



UFG

**UNIVERSIDADE FEDERAL DE GOIÁS
ESCOLA DE AGRONOMIA
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA E
MELHORAMENTO DE PLANTAS**

**CARACTERIZAÇÃO PARCIAL E TAMANHO DO
GENOMA DE *Stryphnodendron adstringens*
(LEGUMINOSAE)**

UERIC JOSÉ BORGES DE SOUZA

Orientadora:
Profa. Dra. Mariana Pires de Campos Telles

**TERMO DE CIÊNCIA E DE AUTORIZAÇÃO PARA DISPONIBILIZAR
VERSÕES ELETRÔNICAS DE TESES E DISSERTAÇÕES
NA BIBLIOTECA DIGITAL DA UFG**

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), regulamentada pela Resolução CEPEC nº 832/2007, sem ressarcimento dos direitos autorais, de acordo com a Lei nº 9610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou *download*, a título de divulgação da produção científica brasileira, a partir desta data.

1. Identificação do material bibliográfico: ☐ Dissertação ☒ Tese

2. Identificação da Tese ou Dissertação:

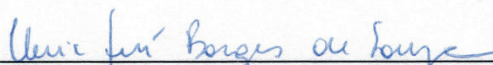
Nome completo do autor: Ueric José Borges de Souza

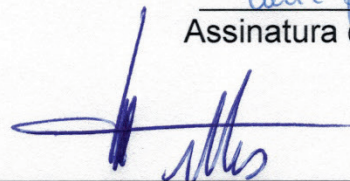
Título do trabalho: Caracterização parcial e tamanho do genoma de *Stryphnodendron adstringens* (LEGUMINOSAE).

3. Informações de acesso ao documento:

Concorda com a liberação total do documento ☒ SIM ☐ NÃO¹

Havendo concordância com a disponibilização eletrônica, torna-se imprescindível o envio do(s) arquivo(s) em formato digital PDF da tese ou dissertação.


Assinatura do(a) autor(a)² Ciente e de acordo:


Assinatura do(a) orientador(a)²

Data: 25/01/2019

¹ Neste caso o documento será embargado por até um ano a partir da data de defesa. A extensão deste prazo suscita justificativa junto à coordenação do curso. Os dados do documento não serão disponibilizados durante o período de embargo.

Casos de embargo:

- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro;
- Publicação da dissertação/tese em livro.

² A assinatura deve ser escaneada.

UERIC JOSÉ BORGES DE SOUZA

**CARACTERIZAÇÃO PARCIAL E TAMANHO DO
GENOMA DE *Stryphnodendron adstringens* (LEGUMINOSAE)**

Tese apresentada ao Programa de Pós-Graduação
em Genética e Melhoramento de Plantas, da
Universidade Federal de Goiás, como exigência
para obtenção do título de Doutor em Genética e
Melhoramento de Plantas.

Orientadora:
Profa. Dra. Mariana Pires de Campos Telles

Goiânia, GO – Brasil

2019

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Souza, Ueric José Borges de
Caracterização parcial e tamanho do genoma de *Stryphnodendron
adstringens* (Leguminosae) [manuscrito] / Ueric José Borges de
Souza. - 2019.
117 f.

Orientador: Profa. Dra. Mariana Pires de Campos Telles .
Tese (Doutorado) - Universidade Federal de Goiás, Escola de
Agronomia e Engenharia de Alimentos (EAEA), Programa de Pós
Graduação em Genética & Melhoramentos de Plantas , Goiânia, 2019.

1. Barbatimão. 2. Elementos repetitivos. 3. Genes. 4. Genoma
nuclear. 5. Genoma cloroplastidial. I. , Mariana Pires de Campos
Telles, orient. II. Título.

CDU 575

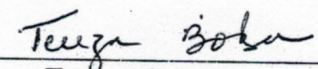


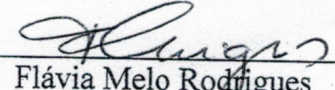
MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE GOIÁS
ESCOLA DE AGRONOMIA
PROGRAMA DE PÓS-GRADUAÇÃO
EM GENÉTICA E MELHORAMENTO DE PLANTAS

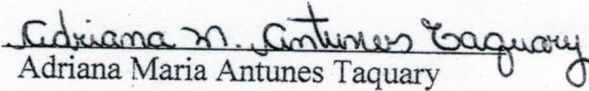


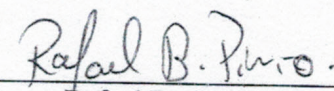
ATA DE DEFESA PÚBLICA DE TESE DE UERIC JOSÉ BORGES DE SOUZA. Aos vinte e cinco dias do mês de janeiro do ano de dois mil e dezenove (25/01/2019), às 08:00, no Auditório do ICB I da UFG, reuniram-se os membros da Banca Examinadora: Dr^a. Mariana Pires de Campos Telles (Presidente/Orientadora), Dr^a. Flávia de Melo Rodrigues, Dr^a. Tereza Cristina de Oliveira Borba, Dr^a. Adriana Maria Antunes Taquary e Dr. Rafael Barbosa Pinto. Sob a presidência da Orientadora, em sessão pública, procedeu-se à avaliação da defesa da Tese intitulada: **CARACTERIZAÇÃO PARCIAL E TAMANHO DO GENOMA DE *Stryphnodendron adstringens* (LEGUMINOSAE)**, de autoria de **UERIC JOSÉ BORGES DE SOUZA**, discente do Programa de Pós-Graduação em Genética e Melhoramento de Plantas, no nível de Doutorado, da Universidade Federal de Goiás. A sessão foi aberta pela presidente da Banca Examinadora, Dr^a. Mariana Pires de Campos Telles, que fez a apresentação formal dos membros da Banca. Em seguida, a palavra foi concedida ao autor da tese que, em 40 minutos, apresentou o seu trabalho. Terminada a apresentação, cada membro da Banca arguiu o doutorando; tendo-se adotado o sistema de diálogo sequencial. Ao final, a Banca reunida em separado procedeu à avaliação da defesa. A Tese foi considerada aprovada pela Banca Examinadora, cumprindo integralmente este requisito para fins de obtenção do título de DOUTOR EM GENÉTICA E MELHORAMENTO DE PLANTAS, pela Universidade Federal de Goiás, em conformidade com o estabelecido pela Resolução nº 1403/2016, do Conselho de Ensino, Pesquisa, Extensão e Cultura da UFG (CEPEC/UFG), que regulamenta o Programa de Pós-Graduação em Genética e Melhoramento de Plantas. Para fins de publicação eletrônica, o doutorando poderá efetuar as modificações eventualmente sugeridas pela Banca Examinadora e encaminhar à Secretaria do PGMP, respeitando-se o prazo máximo de 30 dias após a data da Defesa. A conclusão do curso e a emissão do diploma dar-se-ão em conformidade com o estabelecido pela Resolução CEPEC nº 1403/2016. Cumpridas as formalidades de pauta, às 11:53 a presidência da mesa encerrou esta sessão de defesa da tese e, para constar eu, Jéssica Almeida Silva, Secretária do PGMP, lavrei a presente Ata que depois de lida e aprovada, segue assinada pelos membros da Banca Examinadora, em duas vias de igual teor.


Mariana Pires de Campos Telles
Presidente/Orientadora


Tereza Cristina de Oliveira Borba
Membro Externo


Flávia de Melo Rodrigues
Membro Externo


Adriana Maria Antunes Taquary
Membro Externo


Rafael Barbosa Pinto
Membro Externo

*Aos meus queridos pais
Eurico (in memoriam) e Doracina
por todo amor e incentivo*

DEDICO

AGRADECIMENTOS

Agradeço aos os órgãos de fomento que viabilizaram e tornaram possível o desenvolvimento e realização deste trabalho. Ao Prof. Dr. José Alexandre Felizola Diniz Filho pela coordenação dos projetos financiados pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, Processo: nº 402178/2016-5) e Fundação de Amparo à Pesquisa do Estado de Goiás (FAPEG; Chamada: nº 07/2014, Processo: nº 201410267001736), pela oportunidade de realização deste trabalho com os recursos para custeio que deram origem aos dados trabalhados na tese.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela concessão da bolsa de estudo durante três anos de doutorado. Agradeço também à equipe do INCT em Ecologia, Evolução e Conservação da Biodiversidade (EECBio) pela oportunidade da bolsa DTI que permitiu a finalização do doutorado.

Quero agradecer à minha orientadora Profa. Dra. Mariana Pires de Campos Telles, pela oportunidade e confiança inicial concedida para trabalhar neste projeto. Pela sua orientação, apoio, paciência e disponibilidade para discussões e sugestões, que foram de grande valia para conclusão deste trabalho. Agradeço por ser uma líder inspiradora que participou efetivamente da minha formação.

À toda equipe do Laboratório de Genética & Biodiversidade (LGBio), que de forma direta ou indireta tiveram contribuições fundamentais para realização deste trabalho e da minha formação. À Dra. Cíntia Pelegrineti Targueta de Azevedo Brito pelo apoio na condução das atividades técnicas no LGBio e ao Dr. Fernando Roa pela disponibilidade em auxiliar na etapa de citometria de fluxo.

Aos amigos que a pós-graduação me trouxe, em especial: Stella Cristina, Luciana Vitorino, Edivaldo Júnior, Rhewter Nunes, Fernanda Fraga, Tatianne Piza, Ramilla Braga, Ariany Rosa, Rejane Araújo, Stela Ribeiro, Vanessa Bernardes, Josimar Moraes, e Kássia Marques.

Um agradecimento especial à minha família, principalmente minha mãe Doracina por todo seu incentivo, amor e compreensão sempre presentes. Ao meu irmão Ueslei e minha irmã de coração Poliana, que sempre acreditaram e torceram para que esse momento se concretizasse. Um agradecimento especial também ao Frederico Cesário, meu companheiro que esteve ao meu lado durante esses quatro anos.

SUMÁRIO

RESUMO GERAL.....	10
LISTA DE FIGURAS.....	12
LISTA DE TABELAS.....	14
LISTA DE MATERIAL SUPLEMENTAR.....	15
1 INTRODUÇÃO GERAL.....	16
2 THE COMPLETE CHLOROPLAST GENOME OF <i>Stryphnodendron</i> <i>adstringens</i> (LEGUMINOSAE - CAESALPINIOIDEAE): COMPARATIVE ANALYSIS WITH RELATED MIMOSOID SPECIES.....	20
ABSTRACT.....	21
2.1 INTRODUCTION.....	22
2.2 MATERIAL AND METHODS.....	23
2.2.1 DNA extraction and chloroplast genome sequencing.....	23
2.2.2 Chloroplast Genome Assembly and Annotation.....	24
2.2.3 Characterization of repeat sequences.....	25
2.2.4 Nucleotide Diversity and Synonymous (Ks) and non-synonymous (Ka) substitution rate analysis.....	25
2.2.5 Comparative analysis of genome structure.....	25
2.2.6 Phylogenetic analyses.....	26
2.3 RESULTS AND DISCUSSION.....	26
2.3.1 Genome Assembly and Annotation.....	26
2.3.2 Repeat sequence analysis.....	31
2.3.3 Nucleotide Diversity and Ka/Ks Ratio.....	32
2.3.4 Comparative analysis of genome structure.....	37
2.3.5 Phylogenetic relationships.....	41
2.4 CONCLUSION.....	43
2.5 REFERENCES.....	44
3 TAMANHO DO GENOMA, MONTAGEM “DE NOVO” E ANOTAÇÃO DO GENOMA NUCLEAR DE <i>Stryphnodendron</i> <i>adstringens</i> (LEGUMINOSAE).....	65

RESUMO.....	66
3.1 INTRODUÇÃO.....	67
3.2 MATERIAL E MÉTODOS.....	69
3.2.1 Extração do DNA genômico e sequenciamento.....	69
3.2.2 Tamanho do genoma.....	69
3.2.2.1 Estimativa do tamanho do genoma por Citometria de fluxo.....	69
3.2.2.2 Estimativa do tamanho do genoma por análise de <i>k-mers</i>	70
3.2.3 Montagem do genoma.....	71
3.2.4 Anotação de elementos repetitivos.....	71
3.2.4.1 Caracterização de Elementos Transponíveis.....	71
3.2.4.2 Caracterização de regiões microssatélites.....	72
3.2.5 Predição e anotação funcional de genes.....	73
3.3 RESULTADOS E DISCUSSÃO.....	73
3.3.1 Tamanho do genoma.....	73
3.3.2 Montagem do genoma.....	76
3.3.3 Caracterização de Elementos Transponíveis.....	78
3.3.4 Caracterização de regiões microssatélites.....	85
3.3.5 Predição e anotação de genes.....	89
3.3.5.1 Predição de genes.....	89
3.3.5.2 Anotação funcional dos transcritos de <i>Stryphnodendron adstringens</i>	93
3.4 CONCLUSÃO.....	101
3.5 REFERÊNCIAS.....	102
4 CONCLUSÕES GERAIS.....	115
5 REFERÊNCIAS.....	116

RESUMO

SOUZA, U. J. B. **CARACTERIZAÇÃO PARCIAL E TAMANHO DO GENOMA DE *Stryphnodendron adstringens* (LEGUMINOSAE)**. 2019. 114 f. Tese (Doutorado em Genética e Melhoramento de Plantas) - Escola de Agronomia, Universidade Federal de Goiás, Goiânia, 2019.¹

A espécie *Stryphnodendron adstringens* (Mart.) Coville, conhecida popularmente como barbatimão, é amplamente utilizada na medicina popular como fitoterápico, para tratamentos de diversas doenças. Devido seu alto potencial de aplicação na medicina, a caracterização do genoma desta espécie, com dados de sequenciamento em larga escala, pode contribuir grandemente, quando consideramos que existem ainda poucos estudos ao nível de genoma em espécies de plantas medicinais. Os objetivos deste trabalho foram estimar o tamanho do genoma de *S. adstringens* e realizar a montagem e anotação dos genomas cloroplastidial e nuclear usando dados de sequenciamento *Illumina*. O genoma cloroplastidial de *S. adstringens* foi montado em uma única sequência contendo 162.169 bp, sendo identificados 77 genes codificadores de proteínas, 30 genes de tRNA e 4 genes de rRNA. Um total de 137 microssatélites e 42 estruturas de repetição foram identificadas, com a maior proporção na região LSC (*Large Single Copy*). A comparação da estrutura genoma cloroplastidial de *S. adstringens* com outras espécies do clado mimosoide demonstrou que regiões gênicas são altamente conservadas. A reconstrução filogenética usando 73 genes codificantes de proteínas sugere que a família Leguminosae-Caesalpinoideae é parafilética. O tamanho do genoma nuclear de *S. adstringens* foi estimado em $1C = 0.684$ pg, o que corresponde a aproximadamente 669 Mpb. O genoma nuclear foi montado usando estratégias de novo em 63.320 *contigs* com N50 igual a 14.541 pb, o que corresponde a 495 Mb e 76,3% do tamanho estimado do genoma. Um total de 20.124 regiões microssatélites foi identificado no genoma nuclear, sendo a repetição AT/TA mais frequente (44,16%). Elementos transponíveis foram encontrados em 44,58% do genoma e os retroelementos com repetições terminais longas (LTR) foram os mais abundantes. Foram anotados 31.636 genes codificadores de proteínas e o número estimado de mRNAs transcritos pelo conjunto de genes nucleares foi igual a 95.811, com uma média de 3,03 transcritos por gene. Do total de transcritos preditos, 60.798 foram classificados funcionalmente, nos termos do Gene Ontology. Cerca de 1.521 transcritos (1,58%) foram similares com sequências de proteínas envolvidas nas vias de biossíntese de taninos. Este trabalho representa um grande avanço nos conhecimentos sobre a estrutura e a organização dos genomas nuclear e cloroplastidial de *S. adstringens*. Além disso, os resultados obtidos fornecem recursos genômicos e abrem caminho para novas investigações genéticas e biotecnológicas para a espécie.

Palavras-chave: barbatimão, elementos repetitivos, genes, genoma nuclear, genoma cloroplastidial, mimosoide.

¹ Orientadora: Profa. Dra. Mariana Pires de Campos Telles. ICB – UFG.

ABSTRACT

SOUZA, U. J. B. **PARTIAL CHARACTERIZATION AND GENOME SIZE OF *Stryphnodendron adstringens* (LEGUMINOSAE).** 2019. 114 f. Thesis (PhD in Genetics and Plant Breeding) – Escola de Agronomia, Universidade Federal de Goiás, Goiânia, 2019.¹

Stryphnodendron adstringens (Mart.) Coville (Barbatimão) is one of the plant species from Cerrado, locally known as 'barbatimão', widely used in folk medicine for treatments of various diseases. Due to its high potential for application in medicine, the genomic characterization of this species with large-scale sequencing data can contribute greatly when we consider that there are few studies at the genome level in medicinal plant species. The purpose of this work were to estimate the genome size of *S. adstringens* and to perform the assembly and annotation of the chloroplast and nuclear genomes using Illumina sequencing data. The chloroplast genome of *S. adstringens* was assembled in a single sequence with 162,169 base pairs, comprising 77 protein-coding genes, 30 tRNA genes and four ribosomal RNA genes. A total of 137 microsatellites and 42 repeat structures were identified, with the highest proportion in the LSC (Large Single Copy) region. The comparison of the chloroplast genome structure of *S. adstringens* with other species of the mimosoid clade demonstrated that gene regions are highly conserved. Phylogenetic reconstruction using 73 protein-coding genes suggests that the Leguminosae-Caesalpinoideae family is paraphyletic. The nuclear genome size of *S. adstringens* was estimated at 1C = 0.684 pg, which corresponds to approximately 669 Mpb. The nuclear genome was assembled using de novo genome assembly in 63,320 contigs with N50 equal to 14,541 bp, corresponding to 495 Mb and 76.3% of estimated genome size. A total of 20,124 microsatellite regions were identified in the nuclear genome, with AT/TA repeat being more frequent (44.16%). Transposable elements were found in 44.58% of the genome and the retroelements containing long terminal repetitions (LTR) were the most abundant. It was predicted 31,636 protein-coding nuclear genes and the estimated number of mRNAs transcribed was 95,811, with a mean of 3.03 transcripts per gene. Of the total number of predicted transcripts, 60,798 were classified functionally, according to Gene Ontology. About 1,521 transcripts (1.58%) were similar with protein sequences involved in the tannin biosynthesis pathways. This work represents a great advance in the knowledge on the structure and organization of the nuclear and chloroplast genomes of *S. adstringens*. In addition, the results obtained provide genomic resources and open the way to new genetic and biotechnological investigations for the species.

Keywords: barbatimão, repetitive elements, genes, nuclear genome, chloroplast genome, mimosoid.

¹Advisor: Profa. Dra. Mariana Pires de Campos Telles. ICB – UFG.

LISTA DE FIGURAS

CAPÍTULO 1

- Figure 1. Gene map of the *Stryphnodendron adstringens* chloroplast genome. The genes drawn outside and inside of the circle are transcribed in clockwise and counterclockwise directions, respectively. Genes were colored based on their functional groups. The inner circle shows the quadripartite structure of the chloroplast: small single copy (SSC), large single copy (LSC) and a pair of inverted repeats (IRa and IRb). The gray ring marks the GC content with the inner circle marking a 50% threshold. Asterisks mark genes that have introns. 28
- Figure 2. Number and type of simple sequence repeats in *Stryphnodendron adstringens* chloroplast genome. 32
- Figure 3. Sliding window analysis of Mimosoid chloroplast genome..... 33
- Figure 4. The Ka/Ks ratio of Mimosoid chloroplast genomes for individual genes..... 34
- Figure 5. Visualization of genome alignment of six Mimosoid chloroplast genomes using *Stryphnodendron adstringens* as reference. The vertical scale indicates the percent of identity, ranging from 50% to 100%. Coding regions are marked in purple, and non-coding as red. The horizontal axis indicates the coordinates within the chloroplast genome. 39
- Figure 6. Comparison of the junction sites between the Long Single Copy (LSC, light blue), Short Single Copy (SSC, light green) and Inverted Repeat (IRa and IRb, orange) regions among the six Mimosoid chloroplast genomes. JLB (IRb/LSC), JSB (IRb/SSC) JSA (SSC/IRa) and JLA (IRa/LSC) denote the junction sites between each corresponding regions on the genome. 40
- Figure 7. Maximum credibility tree reconstructed based on 73 conserved coding-protein genes from twenty-one species. All nodes of the phylogenetic tree are supported by 1.00 Bayesian inference posterior probability. 42

LISTA DE FIGURAS

CAPÍTULO 2

Figura 1. Histogramas de fluorescência relativa (unidades arbitrárias) de núcleos corados com iodeto de propídio obtidos por Citometria de Fluxo. Sa: <i>Stryphnodendron adstringens</i> . Sl: <i>Solanum lycopersicum</i> . Rs: <i>Raphanus sativus</i>	74
Figura 2. Distribuição de <i>k-mers</i> nos reads de sequenciamento da espécie <i>Stryphnodendron adstringens</i>	75
Figura 3. Distribuição do tamanho final do contigs obtidos na montagem “de novo” do genoma de <i>Stryphnodendron adstringens</i>	77
Figura 4. Distribuição do total de microssatélites encontrados no genoma da espécie <i>Stryphnodendron adstringens</i>	85
Figura 5. Número de repetições em tandem para cada um dos motivos de repetição encontrados no genoma da espécie <i>Stryphnodendron adstringens</i>	87
Figura 6. Frequência relativa da ocorrência de genes de acordo com o número de contigs obtidos com a montagem do genoma de <i>Stryphnodendron adstringens</i>	90
Figura 7. Distribuição da frequência relativa dos genes preditos na montagem do genoma de <i>Stryphnodendron adstringens</i> de acordo com seus tamanhos em pares de base.	91
Figura 8. Distribuição da frequência relativa do número de transcritos por gene na montagem do genoma de <i>Stryphnodendron adstringens</i>	92
Figura 9. Distribuição da frequência relativa do tamanho dos transcritos na montagem do genoma de <i>Stryphnodendron adstringens</i>	92
Figura 10. Número de transcritos com BLAST-HITS e com anotação nos termos Gene Ontology.	94
Figura 11. Distribuição das espécies Top-Hits do Blast.....	95
Figura 12. Anotação funcional dos transcritos encontrados no genoma de <i>Stryphnodendron adstringens</i>	98
Figura 13. Número total de transcritos no genoma de <i>Stryphnodendron adstringens</i> com similaridade com proteínas envolvidas nas vias de biossíntese de taninos.	99
Figura 14. Via biosintética de taninos condensados. PAL (phenylalanine ammonialyase); C4H (cinnamate 4-hydroxylase); 4CL (4-coumarate CoA ligase); CHS (chalcone syntase); DFR (dihydroflavonol 4-reductase); LAR (leucoanthocyanidin reductase); ANR (anthocyanidin reductase). Adaptada de Chen et al. (2018).....	100

LISTA DE TABELAS

CAPÍTULO 1

Table 1. Chloroplast genome information from sampled Mimosoid species and the newly assembled <i>Stryphnodendron adstringens</i>	27
Table 2. List of genes in the chloroplast genome of <i>Stryphnodendron adstringens</i>	30
Table 3. The Ka, Ks and Ka/Ks ratio of Mimosoid chloroplast genome for individual genes and region.	35

CAPÍTULO 2

Tabela 1. Métricas utilizadas para avaliar a montagem do genoma da espécie <i>Stryphnodendron adstringens</i>	78
Tabela 2. Anotação do total de elementos transponíveis no genoma de <i>Stryphnodendron adstringens</i> usando o RepeatMasker.....	82
Tabela 3. Anotação do total de elementos transponíveis no genoma de <i>Stryphnodendron adstringens</i> usando o RepeatModeler.	83

LISTA DE MATERIAL SUPLEMENTAR

CAPÍTULO 1

Table S1. Accession number and sampled chloroplast genomes obtained from GenBank.	50
Table S2. Models, Gamma Distribution, Proportion of Invariant Sites and Kappa, as Estimated by jModelTest for Each Gene Alignment.	51
Table S3. Forward, Reverse and palindromic repeat sequences in the <i>Stryphnodendron adstringens</i> chloroplast genome.	53
Table S4. Distribution of simple sequence repeat (SSR) in the <i>Stryphnodendron adstringens</i> chloroplast genome.	58

CAPÍTULO 2

Tabela suplementar 1. Tipos de motivos de repetição encontrados no genoma da espécie <i>Stryphnodendron adstringens</i>	110
--	-----

1 INTRODUÇÃO GERAL

As metodologias clássicas de sequenciamento tiveram início na década de 70, quando o método de sequenciamento de Sanger foi proposto, sendo então utilizado para maioria das aplicações de análises de sequências de DNA dos organismos (Van Dijk et al., 2014). Esse método passou por melhorias contínuas ao longo dos anos, culminando na introdução dos sistemas de eletroforese capilar que são utilizados até o presente para o sequenciamento clássico. No entanto, para ampliar de forma significativa a quantidade de dados genômicos e reduzir o custo e o tempo dos projetos de sequenciamento em larga escala, novas metodologias com capacidade de gerar milhares de sequências de DNA em um único processo, o que ficou conhecido como sequenciamento de nova geração (NGS – *Next-Generation Sequencing*) foram desenvolvidas (Metzker, 2010; Mardis, 2011, 2013). Isso permitiu um aumento do volume de projetos de genômica no mundo, viabilizando o sequenciamento de organismos modelos e não modelos (Schuster, 2008; Liu et al., 2012; Mardis, 2013).

O primeiro *draft* de sequências genômicas nucleares de uma espécie não modelo medicinal usando dados de sequenciamento de nova geração, foi para *Cannabis sativa*, conhecida pela concentração de compostos psicoativos (canabinóides) (Van Bakel et al., 2011). Os genes e vias de biossíntese relacionadas com canabinóides foram explorados com a geração de um *draft* de 534 Mb (Van Bakel et al., 2011). Chen et al. (2011) iniciaram um projeto na China, chamado “*Herb Genome Programme*” com o objetivo de realizar o sequenciamento do genoma e estudar as vias de síntese de metabólitos secundários em várias plantas medicinais. Como parte desse projeto, o genoma de algumas ervas medicinais comumente usadas, tais como: *Ganoderma lucidum*, *Salvia miltiorrhiza* e *Catharanthus roseus* foram sequenciadas (Chen et al., 2012; Kellner et al., 2015; Xu et al., 2016). Essas espécies, ricas em componentes farmacêuticos ativos, incluindo triterpenos, quinonas terpênicas e alcalóides indólicos, confirmaram-se como recursos valiosos para o estudo de atividades envolvidas na biossíntese desses metabolitos em espécies herbáceas.

As tecnologias de sequenciamento de nova geração, também têm sido amplamente utilizadas para sequenciar genomas cloroplastidiais. As bibliotecas genômicas de DNA (gDNA) de plantas, são compostas por DNA nuclear e organelar. Com isso, a montagem do genoma cloroplastidial pode ser realizada a partir dos *reads* gerados pelo sequenciamento de bibliotecas de gDNA, sem a necessidade de enriquecimento ou isolamento prévio de

DNA cloroplastidial (Twyford & Ness, 2017). Os principais desafios, no entanto, não estão mais relacionados com a aquisição de dados, mas com metodologias de análises, para extrair e montar o genoma cloroplastidial do total de mega/giga bases de sequências geradas (Tonti-Filippini et al., 2017). A montagem pode ser dividida basicamente em duas abordagens: *de novo* e *reference-guided*. A abordagem *reference-guided* é efetiva quando se têm disponível o genoma cloroplastidial de uma espécie próxima do ponto de vista evolutivo. Quando não existe um genoma referência, a montagem deve ser realizada utilizando abordagem *de novo* (Dierckxsens et al., 2016; Tonti-Filippini et al., 2017; Twyford & Ness, 2017). Em ambos os casos, as estratégias comumente usadas para a montagem, consistem em filtrar as sequências cloroplastidiais, seja mantendo regiões de alta cobertura, ou pelo mapeamento dos *reads* contra um banco de dados com genomas cloroplastidiais de espécies próximas do ponto de vista evolutivo (Dierckxsens et al., 2016).

O Cerrado é um bioma de características únicas no planeta ocupando aproximadamente 2 milhões de km² no território Brasileiro e reconhecido por sua alta biodiversidade (Bridgewater et al., 2004). Dentre a infinidade de recursos vegetais encontrada no Cerrado, destacam-se as plantas utilizadas pelas comunidades tradicionais em diversos tipos de aplicações, tais como, as espécies do gênero *Stryphnodendron* que são amplamente usadas com propósito medicinal (Souza & Felfili, 2006; Feitosa et al., 2014). O extrato da casca do caule é rico em tanino, característica que deu origem ao nome do gênero (*Stryphnos*=adstringente, *dendron*=madeira) (Scalon, 2007). Dentre as espécies do gênero, *Stryphnodendron adstringens* (Mart.) Coville, conhecida popularmente como “barbatimão”, é considerada uma das maiores fontes naturais de tanino. O teor de taninos totais em extratos e frações de cascas *S. adstringens* varia entre 20 e 40% (Corrêa et al., 2012).

Os taninos são compostos fenólicos, classificados como metabólitos secundários, importantes nos mecanismos de defesa da planta contra radiação solar, herbívoros e agentes patogênicos (Carvalho et al., 2014). Esse composto tem atraído o interesse da comunidade científica por suas propriedades químicas, responsáveis pela atividade antioxidante, antimicrobiana, antisséptica e anti-inflamatória (Souza et al., 2007; Souza et al., 2007; Fiori et al., 2013), as quais, fazem do “barbatimão” uma planta amplamente utilizada na medicina popular como fitoterápico no tratamento de diversas doenças (Dias & Laureano, 2010).

Devido à sua importância medicinal, essa espécie nativa do bioma Cerrado está inserida na “Iniciativa Plantas para o Futuro – Região Centro-Oeste”, ação do Ministério do Meio Ambiente (MMA), o qual busca o uso sustentável das espécies nativas da flora

brasileira de valor econômico atual ou potencial, utilizadas local ou regionalmente (Vieira et al., 2016). Muitas espécies medicinais da Região Centro-Oeste, possuem mercado regional amplo e estabelecido (Vieira et al., 2016). Por exemplo, para a espécie *Stryphnodendron adstringens*, pode-se citar a pomada FITOSCAR® fabricada pelo Laboratório Aspen, com pesquisa desenvolvida por uma universidade brasileira (Faculdade de Medicina da Universidade de Ribeirão Preto). Essa pomada foi o primeiro medicamento elaborado a partir de extrato seco de uma planta do Cerrado a ter um registro concedido pela Agência Nacional de Vigilância Sanitária (ANVISA). Segundo a bula, a pomada é indicada como agente cicatrizante em vários tipos de lesões.

Apesar da importância medicinal e comercial do “barbatimão”, estudos na área da genética para essa espécie, ainda são relativamente recentes e restritos a poucos grupos de pesquisa no Brasil. Glasenapp et al. (2014) descreveu a diversidade genética em três populações de *S. adstringens* localizadas em unidades de conservação no Estado de Minas Gerais, usando marcadores isoenzimáticos. Mendonça et al. (2012) descreveu a diversidade genética em 12 acessos de populações naturais localizados nos estados de São Paulo, Minas Gerais e Goiás, usando marcadores AFLP (*Amplified Fragment Length Polymorphism*). Mendonça (2011) testou a transferibilidade de marcadores microssatélites de três espécies (*Anadenanthera colubrina*, *Hymenaea courbaril*, e *Copaifera langsdorffii*) para *S. adstringens* e observou que não houve transferência de microssatélites de *H. courbaril* e *C. langsdorffii*, enquanto que de *A. colubrina* 10 microssatélites foram transferidos, porém somente 1 apresentou polimorfismo.

Aspectos como estes citados acima, geram motivação para o estudo genômico da espécie via sequenciamento de DNA. Isso porque fornece a possibilidade de identificação de regiões microssatélites potenciais para o posterior desenvolvimento de marcadores moleculares, uma vez que, todos os trabalhos realizados com a espécie, até então, utilizaram dados de um conjunto restrito de marcadores. Além disto, fornece informações viáveis sobre os genes que estão envolvidos na via de biossíntese de taninos, compostos que estão diretamente relacionados com seu amplo uso econômico atual e/ou potencial.

Neste contexto, este trabalho teve como objetivo geral realizar a montagem e anotação do genoma cloroplastidial e nuclear de *S. adstringens* a partir dos *reads* gerados pelo processo de sequenciamento com o auxílio de softwares de bioinformática. O primeiro capítulo da tese descreve a montagem e anotação do genoma cloroplastidial da espécie *S. adstringens* e sua comparação com o genoma de outras espécies pertencentes ao clado

mimoside da subfamília Caesalpinoideae-Leguminosae. O segundo capítulo descreve a estimativa de tamanho de genoma, a montagem e a anotação de elementos repetitivos e genes no genoma nuclear desta mesma espécie.

