UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA E BIOLOGIA MOLECULAR

ADRIANA MARIA ANTUNES TAQUARY

Tamanho, montagem *de novo* e anotação do genoma de *Dipteryx alata* (Leguminosae)

Goiânia

2017

sistema de bibliotecas ufg

UFG

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO PARA DISPONIBILIZAR AS TESES E DISSERTAÇÕES ELETRÔNICAS 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** **[X] Tese**

2. Identificação da Tese ou Dissertação

Nome completo do autor: Adriana Maria Antunes Taquary

Título do trabalho: Tamanho, montagem *de novo* e anotação do genoma de *Dipteryx alata* (Leguminosae)

3. Informações de acesso ao documento:

Concorda com a liberação total do documento [X] SIM [ ] NÃO[1]

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

Adriana Maria Antunes Taquary
Assinatura do (a) autor (a)

Data: 08 / 05 / 2017

[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.

ADRIANA MARIA ANTUNES TAQUARY

Tamanho, montagem *de novo* e anotação do genoma de *Dipteryx alata* (Leguminosae)

Orientanda - Doutoranda: Adriana Maria Antunes Taquary

Orientadora: Profa. Dra. Mariana Pires de Campos Telles

Tese apresentada ao Programa de Pós-Graduação em Genética e Biologia Molecular da Universidade Federal de Goiás, como parte dos requisitos para obtenção do título de Doutora.

Março

2017

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

Maria Antunes Taquary, Adriana
Tamanho, montagem de novo e anotação do genoma de Dipteryx alata (Leguminosae) [manuscrito] / Adriana Maria Antunes Taquary. - 2017.
CXXXVII, 137 f.: il.

Orientador: Profa. Dra. Mariana Pires de Campos Telles.
Tese (Doutorado) - Universidade Federal de Goiás, Instituto de Ciências Biológicas (ICB), Programa de Pós-Graduação em Genética e Biologia Molecular, Goiânia, 2017.
Bibliografia. Apêndice.

1. Número cromossômico. 2. Tamanho genômico. 3. Elementos repetitivos. 4. Genes. 5. Polimorfismo de nucleotídeo único. I. Pires de Campos Telles, Mariana, orient. II. Título.

CDU 575

SERVIÇO PÚBLICO FEDERAL
UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA E BIOLOGIA MOLECULAR

ATA DA SESSÃO PÚBLICA DE DEFESA DE TESE DE Nº 004

Aos vinte e quatro dias do mês de abril do ano de dois mil e dezessete, às 08hs00min, no Anfiteatro do ICB II/UFG, reuniram-se os componentes da banca examinadora: **Profa. Dra. Mariana Pires de Campos Telles; Profa. Dra. Tereza Cristina de Oliveira Borba; Profa.Dra. Luciane Madureira Almeida; Prof. Dr. Evandro Novaes e Profa. Dra. Thannya Nascimento Soares** para, em sessão pública presidida pela primeira examinadora citada, procederem à avaliação da defesa de tese intitulada: **"TAMANHO, MONTAGEM DE NOVO E ANOTAÇÃO DO GENOMA DE** *Dipteryx alata* **(Leguminosae)"**, em nível de doutorado, área de concentração em **Genética e Biologia Molecular**, de autoria de **Adriana Maria Antunes**, discente do Programa de Pós-Graduação em Genética e Biologia Molecular da Universidade Federal de Goiás. A sessão foi aberta pela presidente da banca, Profa. Dra. Mariana Pires de Campos Telles, que fez a apresentação formal dos membros da banca. A palavra, a seguir, foi concedida à autora da tese que, em cerca de 35 minutos, procedeu à apresentação de seu trabalho. Terminada a apresentação, cada membro da banca arguiu a examinada, tendo-se adotado o sistema de diálogo sequencial. Terminada a fase de arguição, procedeu-se à avaliação da tese. Tendo-se em vista o que consta na Resolução nº 1294 de 06 de Junho de 2014 do Conselho de Ensino, Pesquisa, Extensão e Cultura (CEPEC), que regulamenta o Programa de Pós-Graduação em Genética e Biologia Molecular, a tese foi aprovada, considerando-se integralmente cumprido este requisito para fins de obtenção do título de Doutor (a) em Genética e Biologia Molecular pela Universidade Federal de Goiás. A conclusão do curso dar-se-á quando da entrega da versão definitiva da tese na secretaria do programa, com as devidas correções sugeridas pela banca examinadora, no prazo de trinta dias a contar da data da

SERVIÇO PÚBLICO FEDERAL
UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA E BIOLOGIA MOLECULAR

defesa. Cumpridas as formalidades de pauta, às 12 horas e 16 minutos, encerrou-se a sessão de defesa e, para constar, eu, Gleizilene Braz Pereira dos Santos, Assistente em Administração da Universidade Federal de Goiás, lavrei a presente ata que, após lida e aprovada, será assinada pelos membros da banca examinadora em três vias de igual teor.

Profa. Dra. Mariana Pires de Campos Telles
Presidente da Banca
ICB/UFG

Profa. Dra. Tereza Cristina de Oliveira Borba
EMBRAPA ARROZ E FEIJÃO/GO

Profa. Dra. Luciane Madureira Almeida
UEG/GO

Prof. Dr. Evandro Novaes
ICB/UFG.

Profa. Dra. Thannya Nascimento Soares
ICB/UFG

AGRADECIMENTOS

Agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela bolsa de estudos de Doutorado. Agradeço também à CAPES, ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e a Fundação de Amparo à Pesquisa do Estado de Goiás (FAPEG) pelo suporte financeiro do projeto no qual esse trabalho está inserido (CAPES e GENPAC-02 e 03 CNPq/Fapeg/PRO-CENTRO-OESTE 563839/2010-4 e 201110267000125).

Agradeço a minha orientadora professora Mariana Pires de Campos Telles pela oportunidade de desenvolver esse trabalho e por todos os ensinamentos. Agradeço a toda à equipe do Laboratório de Genética & Biodiversidade (Intituto de Ciências Biológicas UFG) e do Laboratório de Genética e Genômica de Plantas (Escola de agronomia UFG) que direta ou indiretamente contribuiram para a realização desse trabalho. Agradeço em especial ao Rhewter Nunes, a Lays Karolina, a Stela Ribeiro, a Ivone de Bem, ao Arthur Tavares e ao Fernando Roa, colegas de doutorado que muito me ajudaram e ensinaram. Agradeço ao professor Gustavo Souza, do Laboratório de Citogenética e Evolução Vegetal da Universidade Federal de Pernambuco, pela hospitalidade e importante colaboração.

Agradeço a Deus por me guiar, iluminar e me dar forças para realização desse sonho. Agradeço aos meus pais (Ivan Antunes e Doraci Gonçalves) e ao meu irmão (Alisson Antunes) por todo esforço pela minha formação pessoal e profissional, por me incentivar e torcer pela minha vitória. Agradeço ao meu marido (Nelson Taquary) por dividir comigo as alegrias e angustias, por me apoiar, compreender e por acreditar tanto em mim.

RESUMO

Nos últimos anos houve um rápido aumento na disponibilidade e qualidade dos dados de sequenciamento e com isso ocorreu uma explosão de projetos de sequenciamento dos genomas de plantas. Nesse cenário, as análises genômicas vêm sendo caracterizadas como eficientes para gerar informações genéticas em larga escala, inclusive para espécies não modelos. *Dipteryx alata* é uma espécie de árvore não modelo endêmica do bioma Cerrado pertencente à família Leguminosae. Os objetivos deste trabalho foram estimar o número de cromossomos e o tamanho do genoma de *D. alata*, e também montar e anotar sequências dos genomas organelares e nuclear da espécie usando dados de sequenciamento Illumina. O tamanho do genoma de *D. alata* foi estimado como 1C = 0.825 pg, o que corresponde a um genoma haplóide de 807.2 MB com 2n=16 cromossomos. Foram montadas 275.709 sequências genômicas nucleares com N50 igual a 1598, o que corresponde a 355MB e 44% do genoma inteiro. Nas sequências nucleares foram anotados 21.981 regiões microssatélites, das quais 49,3% possuem motivos dinucleotídeos, 42,7% trinucleotídeo e 4% tetranucleotídeo. Elementos transponíveis (TEs) foram encontrados em 39,29% das sequências analisadas, o que corresponde a 421.701 TEs. Os retrotransposons LTR (*gypsy* e *copia*) foram os TEs mais abundantes nas sequências nucleares. Foram anotados 1.431 genes de RNAs não traduzidos em proteínas, sendo 176 rRNAs, 189 tRNAs, 477 snRNAs, 8 snoRNAs, 466 miRNAs e 115 lncRNAs. Foram anotados também 62.200 genes codificadores de proteínas com tamanho médio de 1.156 pb. O número estimado de mRNAs transcritos pelo conjunto de genes nucleares anotados foi igual a 160.450, dos quais 131.228 apresentaram similaridade significativa com sequências já conhecidas e 84.793 foram classificadas funcionalmente nos termos do Gene Ontology. Um total de 736.787 SNPs e 90.803 InDels foram descobertos nas sequências nucleares. Foi identificada uma média de 1 SNP a cada 189 pb do genoma e a razão entre as mutações de transição (Ts) e transversão (Tv) foi de 1,58. Uma porcentagem de 46,5% dos SNPs ocorreu em contexto gênico e os efeitos dos SNPs foram anotados principalmente em éxons e regiões intergênicas. Foram montados 110 KB de sequências cloroplastidiais com N50 de 2.384 pb e 327 KB de sequências mitocondriais com N50 de 1.784 pb. Foram anotados genes de 3 rRNA, 13 tRNA, 6 miRNA e 20 lncRNA para o cloroplasto e genes de 4 rRNA, 26 tRNA, 7 miRNA e 54 lncRNA para a mitocôndria. Para o cloroplasto foram preditos 20 genes codificantes de proteínas com tamanho médio de 2.374 pb e para a mitocôndria foram preditos 176 genes com tamanho médio de 1.279 pb. O número estimado de mRNAs transcritos por esse conjunto de genes foi igual a 63 e 525 para cloroplasto e mitocôndria, respectivamente. Foram anotados também 39 regiões microssatélites e 4 TEs nas sequências cloroplastidiais e 158 regiões microssatélites e 26 TEs nas sequências mitocondriais. Este trabalho, que pode ser considerado um dos primeiros estudos genômicos para espécies do Cerrado, representa um grande avanço nos conhecimentos sobre a estrutura e a organização do genoma de *D. alata*. Os resultados obtidos abrem caminho para novas investigações genéticas e genômicas para a espécie.

Palavras-chave: número cromossômico, tamanho genômico, elementos repetitivos, genes, polimorfismos de nucleotídeo único.

ABSTRACT

In recent years there has been a rapid increase in the availability and quality of sequencing data and with this an explosion of projects of sequencing of the genomes of plants occurred. In this scenario, genomic analyzes have been characterized as efficient to generate genetic information on a large scale, including for non-model species. *Dipteryx alata* is a non-model tree species endemic to the Cerrado biome belonging to the Leguminosae family. The objectives of this work were to estimate the number of chromosomes and the size of the genome of *D. alata*, and also assemble and annotate sequences of the genomes organelles and nuclear of the species using Illumina sequencing data. The size of the genome of *D. alata* was estimated as 1C = 0.825 pg, which corresponds to a haploid genome of 807.2 MB with 2n = 16 chromosomes. Were assembled 275,709 nuclear genomic sequences with N50 equal to 1598, which corresponds to 355MB and 44% of the whole genome. In the nuclear sequences, 21,981 microsatellite regions were annotated, of which 49.3% had dinucleotide motifs, 42.7% trinucleotide motifs and 4% tetranucleotide motifs. Transposable elements (TEs) were found in 39.29% of the sequences analyzed, corresponding to 421,701 TEs. LTR retrotransposons (*gypsy* and *copy*) were the most abundant TEs in nuclear sequences. Were annotated 1,431 RNA genes non-translated into proteins, being 176 rRNAs, 189 tRNAs, 477 snRNAs, 8 snoRNAs, 466 miRNAs and 115 lncRNAs. Were annotated also 62,200 protein coding genes with an average size of 1,156 bp. The estimated number of mRNAs transcribed by the set of annotated nuclear genes was 160,450, of which 131,228 showed significant similarity with known sequences and 84,793 were classified functionally in the Gene Ontology terms. A total of 736,787 SNPs and 90,803 InDels were discovered in the nuclear sequences. A mean of 1 SNP was identified for each 189 bp of the genome and the ratio between the transition (Ts) and transversion (Tv) mutations was 1.58. A percentage of 46.5% of the SNPs occurs in the genic context and the effects of the SNPs were annotated mainly in exons and intergenic regions. Were assembled 110 KB of chloroplastid sequences with N50 of 2,384 bp and 327 KB of mitochondrial sequences with N50 of 1,784 bp. Were annotated genes of 3 rRNA, 13 tRNA, 6 miRNA and 20 lncRNA for the chloroplast and genes of 4 rRNA, 26 tRNA, 7 miRNA and 54 lncRNA for the mitochondria. For the chloroplast were predicted 20 protein coding genes with a mean size of 2,374 bp and for mitochondria were predicted 176 genes with a mean size of 1,279 bp. The estimated number of mRNAs transcribed by this gene set was 63 and 525 for chloroplast and mitochondria respectively. Were annotated 39 microsatellite regions and 4 TEs in the chloroplastid sequences and 158 microsatellite regions and 26 TEs in the mitochondrial sequences. This work, which can be considered one of the first genomic studies for Cerrado species, represents a great advance in the knowledge on the structure and organization of the *D. alata* genome. The obtained results open the way for further genetic and genomic investigation for the species.

Key words: chromosome number, genomic size, repetitive elements, genes, single nucleotide polymorphisms.

SUMÁRIO

LISTA DE TABELAS.....14
LISTA DE FIGURAS 15

INTRODUÇÃO GERAL.....17
REVISÃO DA LITERATURA.....19
SEQUENCIAMENTO ILLUMINA.....19
MONTAGEM DE GENOMAS.....20
ANOTAÇÃO DO GENOMA DE PLANTAS.....22
Anotação de genes.....22
Anotação de sequências repetitivas.....23
Anotação de polimorfismos de nucleotídeo único (SNPs).....24
OBJETIVO GERAL.....26
OBJETIVOS ESPECIFICOS.....26
REFERÊNCIAS.....27

CAPÍTULO 1: TENDÊNCIAS EM TAMANHO DO GENOMA E EVOLUÇÃO DO NÚMERO DE CROMOSSOMOS NA SUBFAMÍLIA DE PAPILIONOIDEAE (LEGUMINOSAE) COM ÊNFASE NO CLADO BASAL ADA.....31
RESUMO.....31
INTRODUÇÃO.....32
MATERIAL E METÓDOS.....33
Material vegetal.....33
Análise de Citometria de fluxo.....33
Determinação do número de cromossomos de *D. alata*.....34
Frequências e correlações entre 2C e 2n.....34
Análise filogenética.....34
Reconstrução do estado ancestral e inferência do número de cromossomos / tamanho do genoma.....35
RESULTADOS.....35
Tamanho do genoma e número de cromossomos.....35
Variação do tamanho genômico e do número de cromossomos em Leguminosae.....35
Correlação entre os valores 2n e 2C.....37
Análise de ancestralidade do valor 2C e número de cromossomos.....38
DISCUSSÃO.....40
2C e 2n em Leguminosae.....40
O clado ADA e *Dipteryx alata* como um modelo para Papilionoideae genoma análise.....41
CONCLUSÃO.....41
MATERIAL SUPLEMENTAR.....42
REFERÊNCIAS.....42

CAPÍTULO 2: CARACTERIZAÇÃO DE ELEMENTOS REPETITIVOS EM SEQUÊNCIAS GENÔMICAS NUCLEARES DE *Dipteryx alata* (LEGUMINOSAE).....46
RESUMO.....46
INTRODUÇÃO.....47
MATERIAL E MÉTODOS.....49
Material vegetal e extração de DNA.....49
Preparo das bibliotecas de DNA e sequenciamento Illumina.....49
Análise de qualidade e montagem de sequências genômicas nucleares.....49
Anotação de regiões microssatélites e desenho de primers.....50
Anotação de TEs e desenho de primers.....51
RESULTADOS.....52
Sequenciamento e montagem das sequências genômicas.....52
Anotação de regiões microssatélites em *Dipteryx alata*.....53
Desenho de primers para regiões microssatélites em *Dipteryx alata*.....56
Anotação de TEs em *Dipteryx alata*.....56
Desenho de primers para retroelementos LTR *gypsy* e *copia* em *Dipteryx alata*.....60
DISCUSSÃO.....60
CONCLUSÃO.....64
MATERIAL SUPLEMENTAR.....64
REFERÊNCIAS.....65

CAPÍTULO 3: ANOTAÇÃO ESTRUTURAL E FUNCIONAL DE GENES NO GENOMA NUCLEAR DE *Dipteryx alata* (LEGUMINOSAE).....70
RESUMO.....70
INTRODUÇÃO.....71
MATERIAL E MÉTODOS.....72
Extração de DNA e sequenciamento Illumina.....72
Análise de qualidade e montagem das sequências genômicas nucleares.....73
Anotação de genes de RNAs não traduzidos.....73
Anotação de genes codificantes de proteínas.....73
RESULTADOS.....74
Sequenciamento e montagem das sequências genômicas.....74
Anotação de genes codificantes de RNAs não traduzidos.....74
Anotação estrutural de genes codificantes de proteínas.....74
Anotação funcional de genes codificantes de proteínas.....75
DISCUSSÃO.....78
CONCLUSÃO.....80
MATERIAL SUPLEMENTAR.....81
REFERÊNCIAS.....81

CAPÍTULO 4: DESCOBERTA E CARACTERIZAÇÃO DE POLIMORFISMOS DE NUCLEOTÍDEO ÚNICO NO GENOMA NUCLEAR DE *Dipteryx alata* (LEGUMINOSAE).....84
RESUMO.....84
INTRODUÇÃO.....85
MATERIAL E METÓDOS.....86
Extração de DNA e sequenciamento Illumina.....86
Análise de qualidade e montagem de sequências genômicas nucleares.....86
Descoberta e caracterização de SNPs.....87
RESULTADOS.....88
Montagem das sequências.....88
Descoberta de SNPs.....89
Anotação das variantes e seus efeitos.....91
DISCUSSÃO.....92
CONCLUSÃO.....93
MATERIAL SUPLEMENTAR.....93
REFERÊNCIAS.....94

CAPÍTULO 5: MONTAGEM DE NOVO E CARACTERIZAÇÃO DE SEQUÊNCIAS GENÔMICAS CLOROPLASTIDIAIS E MITOCONDRIAIS DE *Dipteryx alata* (LEGUMINOSAE) USANDO DADOS DE SEQUENCIAMENTO DE NOVA GERAÇÃO.....97
RESUMO.....97
INTRODUÇÃO.....98
MATERIAL E METÓDOS.....99
Material vegetal e extração de DNA.....99
Preparo das bibliotecas de DNA e sequenciamento Illumina.....100
Análise de qualidade e montagem de sequências genômicas organelares.....100
Anotação de genes de RNAs não codificantes (ncRNAs).....100
Anotação de genes codificantes de proteínas.....101
Anotação de regiões microssatélites.....101
Anotação de TEs.....101
RESULTADOS.....102
Montagem das sequências.....102
Anotação gênica de RNAs não traduzidos.....103
Anotação gênica de RNAs codificantes de proteínas.....104
Anotação funcional de genes codificantes de proteínas.....105
Anotação de microssatélites.....108
Anotação de elementos transponíveis.....109
DISCUSSÃO.....109
CONCLUSÃO.....112
MATERIAL SUPLEMENTAR.....112
REFERÊNCIAS.....113

CONSIDERAÇÕES FINAIS.....118
APÊNDICE 119

LISTA DE TABELAS

CAPÍTULO 2

Tabela 1 – Características das montagens realizadas para D. alata usando dados de sequenciamento Illumina.....52

Tabela 2 – Frequência dos motivos de repetição em sequências genômicas de *D. alata*.....55

Tabela 3 – Anotação do total de TEs em sequências de *D. alata* usando RepeatModeler e RepeatMasker.....57

Tabela 4 – Anotação dos TEs identificados exclusivamente para *D. alata* usando RepeatModeler e RepeatMasker.....58

Tabela 5 – Tamanhos mínimos, máximos e médios dos TEs identificados em *D. alata*.....59

CAPÍTULO 3

Tabela 1 - Anotação estrutural de genes codificantes de RNAs não traduzidos em *D. alata*..74

Tabela 2 – Anotação estrutural de genes codificantes de proteínas em *D. alata*.....75

CAPÍTULO 4

Tabela 1 - Número de polimorfismos de nucleotídeo único (SNPs) e indels nas sequências de *D. alata* identificados em cada etapa do pipeline do GATK.....90

Tabela 2 - Quantidades de transição e transversão para os SNPs identificados em *D. alata*...91

Tabela 3 – Número de efeitos por tipos identificados na anotação funcional de SNPs em *D. alata*.....92

CAPÍTULO 5

Tabela 1 – Qualidade das montagens de sequências genômicas organelares de *Dipteryx alata*.....102

Tabela 2 - Anotação estrutural de genes codificantes de RNAs não traduzidos em sequências organelares de *D. alata*.....104

Tabela 3 – Anotação estrutural de genes organelares codificantes de proteínas em *D. alata*.....105

Tabela 4 - Quantidade de microssatélites encontrados em sequências organelares de *Dipteryx alata* por tamanho dos motivos de repetição.....108

Tabela 5 - Motivos de repetição mais frequentes em sequências organelares de *D. alata*.....108

Tabela 6 – Anotação de TEs em sequências organelares de *D. alata* usando RepeatMasker.....109

LISTA DE FIGURAS

CAPÍTULO 1

Figura 1 - Histogramas de fluorescência relativa (unidades arbitrárias) de núcleos corados com iodeto de propídio por Citometria de Fluxo. Fig1a Da - *Dipteryx alata*. Gm - *Glycine max*, padrão de referência; Fig1b Prometafase de D. alata 2n = 16 corada com DAPI. Cor invertida. Bar = 5 um; Fig1c Ac - *Amburana cearensis*, Sl - *Solanum lycopersicum*, padrão de referência.....36

Figura 2 - Dot plot distribuição de 2n para Papilionoideae; Y contém nome, três mais freqüentes 2n e tamanho da amostra. Tamanho do ponto em diferentes escalas dependendo do tamanho da amostra.....37

Figura 3. Gráficos de violino mostrando a densidade de 2C em Papilionoideae. O eixo y contém os três picos mais altos em pg e o tamanho da amostra. Largura de banda em função de densidade personalizável nos recursos on-line.....38

Figura 4 - Evolução do tamanho do genoma e número de cromossomos em espécies de Leguminosae. Valores de 2n e 2C são mostrados por cores nos nós e ramos antigos.....39

CAPÍTULO 2

Figura 1 - Distribuição dos tamanhos das sequências nucleares na montagem realizada com software Dipspades.....53

Figura 2 – Distribuição dos microssatélites encontrados em sequências genômicas de *Dipteryx alata* por tamanho de motivo de repetição.....54

Figura 3 – Distribuição dos microssatélites encontrados em *D. alata* por tamanho de motivo de repetição e número de repetições em tandem.....54

Figura 4 - Diagrama de Venn com o número de TEs identificados nas sequências de *D. alata* usando as duas metodologias de anotação.....59

CAPÍTULO 3

Figura 1 – Distribuição das espécies Top-Hits do Blast.....76

Figura 2 – Anotação funcional de sequências gênicas nucleares codificantes de proteínas de *Dipteryx alata*.....77

CAPÍTULO 4

Figura 1 - Distribuição dos tamanhos das sequências nucleares na montagem de *D. alata*.....89

Figura 2 - Diagrama de Venn apresentando o número de SNPs após cada etapa de filtragem. F1 - Filtro após a primeira chamada de variantes. F2 - Filtro após a primeira recalibração. F3 - Filtro após a segunda recalibração.....90

Figura 3 - Distribuição do percentual de efeitos dos SNPs baseada na região genômica onde ocorre em *D. alata*.....91

CAPÍTULO 5

Figura 1 - Distribuição dos tamanhos das sequências organelares cloroplastidiais na montagem de *D. alata* com o software Spades.....102

Figura 2 - Distribuição dos tamanhos das sequências organelares mitocondriais na montagem de *D. alata* com o software Spades.....103

Figura 3 – Distribuição das espécies Top-Hits do Blast das sequências cloroplastidiais.....105

Figura 4 – Distribuição das espécies Top-Hits do Blast das sequências mitocondriais.....106

Figura 5 – Distribuição dos dados referentes à análise funcional de 63 sequências cloroplastidiais e 525 sequências mitocondriais.....106

Figura 6 – Anotação funcional de sequências gênicas cloroplastidiais codificantes de proteínas de *Dipteryx alata*.....107

Figura 7 – Anotação funcional de sequências gênicas mitocondriais codificantes de proteínas de *Dipteryx alata*.....107

INTRODUÇÃO GERAL

O Cerrado é um bioma do tipo savana que ocorre no Brasil e possui grande riqueza de espécies e endemismo (Mendonça et al 2008). Apesar de ser considerado um *hotspot* de biodiversidade, muitas espécies do Cerrado estão ameaçadas devido à fragmentação do habitat e alteração da paisagem natural do bioma causada pela expansão da agricultura (Smith et al 2012; Grecchi et al 2014). Nesse cenário, existe a necessidade, para fins de uso e de conservação, de se obter novos conhecimentos sobre a biologia das espécies de plantas. Entre as espécies endêmicas do Cerrado está *Dipteryx alata*, uma espécie de árvore pertencente à família Leguminosae.

