_ssr12	2175	accaccacttaatcaccaaa	acccaacaaggactattgac	49.64	AG	15	171
B	Sam_ssr13	2991	acatcgctttcttcatcact	caagaaacgagcgagagata	49.72	AG	19	171
B	Sam_ssr14	3858	tcgattgtttgtctgtacgt	agtcaaatcagccggtaaat	49.58	AG	16	178
B	Sam_ssr15	4317	cttcaatcacaactgttgca	ctccgagtttggtaagacat	49.25	AG	17	172
B	Sam_ssr16	4692	agaagttcatagggctctga	tggcaagagatatgttcctg	50.23	AC	15	190
B	Sam_ssr17	4823	atacactgtacaagggcaag	tccaagtcggtgatcttaga	50.42	AC	15	159
B	Sam_ssr18	4875	taggtttagtgtgtgtggg	acacactctctctctctctc	50.26	AACAC	6	163
B	Sam_ssr19	4915	ttgacctttagtggcttgtt	atttgcatgagacgagctta	49.78	AG	17	189
B	Sam_ssr20	666	agaaggtttagctgacagtg	ataagagtgcatcttcctgc	50.33	AG	15	170
B	Sam_ssr21	6769	acaagagctcattggagaag	agccagagaatttgaaggg	50.41	AG	15	154
B	Sam_ssr22	699	ccctttaccactcttcaaca	aatgcaacaaagctgagaac	50.41	AG	16	181
C	Sam_ssr23	10	ggcttccttgagtactttga	cttgaacttgacagactggt	50.35	AG	19	167
C	Sam_ssr24	1051	aaagatcgaattgcctgtct	ttgacaagaaacctcacaca	49.76	AG	17	170
C	Sam_ssr25	10,803	gtcacgatccctttaaactga	gatagtggtggtggatttgt	50.1	AC	20	189
C	Sam_ssr26	1107	cagttcctgttccattctct	atcatcttcaggcgatatgg	50.11	AG	20	165
C	Sam_ssr27	124	agtggcaaactatagtgcat	cctcataaggtactcagctg	49.53	AG	16	182
C	Sam_ssr28	127	agtggtgatactccctttct	gcacattacaactgctgaaa	49.3	AG	19	167
C	Sam_ssr29	1382	aacattgcaagagagccata	ttctccgggaaatgtaactc	49.8	AG	27	181
C	Sam_ssr30	1609	tattgatggtggcgtacaaa	tccattcatctgtgaagcaa	49.72	AG	16	163
C	Sam_ssr31	1630	gatgatagcagcaagaacct	agaagatgtacctcaactgc	50.13	AG	19	187
C	Sam_ssr32	168	cgcacaaacttcccaaataa	agagagagcaaacaaacgat	49.32	AG	15	171
C	Sam_ssr33	178	ggctaagagcataagaccaa	tatttagccgaattgcctct	49.32	AG	16	185