CAPITULO 1

THE COMPLETE CHLOROPLAST GENOME OF *Stryphnodendron adstringens* (LEGUMINOSAE - CAESALPINIOIDEAE): COMPARATIVE ANALYSIS WITH RELATED MIMOSOID SPECIES¹

¹ Capítulo elaborado para submissão no periódico *Scientific Reports*

ABSTRACT

Stryphnodendron adstringens is a medicinal plant belonging to the Leguminosae family, and it is commonly found in the southeastern savannas and endemic to the Cerrado biome. The goal of this study was to assemble and annotate the chloroplast genome of *S. adstringens* and to compare it with previously known genomes of the mimosoid clade within Leguminosae. The chloroplast genome was reconstructed using *de novo* and referenced-based assembly of paired-end reads generated by shotgun sequencing of total genomic DNA. The size of the *S. adstringens* chloroplast genome was 162.169 bp. This genome included a large single-copy (LSC) region of 91.045 bp, a small single-copy (SSC) region of 19.014 bp and a pair of inverted repeats (IRa and IRb) of 26.055 bp each. The *S. adstringens* chloroplast genome contains a total of 113 functional genes, including 77 protein-coding genes, 30 transfer RNA genes, and 4 ribosomal RNA genes. A total of 137 SSRs and 42 repeat structures were identified in *S. adstringens* chloroplast genome, with the highest proportion in the LSC region. A comparison of the *S. adstringens* chloroplast genome with those from other mimosoid species indicated that gene content and synteny are highly conserved in the clade. The phylogenetic reconstruction using 73 conserved coding-protein genes from 19 Leguminosae species was supported to be paraphyletic. Furthermore, the noncoding and coding regions with high nucleotide diversity may supply valuable markers for molecular evolutionary and phylogenetic studies at different taxonomic levels in this group.

Keywords: barbatimão, molecular phylogenomics, plastome, repeats, sequence divergence.

2.1 INTRODUCTION

The chloroplast, which is considered to have originated from free-living cyanobacteria through endosymbiosis, plays an essential role in photosynthesis and in many processes in plant cells^{1–3}. In this evolutionary context, the chloroplast genome of angiosperms exhibit a highly conserved organization with a quadripartite structure, comprising two copies of inverted repeats (IRs), separated by large (LSC) and small (SSC) single-copy regions⁴.

The size of the chloroplast circular genome range between 120 and 160 kb in length⁵, but varies considerably both within and among plant families. For example, in the Geraniaceae, the size of the chloroplast genome ranges from 116.935 bp in *Erodium carvifolium*⁶ to 242.575 bp in *Pelargonium transvaalense* (Accession: NC_031206.1 unpublished). In the Leguminosae, the size ranges from 120.289 bp for *Lathyrus odoratus* (Accession: NC_027150.1 unpublished) up to 178.887 in *Pithecellobium flexicaule*⁷. The variations in size can be attributed mostly to the expansion, contraction or loss of IRs, as well as variation in the length of intergenic spacers^{5,8}.

Most angiosperm chloroplast genome usually contain 100–130 distinct genes, comprising of 80–90 protein coding genes and approximately 30 transfer RNA (tRNA) genes and four ribosomal RNA (rRNA) genes^{9,10}. The IR region comprises a duplicated set of tRNA and rRNA genes, whereas the single copy regions mostly consists of protein-coding genes involved in major functions, which include components of the photosynthetic machinery (such as photosystem I (PSI), photosystem II (PSII), the cytochrome b6/f complex, and the ATP synthase), transcription, and translation. Thus, the two IRs are identical in their nucleotide sequence, so that every gene contained within them is present in two copies per genome which only differ in their relative orientation^{9–11}.

The gene order and content of chloroplast genomes are generally highly conserved along plant evolution and the substitution rates are much lower than that of the nuclear genome¹². This fact, coupled with the non-recombinant nature and maternal inheritance in most angiosperms, makes plant chloroplasts genomes valuable sources of genetic markers for analyzing evolutionary relationships at multiple scales, ranging from short-term phylogeographic patterns up to phylogenetic relationships among large clades^{13,14}.

The first complete chloroplast genome sequences were decoded over 30 years ago, for *Nicotiana tabacum*¹⁵ and *Marchantia polymorpha*¹⁶. However, because the time and cost

associated with the conventional Sanger sequencing, the reconstruction of complete chloroplast genome was impractical for non-model species. More recently, with the advent of next-generation sequencing technology, whole genome sequencing has increased dramatically¹⁷. This offers an alternative way to obtain chloroplast genome based on downstream bioinformatics pipelines that allows distinguishing plastid reads from nuclear and mitochondrial reads¹⁸. Currently, approximately 1.654 eudicotyledons chloroplast genomes have been sequenced and deposited in the NCBI Organelle Genome, out of which 114 belong to the legume family (Leguminosae).

The Leguminosae is the third-largest angiosperm family, with approximately 751 genera and ca. 19.500 species^{19,20}. The Leguminosae was divided into three sub-families, the Caesalpinioideae, Mimosoid and Papilionoideae²⁰. However, a new classification of the legumes has been proposed by The Legume Phylogeny Working Group²¹. This paper uses a *matK* gene-based phylogeny and a wide Leguminosae sample (~90% of genera) to proposes a new organization of this group in six sub-families. The six sub-families are Caesalpinioideae, Cercidoideae, Detarioideae, Dialioideae, Duparquetioideae and Papilionoideae²¹. The traditional subfamily Mimosoid is now recognized as a distinct ‘mimosoid clade’ nested in the recircumscribed Caesalpinioideae²¹. Within the ‘mimosoid clade’, the genus *Stryphnodendron* Mart. includes approximately 21 species and two subspecies, mainly found in the South-American neotropical savannas²². Among them, *Stryphnodendron adstringens* (Mart.) Coville, popularly known as “barbatimão”, is a common tree in the Brazilian Savanna. It’s bark is used in the Brazilian traditional medicine and has several biological actions, including anti-inflammatory, antiulcerogenic and antimicrobial effects^{23–26}.

The goal of this study was to assemble the chloroplast genome of *S. adstringens* from whole genome sequence data, reporting the annotation and its structural characterization providing new genomic resources for this species and close related species. We also used a phylogenetic analysis to evaluate the sequence divergence in chloroplast regions of *S. adstringens* when compared with other known species of the mimosoid clade.

2.2 MATERIALS AND METHODS

2.2.1 DNA extraction and chloroplast genome sequencing

Fresh young leaves of *Stryphnodendron adstringens* were collected in Niquelândia, Goiás, Brazil (*Sisgen Registration: A4EE2BE*). Total genomic DNA was extracted using a CTAB protocol²⁷. DNA quantity were assessed using a NanoDrop spectrophotometer (Thermo Scientific) and DNA quality was evaluated using a 1% agarose gel. In addition, DNA was quantified through fluorometry using Qubit 2.0 (Life Technologies). Genomic library preparation was performed using a Nextera DNA Sample Preparation Kit (Illumina, San Diego, USA), following the manufacturer's instructions. The library was sequenced on a single lane in paired-end mode (2 x 100bp) using the HiSeq 2500 platform and V4 SBS kit (Illumina) at the University of São Paulo (Escola Superior de Agricultura Luiz de Queiroz da Universidade de São Paulo) in Piracicaba, Brazil.

2.2.2 Chloroplast Genome Assembly and Annotation

Paired-end Illumina raw reads were filtered and trimmed using Trimmomatic V.0.36²⁸ using the ILLUMINACLIP: NexteraPE-PE.fa:2:30:10 for adapter trimming, a sliding window of 10 base pairs with a minimum average quality score of 20 (SLIDINGWINDOW:10:20), and a minimum length of 40 bp (MINLEN:40).

The chloroplast genome of *S. adstringens* was reconstructed using a combination of *de novo* and reference-guided assemblies. To obtain the *de novo* chloroplast genome assembly, the paired-end sequence reads were mapped to five Mimosoid plastomes using Bowtie2 v.2.3.4.1²⁹ to exclude reads of nuclear and mitochondrial origins (*Adenanthera microsperma* Teijsm. & Binn. [accession no. NC_034986], *Dichrostachys cinerea* (L.) Wight & Arn. [accession no. NC_035346], *Leucaena trichandra* (Zucc.) Urb. [accession no. NC_028733], *Parkia javanica* (Lam.) Merr. [accession no. NC_034989], *Piptadenia communis* Benth. [accession no. NC_034990]). The obtained putative chloroplast reads were then used for *de novo* assembly using SPAdes 3.6.1 with iterative K-mer sizes of 55, 69 and 87³⁰. Reference guided assembly was performed with YASRA 2.32³¹ using *Piptadenia communis* Benth. as reference chloroplast genome. Contigs with coverage below than 10x were eliminated. The remaining *de novo* contigs were merged with reference-guided contigs in Sequencher 5.4.6 (Genecodes, Ann Arbor, Michigan, USA) based on at least 20 bps overlap and 98% similarity. Any discrepancies between *de novo* and reference-guided contigs were corrected by searching the high quality read pool using the UNIX 'grep' function. A "genome walking" technique, using the Unix "grep" function, was used to find

reads that could fill any gaps between contigs that did not assemble in the initial set of analyses.

Annotation of the chloroplast genome was performed using Verdant³² and Dual Organellar Genome Annotator-DOGMA³³, coupled with manual correction of start and stop codons and intron/exon boundaries. Transfer RNA (tRNA) genes were identified with DOGMA and the tRNAscan-SE program ver. 2.0³⁴ in organellar search mode with default parameters. The circular chloroplast genome map was drawn using OrganellarGenomeDRAW (OGDRAW)³⁵.

2.2.3 Characterization of repeat sequences

The sizes and locations of forward, reverse, palindromic and complementary repeats in the *S. adstringens* chloroplast genome were determined by REPuter³⁶ with a minimal size of 30 bp, hamming distance of 3 and over 90% identity. Simple sequence repeats (SSRs) were detected using the microsatellite identification tool MISA (available online: <http://pgrc.ipk-gatersleben.de/misa/misa.html>). The minimum number of SSRs was set to ten repeat units for mononucleotide, five repeat units for dinucleotide, four repeat units for trinucleotide and three repeat units for tetra-, penta- and hexanucleotide.

2.2.4 Nucleotide Diversity and Synonymous (Ks) and non-synonymous (Ka) substitution rate analysis

The complete chloroplast genome sequence of *S. adstringens* was compared with the chloroplast genome sequences of the five Mimosoid chloroplast genomes used in the assembling. To assess the complete nucleotide diversity (Pi) among the complete chloroplast genome of the six species, the complete chloroplast genome sequences were aligned using MAFFT aligner tool³⁷, and manually adjusted with Bioedit³⁸. We then performed a sliding window analysis to calculate the nucleotide variability (Pi) values using DnaSP 6³⁹ with window length of 600 bp and step size of 200 bp. The 77 protein-coding genes were extracted and aligned separately using MAFFT to estimate the synonymous (Ks) and non-synonymous (Ka) substitution rates. The Ka/Ks for each gene were estimated in DnaSP 6.

2.2.5 Comparative analysis of genome structure

The mVISTA program was applied to compare the complete chloroplast genome of *S. adstringens* against the whole chloroplast genome of the five mimosoid species using the shuffle-LAGAN mode⁴⁰. The annotated *S. adstringens* chloroplast genome was used as reference. The expansion and contraction of the IR regions at junction sites between the six mimosoid species were verified and plotted using IRscope⁴¹.

2.2.6 Phylogenetic analyses

Seventy-three protein-coding genes were recorded from 19 species within the Leguminosae – Caesalpinioideae, as well as from two outgroups (*Cucumis sativus* L. and *Fragaria vesca* L.). All genes sequences were obtained from GenBank (see Table S1 for accession numbers).

The nucleotide sequences were aligned using MAFFT³⁷ with default parameters. The Akaike Information Criterion (AIC) in JModelTest v2.1.10 was used to determine the best-fitting model of molecular evolution for each gene⁴² (models selected can be seen in Table S2). The alignments from the 73 protein-coding genes were concatenated and a Bayesian inference was performed using BEAST v1.10.1⁴³ at the XSEDE Teragrid of the CIPRES science gateway⁴⁴ (available online: www.phylo.org). The Markov chain Monte Carlo (MCMC) was set to run 50.000.000 generations and sampled every 1.000 generations, under a strict clock approach using the Yule speciation tree prior with the evolutionary models. The Convergence of parameters during MCMC runs were assessed by their Effective Sample Size (ESS) > 200 using TRACER v1.7.1⁴⁵. The phylogenetic tree was annotated as a maximum clade credibility tree using TREEANNOTATOR v.2.5.0 (part of the BEAST package), with burn in of 20%. The final tree was produced using FigTree v.1.4.3 (<http://tree.bio.ed.ac.uk/software/figtree/>).

2.3 RESULTS AND DISCUSSION

2.3.1 Genome Assembly and Annotation

A total of 563.117.260 raw Illumina paired-end reads from the *S. adstringens* genome were generated and filtered against Mimosoid chloroplast genomes. The 10.549.708

obtained reads were used to assemble the chloroplast genome with a high mean estimated coverage of 1.033x.

The complete chloroplast genome of *S. adstringens* was assembled with a total of 162.169 bp in size, divided in four regions, which included a large single-copy (LSC) region of 91.045 bp, a small single-copy (SSC) region of 19.014 bp, separated by two inverted repeat (IR) regions of 26.055 bp each (Figure 1; Table 1). Analogous to most angiosperms, the *S. adstringens* chloroplast genome comprises a single circular molecule with a quadripartite structure⁴ and is similar in size from other species from the Mimosoid clade (Leguminosae; Table 1; see also⁷).

Table 1. Chloroplast genome information from sampled Mimosoid species and the newly assembled *Stryphnodendron adstringens*.

Species	Size (bp)	LSC (bp)	SSC (bp)	IR (bp)	GC (%)	Protein	RNA
<i>Adenanthera microsperma</i>	159.389	88.577	18.756	26.028	36.5%	77	34
<i>Dichrostachys cinerea</i>	161.240	90.430	18.526	26.142	35.9%	77	34
<i>Leucaena trichandra</i>	164.692	93.690	18.890	26.056	35.6%	77	34
<i>Parkia javanica</i>	161.681	91.093	18.574	26.007	35.9%	77	34
<i>Piptadenia communis</i>	162.552	91.517	18.941	26.047	35.9%	77	34
<i>Stryphnodendron adstringens</i>	162.169	91.045	19.014	26.055	35,9%	77	34

LSC Large Single Copy, SSC Small Single Copy, IR Inverted Repeat.

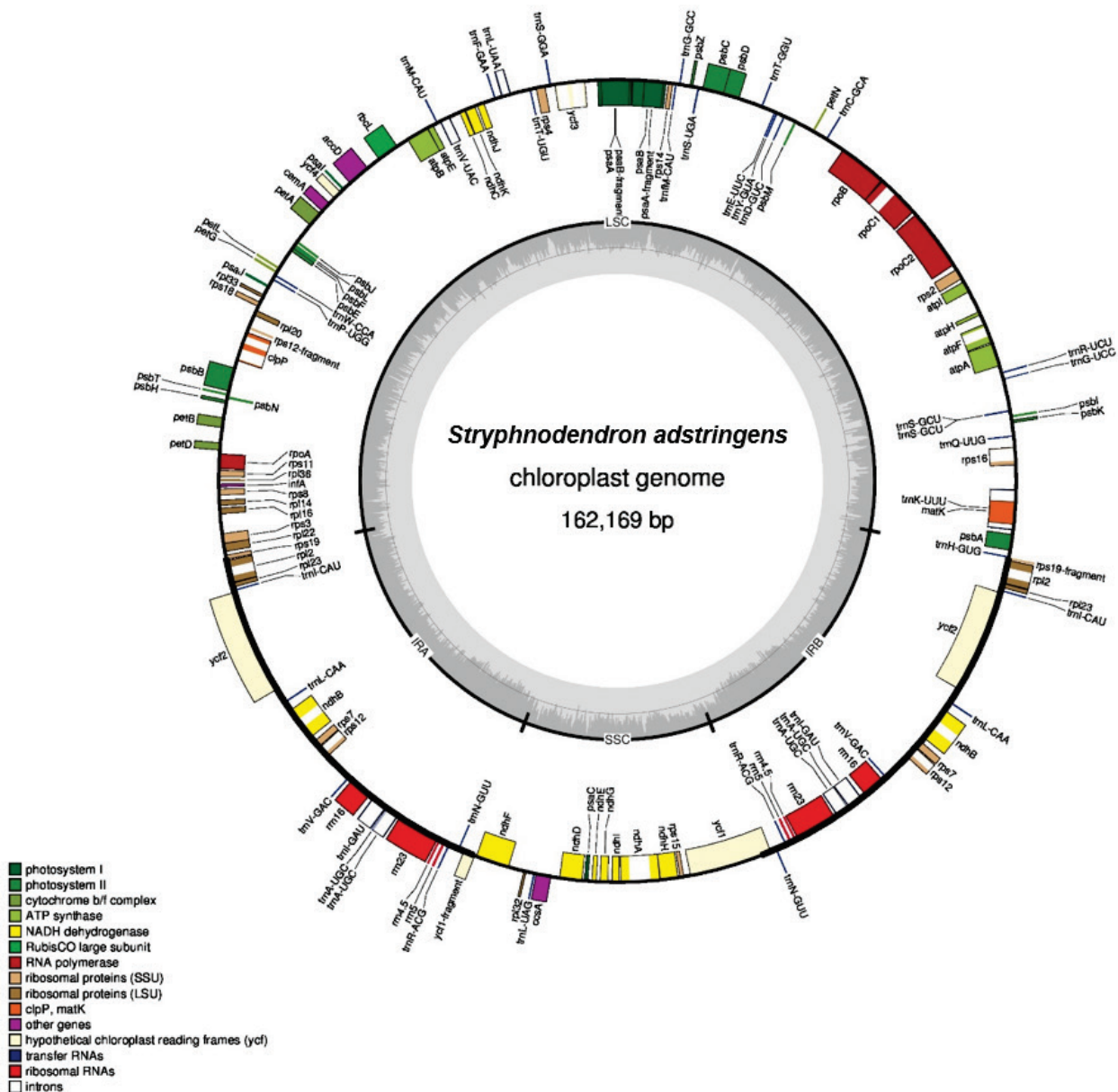


Figure 1. Gene map of the *Stryphnodendron adstringens* chloroplast genome. The genes drawn outside and inside of the circle are transcribed in clockwise and counterclockwise directions, respectively. Genes were colored based on their functional groups. The inner circle shows the quadripartite structure of the chloroplast: small single copy (SSC), large single copy (LSC) and a pair of inverted repeats (IRa and IRb). The gray ring marks the GC content with the inner circle marking a 50% threshold. Asterisks mark genes that have introns.

The GC content of the *S. adstringens* chloroplast genome was 35.9%, which is also consistent with other Mimosoid, whose plastomes ranged from 35.6 to 36.5% overall GC content (Table 1; see also⁷). Among the LSC, SSC and IR regions, the highest GC content was found in the IR regions (42.7%), while the GC content of LSC and SSC was 33.3% and 30.0%, respectively. The high GC content in the IR regions are mainly due to high GC

contents of the four ribosomal RNA (rRNA) genes, *rrn23*, *rrn16*, *rrn5*, *rrn4.5*, with 55.3%, 56.4%, 52.9% and 50%, respectively, that are located in this region.

The assembled chloroplast genome contained 111 different genes, with 77 protein-coding genes, 30 transfer RNA (tRNA) and 4 ribosomal RNA genes (rRNA) (Figure 1; Table 2). A total of 10 protein-coding genes and 6 tRNAs contained a single intron, whereas two genes (*clpP* and *ycf3*) exhibit two introns each. In one gene, *rps12*, a trans-splicing event was observed, with the 5' end located in the LSC region and the duplicated 3' end in the IR region. The *trnK-UUU* has the largest intron encompassing the *matK* gene, with 2.629 bp, whereas the intron of *trnL-UAA* is the smallest (514 bp).

The duplicated IR of the *S. adstringens* chloroplast genome resulted in complete duplication of sixteen genes (including five protein-coding genes [*rpl2*, *rpl23*, *rps7*, *rps12* and *ndhB*], seven tRNAs [*trnA-UGC*, *trnI-CAU*, *trnI-GAU*, *trnL-CAA*, *trnN-GUU*, *trnR-ACG*, *trnV-GAC*], and all four rRNAs [*rrn23*, *rrn16*, *rrn5*, *rrn4.5*], see Table 2; Figure 1) and parts of the 5' end of *ycf1* and *rps19*. These results corroborate the findings of Wang et al. (2017), which all species belonging to tribe Mimoseae had canonical IRs, with a typical gene content and general organization⁷.

It is important to point out that the structure of the chloroplast genome in species from Leguminosae family are highly variable because of either expansion or contraction of the IR^{7,46,47}. This is mainly caused by the gene transfer from single copy regions to inverted repeat and vice versa in the boundaries of IRs, during evolution⁴⁷. With the exception of *Acacia ligulata*⁴⁸, the legume plastomes of all species documented to date within the clade formed by *Ingeae* and *Acacia* species, have IRs ca. 13 kb larger, and a SSC correspondingly smaller, than other legumes^{7,46}. In addition, IRs do not seem to be essential and the size variation of the plastid genome has been explained, at least in part, by the loss of one IR. However, the mechanisms that led to IR loss are still unknown. Examples of variation within the Leguminosae include the loss of the IR in the monophyletic group within the subfamily Papilionoideae (including *Trifolium*, *Pisum*, *Cicer*, *Medicago*, *Glycyrrhiza* and *Vicia*), known as the “inverted repeat lacking clade”^{49,50}.

Table 2. List of genes in the chloroplast genome of *Stryphnodendron adstringens*.

Category	Gene groups	Name of genes
Self-replication	Large subunit of ribosomal proteins	rpl2 ^{1,2} , rpl14, rpl16 ¹ , rpl20, rpl23 ² , rpl32, rpl33, rpl36
	Small subunit of ribosomal proteins	rps2, rps3, rps4, rps7 ² , rps8, rps11, rps12 ^{1,2} , rps14, rps15, rps16 ¹ , rps18, rps19
	DNA-dependent RNA polymerase	rpoA, rpoB, rpoC1 ¹ , rpoC2
	Ribosomal RNA genes	rrn4.5 ² , rrn5 ² , rrn16 ² , rrn23 ²
	Transfer RNA genes	trnA-UGC ^{1,2} , trnC-GCA, trnD- GUC, trnE-UUC, trnF-GAA, trnM-CAU, trnG-UCC ¹ , trnG- UCC, trnH-GUG, trnI-CAU ² , trnI- GAU ^{1,2} , trnK-UUU ¹ , trnL-CAA ² , trnL-UAA ¹ , trnL-UAG, trnM-CAU, trnN-GUU ² , trnP-UGG, trnQ- UUG, trnR-ACG ² , trnR-UCU, trnS- GCU, trnS-UGA, trnS-GGA, trnT- UGU, trnT-GGU, trnV-UAC ¹ , trnV-GAC ² , trnW-CCA, trnY-GUA
Photosynthesis	Photosystem I	psaA, psaB, psaC, psaI, psaJ
	Photosystem II	psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ
NADH dehydrogenase	NADH dehydrogenase	ndhA ¹ , ndhB ^{1,2} , ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK
	Cytochrome b/f complex	petA, petB ¹ , petD ¹ , petG, petL, petN
	ATP synthase	atpA, atpB, atpE, atpF ¹ , atpH, atpI
	RubisCo large subunit	rbcL
Other genes	Maturase K	matK
	Envelope membrane protein	cemA

	Subunit of acetyl-CoA carboxylase	accD
	C-type cytochrome synthesis gene	ccsA
	Protease	clpP ¹
	Conserved hypothetical chloroplast open reading frames	ycf1, ycf2 ² , ycf3 ¹ , ycf4

1 – Gene with introns

2 – Gene completely duplicated in the inverted repeat

2.3.2 Repeat sequences analysis

A total of 42 repeats structures, with lengths ranging from 30 bp to 128 bp, were detected in the *S. adstringens* chloroplast genome. They included 22 (52.38%) forward repeats, 18 (42.86%) palindromic repeats, and two (4.76%) reverse repeats (Table S3), whereas none complementary structure was identified. The forward repeats ranged from 30 bp to 128 bp. The palindromic repeats were 30 bp to 60 bp, whereas the two reverse repeats were 30 bp and 38 bp (Table S3). In the majority of Mimosoid species, the most abundant dispersed repeat identified were forward, then palindromic and the least reverse⁷.

Among the 42 repeats, 64.29% are located in the LSC region, 14.29% in the SSC region and 16.67% in the IR region. Two repeats, which were 39 bp and 52 bp, located in the intron region of *ycf3* and *rpl16*, was found to repeat thrice (LSC/SSC/IR) and twice (LSC/SSC), respectively, as forward repeat. Most of the repeats (80.95%) were found in the intergenic spacer regions (IGS), whereas 19,05% were located in the introns (*ndhA*, *ycf3* and *rpl16*) and only two were located in coding region (*psaB* and *trnS-GCU*) (Table S3).

A total of 137 SSRs, with a length of at least 10 bp and repeated 3 to 21 times, were found in the *S. adstringens* chloroplast genome. Among them, 90 (65.69%) were mono-repeats, 19 (13.87%) were di-repeats, nine (6.57%) were tri-repeats, 11 (8.03%) were tetra-repeats, seven (5.11%) were penta-repeats and one (0.73%) was hexa-repeat. The majority (89 or 98.89%) of the mononucleotide repeats consisted of A/T motifs, while only one was composed of a G/C motif. Likewise, most of the dinucleotides and trinucleotides were AT-rich, being composed of AT/TA (84.21%) and AAT/TTA (77.78%) repeats, respectively (Figure 2). These results showed that the SSRs exhibit a strong AT bias, which is consistent with the observed in other Leguminosae species, such as *Vigna radiata*⁵¹, *Cajanus cajan*¹⁰, *Cajanus scarabaeoides*¹⁰ and *Arachis hypogaea*⁵².

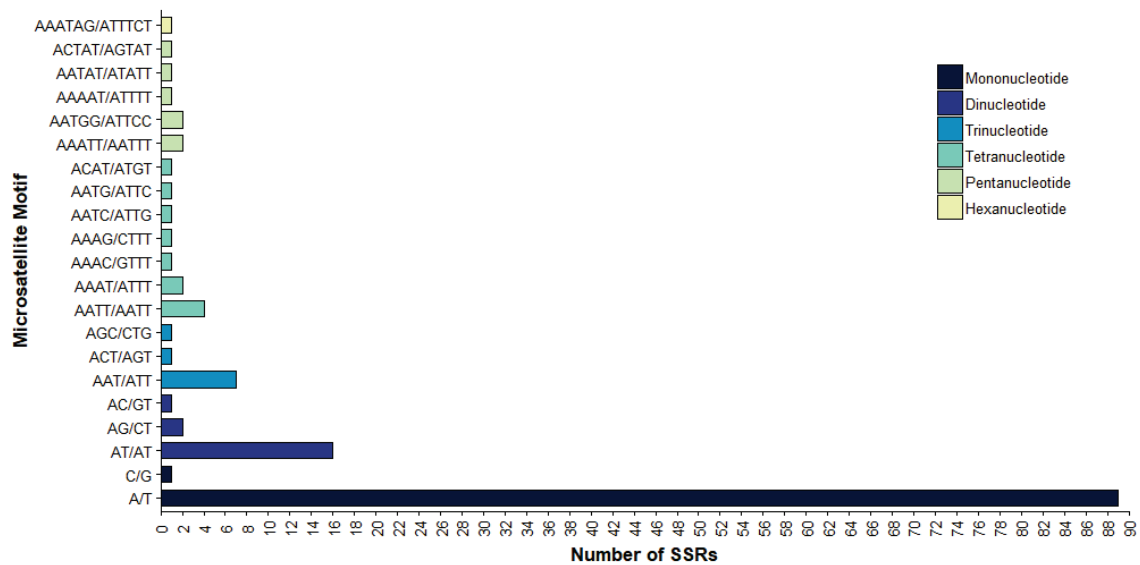


Figure 2. Number and type of simple sequence repeats in *Stryphnodendron adstringens* chloroplast genome.

Also, among the 137 SSRs, 114 (83.21%) were found in the LSC region, 17 (12.41%) in the SSC region and six (4.38%) in the IR region (Table S4). Most of these SSRs were located in intergenic regions (102 or 74.45%), while 17 (12.41%) were in the introns and 18 (13.14%) were in the protein-coding genes (Table S4). The *ycf1* gene contained more SSRs than the other genes (Table S4). In addition, the results found herein are in agreement with those from *Cajanus cajan*¹⁰, *Cajanus scarabaeoides*¹⁰, *Vigna radiata*⁵¹ and *Glycine* species⁵³. The *ycf1* encodes a protein of approximately 1.800 amino acids. The *ycf1* located in the IR region is short and conserved, while the located in SSC region are extremely variable in seed plants^{54,55}. Some studies reported that this region is the most variable locus for the design of primers, as well, as more variable than *matK* in many taxa, and thus suitable for molecular systematics at low taxonomic levels^{54,56}.

2.3.3 Nucleotide Diversity and Ka/Ks Ratio

The average nucleotide variability (P_i) among the five chloroplast genomes of Mimosoid species was estimated to be 0.01771, ranging from 0 to 0.172. The nucleotide variability was higher in the SSC ($P_i = 0.02712$) and LSC ($P_i = 0.02488$), when compared to IR regions, which had a much lower nucleotide diversity ($P_i = 0.00339$). Similar results were obtained in the comparison of chloroplast genome sequences among *Aconitum* L.

species, in which the average Pi in the IR region was 0.00146, whereas in the LSC and SSC regions the diversity estimates were 0.007140 and 0.008368, respectively⁵⁷. Park et al. (2018) analyzed the nucleotide diversity in six *Ipomoea* L. chloroplast genome and also found that the IR regions were more conserved than the LSC and SSC regions, with average Pi values of 0.003 for IR and more than 0.006 for SSC and LSC regions⁵⁸.

Five regions (*trnS-GCU-trnG-UCC*, *trnR-UCU-atpA*, *trnC-GCA-petN*, *psbZ-trnG-UCC*, *ndhC-trnV-UAC*) showed high levels of nucleotide diversity, with Pi values > 0.8 (Figure 3). All of these highly variable regions are found in intergenic spacer from LSC region. Liu et al. (2018) analyzed the nucleotide diversity among seven species from caesalpinoid legumes and reported five regions including *psbZ-trnG* (*trnT-trnL*, *rps3-rps19*, *rpl32* and *ycf1*) with higher Pi values, all with $Pi > 0.12$ ⁵⁹. Highly variable regions among chloroplast genomes can be useful for phylogenetic reconstruction and may be used for further phylogenetic study of the genus *Stryphnodendron*.

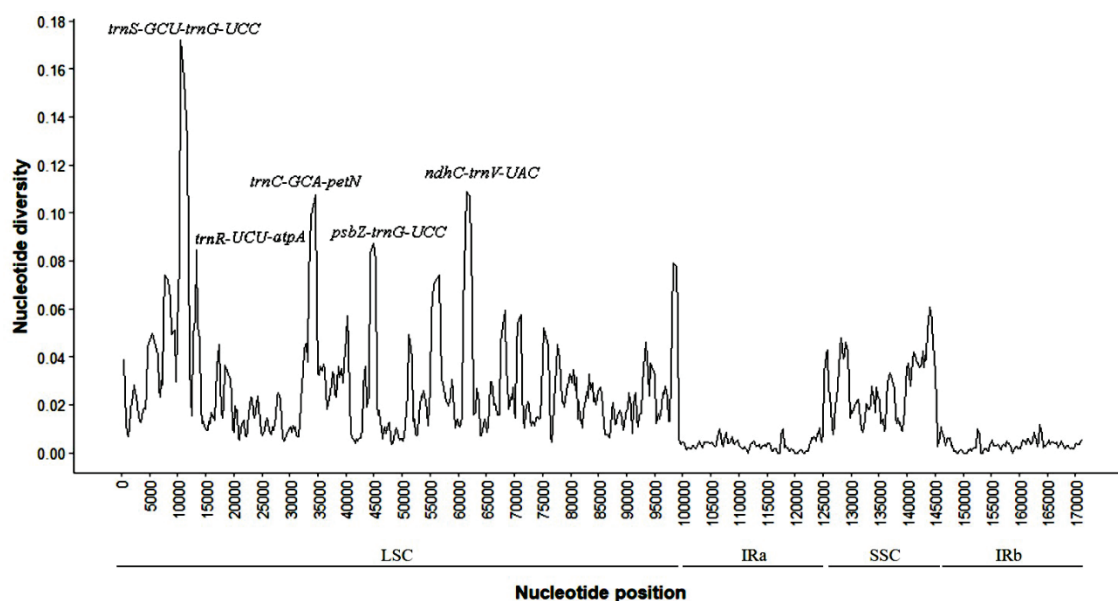


Figure 3. Sliding window analysis of Mimosoid chloroplast genome.

The non-synonymous (Ka) to synonymous (Ks) rate ratio (Ka/Ks) was calculated for the 77 protein-coding genes in common across all six chloroplast genomes (Figure 4 and Table 3). The synonymous substitution does not change the amino acid within a peptide chain, whereas nonsynonymous substitution does. The *rpl32* gene associated with large

subunit of ribosomal proteins had the highest synonymous rate, 0.09263, while the *ycf1* gene with unknown functions, had the highest nonsynonymous rate, 0.03534.

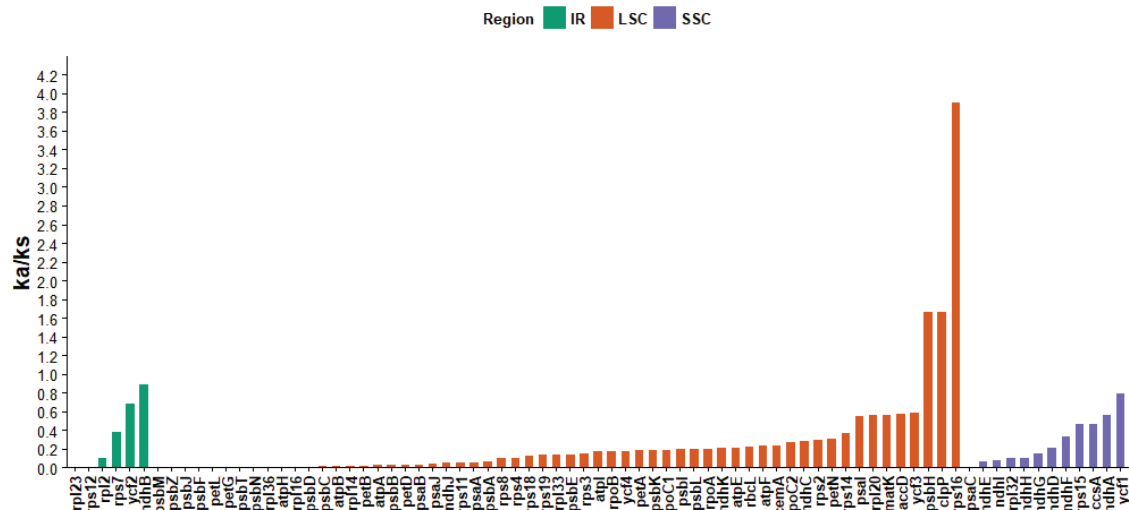


Figure 4. The Ka/Ks ratio of Mimosoid chloroplast genomes for individual genes.