D. alata, conhecida popularmente como Baru, possui uma ampla distribuição geográfica, sendo encontrado em áreas brasileiras de Cerrado nos estados de São Paulo, Minas Gerais, Mato Grosso, Mato Grosso do Sul, Tocantins e Goiás. É encontrada também no Paraguai e Bolívia (Ratter et al 2000). *D. alata* possui porte arbóreo e apresenta altura média de 15 metros e máxima de 25 metros. Possui copa larga com até 11 metros de diâmetro, folhas compostas e alternas com até 12 folíolos, limbo geralmente oblongo, inflorescência do tipo panícula, flores hermafroditas, fruto do tipo drupa com até 5cm de comprimento e semente elipsoide. A floração ocorre de novembro a fevereiro e a formação dos frutos inicia-se em dezembro. Cada árvore de *D. alata* pode produzir ate 5000 frutos por estação. A dispersão dos frutos ocorre por barocoria ou por zoocoria (Sano et al 2004).

D. alata possui importância econômica principalmente devido à polpa (mesocarpo) e as amêndoas (sementes) de seus frutos, os quais são comestíveis e utilizados *in natura* ou na fabricação de sorvetes, licores, castanhas torradas, e outros produtos por pequenas e médias indústrias. Além do uso alimentício, *D. alata* tem utilidade medicinal, forrageira, na recuperação de áreas degradadas, no paisagismo e na extração de madeira. O oléo da semente de *D. alata* é usado na medicina popular como anti-reumático e a casca do tronco na cura de dores na coluna. *D. alata* possui utilidade forrageira visto que a polpa de seus frutos, rica em calorias, é consumida pelos animais e contribui com nutrientes, principalmente nas estações secas quando a disponibilidade de forragem é baixa. *D. alata* pode ser usada no paisagismo, pois a copa larga das árvores fornece boa sombra durante as estações quentes, e na recuperação de áreas degradadas porque possui bom crescimento e baixa exigência de adubação. A madeira de *D. alata* é comercializada e utilizada, por exemplo, na construção civil. Isso porque a madeira é compacta e possui alta densidade e durabilidade (Sano et al 2004). Devido à importância de *D. alata* para as comunidades locais no Brasil Central é relevante conhecer as características genéticas e genômicas da espécie.

São encontrados na literatura estudos genéticos para *D. alata* com foco principalmente no desenvolvimento de marcadores microssatélites em pequena escala e estudos populacionais (Soares et al 2008a; Soares et al 2008b; Soares et al 2012; Collevatti et al 2013; Soares et al 2015). As investigações genômicas disponíveis ainda são raras, apenas 29 sequências encontram-se depositadas no banco de dados GenBank da base NCBI (*National Center for Biotechnology Information*) para essa espécie. Assim, as estratégias de sequenciamento de DNA em larga escala por plataformas de nova geração e a anotação do genoma são promissoras para gerar novos conhecimentos para está espécie.

Os avanços nas tecnologias de sequenciamento, com as plataformas de sequenciamento de nova geração (*Next Generation Sequencing* - NGS), tem permitido gerar sequências de nucleotídeos com alto rendimento, baixas taxas de erro, em um curto espaço de tempo e menor custo por base sequenciada (Mardis 2008; Metzker 2010). Assim, as tecnologias NGS provocaram uma explosão de projetos de sequenciamento dos genomas de plantas e muitos trabalhos foram e ainda são realizados com o objetivo de montar e anotar os genomas das espécies. Milhares de sequências estão sendo publicadas em banco de dados e as análises genômicas vêm sendo caracterizadas como eficientes para geração de informações genéticas em larga escala, inclusive para espécies não modelos como *D. alata* (Mittler et al 2013; Lohse et al 2014).

Este trabalho foi realizado com o propósito de estimar o número de cromossomos e o tamanho do genoma de *D. alata*, sequenciar parte dos genomas nuclear e organelares da espécie usando a tecnologia Illumina, montar rascunhos dos genomas e anotar microssatélites, elementos transponíveis, genes e polimorfismos de nucleotídeo único. Os resultados obtidos, além de gerar novos conhecimentos sobre a arquitetura genômica de *D. alata*, ampliarão as possibilidades de pesquisa para a espécie. Os dados de número cromossômico e tamanho do genoma de *D. alata* contribuirão com novos dados citogenéticos para a família Leguminosae e permitirá a realização de análises comparativas com outras espécies próximas filogeneticamente. As sequências de regiões microssatélites, elementos transponíveis e Polimorfimos de nucleotídeo único poderão ser usados para o desenvolvimento de marcadores moleculares, ferramentas valiosas para análises de genética e genômica de populações desta espécie do Cerrado. As regiões gênicas de *D. alata* identificadas e caracterizadas funcionalmente poderão elucidar redes de regulação de genes e vias metabólicas, eventos de duplicação de todo o genoma e ainda serão utéis para a análise de sintenia baseada na comparação com o genoma de outras espécies da mesma família. Este pode ser considerado um dos primeiros estudos genômicos para plantas nativas do Cerrado brasileiro e os resultados obtidos abrem caminho para novas investigações para *D. alata*.

REVISÃO DA LITERATURA

SEQUENCIAMENTO ILLUMINA

Durante vários anos o sequenciamento Sanger foi o método dominante para análises de sequência de nucleotídeo. No entanto, mesmo com os avanços nos métodos de Sanger, tornou-se necessário o desenvolvimento de novas tecnologias de sequenciamento para se obter de forma eficiente um grande número de sequências (Metzker, 2010). O método de Sanger é considerado uma tecnologia de primeira geração e os novos métodos de sequenciamento são conhecidos como sequenciamento de nova geração (*Next Generation Sequencing* - NGS). O sequenciamento Sanger vem sendo substituído pelas tecnologias NGS devido ao seu custo inferior por bases de DNA sequenciada (Strickler et al 2012). Além disso, o NGS gera rapidamente, a partir de pequenas quantidades de material genético, uma enorme quantidade de dados de sequência (Ekblom e Galindo, 2011). Dessa forma é possível sequenciar em dias ou mesmo horas o que levaria anos pelo sequenciamento Sanger (Strickler et al 2012).

Várias tecnologias muito diferentes constituem o chamado sequenciamento de nova geração, cada uma das quais apresenta o seu conjunto próprio de características (Ekblom e Galindo, 2011). Dentre as várias tecnologias de NGS comercialmente disponíveis, o sequenciamento Illumina vem dominando o mercado. Os métodos da plataforma Illumina envolvem a quebra aleatória do DNA em tamanhos menores e a ligação das extremidades dos fragmentos a adaptadores. Tais etapas, realizadas durante o preparo de bibliotecas genômicas, são importantes porque a fragmentação do DNA minimiza os erros de sequenciamento e a ligação dos fragmentos a adaptadores permite que genomas complexos sejam amplificados por *primers* comuns. A Illumina usa o modelo de DNA amplificado para a preparação de reações NGS, estratégia realizada porque o sistema de imagem não consegue identificar eventos fluorescentes individuais e por isso exige a amplificação dos fragmentos de DNA para maximizar o sinal (Metzker, 2010).

A metodologia de PCR em fase sólida é utilizada pela Illumina para amplificação dos fragmentos. Nesse método os fragmentos de DNA são imobilizados em uma superfície de vidro chamada *flow cell*. Para isso os adaptadores ligados aos fragmentos de DNA de cadeia simples se ligam a oligonucleotideos da *flow cell* por complementariedade de bases, e em seguida os fragmentos se curvam e formam pontes de amplificação com *primers* adjacentes (Fedurco et al 2006). Cada *flow cell* pode possuir até oito canaletas, sendo que cada uma delas pode receber a mesma biblioteca de fragmentos ou bibliotecas diferentes (Mardis, 2008). Durante o processo de amplificação, a DNA polimerase produz múltiplas cópias do fragmento de DNA original e após a PCR os fragmentos imobilizados formam *clusters*, que são aglomerados de fragmentos clonados (Metzker, 2010). Ao final do processo cada *cluster* possui cerca de um milhão de cópias do fragmento original, quantidade suficiente para fornecer a intensidade de sinal para a detecção durante o sequenciamento (Mardis, 2008).

A plataforma Illumina faz o sequenciamento pelo método de ciclo de terminação reversível (Metzker, 2010). No método de ciclo de terminação reversível os quatro nucleotídeos fluorescentes (existem quatro cores, uma para cada nucleotídeo) são adicionados simultaneamente à *flow cell*. Em seguida, uma DNA polimerase incorpora um nucleotídeo

fluorescente à fita que está sendo sintetizada por complementariedade de bases com o fragmento molde imobilizado. Esse nucleotídeo possui um terminador reversível que promove a terminação da síntese, pois bloqueia a extremidade 3'OH que é necessária para a formação da próxima ligação fosfodiéster. Em seguida, os demais nucleotídeos não incorporados são lavados, é feita uma imagem para capturar a fluorescência, e posteriormente ocorre a clivagem do grupo de inibição da síntese e a regeneração do grupo 3'OH, após é feita uma lavagem adicional e o ciclo é repetido (Ju et al 2006; Metzker, 2010). A cada ciclo é feita a leitura de uma base da sequência do fragmento analisado. Dessa forma, a chave para esse método é o terminador reversível (Ju et al 2006). Os ciclos são repetidos até que seja realizada a leitura do comprimento em bases máximo para cada equipamento Illumina. Após a corrida, um algoritmo *base-calling* atribui valores de qualidade para cada base e remove as sequências de baixa qualidade e as sequências dos adaptadores (Mardis, 2008).

A tecnologia Illumina gera grandes quantidades de *reads* curtos em uma única corrida de sequenciamento, que ocorre na escala de dias ou mesmo horas, e apresenta baixas taxas de erros de sequenciamento em comparação a outras plataformas NGS, sendo os principais erros do tipo substituição de nucleotídeos (Ekblom e Galindo, 2011). A Illumina faz o sequenciamento *paired-end*, estratégia de sequenciamento onde as duas extremidades de cada fragmento de DNA são sequenciadas para produzir pares de *reads* (Martin e Wang, 2011; Fun et al 2013). Existem vários equipamentos Illumina, alguns deles são HiSeq 4000, HiSeq 3000, HiSeq 2500, iScan e MiSeq. Esses equipamentos diferem quanto ao custo para compra, tempo de corrida, tamanho dos *reads* sequenciados e quantidade de *reads* por corrida. Por exemplo, o HiSeq 4000 gera até 1500 GB de sequências por corrida (*reads paired end* de 150 pb), mais de 400 GB de rendimento diário, e o tempo de corrida varia de 1 a 3,5 dias. Já o MiSeq gera até 15 GB de sequências por corrida (*reads paired end* de 300 pb) e o tempo máximo de cada corrida é de 56 horas (http://www.illumina.com).

MONTAGEM DE GENOMAS

Após a geração das sequências curtas (*reads*) pelas plataformas de nova geração, as etapas de organização e análise dos dados são feitas usando ferramentas de bioinformática. Inicialmente os *reads* sequenciados passam por uma análise de qualidade e bases de baixa qualidade são removidas. A qualidade das bases é avaliada por meio do score chamado Phred que indica quanto confiável foi o resultado do sequenciamento para cada uma das bases que compõem o *read*. Assim, quanto maior o valor de Phred, maior a confiança de que determinada base teve sua atribuição correta no *read*. Exemplos de softwares para análise de qualidade e cortes de bases de baixa qualidade são o FastQC e o Trimommetic, respectivamente (Andrews, 2010; Bolger et al 2014).

Após os cortes de qualidade os *reads* são usados para a montagem (*assembly*) das sequências genômicas. A montagem consiste em ordenar os milhares de *reads* para obtenção das sequências originais e informativas do genoma (Schlebusch e Illing, 2012). O alinhamento é o princípio por trás da montagem de sequências genômicas e permite agrupar *reads* em *contig*s e *contigs* em *scaffolds*. *Contigs* são sequências consenso resultantes do alinhamento de *reads*. *Scaffolds*, também chamados de *supercontigs* ou *metacontigs*, são formados por *contigs* e por *gaps*, espaços que não foram sequenciados. Uma forma de

diminuir os gaps, que dificultam a montagem do genoma, é aumentar a cobertura ou profundidade de sequenciamento. A cobertura do sequenciamento refere-se à quantidade de vezes que o genoma foi sequenciado. Alta cobertura garante maior precisão no alinhamento dos *reads* e reduz a taxa de erros nas montagens (Miller et al 2010; Schatz et al 2010). A alta cobertura é importante também para lidar com a ambiguidade gerada pelas sequências repetitivas dos genomas. No entanto, algumas repetições genômicas podem ser indistinguíveis mesmo que cada base tenha sido sequenciada com alta cobertura (Nagarajan e Pop, 2013).

Além da alta qualidade das sequências e alta cobertura de sequenciamento, a montagem das sequências genômicas é facilitada pela existência de um genoma de referência para a espécie. Quando existe um genoma de referência os *reads* sequenciados podem ser alinhados a este genoma. A montagem com sequências de referência é útil sempre que existam referências confiáveis, no entanto, uma desvantagem é que essa estratégia não permite resolver grandes variações estruturais (Miller et al 2010). A montagem de sequências genômicas de espécies não modelo, que não possuem genoma de referência, é mais complexa. A estratégia mais eficiente para fazer a montagem de genomas de novo é a que usa como algoritmo os grafos de bruijn (Schatz et al 2010).

Os grafos de bruijn fazem a montagem das sequências através da decomposição dos *reads* em subsequências de tamanho K (K-mer). Os grafos são formados por vários nós que se interconectam, mostrando diferentes caminhos para a montagem da sequência. Em um grafo 5mer, por exemplo, um nó começa com as 4 últimas bases do nó anterior. Em uma situação perfeita a análise dos grafo faz voltar à sequência original, pois permite ler a sequência que gerou o grafo. Alguns grafos possuem muitas encruzilhadas e podem gerar milhões de sequências. Muitas vezes as encruzilhadas são geradas por erros de sequenciamento, que podem ser reconhecidos na estrutura dos grafos. Por exemplo, erros no final de um *read* criam k-mer que ocorrem somente uma vez e, portanto finalizam um caminho no grafo; erros no meio dos *reads* criam caminhos alternativos chamados bolhas. Após a formação dos grafos os softwares montadores executam uma fase de correção em que resolvem essas estruturas nos grafos considerando principalmente a cobertura de sequenciamento, enquanto os caminhos de baixa cobertura são desprezados (Schatz et al 2010).

Existem vários montadores que fazem a montagem de sequências genômicas *de novo* usando como algoritmo a construção de grafos de bruijn, alguns exemplos são Abyss (Simpson et al 2009), Masurca (Zimin et al 2013), SOAPdenovo (Luo et al 2012) e Dipspades (Safonova et al 2015). Apesar de todos esses montadores usarem a estratégia de grafos de bruijn para a montagem *de novo*, cada um lida de maneira diferente, por exemplo, com erros de sequenciamento, com coberturas baixas e não uniformes e com sequências repetitivas que são abundantes em genomas de plantas (Miller et al 2010). Como cada espécie possui uma estrutura genômica própria, é importante testar diferentes montadores para cada conjunto de dados visando escolher o de melhor desempenho (Ištvánek et al 2014).

A qualidade das montagens é avaliada pelo tamanho e pela precisão de seus *contigs* e *scaffolds* (Miller et al 2010). Várias métricas são usadas para avaliar a montagem como número de *scaffolds* gerados, tamanho total dos *scaffolds* em nucleotídeos, tamanho do maior e do menor *scaffold*, número de *scaffolds* maiores que 500nt, 1k, 10k, 100k e 1M,

média do tamanho dos *scaffolds*, N50 dos *scaffolds*, entre outras. O N50 é uma medida estatística de comprimento médio de sequências e é amplamente usado em genômica. No conjunto de sequências de comprimentos diferentes 50% das sequências são menores que o valor de N50 e 50% das sequências são maiores que o valor N50. Assim, a estatística N50 analisa quanto do genoma é coberto por *contigs* grandes (Schatz et al 2010).

ANOTAÇÃO DO GENOMA DE PLANTAS

A obtenção das sequências de DNA não fornece muitas informações biológicas de modo imediato e representa apenas o primeiro passo para a compreensão do genoma. A elucidação dos mecanismos envolvidos no funcionamento do organismo depende da precisão dos processos de anotação (Zieleziński et al 2012; Lohse et al 2014). Anotar o genoma significa tomar as sequências de ácidos nucleicos brutas obtidas pelos projetos de sequenciamento e interpretá-las de forma a extrair o seu significado biológico (Stein 2001). O processo de anotação visa fazer uma descrição detalhada de todos os elementos genéticos que constituem o genoma (Zieleziński et al 2012). Nesse contexto o trabalho de anotação é dividido em duas partes: anotação estrutural e anotação funcional. A anotação estrutural consiste na localização dos elementos que constituem o genoma, tais como os genes e os elementos repetitivos. A anotação funcional por sua vez visa identificar a função biológica da sequência, principalmente atribuir a função bioquímica aos produtos de genes (Aubourg et al 2001). Assim, a anotação das sequências de DNA é um pré-requisito para a exploração racional dos dados genômicos e permite investigar as diversas arquiteturas dos genomas de plantas (Caicedo e Purugganan 2005; Amar et al 2014).

Anotação de genes

O processo de anotação de sequências gênicas é um dos primeiros e mais importantes passos para o entendimento do genoma de uma espécie (Stein 2001). Em eucariotos os genes são divididos em genes codificantes de proteínas e genes não codificantes de proteínas. A maioria das funções biológicas do organismo é atribuída a genes que codificam RNAs mensageiros (mRNA) que são traduzidos em proteínas. No entanto, os genes de RNAs não traduzidos (ncRNA) também desempenham papéis importantes, pois esses RNAs participam dos mecanismos de expressão gênica. Os ncRNAs podem ser divididos em duas categorias com base em suas funções: house-keeping e ncRNAs regulatórios. Os ncRNAs house-keeping desempenham papéis estruturais e catalíticos, e nesse grupo incluem-se os RNAs transportadores (tRNAs) e RNAs ribossomais (rRNAs), ambos envolvidos na tradução, pequenos RNAs nucleares (snRNAs) envolvidos no processo de splicing do mRNA e pequenos RNAs nucleolares (snoRNAs) que participam do splicing do rRNA. Os ncRNAs regulatórios podem ser divididos com base no tamanho em pequenos ncRNAs, que incluem os microRNAs (miRNAs), e em longos ncRNAs (lncRNAs), que possuem comprimento de mais de 200 nucleotídeos. Os ncRNAs regulatórios atuam na regulação da expressão de genes codificadores de proteínas através da interação com fatores de transcrição ou diretamente com a DNA polimerase, ou ainda podem atuar a nível pós transcricional e afetar a estabilidade do mRNA. Os miRNAs, por exemplo, fazem a inibição da expressão gênica por repressão da

tradução ou por degradação do mRNA (Ebert et al 2012; Zieleziński et al 2012; ZhiCheng e Yan, 2013).

Apesar da importância do estudo das regiões gênicas, existem várias dificuldades para a anotação de genes em espécies não modelos. Em plantas, a fração de genes no genoma é pequena, por exemplo, é menor que 30% na espécie modelo *Arabidopsis thaliana*, o que representa uma pequena parcela entre as sequências não codificantes (Zieleziński et al 2012). Os genes são constituídos por sequências codificantes (os éxons) intercaladas a sequências não codificantes (os íntrons). Não existe uma delimitação clara entre as sequências intragênicas, que separam os éxons, e as sequências intergênicas, que separam genes adjacentes (Stein 2001). Além disso, em muitas espécies de plantas grande percentual da sequência de um gene é constituído por íntrons, os quais podem conter poucos pares de bases (Mathé et al 2002). Dessa forma, o número e o comprimento dos íntrons torna difícil a detecção exata das sequências de éxons. Outro fator que contribui para a dificuldade de anotação de sequências gênicas é a alta frequência de splicing alternativo (Stein 2001). Mas, apesar da complexidade da organização do genoma, vários softwares já foram desenvolvidos para conduzir, de forma eficiente, a anotação estrutural e funcional de genomas eucarióticos (Aubourg et al 2001; Ekblom e Wolf, 2014).

A anotação dos genes que constituem o genoma de uma espécie pode ser feita pelo método de previsão *ab initio* e/ou pelo método de similaridade. Os preditores *ab initio* usam características intrínsecas do genoma para discriminar regiões codantes e não codantes, como códons de início e parada, promotores, sinais de splicing, limites de junção de éxons e regiões ricas em GC (Zieleziński et al 2012). Os métodos baseados em similaridade anotam genes por meio da semelhança significativa com sequências gênicas já conhecidas e armazenadas em bancos de dados. A anotação baseada em similaridade considera que dois genes são semelhantes porque são homólogos, ou seja, possui a mesma origem ancestral comum. Sequências homólogas codificam proteínas que compartilham a mesma função bioquímica, assim é possível também fazer a anotação funcional usando a abordagem de similaridades (Stein, 2001).

Anotação de sequências repetitivas

No processo de anotação é importante também estudar as sequências repetitivas, que podem representar até 90% do tamanho do genoma (Mehrotra et al 2014). As sequências repetitivas podem ser encontradas em uma organização em tandem ou dispersas por todo o genoma (Treangen et al 2012; Mehrotra et al 2014). As repetições em tandem são classificadas de acordo com o tamanho dos motivos de repetição em microssatélites (motivos de 1 a 6 nucleotídeos), minissatélites (motivos de 7 a 100 nucleotídeos) e satélites (motivos com mais de 100 nucleotídeos) (Foulongne-Oriol et al 2013). As sequências dispersas pelo genoma incluem os elementos transponíveis (TEs), fragmentos de DNA capazes de mudar de localização no genoma (Feschotte et al 2002). As três classes principais de TEs, os retrotransposons, os DNAs transposons e os *Rolling-circles* (RC) são caracterizadas por diferentes mecanismos de transposição. Microssatélites e TEs são abundantes nos genomas de plantas.

Entre as sequências repetitivas podem ser anotados também pseudogenes e duplicações segmentares. Os pseudogenes são cópias de genes codificadores de proteínas que tiveram a sequência alterada por mutação e com isso perderam função. Existem várias categorias de pseudogenes, incluindo pseudogenes duplicados, pseudogenes processados e pseudogenes unitários. Pseudogenes duplicados resultam de duplicações de genes na qual uma cópia mantém a função e a outra degrada-se em um pseudogene. Pseudogenes processados surgem quando o RNAm de um gene é transcrito de forma reserva e inserido no genoma. Pseudogene unitário surge quando um gene se torna não funcional (Alexander et al 2010). As duplicações segmentares são regiões repetidas do genoma que variam em comprimento entre mil e vários milhões de nucleotídeos e que mantem pelo menos 90% de identidade (Alexander et al 2010). Informações importantes sobre a história evolutiva de uma espécie podem ser reveladas pelo estudo de duplicações segmentares (Stein, 2001).

A anotação das sequências repetitivas dos genomas pode ser realizada com base em similaridade de sequências ou com base em modelos e estatísticas. A descoberta de repetições com base em similaridade envolve a comparação da sequência de interesse com repetições conhecidas e a identificação de sequências que caem acima de um limite mínimo de conservação (Alexander et al 2010). Por exemplo, a ferramenta de anotação RepeatMasker utiliza análise de similaridade para identificar sequência em um genoma alvo que são homólogas às repetições conhecidas (Tarailo-Graovac e Chen, 2009). Outra estratégia de anotação é criar um modelo com base em repetições conhecidas e usar a estatística para examinar os genomas e encontrar sequências que se ajustem ao modelo (Alexander et al 2010).

A anotação de sequências repetitivas gera informações sobre a estrutura e a organização dos genomas e ainda viabiliza o desenvolvimento de marcadores moleculares para as espécies. Os marcadores moleculares são segmentos de DNA representativos de diferenças genômicas (polimorfismos) entre indivíduos e populações (Agarwal et al 2008). Por exemplo, regiões microssatélites e elementos transponíveis descobertos em estudos genômicos podem ser usados para o desenvolvimento de marcadores. Regiões microssatélites possuem elevada taxa de mutação o que resulta em diferenças no número de unidades ou motivos de repetição em cada locus microssatélite em indivíduos e populações (Liu e Cordes, 2004). Elementos transponíveis também podem gerar polimorfismos genômicos porque o mecanismo de transposição dos retrotransposons para novos locais no genoma resulta em alterações de centenas a milhares de bases no local de inserção (Kumar et al 2011). O desenvolvimento de marcadores moleculares pode ser feito através do desenho de *primers* para a região alvo, da amplificação por PCR (Reação em Cadeia da Polimerase) e do sequenciamento do locus microssatélite ou de elementos transponível para investigar a existência de polimorfismos.

