Applications in Plant Sciences 2016 4(10): 1600048

Genomic Resources Note

Chloroplast sequence of treegourd (*Crescentia cujete*, Bignoniaceae) to study phylogeography and domestication[1]

Priscila Ambrósio Moreira[2,6], Cédric Mariac[3], Nora Scarcelli[3], Marie Couderc[3], Doriane Picanço Rodrigues[2,4], Charles R. Clement[2,5], and Yves Vigouroux[3,6]

[2]Post-Graduate Program in Botany, Instituto Nacional de Pesquisas da Amazônia (INPA), Av. André Araújo 2936, Petrópolis, 69067-375 Manaus, Amazonas, Brazil; [3]UMR DIADE, Institut de Recherche pour le Développement (IRD), 393 Avenue Agropolis, Montpellier, Cedex 5, France; [4]Laboratório de Evolução Aplicada, Universidade Federal do Amazonas (UFAM), 69077-000 Manaus, Amazonas, Brazil; and [5]Coordenação de Tecnologia e Inovação, INPA, Manaus, Amazonas, Brazil

- *Premise of the study: Crescentia cujete* (Bignoniaceae) fruit rinds are traditionally used for storage vessels and handicrafts. We assembled its chloroplast genome and identified single-nucleotide polymorphisms (SNPs).
- *Methods and Results:* Using a genome skimming approach, the whole chloroplast of *C. cujete* was assembled using 3,106,928 sequence reads of 150 bp. The chloroplast is 154,662 bp in length, structurally divided into a large single copy region (84,788 bp), a small single copy region (18,299 bp), and two inverted repeat regions (51,575 bp) with 88 genes annotated. By resequencing the whole chloroplast, we identified 66 SNPs in *C. cujete* ($N = 30$) and 68 SNPs in *C. amazonica* ($N = 6$). Nucleotide diversity was estimated at 1.1×10^{-3} and 3.5×10^{-3} for *C. cujete* and *C. amazonica*, respectively.
- *Conclusions:* This broadened *C. cujete* genetic toolkit will be important to study the origin, domestication, diversity, and phylogeography of treegourds in the Neotropics.

Key words: Bignoniaceae; calabash tree; *Crescentia amazonica*; cuia; next-generation sequencing; single-nucleotide polymorphism (SNP).

Crescentia cujete L. (Bignoniaceae) is a diploid species ($2n = 40$) that produces non-edible fruits that have been of great importance to many indigenous and traditional communities of tropical America since pre-Columbian times, especially as drinking cups and storage vessels. Its wild geographic distribution is unknown, but it is found in many areas in the Neotropics in close contact with wild relatives in quite different environments.

There are two hypotheses of its origin of domestication. Gentry (1980) hypothesized an origin in Mesoamerica, where wild populations are found in seasonally flooded savannas. This hypothesis was not confirmed with chloroplast microsatellites in the eastern Yucatán of Mexico (Aguirre-Dugua et al., 2012). Ducke (1946) hypothesized that *C. amazonica* Ducke (described in 1937) gave rise to the cultivated *C. cujete*. This Amazonian species is also found in the Orinoco Basin and elsewhere in northern South America (Gentry, 1980; Wittmann et al., 2006; Díaz, 2009), where it is common in floodplain forests. The distributions of the other four accepted species of *Crescentia* L. are restricted to Central America and the Antilles, leading Gentry (1980) to comment on *C. amazonica*'s distribution outside of the distribution of the other species. Contrary to Ducke (1946), Gentry (1980) suggested that *C. amazonica* was derived from cultivated *C. cujete* "when human selection for large fruits is relaxed." However, using amplified fragment length polymorphism markers and a single accession of *C. amazonica* from the Orinoco Basin, Arango-Ulloa et al. (2009) found no relationship with *C. cujete* from Colombia.

Identification of the origin of domestication of treegourd and its routes of dispersal in the Neotropics remains unclear, and requires a molecular genetic analysis of a broader geographic sample. Using *C. amazonica* and *C. cujete* collections widely distributed along major rivers of the Brazilian Amazon Basin and the assembly of the chloroplast genome, we aim to identify single-nucleotide polymorphisms (SNPs) to compare chloroplast diversity between *C. cujete* and *C. amazonica* in order to evaluate the two hypotheses about the relationships between these species and better understand the domestication history of treegourd.