The Ka/Ks ratio can indicate whether selective pressure is acting on a particular protein-coding gene. A Ka/Ks > 1 indicates that the gene is affected by positive selection, whereas Ka/Ks < 1 indicates that the gene is affected by negative selection or purifying selection. A value of 0, indicates the presence of neutral selection⁶⁰. Concerning the different regions of chloroplast genomes, the Ka/Ks ratio were highest on average in the SSC region (0.2991) and lowest in the IR region (0.2856) and LSC region (0.2708). The lowest Ka/Ks ratio was observed for genes encoding subunits of ATP synthase, subunits of the cytochrome b/f complex, subunits of the large subunit of ribosomal proteins and subunits of photosystem II (Figure 4; Table 3).

Herein, the Ka/Ks ratio was calculated to be 0 for 13 genes, two inside the IR region (*rpl23*, *rps12*), ten in the LSC region (*rpl36*, *psbM*, *psbZ*, *psbJ*, *psbF*, *psbT*, *psbN*, *petL*, *petG*, *atpH*) and one in the SSC region (*psaC*) (Figure 4). However, this occurred because the Ka or Ks is 0 or extremely low, thus Ka/Ks ratio could not be calculated^{61,62}. Among the 77 protein-coding genes, in 62 of them the Ka/Ks indicates purifying selection (Figure 4; Table 3). The Ka/Ks ration indicates positive selection for three genes analyzed, one of it is associated with small subunit of ribosomal proteins (*rps16*), the other is associated with Photosystem II (*psbH*) and the third with the *clpP* proteases (Figure 4; Table 3). Liu et al. (2018) reported four genes with Ka/Ks ratio more than 1, indicating positive selection, *ndhD*,

ycf1, *infA* and *rpl23* in caesalpinoid legumes⁵⁹, whereas, Park et al. (2018) observed positive selection in three genes, *accD*, *cemA*, and *ycf2*, among six *Ipomoea* species⁵⁸. In addition, Tian et al. (2018) reported only one gene (*rps12*) with positive selection among nine *Araceae* species⁶³.

Table 3. The *Ka*, *Ks* and *Ka/Ks* ratio of Mimosoid chloroplast genome for individual genes and region.

Gene groups	Gene	Ka	Ks	Ka/Ks	Region
Large subunit of ribosomal proteins	rpl33	0.00652	0.0445	0.146517	LSC
	rpl20	0.01379	0.02411	0.571962	LSC
	rpl36	0.00000	0.03904	0.00000	LSC
	rpl14	0.00120	0.04091	0.029333	LSC
	rpl16	0.00111	0.06961	0.015946	LSC
	rpl2	0.00108	0.00989	0.109201	IR
	rpl23	0.00156	0.00000	0.00000	IR
	rpl32	0.01046	0.09263	0.112922	SSC
Small subunit of ribosomal proteins	rps16	0.02246	0.00574	3.912892	LSC
	rps2	0.00486	0.01590	0.305660	LSC
	rps14	0.00577	0.01520	0.379605	LSC
	rps4	0.00291	0.02525	0.115248	LSC
	rps18	0.00652	0.04801	0.135805	LSC
	rps11	0.00217	0.03277	0.066219	LSC
	rps8	0.00554	0.05103	0.108564	LSC
	rps3	0.00567	0.03538	0.16026	LSC
	rps19	0.00598	0.04149	0.144131	LSC
	rps7	0.00343	0.00889	0.385827	IR
	rps12	0.00000	0.01458	0.000000	IR
	rps15	0.01777	0.0375	0.473867	SSC
DNA-dependent RNA polymerase	rpoC2	0.00912	0.03179	0.286883	LSC
	rpoC1	0.00497	0.02452	0.202692	LSC
	rpoB	0.00423	0.02317	0.182564	LSC
	rpoA	0.00903	0.04328	0.208641	LSC
Photosystem I	psaB	0.00138	0.02978	0.046340	LSC

	psaA	0.00175	0.0248	0.070565	LSC
	psaI	0.02987	0.05379	0.555308	LSC
	psaJ	0.00338	0.05941	0.056893	LSC
	psaC	0.00000	0.05761	0.000000	SSC
Photosystem II	psbA	0.00231	0.02972	0.077725	LSC
	psbK	0.00980	0.04920	0.199187	LSC
	psbI	0.00416	0.02023	0.205635	LSC
	psbM	0.00000	0.04284	0.000000	LSC
	psbD	0.00041	0.02315	0.017711	LSC
	psbC	0.00062	0.02620	0.023664	LSC
	psbZ	0.00000	0.02609	0.000000	LSC
	psbJ	0.00000	0.00000	0.000000	LSC
	psbL	0.00416	0.02023	0.205635	LSC
	psbF	0.00000	0.01118	0.000000	LSC
	psbE	0.00175	0.01173	0.149190	LSC
	psbB	0.00112	0.02898	0.038647	LSC
	psbT	0.01757	0.00000	0.000000	LSC
	psbN	0.00000	0.01047	0.000000	LSC
	psbH	0.02040	0.01214	1.680395	LSC
NADH dehydrogenase	ndhJ	0.00182	0.02749	0.066206	LSC
	ndhK	0.00641	0.02938	0.218176	LSC
	ndhC	0.00928	0.03154	0.294230	LSC
	ndhB	0.00162	0.00179	0.905028	IR
	ndhF	0.01270	0.03726	0.340848	SSC
	ndhD	0.00725	0.03343	0.216871	SSC
	ndhE	0.00419	0.05561	0.075346	SSC
	ndhG	0.00721	0.04555	0.158288	SSC
	ndhI	0.00324	0.03583	0.090427	SSC
	ndhA	0.00963	0.01685	0.571513	SSC
	ndhH	0.00388	0.03389	0.114488	SSC
Cytochrome b/f complex	petN	0.00512	0.01621	0.315854	LSC
	petA	0.00614	0.03184	0.192839	LSC
	petL	0.00000	0.03487	0.000000	LSC

	petG	0.00000	0.01184	0.000000	LSC
	petB	0.00137	0.04232	0.032372	LSC
	petD	0.00092	0.02175	0.042299	LSC
ATP synthase	atpA	0.00159	0.0428	0.037150	LSC
	atpF	0.00675	0.02774	0.243331	LSC
	atpH	0.00000	0.02566	0.000000	LSC
	atpI	0.00370	0.02032	0.182087	LSC
	atpE	0.00530	0.02428	0.218287	LSC
	atpB	0.00089	0.03743	0.023778	LSC
RubisCo large subunit	rbcL	0.01105	0.04678	0.236212	LSC
Maturase K	matK	0.01494	0.02608	0.572853	LSC
Envelope membrane protein	cemA	0.00561	0.02283	0.245729	LSC
Subunit of acetyl-CoA carboxylase	accD	0.01713	0.02936	0.583447	LSC
C-type cytochrome synthesis gene	ccsA	0.01181	0.02449	0.482238	SSC
Protease	clpP	0.02109	0.01255	1.680478	LSC
	ycf3	0.00522	0.00869	0.600690	LSC
Conserved hypothetical	ycf4	0.00380	0.02053	0.185095	LSC
chloroplast open reading frames	ycf2	0.00383	0.00551	0.695100	IR
	ycf1	0.03534	0.04383	0.806297	SSC

Ka non-synonymous substitution, *Ks* synonymous, LSC Large Single Copy, SSC Small Single Copy, IR Inverted Repeat.

2.3.4 Comparative analysis of genome structure

The structural characteristics in chloroplast genome among Mimosoid species revealed that gene coding regions were more conserved than the noncoding regions, and IRs were more conserved than LSC and SSC regions (Figure 5). This result is consistent with the pattern revealed in other Leguminosae species, such as *Glycine* species⁶⁴ and species from Caesalpinioideae subfamily⁷. Additionally, it was also observed that the intergenic spacers regions between several pairs of genes varied greatly, for example, between *matK-rps16*, *rps16-psbK*, *trnS-GCU-trnG-UCC*, *atpH-atpI*, *trnC-GCA-petN*, *psbZ-trnG*, *trnT-*

UGU-*trnL*-UAA, *ndhC-trnV*-UAC and *rps3-rps19*. Some of these intergenic spacers regions had also the highest level of nucleotide diversity.

The expansion and contraction of the IR region and the single-copy boundary regions result in change in the position of the junction sites, which is considered as a primarily mechanism causing length variation of chloroplast genomes in higher plants⁶⁵. The length of the IR regions was similar, ranging from 26.006 bp in *Parkia javanica* to 26.143 bp in *Dichrostachys cinerea* (Figure 6). Wang et al. (2017) observed that species belonging to tribe Mimoseae had canonical IRs, whereas species belonging to tribes Ingae and Acacieae had much long IRs, as all of them experienced ca. 13 kb IR expansion into SSC previously reported by Dugas et al. (2015).

The endpoint of the Mimoseae JLA (Junction between IRa and LSC) was located upstream of the *rps19* and downstream of the *trnH*-GUG. Similar patterns were observed by Amiryousefi et al. (2018b) in Solanaceae chloroplast genomes⁶⁶. The junction between IRb and SSC region (JSB) was located in the intergenic *ycf1/ndhF*, and the distance between the *ndhF* end to the junction of the IRb/SSC differs by 11 bp in *Dichrostachys cinerea* to 150 bp in *Adenanthera microsperma*. The junction between IRa and SSC (JSA) was located within the *ycf1* gene, and the fragment located at the IRa region ranged from 692 bp to 779 bp (Figure 6). The gene *rps19* crossed the LSC/IRb region and the extent of the IR expansion to *rps19* slightly varies among the Mimoseae species ranging from 101 bp to 105 bp.

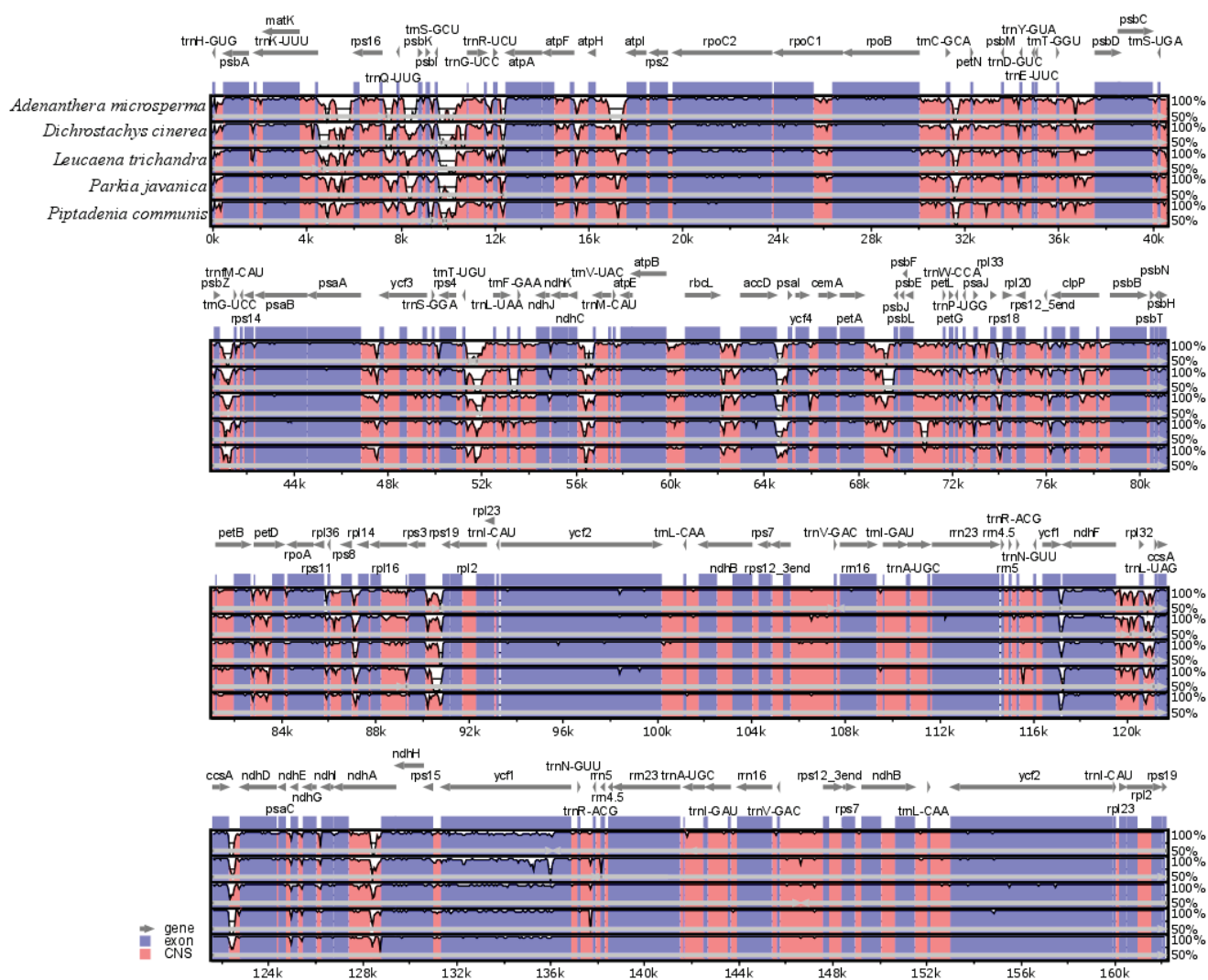


Figure 5. Visualization of genome alignment of six Mimoid chloroplast genomes using *Stryphnodendron adstringens* as reference. The vertical scale indicates the percent of identity, ranging from 50% to 100%. Coding regions are marked in purple, and non-coding as red. The horizontal axis indicates the coordinates within the chloroplast genome.

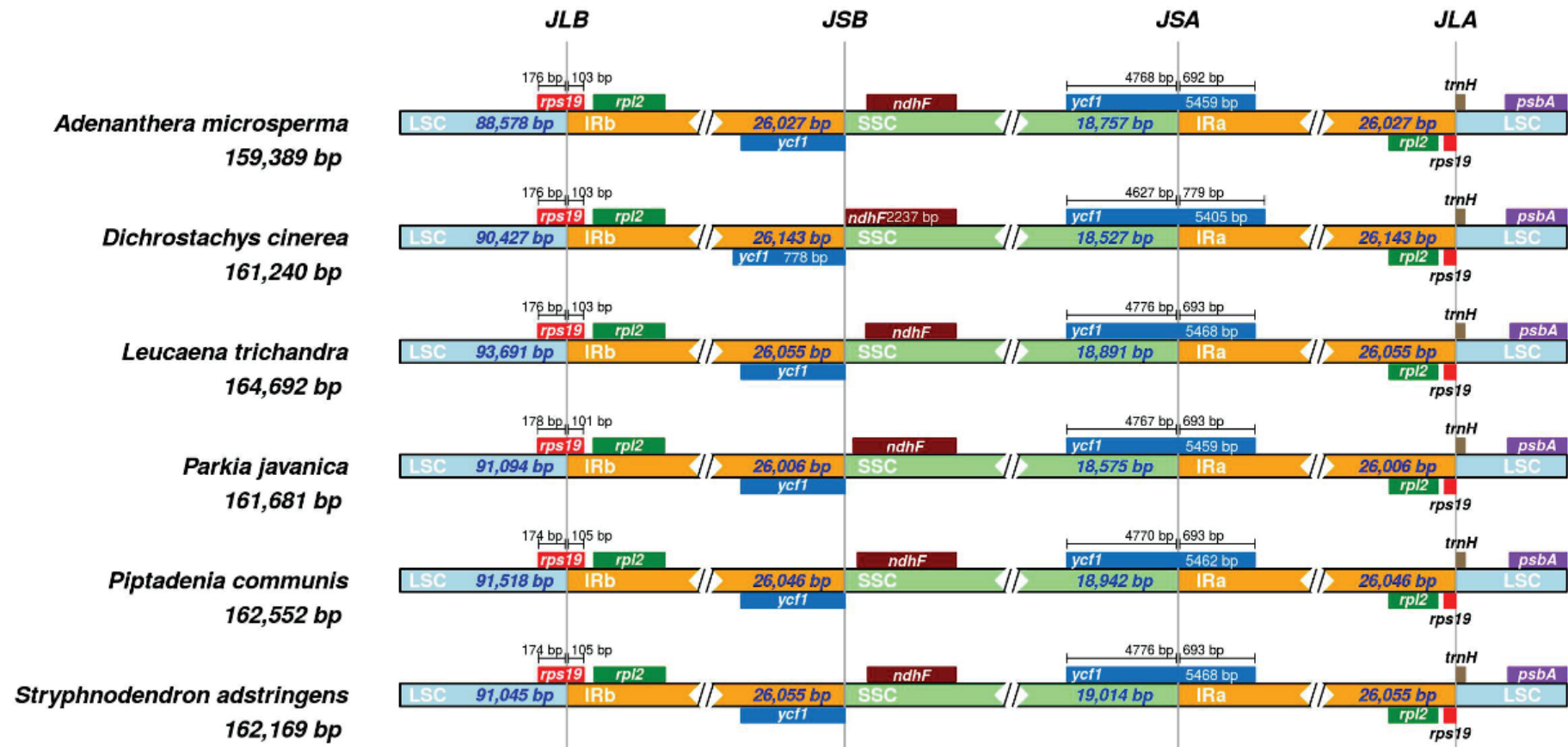


Figure 6. Comparison of the junction sites between the Long Single Copy (LSC, light blue), Short Single Copy (SSC, light green) and Inverted Repeat (IRa and IRb, orange) regions among the six Mimosoid chloroplast genomes. JLB (IRb/LSC), JSB (IRb/SSC) JSA (SSC/IRa) and JLA (IRa/LSC) denote the junction sites between each corresponding regions on the genome.

2.3.5 Phylogenetic relationships

In this study, 19 species from Caesalpinioideae and two outgroups were analysed, based on 73 protein-coding genes in their chloroplast genomes. The total concatenated alignment length from the 73 protein-coding genes was 60,736 bp. The reconstructed phylogeny indicated that Caesalpinioideae was paraphyletic and that the species from tribe Mimoseae (*Adenantha microsperma*, *Dichrostachys cinerea*, *Leucaena trichandra*, *Parkia javanica*, *Piptadenia communis* and *Stryphnodendron adstringens*) were deemed non-monophyletic (Figure 7). All nodes were strongly supported, given the Bayesian posterior probability (Figure 7). These results are consistent with those from Wang et al. (2017) and support the new classification system proposed for the Leguminosae²¹.

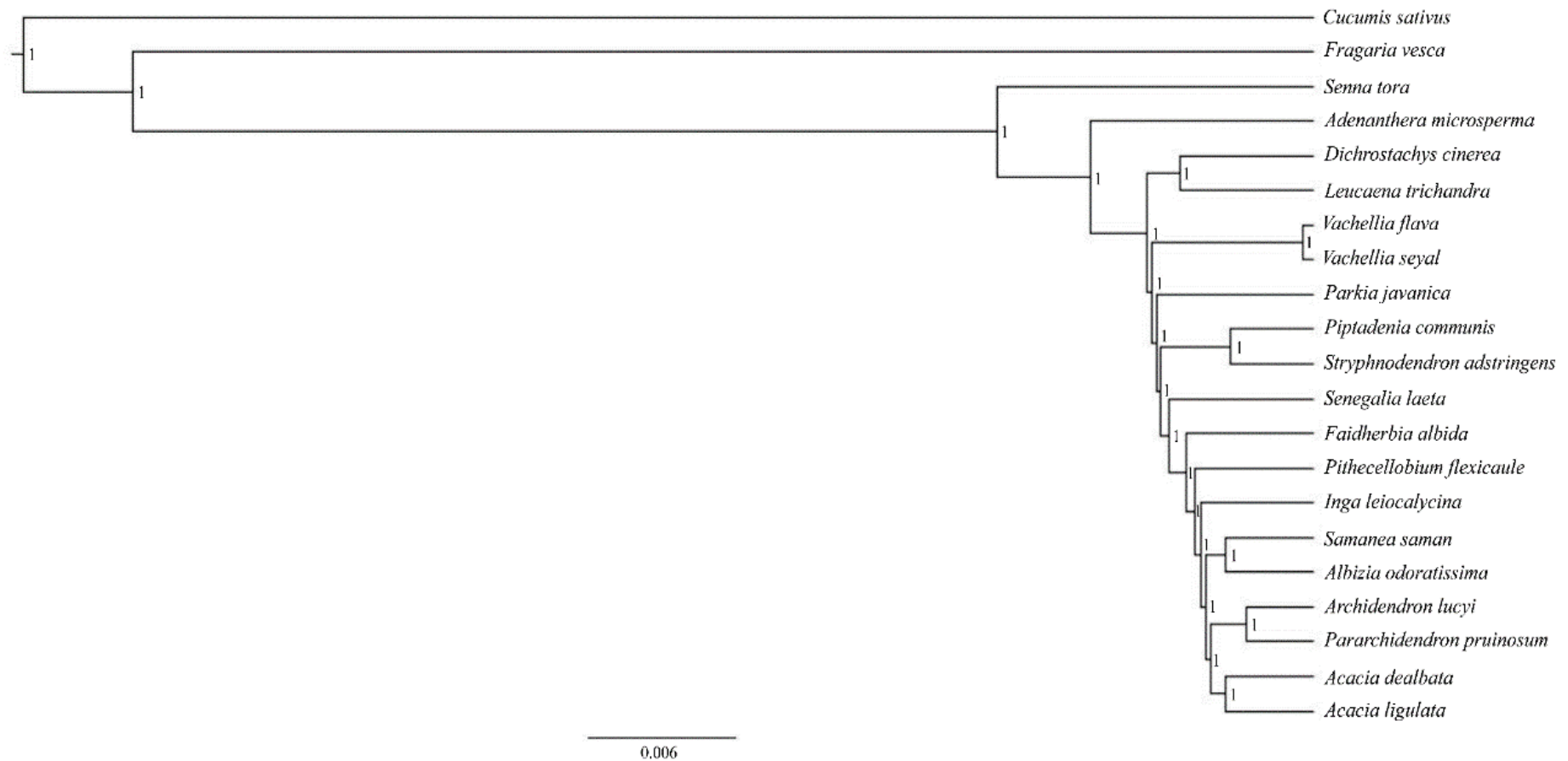


Figure 7. Maximum credibility tree reconstructed based on 73 conserved coding-protein genes from twenty-one species. All nodes of the phylogenetic tree are supported by 1.00 Bayesian inference posterior probability.

2.4 CONCLUSION

In this work, we assemble the complete chloroplast genome of *S. adstringens* with 162,169 bp. Genome gene contents and orientation are similar to those found in the chloroplast genome of other Mimosoid (Leguminosae) species. This study also revealed the distribution and location of repeated structures and microsatellites along the chloroplast genome of *S. adstringens*. We also generated important genomic resources for Mimosoid group. Moreover, the Ka/Ks ratio was lower in the LSC region compared to SSC region. As expected, the comparison with other five Mimosoid species revealed that the coding regions are more conserved than non-coding regions, and IRs more conserved than LSC and SSC regions. Finally, the phylogenetic relationships built for 19 species of Caesalpinioideae, including the new data from *S. adstringens* and two outgroups, were fully resolved with high supports based on 73 conserved protein-coding genes. The maximum credibility tree revealed that the tribe Mimoseae is paraphyletic, consistent with the new classification proposed for the Leguminosae.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA ARCHIVING STATEMENT

Data will be deposited in the National Center for Biotechnology Information (NCBI) (<https://www.ncbi.nlm.nih.gov/>) after acceptance for review and the accession numbers will be supplied once available.

2.5 REFERENCES

1. Leister, D. Chloroplast research in the genomic age. *Trends in Genetics* **19**, 47–56 (2003).
2. Neuhaus, H. E. & Emes, M. J. Nonphotosynthetic Metabolism in Plastids. *Annual Review of Plant Physiology and Plant Molecular Biology* **51**, 111–140 (2000).
3. Rodríguez-Ezpeleta, N. *et al.* Monophyly of Primary Photosynthetic Eukaryotes: Green Plants, Red Algae, and Glaucophytes. *Current Biology* **15**, 1325–1330 (2005).
4. Jansen, R. K. & Ruhlman, T. A. in *Genomics of Chloroplasts and Mitochondria* 103–126 (Springer, Dordrecht, 2012). doi:10.1007/978-94-007-2920-9_5
5. Palmer, J. D. Comparative Organization of Chloroplast Genomes. *Annual Review of Genetics* **19**, 325–354 (1985).
6. Blazier, J. C., Guisinger, M. M. & Jansen, R. K. Recent loss of plastid-encoded *ndh* genes within *Erodium* (Geraniaceae). *Plant Molecular Biology* **76**, 263–272 (2011).
7. Wang, Y.-H., Qu, X.-J., Chen, S.-Y., Li, D.-Z. & Yi, T.-S. Plastomes of Mimosoideae: structural and size variation, sequence divergence, and phylogenetic implication. *Tree Genetics & Genomes* **13**, 41 (2017).
8. Raubeson, L. A. & Jansen, R. K. in *Plant diversity and evolution: genotypic and phenotypic variation in higher plants*. (ed. Henry, R.) 45–68 (Cambridge, MA: CABI, 2005).
9. Bock, R. in *Cell and Molecular Biology of Plastids* (ed. Bock, R.) 29–63 (Springer-Verlag Berlin Heidelberg, 2007).
10. Kaila, T. *et al.* Chloroplast Genome Sequence of Pigeonpea (*Cajanus cajan* (L.) Millspaugh) and *Cajanus scarabaeoides* (L.) Thouars: Genome Organization and Comparison with Other Legumes. *Frontiers in Plant Science* **7**, 1847 (2016).
11. Li, B. & Zheng, Y. Dynamic evolution and phylogenomic analysis of the chloroplast genome in Schisandraceae. *Scientific Reports* **8**, 9285 (2018).
12. Wolfe, K. H., Li, W. H. & Sharp, P. M. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proceedings of the National Academy of Sciences* **84**, 9054–9058 (1987).
13. Provan, J., Powell, W. & Hollingsworth, P. M. Chloroplast microsatellites: new tools for studies in plant ecology and evolution. *Trends in Ecology & Evolution*

- 16**, 142–147 (2001).
14. Ravi, V., Khurana, J. P., Tyagi, A. K. & Khurana, P. An update on chloroplast genomes. *Plant Systematics and Evolution* **271**, 101–122 (2008).
 15. Shinozaki, K. *et al.* The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. *The EMBO Journal* **5**, 2043–2049 (1986).
 16. Ohyama, K. *et al.* Chloroplast gene organization deduced from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA. *Nature* **322**, 572–574 (1986).
 17. Nock, C. J. *et al.* Chloroplast genome sequences from total DNA for plant identification. *Plant Biotechnology Journal* **9**, 328–333 (2011).
 18. Twyford, A. D. & Ness, R. W. Strategies for complete plastid genome sequencing. *Molecular Ecology Resources* **17**, 858–868 (2017).
 19. Lewis, G. P., Schrire, B. D., Mackinder, B. & Lock, M. Legumes of the world. *Royal Botanic Gardens, Kew*, 577p (2005).
 20. The Legume Phylogeny Working Group. Legume phylogeny and classification in the 21st century: Progress, prospects and lessons for other species-rich clades. *Taxon* **62**, 217–248 (2013).
 21. Azani, N. *et al.* A new subfamily classification of the Leguminosae based on a taxonomically comprehensive phylogeny – The Legume Phylogeny Working Group (LPWG). *Taxon* **66**, 44–77 (2017).
 22. Souza, V. C. & Gibau, A. *Stryphnodendron* in Flora do Brasil 2020 em construção. *Jardim Botânico do Rio de Janeiro* (2018). Available at: <http://reflora.jbrj.gov.br/reflora/floradobrasil/FB19133>. (Accessed: 14th July 2018)
 23. Audi, E. A. *et al.* Gastric antiulcerogenic effects of *Stryphnodendron adstringens* in rats. *Phytotherapy Research* **13**, 264–266 (1999).
 24. Ishida, K. *et al.* Influence of tannins from *Stryphnodendron adstringens* on growth and virulence factors of *Candida albicans*. *Journal of Antimicrobial Chemotherapy* **58**, 942–949 (2006).
 25. Lima, J. C. S., Martins, D. T. O. & de Souza, P. T. Experimental evaluation of stem bark of *Stryphnodendron adstringens* (Mart.) Coville for antiinflammatory activity. *Phytotherapy Research* **12**, 218–220 (1998).
 26. Luiz, R. L. F. *et al.* Proanthocyanidins polymeric tannin from *Stryphnodendron*

- adstringens* are active against *Candida albicans* biofilms. *BMC Complementary and Alternative Medicine* **15**, 68 (2015).
27. Doyle, J. J. & Doyle, J. L. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* **19**, 11–15 (1987).
 28. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
 29. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357–359 (2012).
 30. Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology* **19**, 455–477 (2012).
 31. Ratan, A. Assembly algorithms for next-generation sequence data. (The Pennsylvania State University, University Park, PA, USA, 2009).
 32. McKain, M. R., Hartsock, R. H., Wohl, M. M. & Kellogg, E. A. Verdant: automated annotation, alignment and phylogenetic analysis of whole chloroplast genomes. *Bioinformatics* **33**, 130–132 (2017).
 33. Wyman, S. K., Jansen, R. K. & Boore, J. L. Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* **20**, 3252–3255 (2004).
 34. Lowe, T. M. & Chan, P. P. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Research* **44**, W54–W57 (2016).
 35. Lohse, M., Drechsel, O. & Bock, R. OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. *Current Genetics* **52**, 267–274 (2007).
 36. Kurtz, S. REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Research* **29**, 4633–4642 (2001).
 37. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* **30**, 772–780 (2013).
 38. Hall, T. H. BioEdit: a user friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp* **41**, 95–98 (1999).
 39. Rozas, J. *et al.* DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. *Molecular Biology and Evolution* **34**, 3299–3302 (2017).
 40. Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M. & Dubchak, I. VISTA: computational tools for comparative genomics. *Nucleic Acids Research* **32**, W273–

- W279 (2004).
41. Amiryousefi, A., Hyvönen, J. & Poczai, P. IRscope: an online program to visualize the junction sites of chloroplast genomes. *Bioinformatics* **34**, 3030–3031 (2018).
 42. Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods* **9**, 772–772 (2012).
 43. Suchard, M. A. *et al.* Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution* **4**, vey016 (2018).
 44. Miller, M. A., Pfeiffer, W. & Schwartz, T. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. in *Gateway Computing Environments Workshop (GCE)* 1–8 (2010).
 45. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Systematic Biology* **67**, 901–904 (2018).
 46. Dugas, D. V. *et al.* Mimosoid legume plastome evolution: IR expansion, tandem repeat expansions, and accelerated rate of evolution in clpP. *Scientific Reports* **5**, 16958 (2015).
 47. Zhu, A., Guo, W., Gupta, S., Fan, W. & Mower, J. P. Evolutionary dynamics of the plastid inverted repeat: the effects of expansion, contraction, and loss on substitution rates. *New Phytologist* **209**, 1747–1756 (2016).
 48. Williams, A. V., Boykin, L. M., Howell, K. A., Nevill, P. G. & Small, I. The Complete Sequence of the *Acacia ligulata* Chloroplast Genome Reveals a Highly Divergent clpP1 Gene. *PLOS ONE* **10**, e0125768 (2015).
 49. Wojciechowski, M. F., Lavin, M. & Sanderson, M. J. A phylogeny of legumes (Leguminosae) based on analysis of the plastid mat K gene resolves many well-supported subclades within the family. *American Journal of Botany* **91**, 1846–1862 (2004).
 50. Sabir, J. *et al.* Evolutionary and biotechnology implications of plastid genome variation in the inverted-repeat-lacking clade of legumes. *Plant Biotechnology Journal* **12**, 743–754 (2014).
 51. Tangphatsornruang, S. *et al.* The Chloroplast Genome Sequence of Mungbean (*Vigna radiata*) Determined by High-throughput Pyrosequencing: Structural Organization and Phylogenetic Relationships. *DNA Research* **17**, 11–22 (2010).
 52. Yin, D. *et al.* Development of chloroplast genome resources for peanut (*Arachis hypogaea* L.) and other species of Arachis. *Scientific Reports* **7**, 11649 (2017).