C	Sam_ssr34	7891	aatacctggagtcagcattg	aatgctggagaagactgaag	50.41	AG	18	187
C	Sam_ssr35	832	caagccttaaatagcgcaaa	tgattctcacaacttggtcc	50.41	AG	19	186
D	Sam_ssr36	1059	gtccttgtcttctcgcttta	ttacaggaagcgtaaacacg	50.33	AG	20	189
D	Sam_ssr37	1100	atttgaggcaagtttctcca	tgctgtgctacactgttaat	49.67	AG	20	187
D	Sam_ssr38	16,895	cgagtggattaatgaaacagc	acagcaacattatcggagtt	49.75	AG	15	176
D	Sam_ssr39	1816	tgagctactgaagatcctga	atttcgtcatgtgttcaggt	49.94	AG	15	158
D	Sam_ssr40	192	taatttgcaggtccaccatt	agaaatgagagggagacaga	50.06	AG	21	170
D	Sam_ssr41	2	ttgcccatcaagctatagc	ggaatatcacaagccctacc	50.73	AG	17	156
D	Sam_ssr42	221	tgttgtcaactgtcagagtt	cagagcttcagaaggtactc	49.48	AC	20	174
D	Sam_ssr43	2894	ggtttgaaagctttgctgaa	tgagtttccactttctctctc	49.79	AG	15	163
D	Sam_ssr44	2952	gttcatagccaatccaaagc	ctcggcttcaatctgtgaa	50.18	AG	22	189
D	Sam_ssr45	3358	ccagtaaagactcggcaata	tggtagagatttagagaaatggg	49.77	AG	17	171
D	Sam_ssr46	3938	tcatctgtccctcttaccttt	tagggaatgtgtcttgtgtg	50.35	AC	15	181
D	Sam_ssr47	394	cgcctaaagatccttcttga	taatcgtcgactcattcacc	50.16	AG	20	177
E	Sam_ssr48	3483	gcagagaagtggagatgaat	aaagcctgcaacagatatga	49.8	AG	16	184
E	Sam_ssr49	5159	actgctttgttcttgagtct	ccaacactcttgaggtaaca	49.57	AC	15	186
E	Sam_ssr50	6003	tgttgagagctacacattgt	accttgactttctctcctct	49.35	AG	15	186
E	Sam_ssr51	636	agcttgatacacttcaccag	actctgtttctaccacaaca	48.98	AG	19	185
E	Sam_ssr52	6598	ctgccaacgatccctaaata	atccaatctctctccaatgc	50.28	AG	19	181
E	Sam_ssr53	6758	tgagctcagcatttacaact	aatccctctttcagcttcag	49.69	AG	15	154
E	Sam_ssr54	676	agaagccacaattcctacag	tcccttgacctccttcttta	50.43	AG	17	175
E	Sam_ssr55	7031	tgcggaaatgacctaaatca	acaggttctctctctctctc	49.77	ACCCC	6	174