Anotação de polimorfismos de nucleotídeo único (SNPs)

Os Polimorfismos de nucleotídeo único (SNPs) são variações pontuais em que apenas um nucleotídeo é alterado, isto é, ocorre a variação de apenas uma base nitrogenada em uma dada sequência de DNA. As variações genéticas dos SNPs surgem por mutações

nucleotídicas que geram substituições ou indels. As substituições consistem na troca de um nucleotídeo por outro e alteram as sequências de purinas e/ou pirimidinas no DNA, dando origem a diferentes alelos contendo bases alternativas (Brumfield et al 2003). Os SNPs ocorrem em regiões codantes e não codantes do genoma. Os SNPs presentes nas regiões gênicas e reguladoras podem alterar a estrutura e a função das proteínas codificadas e modificar a expressão gênica (Caetano, 2009). Os SNPs ocorrem abundantemente nos genomas e portanto, têm sido caracterizados como marcadores moleculares de preferência em vários estudos, pois permitem uma ampla detecção da variação genômica (Syvanen, 2001; Kerstens et al 2009; Ganal et al 2009).

Várias abordagens estão disponíveis para a descoberta de SNPs. O uso do sequenciamento de nova geração (NGS) permite a identificação de SNPs em larga escala (Ganal et al 2009; Clevenger et al 2015). As grandes quantidades de dados geradas pelo sequenciamento de nova geração e o uso das ferramentas de bioinformática permitem descobrir e analisar milhares de SNPs em centenas de indivíduos, mesmo em organismos não modelos com pouca ou nenhuma informação genética disponível (Kerstens et al 2009; Davey et al 2011; Jiang et al 2016).

Uma abordagem para a descoberta de SNPs usando NGS é o sequenciamento e a montagem do genoma da espécie seguido do re-sequenciamento de vários indivíduos visando comparação de sequências e identificação de variantes. Essa abordagem pode ser usada para o genoma inteiro ou para subconjuntos do genoma (Clevenger et al 2015). Outra abordagem para a identificação de SNPs envolve a redução de complexidade do genoma usando enzimas de restrição. Múltiplas amostras de DNA genômico são digeridas com enzimas de restrição, os fragmentos são selecionados por amplificação e sequenciados usando NGS. Entre os métodos de redução de complexidade do genoma destacam-se bibliotecas de representação reduzida (RRL), RAD–seq e Genotipagem por sequenciamento (GBS) (Davey et al 2011). As sequências obtidas pela redução de complexidade do genoma podem ser comparadas para a identificação de SNPs. Em cada estudo, a escolha da estratégia para descoberta de SNPs depende da disponibilidade de recursos e da questão biológica a ser respondida (Clevenger et al 2015).

OBJETIVO GERAL

- Montar e anotar sequências dos genomas organelares e nuclear de *Dipteryx alata* usando dados de sequenciamento Illumina.

OBJETIVOS ESPECÍFICOS

- Estimar o tamanho do genoma e o número de cromossomos de *Dipteryx alata*;
- Montar sequências do genoma nuclear e organelar usando dados de sequenciamento de nova geração;
- Realizar a anotação estrutural e funcional de sequências gênicas;
- Identificar, caracterizar e quantificar microssatélites e elementos transponíveis;
- Desenhar primers para o futuro desenvolvimento de marcadores microssatélites e marcadores baseados em elementos transponíveis;
- Descobrir e caracterizar polimorfismos de nucleotídeo único em sequências genômicas nucleares.

REFERÊNCIAS

Agarwal M; Shrivastava N; Padh H. (2008) Advances in molecular marker techniques and their applications in plant sciences. Plant Cell Rep, 27:617–631.

Alexander RP; Fang G; Rozowsky J; Snyder M; Gerstein MB. (2010) Annotating non-coding regions of the genome. Nature Reviews Genetics, 11:559-571.

Amar D; Frades I; Danek A; Goldberg T; Sharma SK; Hedley PE; Proux-Wera E; Andreasson E; Shamir R; Tzfadia O; Alexandersson E. (2014) Evaluation and integration of functional annotation pipelines for newly sequenced organisms: the potato genome as a test case. BMC Plant Biology, 14:329.

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Aubourg S; Rouzé P. (2001) Genome annotation. Plant Physiol. Biochem. 39:181−193.

Bolger AM; Lohse M; Usadel B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 1.

Brumfield RT; Beerli P; Nickerson DA; Edwards SV. (2003) The utility of single nucleotide polymorphisms in inferences of population history. Trends in Ecology and Evolution, 18(5):249-256.

Caetano AR. (2009) Marcadores SNP: conceitos básicos, aplicações no manejo e no melhoramento animal e perspectivas para o futuro. R. Bras. Zootec., 38:64-71.

Caicedo AL; Purugganan MD. (2005) Comparative Plant Genomics. Frontiers and Prospects. Plant Physiology, 138:545–547.

Clevenger J; Chavarro C; Pearl SA; Akins PO; Jackson SA (2015) Single Nucleotide Polymorphism Identification in Polyploids: A Review, Example, and Recommendations. Molecular Plant, 8:831–846.

Collevatti RG; Telles MPC; Nabout JC; Chaves LJ; Soares TN. (2013) Demographic history and the low genetic diversity in *Dipteryx alata* (Fabaceae) from Brazilian Neotropical savannas. Heredity, 111:97–105

Davey JW; Hohenlohe PA; Etter PD; Boone JQ; Catchen JM; Blaxter ML. (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Genetics, 12.

Ebert MS; Sharp PA. (2012) Roles for microRNAs in conferring robustness to biological processes. 149(3):515-524.

Ekblom R; Galindo J. (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity, 107:1–15.

Ekblom R; Wolf JBW (2014) A field guide to whole-genome sequencing, assembly and annotation. Evolutionary Applications, 7:1026–1042.

Fedurco M; Romieu A; Williams S; Lawrence I; Turcatti G. (2006) BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. Nucleic Acids Res. 34.

Feschotte C; Jiang N; Wessler SR. (2002) Plant transposable elements: where genetics meets genomics. Nature Reviews Genetics, 3.

Foulongne-Oriol M; Murat M; Castanera R; Ramírez L; Sonnenberg ASM. (2013) Genome-wide survey of repetitive DNA elements in the button mushroom *Agaricus bisporus* Fungal Genetics and Biology, .55:6–21.

Fu N; Wang Q; Shen HL. (2013) De Novo Assembly, Gene Annotation and Marker Development Using Illumina Paired-End Transcriptome Sequences in Celery (*Apium graveolens* L.). Plos One, 8.

Ganal Mw; Altmann T; Roder MS. (2009) SNP identification in crop plants, Current Opinion in Plant Biology 12:211–217.

Ištvánek J; Jaroš M; Křenek A; Řepková J. (2014) Genome Assembly and Annotation for Red Clover (*Trifolium Pratense*; Fabaceae). American Journal of Botany, 101(2):327–337.

Jiang Z; Wang H; Michal JJ; Zhou X; Liu B; Woods LS; Fuchs RA. (2016) Genome Wide Sampling Sequencing for SNP Genotyping: Methods, Challenges and Future Development. International Journal of Biological Sciences, 12(1):100-108.

Ju, J et al (2006) Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. Proc. Natl. Acad. Sci. 103:19635–19640.

Kapitonov VK; Jurka J (2007) Helitrons on a roll: eukaryotic rolling-circle transposons. Trends in Genetics, 23(10).

Kerstens HHD; Crooijmans RPMA; Veenendaal A; Dibbits BW; Chin-A-Woeng TFC; Dunnen JTD; Groenen MAM. (2009) Large scale single nucleotide polymorphism discovery in unsequenced genomes using second generation high throughput sequencing technology: applied to turkey. BMC Genomics, 10.

Kumar A; Hirochika H. (2001) Applications of retrotransposons as genetic tools in plant biology. TRENDS in Plant Science, 6(3).

Liu ZL; Cordes JF (2004) DNA marker technologies and their applications in aquaculture genetics. Aquaculture, 238:1–37.

Lohse M; Nagel A; Herter T; May P; Schroda M; Zrenner R; Tohge T; Fernie AR; Stitt M; Usadel B. (2014) Mercator: a fast and simple web server for genome scale functional annotation of plant sequence data. Plant, Cell and Environment, 37:1250–1258.

Lopes ZJ; Bureau TE. (2014) Diversity and evolution of transposable elements in Arabidopsis. Chromosome Res, 22:203-2016.

Luo et al (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience, 1:18.

Mardis ER. (2008) Next-Generation DNA Sequencing Methods. Annu. Rev. Genomics Hum. Genet., 9:387–402.

Martin JA; Wang Z. (2011) Next-generation transcriptome assembly. Nature Reviews Genetics, 12.

Mathé C; Sagot MF; Schiex T; Rouzé P. (2002) Current methods of gene prediction, their strengths and weaknesses. Nucl. Acids Res. 30(19):4103-4117.

Mehrotra S; Goyal V. (2014) Repetitive Sequences in Plant Nuclear DNA: Types, Distribution, Evolution and Function. Genomics Proteomics Bioinformatics, 12:164–171.

Mendonça RC; Felfili JM; Walter BMT; Silva Junior MC; Rezende AV; Filgueiras TS; Nogueira PE; Fagg CW (2008) Flora vascular do Bioma Cerrado: checklist com 12.356 especies. In: Sano SM, Almeida SP, Ribeiro JF (eds) Cerrado: ecologia e flora. Embrapa Cerrados, Brasília, pp 423–1279

Metzker M (2010) Sequencing technologies — the next generation. Nature Reviews Genetics, 11.

Miller JR; Koren S; Sutton G. (2010) Assembly algorithms for next-generation sequencing data. Genomics, 95:315–327.

Mittler R; Shulaev V. (2013) Functional genomics, challenges and perspectives for the future. Physiologia Plantarum, 148:317–321.

Nagarajan N; Pop M. (2013) Sequence assembly demystified. Nature Reviews Genetics, 14:157-167.

Ratter, JA et al (2000) Estudo preliminar da distribuição das espécies lenhosas da fitofisionomia Cerrado sentido restrito nos Estados compreendidos pero Bioma Cerrado. Boletim do Herbário Esechias Paulo Heringer, Brasília, 5: 5-43.

Safonova Y; Bankevich A; Pevzner PA. (2015) dipSPAdes: Assembler for Highly Polymorphic Diploid Genomes. Journal Of Computational Biology, 22(6):528-545.

Sano SM; Ribeiro JF; Brito MA. (2004) Baru: biologia e uso. Planaltina, DF: Embrapa Cerrados, 52p.

Schatz MC; Delcher AL; Salzberg SL (2010) Assembly of large genomes using second-generation sequencing. Genome Research, 20:1165–1173.

Schlebusch S; Illing N (2012) Next generation shotgun sequencing and the challenges of de novo genome assembly. South African Journal of Science, 108.

Simpson JT; Wong K; Jackman SD; Schein JE; Jones SJM.; Birol I. (2009) ABySS: A parallel assembler for short read sequence data. Genome research, 19:1117-1123.

Soares TN; Chaves LJ; Telles MPC; Diniz-Filho JAF; Resende LV. (2008a) Distribuição espacial da variabilidade genética intrapopulacional de *Dipteryx alata*. Pesq. agropec. bras., 43(9):1151-1158.

Soares TN; Chaves LJ; Telles MPC; Diniz-Filho JAF; Resende LV. (2008b) Landscape conservation genetics of *Dipteryx alata* (''baru'' tree: Fabaceae) from Cerrado region of central Brazil. Genetica, 132:9–19.

Soares TN; Melo DB; Resende LV; Vianello RP; Chaves LJ; Collevatti RG; Telles MPC. (2012) Development of microsatellite markers for the neotropical tree species *Dipteryx alata* (Fabaceae). American Journal of Botany.

Soares TN; Diniz-Filho JAF; Nabout JC; Telles MPC; Terribile RC; Chaves LJ. (2015) Patterns of genetic variability in central and peripheral populations of *Dipteryx alata* (Fabaceae) in the Brazilian Cerrado. Plant Syst Evol, 301:1315–1324.

Stein, L.; (2001) Genome Annotation: From Sequence to Biology. Nature Reviews Genetics, 2..

Syvanen AC. (2001) Accessing genetic variation: genotyping Single Nucleotide Polymorphisms. Nature Genetics, 2:930-942.

Tarailo-Graovac M; Chen N. (2009) Using RepeatMasker to Identify UNIT 4.10. Repetitive Elements in Genomic Sequences. Current Protocols in Bioinformatics.

Treangen TJ; Salzberg SL. (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature Reviews Genetics, 13.

Strickler SR; Bombarely A; Mueller LA. (2012) Designing a transcriptome next-generation sequencing project for a nonmodel plant species. American Journal of Botany, 99(2):257–266.

ZhiCheng D; Yan C. (2013) Transcriptomics: Advances and approaches. SCIENCE CHINA Life Sciences, 56(10):960-967.

Zhou MB; Zhong H; Tang DQ. (2011) Isolation and characterization of seventy-nine full-length mariner-like transposase genes in the Bambusoideae subfamily. Journal of Plant Research, 124:607-617.

Zieleziński A; Potarzycki P; Książkiewicz M; Karłowski WM. (2012) Annotating a non-model plant genome – a study on the narrow-leafed lupin. Journal of Biotechnology, Computational Biology and Bionanotechnology. 93(3):318-332.

Zimin AV; Marçais G; Puiu D; Roberts M; Salzberg SL; Yorke JA. (2013) The MaSuRCA genome assembler. Bioinformatics, 29(21):2669-77.

CAPÍTULO 1

TENDÊNCIAS EM TAMANHO DO GENOMA E EVOLUÇÃO DO NÚMERO DE CROMOSSOMOS NA SUBFAMÍLIA PAPILIONOIDEAE (LEGUMINOSAE) COM ÊNFASE NO CLADO BASAL ADA

RESUMO

O tamanho do genoma (valor C) e o número de cromossomos (2n ou n) apresentam uma enorme diversidade de valores em plantas. Apesar das melhorias nos métodos de estudo, as estimativas 1C e 2n estão disponíveis para um número limitado de espécies. Este estudo teve como objetivo investigar a evolução desses parâmetros em um contexto filogenético para Leguminosae, com foco no clado Papilionoideae ADA (Angylocalyceae + Dipterygeae + Amburaneae) e contribuir com novas estimativas para *Dipteryx alata* e *Amburana cearensis*. Foi realizada uma análise filogenética para 349 membros de Leguminosae, com base nos genes plastidial matK e rbcL combinados e no gene nuclear 5,8S rDNA, seguido por uma reconstrução de caracteres para 2n e 2C de todas estas espécies. O presente trabalho estimou o tamanho do genoma de *D. alata* como 1C = 0.825 pg, o que corresponde a um genoma haplóide de 807.2 MB com 2n = 16 e de *A. cearensis* 1C = 0.483pg (1C = 472.3 MB). Esses dados estão em concordância com o observado em outras espécies do clado ADA, que apresentam números cromossômicos baixos e genomas pequenos. O tamanho do genoma e o número de cromossomos reconstruídos na filogenia sugerem que esta observação é uma condição plesiomórfica em Papilionoideae, embora várias mudanças específicas no clado sejam observadas, como um aumento notável do tamanho do genoma em Fabeae. Os valores de 2n e 1C estimados para *Dipteryx alata* e *Amburana cearensis* caracterizam essas espécies como candidatas da família ao sequenciamento do genoma devido a seus pequenos tamanhos genômicos e a suas posições basais no clado Papilionoideae.

Palavras-chave: *Amburana cearensis*, C-valor, citogenética, *Dipteryx alata*, genoma modelo, Leguminosae

INTRODUÇÃO

O tamanho do genoma (valor 1C) e o número de cromossomas (2n ou n) apresentam uma enorme diversidade em plantas. O tamanho do genoma, expresso como valor C em picogramas (pg) para células haplóides não está estritamente correlacionado com a complexidade dos organismos (Huang et al 2015; Hidalgo et al 2015). Entre as angiospermas, a variação no tamanho do genoma é de cerca de 2400 vezes, de 0,06 (*Genlisea margaretae*) a 152,23 pg (*Paris japonica*) (Greilhuber et al 2006; Pellicer et al 2010). O tamanho do genoma é geralmente estável dentro de uma espécie, mas pode diferir fortemente entre os taxa filogenéticos estreitamente relacionados (Nag et al 2011; Pandit et al 2014; Hidalgo et al 2015). A variação do número de cromossomos nas plantas também é grande, o menor número 2n=4 é encontrado em espécies como *Haplopappus gracilis* e *Brachyscome dichromosomatica* e o maior número é encontrado na espécie de samambaia *Ophioglossum reticulatum* que possui 2n = 1400 (Castiglione e Cremonini 2012; Dyer et al 2013).

O tamanho do genoma e o número de cromossomos são informações que podem ser utilizadas no campo da ecologia, fisiologia, melhoramento de plantas, conservação e outras áreas (Bennett et al 2000; Pandit et al 2014). Na prática, o conhecimento do tamanho do genoma é importante para estimar o tempo e o custo de projetos de sequenciamento do genoma completo (Garcia et al 2014, Wang et al 2015). Na perspectiva evolutiva, a medição do conteúdo de DNA nuclear e o número de cromossomos podem ser usados para elucidar padrões evolutivos e mecanismos de adaptação em plantas (Wang et al 2015; Huang et al 2015).

As estimativas do tamanho do genoma estão disponíveis para apenas 7500 angiospermas, isto é, 2% das espécies conhecidas (Wang et al 2015). A família Leguminosae é a terceira maior família de plantas com flores, com 24.505 espécies de 946 gêneros (The Plant List 2013). Leguminosae é classificada, com base em dados filogenéticos, em seis subfamílias: Caesalpinioideae DC, Cercidoideae Grupo de Trabalho de Filogenia de Leguminosas, Detarioideae e Burmeister, Dialioideae Grupo de Trabalho de Filogenia de Leguminosas, Duparquetioideae Grupo de Trabalho de Filogenia de Leguminosas (Stat Nov.), Papilionoideae DC (Azani et al, 2017). As leguminosas Papilionoideae são constituídas por cerca de 13.800 espécies e são um legado devido sua ampla gama ecológica e biogeográfica associada a diversificações em variados ecossistemas (Lavin et al 2005; Lewis et al 2005).

Papilionoideae também são bem conhecidas por sua importância econômica para a agricultura e para a indústria de alimentos. Devido ao seu interesse econômico, várias Papilionoideae são alvo de projetos de sequenciamento genômico, por exemplo *Glycine max*, *Phaseolus vulgaris*, *Cicer arietinum* e *Medicago truncatula* (Schmutz et al 2010; Schmutz et al 2014; Tang et al 2014). As Papilionoideae potencialmente de ramificação mais antiga pertencem ao clado ADA, que inclui as tribos monofiléticas Angylocalyceae, Dipterygeae e Amburanae (Gunn and Service 1983; The Plant List 2013; Cardoso et al 2013; Hinchliff et al 2015). Angylocalyceae possui 21 espécies pertencentes a 5 gêneros, Dipterygeae possui 25 espécies pertencentes a 4 gêneros e Amburaneae possui 29 espécies divididas em 7 gêneros. Angylocalyceae possui gêneros na África, em Madagascar, na Austrália e em florestas tropicais do norte da América do Sul. Os gêneros das tribos Dipterygeae e Amburaneae são

encontrados na região neotropical (Cardoso et al 2013). Entre esses gêneros estão *Dipteryx*, (que possui 12 espécies) e *Amburana* (que possui 2 espécies) a quais pertencem as espécies de árvores de Leguminosae *Dipteryx alata* Vogel e *Amburana cearensis* (Allemão) A. C. Sm, o foco deste estudo.

A diversidade de números de cromossomos e tamanho do genoma também é grande na subfamília Papilionoideae. O número de cromossomos varia de 2n=8 a 2n=180, enquanto que o tamanho do genoma apresenta uma variação de 80 vezes de 2C=0,69 pg a 2C=54,8 pg (Garcia et al 2014; Rice et al 2015). O presente estudo teve como objetivo estimar o tamanho do genoma de *Dipteryx alata* e *Amburana cearensis*, espécies do clado ADA, e o número de cromossomos de *Dipteryx alata,* e reconstruir a evolução desses caracteres em um contexto filogenético.

MATERIAL E METÓDOS

Material vegetal

Para citometria de fluxo foram coletadas folhas de três indivíduos de *Dipteryx alata* e *Amburana cearensis*. As folhas de *D. alata* foram coletadas no banco de germoplasma mantido pela escola de Agronomia da Universidade Federal de Goiás. Folhas de sementes germinadas de *Glycine max* cv. 'Polanka' (Valor 2C=2,5pg), *Pisum sativum* cv. 'Ctirad' (9,76pg) e *Solanum lycopersicum* cv. 'Stupicke' (1,96pg), foram utilizadas como padrões internos para Citometria de Fluxo. As sementes germinadas foram obtidas no Instituto de Botânica Experimental em Olomouc, República Tcheca (Doležel e Greilhuber, 2010). As raízes de *D. alata* para as análises cromossômicas foram coletadas de sementes germinadas.

Análise de Citometria de fluxo

As folhas foram mantidas em papel úmido a baixa temperatura (4-6 ° C) durante um período de dois dias antes do procedimento. A estimativa do tamanho do genoma foi realizada de acordo com o protocolo de Dolezel e Gohde (1995). Pequenos fragmentos de tecido foliar (40 mg) da amostra e padrão foram co-picados com lâmina de barbear em uma placa de Petri contendo 1,5 ml de tampão WPB (Tris-HCl 0,02 M, $MgCl_2.6H_2O$ 4mM, Triton X 100 1%, EDTA Na_2. $2H_2O$ 2 mM, NaCl 86 mM, metabissulfito de sodio 10 mM, 2 PVP-10 1%). As amostras foram fragmentadas rapidamente para minimizar a liberação de compostos citosólicos, passadas através de um filtro com poros de 30 μm para remover os detritos celulares e coradas com 50 μl de iodeto de propídio (1 mg/mL). Esta preparação foi repetida em triplicata para cada planta e foram realizadas três medições em dias consecutivos.

As amostras foram analisadas em um Partec CyFlow Space (Partec, Münster, Alemanha) equipado com um laser operando a 532 nm. Para cada amostra, os picos de fluorescência foram lidos pelo menos 2.000 núcleos 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. Determinou-se o conteúdo de DNA nuclear (valor 2C) de

acordo com a seguinte fórmula para a fluorescência relativa: ([Média do pico da amostra/ Média do pico do padrão] x valor 2C do padrão em pg). O valor C foi multiplicado por 978 para obter o tamanho do genoma em megabases (MB), visto que 1 pg é igual a 978 pb.

Determinação do número de cromossomos de *D. alata*

Pontas radiculares de *D. alata* foram pré-tratadas com 0,002M de 8-hidroxiquinolina à temperatura ambiente durante 5h, fixadas durante a noite em etanol/ácido acético glacial 3:1 (v/v) e armazenadas a -6°C. As pontas de raízes foram digeridas usando 18U/ml de celulase e 18U/ml de pectinase durante 120 min a 37°C. O meristema foi dissecado em ácido acético (45%) e esmagado sob uma lamínula. A coloração foi realizada com 4',6-diamidino-2-fenilindole (DAPI - 2ug/ml)/glicerol a 1:1. Captura foi feita em um Zeiss Axio Scope A1 equipado com uma câmera CCD Cube 1m Cool e o software Cometa. A informação referente ao 2n foi determinada por contagem de cromossomas de células de pelo menos quatro indivíduos.

Frequências e correlações entre 2C e 2n

A base de dados de valor 2C Kew foi consultada para o tamanho do genoma e o número de cromossomos das espécies de Leguminosae, enquanto que a Base de Dados de Contagens de Cromossomas (CCDB) foi consultada para obter o número de cromossomos de 6.071 espécies (Garcia et al 2014; Rice et al 2015) (Material Suplementar online, tabela S1). Histogramas de frequências e estimativas de densidade gaussiana de kernel foram obtidas pelo programa R com as funções: geom_histogram, geom_point, geom_density, enquanto que a correlação de Kendall (não paramétrica) foi utilizada para avaliar a correlação entre 2C e 2n com a função cor.test. O interativo "figuras on-line", citado ao longo do texto é disponível em http://cyto.shinyapps.io/fabaceae e foi desenvolvido com shiny-R (Wickham 2009; Wang et al 2013; Chang et al 2016; R Core Team 2016).

Análise filogenética

As relações filogenéticas foram baseadas em genes plastidiais matK e rbcL combinados e na família de genes nucleares 5,8S rDNA para 349 espécies de Leguminosae (Material suplementar online, Tabela S2). Todas as sequências foram obtidas a partir do Genbank (Material suplementar online, Tabela S2) e foram selecionadas todas as espécies de leguminosas que tiveram dados de número cromossômico e/ou conteúdo de DNA conhecidos. Todos os clados identificados foram nomeados de acordo com Borges et al (2013). Os membros do clado Caesalpinoideae foram utilizados como *outgroups* (Borges et al 2013).

A Inferência Bayesiana (BI) foi realizada com sequências combinadas matK + rbcL + ITS usando o Mr. Bayes v3.1.2 (Ronquist e Huelsenbeck, 2003) sob um modelo particionado como implementado no CIPRES Science Gateway V.3.1 (www.phylo.org). O modelo de evolução de sequências mais apropriado para cada região foi determinado utilizando o critério de informação Akaike (AIC) implementado no programa jModelTest

0.1.1 (Posada, 2008). A análise foi conduzida para duas corridas independentes e para 10.000.000 gerações, amostrando 1.000 árvores. Tracer v 1.0.1 foi utilizado para assegurar que as cadeias de Markov atingiram o estado estacionário e para determinar o número apropriado de corridas (Rambaut e Drummond, 2003). As probabilidades posteriores (PP) foram estimadas construindo uma árvore consenso com as árvores amostradas. As árvores foram visualizadas usando FigTree v1.3.1 (Rambaut, 2009).