METHODS AND RESULTS

DNA was extracted from dried leaves of 36 samples of *C. cujete* and *C. amazonica* from the Brazilian Amazon Basin (Appendix 1). We used the

[1]Manuscript received 15 April 2016; revision accepted 30 August 2016.

This research was supported by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq-473422/2012-3), the Fundação de Apoio à Pesquisa do Estado do Amazonas (FAPEAM 062.03.137/2012), the Agence Nationale de la Recherche (ANR-13-BVS7-0017), and the ARCAD project funded by Agropolis Fondation. P.A.M. thanks the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior for a scholarship (CAPES-99999.010075/2014-03). We thank the Instituto de Desenvolvimento Agrário do Amazonas for field support, and family farmers and the Instituto Brasileiro do Meio Ambiente (IBAMA-14BR015576/DF) for their consent for this research.

[6]Authors for correspondence: pri.ambrosio@hotmail.com; yves.vigouroux@ird.fr

doi:10.3732/apps.1600048

cetyltrimethylammonium bromide (CTAB) 5% extraction protocol (Doyle and Doyle, 1990) with minor modifications; instead of cold isopropanol, NaCl was added to precipitate pellets. Barcoded libraries were constructed following the protocol of Mariac et al. (2014). Briefly, 0.5–1 µg of DNA were fragmented in a Bioruptor Pico sonicator (Diagenode, Liège, Belgium) using a standard protocol including six cycles and on/off conditions set to 30/90 s to reach a target size distribution of 300 bp. After sizing, end repair, ligation, and *Bacillus stearothermophilus* (Bst) DNA polymerase treatment, libraries were amplified with the KAPA HiFi HotStart Real-Time PCR Kit (KAPA Biosystems, Wilmington, Massachusetts, USA) with eight cycles to extend Illumina adapters and quantified by using the KAPA SYBR FAST LightCycler 480 qPCR Kit (KAPA Biosystems). Paired-end sequencing (2 × 150) was conducted on an Illumina MiSeq version 3 and HiSeq 2500 (Illumina, San Diego, California, USA) at the CIRAD facilities (Montpellier, France) and at Genotoul (Toulouse, France), respectively. Twelve picomoles of the bulked libraries with 1% PhiX were loaded in the flow cell. Mean passing filter among the different runs was 84.3%, producing 13 million clusters. The percentage of bases having a quality score above Q30 was 93.7%.