Table 3
Set of multiplexed SSR-Seq primers designed for *S. versicolor*. The difference in melting temperature (ΔTm) of the set primers within each multiplex is: A - 1.33 °C; B - 1.36 °C; C - 0.8 °C; D - 1.57 °C.

Multiplex	Identifier	Contig	Forward	Reverse	Tm	Motif	N_rep	PCR product size
A	Sve-ssr1	116	gcaccctctgctaaagatta	ttattctcagtgctcagacg	50.12	AC	18	168
A	Sve-ssr2	124	ccctgtccatattcttagtgt	ctacctatgcaacccacac	49.62	AG	15	172
A	Sve-ssr3	1334	aaaccctggaacaccaaatt	gccatatcgactgactgtaa	49.94	AG	17	189
A	Sve-ssr4	1337	gctgtaggtgaattcaaagc	cttcacacccacaactcata	50.08	AG	15	188
A	Sve-ssr5	1397	tccaatcttatgtcaccacg	atttgatcagacgaggttgg	50.43	AAGCC	6	152
A	Sve-ssr6	1418	tgtctgtaacaagctgtgaa	tcctttgctttctctgtgtt	49.53	AG	20	175
A	Sve-ssr7	1612	caccgaatgatcactctcat	cccatttgtctccctgtaat	50.23	AG	15	182
A	Sve-ssr8	185	tttagagttgctgtgttgga	acactgcatgaaattgaagc	49.57	AG	20	182
A	Sve-ssr9	1865	ccacttctccacgaattaca	acagcgtttcttcacactta	49.82	AG	21	189
A	Sve-ssr10	846	aatcatccacctgttcgaaa	ttcctcctctacaaggttga	49.71	AG	21	181
A	Sve-ssr11	86	catcccacgtccactaaac	cacagaagagacagagatgg	50.82	AG	19	181
A	Sve-ssr12	973	tttacccaaagtaatcggca	tgttcagacgattggcttt	49.72	AC	17	180
A	Sve-ssr13	6003	agtcaattcccggagtaatg	cccaactttgatctaggctt	50.22	AG	18	181
A	Sve-ssr14	631	aatgcttgacttgttgggta	tgtaagaagttgtcacacgt	49.49	AG	15	187
A	Sve-ssr15	7352	acaatcagctagtaagtggg	cgtatttggatggttggact	49.9	AG	18	182
A	Sve-ssr16	780	gtaaacagagaagacactgga	tgataaacggaaagctagcc	49.92	AG	17	171
B	Sve-ssr17	2059	agcttaacagtgaaatacca	tcacttctgagtctccatct	49.38	AG	18	180
B	Sve-ssr18	2078	gattgactgcaattctgtgg	aagatcatttcatcggacgg	50.09	AC	15	186
B	Sve-ssr19	2121	ttcaccttaaccggtctaac	cttcacaggttggtctcttt	50.06	AC	15	157
B	Sve-ssr20	2583	tgatcccacattaaacagca	actacttacctcgagctctt	49.75	AG	18	187
B	Sve-ssr21	2702	ttaacatgcttagcctaggg	cctttccattgaacttgctc	49.99	AG	15	158
B	Sve-ssr22	280	ccctcaaatcagatggtacc	ccgatataagctaccgacag	50.74	AC	20	164
B	Sve-ssr23	2809	aacgtatagggcgagttaac	agaagtcacacaaagaacga	49.5	AG	17	179
B	Sve-ssr24	2942	taacacacatcatctgagcc	atcatatgggtgcaactctc	50.31	AG	19	181
B	Sve-ssr25	3040	cgcctatccttgaccataaa	caagtcaaggttggagagag	50.28	AG	15	190
B	Sve-ssr26	3188	ctgtcatcctcatacatcca	tccgggaatgaactaatgac	49.44	AG	24	163