53. Ozyigit, I. I., Dogan, I. & Filiz, E. In silico analysis of simple sequence repeats (SSRs) in chloroplast genomes of *Glycine* species. *Plant Omics Journal* **8**, 24–29 (2015).
54. Dong, W., Liu, J., Yu, J., Wang, L. & Zhou, S. Highly Variable Chloroplast Markers for Evaluating Plant Phylogeny at Low Taxonomic Levels and for DNA Barcoding. *PLoS ONE* **7**, e35071 (2012).
55. Dong, W. *et al.* *ycf1*, the most promising plastid DNA barcode of land plants. *Scientific Reports* **5**, 8348 (2015).
56. Neubig, K. M. *et al.* Phylogenetic utility of *ycf1* in orchids: a plastid gene more variable than *matK*. *Plant Systematics and Evolution* **277**, 75–84 (2009).
57. Kong, H., Liu, W., Yao, G. & Gong, W. A comparison of chloroplast genome sequences in *Aconitum* (Ranunculaceae): a traditional herbal medicinal genus. *PeerJ* **5**, e4018 (2017).
58. Park, I. *et al.* The Complete Chloroplast Genomes of Six *Ipomoea* Species and Indel Marker Development for the Discrimination of Authentic Pharbitidis Semen (Seeds of *I. nil* or *I. purpurea*). *Frontiers in Plant Science* **9**, 965 (2018).
59. Liu, W. *et al.* Complete Chloroplast Genome of *Cercis chuniana* (Fabaceae) with Structural and Genetic Comparison to Six Species in Caesalpinioideae. *International Journal of Molecular Sciences* **19**, 1286 (2018).
60. Nei, M. & Kumar, S. *Molecular evolution and phylogenetics*. (Oxford university press, 2000).
61. Redwan, R. M., Saidin, A. & Kumar, S. V. Complete chloroplast genome sequence of MD-2 pineapple and its comparative analysis among nine other plants from the subclass Commelinidae. *BMC Plant Biology* **15**, 196 (2015).
62. Wang, D., Liu, F., Wang, L., Huang, S. & Yu, J. Nonsynonymous substitution rate (Ka) is a relatively consistent parameter for defining fast-evolving and slow-evolving protein-coding genes. *Biology Direct* **6**, 13 (2011).
63. Tian, N., Han, L., Chen, C. & Wang, Z. The complete chloroplast genome sequence of *Epipremnum aureum* and its comparative analysis among eight Araceae species. *PLOS ONE* **13**, e0192956 (2018).
64. Asaf, S. *et al.* Comparative analysis of complete plastid genomes from wild soybean (*Glycine soja*) and nine other *Glycine* species. *PLOS ONE* **12**, e0182281 (2017).
65. Niu, Y.-T. *et al.* Combining complete chloroplast genome sequences with target

- loci data and morphology to resolve species limits in *Triplostegia* (Caprifoliaceae). *Molecular Phylogenetics and Evolution* **129**, 15–26 (2018).
66. Amiryousefi, A., Hyvönen, J. & Poczar, P. The chloroplast genome sequence of bittersweet (*Solanum dulcamara*): Plastid genome structure evolution in Solanaceae. *PLOS ONE* **13**, e0196069 (2018).

Table S1. Accession number and sampled chloroplast genomes obtained from GenBank.

Species	Accession	Genome Size
Caesalpinioideae		
<i>Acacia dealbata</i>	NC_034985	174217 bp
<i>Acacia ligulata</i>	NC_026134	174233 bp
<i>Adenanthera microsperma</i>	NC_034986	159389 bp
<i>Albizia odoratissima</i>	NC_034987	174861 bp
<i>Archidendron lucyi</i>	NC_034988	176870 bp
<i>Dichrostachys cinerea</i>	NC_035346	161240 bp
<i>Faidherbia albida</i>	NC_035347	175646 bp
<i>Inga leiocalycina</i>	NC_028732	175489 bp
<i>Leucaena trichandra</i>	NC_028733	164692 bp
<i>Pararchidendron pruinsum</i>	NC_035348	176692 bp
<i>Parkia javanica</i>	NC_034989	161681 bp
<i>Piptadenia communis</i>	NC_034990	162552 bp
<i>Pithecellobium flexicaule</i>	NC_034991	178887 bp
<i>Samanea saman</i>	NC_034992	176717 bp
<i>Senegalia laeta</i>	NC_036736	162754 bp
<i>Senna tora</i>	NC_030193	162426 bp
<i>Vachellia flava</i>	NC_036734	165829 bp
<i>Vachellia seyal</i>	NC_036735	165383 bp
Outgroup		
<i>Cucumis sativus</i>	NC_007144	155293 bp
<i>Fragaria vesca</i>	NC_015206	155691 bp

Table S2. Models, Gamma Distribution, Proportion of Invariant Sites and Kappa, as Estimated by JModelTest for Each Gene Alignment.

Gene	Model	Gamma	Invariant sites	Kappa
<i>atpA</i>	GTR+I+G	0.849	0.559	N/A
<i>atpB</i>	GTR+G	0.270	N/A	N/A
<i>atpE</i>	HKY+I	N/A	0.705	3.117
<i>atpF</i>	GTR+I	N/A	0.466	N/A
<i>atpH</i>	GTR+I	N/A	0.784	N/A
<i>atpI</i>	GTR+G	0.165	N/A	N/A
<i>ccsA</i>	GTR+G	0.543	N/A	N/A
<i>cemA</i>	GTR+G	1.060	N/A	N/A
<i>clpP</i>	GTR+I+G	0.705	0.050	N/A
<i>matK</i>	GTR+G	1.000	N/A	N/A
<i>ndhA</i>	GTR+G	0.390	N/A	N/A
<i>ndhB</i>	GTR+I	N/A	0.808	N/A
<i>ndhC</i>	GTR+G	0.639	N/A	N/A
<i>ndhD</i>	GTR+G	0.420	N/A	N/A
<i>ndhE</i>	HKY+G	0.600	N/A	5.017
<i>ndhF</i>	GTR+I+G	0.911	0.289	N/A
<i>ndhG</i>	GTR+G	1.063	N/A	N/A
<i>ndhH</i>	GTR+I+G	0.673	0.442	N/A
<i>ndhI</i>	GTR+I	N/A	0.648	N/A
<i>ndhJ</i>	GTR+G	0.239	N/A	N/A
<i>ndhK</i>	GTR+G	0.304	N/A	N/A
<i>petA</i>	GTR+G	0.213	N/A	N/A
<i>petB</i>	HKY+I	N/A	0.763	6.439
<i>petD</i>	HKY+G	0.118	N/A	4.329
<i>petG</i>	HKY+I	N/A	0.765	4.481
<i>petL</i>	HKY+I	N/A	0.625	2.962
<i>petN</i>	GTR	N/A	N/A	N/A
<i>psaA</i>	GTR+I+G	0.832	0.563	N/A
<i>psaB</i>	GTR+G	0.096	N/A	N/A
<i>psaC</i>	GTR+I	N/A	0.760	N/A
<i>psaI</i>	HKY+G	0.504	N/A	5.257
<i>psaJ</i>	HKY+I	N/A	0.421	4.913
<i>psbA</i>	GTR+G	0.120	N/A	N/A
<i>psbB</i>	GTR+I+G	0.677	0.586	N/A
<i>psbC</i>	GTR+I+G	0.762	0.579	N/A
<i>psbD</i>	GTR+I+G	0.777	0.613	N/A
<i>psbE</i>	HKY+I	N/A	0.805	10.311

Gene	Model	Gamma	Invariant sites	Kappa
<i>psbF</i>	GTR	N/A	N/A	N/A
<i>psbH</i>	GTR+G	0.789	N/A	N/A
<i>psbI</i>	HKY	N/A	N/A	9.224
<i>psbJ</i>	GTR+I	N/A	0.593	N/A
<i>psbK</i>	HKY+I	N/A	0.533	4.310
<i>psbL</i>	HKY	N/A	N/A	6.413
<i>psbM</i>	HKY+I	N/A	0.730	7.709
<i>psbN</i>	HKY	N/A	N/A	4.153
<i>psbT</i>	GTR+I	N/A	0.608	N/A
<i>psbZ</i>	HKY+I	N/A	0.694	10.086
<i>rbcL</i>	GTR+I+G	0.748	0.655	N/A
<i>rpl2</i>	HKY	N/A	N/A	4.161
<i>rpl14</i>	HKY+G	0.148	N/A	5.142
<i>rpl16</i>	GTR+G	0.203	N/A	N/A
<i>rpl20</i>	GTR+I+G	0.748	0.470	N/A
<i>rpl23</i>	GTR	N/A	N/A	N/A
<i>rpl32</i>	HKY+G	1.119	N/A	1.981
<i>rpl33</i>	GTR+G	0.338	N/A	N/A
<i>rpl36</i>	HKY+G	0.041	N/A	6.957
<i>rpoA</i>	GTR+G	0.550	N/A	N/A
<i>rpoB</i>	GTR+G	0.293	N/A	N/A
<i>rpoC1</i>	GTR+I+G	0.761	0.431	N/A
<i>rpoC2</i>	GTR+G	0.536	N/A	N/A
<i>rps2</i>	GTR+I+G	0.875	0.481	N/A
<i>rps3</i>	GTR+G	0.555	N/A	N/A
<i>rps4</i>	GTR+I+G	0.826	0.327	N/A
<i>rps7</i>	GTR	N/A	N/A	N/A
<i>rps8</i>	GTR+I	N/A	0.577	N/A
<i>rps11</i>	GTR+G	0.922	N/A	N/A
<i>rps14</i>	HKY+I	N/A	0.603	7.049
<i>rps15</i>	GTR+G	0.943	N/A	N/A
<i>rps18</i>	HKY+G	1.127	N/A	2.797
<i>rps19</i>	GTR+I+G	0.800	0.411	N/A
<i>ycf2</i>	GTR+I+G	0.893	0.560	N/A
<i>ycf3</i>	GTR+I	N/A	0.692	N/A
<i>ycf4</i>	GTR+G	0.522	N/A	N/A

GTR = General Time Reversible; HKY85 = Hasegawa, Kishino and Yano (1985);

Table S3. Forward (F), Reverse (R) and palindromic (P) repeat sequences in the *Stryphnodendron adstringens* chloroplast genome.

No.	Type	Repeat length (bp)	Location	Region	Repeat unit
1	F	30	IGS (<i>trnQ-UUG</i> - <i>psbK</i>)	LSC	TTATTATTAATATTAATTCTATA ATATAAT
2	F	31	IGS (<i>trnP-UGG</i> - <i>psaJ</i>)	LSC	TAATAATAAGAATAAAGAATAA GAATAAAGA
3	F	32	IGS (<i>atpH</i> - <i>atpI</i>)	LSC	TTAATTTATATATTATTATTAATT TATAATTT
4	F	32	IGS (<i>ndhC</i> - <i>trnV-UAC</i>)	LSC	GTTATTTAGTATTTAATAGTATT AGTAAGTAT
5	F	32	IGS (<i>trnR-ACG</i> - <i>trnN-GUU</i>)	IR	AAAAAGCAAAAAAGATATGCGG ATTAGGGCAT
6	F	32	IGS (<i>trnR-ACG</i> - <i>trnN-GUU</i>)	IR	AATATAGTATGCCCTAATCCGCA TATCTTTTT
7	F	33	IGS (<i>trnS-GCU</i> - <i>trnG-UCC</i>)	LSC	TAAACAATATTCTATTTATAGAA AAAAATATTC
8	F	33	<i>ndhA</i> (intron)	SSC	TATAAAGAAGAAAAGAAATAAA GACGAAAAGAA
9	F	34	IGS (<i>accD</i> - <i>psaI</i>)	LSC	TTTTTATTATATTATTTCTATTTT ATTATATTAT

No.	Type	Repeat length (bp)	Location	Region	Repeat unit
10	F	35	IGS (<i>atpH</i> - <i>atpI</i>)	LSC	TTATTATATATATATTATATA TTATTATATATT
11	F	35	IGS (<i>ndhF</i> - <i>rpl32</i>)	SSC	TATATAATATAAAAAGGATAGTT TTTTTGATATAA
12	F	36	IGS (<i>ccsA</i> - <i>ndhD</i>)	SSC	TAATAAAATACAATAGATAATA AAATAAAATACAAT
13	F	36	IGS (<i>rps3</i> - <i>rps19</i>)	LSC	TCTTTGATACTCTTCTTTCCTTTC ATACTCTTCTTC
14	F	37	<i>ndhA</i> (intron)	SSC	CTATATCTATATATTATATATTTC TATATTATATCTA
15	F	39	ycf3 (intron), <i>ndhA</i> (intron), IGS(<i>rps12</i> - <i>trnV</i> -GAC)	LSC/SSC/IR	CCAGAACCGTACGTGAGATTTTC ATCTCATACGGCTCCT
16	F	41	IGS (<i>trnP</i> -UGG - <i>psaJ</i>)	LSC	TAATAAGAATAAAGAATAAGAA TAAAGAATAAGAATAAAGA
17	F	41	<i>psaB</i>	LSC	ATGCAATAGCTAAATGATGATG AGCAATATCAGTCAGCCAT
18	F	42	IGS (<i>rps12</i> - <i>trnV</i> -GAC)	IR	CTACAGAACCGTACATGAGATTT TCACCTCATACGGCTCCTC

No.	Type	Repeat length (bp)	Location	Region	Repeat unit
19	F	43	<i>rpl16</i> (intron)	LSC	ACTACAGAACCGTACATGAGAT TTTCACCTCATACGGCTCCTC
20	F	43	IGS (<i>rrn5</i> - <i>trnR-ACG</i>)	IR	CTCATTCTTATTACTTTTTTCATAT TGAAAAAGTAATAAGAATG
21	F	52	<i>rpl16</i> (intron), <i>ndhA</i> (intron)	LSC/SSC	TCCATCTCTACTACAGAACCGTA CATGAGATTTTCACCTCATACGG CTCCTC
22	F	128	IGS (<i>trnC-GCA</i> - <i>petN</i>)	LSC	TTAGAATATAAAATATAAAATAG AATATAAAATATAAAATAGAATA TAAAATATAAAATAGAATATAAA ATATAAATAGAATATAAAATAT AAATAGAATATAAAATATAAAT AGAATATAAAATATAAAT
23	R	30	IGS (<i>psbZ</i> - <i>trnG-UCC</i>)	LSC	ATAGATATAGATAGATATAGAT AGATATAG
24	R	38	IGS (<i>rbcl</i> - <i>accD</i>)	LSC	CTATTTTATATTTCTATATATTTT TATATTTCTATATA
25	P	30	IGS (<i>trnQ-UUG</i> - <i>psbK</i>)	LSC	TTAATAATTTAATTAATTAATTA AATTATT

No.	Type	Repeat length (bp)	Location	Region	Repeat unit
26	P	31	IGS (trnQ-UUG - psbK)	LSC	ATTAATTCTATAATATAATTAGA ATATTAAT
27	P	31	IGS (trnT-UGU - trnL-UAA)	LSC	ATATATCTAAATAAATAGTTATC TATAAAAT
28	P	31	IGS (trnK-UUU - rps16)	LSC	GGAACGGGACTAAATGGATCCA TTTAGTCCC
29	P	32	IGS (rps3 - rps19)	LSC	AATAATAAATATTTATATATTAA TATTTATTA
30	P	32	IGS (trnR-ACG - trnN-GUU)	IR	AAAAAGCAAAAAAGATATGCGG ATTAGGGCAT
31	P	35	trnS-GCU	LSC	GGGAAAGAGAGGGATTTCGAACC CTCGGTACGAAAA
32	P	37	IGS (ccsA - ndhD)	SSC	AATTTTATTAAAATTAATTAAAT TAAATTTAATAAAA
33	P	39	ycf3(intron)	LSC	CCAGAACCGTACGTGAGATTTTC ATCTCATACGGCTCCT
34	P	41	IGS (trnQ-UUG - psbK)	LSC	TTTTATTAATATTAATTCTATAAT ATAATTAGAATATTAAT

No.	Type	Repeat length (bp)	Location	Region	Repeat unit
35	P	41	IGS (accD - psal)	LSC	ATTTAGAATAGTATATTTAGAAT CTAATTATACTATTCTAA
36	P	42	ndhA (intron)	SSC	TTACAGAACCGTACATGAGATTT TCACCTCATACGGCTCCTC
37	P	43	rpl16 (intron), IGS (trnV-GAC - rps12)	LSC	ACTACAGAACCGTACATGAGAT TTTCACCTCATACGGCTCCTC
38	P	43	IGS (rrn5 - trnR-ACG)	IR	CTCATTCTTATTACTTTTTTCATAT TGAAAAAGTAATAAGAATG
39	P	43	IGS (trnR-ACG - rrn5)	IR	TTCATTCTTATTACTTTTTCAATA TGAAAAAGTAATAAGAATG
40	P	44	IGS (psbT - psbN)	LSC	ATTGAAGTAATGAGCCTACCAAT ATTGGCAGGCTCATTACTTCA
41	P	49	IGS (petN - psbM)	LSC	AGATAGTATGGTAGAAAGAAAT AAAATCTATTTCTTTCTACCATA CTAT
42	P	60	IGS (rbcl - accD)	LSC	AATTTCAATTCTATAATATTCAA TTCTATTATTATAGAATTGAATA TTATAGAATTGAAA

Table S4. Distribution of simple sequence repeat (SSR) in the *Stryphnodendron adstringens* chloroplast genome.

No. SSR	SSR type	SSR	Size	Start	End	Location
1	p1	(T)10	10	215	224	LSC
2	p3	(CAG)4	12	1065	1076	LSC (psbA)
3	p4	(TTCA)3	12	1721	1732	LSC
4	p1	(T)13	13	2067	2079	LSC
5	p1	(A)14	14	3897	3910	LSC
6	p1	(A)13	13	4439	4451	LSC
7	p1	(A)11	11	4553	4563	LSC
8	p1	(A)13	13	4990	5002	LSC
9	p2	(AT)6	12	5302	5313	LSC
10	c	(G)10agataagagagaaaaggattatataaatctacataaaagtcac cacaccctctttttctttttga(ATT)3	91	5900	5990	LSC
11	c	(T)12cttattctattctttcccg(A)10	41	6399	6439	LSC (*rps16)
12	p1	(T)10	10	6833	6842	LSC (*rps16)
13	p1	(A)13	13	7674	7686	LSC
14	c	(ATT)4tcttatttc(T)14attattatttattaattaataat(TTAA)4 attattattaatattaattctataataattttattaatattaattctataatata attagaatattaattaatagataaatcataaatgaaat(A)10ttcaatta aaaattcatttattagaataattcaga(AT)6aattaataat(AATAG A)3	255	8357	8611	LSC

No. SSR	SSR type	SSR	Size	Start	End	Location
15	p1	(A)13	13	8953	8965	LSC
16	p4	(CAAA)3	12	9386	9397	LSC
17	c	(AT)6ttctataatattctatagatattctatagaataataaaattagaata ctaataataaaattagaataactaata(AAT)4atttcagaatttatatt aat(ATT A)3	131	9635	9765	LSC
18	c	(A)11tacttgattgaacaagatcaaacataagacataagaatgtca tttctattcctctttctttttccgctaaaatattctaaaagaaaaag(A)1 3	119	10459	10577	LSC
19	c	(AT)6tttttagctcgatgaaaatcaaagcttttcttaaggattctcttaa atagaaatagagaacgaagtaactag(A)18	103	11234	11336	LSC
20	p1	(T)12	12	14736	14747	LSC (*atpF)
21	p1	(T)10	10	14911	14920	LSC (*atpF)
22	p3	(TTA)4	12	15477	15488	LSC
23	p1	(A)14	14	16346	16359	LSC
24	p1	(T)10	10	16972	16981	LSC
25	c	(TA)6ttatatattattatat(ATTAA)3	43	17214	17256	LSC
26	c	(T)10acaattccctggtaactctttttacttttagacaaaatcgatactta ttttatttagaccttattgcactttc(TTATT)3	101	17421	17521	LSC
27	c	(T)11caaat(A)10gaatttg(A)12	45	19338	19382	LSC
28	p1	(T)13	13	19681	19693	LSC (rpoC2)
29	c	(T)12accgttccgagtggtatcaagatgccgctgtgtcgggatatctt atccatctctccaggaaaatagatatcccgagaaaagattttaattcaat cc(T)10	119	21486	21604	LSC (rpoC2)

No. SSR	SSR type	SSR	Size	Start	End	Location
30	p1	(T)10	10	22129	22138	LSC (rpoC2)
31	p2	(TA)5	10	22973	22982	LSC (rpoC2)
32	c	(A)13gaagtc(T)12	31	26103	26133	LSC (*rpoC1)
33	p1	(T)10	10	29375	29384	LSC (rpoB)
34	p1	(T)10	10	30113	30122	LSC
35	p1	(A)10	10	31165	31174	LSC
36	p4	(TTCT)3	12	31750	31761	LSC
37	p2	(TA)5	10	31887	31896	LSC
38	p4	(TTAA)3	12	32404	32415	LSC
39	p1	(T)11	11	33886	33896	LSC
40	p2	(TG)5	10	34122	34131	LSC
41	p1	(A)14	14	35274	35287	LSC
42	p1	(T)11	11	35996	36006	LSC
43	c	(A)15tagactttggagtttttttattaatctgaaagaaaggaaaatataa caaagaagacaataccaataaaag(A)16	102	36942	37043	LSC
44	c	(TA)5catgcaggatccagcatgcccctttgtcaagt(A)10	52	40405	40456	LSC
45	c	(ACTAT)3aataactataactatataactataat(TATAA)3	57	47494	47550	LSC
46	c	(T)10ctcgagtt(A)15	33	47896	47928	LSC (*ycf3)

No. SSR	SSR type	SSR	Size	Start	End	Location
47	p1	(T)10	10	50848	50857	LSC
48	p2	(TA)5	10	50981	50990	LSC
49	p3	(AAT)4	12	51856	51867	LSC
50	p1	(A)11	11	52224	52234	LSC
51	p1	(T)11	11	53472	53482	LSC
52	p2	(AT)5	10	56199	56208	LSC
53	p1	(T)15	15	56362	56376	LSC
54	p4	(TACA)3	12	56489	56500	LSC
55	p1	(T)10	10	59779	59788	LSC (atpB)
56	p1	(T)10	10	60222	60231	LSC
57	p1	(A)14	14	60393	60406	LSC
58	p3	(TTA)4	12	62856	62867	LSC
59	p4	(AATA)3	12	64584	64595	LSC
60	p1	(T)13	13	64737	64749	LSC
61	p1	(A)12	12	65311	65322	LSC
62	c	(A)10tgaactgaaaactgaatggaatgctttctt(A)13	53	65893	65945	LSC
63	p1	(T)14	14	68432	68445	LSC
64	c	(A)10tagaaaaaataaatagatactatatac(TA)5	47	69039	69085	LSC

No. SSR	SSR type	SSR	Size	Start	End	Location
65	p1	(A)10	10	70797	70806	LSC
66	p4	(TGAT)3	12	71479	71490	LSC
67	c	(TA)5gcataatc(TA)5gaactctcatc(AT)5	49	72289	72337	LSC
68	p3	(ATA)4	12	72642	72653	LSC
69	p1	(T)10	10	73160	73169	LSC
70	p2	(TA)5	10	73924	73933	LSC
71	p1	(A)13	13	74461	74473	LSC (rps18)
72	p1	(T)14	14	75780	75793	LSC
73	p1	(T)11	11	76780	76790	LSC (*clpP)
74	c	(A)10tcaatgtgtagatttcaacctctttc(T)12	49	77417	77465	LSC (*clpP)
75	c	(T)11caaggagttagaagttataataaatcgaaagatttgattctttca aactgctgcttatcc(T)12	84	81377	81460	LSC (*petB)
76	c	(T)12ag(A)13	27	84249	84275	LSC
77	c	(T)23gtgcatccgggtcaagcccc(T)20	63	85878	85940	LSC
78	p3	(ATT)4	12	87059	87070	LSC
79	c	(A)14gaaatattgtgtatc(A)10	39	87709	87747	LSC
80	p1	(T)12	12	89093	89104	LSC (*rpl16)
81	c	(ATTA)3a(TAC)4	25	89224	89248	LSC (*rpl16)

No. SSR	SSR type	SSR	Size	Start	End	Location
82	p1	(T)10	10	90173	90182	LSC
83	p1	(T)10	10	90371	90380	LSC
84	p1	(T)10	10	91165	91174	IRA
85	p5	(AATGG)3	15	101485	101499	IRA
86	p2	(AG)5	10	115211	115220	IRA
87	p5	(TTAAT)3	15	117210	117224	SSC
88	p1	(T)10	10	118022	118031	SSC (ndhF)
89	p1	(A)14	14	119489	119502	SSC
90	p1	(T)12	12	119714	119725	SSC
91	p1	(A)11	11	120298	120308	SSC
92	p2	(TA)5	10	120715	120724	SSC
93	p1	(T)10	10	120843	120852	SSC
94	p1	(T)11	11	122704	122714	SSC
95	p1	(T)13	13	130963	130975	SSC
96	p1	(T)11	11	132548	132558	SSC (ycf1)
97	p1	(T)11	11	132755	132765	SSC (ycf1)
98	p1	(T)12	12	132906	132917	SSC (ycf1)
99	p1	(T)11	11	133696	133706	SSC (ycf1)

No. SSR	SSR type	SSR	Size	Start	End	Location
100	p1	(T)10	10	134082	134091	SSC (ycf1)
101	c	(T)13agcttcggaatccacctttcatgttctcgaaataaataaagtcct ttaaattgaaattggatactgattttgcagaaaattgactgatcaagtt(A)10	119	135144	135262	SSC (ycf1)
102	p1	(T)11	11	135967	135977	SSC (ycf1)
103	p2	(CT)5	10	137994	138003	IRB
104	p5	(TTCCA)3	15	151713	151727	IRB
105	p1	(A)10	10	162040	162049	IRB

CAPÍTULO 2

TAMANHO DO GENOMA, MONTAGEM “DE NOVO” E ANOTAÇÃO DO GENOMA NUCLEAR DE *Stryphnodendron adstringens* (LEGUMINOSAE)

RESUMO

A espécie *Stryphnodendron adstringens* (Mart.) Coville, conhecida popularmente como barbatimão é amplamente utilizada na medicina popular como fitoterápico. Em função do seu alto potencial de utilização como planta medicinal, a caracterização do genoma da espécie pode contribuir para o desenvolvimento de ferramentas genômicas no âmbito relativo à conservação e uso para gerar produtos biotecnológicos. Neste contexto, o objetivo deste trabalho foi estimar o tamanho do genoma e realizar a montagem e anotação do genoma nuclear de *S. adstringens*. O conteúdo médio de DNA nuclear obtido por citometria de fluxo foi de $1C = 0.684$ pg e o tamanho do genoma haploide foi estimado em 669 Mpb. Por outro lado, utilizando a distribuição de *k-mers* o tamanho do genoma foi estimado em 624 Mbp. A montagem “de novo” gerou um total de 63.320 *contigs* com N50 de 14.541 pb. Um total de 20.124 microssatélites foi identificado, dos quais 16910 são di (84,03%), 1822 tri (9,05%), 880 tetra (4,37%), 322 penta (1,60%) e 190 hexanucleotídeos (0,94%). O motivo de repetição AT/TA foi o mais frequente (44,16%), seguido dos motivos AG/TC (12,44%) e CT/GA (11,78%). Foi possível identificar que 44,58% do genoma é composto de elementos transponíveis. A maior parte dos elementos transponíveis identificado (64,20%) não foram classificados em nenhuma categoria. Como já observado nos genomas de plantas, os retroelementos foram a classe mais comum de elementos transponíveis encontrado no genoma de *S. adstringens*, dentre estes, as repetições terminais longas (LTR) foi a família mais abundante. Foram preditos 31.636 genes e 95.811 transcritos com um média de 3,03 transcritos por gene. 60.798 transcritos foram classificados funcionalmente, nos termos do *Gene Ontology*. Foram identificados 1.521 transcritos (1,58%) com similaridade com proteínas envolvidas nas vias de biossíntese de taninos. Com isso, a caracterização do genoma da espécie *S. adstringens* fornece um recurso genômico valioso para o futuro desenvolvimento de marcadores moleculares e aplicações biotecnológicas.

Palavras-chaves: barbatimão, elementos repetitivos, genes, taninos, transcritos.

3.1 INTRODUÇÃO

A espécie *Stryphnodendron adstringens* (Mart.) Coville, conhecida popularmente como barbatimão é nativa dos cerrados do Centro Oeste e Sudeste do Brasil, com ampla distribuição geográfica e ocorrência confirmada nos estados do Paraná, Minas Gerais, São Paulo, Bahia, Tocantins, Distrito Federal, Goiás, Mato Grosso e Mato Grosso do Sul (Souza & Gibau, 2018). Essa espécie possui vários compostos produzidos pelo seu metabolismo secundário, como, os alcaloides, terpenos, flavonoides, esteroides e taninos, sendo este último o seu constituinte predominante que lhe agrega valor terapêutico (Panizza et al., 1988). Os taninos estão presentes em toda a planta, porém se concentram principalmente nas cascas presentes no caule, com cerca de 20 a 40% de taninos em extrato aquoso (Panizza et al., 1988; Lopes et al., 2009; Corrêa et al., 2012). Vários estudos já avaliaram as propriedades medicinais das cascas do caule para uso em tratamentos antibacteriano (Audi et al., 2004), anti-inflamatório (Lima et al., 1998), antifúngico (Ishida et al., 2006; Luiz et al., 2015), antinociceptiva (Melo et al., 2007), adstringente e cicatrizante (Lima et al., 2016), além de efeito protetor direto da mucosa do estômago (Audi et al., 1999).

Apesar da importância do uso de *S. adstringens* na medicina fitoterápica, estudos na área da genética ainda são relativamente recentes e restritos a poucos grupos de pesquisa no Brasil. Porém, podem-se encontrar alguns estudos que utilizaram marcadores moleculares (Isoenzimas e AFLP - *Amplified Fragment Length Polymorphism*) para elucidar a variabilidade genética entre e dentro de populações naturais de *S. adstringens* (Glasenapp et al., 2012, 2014; Mendonça et al., 2012).

Com o advento das tecnologias de sequenciamento de nova geração, o sequenciamento de genomas inteiros passou a ser conduzido de maneira rápida e com custos mais acessíveis por giga base de dados gerados, quando comparados ao sequenciamento Sanger (Metzker, 2010; Goodwin et al., 2016). Além disso, com o aprimoramento constante de ferramentas computacionais para análise dos dados, o sequenciamento de alta performance passou então a ser aplicado em organismos não-modelos (Unamba et al., 2015), incluindo, as plantas medicinais (Chen et al., 2011; Kellner et al., 2015; Mochida et al., 2017).

Nos dias atuais, genomas eucarióticos grandes e complexos podem ser montados em pouco tempo, com ou sem um genoma de referência, utilizando abordagem “*de novo*” (Alhakami et al., 2017). Após a montagem do genoma, a caracterização de sequências

repetitivas de DNA é o primeiro passo para garantir eficiência e acurácia na anotação desses genomas (Yandell & Ence, 2012). O número e a complexidade de sequências de DNA repetitivo variam entre espécies, porém, de maneira geral, espécies com tamanhos de genoma maiores possuem mais elementos repetitivos (Hawkins et al., 2008). Os elementos repetitivos são basicamente representados por repetições em tandem, tais como microsatélites, e os elementos transponíveis (López-Flores & Garrido-Ramos, 2012). Os elementos transponíveis se dividem em dois grupos principais: transposons de DNA e retrotransposons (López-Flores & Garrido-Ramos, 2012).

Para aproveitar todo o potencial das sequências genômicas geradas pelas plataformas de sequenciamento de nova geração, após a montagem, é necessário realizar a anotação buscando atribuir informações biológicas, estruturais e funcionais para um genoma (Yandell & Ence, 2012; Ekblom & Wolf, 2014). Uns dos objetivos de uma anotação é identificar as principais características do genoma, em particular os genes e seus produtos (Ekblom & Wolf, 2014).

Inicialmente, programas com algoritmos *ab initio*, que utilizam apenas sequências genômicas como base de dados, podem ser usados como predição base para encontrar genes codificadores de proteínas (Yandell & Ence, 2012; Ekblom & Wolf, 2014). Após predição das sequências de aminoácidos codificantes de proteínas é importante associá-las com sua função biológica. A anotação funcional automática é uma forma interessante de resolver este problema (Yandell & Ence, 2012). Na anotação automática, todas as sequências gênicas preditas em um genoma são anotadas por comparação com sequências disponíveis em bancos de dados públicos, como por exemplo, o NCBI. Posteriormente essas sequências são classificadas em categorias dos termos GO (*Gene Ontology*) para determinação funcional gênica (Conesa et al., 2005).

A obtenção de informações ao nível de genoma para plantas medicinais é importante para a compreensão do envolvimento dos genes em vias metabólicas específicas e pode fornecer um recurso genético valioso para prospecção e descoberta de medicamentos baseados em produtos naturais (Rai et al., 2017). Desde a publicação de genes associados com a biossíntese de benzoxazinoide em milho (*Zea mays*, Jonczyk et al., 2008) diversos outros genes associados com vias de metabólitos secundários têm sido reportados, tais como genes associados com vias de biossíntese de talianol sintase em *Arabidopsis thaliana* (Field & Osbourn, 2008), terpenos em *Solanum lycopersicum* (Matsuba et al., 2013), cucurbitacina

em *Cucumis sativus* (Shang et al., 2014) e taninos condensados em *Rhus gallus* (Chen et al., 2018).

Neste contexto, o objetivo deste trabalho foi estimar o tamanho do genoma e realizar a montagem e anotação de elementos repetitivos e genes no genoma de *S. adstringens*.

3.2. MATERIAL E MÉTODOS

3.2.1 Extração do DNA genômico e sequenciamento

Foram obtidas amostras de tecido foliar a partir de dois indivíduos, SadNIQGO_31 de Niquelândia-Goiás e SadCANMG_16 de Candeias-Minas Gerais. O DNA genômico foi obtido de acordo com o protocolo descrito por Doyle & Doyle (1987). A avaliação da concentração e qualidade do DNA foi feita por eletroforese horizontal em gel de agarose 1% e no equipamento *Qubit® Fluorometer Quantitation* (Life Technologies, 2014).