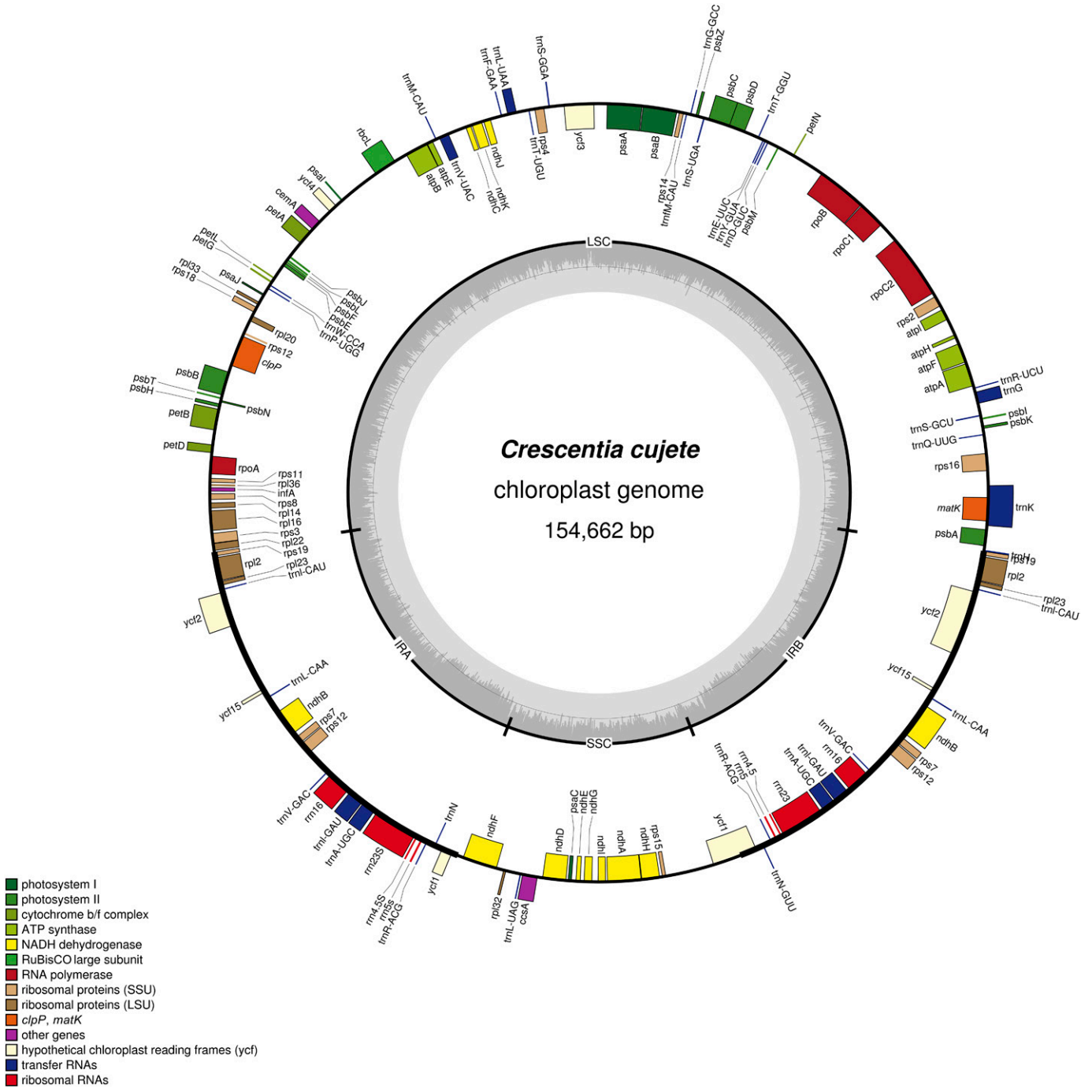

Fig. 1. Circular map of the chloroplast genome of *Crescentia cujete* from Amazonas, Brazil (5.34°S, 60.44°W), deposited in GenBank (accession no. KT182634). Genes drawn within the circle are transcribed clockwise, while genes drawn outside are transcribed counterclockwise. Genes belonging to different functional groups are color coded. Dark bold lines indicate inverted repeats (IRA and IRB) that separate the genome into large (LSC) and small (SSC) single copy regions. Drawn using OrganellarGenomeDraw (Lohse et al., 2013).

Assembly was performed using the chloroplast of *Tanaecium tetragonolobum* (Jacq.) L. G. Lohmann (NC_027955) as a guide sequence for MITObim 1.7 (Hahn et al., 2013). First, MITObim mapped reads to the reference genome using MIRA version 4 (Chevreux et al., 1999) and an initial set of contigs was built (Appendix S1). Then, a second mapping was done on these contigs. Contigs were extended if there was at least a 31-bp overlap with a given read. This process was iterated until a complete de novo genome was achieved. For the assembly, we used the two-step strategy pioneered by Li et al. (2013), because the repetitive nature of the inverted repeat (IR) regions (i.e., IRA and IRB) was difficult to assemble (Li et al., 2013). We first performed an assembly using the sequence large single copy (LSC), IRA, and small single copy (SSC), followed by a second independent assembly using the sequence SSC, IRB, and LSC from *T. tetragonolobum* (NC_027955). From the initial 3,106,862 shotgun reads, 268,499 reads were useful for the de novo chloroplast assembly. The SSC region showed a pairwise identity of 99.6% between the two assemblies, and the LSC region showed 99.7%. The slight differences observed are mainly locally close to repeat regions (mini- and microsatellite), and thus difficult to assemble. The IRs showed a 99.1% pairwise identity. The two fractions were manually aligned using the software Genious Pro 4.8.5 (Drummond et al., 2009), and a consensus *C. cujete* chloroplast sequence was built. The final assembly has a low number of N positions (46 Ns), and 96.7% of reads were properly paired, meaning that both read R1 and R2 were properly mapped. The mean depth of coverage of the sequence was 165×, meaning that for each position we have an average of 165 aligned reads. The final chloroplast genome size was 154,662 bp (Fig. 1, Appendix 1).