B	Sve-ssr27	1118	ggaaccaagcaaagattacg	accttctcctctcttccaat	50.07	AG	23	164
B	Sve-ssr28	9	gagtggtagagatctgaggt	agtgttaatgcagaggtgag	50.4	AG	16	177
B	Sve-ssr29	4448	gggagaaaggacttgaacta	tggtccaagcacaagattaa	49.52	AG	17	188
B	Sve-ssr30	787	gatgcctagtgtgatgaact	cctatccttagacatgcaca	49.7	AG	18	161
C	Sve-ssr31	3193	ttcttgtgctccctagaaac	agagctcatcctctctcttt	50.35	AG	17	159
C	Sve-ssr32	3862	gcccaataactgaaacccta	gaaacagtgaaacagagcag	49.92	AG	15	169
C	Sve-ssr33	3872	agcaaactgttgaatgatgc	gccctagagttgattcttct	49.64	AG	16	157
C	Sve-ssr34	4049	ggcctctgtgaagtaaagaa	agatatgctttgactgaccg	50.35	AG	21	187
C	Sve-ssr35	413	gagaacgaccgatcatctag	ccacatcaccttctctttca	50.36	AG	20	187
C	Sve-ssr36	4297	agggaagcaaagatgtcttt	tgccaatcagaagttagcat	49.71	AG	15	157
C	Sve-ssr37	4305	agtgttatccgcagttcaat	ttcgtcgtccaacttaatgt	49.61	AG	16	152
C	Sve-ssr38	4375	ctacatgcagagacgagaaa	ctcactcgctgccttaatat	50.12	AG	15	152
C	Sve-ssr39	4378	atctccactgttgatcgttc	cccatttcttctttctcctct	50.08	AG	20	166
C	Sve-ssr40	831	tcacactctctctcactcaa	taaatgacttgtgggagacg	50.37	AG	18	158
C	Sve-ssr41	6105	atcgggtactacaagcaatc	attctctctttaccgatgcc	50.24	AG	18	181
C	Sve-ssr42	6514	attcaactctacctgcaacc	ctgctttaaattgcctccc	50.41	AG	21	184
C	Sve-ssr43	695	aagctcatgatcgtactgtc	ataacagcagatggtacgag	49.96	AG	19	160
D	Sve-ssr44	4424	cgttcatgctatcgacctaa	gtctaggttgtgtccacatt	50.23	AC	15	179
D	Sve-ssr45	4593	cagagagtgtgtaagccaaa	ctcctcagcatcttaaagcc	50.58	AG	19	176
D	Sve-ssr46	4651	tgaattgcctttcaccgata	gaccagggtagttacagaga	49.77	AG	18	172
D	Sve-ssr47	4968	cttggaagtctacgtctacg	aagtatctgttgtgaagcgt	49.64	AG	17	174
D	Sve-ssr48	5026	tgactactctgtgcagaaac	cagagagctttcttcaggtt	50.28	AG	18	150
D	Sve-ssr49	537	agttgtttgttcttgaagcg	aatctatacgggccatcaac	49.46	AG	25	179
D	Sve-ssr50	5415	gctaccacccaagtatgatg	catagtcggacaagttgtga	50.33	AG	18	181
D	Sve-ssr51	5539	aagggagggaaataatcacg	accgagcaaactctacattt	49.68	AGCCC	6	168
D	Sve-ssr52	5848	aatgccatcaattgaaagcc	cccaactccaactattccaa	49.79	AG	17	159
D	Sve-ssr53	5854	tctaccggacaagactgtat	ttggagaagcaacaagagtt	49.85	AG	18	162
D	Sve-ssr54	3189	ggagaagaaatgatggacca	gtggataagtggaatttgttgt	49.01	AG	19	180
D	Sve-ssr55	4411	ccgtggtgccatttataact	ctattgggtggtgtttctgt	50.65	AC	17	187
D	Sve-ssr56	5871	aaagaagctcaagtggatgt	gccttctaagctctgatacc	49.67	AC	15	165