Duas bibliotecas *paired-end* de sequenciamento foram preparadas de acordo com o protocolo *Illumina Nextera® XT DNA Library Prep Kit* (Illumina, San Diego, CA). O sequenciamento foi realizado usando a plataforma *Illumina HiSeq2500* (2 x 100 bp com tamanho médio de inserto de 450 pb) no Centro de Genômica Funcional da Universidade de São Paulo (Piracicaba, Brasil). Os dados de sequenciamento foram submetidos à análise de qualidade no programa FastQC v0.11.5 (Andrews, 2010). As bases de baixa qualidade, bem como as sequências de adaptadores foram removidas com o programa *Trimmomatic* v0.36 (Bolger et al., 2014), com os seguintes parâmetros: ILLUMINACLIP: NexteraPE-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:50. O programa *FastqUniq* (Xu et al., 2012) foi posteriormente usado para remover *reads* duplicados em cada uma das bibliotecas de sequenciamento.

3.2.2 Tamanho do genoma

3.2.2.1 Estimativa do tamanho do genoma por Citometria de fluxo

Sementes de *S. adstringens* foram pré-embebidas em solução aquosa de hipoclorito de sódio a 1%, durante 15 minutos, com posterior lavagem em água destilada por 90

segundos. Após secagem em temperatura ambiente as sementes foram semeadas em substrato vermiculita para germinação. Cerca de 50-60 mg do tecido foliar das sementes germinadas de três indivíduos de *S. adstringens*, foram coletadas e utilizadas para citometria de fluxo. A mesma quantidade de tecido foliar jovem de sementes germinadas de *Solanum lycopersicum* ‘Stupicke’ (2C = 1,96 pg ou 1.917 Mbp); (Doležel et al., 1992) e *Raphanus sativus* ‘Saxa’ (2C= 1.11 pg or 1.086 Mbp); (Doležel et al., 1992) foram coletadas e utilizadas como padrões internos na citometria de fluxo. As sementes germinadas dos padrões internos foram obtidas no Instituto de Botânica Experimental em Olomouc, República Tcheca.

Foram analisadas três amostras por tratamento (delineamento inteiramente ao acaso composto por três repetições). As amostras de tecido foliar foram trituradas com auxílio de uma lâmina, em placas de Petri adicionando-se 1,5 mL de tampão de extração de núcleos WPB (*Woody Plant Buffer*, Loureiro et al., 2007). Em seguida, a suspensão foi passada através de um filtro com poros de 30 µm, de forma a eliminar a maior parte dos resíduos. Esta suspensão foi então corada com 50 µl de iodeto de propídio (1 mg/mL).

As amostras foram analisadas em um citômetro de fluxo *Partec CyFlow Space* (Partec, Münster, Alemanha) equipado com um laser de luz operando a 532 nm. Para cada amostra, pelo menos 2.000 núcleos foram medidos em cada leitura e analisados utilizando o software *Partec FloMax* (Versão 2.4). Os picos de amostra e padrão foram manualmente delimitados nos histogramas. As amostras com um pico de coeficiente de variação (HPCV) superior a 5% não foram consideradas para análise. O conteúdo de DNA nuclear (pg) de *S. adstringens* foi estimado por comparação com base na fluorescência média emitida pelos núcleos em fase G1 do ciclo celular, do padrão interno de referência, de acordo com a fórmula adaptada de Dolezel & Bartos (2005).

Conteúdo de DNA 2C da amostra (pg)

$$= \frac{\text{Pico G1 amostra}}{\text{Pico G1 padrão}} \times \text{Conteúdo de DNA 2C do padrão (pg)}$$

O tamanho médio do genoma foi apresentado como valores 1C em picogramas (pg), e em mega pares de base com a conversão de 1pg = 978 Mbp.

3.2.2.2 Estimativa do tamanho do genoma por análise de *k-mers*

O tamanho do genoma de *S. adstringens* também foi estimado usando os *reads* das duas bibliotecas de sequenciamento após o controle de qualidade, com o valor de *k-mer* de 31. A frequência de distribuição de *k-mers* foi obtida com o programa *Jellyfish* v.2.2.6 (Marçais & Kingsford, 2011), usando o comando “*jellyfish count*”. O comando “*jellyfish histo*” foi posteriormente usado para obter o histograma de frequência de *k-mers* e determinar o “pico” da curva de distribuição de frequência. O tamanho do genoma de *S. adstringens* foi então estimado usando o script em perl “*estimate_genome_size.pl*” v.0.03 (Ryan, 2013) (https://github.com/josephryan/estimate_genome_size.pl).

3.2.3 Montagem do genoma

A montagem “de novo” do genoma foi realizada utilizando as duas bibliotecas com o programa MaSurCA (Zimin et al., 2013), que combina a eficiência dos *grafos de Bruijn* e o *Overlap-Layout-Consensus* para a montagem dos *contigs*. O pipeline *Redundans* (Pryszcz & Gabaldón, 2016) foi utilizado para identificar e combinar *contigs* redundantes, advindo da montagem de diferentes regiões heterozigotas do genoma em *contigs* separados.

Visando melhorar a qualidade da montagem obtida, *contigs* composto por menos de 500 bp foram removidos. Posteriormente, os *contigs* foram submetidos à uma análise de similaridade com o BLASTN, para identificar sequências com níveis de similaridade com sequências conhecidas de DNA organelar. Para isso, foram criados dois bancos de dados, um para cloroplasto e outro para mitocôndria, usando sequências organelares de angiospermas depositadas no *RefSeq* do *GenBank*. Sequências com *hits* com *e-value* menor que 1e-15 e cobertura do *query* menor que 90% foram removidos da montagem. As estatísticas descritivas das montagens foram obtidas usando o programa QUAST (Gurevich et al., 2013).

3.2.4 Anotação de elementos repetitivos

3.2.4.1 Caracterização de Elementos Transponíveis

A identificação de TEs no genoma de *S. adstringens* foi realizada pelos métodos de similaridade de sequências e “de novo”. A identificação de TEs baseados em similaridade foi realizada com o auxílio do programa *RepeatMasker* v.4.0.7 (Smit et al., 2017), o qual

utiliza o banco de dados *RepBase Update* mantido pelo *Genetic Information Research Institute* (GIR). Essa análise foi realizada inicialmente com as sequências nucleares montadas e o banco de dados de plantas eudicotiledôneas para identificar repetições comuns entre *S. adstringens* e outras espécies de plantas com genomas já anotados.

Para a identificação de TEs “de novo” foi criado um banco de dados de repetições para *S. adstringens*, e em seguida foi realizada uma análise com as sequências nucleares e o banco de dados criado utilizando o programa *RepeatModeler* (Smit & Hubley, 2012). O total de TEs identificados “de novo” foi obtido com o programa *RepeatMasker* usando o arquivo de saída do *RepeatModeler* com a extensão “*classified*” e o arquivo de saída da primeira análise realizada no *RepeatMasker* “*fasta.masked*”, contendo as sequências nucleares já “mascaradas” pelo método de similaridade.

3.2.4.2 Caracterização de regiões microssatélites

A identificação de regiões microssatélites foi realizada por meio dos módulos do programa QDD v.3 (Megléc et al., 2014). O pipeline deste programa envolve quatro módulos que vão desde a identificação de regiões microssatélites, análise de similaridade entre os microssatélites identificados por BLAST de modo a manter somente sequências únicas (*singletons*), desenho de iniciadores e a verificação de uma possibilidade de um envolvimento dos microssatélites no contexto de regiões de elementos transponíveis (TEs) conhecidos com o programa (Megléc et al., 2014).

O número mínimo de repetições para cada tamanho de motivo usado na busca dos microssatélites foram: dez para di-nucleotídeos, oito para tri, seis para tetra-nucleotídeos, e cinco para penta e hexa-nucleotídeos. Visando o futuro desenvolvimento de marcadores microssatélites para a espécie *S. adstringens*, o desenho dos primers foi realizado utilizando o programa *Primer3* (Rozen & Skaletsky, 2000) no módulo 3 do QDD. Para o desenho foram considerados os seguintes parâmetros: tamanho do *amplicon* 150-400pb, teor de GC de 30-60%, o valor de TM de 56-62°C, e comprimentos dos primers de 20-25pb. Por fim, foi verificada a associação das regiões microssatélites com elementos transponíveis usando o *RepeatMasker* (Smit et al., 2017). Essa etapa é um diferencial deste programa, pois permite uma maior eficiência no desenho de primers, uma vez que se as regiões microssatélites no contexto de TEs forem evitadas, então pode-se minimizar a ocorrência de alelos nulos na etapa de genotipagem.

3.2.5 Predição e anotação funcional de genes

A predição de genes foi realizada usando uma abordagem de predição *ab initio* no arquivo final de saída do *RepetMasker*, contendo as sequências repetitivas “mascaradas”. A predição de genes foi realizada usando o programa AUGUSTUS (Stanke et al., 2006). Ao final dessa etapa, foi fornecido pelo programa um arquivo no formato GFF3, que contém as informações de posição de genes, éxons, íntrons e de transcritos para cada gene (Stanke et al., 2006). A partir de um *script* fornecido pelo próprio programa (*getAnnoFast.pl*), foi possível obter as sequências de proteínas preditas.

As sequências de proteínas preditas no genoma de *S. adstringens* foram então comparadas com um banco de dados de sequências de proteínas de outras espécies de embriófitas utilizando a ferramenta BLASTP (*E-value* $1e-5$) (Altschul et al., 1997). O arquivo de saída fornecido pelo BLASTP foi utilizado para anotação funcional dos genes, de acordo com as categorias de ontologia do gene (*Gene Ontology* – GO; <http://www.geneontology.org/>): “Componente Celular”, “Função Molecular” e “Processo Biológico”. O agrupamento dos genes foi realizado considerando o segundo nível hierárquico. Essa análise foi realizada por meio do programa Blast2GO v. 5.2.5 (Conesa et al., 2005).

Adicionalmente, foi criado um banco de dados de sequências de proteínas envolvidas na via de biossíntese de taninos (veja material suplementar disponível em: https://www.dropbox.com/sh/uxkknirf6t7hcyg/AAAqysgdVk_YhItK4l85OkwOa?dl=0). Essas sequências foram obtidas através do *UniProt-Swiss-Prot* (<https://www.uniprot.org/>). Posteriormente, foi então realizada uma busca de similaridade das sequências de proteínas preditas no genoma de *S. adstringens* com o banco de dados criado, usando o programa BLASTP (*E-value* $1e-5$).

3.3. RESULTADOS E DISCUSSÃO

3.3.1 Tamanho do genoma

O conteúdo médio de DNA nuclear obtido por citometria de fluxo foi de $1C = 0.684$ pg e o tamanho do genoma haploide foi estimado em aproximadamente 669 Mpb. Pode-se verificar que a determinação dos valores relativos para a quantidade de DNA por citometria

de fluxo, utilizando o fluorocromo iodeto de propídio, gerou histogramas com boa resolução, com picos de núcleos G1/G2, com coeficiente de variação de 3,91% para *Raphanus sativus* e 4,02% para *Solanum lycopersicum* (Figura 1).

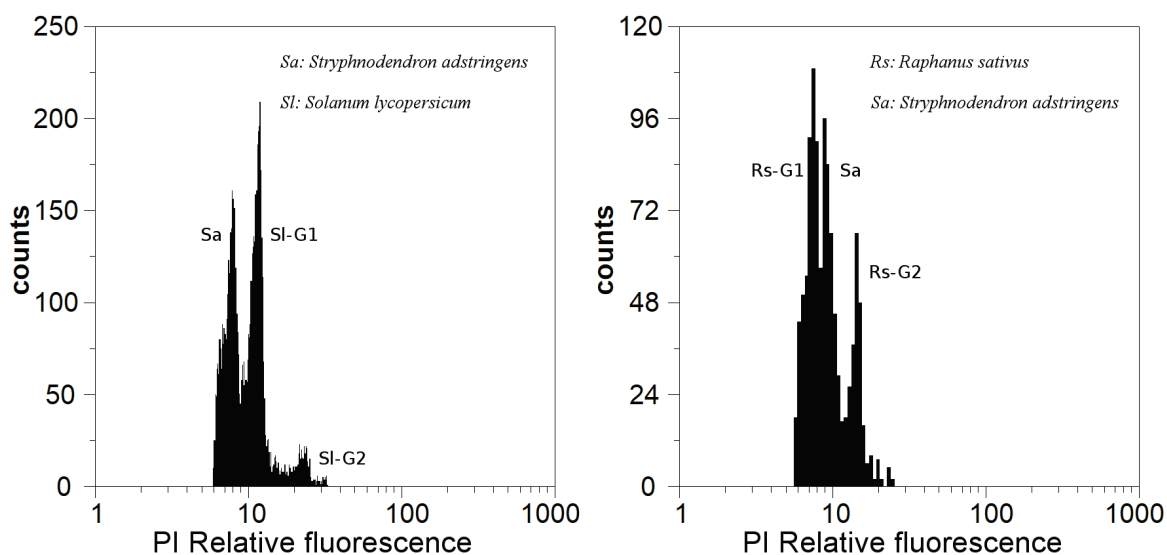


Figura 1. Histogramas de fluorescência relativa (unidades arbitrárias) de núcleos corados com iodeto de propídio obtidos por Citometria de Fluxo. Sa: *Stryphnodendron adstringens*. Sl: *Solanum lycopersicum*. Rs: *Raphanus sativus*.

O tamanho estimado do genoma para *S. adstringens* baseado na distribuição de 31-mer foi de 624 Mpb, valor bem próximo do que foi obtido por citometria de fluxo (Figura 2).

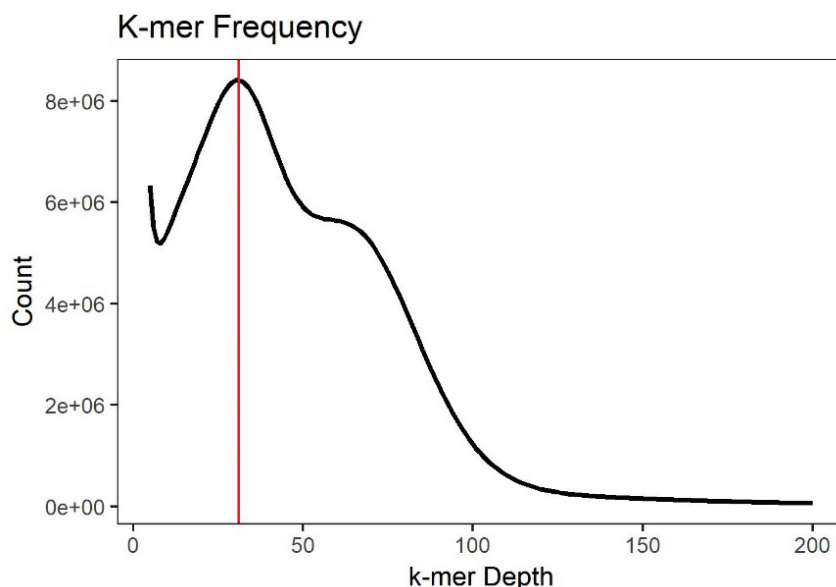


Figura 2. Distribuição de *k-mers* nos *reads* de sequenciamento da espécie *Stryphnodendron adstringens*

Antes do início de um projeto de sequenciamento de genoma, uma das principais informações é o tamanho do genoma da espécie de interesse. Essa informação é fundamental, uma vez que, quanto maior o tamanho do genoma da espécie, mais esforço de sequenciamento de DNA deve ser feito para garantir cobertura adequada para se conhecer as diferentes regiões do genoma. Os métodos mais usados para obtenção de estimativas de tamanho de genoma são *densitometria de Feulgen* e citometria de fluxo. Esses dois métodos têm sido extensivamente aprimorados, e várias fontes de erros identificadas e minimizadas (Doležel et al., 2007; Hare & Johnston, 2012). Contudo, estimativas baseadas por métodos de bioinformática, por exemplo por distribuição de *k-mers* usando dados de sequenciamento de nova geração, tem sido utilizadas em vários projetos de sequenciamento de genomas de espécies de plantas (Guo et al., 2018; Dong et al., 2018).

A determinação do conteúdo de DNA foi primeiramente estudada em 1973, na espécie *Vicia faba* (Heller, 1973). Atualmente, se encontram depositadas no banco de dados “*Plant DNA C-values*” estimativas de tamanhos de genomas para 8.510 espécies (Bennett & Leitch, 2012). Deste total, 722 são espécies pertencentes à família Leguminosae. As espécies dessa família possuem uma grande variação no tamanho do genoma, variando de 298 Mbp (1C = 0.31) à 27 Gbp (1C = 27.40 pg) para *Leucaena macrophylla* e *Vicia faba*, respectivamente (Valor médio 1C = 2.129 ± 2.526 pg, Mediana = 1.174 pg, n = 722).

Entretanto, a estimativa do tamanho do genoma de *S. adstringens* se mostra próximo ao que tem sido observado em algumas espécies da subfamília Caesalpinioideae - clado mimosoide, como por exemplo, *Leucaena esculenta* (636 Mbp; Palomino et al., 1995), *Acacia mangium* (636 Mbp; Blakesley et al., 2002), *Mimosa invisa* (660 Mbp; Bennett et al., 1998), *Adenanthera pavonina* (665 Mbp; Ohri et al., 2004), e *Adenanthera microsperma* (680 Mbp; Ohri et al., 2004). Esse é o primeiro trabalho onde se obteve uma estimativa tamanho do genoma para uma espécie do gênero *Stryphnodendron*.

3.3.2 Montagem do genoma

O sequenciamento gerou um total de 554.143.303 *short-reads* do tipo *paired-end* (135 Gb), com tamanho médio de 101 pb, compreendendo um total de 111.936.947.206 pb. Após filtragem esse número foi reduzido para 511.837.144 *short-reads* do tipo *paired-end* (122 Gb) com tamanho médio de 99,02 pb, compreendendo um total de 100.919.954.035 pb o que corresponde aproximadamente 150x de profundidade de sequenciamento, assumindo-se um tamanho de genoma de 669 Mbp.

A montagem inicial usando o programa MaSurCA resultou em um total de 212.068 *contigs* com N50 de 10.290 pb (Tabela 1). A montagem de genomas usando “*short reads*” compreende um processo computacional complexo, pois os dados gerados no sequenciamento possuem erros de identificação de bases e diversas regiões de repetição que afetam o processo. Além disso, para organismos diploides e/ou poliploides, que apresentam um genoma altamente polimórfico, os programas geralmente geram uma montagem mais fragmentada com um tamanho maior que o esperado. Isso ocorre, porque os “*short-reads*” geralmente são insuficientes para cobrir com acurácia os haplótipos diferentes em regiões heterozigotas, e essas regiões são reportadas como *contigs* alternativos. Por outro lado, as regiões homozigotas, ou com baixa heterozigozidade entre dois cromossomos homólogos, são unidas em um único *contig* (Pryszcz & Gabaldón, 2016). Levando em consideração as características do genoma da espécie *S. adstringens* (i.e. tamanho de genoma e número de elementos repetitivos), além do resultado inicial da montagem bastante fragmentado, o programa *Redundans*, foi usado para minimizar os *contigs* redundantes, reportados como alternativos. Com a identificação e remoção/união de *contigs* redundantes, o número de *contigs* diminuiu para 72.410 (Tabela 1). Observou-se, assim, uma melhora nas métricas de

avaliação da montagem, como no tamanho do maior *contig*, tamanho médio de *contigs*, no número de *contigs* maiores que 25kpb, 50kpb e N50 (Tabela 1).

O *draft* do genoma de *S. adstringens*, após remoção de sequências organelares e menores que 500pb, apresentou cerca de 76,3% (~495 Mb) do valor estimado para o genoma de *S. adstringens* realizado por citometria de fluxo. Nessa montagem, o número de *contigs* foi de 63.320, dos quais 58.873 e 16.142 foram maiores que 1kpb e 10kpb (Figura 3), respectivamente. O tamanho médio dos *contigs* foi de 6.960 pb variando de 500 pb a 138.079 pb com N50 de 14.541 pb (Tabela 1). A cobertura média estimada da montagem final foi de 152x, de acordo com o que era esperado, levando em consideração o número de pares de base que foi sequenciado.

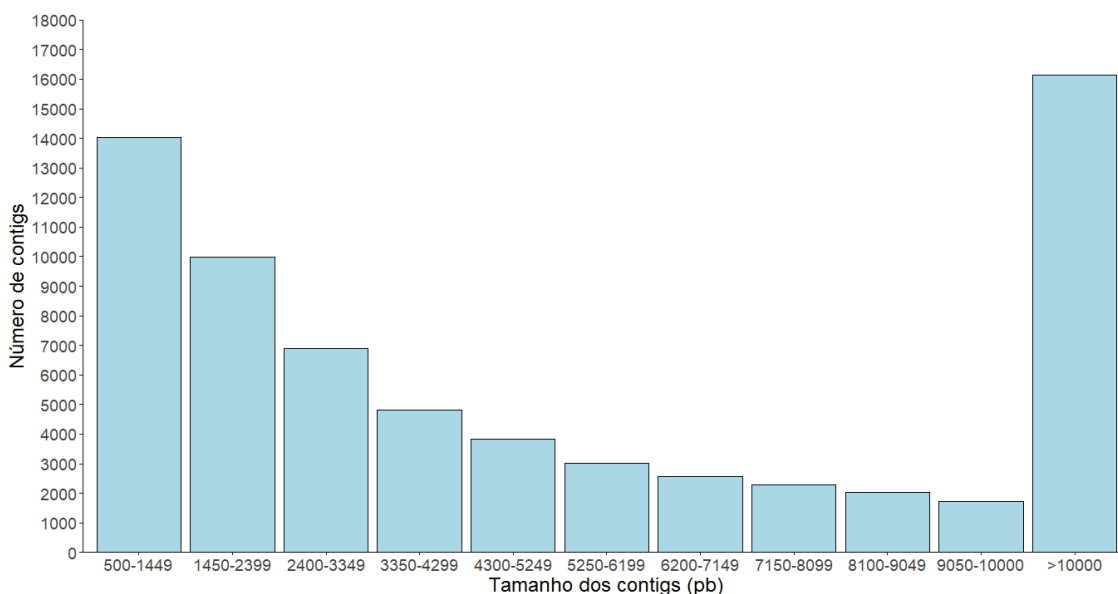


Figura 3. Distribuição do tamanho final do *contigs* obtidos na montagem “*de novo*” do genoma de *Stryphnodendron adstringens*

Tabela 1. Métricas utilizadas para avaliar a montagem do genoma da espécie *Stryphnodendron adstringens*

Métrica	MaSuRCA	MaSuRCA + Redundans	Montagem Final
Nº <i>contigs</i>	212.068	72.410	63.320
<i>contigs</i> (>= 1000 bp)	102.046	59.879	58.873
<i>contigs</i> (>= 5000 bp)	32.430	29.366	28.662
<i>contigs</i> (>= 10000 bp)	16.971	16.655	16.142
<i>contigs</i> (>= 25000 bp)	3.533	3.689	3.507
<i>contigs</i> (>= 50000 bp)	322	365	330
Tamanho médio de <i>contigs</i>	3.090	7.065	6.960
Menor <i>contig</i>	79	204	500
Maior <i>contig</i>	98.315	175.783	138.079
Tamanho total	655.224.604	511.548.318	495.022.034
% do genoma (669 Mbp)	98%	76,5%	76,3%
N50	10.290	14.755	14.541
N75	3.331	7.281	7.222
L50	16.428	10.214	10.021
L75	44.239	22.457	21.996

3.3.3 Caracterização de elementos transponíveis

A partir das análises verificou-se que 220.680.802 pb do genoma montado de *S. adstringens* é composto por elementos transponíveis (44,58%), representando cerca de metade das sequências obtidas. Do total de elementos repetitivos identificados, 187.534.038 pb (37,88%) foram anotados pelo método “*de novo*”, ou seja, são exclusivos do genoma de *S. adstringens*. Enquanto 33.146.764 pb (6,70%) foram anotados pelo método de similaridade com elementos repetitivos de eudicotiledôneas depositadas em bancos de dados.

Considerando os elementos identificados pelo método de similaridade, os elementos transponíveis da classe I, ou retrotransposons foram a classe mais abundante (74,42%) no genoma de *S. adstringens* (Tabela 2). Dentre os elementos da classe I, os LTR retrotransposons foram a ordem mais frequente, abrangendo 57,06% do total de elementos encontrados (Tabela 2), sendo predominantemente representados pelas superfamílias Ty3-

Gypsy (31,93% do total de elementos) e *Ty1-Copia* (22,61% do total de elementos). Para os elementos transponíveis da classe II ou transposons de DNA foi observada uma frequência de 21,96% entre o total de elementos identificados, enquanto que para 0,048% não foi possível determinar a classe (Tabela 2).

Recentemente, a caracterização de elementos transponíveis usando métodos “de novo” em sequências genômicas tem ganhando notoriedade como uma abordagem eficaz para identificar superfamílias de elementos ainda desconhecidas, conforme observado no sequenciamento do genoma das espécies *Glycine latifolia* (Liu et al., 2018), *Handroanthus impetiginosus* (Silva-Junior et al., 2018) e *Coffea arabica* (Tran et al., 2018).

Um total de 347.416 (64,16%) elementos transponíveis identificados pelo método “de novo” não foram agrupados em nenhuma classe. Essa grande proporção de elementos não classificados sugere que o genoma de *S. adstringens* possui elementos repetitivos únicos, que ainda não estão representados no banco de dados *RepBase*. Dentre os elementos classificados, os LTR retrotransposons foram a ordem mais frequente (18,01% do total de elementos; Tabela 3), sendo as superfamílias *Ty3-Gypsy* (10,96% do total de elementos; Tabela 3) e *Ty1-Copia* (5,55% do total de elementos; Tabela 3) os mais abundantes no genoma, assim como foi observado no método por similaridade. Os transposons de DNA foram observados com uma frequência de 11,23%, seguida dos retrotransposons sem LTR: LINEs (4,69%) e SINEs (1,03%), respectivamente.

Sequências repetitivas de DNA, principalmente os elementos transponíveis, compreendem uma grande fração do genoma de plantas e têm um papel importante na estruturação dos genomas ao longo do processo evolutivo. A enorme variação relatada no tamanho, função e estrutura de genomas entre organismos tem sido frequentemente associada ao acúmulo de sequências repetidas (Civán et al., 2011). Alguns estudos têm avaliado a existência de correlação linear entre o tamanho de genomas e número de elementos transponíveis (Bennetzen et al., 2005; Hawkins et al., 2008). Elementos transponíveis em genomas menores, tais como de *Arabidopsis thaliana* (tamanho estimado do genoma=125 Mpb) representam aproximadamente 15% (Arabidopsis Genome Initiative, 2000). Por outro lado, espécies com genomas maiores, geralmente possuem uma grande abundância de elementos repetitivos. Por exemplo, aproximadamente 85% do genoma do milho (*Zea mays*; tamanho estimado do genoma = 2.3 Gpb), é composto por elementos transponíveis (Schnable et al., 2009). Resultado semelhante foi observado para a Cevada (*Hordeum vulgare*; tamanho estimado do genoma = 5.1 Gbp), em que 58,45% do genoma é

composto por elementos transponíveis (Consortium, 2012). Dentre as espécies de leguminosas com genoma já sequenciado, *Glycyrrhiza uralensis* possui o menor tamanho estimado de genoma (400 Mpb), com aproximadamente 36,48% constituído de elementos transponíveis (Mochida et al., 2017). Por outro lado, *Arachis ipaensis* (1.56 Gpb) e *Arachis duranensis* (1.25 Gpb) apresentam os maiores tamanhos de genoma entre as leguminosas sequenciadas. Os elementos transponíveis abrangem aproximadamente 68,5% e 61,7% do genoma de *A. ipaensis* e *A. duranensis*, respectivamente (Bertioli et al., 2016). Os resultados observados para o genoma de *S. adstringens*, com 44,58% das sequências de DNA compostas de elementos transponíveis, se apresenta intermediário entre essas espécies, e corrobora com a hipótese de correlação entre tamanho de genoma e número de elementos repetitivos.

No genoma de *S. adstringens*, os retrotransposons foram a classe predominante de elementos transponíveis dentre os identificados. Os retrotransposons são conhecidamente os elementos transponíveis mais numerosos, na maioria dos genomas de plantas (Chaux et al., 2012). Nas espécies *Glycyrrhiza uralensis*, *Hordeum vulgare*, *Arachis ipaensis*, *Arachis duranensis* e *Zea mays* os retrotransposons foram os mais abundantes e representam 11,9%, 52,7%, 57,6% e 52,6%, 75,6% (Schnable et al., 2009; Bertioli et al., 2016; Mochida et al., 2017). Esses elementos são os que possuem maior influência na variação do tamanho do genoma em plantas, devido sua transposição replicativa, por meio do mecanismo de “cópia e colagem” (Sahebi et al., 2018). Além disso, exercem função no rearranjo genômico, podendo gerar novos pontos de recombinação, duplicações, deleções, inversões e translocações, além de ativar genes previamente inativos ou alterar seu padrão de expressão (Sahebi et al., 2018). Assim, a caracterização e conhecimento desses elementos mostram-se importantes na imediata aplicação de montagem e anotação do genoma da espécie, bem como para futuros estudos de filogenia e aplicação como marcadores moleculares.

A maior superfamília dos retrotransposon são Ty1-*Copia* e Ty3-*Gypsy*. Ambos os grupos são amplamente encontrados no reino vegetal, desde briófitas a angiospermas (Kapitonov & Jurka, 2007; Civián et al., 2011). Transposons de DNA são também encontrados em genomas de plantas, contudo em frequências menores que os retrotransposons (Civián et al., 2011; Sahebi et al., 2018). Nesse estudo, foi observado que os retrotransposons com LTRs, são predominantes no genoma de *S. adstringens*, principalmente a superfamília Ty3-*Gypsy*. A abundância de retrotransposons tem sido descrito no genoma de diversas outras espécies de leguminosas. A superfamília Ty3-*Gypsy*

também foram a superfamília mais abundante nos genomas das seguintes espécies de leguminosas: *Lotus japonicus* (Sato et al., 2008), *Medicago truncatula* (Young et al., 2011), *Glycine max* (Schmutz et al., 2010), *Cajanus cajan* (Varshney et al., 2012), *Phaseolus vulgaris* (Schmutz et al., 2014), *Vigna radiata* (Kang et al., 2014), *Vigna angularis* (Kang et al., 2015), *Arachis duranensis* (Bertioli et al., 2016) e *Ammopiptanthus nanus* (Gao et al., 2018). Enquanto que Ty1-Copia foram mais abundantes em quatro espécies: *Trifolium pratense* (De Vega et al., 2015), *Trifolium subterraneum* (Hirakawa et al., 2016), *Arachis ipaensis* (Bertioli et al., 2016) e *Glycyrrhiza uralensis* (Mochida et al., 2017).

A prevalência de um grupo específico de elementos transponíveis pode estar relacionado com eventos de amplificação de linhagens específicas de determinada superfamília e da atividade de transposição de tais elementos (Feschotte et al., 2002; Sahebi et al., 2018). Por exemplo, as espécies do gênero *Gossypium*, com tamanhos de genomas menores (i.e. *Gossypium raimondii*; ~880 Mpb) possuem uma alta densidade de elementos Ty1-Copia, enquanto, as espécies com tamanhos de genomas maiores (i.e. *Gossypium herbaceum*; ~1.7 Gb e *Gossypium exiguum*; ~2.5 Gb) possuem um aumento no número de cópias de elementos Ty3-Gypsy, sendo a maioria representadas por somente uma linhagem, Gorge3. Assim, tais eventos de amplificação podem ser linhagem-específico, os quais contribuem com grande parte das variações no tamanho do genoma de plantas (Hawkins et al., 2006; Hu et al., 2010).

Tabela 2. Anotação do total de elementos transponíveis no genoma de *Stryphnodendron adstringens* usando o *RepeatMasker*.

Classe	Ordem	Superfamília	Número de elementos	% de elementos	Tamanho ocupado no genoma (pb)	% de sequências no genoma
Classe I (Retrotransposons)	SINEs		709	0,945	76345	0,015
	LINEs	CRE	2	0,003	98	0,00002
		L1	5667	7,551	2130616	0,430
		RTE-BovB	6053	8,066	2428460	0,491
	LTR	Cassandra	581	0,774	31169	0,006
		Caulimovirus	1302	1,735	963789	0,195
		Copia	16972	22,615	12206207	2,466
		Gypsy	23964	31,932	9073314	1,833
	Não classificados		599	0,798	46058	0,009
Classe II (DNA Transposons)	TIRs	CMC-EnSpm	5377	7,165	2621521	0,530
		MULE-MuDR	2771	3,692	481233	0,097
		PIF-Harbinger	1606	2,140	495555	0,100
		TcMar-Pogo	5	0,007	297	0,0001
		TcMar-Stowaway	91	0,121	23628	0,005
		hAT	1	0,001	50	0,00001
		hAT-Ac	4138	5,514	845602	0,171
		hAT-Charlie	151	0,201	13169	0,003
		hAT-Tag1	1966	2,620	442079	0,089
		hAT-Tip100	218	0,290	36566	0,007
	Não classificados		156	0,208	12579	0,003
Rolling-circles	Helitron		2681	3,572	1214383	0,245
Não classificados			36	0,048	4046	0,001

Tabela 3. Anotação do total de elementos transponíveis no genoma de *Stryphnodendron adstringens* usando o *RepeatModeler*.