The *C. cujete* chloroplast genome was aligned with reference annotated genomes using the mauve algorithm implemented in Geneious Pro 4.8.5 (Drummond et al., 2009). For annotation, we used *T. tetragonolobum* (NC_027955; Bignoniaceae) as reference, and complemented it with *Olea europaea* L. (NC_013707; Oleaceae) and *Capsicum chinense* Jacq. (NC_KU041709; Solanaceae) to validate some tRNA orientations and add some introns lacking in the *Tanaecium* genome (*rpl16*, *rps12*). The correspondences of gene positions were identified and annotated manually. The *C. cujete* chloroplast sequence was deposited in GenBank (accession number KT182634). The assembled chloroplast genome of *C. cujete* was used to map another 30 *C. cujete* samples from home gardens and six *C. amazonica* samples from flooded forests with BWA 0.6.2 (Li and Durbin, 2009). Using SAMtools 0.1.7 (Li et al., 2009) and VarScan 2.3.7 (Koboldt et al., 2012), we generated and filtered the variant call format (VCF) files, following the Scarcelli et al. (2016) pipeline. The average number of chloroplast mapped reads was 54,747, equivalent to a 54× depth of coverage (Appendix 1). The minimum coverage was 15,324 reads, so even for this sample, each nucleotide was sequenced 15 times. We only have 0.7% of missing data in our 36 samples. Diversity analysis of the 30 *C. cujete* and six *C. amazonica* samples was done using DnaSP 5.10.1 (Librado and Rozas, 2009).

The size of the reconstructed chloroplast genome of *C. cujete* is 154,662 bp, structurally divided into four distinct regions: large single copy region (LSC: 84,788 bp), small single copy region (SSC: 18,299 bp), and a pair of inverted repeat regions (IR: 51,575 bp) (Table 1, Fig. 1). We identified 88 coding genes, of which nine were duplicated within IR regions, four rRNAs duplicated in IRA and IRB, 30 tRNAs, of which six were duplicated within IR regions. The *C. cujete* chloroplast genome size (bp) and GC content are comparable to *T. tetragonolobum* (Table 1), and within the variation observed in the order Lamiales, where genome lengths vary from 153,493 to 155,889 bp and GC content from 37.6% to 38.3% (Nazareno et al., 2015). The *rps19* and *rpl2* gene positions duplicated in the boundaries of IR (Fig. 1) agree with expectations from other angiosperms (Wang et al., 2008).

Table 1. Comparison of chloroplast genomes between two species of Bignoniaceae.

Characteristics	*Crescentia cujete*	*Tanaecium tetragonolobum*
Size (bp)	154,662	153,776
LSC length (bp)	84,788	84,612
SSC length (bp)	18,299	17,586
IRA, IRB combined length (bp)	51,575	51,578
GC content (%)	38.3	38.3
No. of genes	132	121
Protein-coding genes	88	85
Structure RNAs	30	35
Genes with intron(s)	24	13
Coding rRNAs genes (% bp)	5.85	5.85
Coding tRNAs genes (% bp)	1.78	1.81
Protein-coding genes (% bp)	43.78	51.21
Noncoding regions (% bp)	48.59	41.13
Reference	This study	Nazareno et al., 2015

Note: IRA = inverted repeat region A; IRB = inverted repeat region B; LSC = large single copy; SSC = small single copy.

We found 66 SNPs in 30 individuals of *C. cujete* with 24 haplotypes, and 68 SNPs in six individuals of *C. amazonica* with six haplotypes. Haplotype diversity (h) was 0.98 and 1.00, nucleotide diversity (π) was 1.1×10^{-3} and 3.5×10^{-3}, and Watterson's estimator per site (θ_W) was 2.3×10^{-3} and 4.1×10^{-3} for *C. cujete* and *C. amazonica*, respectively. Diversity was about twice as high in *C. amazonica* compared to *C. cujete*. If *C. amazonica* was simply derived from *C. cujete*, as suggested by Gentry (1980), diversity should be comparable or potentially even slightly lower. Consequently, we rule out the hypothesis that *C. amazonica* is derived from *C. cujete*. However, at this point we cannot rule out either that domestication of *C. amazonica* led to *C. cujete* or that *C. cujete* is derived from other wild species from Central America.

CONCLUSIONS

Next-generation sequencing provided data to have a sufficient number of reads to perform a de novo assembly of the *C. cujete* chloroplast genome, the first assembled chloroplast in the *Crescentia* genus. The reconstructed *C. cujete* genome allowed the identification of SNPs in *C. amazonica* and *C. cujete* that produced diversity estimates that refuted the hypothesis that *C. amazonica* is derived from *C. cujete*, and will be useful in further studies about the origin, diversity, and spread of treegourds in the Neotropics.