Table 4
Set of SSR-Seq primers designed for *S. amara* (Table S1), which showed cross-species amplification in the genome of *S. versicolor* (*in silico* primer transferability).

Num	Identifier	Contig	Forward	Reverse	Tm	Motif	N_rep	PCR product size
1	Sam_ngs_P1	10	ggcttccttgagtactttga	cttgaacttgacagactggt	50.35	AG	19	167
2	Sam_ngs_P15	1630	gatgatagcagcaagaacct	agaagatgtacctcaactgc	50.13	AG	19	187
3	Sam_ngs_P21	190	cgcacaaacttcccaaataa	agagagagcaaacaaacgat	49.32	AG	15	171
4	Sam_ngs_P23	2	ttgcccatcaagctatagc	ggaatatcacaagccctacc	50.73	AG	17	156
5	Sam_ngs_P25	2156	agaacaaataacaagcagcg	tgaagagacctcgtatctga	49.34	AG	18	174
6	Sam_ngs_P26	2175	accaccacttaatcaccaaa	acccaacaaggactattgac	49.64	AG	15	171
7	Sam_ngs_P30	2281	ctctcctctgcttcatcatc	aagtttagatcatgctggca	49.8	AG	16	176
8	Sam_ngs_P40	3115	gctcgctatactgaatcgaa	tggttaacccataatcgcaa	49.79	AG	15	187
9	Sam_ngs_P42	3237	tctctctcacatctgctaca	tgagagagagagagtgaaaga	50.02	AG	15	176
10	Sam_ngs_P48	370	tgcatggcctactttattgt	agcagatgattctctttcca	48.95	AG	16	160
11	Sam_ngs_P53	4025	tactgcaacttcggatttga	aaatgaaggtgaaccctacc	49.66	AG	15	163
12	Sam_ngs_P55	4692	agaagttcatagggctctga	tggcaagagatatgttcctg	50.23	AC	15	190
13	Sam_ngs_P56	4823	atacactgtacaagggcaag	tccaagtcggtgatcttaga	50.42	AC	15	159
14	Sam_ngs_P57	4912	ttggctttggcttatcagat	ccctttatggtcggctattt	49.83	AG	16	187
15	Sam_ngs_P58	4915	ttgacctttagtggcttgtt	atttgcatgagacgagctta	49.78	AG	17	189
16	Sam_ngs_P59	5159	actgctttgttcttgagtct	ccaacactcttgaggtaaca	49.57	AC	15	186
17	Sam_ngs_P62	6003	tgttgagagctacacattgt	accttgactttctctcctct	49.35	AG	15	186
18	Sam_ngs_P68	676	agaagccacaattcctacag	tcccttgacctccttcttta	50.43	AG	17	175
19	Sam_ngs_P70	6883	gaccctttactgcaagctat	atttatttggagtcggctgt	49.78	AG	18	150
20	Sam_ngs_P71	699	ccctttaccactcttcaaca	aatgcaacaaagctgagaac	49.58	AG	16	181
21	Sam_ngs_P74	7672	tttctgtcaatggagtagcc	gcaaattagcaagtgggatc	50.18	AG	19	177
22	Sam_ngs_P75	7829	tgcaaccaccaacaatttac	tatgggtaggagggagaaag	49.55	AG	17	176
23	Sam_ngs_P76	7891	aatacctggagtcagcattg	aatgctggagaagactgaag	50.41	AG	18	187
24	Sam_ngs_P79	810	gttgttctcccttgatacca	taatagggccttacaccaga	50.12	AG	21	159
25	Sam_ngs_P84	832	caagccttaaatagcgcaaa	tgattctcacaacttggtcc	49.43	AG	19	186
26	Sam_ngs_P86	931	gtacaggcgatgtccatatt	cgattgaggtgaaagtgttg	49.98	AG	17	168
27	Sam_ngs_P87	9580	cccgaaccttctagaaacaa	tcatcaagaacgaaagctca	49.64	AG	19	178

(eudicots_odb10). The genomes of the remaining species were retrieved from public databases (Table S3). The depth of coverage of the reads in the genome was done by first aligning the reads in the genome using the Burrows-Wheeler (BWA) aligner v.0.7.17 [11] and then estimating the average depth of coverage using the Samtools v. 1.15 (available in https://www.htslib.org/). The best Kmer value and genome size were estimated using the kmergenie tool v.1.70 [12].

4.2. Development of SSR-Seq primers, multiplexing and cross-species in silico amplification

Identification and extraction of microsatellite repeat regions with minimum repeat motifs of 10 for dinucleotide, 8 for trinucleotide, 6 for tetranucleotide and pentanucleotide, was performed using the QDD tool v.3 [13]. The primers were then developed using the Primer3 tool implemented in the QDD tool v.3 [15] whose parameters were: (i) PCR product size between 120 and 200 bp; (ii) Primer size (minimum - optimal - maximum) of 18 bp - 20 bp - 23 bp; Melting temperature (minimum - optimal - maximum) of 48 °C - 55 °C - 62 °C; (iv) Primer GC content (minimum - optimal - maximum) of 20% - 50% - 80%; (v) Maximum melting temperature difference 1 °C.

Based on the results, additional filtering criteria were applied to exclude microsatellite regions with: (i) trinucleotide SSR motif; (ii) SSR with repeats with 3 or more adenines in a row (AAA ∗); (iii) SSR motif with 100% AT content; (iv) SRR with distance <20 bp from primers; (v) SRR in the context of transposable elements; (vi) compound SSR; (vii) SSR repeats <30 bp in size. We also only consider PCR product sizes between 150 and 190 bp. After applying the filters, a set of 143 and 135 primers were obtained from *S. amara* and *S. versicolor*, respectively.