Classe	Ordem	Superfamília	Número de elementos	% de elementos	Tamanho ocupado no genoma (pb)	% de seqüências no genoma
Classe I (Retrotransposons)	SINE		5598	1,034	962072	0,002
	LINEs	L1	15167	2,801	7863061	0,016
		L1-Tx1	3567	0,659	1278432	0,003
		L2	4444	0,821	2273382	0,005
		RTE-BovB	2241	0,414	395020	0,001
	LTR	Cassandra	6235	1,152	3496838	0,007
		Caulimovirus	993	0,183	454353	0,001
		Copia	30069	5,553	13686903	0,028
		ERVk	258	0,048	34561	0,000
		Gypsy	59326	10,957	43015209	0,087
		Ngaro	582	0,107	67542	0,0001
		Pao	34	0,006	4398	0,00001
	Não classificados		2001	0,370	505245	0,001
Classe II (DNA Transposons)	TIR	CMC-EnSpm	26870	4,963	11217677	0,023
		IS3EU	1054	0,195	397996	0,001
		Kolobok-T2	587	0,108	100141	0,0002
		MULE-MuDR	3874	0,715	1173092	0,002
		MuLE-MuDR	2018	0,373	935461	0,002
		P	754	0,139	122651	0,000
		PIF-Harbinger	6807	1,257	1514129	0,003
		TcMar-Mariner	130	0,024	50265	0,0001
		TcMar-Stowaway	196	0,036	27295	0,0001
		Zisupton	824	0,152	111510	0,0002
		hAT-Ac	11280	2,083	2848254	0,006
		hAT-Charlie	634	0,117	113055	0,000

	hAT-Tag1	3735	0,690	899841	0,002
	hAT-Tip100	98	0,018	9094	0,00002
	Não classificados	1939	0,358	461988	0,001
Rolling-circles	Helitron -	2716	0,502	941641	0,002
Não Classificados		347416	64,164	92572932	0,187

3.3.4 Caracterização de regiões microssatélites

Do total de *contigs* obtidos pela montagem de novo do genoma de *S. adstringens*, 14.453 apresentaram regiões microssatélites, dentre os quais 3.957 (27,38%) possuem duas ou mais regiões. Um total de 20.124 microssatélites foram identificados nos *contigs*. Os microssatélites mais frequentes foram aqueles contendo repetições do tipo dinucleotídeos (84,03%), seguidos pelo trinucleotídeos (9,05%) e tetranucleotídeos (4,37%). Os microssatélites baseados em repetições do tipo pentanucleotídeos (1,60%) e hexanucleotídeos (0,94%) foram os menos abundantes (Figura 4).

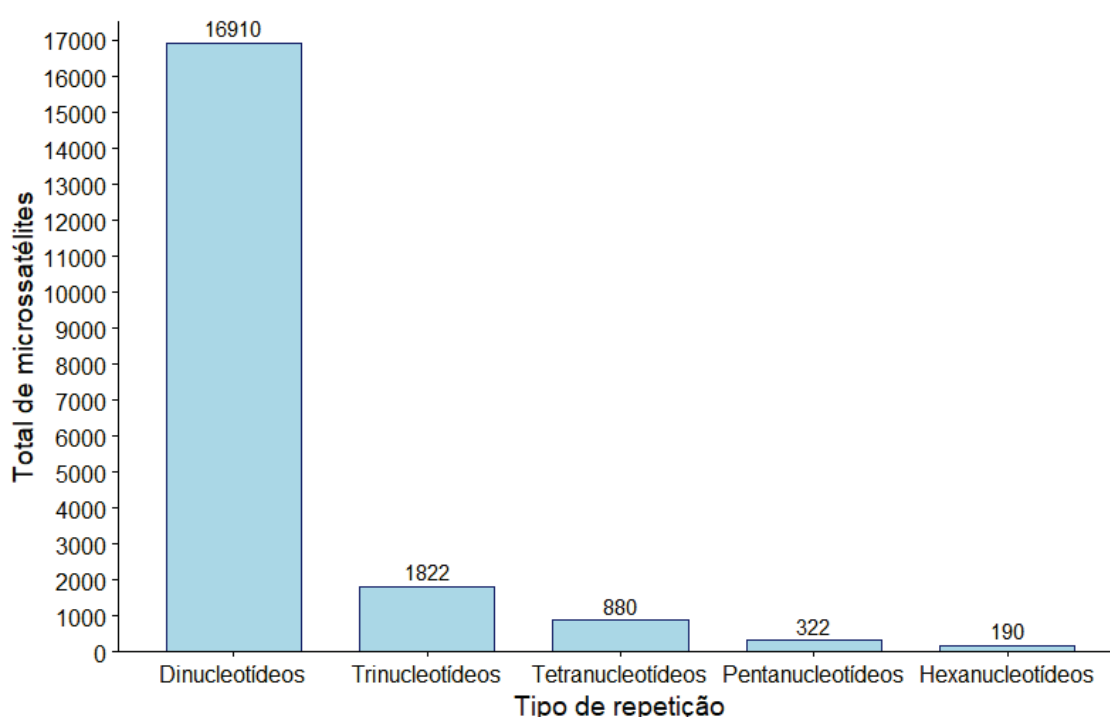


Figura 4. Distribuição do total de microssatélites encontrados no genoma da espécie *Stryphnodendron adstringens*.

Nesse estudo, não foram conduzidas buscas de repetições mononucleotídicas, devido à sua grande prevalência em genoma eucarióticos quando comparadas com os outros tipos de repetição, bem como uma predominância de motivos poli-(A/T) (Tóth et al., 2000; Sharma et al., 2007). Além disso, sequências contendo repetições em tandem de um único nucleotídeo tendem a formar *stutters*, um dos principais complicadores da análise dessas regiões no contexto de marcadores moleculares. Portanto, a identificação

de microssatélites mononucleotídeos não se mostraria útil para o futuro desenvolvimento de ferramentas moleculares para a espécie.

Cavagnaro et al. (2010) realizou uma análise de microssatélites em diferentes espécies de plantas, e observou que, repetições de trinucleotídeos foram as mais abundantes em *Arabidopsis*, soja, arroz e sorgo, enquanto repetições de tetranucleotídeos foram mais abundantes em *Populus* e uva. Neste estudo, repetições de dinucleotídeos foram as mais abundantes, abrangendo 84,03% do genoma. Desconsiderando os microssatélites do tipo mononucleotídeos, uma maior proporção de dinucleotídeos, conforme observado nesse estudo, também foi encontrado no genoma de outras espécies da família leguminosae, tais como, Feijão-guandu (*Cajanus cajan*; Varshney et al., 2012), Grão-de-bico (*Cicer arietinum*; Varshney et al., 2013) e Feijão-Mungo (*Vigna radiata*; Kang et al., 2014). Tais observações sugerem claramente que os tipos de repetição mais abundantes de microssatélites no genoma de plantas variam consideravelmente entre diferentes espécies.

Os microssatélites identificados foram repetidos em tandem entre 5 e 45 vezes, com o tamanho variando de 19 e 110 pares de base. Considerando o tamanho estimado do genoma em 669 Mpb, a densidade de microssatélites nas sequências de DNA de *S. adstringens* é de ~30,1 microssatélites por Mb. Essa densidade de microssatélites observada para *S. adstringens* é bastante inferior ao que tem sido descrito para outras espécies de plantas. Por exemplo, a densidade de microssatélites no genoma de *Arabidopsis*, arroz e pepino foram de 874, 807 e 551 microssatélites por Mpb, respectivamente (Lawson & Zhang, 2006; Cavagnaro et al., 2010). Isso se deve principalmente aos parâmetros mais rigorosos para a busca dos microssatélites usados neste estudo, do que os usados nesses trabalhos. Portanto, vale ressaltar que, a frequência, a distribuição e a abundância dos microssatélites descrita nos genomas são também dependentes dos critérios de busca, do tamanho do conjunto de dados e da ferramenta de bioinformática utilizada na mineração de dados.

Ao analisar o tipo de repetição pelo número de repetições em tandem, foi possível verificar que o número de microssatélites identificados tende a diminuir à medida que o número de unidades repetidas aumenta (Figura 5). Isso foi observado principalmente para as repetições do tipo pentanucleotídeo e hexanucleotídeo, em que o número de repetições em tandem não foi maior que 10 e 13 pares de base, respectivamente (Figura 5). Enquanto para as repetições do tipo dinucleotídeo, trinucleotídeo e

tetranucleotídeo foram encontrados microssatélites com número de repetições em tandem maiores que 20 pares de base (Figura 5).

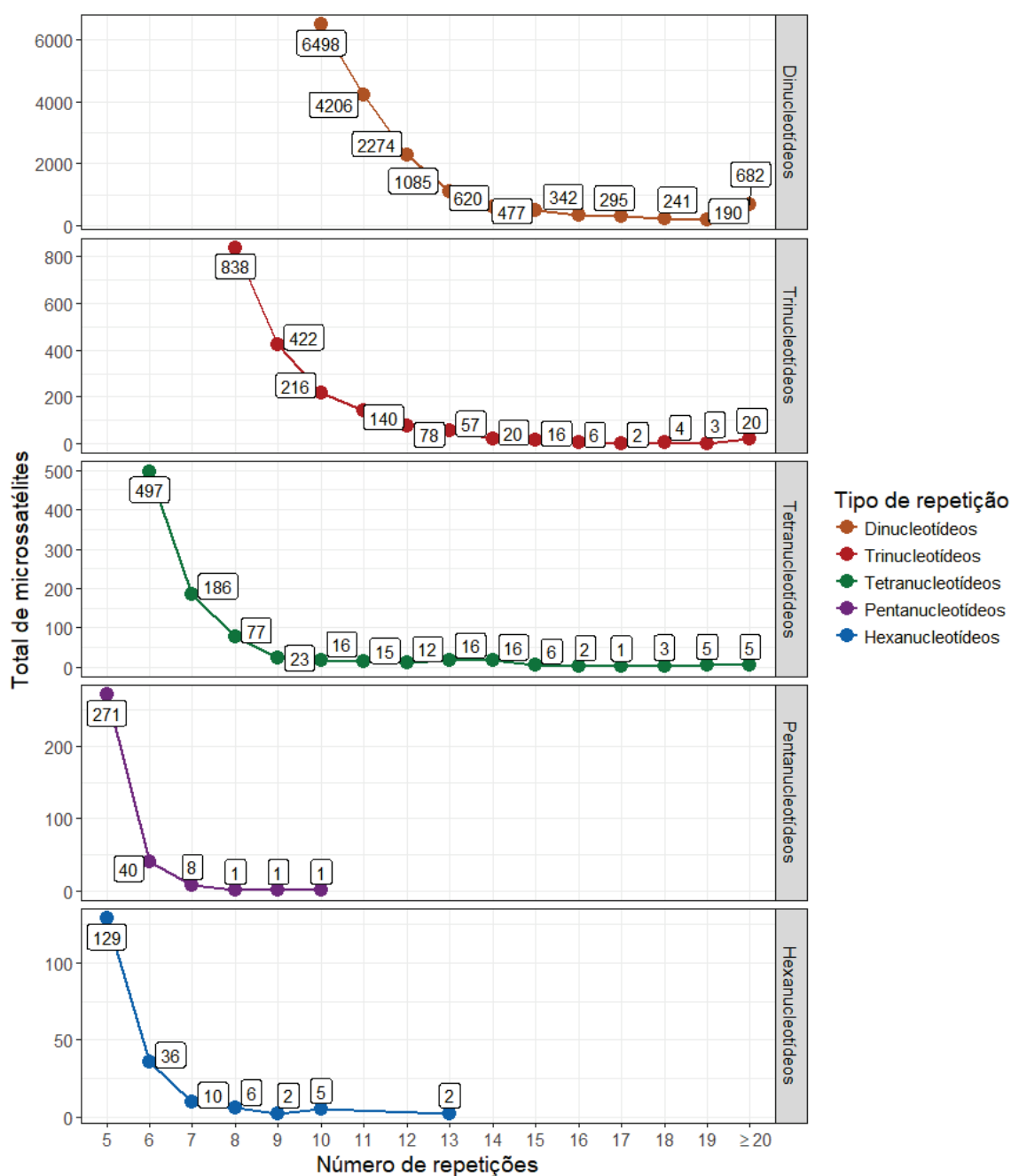


Figura 5. Número de repetições em *tandem* para cada um dos motivos de repetição encontrados no genoma da espécie *Stryphnodendron adstringens*.

Os microssatélites com repetições do tipo dinucleotídeos compostos por motivos AT/TA foram os encontrados com maior frequência (44,16%) entre o total de microssatélites identificados no genoma, seguidos por AG/TC (12,44% do total) e CT/GA

(11,78% do total) (Tabela suplementar 1). Em relação aos trinucleotídeos, os motivos AAT/TTA (1,90%), ATT/TAA (1,19%) e AAC/TTG (1,01%) foram os mais frequentes (Tabela suplementar 1). Os motivos mais frequentes de tetranucleotídeos foram AAAT/TTTA (0,56%), ATAC/TATG (0,52%) e AAAG/TTTC (0,39%), de pentanucleotídeos, AAAAG/TTTTTC (0,24%), AAAAT/TTTTA (0,20%) e AACAT/TTGTA (0,10%) e para hexanucleotídeos AAAAAT/TTTTTA (0,02%) e CTTCTC/GAAGAG (0,02%) (Tabela suplementar 1).

As repetições mais abundantes no genoma de plantas são geralmente ricas em AT, enquanto motivos de repetição ricas em GC são menos frequentes (Morgante & Olivieri, 1993; Kalia et al., 2011). Uma grande proporção de repetições ricas em AT foi encontrado no genoma de *S. adstringens*. Por exemplo, as repetições AT/TA foram predominantes entre os dinucleotídeos, bem como, entre todas repetições identificadas, enquanto, somente 6 repetições GC/CG foram encontradas. Da mesma forma, para tri, tetra, e hexanucleotídeos repetições ricas em AT, foram as mais abundantes. Este resultado, se mostra de acordo com observações anteriores de que microssatélites ricos em AT são predominantes em algumas dicotiledôneas como *Arabidopsis* (Tamanna & Khan, 2005), mamão (Lai et al., 2006), soja (Shultz et al., 2007) e pepino (Cavagnaro et al., 2010), enquanto regiões ricas em GC são mais predominantes em monocotiledôneas, e abundantes em regiões codificantes.

Um total de 38.482 pares de primers foram desenhados para 1.739 regiões, o que corresponde em média 22 pares de primers por região. Essas regiões consistem de sequências únicas (*singleton*) identificadas pelo *pipe2.pl* do programa QDD e que estavam de acordo com os parâmetros *in silico* para desenho dos primers (*pipe3.pl*). Dentre as 1.739 regiões, 1.039 apresentaram similaridade com outros tipos de elementos repetitivos (transposons de DNA, retrotransposons, snRNA, tRNA e entre outros), o que resultou na exclusão de 25.389 pares de primers.

Para o conjunto de 700 regiões (13.093 pares de primers) sem associação com outros tipos de elementos repetitivos, 537 eram dinucleotídeos, 89 trinucleotídeos, 45 tetranucleotídeos, 22 pentanucleotídeos e 7 hexanucleotídeos. Para seleção dos melhores primers foram adotados alguns critérios sugeridos pelo programa QDD: 1. foram descartadas as regiões ricas em AT/TA com o intuito de evitar a formação de estruturas secundárias denominadas “*hairpins*”, as quais podem afetar o anelamento do primer e consequentemente a amplificação. 2. Visando o futuro desenvolvimento de marcadores genéticos neutros, os motivos do tipo trinucleotídeo e hexanucleotídeo, foram excluídos.

Motivos do tipo di-, tetra- e pentanucleotídeos são mais abundantes em regiões não codificantes do genoma, enquanto, as repetições do tipo tri- e hexanucleotídeos são mais abundantes em regiões codificantes, regiões sob pressão de seleção (Morgante et al., 2002). 3. Foi mantido somente um *primer* por *contig*, levando em consideração a seleção automática do programa. 4. foram escolhidos preferencialmente os microssatélites com um maior número de repetição, nas quais o primer estava a uma distância de no mínimo 20 pares de base do microssatélite. A partir desses critérios, foram mantidos 188 pares de primers com alto potencial para o desenvolvimento de marcadores moleculares, os quais poderão ser testados quanto sua aplicabilidade, e usados em estudos genético-populacionais futuros para a espécie *S. adstringens*.

Recentemente, Barateli (2018) testou 20 pares de *primers* do total de 188 que foram mantidos. Entre os 20 pares de *primers*, 16 apresentaram produtos de amplificação polimórficos e quatro monomórficos, disponibilizando um painel de marcadores informativo e indicado para estudos genéticos populacionais para a espécie (Barateli, 2018). O uso dessas regiões microssatélites espécie-específico como marcadores moleculares, podem complementar informações genéticas já existentes, obtidas com trabalhos anteriores que utilizaram marcadores RAPD (Camilo et al., 2001), Isoenzimas (Glasenapp et al., 2012, 2014), AFLP (Mendonça et al., 2012) e microssatélites transferidos de outras espécies arbóreas (Branco et al., 2010; Mendonça, 2011).

A plataforma de sequenciamento *Illumina* juntamente com o programa QDD, foram utilizados recentemente com o objetivo de desenvolver e caracterizar marcadores microssatélites para diversas espécies de plantas, tais como, *Terminalia superba* (Demenou et al., 2015), *Acacia adinophylla* (Nevill & Wardell-Johnson, 2016), *Betula chichibuensis* (Igarashi et al., 2017), *Anthonotha macrophylla* (Demenou & Hardy, 2017), *Guibourtia ehie* (Tosso et al., 2017), *Primula tosaensis* (Yamamoto et al., 2018). Todos estes trabalhos disponibilizaram um conjunto de marcadores microssatélites para uso em estudos de genética populacional destas espécies.

3.3.5. Predição e anotação de genes

3.3.5.1 Predição de genes

A partir do número total de *contigs* (63.320) obtidos com a montagem do genoma de *S. adstringens*, em 20.654 *contigs* (32,62%) foram encontrados genes

putativos codificantes de proteínas. O número médio de genes por *contigs* foi de 1,53 com um número máximo de 15 genes por *contig*, com 70,44% dos *contigs* (14.548) apresentando apenas 1 gene (Figura 6). Considerando a estimativa do tamanho do genoma da espécie (669 Mpb), observa-se a ocorrência de 1 gene a cada 21 kb.

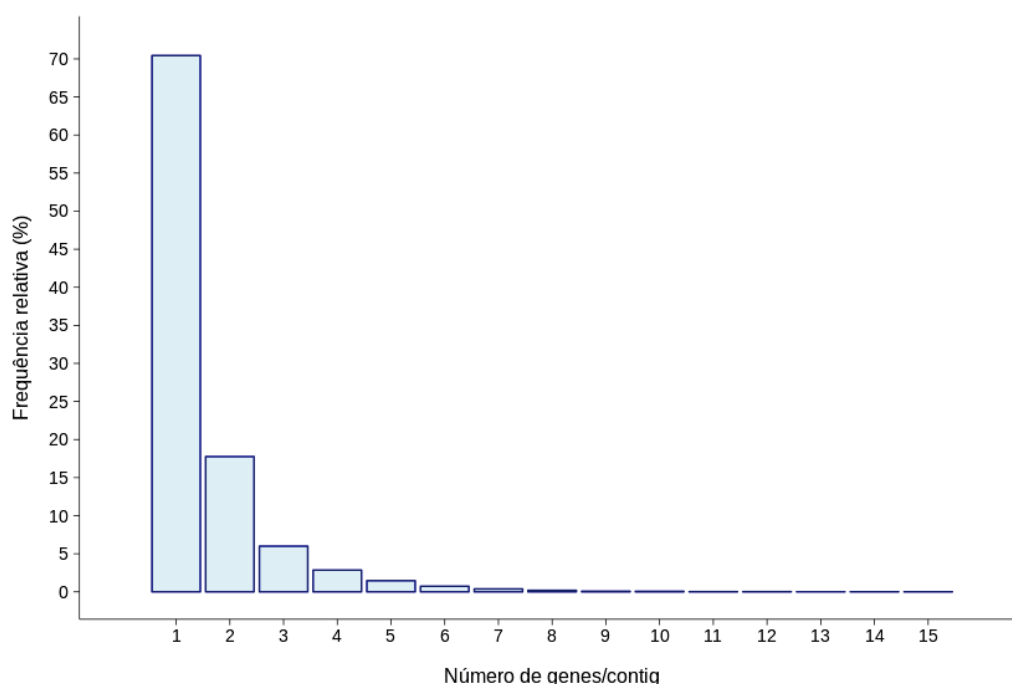


Figura 6. Frequência relativa da ocorrência de genes de acordo com o número de *contigs* obtidos com a montagem do genoma de *Stryphnodendron adstringens*.

O número de genes codificantes de proteínas encontrados no genoma de *S. adstringens* abrange cerca de 20,17% do genoma (669 Mpb). A proporção do genoma de plantas que é composto por genes codificantes de proteínas varia consideravelmente entre diferentes táxons. Por exemplo, a espécie *Picea abies*, que tem uma estimativa de tamanho de genoma de ~19.6 Gb, possui aproximadamente 2,4% do seu genoma constituído por genes codificantes de proteínas (Nystedt et al., 2013). Enquanto a espécie *Utricularia gibba*, que tem uma estimativa de tamanho de genoma de ~83 Mpb, possui aproximadamente 98% do seu genoma constituído por genes codificantes de proteínas (Ibarra-Laclette et al., 2013).

Um total de 31.636 genes putativos codificantes de proteínas foram preditos no genoma de *S. adstringens*. Usualmente, o número de genes codificantes de proteínas nos genomas de plantas varia entre aproximadamente 20.000 e 45.000 e as diferenças que são encontradas no número de genes entre espécies da mesma família, geralmente é resultado de perdas ou ganhos (eventos de duplicação) de genes que ocorreram no decorrer do

processo evolutivos das espécies (Sterck et al., 2007). Considerando as espécies da família Leguminosae o número de genes observados no genoma de *S. adstringens* foi maior do que o encontrado para as espécies *Lablab purpureus* (Chang et al., 2018), *Phaseolus vulgaris* (Schmutz et al., 2014), *Faidherbia albida* (Chang et al., 2018) e *Lotus japonicus* (Sato et al., 2008) com 20.946, 27.197, 28.979 e 30.799 genes, respectivamente. Por outro lado, é menor do que foi relatado para as espécies *Ammopiptanthus nanus* (37.173; Gao et al., 2018) *Glycyrrhiza uralensis* (38.135; Mochida et al., 2017) e *Medicago truncatula* (47.845; Young et al., 2011).

O tamanho médio dos genes identificados foi de 2.910 pb, com tamanho mínimo e máximo de 130 pb e 39.841 pb respectivamente (Figura 7). O tamanho médio dos genes é bastante similar ao que foi observado para *Medicago truncatula* (2.211 pb; Young et al., 2011) e *Lotus japonicus* (2.917 pb; Sato et al., 2008). Os 31.636 genes putativos são compostos por 709.877 íntrons, com tamanho médio de 338 pb, sendo que 94,04% possuem tamanho entre 100 e 1000 pb. Por outro lado, o número de éxons foi de 814.014, com tamanho médio de 213 pb, dentre os quais, 91,13% são compostos por até 500 pb.

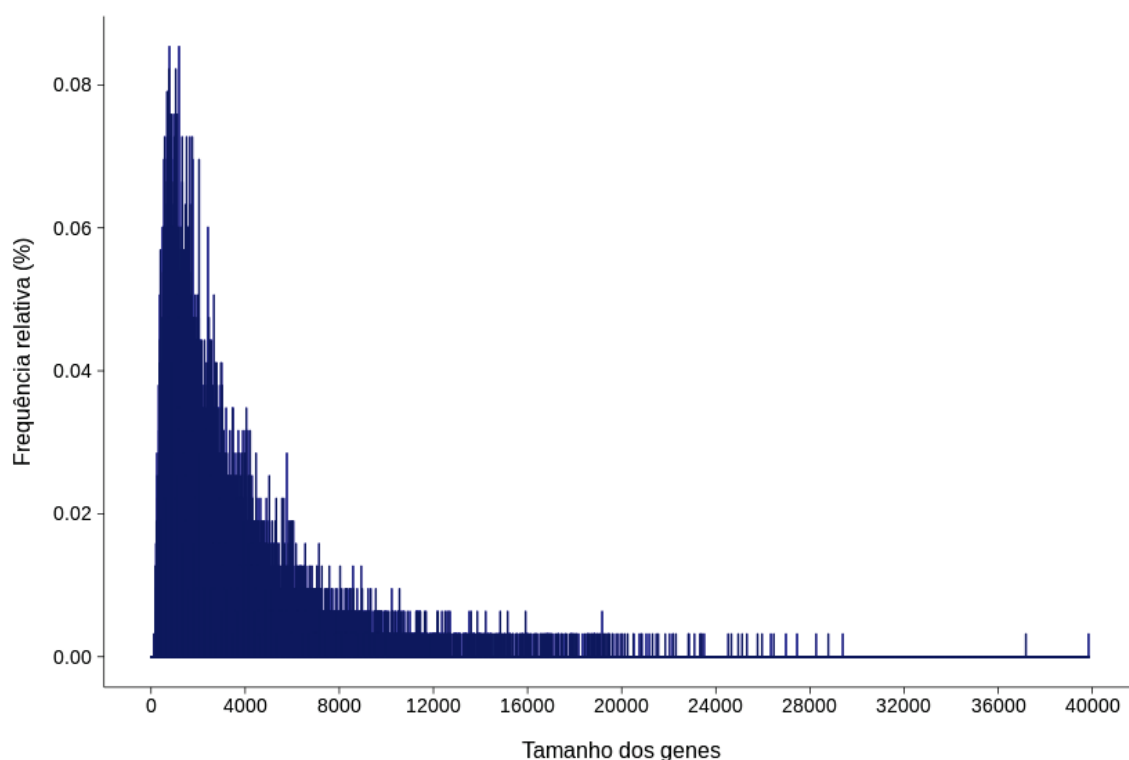


Figura 7. Distribuição da frequência relativa dos genes preditos na montagem do genoma de *Stryphnodendron adstringens* de acordo com seus tamanhos em pares de base.

A partir do total de genes putativos preditos no genoma de *S. adstringens*, foram observados um total de 95.811 transcritos, com uma média de 3,03 transcritos por gene, dentre os quais, 55,76% (17.640) dos genes possuem apenas um transcrito (Figura 8). O tamanho médio dos transcritos observados no genoma foi de 4.603 pb, sendo que 48,98% possuem tamanho entre 130 e 4.000 pb (Figura 9).

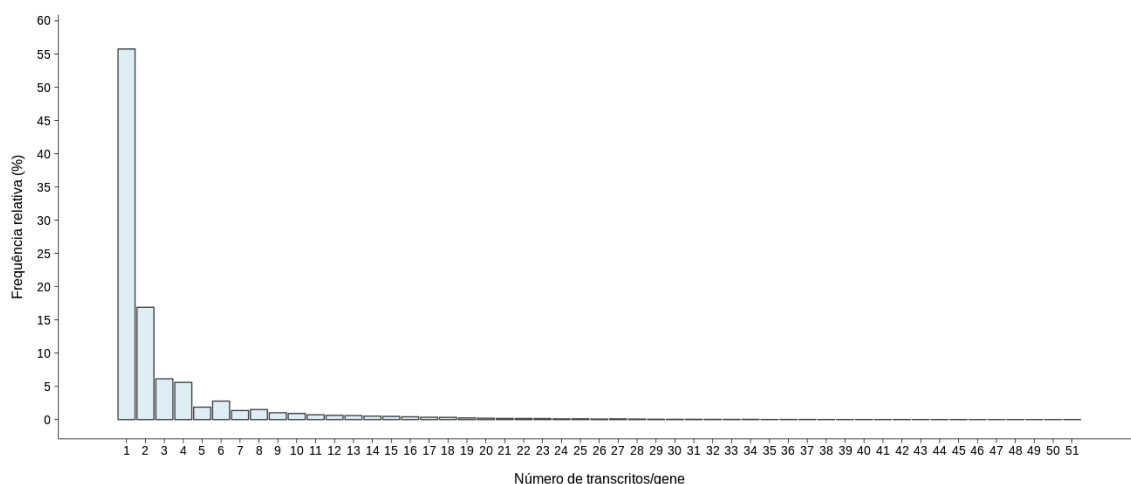


Figura 8. Distribuição da frequência relativa do número de transcritos por gene na montagem do genoma de *Stryphnodendron adstringens*

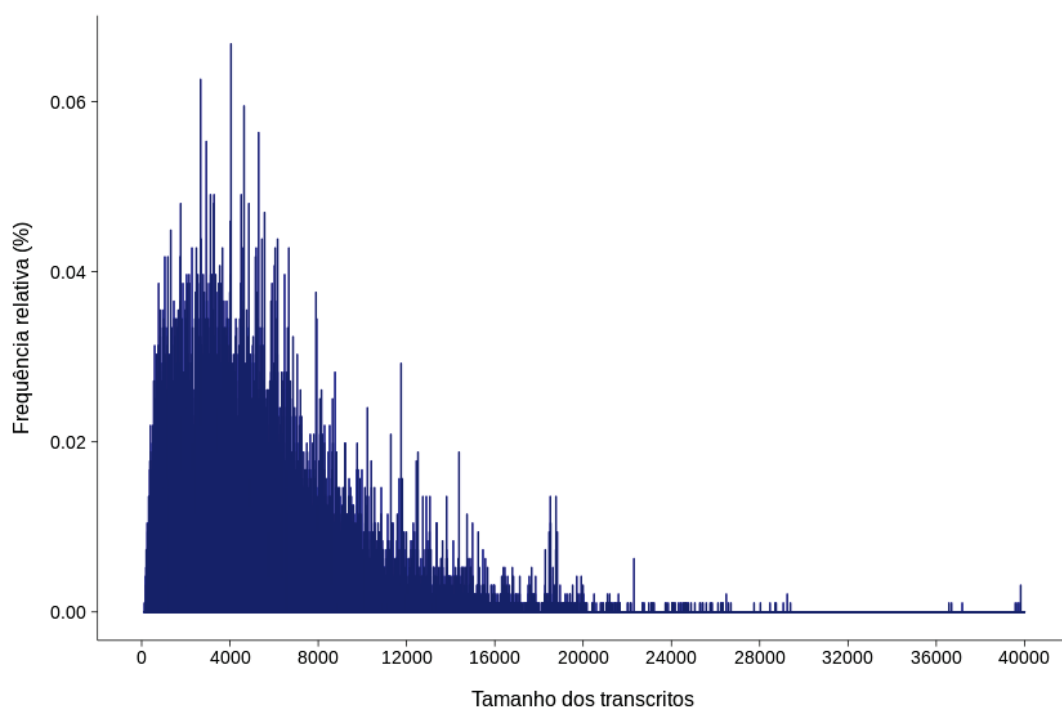


Figura 9. Distribuição da frequência relativa do tamanho dos transcritos na montagem do genoma de *Stryphnodendron adstringens*.

3.3.5.2 Anotação funcional dos transcritos de *Stryphnodendron adstringens*

A busca por similaridade utilizando o programa BLASTP entre as sequências de aminoácidos codificadas pelos transcritos de *S. adstringens* e o banco de dados de proteínas de outras espécies de embriófitas, resultou na identificação de 86.892 transcritos (90,69% do total de transcritos) com similaridade em sequências já conhecidas (Figura 10). As espécies *Glycine max*, *Cicer arietinum*, *Phaseolus vulgaris*, *Medicago truncatula* foram as que apareceram com maior frequência entre os Top-Hits do BLAST (Figura 11).

Todas as espécies com maior frequência de Top-Hits são representantes da família Leguminosae, e, portanto, são próximas do ponto de vista evolutivo de *S. adstringens*. Espécies pertencentes a grupos taxonômicos próximos geralmente apresentam uma alta similaridade genômica, com sequências gênicas conservadas e elevada sintenia (Young & Bharti, 2012).

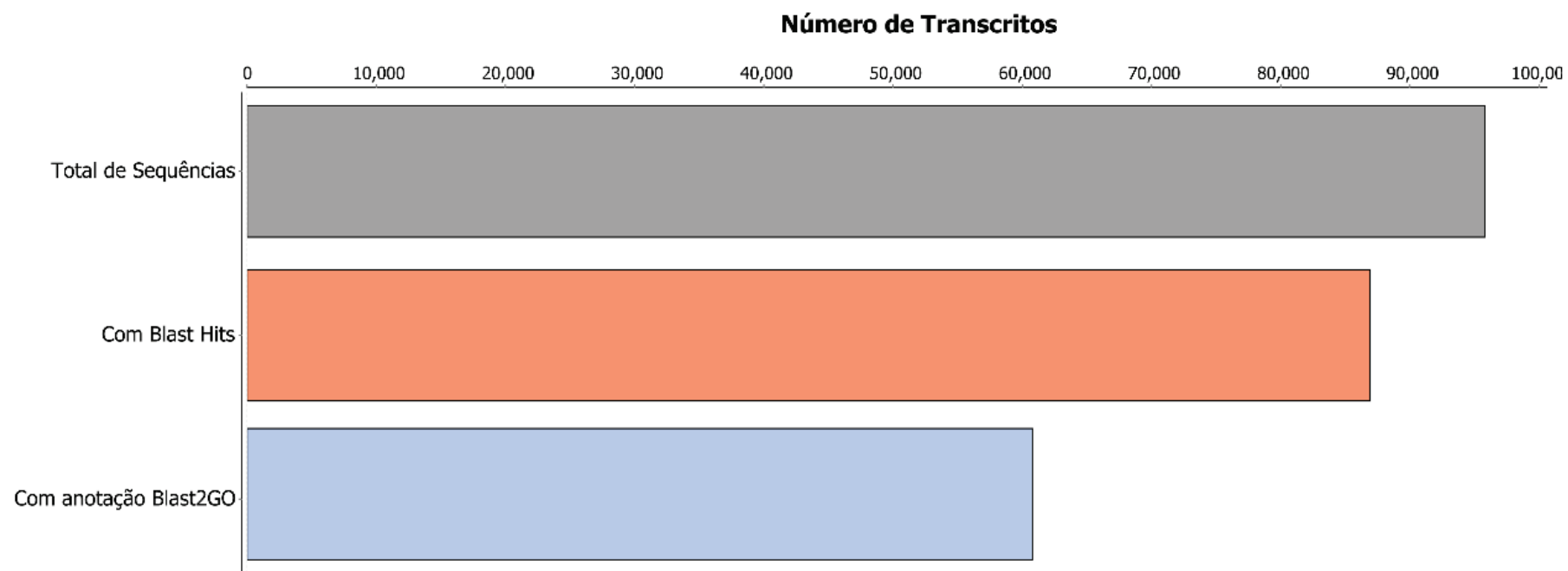


Figura 10. Número de transcritos com BLAST-HITS e com anotação nos termos *Gene Ontology*.