Reconstrução do estado ancestral e inferência do número de cromossomos / tamanho do genoma

Para as espécies com mais de um nível de ploidia ou número de cromossomos, apenas o menor foi utilizado para a reconstrução do número diploide ancestral. Para reconstruir o número de cromossomos diploides ancestrais (2n) ao longo da filogenia, utilizamos o modelo de máxima verossimilhança Markov k-state de um parâmetro (Mk1) em Mesquite (v.2.75; Maddison e Maddison, 2011). Os dados de tamanho do genoma foram considerados como caracteres contínuos e analisados sob máxima parcimônia em Mesquite.

RESULTADOS

Tamanho do genoma e número de cromossomos

O tamanho do genoma de *D. alata* foi calculado como 1C=0.825 pg que correspondente a um genoma haplóide de 807.2 MB (Figura 1a), enquanto que *A. cearensis* mostrou 1C=0.483 pg (1C=472.3 MB). *Glycine max* e *Solanum lycopersicum* foram selecionados como padrões de referência por terem tamanhos de genoma próximos sem sobreposição de picos. Em todas as prometafases analisadas, *D. alata* mostrou 2n = 16 cromossomos (Figura 1b). *A. cearenses* possui 2n = 22 cromossomos (Darlington e Wylie, 1956; Fedorov 1969).

Variação do tamanho genômico e do número de cromossomos em Leguminosae

Dividindo a amostra de números cromossômicos conhecidos de Leguminosae em tercis (cada um com 33,3%), foram definidos três grupos: baixo 2n (8 - 16), intermediário 2n (16 - 22) e alto 2n (22 - 180). A maioria dos clados de Leguminosae e Papilionoideae, mesmo o clade ADA, tinha espécies em cada categoria. A distribuição de 2n é parcialmente descontínua (multimodal) uma vez que os eventos de poliploidia podem ser observados como picos num histograma 2n (Figura 2). Os números mais frequentes indicam poliploidia recente em Baphieae, Dalbergioids, Genistoids, Hologalegina, Indigofereae e MPA de Papilionoideae. Nos clados de Papilionoideae, 2n = 16 (baixo-intermediário 2n), 2n = 14 (baixo) e 2n = 22 (intermediário-alto 2n) estão entre os três números mais frequentes para cada clado nessa ordem, enquanto para outros clados Leguminosae os mais comuns são 2n = 28 e 24 (alto 2n).

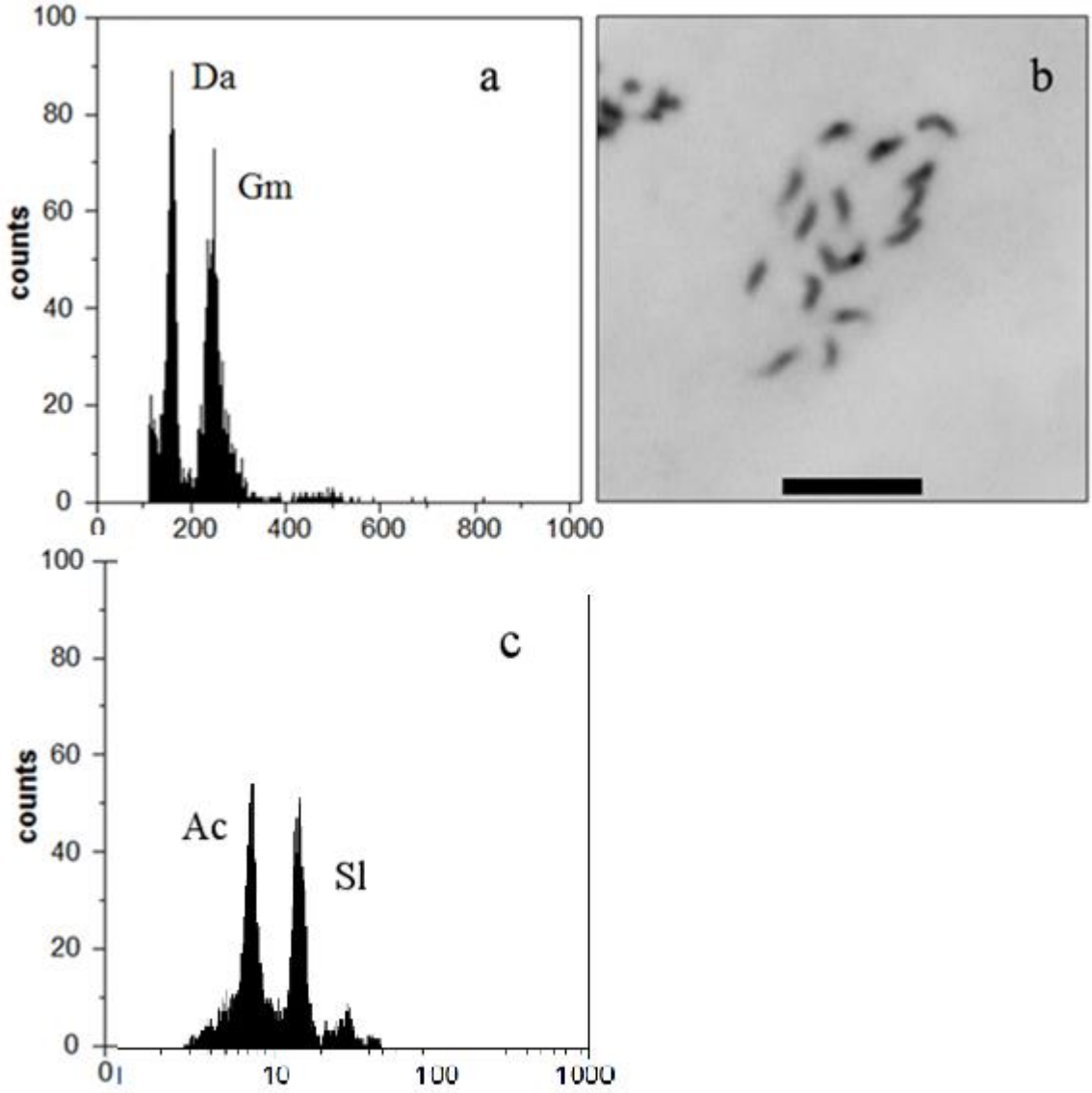

Figura 1 - Histogramas de fluorescência relativa (unidades arbitrárias) de núcleos corados com iodeto de propídio por Citometria de Fluxo. Fig1a Da - *Dipteryx alata*. Gm - *Glycine max*, padrão de referência; Fig1b Prometáfase de *D. alata* 2n = 16 corada com DAPI. Cor invertida. Bar = 5 um; Fig1c Ac - *Amburana cearensis*, Sl - *Solanum lycopersicum*, padrão de referência.

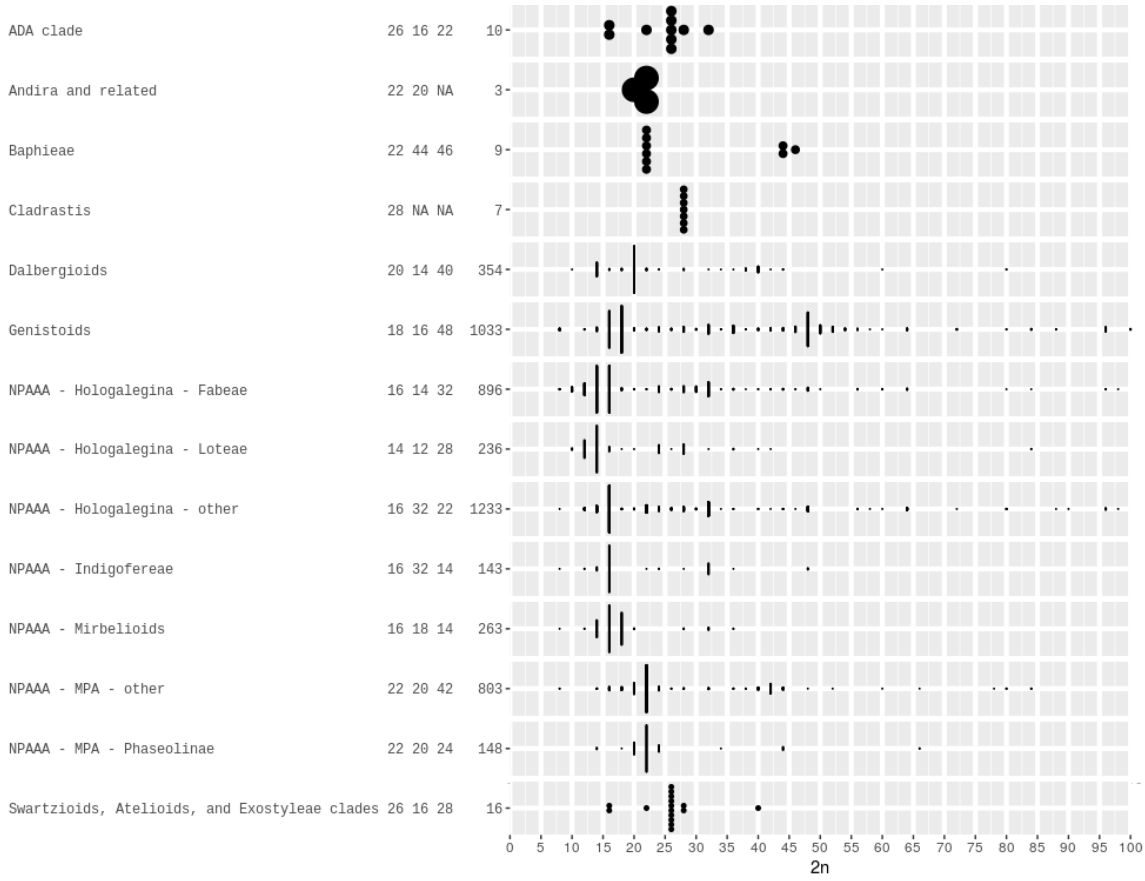

Figura 2 - Dot plot distribuição de 2n para Papilionoideae; Y contém nome, três mais freqüentes 2n e tamanho da amostra. Tamanho do ponto em diferentes escalas dependendo do tamanho da amostra.

A maioria dos clados de Leguminosae e Papilionoideae tinha espécies em todas as três categorias estabelecidas dividindo a amostra de tamanhos de genoma conhecidos em tercis: baixo 2C (0,6 - 1,6 pg), intermediário 2C (1,6 - 3,5 pg) e alto 2C (3,5 - 54,8 pg). Algumas exceções são o clado ADA com baixo 2C, e outros clados pouco estudados. A distribuição de 2C é multimodal como mostrado pela densidade de kernel gaussiano para 2C na Figura 3. O primeiro pico correspondeu à baixa categoria 2C para a maioria dos clados.

Correlação entre os valores 2n e 2C

À primeira vista pode parecer que todas as Leguminosae ou pelo menos as Papilionoideae mostraram uma restrição para altos valores de 2C em espécies com 2n alto. Esse fato foi confirmado para clados ricos como Fabeae, que apresentou correlação negativa entre 2C e 2n. Consequentemente, na Figura 4, o clado Fabeae com baixo 2n apresenta tamanhos de genomas grandes. No entanto, para o clado Hologalegina-Loteae e o clado NPAAA-MPA uma correlação positiva poderia ser estabelecida. Nesses grupos, existe uma restrição para valores baixos de 2C em espécies com 2n alto. De acordo com isso, na filogenia, o clado Erythrininae (NPAAA-MPA) com tamanhos de genoma intermediários (> 2 pg) mostra alto número de cromossomos (Figura 4). Os clados Genistoids e Dalbergioids são

os Papilionoideae mais amostrados fora do clado NPAAA (que inclui Hologalegina) e têm números cromossômicos principalmente nas categorias intermediária e alta 2n. No entanto, é evidente a partir da filogenia e dos histogramas, que Genistoids tem muito mais cariótipos com 2n alto, mas genomas menores. Em contraste, os genomas Dalbergioids tendem a ser maiores do que os genomas de Genistoids, mas com 2n menor, mostrando tendências opostas.

Análise de ancestralidade do valor 2C e número de cromossomos

A reconstrução de máxima verossimilhança 2C em nodos antigos mostrou um pequeno tamanho do genoma para o antepassado Papilionoideae com um máximo de 2 pg (azul na figura 4). Em algum ponto na evolução, Fabeae ganhou um genoma enorme de 12 pg (verde na figura 4), enquanto outros nós ancestrais mostraram até 4 pg, como em Dalbergioids (azul claro). Da reconstrução de n, um processo que descarta qualquer informação de ploidia (recente), o ancestral 2n de Papilionoideae corresponde a um cariótipo de tamanho médio de 2n = 21 a 33 (amarelo). Duas tendências conspícuas são observadas nos ramos, um pequeno número de cromossomos no antepassado do clado Hologalegina e um maior número de cromossomos no antepassado de outros clados NPAAA e os clados basais ADA, Genistoids e Dalbergioids.

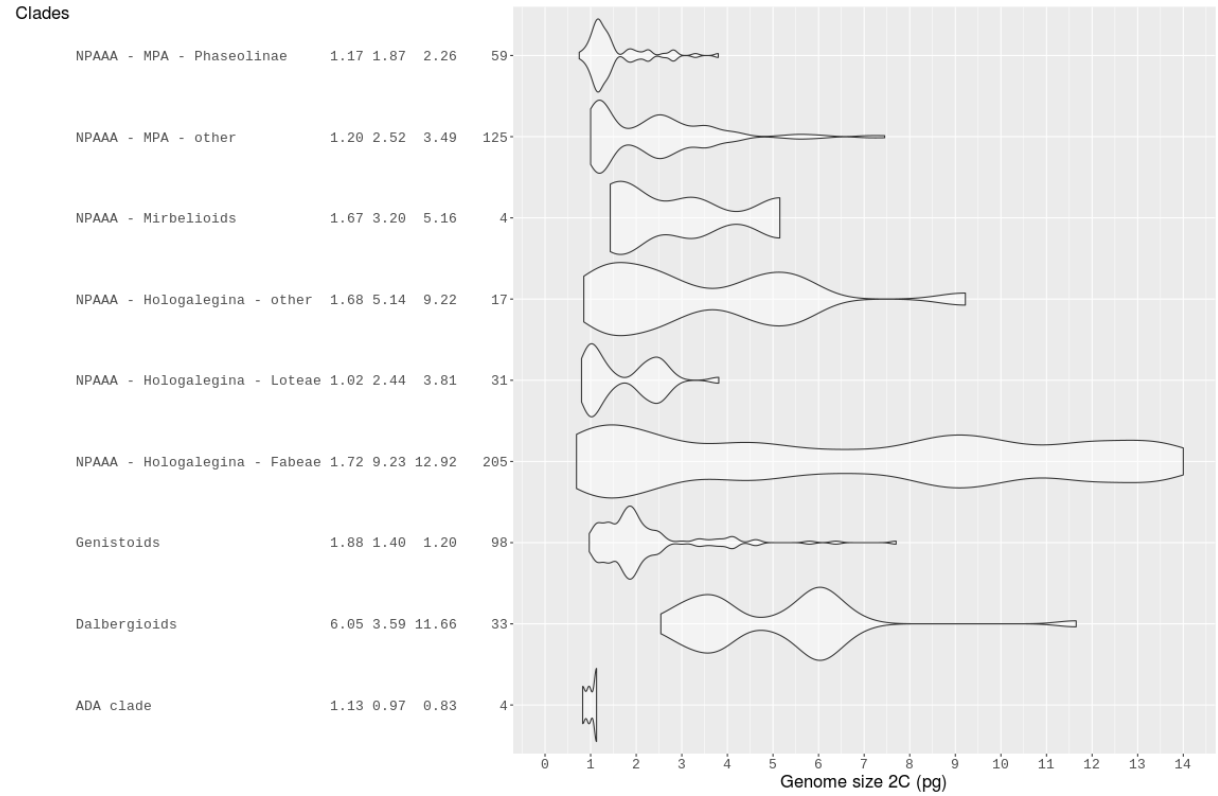

Figura 3. Gráficos de violino mostrando a densidade de 2C em Papilionoideae. O eixo y contém os três picos mais altos em pg e o tamanho da amostra. Largura de banda em função de densidade personalizável nos recursos on-line.

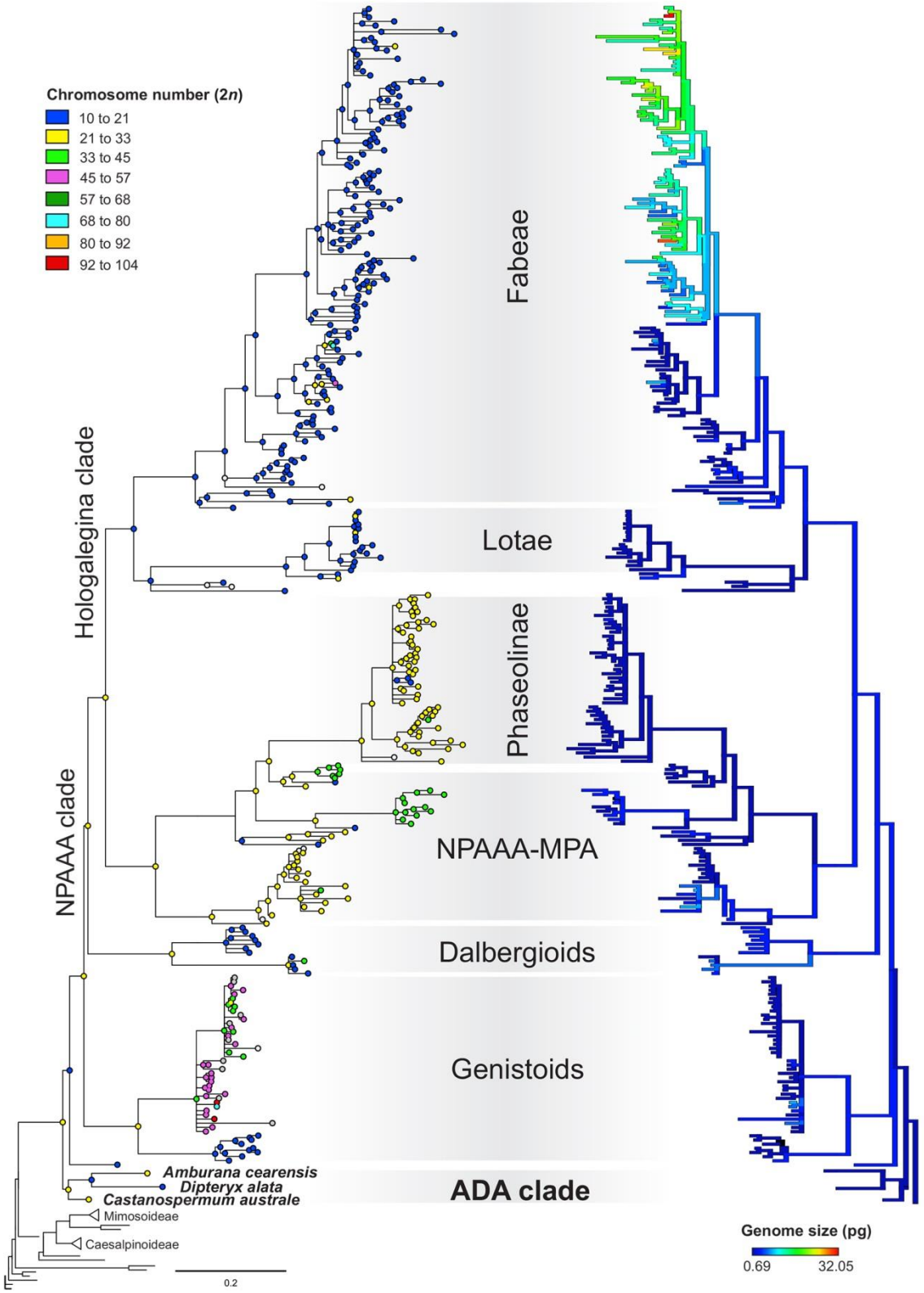

Figura 4 - Evolução do tamanho do genoma e número de cromossomos em espécies deLeguminosae. Valores de 2n e 2C são mostrados por cores nos nós e ramos antigos.

DISCUSSÃO

Evolução de 2C e 2n em Leguminosae

Análises anteriores de milhares de genes homólogos de rosídeos mostraram que o antepassado de leguminosas foi um hexaploide, com provavelmente n = 21 cromossomos, que se rearranjou para produzir n = 6 através de pelo menos 16 fusões e 1 fissão. Posteriormente, o ancestral de todos os Papilionoideae sofreu um evento de duplicação do genoma total (WGD) apresentando n = 12 (Cannon et al 2015). Doyle (2012) sugere um ancestral para Papilionoideae com n = 14. Historicamente, as frequências foram usadas para propor os números ancestrais para os táxons, no entanto, atualmente não é aceito desconsiderar a informação filogenética e genômica disponível. Métodos heurísticos são usados com mais frequência para reconstruir os números ancestrais, mostrando uma gama de valores ancestrais ou uma probabilidade para um ou mais números ancestrais nos nós (Mayrose et al 2010; Peruzzi 2013; Sousa et al 2014, Macas et al 2015). O intervalo ancestral (2n = 21-33) para Papilionoideae obtido aqui corrobora o anteriormente citado, e se encaixa bem com os 2n encontrados nos clados basais ADA, Swarzioids, Atelioids e Cladrastis (modos 2n = 16-28).

Entre Leguminosae, eventos de duplicação ocorreram em antepassados de Cercidae e Detarieae. Dentro de Papilionoideae, antepassados de Genistoides e soja (*Glycine*) um paleopoliploide, sofreram WGD como mostrado pela comparação de segmentos ortólogos. No entanto, não há evidências desses eventos no tamanho atual do genoma em Papilionoideae. A reconstrução dos valores C ancestrais começa excluindo qualquer valor poliploide possível (recente), pois modela apenas variações contínuas ou graduais.

A dinâmica da evolução do tamanho do genoma rompe as mudanças drásticas no cariótipo após um WGD ao longo de longos períodos de tempo. Além de WGD, vários mecanismos são responsáveis pela expansão e contração do tamanho do genoma. Elementos repetitivos, especialmente elementos transponíveis, em particular dos retrotransposons, são os principais fatores que explicam diferenças no tamanho genômico de plantas, uma vez que podem explicar até 85% de diferenças no tamanho do genoma, conforme estudado em espécies da tribo Fabeae (Macas et al 2015). As repetições em tandem e a duplicação segmentar do genoma são comuns em plantas e também estão positivamente correlacionadas com a quantidade de DNA (Bennett et al 2000, Michael et al 2014). Por outro lado, a redução do tamanho do genoma pode ocorrer através de deleções, tais como erros na replicação do DNA e recombinação ilegítima ou não homóloga (Whitney et al 2010).

A presente reconstrução dos valores de C mostra dois clados claramente com um aumento ancestral do tamanho do genoma, o Dalbergioids e NPAAA-Hologalegina-Fabeae, como explicado, independentemente de eventos WGD. Além disso, é claro que vários eventos de poliploidia recente são comuns na família, especialmente em vários clados de Papilionoideae, como evidenciado pela distribuição multimodal de 2n, mas eles têm que ser transformados ou removidos na reconstrução de ancestral 2n como explicado para a reconstrução do tamanho do genoma.

Tendências contrastantes do tamanho do genoma vs. número de cromossomos foram encontradas para clados de Leguminosae. Embora se possa esperar *a priori* uma

correlação positiva entre eles, como observado nos clados NPAAA-MPA e Hologalegina (excluindo Fabeae), várias características ecológicas conduzidas modelam essas características de formas especiais (Bennett 1987; Pandit et al 2014; Suda Et al 2015). Dependendo do hábito e invasividade da espécie, seu genoma se adapta para não se expandir, enquanto o número de cromossomos pode variar por poliploidia conferindo nova matéria-prima para a evolução (Pandit et al 2014). Mecanismos de constrangimento do tamanho do genoma, despertados pela seleção natural, podem permitir estratégias ecológicas que possam explicar a ausência de genomas grandes em espécies de alto 2n. Na ausência de tais mecanismos, pode ocorrer uma expansão ilimitada do genoma (Kelly et al 2015).