These primer sets were filtered according to the desired physicochemical properties in an *in silico* PCR using the openPrimeR tool v1.11.4 [3] with the following constraints: (i) Primer length (min-max): 18–23 bp; (ii) GC ratio (min-max): 0.2–0.8; (ii) CG clamp (min-max): 0–3; (iii) Melting temperature (min-max): 48 °C-62 °C; (iv) Run length (min-max): 0–4; (v) Secondary structure: −1 kcal; (vi) Self dimerization: −6 kcal. A set of 87 and 77 SSR-Seq primer pairs for *S. amara* and *S. versicolor*, respectively, survived the applied filters (Table S1 and Table S2). From the obtained SSR-Seq primer sets, PCR multiplexes for both species were organized and tested *in silico* PCR using the openPrimeR tool v1.11.4 [3] considering the minimum energy value for cross-dimerization of −5 (i.e., $\Delta G > -5$) and the difference in melting temperature between each primer pair (ΔTm) of up to 2 °C (Table 2 and Table 3).

In silico cross-species amplification assays were performed using a panel of 87 SSR-Seq primers initially designed for *S. amara* (Table S1), with the aim of assessing their applicability to *S. versicolor.* PCR products corresponding to the 87 pairs of SSR-Seq primers from *S. amara* were aligned to the *S. versicolor* genome using the BWA aligner [13], and the alignments were visualized using the Tablet tool v.1.21.02.08 [14]. Alignments with conserved microsatellite regions were selected after visual inspection resulting in a total of 34 PCR products. The SSR-Seq primers corresponding to the PCR products were then tested for coverage in the *S. versicolor* genome by an *in silico* PCR using the openPrimeR tool v1.11.4 [3]. To reduce the presence of non-specific amplification, we limited the number of mismatches in the primer region to a maximum of 5 base pairs. At the end of the analyses, the *in silico* cross-species amplification of 27 pairs of SSR-Seq primers in *S. versicolor*, previously developed from *S. amara*, was successful (Table 4). The set of SSR-Seq primers developed for *S. versicolor* was tested on the draft genome of *S. amara*; however, satisfactory results were not obtained in the *in silico* amplification of the primers or in the alignment of the corresponding microsatellite regions.

Limitations

The present study has limitations inherent to the validation process, as the generated datasets were not subjected to laboratory validation. Although *in silico* validation is an initial and valid strategy for efficient screening of SSR-Seq primers, experimental validation in the laboratory with multiple individuals is essential to confirm the practical usefulness of the primers developed. This step is essential for evaluating polymorphisms in microsatellite regions, and ensures the robustness, specificity, and reproducibility of the data for future use.

Ethics Statement

The authors have read the ethical requirements for publication in Data Brief and certify that the current paper does not involve human subjects, animal testing, or data collected from social media platforms.

The collection and management of biological samples were carried out following the legal guidelines of SisGen—National System for the Management of Genetic Heritage and associated traditional knowledge and are available under the access code AFBD2DB.

CRediT Author Statement

Marla A. Almeida-Silva: Investigation, Methodology, Data Curation, Formal analysis, Writing-Original Draft, Visualization. **Leonardo C. J. Corvalan:** Data Curation, Formal analysis, Writing-Review & Editing. **Ramilla S. Braga-Ferreira** Investigation, Methodology, Writing-Review & Editing. **Cíntia P. Targueta:** Investigation, Methodology, Writing-Review & Editing. **Edivani V. Franceschinelli, Carlos M. Silva-Neto, Thannya S. Nascimento Rhewter Nunes:** Conceptualization, Investigation, Methodology, Supervision, Writing-Original Draft, Project administration.

Mariana P. C. Telles: Conceptualization, Writing-Original Draft, Supervision, Resources, Funding acquisition, Project administration.