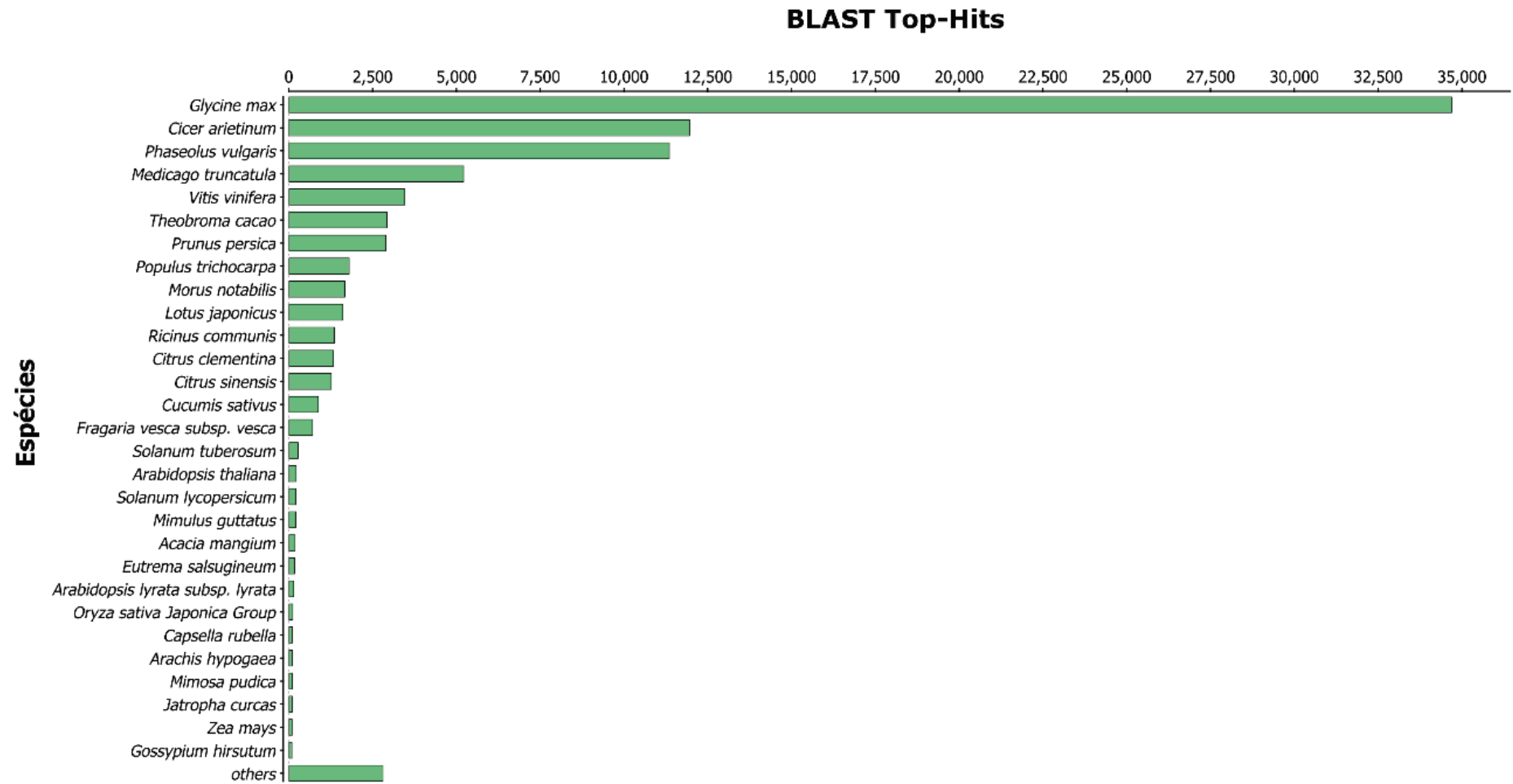


Figura 11. Distribuição das espécies Top-Hits do Blast.

As análises realizadas com o Blast2GO das sequências que apresentaram similaridades com proteínas de outras espécies de embriófitas por meio do BLASTP, resultaram na designação de pelo menos um termo de ontologia gênica (do inglês, *Gene Ontology term* – *GO term*) em 60.798 transcritos (63,46% do total de transcritos) que foram preditos no genoma de *S. adstringens* (Figura 10). O maior número de anotações nos termos GO foi obtido para a categoria “Função Molecular” (49.446 transcritos), seguido das categorias “Processos Biológicos” (44.592 transcritos) e “Componentes Celulares” (40.960 transcritos).

A Ontologia gênica define o universo dos conceitos para funções de genes (“termos GO”) e como essas funções estão relacionadas entre si. O banco de dados do *Gene Ontology* é constantemente revisado e expandido à medida que o conhecimento biológico se acumula (Consortium, 2004). Com isso, ele se apresenta como um recurso valioso para os estudos de anotação de genomas, uma vez que, fornece informações de atividades moleculares de genes, incluindo sua função molecular, seu papel em um processo biológico e sua localização em um componente celular. Além disso, geralmente podem atribuir múltiplos termos GO para uma única sequência *query* (Consortium, 2004).

Considerando a anotação nos termos GO para “Função Molecular”, a maioria dos genes preditos no genoma de *S. adstringens* estão envolvidos em atividades catalíticas (34.261 transcritos ou 56,35% dos transcritos com anotação GO) e ligação (27.647 transcritos ou 45,47% dos transcritos com anotação GO) (Figura 12). Li et al. (2015) utilizando duas espécies do gênero *Phaseolus* (*Phaseolus vulgaris* e *Phaseolus lunatus*) demonstraram resultados similares ao observado em *S. adstringens* para “Função Molecular”, em que, a maior parte dos genes categorizados estão envolvidos em atividades de ligação e catalítica.

Entre os termos GO para “Processos Biológicos”, os mais frequentes foram processos metabólicos (31.983 transcritos ou 52,60% dos transcritos com anotação GO) e processos celulares (31.210 transcritos ou 51,34% dos transcritos com anotação GO) (Figura 12). Resultados similares foram descritos por Han et al. (2016), em trabalhos com transcriptoma de *Gleditsia sinensis* em que, 49,1% e 48,1% dos genes anotados estão relacionados com processos celulares e processos metabólicos, respectivamente.

Na categoria “Componente Celular” pode-se verificar que a maior parte dos transcritos foram anotados como pertencentes às categorias ontológicas relacionadas com a estrutura da célula (célula: 28.603 transcritos ou 47,05% dos transcritos com anotação GO; parte celular: 28.486 transcritos ou 46,85% dos transcritos com anotação GO) e da

membrana (14.749 transcritos ou 24,26 dos transcritos com anotação GO) (Figura 12). Esses resultados também são similares ao obtido recentemente por Zhang et al. (2018) e Han et al. (2016) no transcriptoma das espécies *Trifolium repens* e *Gleditsia sinensi*, respectivamente. Em ambos os trabalhos para a categoria “Componente Celular”, os três principais termos GO observados estão relacionadas com a célula, parte celular e membrana respectivamente.

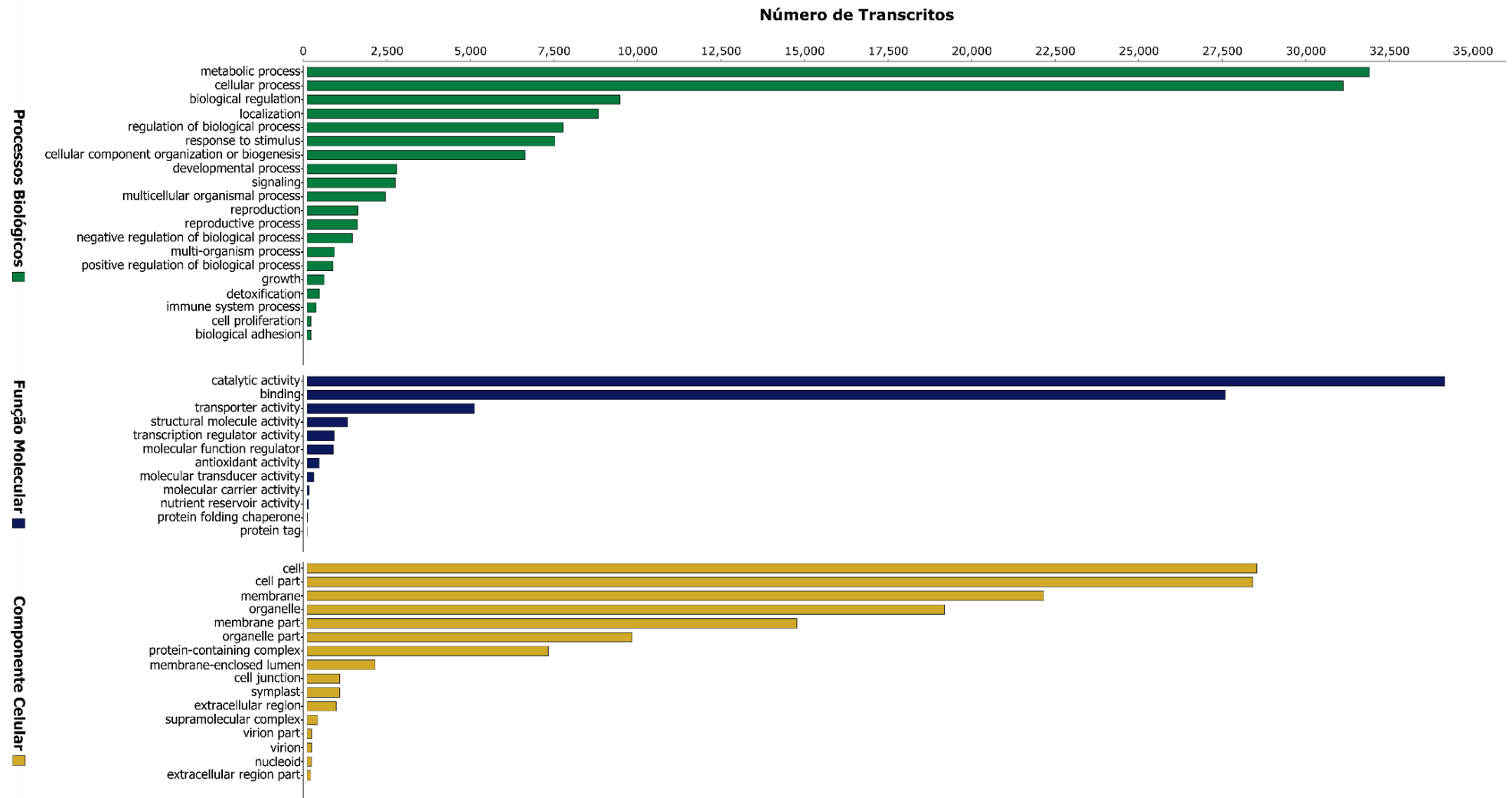


Figura 12. Anotação funcional dos transcritos encontrados no genoma de *Stryphnodendron adstringens*.

Do total de 95.811 transcritos encontrados no genoma de *S. adstringens*, 1.521 transcritos (1,58%) demonstraram similaridade com sequências de proteínas envolvidas nas vias de biossíntese de taninos em outras espécies. A maior parte dos transcritos encontrados (771 transcritos) estão relacionados com vias de biossíntese de flavonas e flavonóis (*Flavonoid 3'-monooxygenase*=516; *4-coumarate--CoA ligase*=191; *Quercetin 3-O-methyltransferase*=64), seguido das vias de biossíntese de flavonoides com 431 transcritos (*Leucoanthocyanidin dioxygenase*=249; *Bifunctional dihydroflavonol 4-reductase/flavanone 4-reductase*=144; *Leucoanthocyanidin reductase*=38), vias de biossíntese de antocianinas com 256 transcritos (*Anthocyanidin 3-O-glucyltransferase*=248; *Phenylalanine ammonia-lyase*=8) e por fim com a via de biossíntese de isoflavonoides com 63 transcritos (*2-hydroxyisoflavanone dehydratase*=63) (Figura 13; Veja material suplementar disponível em: https://www.dropbox.com/sh/x95wr9vwwgceqq/AABYATuwlTNn2fkb9jLhu_vta?dl=0).

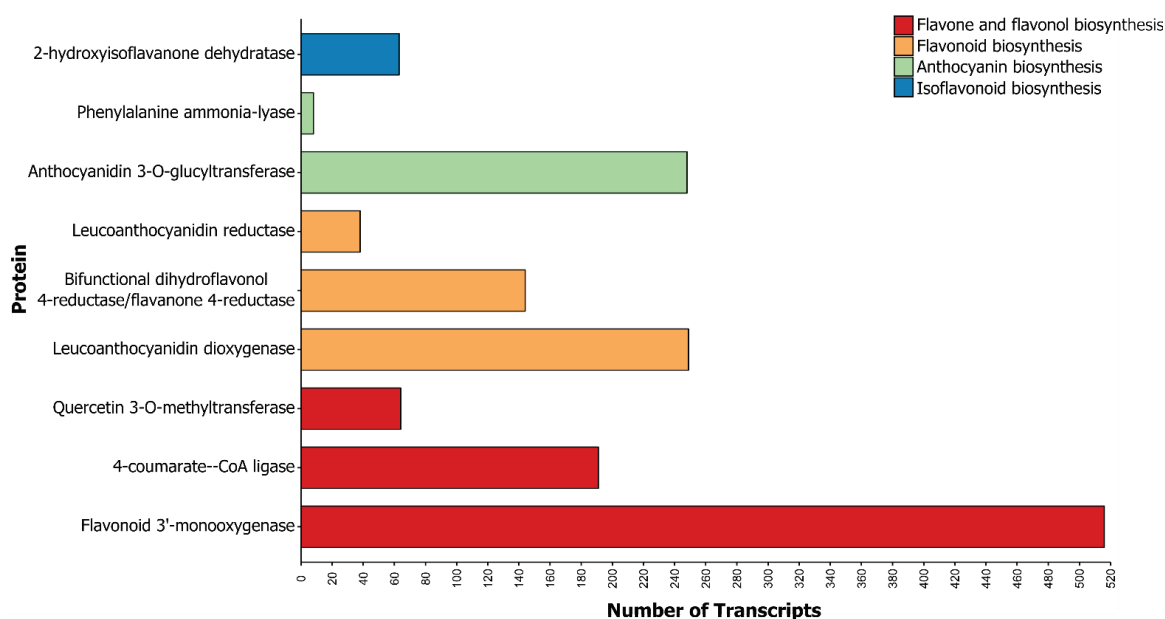


Figura 13. Número total de transcritos no genoma de *Stryphnodendron adstringens* com similaridade com proteínas envolvidas nas vias de biossíntese de taninos.

Os metabólitos secundários, amplamente distribuídos em plantas, são classificados em seis grandes subgrupos: chalconas, flavonas, flavonóis, flavononas, antocianinas e proantocianidinas ou taninos condensados, além de um sétimo grupo denominado auronas,

que é encontrado em algumas espécies (Winkel, 2006; Ferreyra et al., 2012). Os taninos condensados ou proantocianidinas são biossintetizados através da via do ácido chiquímico ou chiquimato, uma via que vem sendo investigada em estudos bioquímicos e genéticos (Ferreyra et al., 2012; Chen et al., 2018) (Figura 14). A compreensão atual é que a biossíntese de proantocianidinas é parte da via biosintética de flavonoides que também produz antocianinas e flavonóis (Chen et al., 2018).

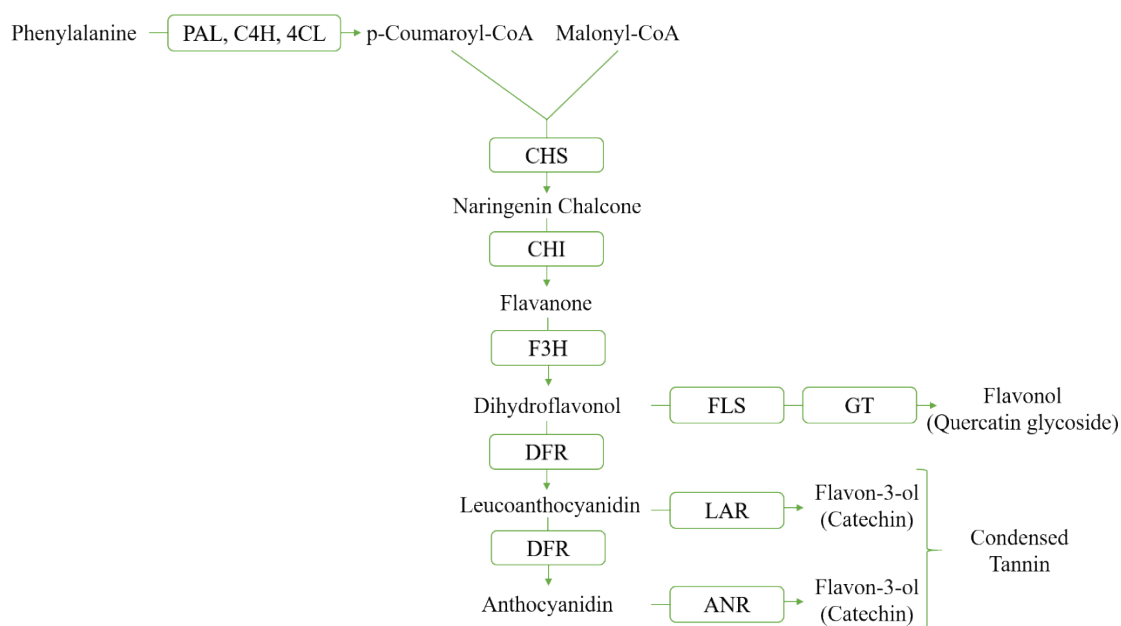


Figura 14. Via biosintética de taninos condensados. PAL (*phenylalanine ammonialyase*); C4H (*cinnamate 4-hydroxylase*); 4CL (*4-coumarate CoA ligase*); CHS (*chalcone syntase*); DFR (*dihydroflavonol 4-reductase*); LAR (*leucoanthocyanidin reductase*); ANR (*anthocyanidin reductase*). Adaptada de Chen et al. (2018).

Apesar da importância das proantocianidinas, pouco conhecimento se tem sobre a contribuição e regulação dos genes na sua biossíntese (Bogs et al., 2005; Chen et al., 2018). Bogs et al. (2005) demonstraram que ambos genes codificadores de LAR - *Leucoanthocyanidin* e ANR - *Anthocyanidin* são expressos durante a biossíntese de proantocianidinas em uva e são regulados de maneira tecido-específico e temporal. Como por exemplo, as uvas acumulam níveis significativos de proantocinidinas em seus frutos e folhas, oferecendo assim a oportunidade de examinar a contribuição desses genes em diferentes tecidos na mesma planta (Bogs et al., 2005). Já em *Arabidopsis* o gene *BANYULS* codifica a proteína ANR e é envolvido na biossíntese de proantocianidinas apenas em suas sementes (Xie et al., 2003). Chen et al., 2018 recentemente observou também uma

abundância maior de proantocianidinas nas folhas de *Rhus chinensis* que em suas noz-de-galha.

Assim, apesar da observação de transcritos de *S. adstringens* apresentando similaridade com proteínas envolvidas na biossíntese de proantocianidinas, para desvendar os detalhes dessa via de biossíntese na espécie, análises ao nível de transcriptoma e a confirmação dos genes envolvidos em tais vias são recomendadas para estudos futuros. Além disso, uma vez que a biossíntese de proantocianidinas é regulada de maneira tecido-específico e temporal, é importante considerar para trabalhos futuros, extração de material genético das cascas do caule, parte da planta que conhecidamente apresenta altas concentrações de tanino.

3.4 CONCLUSÃO

- i. O tamanho do genoma obtido por citometria de fluxo da espécie *S. adstringens* foi de aproximadamente 669 Mpb e o estimado por distribuição de *k-mers* foi de aproximadamente 624 Mpb;
- ii. O sequenciamento de duas bibliotecas *Illumina* possibilitou recuperar 76,3% do tamanho estimado para o genoma de *S. adstringens*;
- iii. O genoma de *S. adstringens* apresenta uma grande quantidade de elementos repetitivos, e os elementos transponíveis constituem a maior parte das repetições identificadas (~44,58% do *draft*), o que corrobora as informações descritas na literatura para genomas de plantas;
- iv. Foi possível identificar uma grande quantidade de regiões microssatélites (20.124), bem como identificar *primers* visando estudos futuros de genética de populações para a espécie *S. adstringens*.
- v. Foram preditos no genoma de *S. adstringens* 31.636 genes e 95.811 transcritos com uma média de 3,03 transcritos por gene.
- vi. 60.798 transcritos (63,46%) foram classificados funcionalmente nos termos do *Gene Ontology*. Além disso, 1.521 transcritos demonstraram similaridade com sequências de proteínas envolvidas nas vias de biossíntese de taninos.

3.5. REFERÊNCIAS

- ALHAKAMI, H.; MIREBRAHIM, H.; LONARDI, S. A comparative evaluation of genome assembly reconciliation tools. **Genome biology**, v. 18, n. 1, p. 93, 2017.
- ALTSCHUL, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. **Nucleic Acids Res**, v. 25, 1997.
- ANDREWS, S. FastQC: A quality control tool for high throughput sequence data. .
- ARABIDOPSIS GENOME INITIATIVE. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. **Nature**, v. 408, n. 6814, p. 796–815, 2000.
- AUDI, E. A. et al. Gastric antiulcerogenic effects of *Stryphnodendron adstringens* in rats. **Phytotherapy Research**, v. 13, n. 3, p. 264–266, 1999.
- AUDI, E. A. et al. Biological activity and quality control of extract and stem bark from *Stryphnodendron adstringens*. **acta farmacéutica bonaerense**, v. 23, p. 328–333, 2004.
- BARATELI, L. O. **Desenvolvimento de marcadores microssatélites para *Stryphnodendron adstringens* (barbatimão - Fabaceae)**. 2018. 63 f. Dissertação (Mestrado em Genética e Melhoramento de Plantas) – Escola de Agronomia, Universidade Federal de Goiás, Goiânia, 2018.
- BENNETT, M. D.; LEITCH, I. J. Plant DNA C-values database. Disponível em: <http://data.kew.org/cvalues>. Acesso em: 14/7/2018.
- BENNETT, M. D.; LEITCH, I. J.; HANSON, L. DNA amounts in two samples of angiosperm weeds. **Annals of Botany**, v. 82, n. SUPPL. A, p. 121–134, 1998. Oxford University Press.
- BENNETZEN, J. L.; MA, J.; DEVOS, K. M. Mechanisms of recent genome size variation in flowering plants. **Annals of botany**, v. 95, n. 1, p. 127–32, 2005. Oxford University Press.
- BERTIOLI, D. J. et al. The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. **Nature Genetics**, v. 48, n. 4, p. 438–446, 2016.
- BLAKESLEY, D. et al. Natural and induced polyploidy in *Acacia dealbata* Link. and *Acacia mangium* Willd. **Annals of Botany**, v. 90, n. 3, p. 391–398, 2002.
- BOGS, J. et al. Proanthocyanidin synthesis and expression of genes encoding leucoanthocyanidin reductase and anthocyanidin reductase in developing grape berries and grapevine leaves. **Plant Physiology**, v. 139, n. 2, p. 652–663, 2005.
- BOLGER, A. M.; LOHSE, M.; USADEL, B. Trimmomatic: a flexible trimmer for Illumina sequence data. **Bioinformatics**, v. 30, n. 15, p. 2114–2120, 2014.
- BRANCO, L.E.; ZIMBACK, L.; LIMA, A. B.; MORI, E.S.; AOKI, H. Estrutura genética de populações de *Stryphnodendron adstringens* (Mart.). **Instituto Florestal**, Sér. Reg., n. 42 p. 31- 37 jul. 2010.
- CAVAGNARO, P. F. et al. Genome-wide characterization of simple sequence repeats in cucumber (*Cucumis sativus* L.). **BMC Genomics**, v. 11, n. 1, p. 569-, 2010.
- CAMILLO, J.; CIAMPI, A.; VIEIRA, R. F. Análise da variabilidade genética em barbatimão

(*Stryphnodendron adstringens*) usando marcadores RAPD. In: **Anais** do Encontro do Talento Estudantil da Embrapa Recursos Genéticos e Biotecnologia, 6, Brasília, DF. Embrapa Recursos Genéticos e Biotecnologia, Brasília, p.5, 2001.

CHANG, Y. et al. The draft genomes of five agriculturally important African orphan crops. **GigaScience**, 2018.

CHAUX, N. et al. The predominantly selfing plant *Arabidopsis thaliana* experienced a recent reduction in transposable element abundance compared to its outcrossing relative *Arabidopsis lyrata*. **Mobile DNA**, v. 3, n. 1, p. 2, 2012.

CHEN, H. et al. Molecular mechanisms of tannin accumulation in *Rhus galls* and genes involved in plant-insect interactions. **Scientific reports**, v. 8, n. 1, p. 9841, 2018..

CHEN, S. et al. An introduction to the medicinal plant genome project. **Frontiers of medicine**, v. 5, n. 2, p. 178–184, 2011.

CIVÁŇ, P.; ŠVEC, M.; HAUPTVOGEL, P. On the Coevolution of Transposable Elements and Plant Genomes. **Journal of Botany**, v. 2011, p. 1–9, 2011. Hindawi.

CONESA, A. et al. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. **Bioinformatics**, v. 21, n. 18, p. 3674–3676, 2005.

CONSORTIUM, G. O. The Gene Ontology (GO) database and informatics resource. **Nucleic acids research**, v. 32, n. suppl_1, p. D258–D261, 2004. Oxford University Press.

CONSORTIUM, T. I. B. G. S. A physical, genetic and functional sequence assembly of the barley genome. **Nature**, v. 491, n. 7426, p. 711–716, 2012.

CORRÊA, V. S. et al. Geographical variation and quality assessment of *Stryphnodendron adstringens* (Mart.) Coville within Brazil. **Genetic resources and crop evolution**, v. 59, n. 7, p. 1349–1356, 2012.

DEMENOU, B. B. et al. Development and Characterization of Microsatellite Markers in the African Deciduous Tree *Terminalia superba* (Combretaceae). **Applications in Plant Sciences**, v. 3, n. 12, p. 1500070, 2015.

DEMENOU, B. B.; HARDY, O. J. Development, Characterization, and Cross-Amplification of Microsatellite Markers in the Understudied African Genus *Anthonothea* (Fabaceae). **Applications in Plant Sciences**, v. 5, n. 1, p. 1600120, 2017.

DOLEŽEL, J.; GREILHUBER, J.; SUDA, J. Estimation of nuclear DNA content in plants using flow cytometry. **Nature Protocols**, v. 2, n. 9, p. 2233–2244, 2007.

DOLEŽEL, J.; SGORBATI, S.; LUCRETTI, S. Comparison of three DNA fluorochromes for flow cytometric estimation of nuclear DNA content in plants. **Physiologia Plantarum**, v. 85, n. 4, p. 625–631, 1992.

DONG, A.-X. et al. High-quality assembly of the reference genome for scarlet sage, *Salvia splendens*, an economically important ornamental plant. **GigaScience**, v. 7, n. 7, 2018.

DOYLE, J. J.; DOYLE, J. L. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. **Phytochemical Bulletin**, v. 19, p. 11–15, 1987.

EKBLOM, R.; WOLF, J. B. W. A field guide to whole genome sequencing and annotation. **Evolutionary applications**, v. 7, n. 9, p. 1026–1042, 2014.

- genome sequ

- FERREYRA, M. L. F.; RIUS, S. P.; CASATI, P. Flavonoids: biosynthesis, biological functions, and biotechnological applications. **Frontiers in Plant Science**, v. 3, p. 222, 2012.
- FESCHOTTE, C.; JIANG, N.; WESSLER, S. R. Plant transposable elements: where genetics meets genomics. **Nature Reviews Genetics**, v. 3, n. 5, p. 329–341, 2002.
- FIELD, B.; OSBOURN, A. E. Metabolic diversification—-independent assembly of operon-like gene clusters in different plants. **Science**, v. 320, n. 5875, p. 543–547, 2008.
- GAO, F. et al. Long-read sequencing and de novo genome assembly of *Ammopiptanthus nanus*, a desert shrub. **GigaScience**, 2018.
- GLASENAPP, J. S. et al. Allozyme variation in a natural population of *Stryphnodendron adstringens* in the Rio Preto State Park, southeastern Brazil. **Brazilian Journal of Botany**, v. 35, n. 2, p. 137–144, 2012.
- GLASENAPP, J. S. et al. Descrição da diversidade genética de populações naturais de barbatimão *Stryphnodendron adstringens* (mart.) coville em unidades de conservação de Minas Gerais. **Revista Árvore**, v. 38, n. 1, p. 103–112, 2014.
- GOODWIN, S.; MCPHERSON, J. D.; MCCOMBIE, W. R. Coming of age: ten years of next-generation sequencing technologies. **Nature Reviews Genetics**, v. 17, n. 6, p. 333, 2016.
- GUO, X. et al. The genomes of two *Eutrema* species provide insight into plant adaptation to high altitudes. **DNA Research**, v. 25, n. 3, p. 307–315, 2018.
- GUREVICH, A. et al. QUAST: quality assessment tool for genome assemblies. **Bioinformatics**, v. 29, n. 8, p. 1072–1075, 2013.
- HAN, S. et al. De novo assembly and characterization of *Gleditsia sinensis* transcriptome and subsequent gene identification and SSR mining. **Genetics and Molecular Research**, v. 15, n. 1, 2016.
- HARE, E. E.; JOHNSTON, J. S. Genome Size Determination Using Flow Cytometry of Propidium Iodide-Stained Nuclei. **Methods in molecular biology (Clifton, N.J.)**. v. 772, p.3–12, 2012.
- HAWKINS, J. S. et al. Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. **Genome Research**, v. 16, n. 10, p. 1252–1261, 2006.
- HAWKINS, J. S.; GROVER, C. E.; WENDEL, J. F. Repeated big bangs and the expanding universe: Directionality in plant genome size evolution. **Plant Science**, v. 174, n. 6, p. 557–562, 2008.
- HELLER, F. O. DNS ~~Bestimmung mit Hilfe~~ ^{Bestimmung mit Hilfe} von Impulscytophotometrie. **Berichte der Deutschen Botanischen Gesellschaft**, v. 86, n. 5–9, p. 437–441, 1973.
- HIRAKAWA, H. et al. Draft genome sequence of subterranean clover, a reference for genus *Trifolium*. **Scientific Reports**, v. 6, n. 1, p. 30358, 2016.
- HU, G. et al. The history and disposition of transposable elements in polyploid *Gossypium*. **Genome**, v. 53, n. 8, p. 599–607, 2010.

IBARRA-LACLETTE, E. et al. Architecture and evolution of a minute plant genome. **Nature**, v. 498, n. 7452, p. 94, 2013.

IGARASHI, Y. et al. Development and Evaluation of Microsatellite Markers for the Critically Endangered Birch *Betula chichibuensis* (Betulaceae). **Applications in Plant Sciences**, v. 5, n. 5, p. 1700016, 2017.

ISHIDA, K. et al. Influence of tannins from *Stryphnodendron adstringens* on growth and virulence factors of *Candida albicans*. **Journal of Antimicrobial Chemotherapy**, v. 58, n. 5, p. 942–949, 2006.

JONCZYK, R. et al. Elucidation of the final reactions of DIMBOA-glucoside biosynthesis in maize: characterization of Bx6 and Bx7. **Plant physiology**, v. 146, n. 3, p. 1053–1063, 2008.

KALIA, R. K.; RAI, M. K.; KALIA, S.; SINGH, R.; DHAWAN, A. K. Microsatellite markers: an overview of the recent progress in plants. **Euphytica**, v. 177, n. 3, p. 309–334, 2011.

KANG, Y. J. et al. Genome sequence of mungbean and insights into evolution within *Vigna* species. **Nature Communications**, v. 5, n. 1, p. 5443, 2014.

KANG, Y. J. et al. Draft genome sequence of adzuki bean, *Vigna angularis*. **Scientific Reports**, v. 5, n. 1, p. 8069, 2015.

KAPITONOV, V. V.; JURKA, J. Helitrons on a roll: eukaryotic rolling-circle transposons. **Trends in Genetics**, v. 23, n. 10, p. 521–529, 2007.

KELLNER, F. et al. Genome
The Plant Journal, v. 82, n. 4, p. 680–692, 2015.

-guided inves

LAI, C. W. J. et al. Analysis of papaya BAC end sequences reveals first insights into the organization of a fruit tree genome. **Molecular Genetics and Genomics**, v. 276, n. 1, p. 1–12, 2006.