O clado ADA e as espécies candidatas para análise do genoma de Papilionoideae

O número de cromossomos e o tamanho do genoma apresentam valores variáveis entre as espécies do clado ADA. *A. cearensis* possui número cromossômico intermédiario (2n = 22) e tamanho do genoma baixo (1C = 0.483 pg). Já *D. alata* possui número cromossômico baixo (2n = 16) e genoma pequeno (1C = 0.825 pg). Os tamanhos do genoma de *D. alata* e *A. cearensis* estão entre os menores da família Leguminosae, principalmente após a espécie do género *Trifolium* (Fabeae), mas ao contrário de outros gêneros com genoma pequeno, são parte de um clado basal de Papilionoideae. Ter um conteúdo de DNA muito pequeno é importante na escolha das espécies candidatas para o sequenciamento do genoma. As plataformas de sequenciamento de nova geração (NGS) permite a análise de sequências de nucleotídeos com alto desempenho e custo reduzido por pares de bases sequenciadas. Apesar dos avanços na tecnologia de sequenciamento, genomas de tamanhos menores são priorizados devido a menores custos para a obtenção das sequências totais estabelecidas. Além disso, pequenos genomas facilitam a montagem das sequências, um trabalho computacional complexo e difícil para genomas grandes (Kelly et al 2012). Nesse sentido, há suporte para colocar *D. alata* e *A. cearensis* como candidatas do Cerrado para estudos de sequenciamento do genoma. O estudo genômico destas espécies permitirá a realização de análise comparativa com outras Papilionoideae que já tem genoma sequenciado, como *Glycine max* e *Phaseolus vulgaris*. Compreender as diferenças moleculares entre os genomas das espécies de Papilionoideae é importante para elucidar os mecanismos que regulam as modificações de valor C e de número de cromossomos.

CONCLUSÃO

Este estudo contribuiu com novas estimativas de tamanho do genoma e número de cromossomos para a família Leguminosae. Os tamanhos dos genomas haplóide de *D. alata* e *A. cearensis*, do clado basal ADA, estão entre os menores em Leguminosae ou Papilionoideae. Devido a essas características, propomos as espécies como candidatas para projetos de sequenciamento do genoma. Isto permitirá uma análise evolutiva aprofundada através de comparações de genoma completo com outras espécies modelo da família como Lotus (Sato et al 2008), soja (Schmutz et al 2010), grão de bico (Varshney et al 2013) e feijão (Schmutz et al 2014). Caracterizar a evolução dos traços citogenéticos de Leguminosae é o

passo inicial antes de elucidar os mecanismos que regulam mudanças no valor C e rearranjos que alterar o número de cromossomos.

MATERIAL SUPLEMENTAR

https://www.dropbox.com/sh/i6u8roglrx5dusw/AACVy6CoglMGYSyDZj8J9NEPa?dl=0

REFERÊNCIAS

Azani N et al (2017) A new subfamily classification of the Leguminosae based on a taxonomically comprehensive phylogeny. The Legume Phylogeny Working Group (LPWG). Taxon, 66(1):44-77.

Bennett MD (1987) Variations in genome form in plants and its ecological implications. New Phytol 106:177–200.

Bennett MD, Bhandol P, Leitch, IJ (2000) Nuclear DNA Amounts in Angiosperms and their Modern Uses - 807 New Estimates. Annals of Botany 86:859-909.

Borges L, Bruneau A, Cardoso D, et al (2013) Towards a new classification system for legumes: Progress report from the 6th International Legume Conference. South African J Bot 89:3–9.

Cannon SB, McKain MR, Harkess A, et al (2015) Multiple polyploidy events in the early radiation of nodulating and nonnodulating legumes. Mol Biol Evol 32:193–210.

Cardoso D, Pennington RTT, de Queiroz LPP, et al (2013) Reconstructing the deep-branching relationships of the papilionoid legumes. South African J Bot 89:58–75.

Castiglione MR; Cremonini R. (2012) A fascinating island: 2n = 4. New Trends in Plant Cytogenetics and Cytoembryology, 146:711-726.

Chang W, Cheng J, Allaire JJ, Xie Y, McPherson J (2016) shiny: Web Application Framework for R.

Cusimano N, Sousa A, Renner SS (2012) Maximum likelihood inference implies a high, not a low, ancestral haploid chromosome number in Araceae, with a critique of the bias introduced by “x”. Ann Bot 109:681–92.

Darlington CD, Wylie AP. (1956) Chromosome atlas of flowering plants. Chromosome atlas of flowering plants. George Allen and Unwin Ltd., London, UK.

Doležel J, Göhde W. (1995) Sex determination in dioecious plants Melandrium album and M. rubrum using high resolution flow cytometry. Cytometry, v.19, p.103–106.

Doležel J, Greilhuber J (2010) Nuclear genome size: Are we getting closer? Cytom Part A 77:635–642.

Doyle JJ (2012) Polyploidy in Legumes. Polyploidy Genome Evol. Springer Berlin Heidelberg, Berlin, Heidelberg, 47–180

Dyer RJ; Pellicer J; Savolainen V; Leitch IJ; Schneider H. (2013) Genome size expansion and the relationship between nuclear DNA content and spore size in the *Asplenium monanthes* fern complex (Aspleniaceae). BMC Plant Biology, 13:219.

Fedorov ANA. (1969) Chromosome numbers of flowering plants. Academy of science of theUSSR Komarov. Botanical Institute Leningrad (Reprint, 1974).

Garcia S, Gálvez F, Gras A, Kovařík A, Garnatje T (2014) Plant rDNA database: update and new features. Database (Oxford) 2014:1–7.

Greilhuber J, Borsch T, MAller K, Worberg A, Porembski S, Barthlott W (2006) Smallest angiosperm genomes found in Lentibulariaceae with chromosomes of bacterial size. Plant Biology 8:770-777.

Gunn CR, Service USAR (1983) A nomenclator of legume (Fabaceae) genera. U.S. Dept. of Agriculture, Agricultural Research Service

Hidalgo O, Garcia S, Garnatje T, Mumbru M, Patterson A, Vigo J, Valles J (2015) Genome size in aquatic and wetland plants: fitting with the large genome constraint hypothesis with a few relevant exceptions. Plant Syst Evol 31.

Hinchliff CE, Smith SA, Allman JF, et al (2015) Synthesis of phylogeny and taxonomy into a comprehensive tree of life. Proc Natl Acad Sci 112:12764–12769.

Hiremath SC, Nagasampige MH (2004) Genome size variation and evolution in some species of Dalbergia Linn.f. (Fabaceae). Caryologia 57:367-372.

Huang H, Tong Y, Zhang QJ, Gao LZ (2015) Genome Size Variation among and within Camellia Species by Using Flow Cytometric Analysis. Plos Onde 8.

Kelly LJ, Leitch AR, Fay MF, et al (2012) Why size really matters when sequencing plant genomes. Plant Ecol Divers 5:415–425.

Kelly LJ, Renny-Byfield S, Pellicer J, et al (2015) Analysis of the giant genomes of *Fritillaria* (Liliaceae) indicates that a lack of DNA removal characterizes extreme expansions in genome size. New Phytol 208:596–607.

Lavin M, Herendeen P, Wojciechowski M (2005) Evolutionary Rates Analysis of Leguminosae Implicates a Rapid Diversification of Lineages during the Tertiary. Syst Biol 54:575–594.

Lewis, G., Schrire, B., Mackinder, B., Lock, M. (eds.) (2005). Legumes of the World. Royal Botanic Gardens, Kew. 577 pp.

Macas JJ, Novak P, Pellicer J, et al (2015) In Depth Characterization of Repetitive DNA in 23 Plant Genomes Reveals Sources of Genome Size Variation in the Legume Tribe Fabeae. PLoS One 10:e0143424.

Maddison WP, Maddison DR. (2011) Mesquite: a modular system for evolutionary analysis. Version 2.75. http://mesquiteproject.org.

Mayrose I, Barker MS, Otto SP (2010) Probabilistic models of chromosome number evolution and the inference of polyploidy. Syst Biol 59:132–144.

Michael TP (2014) Plant genome size variation: bloating and purging DNA. Briefings In Functional Genomics 13:308-317.

Murat F, Zhang R, Guizard S, et al (2015) Karyotype and gene order evolution from reconstructed extinct ancestors highlight contrasts in genome plasticity of modern rosid crops. Genome Biol Evol 7:735–749.

Nag A, Chanda S, Rajkumar S (2011) Estimation of nuclear genome size of important medicinal plant species from western Himalaya using flow cytometry. Journal of Cell & Plant Sciences 2:19-23.

Pandit MK, White SM, Pocock MJO (2014) The contrasting effects of genome size, chromosome number and ploidy level on plant invasiveness: A global analysis. New Phytol 203:697–703.

Pellicer J, Fay MF, Leitch IJ (2010) The largest eukaryotic genome of them all? Botanical Journal of the Linnean Society 164:10-15.

Peruzzi L (2013) " x " is not a bias, but a number with real biological significance. Plant Biosyst - An Int J Deal with all Asp Plant Biol 147:1238–1241.

Posada D. (2008) jModelTest: phylogenetic model averaging. Molecular Biology and Evolution, 25(7):1253.

R Core Team (2016) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

Rambaut A, Suchard MA, Xie D, Drummond AJ. (2014) Tracer v1.6, Available from http://beast.bio.ed.ac.uk/Tracer

Rambaut A. (2009) FigTree version 1.3.1, Available from http://tree.bio.ed.ac.uk.

Rice A, Glick L, Abadi S, et al (2015) The Chromosome Counts Database (CCDB) - a community resource of plant chromosome numbers. New Phytol 206:19–26.

Roa and Telles. The Cerrado Plant cytogenetics database in prep

Ronquist F, Huelsenbeck JP. (2003) MRBAYES 3: Bayesian phylogenetic inference under mixed models. Bioinformatics, v19, p1572-1574.

Sato S, Nakamura Y, Kaneko T, et al (2008) Genome Structure of the Legume, Lotus japonicus. DNA Res 15:227–239.

Schmutz J, Cannon SB, Schlueter J, et al (2010) Genome sequence of the palaeopolyploid soybean. Nature 463:178–183.

Schmutz J, McClean PE, Mamidi S, et al (2014) A reference genome for common bean and genome-wide analysis of dual domestications. Nat Genet 46:707–713.

Sousa A, Cusimano N, Renner SS (2014) Combining FISH and model-based predictions to understand chromosome evolution in *Typhonium* (Araceae). Ann Bot 113:669–680.

Suda J, Meyerson LA, Leitch IJ, Pyšek P (2015) The hidden side of plant invasions: The role of genome size. New Phytol 205:994–1007.

Tang, H.; Krishnakumar, V.; Bidwell, S.; Rosen, B.; Chan, A.; Zhou, S.; Gentzbittel, L.; Childs, K.L.; Yandell, M.; Gundlach, H.; Mayer, K.F.X.; Schwartz, D.C.; Town, C.D. (2014) An improved genome release (version Mt4.0) for the model legume *Medicago truncatula*. BMC Genomics, v.15:312.

The Plant List (2013) Published on the Internet, Version 1.1. http://www.theplantlist.org/. Accessed 3 Sep 2016

Varshney RK, Song C, Saxena RK, et al (2013) Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. Nat Biotechnol 31:240–246.

Wang, X.; Zhang, Z.; Wen, Z.; Xiao, H.; Yang, Z.; Chen, G.; Zhao, X. (2011) The chromosome number, karyotype and genome size of the desert plant diploid *Reaumuria soongorica* (Pall.) Maxim. Plant Cell Rep 30:955–964.

Wang S, Deng Z, Chung F, Hu W (2013) From Gaussian kernel density estimation to kernel methods. Int J Mach Learn Cybern 4:119–137.

Wang J, Liu J, Kang M (2015) Quantitative testing of the methodology for genome size estimation in plants using flow cytometry: a case study of the *Primulina* genus. Frontiers in Plant Science 6.

Whitney KD, Baack EJ, Hamrick JL, Godt MJW, Barringer BC, Bennett MD, Eckert CG, Goodwillie C, Kalisz S, Leitch IJ, Ross-Ibarra J (2010) A role for nonadaptive processes in plant genome size evolution? Evolution 2097-2109.

Wickham H (2009) ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York

CAPÍTULO 2

CARACTERIZAÇÃO DE ELEMENTOS REPETITIVOS EM SEQUÊNCIAS GENÔMICAS NUCLEARES DE *Dipteryx alata* (LEGUMINOSAE)

RESUMO

As regiões repetitivas são componentes abundantes dos genomas das plantas e exercem papéis na organização e evolução genômica. Esse estudo teve como objetivo identificar, caracterizar e quantificar regiões microssatélites e elementos transponíveis (TEs) em sequências genômicas nucleares de *Dipteryx alata* (Leguminosae), uma espécie de árvore nativa do cerrado brasileiro. Para isso foram extraídas amostras de DNA de folhas de quatro indivíduos de *D. alata*. As bibliotecas genômicas foram preparadas de acordo com o protocolo Nextera e o sequenciamento foi realizado em equipamento Illumina MiSeq. Reads que passaram por filtros de qualidade foram montados usando o software Dipspades em 275709 scaffolds com N50 igual a 1598, o que corresponde a 355MB. Microssatélites e TEs foram identificados usando os softwares QDD, RepeatMasker e RepeatModeler. Foram detectadas 21.981 regiões microssatélites, das quais 49,3% possuem motivos dinucleotídeos, 42,7% trinucleotídeo e 4% tetranucleotídeo. O tipo de motivo mais frequente foi AT. Os motivos microssatélites foram repetidos em tandem entre 5 e 29 vezes. Foram desenhados 40.980 pares de primers para regiões microssatélites, das quais 120 foram selecionadas com alto potencial para o desenvolvimento de marcadores moleculares. Foram encontrados TEs em 39,29% das sequências analisadas, o que corresponde a 421701 TEs. Os retrotransposons LTR (*gypsy* e *copia*) foram os TEs mais abundantes nas sequências. Foram desenhados primers para 100 TEs (*gypsy* e *copia*) visando o futuro desenvolvimento de marcadores baseados em transposons para *D. alata*. O grande número de microssatélites e TEs identificados poderão ser utilizados em estudos de genética de populações e evolução que contribuem para o desenvolvimento de programas de conservação e domesticação de Baru.

Palavras-chaves: Baru, elementos repetitivos, microssatélites, TEs.

INTRODUÇÃO

Um componente importante dos genomas das plantas são as sequências de DNA repetitivas, que podem representar até 90% do tamanho do genoma (Mehrotra et al 2014). As sequências nucleotídicas de genes e regiões reguladoras estão presentes em pequena proporção na maioria das espécies de plantas, enquanto elementos de DNA repetitivos representam uma fração significativa do genoma (Kubis et al 1998; Roy et al 2015). Presentes em várias cópias, as sequências repetitivas podem ser encontradas em locais específicos em uma organização em tandem, onde as cópias estão dispostas adjacentes umas as outras na sequência de DNA, ou podem ser encontradas dispersas por todo o genoma, com uma cópia separada das outras por até milhares de nucleotídeos (Treangen et al 2012; Mehrotra et al 2014).

As repetições em tandem são classificadas de acordo com o tamanho das unidades ou motivos de repetição. Os microssatélites, também conhecidos como repetições de sequências simples (SSRs), consistem em repetições curtas em tandem de motivos com 1 a 6 nucleotídeos, enquanto os minissatélites são repetições de motivos com 7 a 100 nucleotídeos, e os satélites repetição maiores, com mais de 100 nucleotídeos (Foulongne-Oriol et al 2013). Os microssatélites são abundantes nos genomas das plantas e surgem durante o processo de replicação do DNA por mutações de inserção ou deleção de um ou mais nucleotídeos. Embora eles sejam distribuídos aleatoriamente ao longo do genoma, os microssatélites são mais comuns em regiões não codificantes (Pandey e Sharma, 2012).

As sequências dispersas pelo genoma compreendem os elementos transponíveis (TEs) (Kubis et al 1998; Treangen et al 2012). Os TEs são fragmentos de DNA capazes de transpor, ou seja, de serem inseridos em novas localizações cromossômicas (Feschotte et al 2002). Os mecanismos de transposição caracterizam três classes principais de TEs, os retrotransposons, os DNAs transposons e os *Rolling-circles* (RC). Entre os retrotransposons o DNA é transcrito em RNA e em seguida transcrito de forma reversa em DNA complementar (cDNA), essa cópia é re-inserida em outra localização do genoma. Na classe de DNA transposons uma transposase, codificada pelo próprio TE, cataliza a clivagem do DNA dupla fita e sua inserção em outra localização genômica (Zhou et al 2011; Lopes e Bureau, 2014). Na classe RC, os transposons codificam uma proteína com domínio Rep que se associa a DNA helicases e a proteínas de ligação de DNA (SSBs). Este complexo molecular faz a clivagem endonucleotídica da sequência do TE e a transferência para outro local genômico. Ao contrário das outras classes, em que os TEs são transpostos como fitas duplas de DNA, os RC usam um mecanismo diferente, em que o TE é deslocado como fita simples do local de origem para o local de destino e forma uma estrutura circular. Entre os RC, a transposição segue um modelo semi-replicativo, pois o TE serve de molde para a síntese de DNA catalizada pela maquinaria enzimática do local de inserção (Kapitonov e Jurka, 2007). Assim como os microssatélites, os TEs são abundantes nos genomas de plantas.

Apesar da sua abundância, os componentes repetitivos do genoma já foram considerados "DNA lixo" ou “DNA parasitário” devido ao desconhecimento de suas funções (Orgel e Crick, 1980; Mehrotra et al 2014). Contudo, os elementos repetitivos são componentes importantes que influenciam o tamanho, a organização e a evolução do genoma

(Foulongne-Oriol et al 2013). As diferenças entre os tamanhos dos genomas da planta são diretamente influenciados pela variação quantitativa no DNA repetitivo, visto que sequências de genes e regiões reguladoras são geralmente semelhantes entre espécies diferentes (Mehrotra et al 2014). Além de contribuir para o aumento do tamanho dos genomas, os elementos repetitivos desempenham papéis na organização genômica, uma vez que repetições em tandem e TEs são os principais constituintes das regiões de centrômero e telômero dos cromossomos eucarióticos. A grande quantidade de elementos repetitivos dessas regiões heterocromáticas sugere que o DNA repetitivo possui papéis estruturais e funcionais como a estabilidade cromossômica e a participação no ciclo celular (Foulongne-Oriol et al, 2013; Lopes e Bureau, 2014).

Os TEs estão associados com a organização do genoma porque os mecanismos de transposição produzem quebras da dupla fita do DNA que implicam na indução de rearranjos cromossômicos, evento que influencia as características cariotípicas. Indícios científicos indicam que genomas ricos em TEs possuem mais eventos de rearranjos e por isso um cariótipo com menor número de cromossomos (Lopes e Burreau, 2014). Nesse sentido os TEs exercem papéis na evolução do genoma, já que os rearranjos cromossômicos favorecem a geração de variação genética, e podem contribuir para a origem de novos genes e redes reguladoras no genoma. Adicionalmente, as regiões com repetições em tandem mostram taxas de mutação mais elevadas, sendo uma chave fundamental para o desenvolvimento da teoria neutra da evolução molecular. Neste sentido, as regiões microssatélites mostram altos níveis de polimorfismo intra e inter populações (Kubis et al 1998; Tóth et al 2000). Devido aos vários papéis desempenhados pelas sequências repetitivas de DNA, é importante a identificação e caracterização destas regiões a fim de compreender a estrutura, o funcionamento e a evolução do genoma (Vaidya et al 2012).

O desenvolvimento das plataformas de sequenciamento de nova geração (Next Generation Sequencing - NGS) e os avanços das ferramentas de bioinformática facilitam estudos *de novo* e permitem um avanço notável nos conhecimentos acerca da abundância e distribuição dos elementos repetitivos nos genomas (Mehrotra et al 2014). As tecnologias de alto rendimento revolucionaram a capacidade de produzir sequências de nucleotídeos. Grandes quantidades de sequências de DNA ou de RNA podem ser produzidas em um curto período de tempo e com baixo custo (Mardis 2008; Metzker 2010). Com o desenvolvimento de ferramentas de bioinformática, grandes e complexos genomas eucarióticos podem ser montados em uma questão de dias, com ou sem uma sequência de referência (Schatz et al 2010). As sequências montadas podem ser anotadas visando investigar o DNA repetitivo em genomas de espécies modelos e não modelos, como é o caso de *Dipteryx alata*.

Dipteryx alata (Baru, Leguminosae) é uma espécie de árvore endêmica do bioma Cerrado brasileiro com importante potencial econômico, principalmente devido à polpa e as sementes de seus frutos, que são utilizados *in natura* ou na fabricação de sorvetes, licores, castanhas torradas, e outros produtos por pequenas e médias indústrias. Além do uso alimentício, *D. alata* tem utilidade medicinal, forrageira, na recuperação de áreas degradadas, no paisagismo e na extração de madeira. Essa árvore, pertencente à família Leguminosae, possui madeira de alta densidade e durável que também pode ser comercializada (Sano et al 2004).

Para *D. alata* já foram publicados artigos científicos com foco em desenvolvimento de marcadores moleculares em pequena escala e estudos populacionais (Soares et al 2008a; Soares et al 2008b; Collevatti et al 2013; Soares et al 2015; Diniz-Filho et al 2016a; Diniz-Filho et al 2016b). No entanto, as investigações genômicas disponível para a espécie ainda são raras e uma maior quantidade de informação genômica é necessária para a compreensão da biologia de *D. alata.*

Assim, os objetivos deste estudo, que pode ser considerado um dos primeiros estudos genômicos para espécies de árvores nativas do cerrado brasileiro, foram identificar, caracterizar e quantificar regiões microssatélites e TEs em sequências genômicas nucleares de *Dipteryx alata* usando sequenciamento de nova geração.

MATERIAL E MÉTODOS

Material vegetal e extração de DNA

Para a extração de DNA foram coletadas folha frescas de quatro indivíduos adultos de *Dipteryx alata* provenientes de diferentes regiões do bioma Cerrado: Natividade-TO, Camapuã-MS, Aquidauana-MS e Cáceres-MT. As amostras foram imediatamente armazenadas em freezer -80°C e os DNA de cada indivíduo foram extraídos usando o método brometo de cetiltrimetilamônio (CTAB) descrito por Doyle e Doyle (1987). A qualidade do DNA extraído foi verificada por eletroforese horizontal em gel de agarose e a quantificação foi realizada utilizando o ensaio Qubit® DNA Fluorometer (Invitrogen).

Preparo das bibliotecas de DNA e sequenciamento Illumina

As etapas de preparação das bibliotecas foram realizadas para cada amostra de DNA, seguindo o protocolo do kit Nextera da Illumina. As amostras de DNA foram fragmentadas, ligadas a adaptadores, amplificadas por PCR e marcadas com códigos de barras. A validação das bibliotecas foi realizada por meio de análises qualitativa com o Agilent Technologies 2100 Bioanalyzer usando o High Sensitivity DNA kit. As bibliotecas foram quantificadas por meio de PCR em tempo real (Applied Biosystems StepOne), usando o kit Kapa (Kapa Biosystems). As bibliotecas foram agrupadas e sequenciadas usando a plataforma MiSeq da Illumina e o kit de sequenciamento Miseq v3 de 600 ciclos (2x250bp, reads paired-end). Foram realizadas duas corridas de sequenciamento, na primeira foram sequenciadas as quatro bibliotecas juntas e na segunda corrida foi sequenciada apenas a biblioteca de Camapuã-MS.

Análise de qualidade e montagem de sequências genômicas nucleares

A qualidade dos reads foi avaliada usando o software fastQC (Andrews, 2010) e bases de baixa qualidade, bem como sequências de adaptadores, foram removidas usando o software Trimmomatic (Bolger et al 2014). Foram usadas as funções HEADCROP (18), CROP (250) e ILLUMINACLIP (2:30:10) do Trimmomatic. Os reads de cada biblioteca foram cortados usando os valores de phred 10, 15 e 20 como limiar mínimo de qualidade. Para isso, esses valores de phred foram usados nas funções LEADING, TRAILING e

SLIDINGWINDOW do Trimmomatic. Os três conjuntos de reads filtrados foram alinhados ao genoma da espécie *Medicago Truncatula*, utilizado como genoma de referência nessa etapa, usando o software bowtie2 (Langmead et al 2009), e o conjunto com maior número de reads alinhados foi selecionado para montagem das sequências genômicas. Dessa forma, para a biblioteca de Camapuã-MS foi usado o filtro com valor de phred15 e para as bibliotecas de Natividade-TO, Aquidauana-MS e Cáceres-MT foi usado o filtro com valor de phred 20, valores estabelecidos como ótimos de filtragem. Reads menores do que 20 pares de bases foram eliminados usando a função MILEN do Trimmomatic. No final da análise de qualidade, foram produzidos dois tipos de arquivos: paired reads (PE) e reads órfãos chamados single reads (SR).

A montagem dos reads foi realizada usando algoritmos baseados na construção de grafos de bruijn. Três montadores foram testados para a montagem das sequências de *D. alata*: MaSuRCA 3.1.3 (Zimin et al 2013), Dipspades 1.0 (Safonova et al 2015) e ABySS (Simpson et al 2009). Nas três montagens, as sequências organelares, bem como as sequências menores que 500pb, foram excluídas. As sequências organelares foram identificadas por meio da comparação com sequências conhecidas armazenadas em banco de dados. Para isso foram criados dois bancos de dados, um para cloroplasto e outro para mitocôndria, usando as sequências organelares de angiospermas depositadas no RefSeq do GenBank. A análise de similaridade foi realizada com o blast (*Basic Local Alignment Search Tool*) (Altschul et al 1990). Sequências com hits com *e-value* menor que $1e^{-16}$, cobertura do query menor que 85% e identidade menor que 95% foram retirados da montagem. A qualidade das montagens foi avaliada usando o script "assemblathon_stats.pl", desenvolvido por Keith Bradnam (2011). A montagem usada para anotação de regiões repetitivas foi selecionada com base na comparação dos seguintes parâmetros: volume de dados montados, número de scaffolds, tamanho do maior scaffold e N50. A cobertura da montagem foi estimada utilizando a ferramenta genomeCoverageBed da suíte bedtools (Quinlan e Hall, 2010).