LITERATURE CITED

Aguirre-Dugua, X., L. E. Eguiarte, A. González-Rodríguez, and A. Casas. 2012. Round and large: Morphological and genetic consequences of artificial selection on the gourd tree *Crescentia cujete* by the Maya of the Yucatan Peninsula, Mexico. *Annals of Botany* 109: 1297–1306.

Arango-Ulloa, J., A. Bohorquez, M. C. Duque, and B. L. Maass. 2009. Diversity of the calabash tree (*Crescentia cujete* L.) in Colombia. *Agroforestry Systems* 76: 543–553.

Chevreux, B., T. Wetter, and S. Suhai. 1999. Genome sequence assembly using trace signals and additional sequence information. *Proceedings of the German Conference on Bioinformatics* 99: 45-56.

Díaz, W. 2009. Composición florística de las comunidades vegetales aledañas al tercer puente sobre el río Orinoco, Venezuela. *Boletín del Centro de Investigaciones Biológicas* 43: 337–354.

Doyle, J. J., and J. L. Doyle. 1990. Isolation of plant DNA from fresh tissue. *Focus (San Francisco, Calif.)* 12: 13–15.

Drummond, A. J., B. Ashton, M. Cheung, J. Heled, M. Kearse, R. Moir, S. Stones-Havas, et al. 2009. Geneious version 4.8.5 for Windows. Computer program and documentation distributed by the author. Website http://www.geneious.com [accessed 15 April 2015].

Ducke, A. 1946. Plantas de cultura precolombiana na Amazônia brasileira. Notas sôbre as espécies ou formas espontâneas que supostamente lhes teriam dado origem. Instituto Agronômico do Norte, Belém, Brazil.

Gentry, A. H. 1980. Bignoniaceae, Part I (Crescentieaea and Tourrettieae). Flora Neotropica, monograph 25. New York Botanical Garden Press, Bronx, New York, USA.

Hahn, C., L. Bachmann, and B. Chevreux. 2013. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. *Nucleic Acids Research* 41: e129.

Koboldt, D. C., Q. Zhang, D. E. Larson, D. Shen, M. D. McLellan, L. Lin, C. A. Miller, E. R. Mardis, et al. 2012. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research* 22: 568–576.

Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* 25: 1754–1760.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, et al. 2009. The sequence alignment/map format and SAM tools. *Bioinformatics (Oxford, England)* 25: 2078–2079.

Li, X., T.-C. Zhang, Q. Qiao, Z. Ren, J. Zhao, T. Yonezawa, M. Hasegawa, et al. 2013. Complete chloroplast genome sequence of holoparasite *Cistanche deserticola* (Orobanchaceae) reveals gene loss and horizontal gene transfer from its host *Haloxylon ammodendron* (Chenopodiaceae). *PLoS ONE* 8: e58747.

Librado, P., and J. Rozas. 2009. DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics (Oxford, England)* 25: 1451–1452.

Lohse, M., O. Drechsel, S. Kahlau, and R. Bock. 2013. OrganellarGenomeDRAW: A suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. *Nucleic Acids Research* 41: W575–W581.

Mariac, C., N. Scarcelli, J. Pouzadou, A. Barnaud, C. Billot, A. Faye, S. Santoni, et al. 2014. Cost-effective enrichment hybridization capture of chloroplast genomes at deep multiplexing levels for population genetics and phylogeography studies. *Molecular Ecology Resources* 14: 1103–1113.

Nazareno, A. G., M. Carlsen, and L. G. Lohmann. 2015. Complete chloroplast genome of *Tanaecium tetragonolobum*: The first Bignoniaceae plastome. *PLoS ONE* 10: e0129930.

Scarcelli, N., C. Mariac, T. L. P. Couvreur, A. Faye, D. Richard, F. Sabot, C. Berthouly-Salazar, and Y. Vigouroux. 2016. Intra-individual polymorphism in chloroplasts from NGS data: Where does it come from and how to handle it? *Molecular Ecology Resources* 16: 434–445.

Wang, R.-J., C.-L. Cheng, C.-C. Chang, C.-L. Wu, T.-M. Su, and S.-M. Chaw. 2008. Dynamics and evolution of the inverted repeat–large single copy junctions in the chloroplast genomes of monocots. *BMC Evolutionary Biology* 8: 36.

Wittmann, F., J. Schöngart, J. C. Montero, T. Motzer, W. J. Junk, M. T. Piedade, L. H. Queiroz, and M. Worbes. 2006. Tree species composition and diversity gradients in white-water forests across the Amazon Basin. *Journal of Biogeography* 33: 1334–1347.