LAWSON, M.; ZHANG, L. Distinct patterns of SSR distribution in the *Arabidopsis thaliana* and rice genomes. **Genome Biology**, v. 7, n. 2, p. R14, 2006.

LI, F. et al. Transcriptome sequencing of lima bean (*Phaseolus lunatus*) to identify putative positive selection in Phaseolus and legumes. **International journal of molecular sciences**, v. 16, n. 7, p. 15172–15187, 2015.

LIMA, C. R. DE O. et al. Barbatiman and chitosan creams as adjuvants in rabbit skin wound healing. **Semina: Ciências Agrárias**, v. 37, n. 3, p. 1317–1326, 2016.

LIMA, J. C. S.; MARTINS, D. T. O.; DE SOUZA, P. T. Experimental evaluation of stem bark of *Stryphnodendron adstringens* (Mart.) Coville for antiinflammatory activity. **Phytotherapy Research**, v. 12, n. 3, p. 218–220, 1998.

LIU, Q. et al. Assembly and annotation of a draft genome sequence for *Glycine latifolia*, a perennial wild relative of soybean. **The Plant Journal**, v. 95, n. 1, p. 71–85, 2018.

LOPES, G. C. et al. Determinação quantitativa de taninos em três espécies de *Stryphnodendron* por cromatografia líquida de alta eficiência. **Brazilian Journal of Pharmaceutical Sciences**, v. 45, n. 1, p. 135–143, 2009.

LÓPEZ-FLORES, I.; GARRIDO-RAMOS, M. A. The repetitive DNA content of eukaryotic genomes. **Repetitive DNA**, v. 7, p.1–28, 2012.

LOUREIRO, J. et al. Two New Nuclear Isolation Buffers for Plant DNA Flow Cytometry: A Test with 37 Species. **Annals of Botany**, v. 100, n. 4, p. 875–888, 2007.

LUIZ, R. L. F. et al. Proanthocyanidins polymeric tannin from *Stryphnodendron adstringens* are active against *Candida albicans* biofilms. **BMC Complementary and Alternative Medicine**, v. 15, n. 1, p. 68, 2015.

MARÇAIS, G.; KINGSFORD, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. **Bioinformatics**, v. 27, n. 6, p. 764–770, 2011.

MATSUBA, Y. et al. Evolution of a complex locus for terpene biosynthesis in *Solanum*. **The Plant Cell**, p. tpc-113, 2013.

MEGLÉCZ, E. et al. 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**, 2014.

MELO, J. O. DE et al. Effect of *Stryphnodendron adstringens* (barbatimão) bark on animal models of nociception. **Revista Brasileira de Ciências Farmacêuticas**, v. 43, n. 3, p. 465–469, 2007.

MENDONÇA, P.C. **Caracterização da diversidade genética de *Stryphnodendron adstringens* (Mart.) Coville por marcador molecular AFLP e transferência de microssatélites**. 2011. 87f.. Tese (Doutorado em Agronomia) - Universidade Estadual Paulista, Botucatu, São Paulo, 2011.

MENDONÇA, P. C. et al. Genetic diversity of *Stryphnodendron adstringens* (Mart.) Coville determined by AFLP molecular markers. **Biochemical systematics and ecology**, v. 41, p. 16–20, 2012.

METZKER, M. L. Sequencing technologies—the next generation. **Nature reviews genetics**, v. 11, n. 1, p. 31, 2010.

MOCHIDA, K. et al. Draft genome assembly and annotation of *Glycyrrhiza uralensis*, a medicinal legume. **The Plant Journal**, v. 89, n. 2, p. 181–194, 2017.

MORGANTE, M.; HANAFEY, M.; POWELL, W. Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. **Nature Genetics**, v. 30, n. 2, p. 194–200, 2002.

MORGANTE, M.; OLIVIERI, A. M. PCR-amplified microsatellites as markers in plant genetics. **The Plant Journal**, v. 3, n. 1, p. 175–182, 1993.

NEVILL, P. G.; WARDELL-JOHNSON, G. Microsatellite Primers for the Rare Shrub *Acacia adinophylla* (Fabaceae). **Applications in Plant Sciences**, v. 4, n. 11, p. 1600084, 2016.

NYSTEDT, B. et al. The Norway spruce genome sequence and conifer genome evolution. **Nature**, v. 497, n. 7451, p. 579, 2013.

OHRI, D.; BHARGAVA, A.; CHATTERJEE, A. Nuclear DNA amounts in 112 species of tropical hardwoods - New estimates. **Plant Biology**, v. 6, n. 5, p. 555–561, 2004.

- PALOMINO, G.; ROMO, G.; ZÁRATE, S. Chromosome Numbers and DNA Content in Some Taxa of *Leucaena* (Fabaceae Mimosoideae). **CYTOLOGIA**, v. 60, n. 1, p. 31–37, 1995.
- PANIZZA, S. et al. *Stryphnodendron barbadetiman* (Vellozo) Martius: teor em Tannino na casca e sua propriedade cicatrizante. **Rev. ciênc. farm**, v. 10, p. 101–106, 1988.
- PRYSZCZ, L. P.; GABALDÓN, T. Redundans: an assembly pipeline for highly heterozygous genomes. **Nucleic Acids Research**, v. 44, n. 12, p. e113–e113, 2016.
- RAI, A.; SAITO, K.; YAMAZAKI, M. Integrated omics analysis of specialized metabolism in medicinal plants. **The Plant Journal**, v. 90, n. 4, p. 764–787, 2017.
- ROZEN, S.; SKALETSKY, H. J. Primer3 on the WWW for General Users and for Biologist Programmers. **Bioinformatics Methods and Protocols Methods in Molecular Biology**, v. 132, p. 365–386, 2000.
- SAHEBI, M. et al. Contribution of transposable elements in the plant's genome. **Gene**, v. 665, p. 155–166, 2018.
- SATO, S. et al. Genome Structure of the Legume, *Lotus japonicus*. **DNA Research**, v. 15, n. 4, p. 227–239, 2008.
- SCHMUTZ, J. et al. Genome sequence of the palaeopolyploid soybean. **Nature**, v. 463, n. 7278, p. 178–183, 2010.
- SCHMUTZ, J. et al. A reference genome for common bean and genome-wide analysis of dual domestications. **Nature Genetics**, v. 46, n. 7, p. 707–713, 2014.
- SCHNABLE, P. S. et al. The B73 Maize Genome: Complexity, Diversity, and Dynamics. **Science**, v. 326, n. 5956, p. 1112–1115, 2009.
- SHANG, Y. et al. Biosynthesis, regulation, and domestication of bitterness in cucumber. **Science**, v. 346, n. 6213, p. 1084–1088, 2014.
- SHARMA, P. C.; GROVER, A.; KAHL, G. Mining microsatellites in eukaryotic genomes. **Trends in Biotechnology**, nov. 2007.
- SHULTZ, J. L. et al. The development of BAC-end sequence-based microsatellite markers and placement in the physical and genetic maps of soybean. **Theoretical and Applied Genetics**, v. 114, n. 6, p. 1081–1090, 2007.
- SILVA-JUNIOR, O. B. et al. Genome assembly of the Pink Ipê (*Handroanthus impetiginosus*, Bignoniaceae), a highly valued, ecologically keystone Neotropical timber forest tree. **GigaScience**, v. 7, n. 1, p. 1–16, 2018.
- SMIT, A. F. A.; HUBLEY, R.; GREEN, P. RepeatMasker Open-3.0. , 2017.
- SMIT, A.; HUBLEY, R. RepeatModeler-1.0. 5. **Institute for Systems Biology**, 2012.
- SOUZA, V. C.; GIBAU, A. *Stryphnodendron* in Flora do Brasil 2020 em construção. Disponível em: <<http://reflora.jbrj.gov.br/reflora/floradobrasil/FB19133>>. Acesso em: 14/7/2018.
- STANKE, M. et al. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. **BMC bioinformatics**, v. 7, n. 1, p. 62, 2006.

- STERCK, L. et al. How many genes are there in plants (... and why are they there)? **Current opinion in plant biology**, v. 10, n. 2, p. 199–203, 2007.
- TAMANNA, A.; KHAN, A. U. Mapping and analysis of simple sequence repeats in the *Arabidopsis thaliana* genome. **Bioinformation**, v. 1, n. 2, p. 64–8, 2005.
- TOSSO, F. et al. Characterization of Microsatellite Markers in the African Tropical Tree Species *Guibourtia ehie* (Fabaceae, Detarioideae). **Applications in Plant Sciences**, v. 5, n. 7, p. 1700023, 2017.
- TÓTH, G.; GÁSPÁRI, Z.; JURKA, J. Microsatellites in different eukaryotic genomes: Surveys and analysis. **Genome Research**, v. 10, n. 7, p. 967–981, 2000.
- TRAN, H. T. M. et al. Use of a draft genome of coffee (*Coffea arabica*) to identify SNPs associated with caffeine content. **Plant Biotechnology Journal**, 13. abr. 2018.
- UNAMBA, C. I. N.; NAG, A.; SHARMA, R. K. Next generation sequencing technologies: the doorway to the unexplored genomics of non-model plants. **Frontiers in plant science**, v. 6, p. 1074, 2015.
- VARSHNEY, R. K. et al. Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. **Nature Biotechnology**, v. 30, n. 1, p. 83–89, 2012.
- VARSHNEY, R. K. et al. Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. **Nature Biotechnology**, v. 31, n. 3, p. 240–246, 2013.
- DE VEGA, J. J. et al. Red clover (*Trifolium pratense* L.) draft genome provides a platform for trait improvement. **Scientific Reports**, v. 5, n. 1, p. 17394, 2015.
- WINKEL, B. S. J. The biosynthesis of flavonoids. **The science of flavonoids**. p.71–95, 2006.
- XIE, D.-Y. et al. Role of anthocyanidin reductase, encoded by BANYULS in plant flavonoid biosynthesis. **Science**, v. 299, n. 5605, p. 396–399, 2003.
- XU, H. et al. FastUniq: A Fast De Novo Duplicates Removal Tool for Paired Short Reads. **PLoS ONE**, v. 7, n. 12, p. e52249, 2012.
- YAMAMOTO, M. et al. Development and characterization of 24 microsatellite markers in *Primula tosaensis*, an endangered primrose, using MiSeq. **Plant Species Biology**, v. 33, n. 1, p. 77–80, 2018.
- YANDELL, M.; ENCE, D. A beginner's guide to eukaryotic genome annotation. **Nature Reviews Genetics**, v. 13, n. 5, p. 329–342, 2012.
- YOUNG, N. D. et al. The *Medicago* genome provides insight into the evolution of rhizobial symbioses. **Nature**, v. 480, n. 7378, p. 520–524, 2011.
- YOUNG, N. D.; BHARTI, A. K. Genome-enabled insights into legume biology. **Annual review of plant biology**, v. 63, p. 283–305, 2012.
- ZALAPA, J. E. et al. Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. **American Journal of Botany**, v. 99, n. 2, p. 193–208, 2012.
- ZHANG, H. et al. Transcriptome analysis reveals potential genes involved in flower

pigmentation in a red-flowered mutant of white clover (*Trifolium repens* L.). **Genomics**, v. 110, n. 3, p. 191–200, 2018.

ZIMIN, A. V. et al. The MaSuRCA genome assembler. **Bioinformatics**, v. 29, n. 21, p. 2669–2677, 2013.

Tabela suplementar 1. Tipos de motivos de repetição encontrados no genoma da espécie *Stryphnodendron adstringens*.

Motivo	Frequência Absoluta	Frequência Relativa	Motivo	Frequência Absoluta	Frequência Relativa
Dinucleotídeos			Hexanucleotídeos		
AT/TA	8886	44,156	AAAAAC/TTTTTA	2	0,010
AC/TG	2065	10,261	AAAAAG/TTTTTC	1	0,005
AG/TC	2504	12,443	AAAAAT/TTTTTA	4	0,020
CA/GT	1078	5,357	AAAAGA/TTTTCT	1	0,005
CG/GC	6	0,030	AAAATT/TTTAA	1	0,005
CT/GA	2371	11,782	AAAACA/TTTTGT	1	0,005
Trinucleotídeos			AAATAA/TTTATT	1	0,005
ACC/TTG	204	1,014	AAATTT/TTTAAA	1	0,005
AAG/TTC	153	0,760	AACACT/TTGTGA	1	0,005
AAT/TTA	382	1,898	AACAGG/TTGTCC	1	0,005
ACA/TGT	87	0,432	AACTAC/TTGATG	1	0,005
ACC/TGG	12	0,060	AAGAAA/TTCTTT	1	0,005
ACG/TGC	6	0,030	AAGATT/TTCTAA	1	0,005
ACT/TGA	21	0,104	AAGCCT/TTCCGA	1	0,005
AGA/TCT	69	0,343	AAACAA/TTTGTT	1	0,005
AGC/TCG	6	0,030	AAATAT/TTTATA	1	0,005
AGG/TCC	20	0,099	AAGTGG/TTCCACC	1	0,005
AGT/TCA	23	0,114	AAGTGT/TTCACA	1	0,005
ATA/TAT	165	0,820	AACAAT/TTGTTA	3	0,015
ATC/TAG	35	0,174	AACACC/TTGTGG	1	0,005
ATG/TAC	33	0,164	AATAAC/TTATTG	3	0,015
ATT/TAA	240	1,193	AATACT/TTATGA	1	0,005
CAA/GTT	91	0,452	AATTAT/TTAATA	1	0,005
CAC/GTG	14	0,070	AATTGC/TTAACG	1	0,005
CAG/GTC	7	0,035	AATTTT/TTAAAA	1	0,005
CAT/GTA	34	0,169	AATAAG/TTATTC	1	0,005
CCA/GGT	16	0,080	AATATA/TTATAT	1	0,005
CCG/GGC	2	0,010	AATCAA/TTAGTT	2	0,010
CCT/GGA	28	0,139	AATGTG/TTACAC	1	0,005
CGA/GCT	3	0,015	ACAACC/TGTTGG	1	0,005
CGC/GCG	5	0,025	ACATAT/TGTATA	2	0,010
CGG/GCC	5	0,025	ACAAAG/TGTTTC	1	0,005
CGT/GCA	4	0,020	ACACAG/TGTGTC	1	0,005
CTA/GAT	23	0,114	ACACGC/TGTGCG	1	0,005
CTC/GAG	28	0,139	ACACTG/TGTGAC	1	0,005
CTG/GAC	5	0,025	ACCGTC/TGGCAG	1	0,005
CTT/GAA	101	0,502	ACCTTC/TGGAAG	1	0,005
Tetranucleotídeos			ACGAAG/TGCTTC	1	0,005

Motivo	Frequência Absoluta	Frequência Relativa	Motivo	Frequência Absoluta	Frequência Relativa
Tetranucleotídeos			Hexanucleotídeos		
AAAC/TTTG	10	0,050	ACGCAA/TGCGTT	1	0,005
AAAG/TTTC	78	0,388	ACTAAA/TGATTT	1	0,005
AAAT/TTTA	113	0,562	ACTACC/TGATGG	1	0,005
AACA/TTGT	3	0,015	ACTACG/TGATGC	1	0,005
AACT/TTGA	2	0,010	ACTATA/TGATAT	1	0,005
AAGA/TTCT	20	0,099	ACTCTA/TGAGAT	1	0,005
AAGC/TTCG	1	0,005	ACTCTC/TGAGAG	1	0,005
AAGG/TTCC	4	0,020	AGAATA/TCCTTAT	1	0,005
AAGT/TTCA	1	0,005	AGAGGG/TCTCCC	1	0,005
ATAA/TTAT	54	0,268	AGATAC/TCTATG	1	0,005
AATC/TTAG	3	0,015	AGATCC/TCTAGG	1	0,005
AATG/TTAC	1	0,005	AGATGG/TCTGCC	1	0,005
AATT/TTAA	27	0,134	AGACAC/TCTGTG	1	0,005
ACGT/TGCA	1	0,005	AGCAGG/TCGTCC	2	0,010
ACTC/TGAG	1	0,005	AGGAAG/TCCTTC	1	0,005
ACTT/TGAA	1	0,005	AGGAGC/TCCTCG	1	0,005
ACAA/TGTT	3	0,015	AGGGAG/TCCCTC	1	0,005
ACAT/TGTA	50	0,248	AGGGGA/TCCCCT	1	0,005
ACGC/TGCG	5	0,025	AGGTGA/TCCACT	2	0,010
AGAA/TCCT	20	0,099	AGCGGC/TCGCCG	2	0,010
AGAC/TCTG	1	0,005	AGGAGA/TCCTCT	1	0,005
AGAT/TCTA	2	0,010	AGGAGC/TCCTCG	1	0,005
AGCA/TCGT	1	0,005	AGTAGC/TCATCG	1	0,005
AGCT/TCGA	1	0,005	AGTCAG/TCAGTC	1	0,005
AGGA/TCCT	3	0,015	AGTCGA/TCAGCT	1	0,005
AGGG/TCCC	9	0,045	ATAAAA/TATTTT	3	0,015
AGGT/TCCA	1	0,005	ATAATG/TATTAC	3	0,015
AGTT/TCAA	1	0,005	ATAATC/TATTAG	1	0,005
TCAT/AGTA	1	0,005	ATACTA/TATGAT	1	0,005
TCAC/AGTG	2	0,010	ATATAC/TATATG	1	0,005
ATAA/TATT	24	0,119	ATATTA/TATAAT	1	0,005
ATAC/TATG	105	0,522	ATATTG/TATAAC	1	0,005
ATAG/TATC	7	0,035	ATAATT/TATTAA	2	0,010
ATCC/TAGG	1	0,005	ATACAT/TATGTA	2	0,010
ATCA/TAGT	2	0,010	ATAGAT/TATCTA	1	0,005
ATCT/TAGA	3	0,015	ATATAC/TATATG	1	0,005
ATGA/TACT	2	0,010	ATCCGT/TAGGCA	1	0,005
ATGC/TACG	11	0,055	ATCAAA/TAGTTT	1	0,005
ATGT/TACA	37	0,184	ATCAAT/TAGTTA	1	0,005
ATTA/TAAT	42	0,209	ATCATA/TAGTAT	1	0,005

Motivo	Frequência Absoluta	Frequência Relativa	Motivo	Frequência Absoluta	Frequência Relativa
Tetranucleotídeos			Hexanucleotídeos		
ATTC/TAAG	1	0,005	ATCATT/TAGTAA	1	0,005
ATTG/TAAC	4	0,020	ATGACA/TACTGT	1	0,005
ATTT/TAAA	26	0,129	ATGATA/TACTAT	2	0,010
CAAA/GTTT	4	0,020	ATGTAT/TACATA	1	0,005
CACG/GTGC	8	0,040	ATGCGA/TACGCT	1	0,005
CAAC/GTTG	1	0,005	ATTATC/TAATAG	1	0,005
CACT/GTGA	6	0,030	ATTACC/TAATGG	1	0,005
CATA/GTAT	68	0,338	ATTCAC/TAAGTG	1	0,005
CATG/GTAC	4	0,020	ATTCTG/TAAGAC	1	0,005
CCCT/GGGA	8	0,040	ATTCTT/TAAGAA	1	0,005
CATC/GTAG	2	0,010	ATTTCA/TAAAGT	1	0,005
CCTA/GGAT	2	0,010	ATTTTA/TAAAAT	2	0,010
CCTC/GGAG	7	0,035	ATTTTG/TAAAAC	1	0,005
CCTT/GGAA	2	0,010	ATTTTT/TAAAAA	3	0,015
CGAC/GCTG	1	0,005	CAAAAA/GTTTTT	2	0,010
CATT/GTAA	2	0,010	CAAAAG/GTTTTTC	2	0,010
CCAA/GGTT	1	0,005	CACACG/GTGTGC	2	0,010
CCAT/GGTA	3	0,015	CACCCA/GTGGGT	1	0,005
CGCA/GCGT	2	0,010	CACGCG/GTGCGC	1	0,005
CGTA/GCAT	3	0,015	CACTAT/GTGATA	1	0,005
CGTT/GCAA	1	0,005	CAGACG/GTCTGC	1	0,005
CTAA/GATT	1	0,005	CAGAGC/GTCTCG	1	0,005
CTAT/GATA	1	0,005	CAGGGG/GTCCCC	1	0,005
CTAG/GATC	1	0,005	CATCAC/GTAGTG	1	0,005
CTCA/GAGT	11	0,055	CCACTC/GGTGAG	1	0,005
CTCC/GAGG	25	0,124	CCGCCA/GGCGGT	2	0,010
CTCG/GAGC	2	0,010	CCTCCA/GGAGGT	1	0,005
CTTC/GAAG	4	0,020	CCTCCG/GGAGGC	1	0,005
CTTT/GAAA	25	0,124	CCTTTA/GGAAAT	1	0,005
Pentanucleotídeos			CGACGG/GCTGCC	1	0,005
AAAAC/TTTTG	7	0,035	CGATGG/GCTACC	1	0,005
AAAAG/TTTTTC	49	0,243	CGGCAT/GCCGTA	1	0,005
AAAAT/TTTTA	40	0,199	CGGGAC/GCCCTG	1	0,005
AAAGA/TTTCT	14	0,070	CGGGTC/GCCCAG	1	0,005
AAACA/TTTGT	1	0,005	CTACTG/GATGAC	2	0,010
AAATA/TTTAT	7	0,035	CTCACT/GAGTGA	1	0,005
AAATG/TTTAC	1	0,005	CTCCCT/GAGGGA	2	0,010
AAATT/TTTAA	4	0,020	CTCTAT/GAGATA	1	0,005
AACAT/TTGTA	20	0,099	CTCTCA/GAGAGT	2	0,010
AACGA/TTGCT	1	0,005	CTGCTA/GACGAT	1	0,005

Motivo	Frequência Absoluta	Frequência Relativa	Motivo	Frequência Absoluta	Frequência Relativa
Pentanucleotídeos			Hexanucleotídeos		
AAGGG/TTCCC	1	0,005	CTTCAT/GAAGTA	1	0,005
AAGAA/TTCTT	6	0,030	CTTCGT/GAAGCA	1	0,005
AATAA/TTATT	6	0,030	CTTCTA/GAAGAT	1	0,005
AATAG/TTATC	1	0,005	CTTCTC/GAAGAG	4	0,020
ACAAA/TGTTT	2	0,010	CTTTTC/GAAAAG	1	0,005
ACAAC/TGTTG	1	0,005	CTTTTT/GAAAAA	2	0,010
ACTAA/TGATT	1	0,005	CAAAAC/GTTTTG	1	0,005
ACTAT/TGATA	1	0,005	CAACTA/GTTGAT	1	0,005
ACTCC/TGAGG	1	0,005	CAATCC/GTTAGG	1	0,005
AATAT/TTATA	3	0,015	CACACT/GTGTGA	2	0,010
AATGA/TTACT	1	0,005	CACCGC/GTGGCG	1	0,005
AATTA/TTAAT	1	0,005	CACGCA/GTGCGT	1	0,005
AATTC/TTAAG	1	0,005	CATATA/GTATAT	1	0,005
AATTT/TTAAA	2	0,010	CATGTG/GTACAC	1	0,005
ACATC/TGTAG	1	0,005	CCACCT/GGTGGA	1	0,005
ACCGG/TGGCC	1	0,005	CCACGA/GGTGCT	1	0,005
ACGGG/TGCCC	1	0,005	CCCCTT/GGGGAA	1	0,005
AGAAA/TCTTT	6	0,030	CCCTAT/GGGATA	1	0,005
AGGGT/TCCCA	1	0,005	CCCTCA/GGGAGT	1	0,005
AGGTA/TCCAT	1	0,005	CCCTCT/GGGAGA	1	0,005
AGAAG/TCTTC	2	0,010	CCTCTC/GGAGAG	1	0,005
AGAGA/TCTCT	4	0,020	CCTCTT/GGAGAA	1	0,005
AGAGT/TCTCA	2	0,010	CGAGAG/GCTCTC	1	0,005
AGGGG/TCCCC	1	0,005	CGCAAC/GCGTTG	1	0,005
ATACA/TATGT	2	0,010	CGTACA/GCATGT	2	0,010
ATATA/TATAT	2	0,010	CGTAGT/GCATCA	1	0,005
ATCTC/TAGAG	2	0,010	CGTCCA/GCAGGT	1	0,005
AGTGG/ACACC	1	0,005	CTAAAA/GATTTT	1	0,005
AGTTG/TCAAC	1	0,005	CTACTC/GATGAG	1	0,005
ATAAA/TATTT	11	0,055	CTACTT/GATGAA	1	0,005
ATAAC/TATTG	1	0,005	CTATTA/GATAAT	1	0,005
ATAAT/TATTA	3	0,015	CTCTCC/GAGAGG	1	0,005
ATATG/TATAC	1	0,005	CTGCCT/GACGGA	1	0,005
ATATT/TATAA	3	0,015	CTTATA/GAATAT	1	0,005
ATGAT/TACTA	1	0,005	CTTCCA/GAAGGT	1	0,005
ATGTT/TACAA	8	0,040	CTTTCC/GAAAGG	1	0,005
ATTAA/TAATT	1	0,005			
ATTAT/TAATA	10	0,050			
ATTTA/TAAAT	3	0,015			
ATTTT/TAAAA	16	0,080			

Motivo	Frequência Absoluta	Frequência Relativa
Pentanucleotídeos		
CAAAA/GTTTT	3	0,015
CAACA/GTTGT	1	0,005
CAATT/GTTAA	1	0,005
CAAAT/GTTTA	1	0,005
CAAGA/GTTCT	1	0,005
CACCG/GTGGC	1	0,005
CACCT/GTGGA	1	0,005
CAGAA/GTCTT	1	0,005
CCAAG/GGTTC	1	0,005
CCAGA/GGTCT	1	0,005
CAGGC/GTCCG	6	0,030
CATTC/GTAAG	1	0,005
CATTT/GTAAA	1	0,005
CCACT/GGTGA	1	0,005
CCAGG/GGTCC	1	0,005
CCCGT/GGGCA	1	0,005
CCCTC/GGGAG	2	0,010
CCTCC/GGAGG	4	0,020
CCTCT/GGAGA	1	0,005
CGGAC/GCCTG	1	0,005
CTACC/GATGG	1	0,005
CTAGC/GATCG	1	0,005
CTTAA/GAATT	1	0,005
CTTCC/GAAGG	1	0,005
CGGGA/GCCCT	2	0,010
CTAAC/GATTG	1	0,005
CTCCC/GAGGG	1	0,005
CTCTA/GAGAT	1	0,005
CTCTT/GAGAA	1	0,005
CTTCT/GAAGA	3	0,015
CTTGT/GAACA	1	0,005
CTTTT/GAAAA	19	0,094

4 CONCLUSÕES GERAIS

O tamanho do genoma de *Stryphnodendron adstringens* foi estimado em aproximadamente 669 Mbp por citometria de fluxo e 624 Mbp pela análise de distribuição de *k-mers*. O sequenciamento de DNA de duas bibliotecas usando a plataforma Illumina HiSeq 2500 permitiu a montagem de aproximadamente 495 Mbp de sequências nucleares. O uso de uma biblioteca de sequenciamento se mostrou eficiente para a montagem do genoma cloroplastidial em uma única sequência com 162.169 pares de base. O *draft* do genoma nuclear permitiu a identificação e caracterização de regiões microssatélites e elementos transponíveis, bem como a identificação de genes e transcritos com similaridade com as via de biossíntese de taninos. A montagem do genoma cloroplastidial possibilitou a anotação de microssatélites e estruturas de repetição, anotação de genes e comparação com o genoma cloroplastidial de outras espécies da subfamília Caesalpinioideae: clado mimosoide. Os resultados aqui obtidos podem ser de grande utilidade para estudos futuros, e podem ser considerados para o desenvolvimento de ferramentas genômicas e genéticas para estudos de genética de populações e filogenéticos para a espécie *S. adstringens*.

5 REFERÊNCIAS

- BRIDGEWATER, S.; RATTER, J. A.; RIBEIRO, J. F. Biogeographic patterns, β -diversity and dominance in the cerrado biome of Brazil. **Biodiversity & Conservation**, v. 13, n. 12, p. 2295–2317, 2004.
- CARVALHO, A. G. et al. Use of tannin adhesive from *Stryphnodendron adstringens* (Mart.) Coville in the production of OSB panels. **European Journal of Wood and Wood Products**, v. 72, n. 4, p. 425–432, 2014.
- CHEN, S. et al. An introduction to the medicinal plant genome project. **Frontiers of medicine**, v. 5, n. 2, p. 178–184, 2011.
- CHEN, S. et al. Genome sequence of the model medicinal mushroom *Ganoderma lucidum*. **Nature Communications**, v. 3, n. 1, p. 913, 2012.
- CORRÊA, V. S. et al. Geographical variation and quality assessment of *Stryphnodendron adstringens* (Mart.) Coville within Brazil. **Genetic resources and crop evolution**, v. 59, n. 7, p. 1349–1356, 2012.
- DIAS, J. E.; LAUREANO, L. C. Farmacopeia Popular do Cerrado / Coordenação: Jaqueline Evangelista Dias e Lourdes Cardozo Laureano. Goiás: **Articulação Pacari (Associação Pacari)**, 1ª Edição, 1ª Reimpressão, 2010, 352 p.
- DIERCKXSENS, N.; MARDULYN, P.; SMITS, G. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. **Nucleic acids research**, v. 45, n. 4, p. e18–e18, 2016.
- FEITOSA, I. S.; ALBUQUERQUE, U. P.; MONTEIRO, J. M. Knowledge and extractivism of *Stryphnodendron rotundifolium* Mart. in a local community of the Brazilian Savanna, Northeastern Brazil. **Journal of ethnobiology and ethnomedicine**, v. 10, n. 64, p. 64, 2014.
- GLASENAPP, J. S. et al. Characterization of diversity and genetic structure in natural populations of *Stryphnodendron adstringens* (Mart.) Coville by means of allozyme markers. **Revista Brasileira de Plantas Mediciniais**, v. 16, n. 2, p. 216–224, 2014.
- KELLNER, F. et al. Genome guided investigation of plant natural product biosynthesis. **The Plant Journal**, v. 82, n. 4, p. 680–692, 2015.
- LANCHOTI FIORI, G. M. et al. Antimicrobial Activity and Rates of Tannins in *Stryphnodendron adstringens*; Mart. Accessions Collected in the Brazilian Cerrado. **American Journal of Plant Sciences**, v. 04, n. 11, p. 2193–2198, 2013.
- LIU, L. et al. Comparison of Next-Generation Sequencing Systems. **Journal of Biomedicine and Biotechnology**, v. 2012, p. 1–11, 2012.
- MARDIS, E. R. A decade's perspective on DNA sequencing technology. **Nature**, v. 470, n. 7333, p. 198–203, 2011. Nature Publishing Group.
- MARDIS, E. R. Next-generation sequencing platforms. **Annual review of analytical chemistry**, v. 6, p. 287–303, 2013.
- MENDONÇA, P.C. **Caracterização da diversidade genética de *Stryphnodendron adstringens* (Mart.) Coville por marcador molecular AFLP e transferência de**

microssatélites. 2011. 87f. Tese (Doutorado em Agronomia) - Universidade Estadual Paulista, Botucatu, São Paulo, 2011.

MENDONÇA, P. C. et al. Genetic diversity of *Stryphnodendron adstringens* (Mart.) Coville determined by AFLP molecular markers. **Biochemical systematics and ecology**, v. 41, p. 16–20, 2012.

METZKER, M. L. Sequencing technologies—the next generation. **Nature reviews genetics**, v. 11, n. 1, p. 31, 2010.

SCALON, V. R. **Revisão taxonômica do gênero *Stryphnodendron* Mart.(leguminosae-mimosoideae)**, 2007. Universidade de São Paulo.

SCHUSTER, S. C. Next-generation sequencing transforms today's biology. **Nature methods**, v. 5, n. 1, p. 16–18, 2008.

SOUZA, C. D.; FELFILI, J. M. Uso de plantas medicinais na região de Alto Paraíso de Goiás, GO, Brasil. **Acta Botanica Brasilica**, v. 20, n. 1, p. 135–142, 2006.

SOUZA, T. M. et al. Bioprospecção de atividade antioxidante e antimicrobiana da casca de *Stryphnodendron adstringens* (Mart.) Coville (Leguminosae-Mimosoidae). **Revista de Ciencias Farmaceuticas Basica e Aplicada**, v. 28, n. 2, p. 221–226, 2007.

SOUZA, T. M. et al. Avaliação da atividade anti-séptica de extrato seco de *Stryphnodendron adstringens* (Mart.) Coville e de preparação cosmética contendo este extrato. **Brazilian Journal of Pharmacognosy**, v. 17, n. 1, p. 71–75, 2007.

TONTI-FILIPPINI, J. et al. What can we do with 1000 plastid genomes? **Plant Journal**, v. 90, n. 4, p. 808–818, 2017.

TWYFORD, A. D.; NESS, R. W. Strategies for complete plastid genome sequencing. **Molecular Ecology Resources**, v. 17, n. 5, p. 858–868, 2017.

VAN BAKEL, H. et al. The draft genome and transcriptome of *Cannabis sativa*. **Genome Biology**, v. 12, n. 10, p. R102, 2011.

VAN DIJK, E. L. et al. Ten years of next-generation sequencing technology. **Trends in Genetics**, v. 30, n. 9, p. 418–426, 2014.

VIEIRA, R. F.; CAMILLO, J.; CORADIN, L. Espécies nativas da flora brasileira de valor econômico atual ou potencial: plantas para o futuro: Região Centro-Oeste. **Embrapa Recursos Genéticos e Biotecnologia-Livro científico (ALICE)**, 2016.

XU, H. et al. Analysis of the Genome Sequence of the Medicinal Plant *Salvia miltiorrhiza*. **Molecular Plant**, 6. jun. 2016.