Anotação de regiões microssatélites e desenho de primers

As sequências nucleares montadas foram investigadas para identificação de regiões microssatélites perfeitas utilizando o software QDD versão 3.1.2 instalado em VirtualBox 5.0.14 para hospedagem em Windows (Meglécz et al 2010). O QDD é um software que identifica regiões microssatélites e desenha primers em grandes projetos de sequenciamento de genomas. As análises foram realizadas em quatro etapas: na etapa 1 foi efetuada a identificação e extração dos microssatélites a partir do arquivo fasta da montagem; na etapa 2 foi feita a detecção de similaridade entre as sequências por meio da comparação de todos contra todos do Blast e somente sequências únicas (singletons) foram mantidas para o desenho de primers; na etapa 3 foram desenhados primers para as sequência selecionadas usando o software Primer3; na etapa 4 as sequências foram analisadas com o software RepeatMasker para verificar a associação das regiões microssatélites com TEs. Os softwares Primer3 e RepeatMasker foram utilizados pelo pipeline do QDD (Rozen e Skaletsky, 2000; Tarailo-Graovac e Chen 2009; Untergasser et al 2012).

O número mínimo de repetições para cada tamanho de motivo usado na busca dos microssatélites foram: dez para di-nucleotídeos, seis para tri e tetra-nucleotídeos e cinco para microssatélites penta e hexa-nucleotídeos. O desenho de primers foi realizado usando 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 22-25 pb. Visando o futuro desenvolvimento de marcadores microssatélites foram selecionados 40 pares de primers para dinucleotídeos, 40 para tetranucleotídeos e 40 para pentanucleotídeos. Foram selecionados primers para regiões microssatélites que não estavam associadas a TEs e com os maiores números de repetições em tandem, sendo no mínimo 14 repetições para dinucleotídeos, 7 repetições para tetranucleotídeos e 6 repetições para pentanucleotídeos. Para cada região microssatélite foi escolhido apenas um par de primer e este foi selecionado com base em alguns critérios sugeridos no manual do QDD: menor score de alinhamento com sequências diferentes do sítio de anelamento, distância maior de 20 pb entre os primers e o microssatélite alvo e seleção de marcadores com diferentes comprimentos de produto de PCR para facilitar a multiplexação.

Anotação de TEs e desenho de primers

As repetições transposons, retrotransposons e *Rolling-circles* nas sequências do genoma nuclear de *D. alata* foram detectadas e anotadas usando os softwares RepeatMasker e RepeatModeler (Tarailo-Graovac e Chen, 2009). O RepeatMasker rastreia elementos repetitivo em sequências de DNA por meio da comparação contra bibliotecas de repetições de espécies com genoma conhecido armazenado em banco de dados. O RepeatModeler emprega métodos computacionais *de novo* para a identificação e modelagem de famílias de repetição.

A identificação dos TEs baseados em similaridade foi realizada por meio da análise do RepeatMasker usando o arquivo fasta com as sequências nucleares montadas e o banco de dados de eudicotyledoneas. Além dos arquivos com os resultados dos TEs identificados, o RepeatMasker gerou um arquivo de saída ".fasta.masked" contendo uma cópia das sequências nucleares montadas mascaradas, em que as bases das sequências dos TEs foram substituídas por N. Para a identificação dos TEs *de novo* foi criado um banco de dados do RepeatModeler com as sequências de *D. alata*, e em seguida realizada a análise do RepeatModeler usando o arquivo fasta com as sequências nucleares montadas e esse banco criado. A obtenção do total de TEs identificados *de novo* foi feito através de uma segunda análise do RepeatMasker usando o arquivo de saída do RepeatModeler com extensão ".classifield" e o arquivo fasta com a montagem das sequências nucleares. Para analisar os TEs que foram identificados pelas duas metodologias, baseada em similaridade e *de novo*, foi realizada um terceira corrida do RepeatMasker usando o arquivo de saída do RepeatModeler ".classifield" e o arquivo de saída da primeira análise realizada no RepeatMasker ".fasta.masked". Nessa terceira análise com o RepeatMasker foram identificados os TEs anotados apenas *de novo*, visto que os elementos identificados por similaridade estavam mascarados.

Retroelementos foram selecionados para o desenho de primer visando o futuro desenvolvimento de marcadores RBIP (*Retrotransposon Based Insertion Polymorphism*) para *D. alata*. Foram selecionados elementos LTR *copia* e LTR *gypsy* com tamanho máximo de

1100pb e localizados em sítios onde não havia sobreposição de repetições. O desenho de primers foi realizado com o software Primer3 versão web, usando regiões da sequência que flanqueiam os TEs, e com os seguintes parâmetros: teor de GC de 30-60%, o valor de Tm de 52-62 °C, e comprimentos dos primers de 20-25 pb (Rozen e Skaletsky, 2000). As sequências dos transposons usadas para o desenho de primers foram submetidas a uma análise com o Blastx (*Basic Local Alignment Search Tool*), usando um E-valor menor que 10^{-6}, visando à identificação de domínios proteicos conservados e de TEs em contexto gênico (Altschul at al 1990).

RESULTADOS

Sequenciamento e montagem das sequências genômicas

As duas rodadas de sequenciamento com a plataforma MiSeq da Illumina geraram 22 GB de dados que correspondem a 61,62 milhões de reads paired-end. Após os cortes de qualidade foram mantidos 60,55 milhões de reads (14.6 Gb) e estes foram usados para montagem das sequências genômicas de *D. alata* (Tabela 1).

Tabela 1 – Características das montagens realizadas para *D. alata* usando dados de sequenciamento Illumina.

Parâmetros	MaSuRCA	Dipspades	ABySS
Volume de dados montados (MB)	144	355	156
Número total de scaffolds	152.905	275.709	189.041
Tamanho total dos scaffold (pb)	145.859.029	363.927.807	158.625.771
Tamanho do maior scaffold (pb)	10.303	87.189	6.998
Tamanho do menor scaffold (pb)	500	500	500
N50 dos scaffolds	1.002	1.598	846
Scaffolds > 1k nucleotídeos	48.114	137.850	41.588
Scaffolds > 10k nucleotídeos	1	140	0

A montagem realizada com o software Dipspades foi selecionada para anotação de regiões repetitivas porque atingiu os melhores resultados entre os montadores, gerou a maior quantidade de sequências montadas e com maior N50. Nessa montagem, o tamanho do maior e do menor scaffold, bem como o tamanho total dos scaffolds, foram de 87189, 500 e 363927807 pb, respectivamente. Dos scaffolds montados 137850 eram maiores que 1k nucleotídeos e 140 scaffolds maiores que 10k nucleotídeos (Figura 1). A montagem final possui 355MB de sequências do genoma nuclear de Baru. A cobertura da montagem foi estimada em 12.78X.

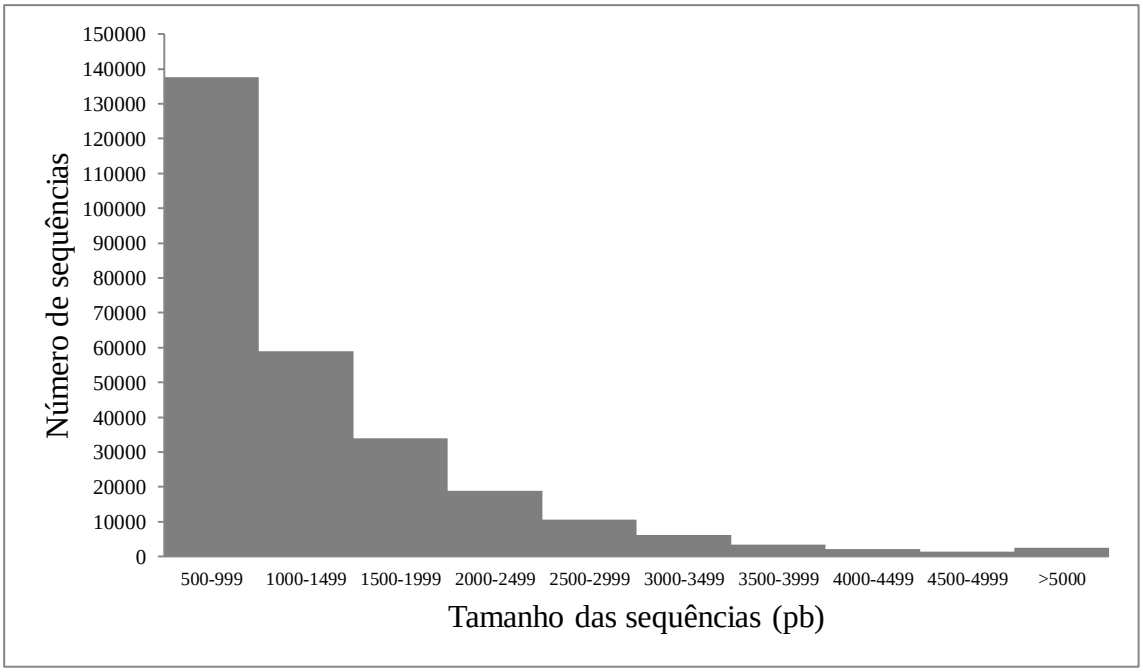

Figura 1 - Distribuição dos tamanhos das sequências nucleares na montagem realizada com software Dipspades.

Anotação de regiões microssatélites em *Dipteryx alata*

Entre o conjunto de sequências genômicas montadas para *D. alata* 19.945 apresentaram regiões microssatélites, sendo que 18155 sequências possuem apenas uma região, enquanto 1.602 possuem 2, 154 possuem 3, 18 possuem 4, 5 possuem 5 e 5 possuem mais de 5 regiões microssatélites. Nessas sequências foram identificadas um total de 21.981 regiões microssatélites, das quais 49,3% apresentaram tamanho dinucleotídeos, 42,7% trinucleotídeo, 4% tetranucleotídeo, 2,7% pentanucleotídeo e 1,3% hexanucleotídeo (Figura 2). Dessa forma, o tamanho de motivo de repetição mais comum foi o dinucleotídeos e o motivo de repetição mais frequente foi a AT (Tabela 2). Considerando o total de regiões microssatélites identificadas e todos os tamanhos de motivos, os mais frequentes foram: para dinucleotídeos AT (24%), AG (8,6%) e CT (8,3%), para trinucleotídeos AAT (10,9%), ATT (10,5%), AAG (5,5%), para tetranucleotídeos TATT (0,75%), AAAT (0,73%) e CTTT (0,45%), para pentanucleotídeos ATTTT (0,4%), AAAAT (0,3%) e CTTTT (0,3%) e para hexanucleotídeos AAAAAT (0,10%), TATTTT (0,08%) e AAAAAC (0,06%). Os motivos microssatélites foram repetidos em tandem entre 5 e 29 vezes (Figura 3). O total de microssatélites identificados ocuparam 516,7 KB das sequências de *D. alata*.

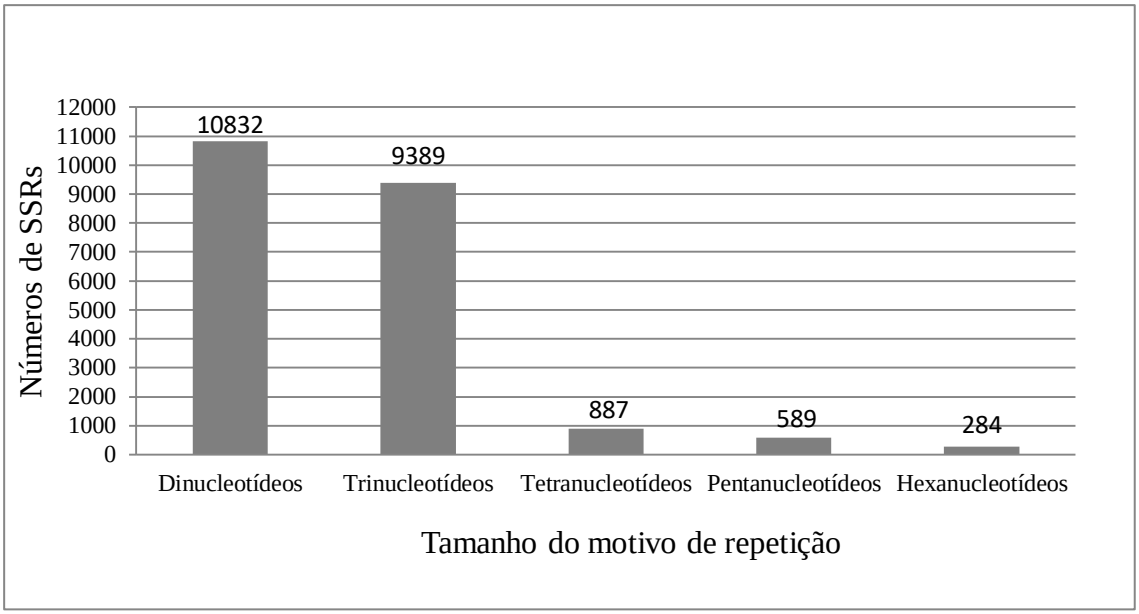

Figura 2 – Distribuição dos microssatélites encontrados em sequências genômicas de *Dipteryx alata* por tamanho de motivo de repetição.

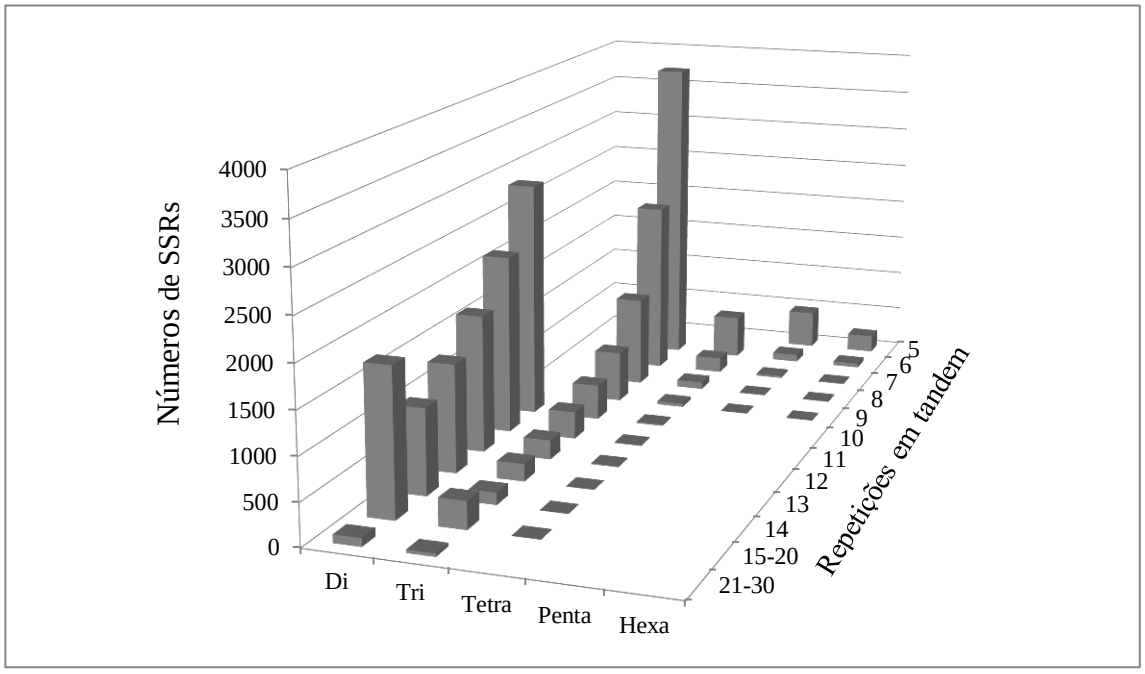

Figura 3 – Distribuição dos microssatélites encontrados em *D. alata* por tamanho de motivo de repetição e número de repetições em tandem.

Tabela 2 – Frequência dos motivos de repetição em sequências genômicas de *D. alata*.

Motivo	Frequência absoluta	Frequência relativa (%)	Motivo	Frequência absoluta	Frequência relativa (%)
AT	5290	24,06	AATAT	22	0,10
AAT	2396	10,90	AATTT	21	0,09
ATT	2312	10,51	AAATT	20	0,09
AG	1893	8,61	AAAAC	19	0,08
CT	1831	8,32	ATTT	18	0,08
AAG	1203	5,47	GTTT	18	0,08
CTT	1151	5,23	TATTTT	18	0,08
GT	927	4,21	ATCT	17	0,07
AC	885	4,02	GATA	16	0,07
AAC	565	2,57	ATTTG	16	0,07
GTT	543	2,47	GTTTT	15	0,06
ATC	233	1,06	ATTTT	15	0,06
ATG	190	0,86	AAAAAC	15	0,06
TATT	165	0,75	TCTTTT	14	0,06
AAAT	161	0,73	AAAC	13	0,05
CTTT	103	0,46	AAAAGA	13	0,05
AAAG	99	0,45	CTCTT	9	0,04
ACC	89	0,40	CTCCTT	8	0,03
ATTTT	89	0,40	CCTT	7	0,03
ACAT	88	0,40	CAGAA	7	0,03
GGT	87	0,39	CG	6	0,02
CCT	86	0,39	AAGG	6	0,02
CTG	86	0,39	AGTG	6	0,02
AGC	82	0,37	AAATC	6	0,02
AGG	81	0,36	GAGAA	6	0,02
AAAAT	81	0,36	TCTCTT	6	0,02
ATGT	76	0,34	AATC	5	0,02
CTTTT	73	0,33	ATTC	5	0,02
AAAAG	64	0,29	AATTG	5	0,02
AATT	46	0,20	AATCC	5	0,02
CGG	41	0,18	GGAAG	5	0,02
CTC	41	0,18	TTGTTT	5	0,02
CCG	38	0,17	TTTAAT	5	0,02
GTC	38	0,17	AAATTT	5	0,02
GAG	37	0,16	ATATTT	5	0,02
ACG	34	0,15	AACT	4	0,01
CTA	30	0,13	CGTG	4	0,01
AGT	26	0,11	TGGTT	4	0,01
ATATT	26	0,11	Outros	274	1,28
AAAAAT	23	0,10	TOTAL	21981	100

Desenho de primers para regiões microssatélites em *Dipteryx alata*

Foram desenhados um total de 40.980 pares de primers para 8.902 regiões microssatélites, o que corresponde em média a 4 pares de primers por região (Tabela Suplementar online 1). Dos primers desenhados, 4.379 foram desenvolvidos para regiões com motivos dinucleotídeos, 3.845 para trinucleotídeos, 363 para tetranucleotídeos, 218 para pentanucleotídeos e 97 para hexanucleotídeos. Entre as 8.902 regiões com primers, 4.903 não estavam associadas a TEs. Para o conjunto de regiões sem associação com TEs, 1.838 eram regiões dinucleotídeos, 2.533 eram trinucleotídeos, 271 eram tetranucleotídeos, 187 eram pentanucleotídeos e 74 eram hexanucleotídeos. Foram desenhados 120 pares de primers com alto potencial para o desenvolvimento de marcadores moleculares (Tabela Suplementar online 2).

Anotação de TEs em *Dipteryx alata*

Entre as sequências de *D. alata* analisadas com o RepeatMasker e o RepeatModeler, um total de 39,29% bases foram mascaradas, o que correspondeu a 143.004.122 pb. Os retrotransposons LTR (*gypsy*/DIRS1 e Ty1/*copia*) foram os TEs mais abundantes nas sequências de baru, seguidos dos retrotransposons LINE/L1, dos DNA transposons CMC-EnSpm e dos retrotransposons LINE RTE-BovB (Tabela 3). Um total de 45,8% dos TEs identificados não tiveram a classe determinada.

Tabela 3 – Anotação do total de TEs em sequências de *D. alata* usando RepeatModeler e RepeatMasker.

	Ordem	Superfamília	Número de elementos	Porcentagem de elementos (%)	Sequências ocupadas (bp)	Percentual de sequências no genoma (%)
Retroelements	LTR elementos	Caulimovirus	5.477	1,29	2.617.339	0,71
		copia	52.388	12,42	22.676.307	6,23
		gypsy	120.642	28,60	50.992.893	14,01
		Não classificados	657	0,15	166.344	0,04
	LINE	LINE/L1	10.887	2,58	3.719.137	1,02
		LINE/L2	10	0,002	415	0,0001
		RTE-BovB	6.966	1,65	2.120.408	0,58
	SINE		808	0,19	101.495	0,02
DNA transposons	TIR	CMC-EnSpm	7.451	1,76	2.738.611	0,75
		Crypton	261	0,06	91.752	0,02
		hAT-Ac	2.778	0,65	758.007	0,20
		hAT-Charlie	72	0,01	8.773	0,002
		hAT-Tag1	4.295	1,01	1.382.478	0,37
		hAT-Tip100	1.016	0,24	302.076	0,08
		MuLE-MuDR	6.133	1,45	1.942.663	0,53
		PIF-Harbinger	2.678	0,63	628.614	0,17
		TcMar-Fot1	112	0,02	32.231	0,008
		TcMar-Stowaway	257	0,06	44.481	0,01
		Não classificados	163	0,03	11.485	0.003
Rolling-circles		Helitron	5.257	1,24	2.178.518	0,59
Não Classificados			193.393	45,86	50.490.095	13,87
Total			421.701	100	143.004.122	39,29

Entre as 39,29% de sequências de *D. alata* identificadas como TEs, 26,67% foram anotadas exclusivamente pela metodologia *de novo*, ou seja, não foram encontrados em sequências de outras espécies armazenadas em bancos de dados (Tabela 4). Assim, para os 421.701 TEs identificados, 76% foram identificados exclusivamente por análise *de novo*, 8,4% foram identificados exclusivamente por análise de similaridade e 15,6% foram identificados por ambas às metodologias (Figura 4).

Tabela 4 – Anotação dos TEs identificados exclusivamente para *D. alata* usando RepeatModeler e RepeatMasker.

	Ordem	Superfamília	Número de elementos	Porcentagem de elementos (%)	Sequências ocupadas (bp)	Percentual de sequências no genoma (%)
Retroelements	LTR elementos	Caulimovirus	3.172	0,99	1.333.842	0,36
		copia	23.877	7,45	7.783.079	2,13
		gypsy	73.669	22,99	29.271.929	8,04
		Não classificados	539	0,16	153.647	0,04
	LINE	LINE/L1	5.551	1,73	1.810.371	0,49
		LINE/L2	10	0,003	415	0,0001
		RTE-BovB	906	0,28	132.324	0,03
DNA transposons	TIR	CMC-EnSpm	4.190	1,30	1.372.393	0,37
		Crypton	261	0,08	91.752	0,02
		hAT-Ac	2.057	0,64	588.520	0,16
		hAT-Charlie	72	0,02	8.773	0,002
		hAT-Tag1	3.120	0,97	890.825	0,24
		hAT-Tip100	541	0,16	137.546	0,03
		MuLE-MuDR	3.849	1,20	1.359.835	0,37
		PIF-Harbinger	1.672	0,52	378.148	0,10
		TcMar-Fot1	112	0,03	32.231	0,008
		TcMar-Stowaway	236	0,07	41.966	0,01
Rolling-circles		Helitron	3.145	0,98	1.198.621	0,32
Não Classificados			193.375	60,36	50.488.236	1,87
Total			320.354	100	97.074.453	26,67

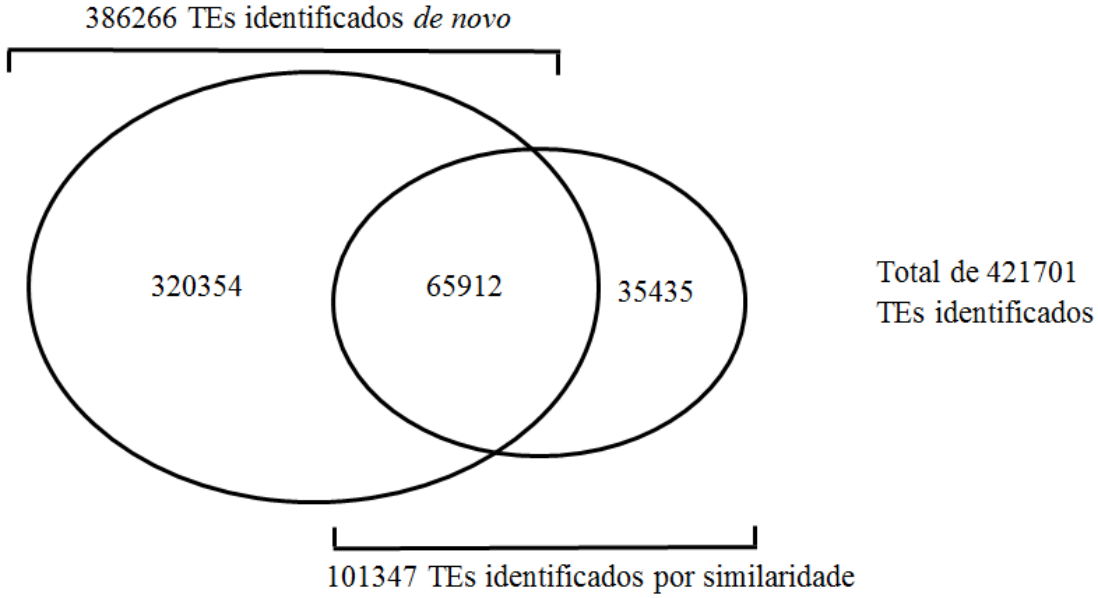

Figura 4 - Diagrama de Venn com o número de TEs identificados nas sequências de *D. alata* usando as duas metodologias de anotação.