Appendix 1. The number of mapped reads and geographical information for the 30 *Crescentia cujete* and six *C. amazonica* samples from the Brazilian Amazon Basin used to analyze chloroplast diversity in this study.

Species	Sample	No. of reads (bp)	Municipality, State	Geographic coordinates
Crescentia cujete L.	I2R26T92	23,923	Barcelos, Amazonas	0°58′12″S, 62°55′12″W
	I8R26T23	16,574	Barcelos, Amazonas	0°06′36″S, 64°01′48″W
	I10R26T58	64,404	Barcelos, Amazonas	0°59′24″S, 62°55′48″W
	R21T35	37,991	Caracaraí, Roraima	1°44′24″N, 61°08′24″W
	I2R26T102	46,813	Caracaraí, Roraima	1°28′12″N, 60°53′24″W
	I8R26T10	15,324	Fonte Boa, Amazonas	2°31′12″S, 65°55′48″W
	I8R26T24	25,859	Fonte Boa, Amazonas	2°28′48″S, 65°58′48″W
	I11R26T37	268,499	Novo Aripuanã, Amazonas*	5°20′24″S, 60°26′24″W
	I1R26T28	33,170	Manaus, Amazonas	2°47′24″S, 60°02′24″W
	I2R26T87	195,654	Manaus, Amazonas	3°10′12″S, 59°54′36″W
	I1R26T12	41,714	Manicoré, Amazonas	5°51′36″S, 61°19′12″W
	I2R26T91	22,462	Manicoré, Amazonas	5°58′12″S, 61°28′12″W
	I1R26T10	121,056	Novo Aripuanã, Amazonas	5°19′48″S, 60°25′48″W
	I2R26T99	79,280	Parintins, Amazonas	2°33′36″S, 56°53′24″W
	I8R26T18	28,994	Parintins, Amazonas	2°33′S, 56°54′W
	I8R26T40	25,895	Parintins, Amazonas	2°33′36″S, 56°53′24″W
	R21T29	48,508	Santarém, Pará	2°08′24″S, 54°44′24″W
	I2R26T76	33,679	Santarém, Pará	2°07′12″S, 54°43′12″W
	I10R26T90	65,648	Santarém, Pará	2°28′12″S, 54°46′48″W
	I8R26T7	46,862	São Gabriel da Cachoeira, Amazonas	0°46′12″N, 67°14′24″W
	I10R26T92	68,994	São Gabriel da Cachoeira, Amazonas	0°46′12″N, 67°14′24″W
	I2R26T101	56,728	São Luís do Anauá, Amazonas	1°04′48″N, 60°11′24″W
	I1R26T35	36,852	São Paulo de Olivença, Amazonas	3°24′S, 68°39′36″W
	I2R26T67	26,431	Tabatinga, Amazonas	4°13′12″S, 69°54′36″W
	I10R26T64	22,805	Tabatinga, Amazonas	4°11′24″S, 69°54′36″W
	I1R26T19	108,320	Tefé, Amazonas	3°24′36″S, 64°33′W
	I1R26T42	45,478	Tefé, Amazonas	3°17′24″S, 64°41′24″W
	I2R26T80	84,319	Tefé, Amazonas	3°24′36″S, 64°32′24″W
	I10R26T82	28,696	Tefé, Amazonas	3°17′24″S, 64°41′24″W
	I10R26T61	81,698	Tefé, Amazonas	2°28′48″S, 64°45′W
Crescentia amazonica Ducke[a]	I11R26T51	49,100	Borba, Amazonas	4°19′48″S, 59°42′36″W
	R18T23	33,401	Manaus, Amazonas	3°14′24″S, 59°57′W
	I1R26T45	17,428	Manaus, Amazonas	3°15′S, 59°57′36″W
	I1R26T32	21,302	Santarém, Pará	2°07′12″S, 54°43′48″W
	R21T13	24,231	São Paulo de Olivença, Amazonas	3°21′S, 68°37′48″W
	I1R26T44	20,820	São Paulo de Olivença, Amazonas	3°27′36″S, 69°02′24″W

**Crescentia cujete* sample used to reconstruct the chloroplast sequence in this study.

[a]Vouchers of *Crescentia amazonica* from Borba and Santarém were deposited in the Instituto Nacional de Pesquisas da Amazônia (INPA) Herbarium (numbers 255.829 and 266.725, respectively).