Data Availability

Simarouba amara isolate SamPIRGO, whole genome shotgun sequencing project (Original data) (NCBI)
Simarouba versicolor isolate Sve-GOIGO, whole genome shotgun sequencing project (Original data) (NCBI)

Acknowledgements

This work was supported by the National Institute of Science and Technology (INCT) in Ecology, Evolution, and Biodiversity Conservation funded by CNPq (409197/2024–6). This paper is also a contribution of the "PPBio Araguaia" project supported by CNPq (proc. 441114/2023–7) and "Araguaia Vivo 2030" program developed under the agreement between the Tropical Alliance Water Research (TWRA) and FAPEG (proc. 202210267000536). This work was carried out within the context of the Neotropical BioGenomes Network. M.A.A.S has financial support from the State University of Piauí-PI-Brazil, RN has a PDCTR scholarship from CNPq/FAPEG (process number: 2021102670 0 0863). M.P.C. Telles and C.M. Silva-Neto were supported by productivity fellowships from CNPq.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary Materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.dib.2026.112855.

References

[1] F. Pustahija, et al., Small genomes dominate in plants growing on serpentine soils in West Balkans, an exhaustive study of 8 habitats covering 308 taxa, Plant Soil. 373 (1–2) (2013) 427–453, doi:10.1007/s11104-013-1794-x.
[2] E. Meglécz et al., 'QDD: a user-friendly program to select microsatellite markers and design primers from large sequencing projects', 2010, Oxford University Press. doi: 10.1093/bioinformatics/btp670.
[3] C. Kreer, et al., openPrimeR for multiplex amplification of highly diverse templates, J. Immunol. Methods 480 (2020) 112752, doi:10.1016/j.jim.2020.112752.
[4] J.S. Chamberlain, R.A. Gibbs, J.E. Ranier, P. Nga Nguyen, C.T. Caskey, Deletion screening of the Duchenne muscular dystrophy locus via multiplex DNA amplification, Nucleic. Acids. Res. 16 (1988) 11141–11156 [Online]. Available http://nar.oxfordjournals.org/ .
[5] E.M. Elnifro, A.M. Ashshi, R.J. Cooper, P.E. Klapper, Multiplex PCR: optimization and application in Diagnostic virology, Clin. Microbiol. Rev. 13 (4) (2000) 559–570.
[6] T. Barbará, C. Palma-Silva, G.M. Paggi, F. Bered, M.F. Fay, and C. Lexer, 'Cross-species transfer of nuclear microsatellite markers: potential and limitations', 2007. doi: 10.1111/j.1365-294X.2007.03439.x.
[7] J.J. Doyle, J.L. Doyle, A rapid DNA isolation procedure for small quantities of fresh leaf tissue, 1987. [Online]. Available https://webpages.uncc.edu/~jweller2/pages/BINF8350f2011/BINF8350_Readings/Doyle_plantDNAextractCTAB_1987.pdf.
[8] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics 30 (15) (2014) 2114–2120, doi:10.1093/bioinformatics/btu170.

[9] A. Bankevich, et al., SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing, J. Comput. Biol. 19 (5) (2012) 455–477, doi:10.1089/cmb.2012.0021.
[10] F.A. Simão, R.M. Waterhouse, P. Ioannidis, E.V. Kriventseva, E.M. Zdobnov, 'BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs', Bioinformatics 31 (19) (2015) 3210–3212, doi:10.1093/bioinformatics/btv351.
[11] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform, Bioinformatics 25 (14) (2009) 1754–1760, doi:10.1093/bioinformatics/btp324.
[12] R. Chikhi, P. Medvedev, Informed and automated k-mer size selection for genome assembly, Bioinformatics 30 (1) (2014) 31–37, doi:10.1093/bioinformatics/btt310.
[13] E. Meglécz et al., 'QDD: a user-friendly program to select microsatellite markers and design primers from large sequencing projects', 2010, Oxford University Press. doi: 10.1093/bioinformatics/btp670.
[14] I. Milne, et al., Using tablet for visual exploration of second-generation sequencing data, Brief. Bioinform. 14 (2) (2013) 193–202, doi:10.1093/bib/bbs012.