Os TEs com maior comprimento máximo de sequência em *D. alata* foram os LTR *gypsy* (5.684pb), seguidos dos LTR Caulimovirus (5.300pb) e dos LTR *copia* (5.100pb) (Tabela 5). Dessa forma, os elementos *gypsy* foram os mais abundantes e com maior comprimento máximo de sequências em *D. alata*.

Tabela 5 – Tamanhos mínimos, máximos e médios dos TEs identificados em *D. alata*.

	Tamanho mínimo (pb)	Tamanho máximo (pb)	Tamanho médio (pb)
Caulimovirus	12	5.300	477
copia	10	5.100	432
gypsy	10	5.684	422
Helitron	10	1.330	414
CMC-EnSpm	10	3.297	367
Crypton	36	904	351
LINE/L1	10	2.465	341
hAT-Tag1	12	2.397	321
MuLE-MuDR	10	3.703	316
RTE-BovB	12	3.103	304
hAT-Tip100	17	2.052	297
TcMar-Fot1	27	1.986	287
hAT-Ac	24	1.989	273
PIF-Harbinger	11	2.676	234
TcMar-Stowaway	23	716	173
SINE	12	186	125
hAT-Charlie	30	299	121
LINE/L2	30	37	31

Desenho de primers para retroelementos LTR *gypsy* e *copia* em *Dipteryx alata*

Foram projetados 60 pares de primers para elementos LTR/*gypsy* e 40 pares de primers para elementos LTR/*copia* (Tabela Suplementar online 3). Os tamanhos do maior e do menor elemento *gypsy* usado como região alvo para o desenho de primers foram 1.036pb e 50pb, respectivamente. O tamanho do maior e do menor elemento Cópia usado no projeto dos primers foi 831pb e 42pb, respectivamente. Um percentual de 45% dos TEs usados no desenho de primers tiveram similaridade significativa com sequências de proteínas armazenadas em banco de dados para outras espécies de plantas, incluindo espécies da família Leguminosae tais como *Glycine max*, *Cicer arietinum* e *Vigna radiata* (Tabela Suplementar online 4).

DISCUSSÃO

A quantidade de dados de sequências de DNA gerada para *Dipteryx alata* usando a plataforma MiSeq Illumina permitiu montar um rascunho de parte do genoma da espécie. Os três montadores testados para *D. alata* forneceram resultados distintos. Apesar dos três montadores usarem a estratégia de grafos de bruijn para fazer a montagem *de novo*, cada um possui métodos diferentes, por exemplo, para lidar com erros de sequenciamento, para lidar com coberturas baixas e não uniformes e para resolver sequências repetitivas que são abundantes em genomas de plantas (Miller et al 2010). Sequências repetitivas estão presentes em grande quantidade no genoma de *D. alata*. Dessa forma, como cada espécie possui estrutura genômica própria, o resultado de um montador pode variar entre espécies. Por isso é importante testar diferentes montadores para cada conjunto de dados visando escolher o de melhor desempenho (Ištvánek et al 2014). O montador Dipspades gerou o melhor resultado de montagem para *D. alata* visto que montou um volume de sequências maior (355MB) e com maior N50 (1598).

A montagem obtida com o software dipspades representa 44% do genoma, visto que o tamanho do genoma de *D. alata* foi estimado em 807,2 MB usando citometria de fluxo (Capítulo 1). O conjunto de sequências montadas permitiu uma caracterização abrangente de elementos repetitivos para a espécie. Este resultado sugere que investigações biológicas significativas podem ser realizadas sem a necessidade de sequenciar todas as bases do genoma. A caracterização dos elementos repetitivos usando os métodos de sequenciamento de nova geração não depende de uma elevada cobertura do genoma (Yu et al 2011). O número de sequências obtidas para *D. alata* permitiu a anotação de grande número de microssatélites e TEs e gerou informações genômica até então não disponível para esta espécie do Cerrado brasileiro.

O percentual de regiões microssatélites encontrada em *D. alata* (516.7 KB) corresponde a 0,14% das sequências investigadas, valor semelhante ao encontrado no genoma da espécie *Medicago truncatula* onde 0,18% das 246 MB de sequências foram anotadas como microssatélites (Young et al 2011). Na amostra do genoma de *D. alata* os tamanhos de motivo de repetição microssatélite mais comuns foram os dinucleotídeos (49,3%) e trinucleotídeos (42,7%). Este resultado é semelhante ao encontrado no genoma de *Cicer arietinum*, onde os dinucleotídeos e trinucleotídos representavam 57,6% e 35,9% do total de 81.845 regiões

microssatélites encontradas em 532 MB de sequências investigadas (Varshney et al 2013). Nesse sentido, Yang et al (2012) afirmam que para várias espécies repetições di e tri-nucleotídeos constituem a maior proporção de microssatélites, enquanto mono, tetra, penta e hexa-repetições são consideravelmente menos frequentes. As regiões microssatélites de *D. alata* com maior número de repetições em tandem foram as di e trinucleotídeo. Segundo Zalapa et al (2012) em plantas os microssatélites com motivos mais curtos geralmente têm maior número de repetições do que microssatélites com motivos mais longos como penta ou hexanucleotídeos. Quanto ao tipo de repetição, o motivo AT foi o que ocorreu com maior frequência nas sequências do genoma de *D. alata*. Esse resultado é correspondente ao encontrado na espécie modelo *Arabidopsis thaliana* onde o motivo de repetição AT também é o mais frequente no genoma (Yu et al, 2011).

Este estudo permitiu a identificação de um grande número de regiões microssatélites para *D. alata* que são potencialmente úteis para o desenvolvimento de marcadores moleculares. De acordo com Yu et al (2011) testes para o desenvolvimento de marcadores moleculares podem ser feitos através do desenho de primers para a região alvo, da amplificação por PCR (Reação em Cadeia da Polimerase) e do sequenciamento do locus microssatélite para confirmar a presença da sequência no genoma. O software QDD permitiu o desenho de primers para 8902 regiões microssatélites de *D. alata*, ou seja, para 40% das regiões identificadas. Desse conjunto foram selecionados pares de primers para 120 regiões com alto potencial para marcadores moleculares. Informações da literatura foram usadas para selecionar as regiões microssatélites de *D. alata* com características favoráveis ao desenvolvimento de marcadores. Foram escolhidas regiões com maiores números de repetições em tandem uma vez que microssatélites que tem um maior número de unidades de repetição são mais susceptíveis de ser associados com grande variação genotípica (Zalapa et al 2012). Não foram selecionados primers para regiões mono, tri e hexanucleotídeo. Esses tamanhos de motivos de repetição não são recomendados para o desenvolvimento de marcadores, uma vez que mononucleotídeos podem resultar de erros de sequenciamento e tri e hexanucleotídeos são frequentes em regiões codantes do genoma, regiões sob seleção (Yu et al 2011). Microssatélites com motivo AT, apesar de serem os mais frequentes no genoma, também não foram selecionadas para o desenvolvimento de marcadores porque podem formar *hairpin* diminuindo a eficiência da amplificação durante a PCR. Regiões microssatélites associadas a TEs também não foram selecionadas para o desenvolvimento de marcadores microssatélites devido à possibilidade de alelos nulos gerados pelos eventos de transposição. Dessa forma, nesse estudo foi feito o desenho de primers em larga escala para regiões microssatélites de *D. alata* e foram selecionadas um conjunto de regiões com várias características potenciais para gerar marcadores moleculares.

Alguns estudos têm feito o desenvolvimento de marcadores microssatélites utilizando como abordagem metodológica o sequenciamento Illumina e montagem de parte do genoma. Um exemplo é o trabalho de Silva et al (2013) com a espécie *Brachiaria ruziziensis*. Silva et al (2013) identificaram 85.567 loci microssatélites no genoma parcial da espécie (277MB), das quais 500 foram selecionados para a desenho e síntese de primers e 198 foram confirmados polimórficos em 11 indivíduos *B. ruziziensis*. Tal como no estudo de Silva

et al (2013), as sequências de primers desenhadas para *D. alata* podem ser testadas a fim de caracterizar marcadores polimórficos e informativas para a espécie.

Esta estratégia baseada no sequenciamento de nova geração pode facilitar estudos genéticos em populações de *D. alata*, uma espécie que ainda tem poucos marcadores microssatélites disponíveis atualmente (Soares et al 2012). Existem várias dificuldades no desenvolvimento de marcadores microssatélites utilizando métodos convencionais baseados no uso de sondas artificiais, na clonagem de fragmentos e no sequenciamento Sanger. Estes métodos são trabalhosos, requerem um longo tempo de execução, tem alto custo e ainda apresentam baixa taxa de sucesso porque geram uma porcentagem de microssatélites baixas (Yu et al 2011; Pandey e Sharma, 2012). Além destas dificuldades metodológicas, *D. alata* tem baixa diversidade genética dentro das populações o que torna ainda mais difícil desenvolver marcadores microssatélites para a espécie (Collevatti et al 2013). Neste cenário, a caracterização de regiões microssatélites candidatos em larga escala seguido do desenvolvimento de marcadores microssatélites utilizando processos mais rápidos e mais baratos do que protocolos baseados em Sanger, vai facilitar muitas investigações biológicas em *Dipteryx alata*.

Os marcadores microssatélites são ferramentas valiosas em vários estudos em plantas como estudos de diversidade genética, análise da estrutura populacional, estudo de parentesco, genética da conservação, investigações forenses, estudos evolutivos, estudos filogenéticos em populações naturais de espécies ameaçadas de extinção, entre outros (Matsumoto et al 2010; Yu et al 2011; Zalapa et al 2012; Yahya et al 2014). Microssatélites emergiram como um importante marcador em muitas áreas de genética molecular. Por conseguinte, os resultados obtidos para *D.alata* poderão ser usadas para o desenvolvido de marcadores moleculares e com isso expandirão as possibilidades de pesquisa genética para a espécie.

Além da caracterização de regiões microssatélites, este estudo permitiu a identificação de 421.701 TEs em *Dipteryx alata*, o que corresponde a 39,29% das sequências analisadas. De acordo com Lopes e Bureau (2014) o teor de TEs varia consideravelmente entre as espécies. Em *Arabidopsis thaliana* cerca de 15% do genoma é composto por TEs, enquanto em *Zea mays* a porcentagem de TEs excede 85% do genoma (Schnabel et al 2009; De La Chaux et al 2012.). Normalmente, existe uma correlação entre o tamanho do genoma e o conteúdo de TEs, genomas maiores têm uma maior porcentagem destes elementos repetitivos. Entre as espécies da família Leguminosae, por exemplo, *Glycine max* possui 1335 MB de genoma e 58,7% de TEs, *Phaseolus vulgaris* possui 773 MB de genoma e 45,4% de TEs e *Medicago truncatula* possui 479 MB de genoma e 30,5% de TEs (Schmutz et al 2010; Bennett e Leitch 2011; Young et al 2011; Schmutz et 2014). Os TEs são os principais responsáveis pela variação nos tamanhos de genoma entre as espécies (Grover e Wendel, 2010).

Nas sequências do genoma de *D. alata* os retroelementos foram os TEs mais abundantes e representam 47% do total dos TEs identificados. Nas espécies *Glycine max*, *Phaseolus vulgaris* e *Medicago truncatula* os retroelementos também foram os TEs mais numerosos e representam 51,4%, 63,4% e 86,7% do total de TEs anotados, respectivamente (Schmutz et al 2010; Young et al 2011; Schmutz et al 2014). Segundo De La Chaux et al

(2012) os retrotransposons são os TEs mais predominantes na maioria dos genomas de plantas. Em *D. alata* os retroelementos LTR *copia* e *gypsy* foram os mais abundantes e com maiores comprimentos de sequências. Os elementos LTR são assim chamados porque possuem uma longa repetição terminal e são subclassificados em Ty1-*copia* e *gypsy*, superfamílias que diferem uma da outra pelo grau de similaridade das sequências e pelos produtos gênicos codificados (De La Chaux et al 2012). Os elementos Ty1-*copia* são encontrados dispersos por todo o genoma, enquanto os elementos *gypsy* são mais frequentes em regiões heterocromáticas (Cruz et al 2012).

No genoma de *D. alata* os elementos *gypsy* são mais abundante do que os elementos *copia*. Esta relação entre a TEs *gypsy* e *copia* é específico da espécie. Em espécies como *Oryza sativa, Carica papaya, Sorghum bicolor, Helianthus annuus* e *Olea europaea* a quantidade de elementos *gypsy* no genoma também é maior do que a quantidade de elementos *copia* (International Rice Genome Sequencing Projeto 2005; Ming et al 2008; Paterson et al 2009; Cavallini et al 2010; Barghini et al 2014). No genoma de *Vitis vinifera* os elementos *copia* são mais abundantes do que *gypsy* (Jaillon et al 2007). Em *Zea mays* e *Populus trichocarpa* as duas famílias de TEs têm abundâncias semelhantes (Meyers et al 2001; Cossu et al 2012). Entre as Leguminosae, nos genomas de *Glycine max, Phaseolus vulgaris* e *Cicer arietinum* os elementos *gypsy* também são mais abundante do que os elementos *copia*, mas, em contrapartida, no genoma de *Medicago truncatula* os elementos *copia* são mais abundantes que *gypsy* (Schmutz et al 2010; Young et al 2011; Varshney et al 2013; Schmutz et 2014). A alta frequência de *gypsy* e de *copia* em *D. alata* indica que estas superfamílias de TEs foram submetidos à taxa de amplificação elevada durante a evolução do genoma da espécie.

Os elementos LTR geralmente são silenciados durante o desenvolvimento das plantas e a ativação da transcrição, assim como a transposição, são induzidas principalmente por estresses bióticos e abióticos, o que pode ser considerado uma resposta adaptativa do genoma (Xiao et al 2007; Foulonne-Oriol et al 2013; Fan et al 2014). Por exemplo, em células expostas a condições de stress Ty1/Cópia desempenha um importante papel na reorganização de genes e na otimização da expressão génica (Xiao et al 2007). Assim, as diferentes composições de TEs nos genomas refletem as distintas histórias evolutivas das espécies. A identificação de TEs em sequências de *D. alata* contribui com informações sobre a estrutura do genoma e abre a possibilidade de utilizar os TEs como marcadores moleculares em estudos de evolução molecular.

Características específicas, tais como a onipresença e a irreversibilidade fazem dos TEs excelentes fontes para o desenvolvimento de marcadores moleculares. Retroelementos são onipresentes, são encontrados em vários lugares ao mesmo tempo no genoma de organismos eucarióticos e geram polimorfismos irreversíveis (Kalendar et al 2011; Alipour et al 2013; Guo et al 2014). Os polimorfismos surgem porque o mecanismo de transposição dos retrotransposons para novos locais no genoma resulta em alterações de centenas a milhares de bases no local de inserção. Assim, os chamados polimorfismos de inserção são gerados por alteração genômica de grande escala. A este respeito, os marcadores baseados em TEs diferem de outros marcadores, como microssatélites, que são baseados em mudança genômico de pequena escala, envolvendo uma ou algumas dezenas de nucleótidos, e

que geram polimorfismos reversíveis (Kumar et al 2011). TEs ativos possuem sequências conservadas, enquanto os TEs mais antigos podem ter variação nas sequências.

O desenvolvimento de marcadores baseados em retrotransposons é possível devido a existência de subdomínios LTR conservados perto do local de inserção para os quais primers podem ser concebidos (Kalendar et al 2011; Roy et al 2015). Assim, a maioria dos métodos de marcadores baseados em retrotransposons depende da reação em cadeia da polimerase (PCR), que permite a amplificação dos retrotranspons e sequências adjacentes ao local de inserção. Para *D. alata* foram projetados 100 pares de primers visando o futuro desenvolvimento de marcadores RBIP. Do total de retrotransposons usados para projetar os primers, 45% tiveram similaridade com sequências de proteínas armazenadas em bancos de dados, o que permite sugerir que esses transposons estão em contexto gênico e que os mecanismos de transposição podem alterar o número e a organização das sequências codantes no genoma. Nesse sentido, os marcadores baseados em TEs são poderosas ferramentas genéticas para analisar o genoma de plantas.

Marcadores com base em TEs tem aplicações importantes como a avaliação da diversidade genética e determinação da filogenia (Kumar et al 2001). Os TEs são muito informativos nestes estudos porque os mecanismos de transposição geram alterações genotípicas. Além disso, cada TE tem a sua própria história transposicional (tempo e taxa de transposições), de modo que a sequências temporal de eventos de inserção podem ser usadas para determinar filogenias (Kumar el al 2001). Marcadores com base em marcadores TEs fornecem um novo local para análise genética, contribuindo para o já poderoso conjunto de ferramentas genéticas disponíveis para estudar a biologia das plantas (Roy et al 2015; Kumar et al 2001). Os resultados deste estudo podem ser considerado o primeiros passo para a exploração das sequências de TEs em *D. alata* e o futuro desenvolvimento de marcadores baseados em TEs para a espécie.

CONCLUSÃO

Este trabalho representa um grande avanço na identificação de elementos repetitivos no genoma *D. alata* por aplicação de sequenciamento Illumina. A anotação dos microssatélites e TEs geraram conhecimentos sobre a estrutura e a organização do genoma dessa espécie do Cerrado brasileiro. Além disso, os resultados obtidos permitem a exploração das sequências de microssatélites e TEs visando o desenvolvimento de novos marcadores moleculares para a espécie, valiosos em estudos de genética de populações.

MATERIAL SUPLEMENTAR

https://www.dropbox.com/sh/i6u8roglrx5dusw/AACVy6CoglMGYSyDZj8J9NEPa?dl=0

REFERÊNCIAS

Alipour A; Tsuchimoto S; Sakai H; Ohmido N; Fukui K. (2013) Structural characterization of *copia*-type retrotransposons leads to insights into the marker development in a biofuel crop, *Jatropha curcas* L. Biotechnology for Biofuels, 6:129.

Altschul SF; Gish W; Miller W; Myers EW; Lipman DJ. (1990) Basic local alignment search tool. Journal of Molecular Biology, 23(3):403-410.

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Barghini E; Natali L; Cossu RM; Giordani T; Pindo M; Cattonaro F; Scalabrin S; Velasco R; Morgante M; Cavallini A. (2014) The Peculiar Landscape of Repetitive Sequences in the Olive (*Olea europaea* L.) Genome. Genome Biol. 6(4):776–791.

Bennett MD; Leitch IJ. (2011) Nuclear DNA amounts in angiosperms: targets, trends and tomorrow. Annals of Botany, 107:467–590.

Bolger AM; Lohse M; Usadel B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 1.

Cavallini A et al (2010) Analysis of transposons and repeat composition of the sunflower (*Helianthus annuus* L.) genome. Theor Appl Genet, 120:491–508.

Collevatti RG; Telles MPC; Nabout JC; Chaves LJ; Soares TN. (2013) Demographic history and the low genetic diversity in *Dipteryx alata* (Fabaceae) from Brazilian Neotropical savannas. Heredity, 111:97–105

Cossu RM; Buti M; Giordani T; Natali L; Cavallini A. (2012) A computational study of the dynamics of LTR retrotransposons in the Populus trichocarpa genome. Tree Genet Genomes, 8:61–75.

Cruz GMQ; Alves CS; Domingues DS; Metcalfe CJ; Nogueira FTS; Sluys MAV. (2012) Analysis of plant LTR-retrotransposons at the fine-scale family level reveals individual molecular patterns. BMC Genomics, 13.

De La Chaux N; Tsuchimatsu T; Shimizu Kk; Wagner A. (2012) The predominantly selfing plant *Arabidopsis thaliana* experienced a recent reduction in transposable element abundance compared to its outcrossing relative Arabidopsis lyrata. Mob DNA, 3(2).

Diniz-Filho JAF; Melo DB; de Oliveira G; Collevatti, RG; Soares, TN; Nabout JC; Dobrovolski R ; Chaves LJ; Loyola RD; Telles MPC (2012) Planning for optimal conservation of geographical genetic variability within species. Conservation Genetics, 13:1085-1093.

Diniz-Filho JAF; Soares TN; Telles MPC. (2016a) Geographically weighted regression as a generalized Wombling to detect barriers to gene flow. Genetica ('s-Gravenhage), 144: 425-433.

Diniz-Filho JAF; Diniz JVBPL; Telles MPC (2016b) Exhaustive search for conservation networks of populations representing genetic diversity. Genetics and Molecular Research , 15:1-10.

Doyle JJ; Doyle JL. (1987). Isolation of plant DNA from fresh tissue. Focus (San Francisco, Calif.), 12:13-15.

Fan F; Cui B; Zhang T; Ding G; Wen X. (2014) LTR-retrotransposon activation, IRAP marker development and its potential in genetic diversity assessment of masson pine (*Pinus massoniana*). Tree Genetics & Genomes, 10:213–222.

Feschotte C; Jiang N; Wessler SR. (2002) Plant transposable elements: where genetics meets genomics. Nature Reviews Genetics, 3.

Foulongne-Oriol M; Murat M; Castanera R; Ramírez L; Sonnenberg ASM. (2013) Genome-wide survey of repetitive DNA elements in the button mushroom *Agaricus bisporus* Fungal Genetics and Biology, 55:6–21.

Grover CE; Wendel JF. (2010) Recent Insights intoMechanisms of Genome Size Change in Plants. Journal of Botany.

Guo DL; Hou XG; Zhang X. (2014) A Simple and Efficient Method to Isolate LTR Sequences of Plant Retrotransposon, BioMed Research International, 1-8.

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. Nature, 436.

Ištvánek J; Jaroš M; Křenek A; Řepková J. (2014) Genome Assembly and Annotation for Red Clover (*Trifolium Pratense*; Fabaceae). American Journal of Botany, 101(2):327–337.

Jaillon O et al (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature, 449:463–467.

Kalendar R. (2011) The Use of Retrotransposon-Based Molecular Markers to Analyze Genetic Diversity. Genetics and Breeding, 48:261-274.

Kapitonov VK; Jurka J. (2007) Helitrons on a roll: eukaryotic rolling-circle transposons. Trends in Genetics, 23(10).

Kubis S; Schmidtj T; Seymour J; Harrison H. (1998) Repetitive DNA Elements as a Major Component of Plant Genomes. Annals of Botany, 45-55.

Kumar A; Hirochika H. (2001) Applications of retrotransposons as genetic tools in plant biology. TRENDS in Plant Science, 6(3).

Langmead B; Trapnell C; Pop M; Salzberg SL. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology, 10.

Lopes ZJ; Bureau TE. (2014) Diversity and evolution of transposable elements in *Arabidopsis*. Chromosome Res, 22:203-2016.

Mardis ER. (2008) Next-Generation DNA Sequencing Methods. Annu. Rev. Genomics Hum. Genet., 9:387–402.

Matsumoto A; Uchida K; Taguchi Y; Tani N; Tsumura Y. (2010) Genetic diversity and structure of natural fragmented *Chamaecyparis obtusa* populations as revealed by microsatellite markers. Journal of Plant Research, 123:689-699.

Meglécz E; Costedoat C; Dubut V; Gilles A; Malausa T; Pech N; Martin JF. (2010) QDD: a user-friendly program to select microsatellite markers and design primers from large sequencing projects. Bioinformatics Applications Note, 26(3)403-404.

Mehrotra S; Goyal V. (2014) Repetitive Sequences in Plant Nuclear DNA: Types, Distribution, Evolution and Function. Genomics Proteomics Bioinformatics, 12:164–171.

Metzker M. (2010) Sequencing technologies — the next generation. Nature Reviews Genetics, 11.

Meyers BC; Tingey SV; Morgante M. (2001) Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. Genome Res.11:1660–1676.

Miller JR; Koren S; Sutton G. (2010) Assembly algorithms for next-generation sequencing data. Genomics 95:315–327.

Ming R et al (2008) The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). Nature, v:991–997.

Orgel LE; Crick FH. (1980) Selfish DNA: the ultimate parasite. Nature, 284:604–7.

Pandey M; Sharma J. (2012) Efficiency of microsatellite isolation from orchids via next generation sequencing. Open Journal of Genetics, 2:167-172.

Paterson AH et al (2009) The Sorghum bicolor genome and the diversification of grasses. Nature, 457:551–556.

Quinlan A R; Hall IM. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 26(6):841-842.

Roy NS; Choi JY; Lee SI; Kim NS. (2015) Marker utility of transposable elements for plant genetics, breeding, and ecology: a review. Genes Genom, 37:141–151.

Rozen S; Skaletsky H. (2000) Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol, 132:365-86.

Safonova Y; Bankevich A; Pevzner PA. (2015) dipSPAdes: Assembler for Highly Polymorphic Diploid Genomes. Journal Of Computational Biology, 22(6):528-545.

Sano SM; Ribeiro JF; Brito MA. (2004) Baru: biologia e uso. Planaltina, DF: Embrapa Cerrados, 52p.

Schatz MC; Delcher AL; Salzberg SL. (2010) Assembly of large genomes using second-generation sequencing. Genome Research, 20:1165–1173.

Schmutz J et al (2010) Genome sequence of the palaeopolyploid soybean. Nature.

Schmutz J et al (2014) A reference genome for common bean and genome-wide analysis of dual domestications. Nature Gnetics, 46(7).

Schnable PS et al (2009) The B73 maize genome: complexity, diversity, and dynamics. Science, 326:1112–1115.

Silva PIT; Martins AM; Gouvea EG; Pessoa-Filho M; Ferreira ME. (2013) Development and validation of microsatellite markers for *Brachiaria ruziziensis* obtained by partial genome assembly of Illumina single-end reads. BMC Genomics, 14.

Simpson JT; Wong K; Jackman SD; Schein JE; Jones SJM; Birol I. (2009) ABySS: A parallel assembler for short read sequence data. Genome research, 19:1117-1123.

Soares TN; Chaves LJ; Telles MPC; Diniz-Filho JAF; Resende LV. (2008a) Distribuição espacial da variabilidade genética intrapopulacional de *Dipteryx alata*. Pesq. agropec. bras., 43(9):1151-1158.

Soares TN; Chaves LJ; Telles MPC; Diniz-Filho JAF; Resende LV. (2008b) Landscape conservation genetics of *Dipteryx alata* (''baru'' tree: Fabaceae) from Cerrado region of central Brazil. Genetica, 132:9–19.

Soares TN; Melo DB; Resende LV; Vianello RP; Chaves LJ; Collevatti RG; Telles MPC. (2012) Development of microsatellite markers for the neotropical tree species *Dipteryx alata* (Fabaceae). American Journal of Botany.1:2

Soares TN; Diniz-Filho JAF; Nabout JC; Telles MPC; Terribile RC; Chaves LJ. (2015) Patterns of genetic variability in central and peripheral populations of *Dipteryx alata* (Fabaceae) in the Brazilian Cerrado. Plant Syst Evol, 301:1315–1324.

Tarailo-Graovac M; Chen N. (2009) Using RepeatMasker to Identify UNIT 4.10. Repetitive Elements in Genomic Sequences. Current Protocols in Bioinformatics.

Tóth G; Gáspári Z; Jurka J. (2000) Microsatellites in Different Eukaryotic Genomes: Survey and Analysis. Genome Research, 10:967–981.

Treangen TJ; Salzberg SL. (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature Reviews Genetics, 13.

Untergasser A; Cutcutache I; Koressaar T; Ye J; Faircloth BC; Remm MR; Rozen SG. (2012) Primer3—new capabilities and interfaces. Nucleic Acids Researc, 40(15).

Vaidya MB; Sajnani MR; Ramani UV; Tripathi AK; Bhatt VD; Patel JS; Patel MM; Ahir VB; Singh KM; Jakhesara SJ; Koringa PG; Rank DN; Joshi CG. (2012) A preliminary analysis of repetitive sequences organisation in *Bubalus bubalis* genome. Indian Journal os Biotechnology, 11:62-66.

Varshney RK et al (2013) Draft genome sequence of chickpea (Cicer arietinum) provides a resource for trait improvement. Nature biotechnology, 31(3).

Xiao W; Su Y; Sakamoto W; Sodmergen. (2007) Isolation and characterization of Ty1/*copia*-like retrotransposons in mung bean (*Vigna radiata*) Journal of Plant Research, 120:323-328.

Yahya AF; Hyun JO; Lee JH; Kim YY; Lee KN; Hong KN; Kim SC. (2014) Genetic variation and population genetic structure of *Rhizophora apiculata* (Rhizophoraceae) in the greater Sunda Islands, Indonesia using microsatellite markers. Journal of Plant Research, 127:287-297.

Yang T; Bao SY; Ford R; Jia TJ; Guan JP; He YH; Sun XL; Jiang JY; Hao JY; Zhang XY; Zong XX. (2012) High-throughput novel microsatellite marker of faba bean via next generation sequencing. BMC Genomics, 13.

Young ND et al (2011) The Medicago genome provides insight into the evolution of rhizobial symbioses. Nature, 480.

Yu JN; Won C; Jun J; Lim YW; Kwak M. (2011) Fast and Cost-Effective Mining of Microsatellite Markers Using NGS Technology: An Example of a Korean Water Deer Hydropotes inermis argyropus. Plos One, 6.

Wang W; Wu Y; Messing, J. (2012) The Mitochondrial Genome of an Aquatic Plant, *Spirodela polyrhiza*. Plos One, 7(10).

Zalapa JE; Cuevas H; Zhu H; Steffan S; Senalik D; Zeldin E; Mccown B; Harbut R; Simon P. (2012) Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. American Journal of Botany, 99(2):193–208.

Zhou MB; Zhong H; Tang DQ. (2011) Isolation and characterization of seventy-nine full-length mariner-like transposase genes in the Bambusoideae subfamily. Journal of Plant Research, 124:607-617.

Zimin AV; Marçais G; Puiu D; Roberts M; Salzberg SL; Yorke JA. (2013) The MaSuRCA genome assembler. Bioinformatics, 29(21):2669-77.

CAPÍTULO 3

ANOTAÇÃO ESTRUTURAL E FUNCIONAL DE GENES NO GENOMA NUCLEAR DE *Dipteryx alata* (LEGUMINOSAE)

RESUMO

A anotação de genes é um dos primeiros e mais importantes passos para a exploração racional dos dados genômicos obtidos através do sequenciamento de DNA. Esta pesquisa teve como objetivo realizar a anotação gênica estrutural e funcional em amostras de sequências do genoma nuclear de *Dipteryx alata*, espécie arbórea nativa do bioma Cerrado Brasileiro. Para tanto foram extraídas amostras de DNA de folhas de quatro indivíduos coletados em diferentes localidades do Cerrado. Bibliotecas genômicas foram preparadas de acordo com o protocolo Nextera e sequenciadas na plataforma Illumina MiSeq. Reads nucleares foram montados com software Dipspades. A anotação de genes de RNAs não traduzidos em proteínas foi realizada com os softwares RNAmmer, tRNAscan-SE e Blastn usando o banco de dados PNRD (Plant Non-coding RNA Database). A anotação estrutural de genes codificantes de proteínas foi realizada com o software Augustus e anotação funcional com o software Blast2GO. Os reads obtidos para *D. alata* foram montados em 275.709 scaffolds com N50 igual a 1.598. Nessas sequências foram anotados 1.431 genes de ncRNAs, sendo 176 rRNAs, 189 tRNAs, 477 snRNAs, 8 snoRNAs, 466 miRNAs e 115 lncRNAs. Foram anotados também 62.200 genes codificados de proteínas com tamanho médio de 1.156 pb. Os genes de *D. alata* apresentam em média 3,03 éxons e 2,90 íntrons com tamanho médio de 246 e 255 pb, respectivamente. O número estimado de mRNAs transcritos pelo conjunto de genes anotados é igual a 160.450, dos quais 131.228 apresentaram similaridade significativa com sequências já conhecidas e 84.793 foram classificadas funcionalmente, nos termos do Gene Ontology. Genes envolvidos em inúmeras vias bioquímicas e processos biológicos foram anotados nas sequências de *D. alata*. Parte dos genes nucleares identificados não tiveram suas funções atribuídas. De acordo com a literatura, a ausência de sequências homólogas nas bases de dados públicas pode indicar que esses genes não dispõem de domínios funcionais conservados. Os dados genômicos obtidos para *D. alata* permitiram conhecer parte do repertório gênico da espécie e representa uma adição valiosa as informações disponíveis em bancos de dados.

Palavras-chaves: baru, genes, Illumina, MiSeq.

INTRODUÇÃO

O rápido aumento na disponibilidade e qualidade dos dados de sequenciamento tem propiciado muitos avanços nos estudos das espécies de plantas (Amar et al 2014). As plataformas de sequenciamento de nova geração (*Next Generation Sequencing* - NGS) permitiram gerar um grande volume de dados de sequências de nucleotídeo, com baixas taxas de erro, em um curto espaço de tempo e com custo reduzido, quando comparado ao método de sequenciamento clássico (Mardis 2008; Metzker 2010; Lohse et al 2014). Assim, as tecnologias NGS provocaram uma explosão de projetos de sequenciamento dos genomas de plantas, muitos dos quais ainda estão em andamento (Mittler et al 2013; Lohse et al 2014). Milhares de sequências genômicas estão sendo publicadas em banco de dados e estão permitindo conhecer a composição e a ordem dos nucleotídeos nos genomas das espécies. No entanto, a obtenção das sequências de DNA não fornece muitas informações biológicas de modo imediato e representa apenas o primeiro passo para a compreensão do genoma. A elucidação dos mecanismos envolvidos no funcionamento do organismo depende da precisão dos processos de anotação (Zieleziński et al 2012; Lohse et al 2014). A anotação das sequências de DNA permite investigar as diversas arquiteturas dos genomas de plantas (Caicedo e Purugganan 2005; Amar et al 2014).

O processo de anotação de sequências gênicas é um dos primeiros e mais importantes passos para o entendimento dos genomas (Stein 2001). São encontrados na literatura dados de anotação gênica estrutural e funcional para várias espécies de plantas, incluindo espécies da família Leguminosae, a terceira maior família de plantas com flores. Estudos de regiões genômicas codantes foram conduzidos, por exemplo, para espécies como *Glycine max*, *Phaseolus vulgaris*, *Cicer arietinum* e *Medicago truncatula* (Schmutz et al 2010; Varshney et al 2013; Schmutz et al 2014; Tang et al 2014). Muitas dessas espécies possuem potencial econômico, devido principalmente ao valor nutricional, e compõem parte importante da agricultura mundial (Schmutz et al 2010). Quantidade elevada de sequências vem sendo depositadas em banco de dados para *Glycine max*, espécie com alto valor econômico que tem sido alvo de extensa pesquisa genômica. *G. max* é caracterizada como espécie modelo para o estudo de Leguminosae (Young et al 2012). Nesse contexto, o aumento na quantidade de anotações gênicas depositadas em banco de dados facilita investigações genômicas de espécies não modelo por meio de análises de similaridade.

A anotação dos genes que constituem o genoma de uma espécie pode ser feita pelo método de previsão *ab initio* e/ou pelo método de similaridade. Os preditores *ab initio* usam características intrínsecas do genoma para discriminar regiões codantes e não codantes, como códons de início e parada, promotores, sinais de splicing e limites de junção de exons (Zieleziński et al 2012). Os métodos baseados em similaridade anotam genes por meio da semelhança significativa com sequências gênicas já conhecidas e armazenadas em bancos de dados. Alguns projetos de pesquisa têm usado o método de previsão *ab initio* para análise estrutural e o método baseado em similaridade para análise funcional das regiões gênicas (Ištvánek et al 2014). A anotação funcional, usando similaridade de sequências, é feita com base no Gene Ontology (GO), uma iniciativa de bioinformática que visa relacionar genes e produtos de genes para diversas espécies (Conesa et al 2005; Gotz et al 2008). A análise

funcional GO representa as sequências gênicas em três categorias: componentes celulares, função molecular e processos biológicos (Amar et al 2014). A categoria "componentes celulares" permite localizar o produto do gene dentro da célula, em compartimentos subcelulares, ou no meio extracelular. A categoria função molecular descreve as tarefas realizadas pelos produtos de genes individuais como, por exemplo, a sua atividade enzimática ou função estrutural A categoria processo biológico descreve processos mais amplos desempenhados pelos produtos do gene, como o ciclo celular. Assim, as categorias GO permitem descrever aspectos da identidade biológica dos produtos gênicos (Berardini et al 2004). Segundo Amar et al (2014) anotar funcionalmente um gene consiste em colocá-lo dentro de uma ou mais categorias do Gene Ontology.

Os avanços nas metodologias de predição de genes e o aumento do número de sequências conhecidas nos bancos de dados contribuem para a caracterização do genoma de organismos não modelos e permite estudar o genoma de várias espécies, incluindo espécies de plantas nativas do Cerrado, como *Dipteryx alata*. Essa é uma espécie de árvore nativa do Cerrado brasileiro que pertence à família Leguminosae e que possui ampla distribuição geográfica nesse bioma. O potencial econômico da espécie deve-se a poupa e a semente de seus frutos, que são utilizados in natura ou na fabricação de sorvetes, licores, castanha torrada, e outros produtos por pequenas indústrias. *D. alata* possui também utilidade forrageira, medicinal, na recuperação de áreas degradadas, no paisagismo e na extração de madeira (Sano et al 2004). Dessa forma, *D. alata* pode ser considerado um valioso recurso genético para o Cerrado. Nesse cenário, existe a necessidade, para fins de uso e de conservação, de se obter novos conhecimentos sobre a biologia molecular de *Dipteryx alata*, uma vez que a espécie está ameaçada devido à fragmentação do habitat e alteração da paisagem natural do Cerrado, causadas pela expansão da agricultura (Soares et al 2012).

Esta pesquisa teve como objetivo realizar a anotação estrutural e funcional de genes em uma amostra de sequências do genoma nuclear de *Dipteryx alata* usando dados de sequenciamento Illumina. Este pode ser considerado um dos primeiros trabalhos genômicos para espécies arbóreas nativas do Cerrado Brasileiro.

MATERIAL E MÉTODOS

Extração de DNA e sequenciamento Illumina

Conforme descrito no capítulo 2, foi realizada a extração de DNA, usando o método CTAB (brometo de cetiltrimetilamônio) descrito por Doyle e Doyle (1987), das folhas de quatro indivíduos de *Dipteryx alata* coletados em diferentes localidades de Cerrado. As amostras foram preparadas para o sequenciamento de DNA seguindo o protocolo Nextera da Illumina. O sequenciamento foi realizado na plataforma MiSeq da Illumina usando o kit de sequenciamento Miseq v3 de 600 ciclos (2x250bp, reads paired-end).

Análise de qualidade e montagem das sequências genômicas nucleares

A qualidade dos reads foi analisada usando o programa fastQC e sequências com baixa qualidade foram removidas usando o programa Trimmomatic (Andrews, 2010; Bolger et al 2014). A montagem foi realizada usando o software Dipspades 1.0 (Safonova et al 2015). Sequências organelares e menores que 500pb foram excluídas da montagem. A qualidade da montagem foi avaliada usando o script em perl desenvolvido por Keith Bradnam (2011). A cobertura da montagem foi estimada através do alinhamento dos reads contra o fasta da montagem usando o software bowtie2 (Langmead et al 2009) e a ferramenta genomeCoverageBed. Mais detalhes no capítulo 2.

Anotação de genes de RNAs não traduzidos

Genes codificantes de RNAs ribossomais (rRNAs) foram preditos usando o programa RNAmmer (Lagesen et al 2007). RNAmmer é um programa com elevada precisão para anotar rRNA e usa modelos ocultos de Markov treinados em base de dados de RNAs ribossômicos. Genes de RNAs transportadores (tRNAs) foram preditos usando o programa tRNAscan-SE, que usa um modelo altamente seletivo de covariância (Lowe e Eddy, 1997). Genes de outros RNAs não codificantes de proteínas (ncRNAs), tais como pequenos RNA nucleares (snRNAs), pequenos RNA nucleolares (snoRNAs), microRNAs (miRNAs) e longos RNAs não codificante (lncRNAs), foram preditos usando o Blastn (Basic Local Alignment Search Tool) e o banco de dados PNRD (Plant Non-coding RNA Database). O banco de dados PNRD, que é acessível em http://structuralbiology.cau.edu.cn/PNRD/, disponibiliza 25739 sequências de ncRNAs, pertencente a diferentes famílias de RNAs e armazenadas a partir de 150 espécies de plantas (Yi et al 2015). Foram anotados sequência de *D. alata* com similaridade significativa a ncRNAs armazenados no PNRD usando Blastn com E-valor de 10^{-6}.

Anotação de genes codificantes de proteínas

As sequências montadas para *D. alata* foram analisadas com o software preditor de genes Augustus, que fez a previsão *ab initio* de genes completos e parciais codificantes de proteínas usando como matriz de treinamento o genoma de *Arabidopsis* (Stanke et al 2004). A partir dos resultados do Augustus foi obtido um arquivo com as sequências de aminoácidos codificadas por todos os transcritos preditos. O software BLASTp foi usado para comparar as sequências de aminoácidos contra o banco de dados de proteínas de embriófitas (Altschul et al 1990). Os melhores resultados de similaridade foram selecionados, com base no maior valor de score com um E-valor menor que 10^{-6}, e analisados no software Blast2GO v3.0 para determinação e anotação da função dos genes de *D. alata* (Conesa et al 2005). Com o Blast2GO as sequências foram anotadas nas categorias do Gene Ontology (http://www.geneontology.org) componentes celulares, função molecular e processos biológicos. Os dados foram representados usando o nível 2 do Blast2GO, que ilustra categorias funcionais gerais.

RESULTADOS

Sequenciamento e montagem das sequências genômicas

Foram obtidos com o sequenciamento Illumina 22 GB de dados que corresponderam a 61,62 milhões de reads paired-end. Após os cortes de qualidade foram mantidos 60,55 milhões de reads (14,6 GB). Os reads foram montados em 275.709 scaffolds com N50 igual a 1.598. O tamanho do maior e do menor scaffold, bem como o tamanho total dos scaffolds, foram de 87.189, 500 e 363.927.807 pb, respectivamente. A montagem possui 355 MB de sequências do genoma nuclear e cobertura de 12,78X (Capítulo 2).

Anotação de genes codificantes de RNAs não traduzidos

Foram identificados, nas sequências montadas para *D. alata*, 1.431 genes codificantes de RNAs não traduzidos (Tabela 1). Pequenos RNAs nucleares e microRNAs foram os RNAs não traduzidos anotados em maior quantidade. Foram identificadas também sequências de RNAs ribossômicos, RNAs transportadores, pequenos RNAs nucleolares e longos RNAs não codificantes.

Tabela 1 - Anotação estrutural de genes codificantes de RNAs não traduzidos em *D. alata.*

Tipo de RNA	**Número de RNAs**	**Tamanho (pb)**		
		Médio	**Mínimo**	**Máximo**
RNA ribossômico (rRNA)	8S - 174 28S - 2	8S - 113 28S - 3278	8S - 94 28S - 2851	8S - 154 28S - 3706
RNA Transportador (tRNA)	189	74	56	95
Pequeno RNA nuclear (snRNA)	477	108	44	203
Pequeno RNA nucleolar (snoRNA)	8	103	54	127
microRNA (miRNA)	466	112	45	210
Longo RNA não codificante (lncRNA)	115	725	200	2971

Anotação estrutural de genes codificantes de proteínas

Das sequências montadas para *D. alata* 21,3% (58.799 scaffolds) possuem genes putativos codificantes de proteínas. Um total de 62.200 genes completos e parciais foi predito, compostos por 486.242 éxons e 406.544 íntrons (Material suplementar online Figura1A-H). O número máximo de genes por sequência foi 19 e o tamanho médio dos genes foi 1156 pb. O número médio de éxons por gene foi 3.03 e o tamanho médio dos éxons foi 246 pb. O número médio de íntrons por gene foi 2.90 e o tamanho médio dos íntrons foi 255 pb. O número estimado de mRNAs transcritos pelo conjunto de genes anotados foi igual a 160.450. O

número médio de transcritos por gene foi 2.59 e o tamanho médio dos transcritos foi 1425pb (Figura 2).

Tabela 2 – Anotação estrutural de genes codificantes de proteínas em *D. alata*.

Parâmetros	Valores estimados			Frequências relativas
	Médio	Mínimo	Máximo	
Número de genes por scaffold	1	1	19	94% dos scaffols possuem 1 gene 5% dos scaffols possuem 2 genes 1% dos scaffols possui mais de 2 genes
Tamanho dos genes (pb)	1.156	50	19.990	54% dos genes tem até 1000pb 34% dos genes tem entre 1000 e 2000pb 12% dos genes tem mais de 2000pb
Número de éxons por gene	3,03	1	45	69% dos genes tem até 4 éxons 29% dos genes tem entre 4 e 10 éxons 2% dos genes tem mais de 10 éxons
Tamanho dos éxons (pb)	246	2	4.840	60% dos éxons tem até 200pb 38% dos éxons tem entre 200 e 1000pb 2% dos éxons tem mais de 1000pb
Número de íntrons por gene	2,90	1	44	71% dos genes tem até 4 íntrons 27% dos genes tem entre 4 e 10 íntrons 2% dos genes tem mais de 10 íntrons
Tamanho dos íntrons (pb)	255	5	4.842	60% dos éxons tem até 200pb 37% dos éxons tem entre 200 e 1000pb 3% dos éxons tem mais de 1000pb
Número de transcritos por gene	2,58	1	43	40% dos genes possuem 1 transcrito 29% dos genes possuem 2 transcritos 31% dos genes possui mais de 2 transcritos
Tamanho dos transcritos (pb)	1.425	50	19.249	41% dos transcritos tem até 1000pb 37% dos transcritos tem entre 1000 e 2000pb 22% dos genes tem mais de 2000pb

Anotação funcional de genes codificantes de proteínas

A comparação das sequências de aminoácidos codificadas pelos 160.450 transcritos de *D. alata* com base de dados de proteínas permitiu identificar 131.228 sequências que tiveram similaridade significativa com sequências já conhecidas. As espécies que apareceram com maior frequência entre os Top-Hits do Blast foram *Glycine max* (Soja), *Phaseolus vulgaris* (Feijão), *Cicer arietinum* (Grão-de-bico) e *Medicago truncatula* (Figura 1). Entre as sequências identificadas por meio do blast, 84.793 foram classificadas funcionalmente nos termos do Gene Ontology, enquanto 46.435 tiveram similaridade com sequências cujas funções ainda não estão descritas.

O maior número de anotações foi obtida para a categoria processos biológicos (221.451 anotações), seguido das categorias componentes celulares (156.743 anotações) e função molecular (104.829 anotações) (Figura 2). Entre os termos GO para processos biológicos, os mais frequentes foram processos metabólicos (24% dos genes) e processos celulares (21% dos genes). Nessa categoria destacam-se processos de oxidação-redução, fosforilação de proteínas, regulação da transcrição, processos catabólicos de ATP e proteólise. Na categoria função molecular, a maioria dos genes preditos estavam envolvidos com atividades catalíticas (47%) e ligação (41%). As funções moleculares mais abundantes foram ligação a ATP, ligação a íon zinco, ligação a íon metal, atividade transferase e atividade hidrolase. Na categoria de componentes celulares, 21% dos produtos de genes foram anotados com localização celular, 15% com localização organelar e 12% como pertencentes à membrana.

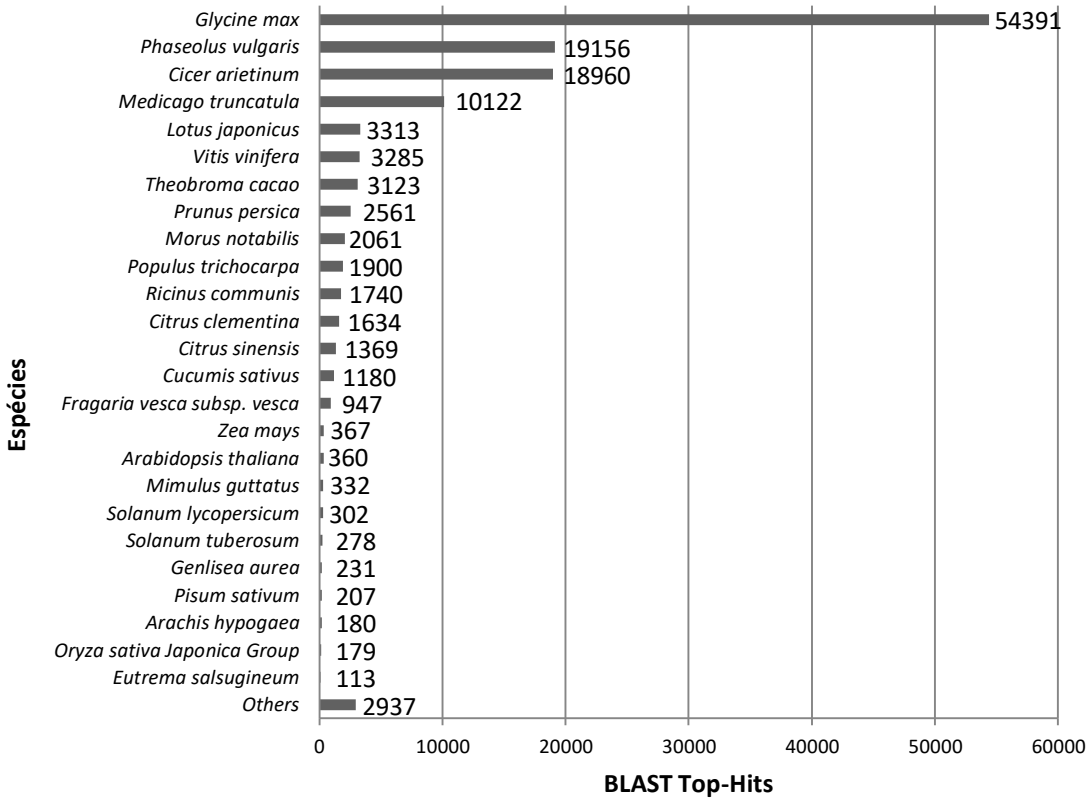

Figura 1 – Distribuição das espécies Top-Hits do Blast.

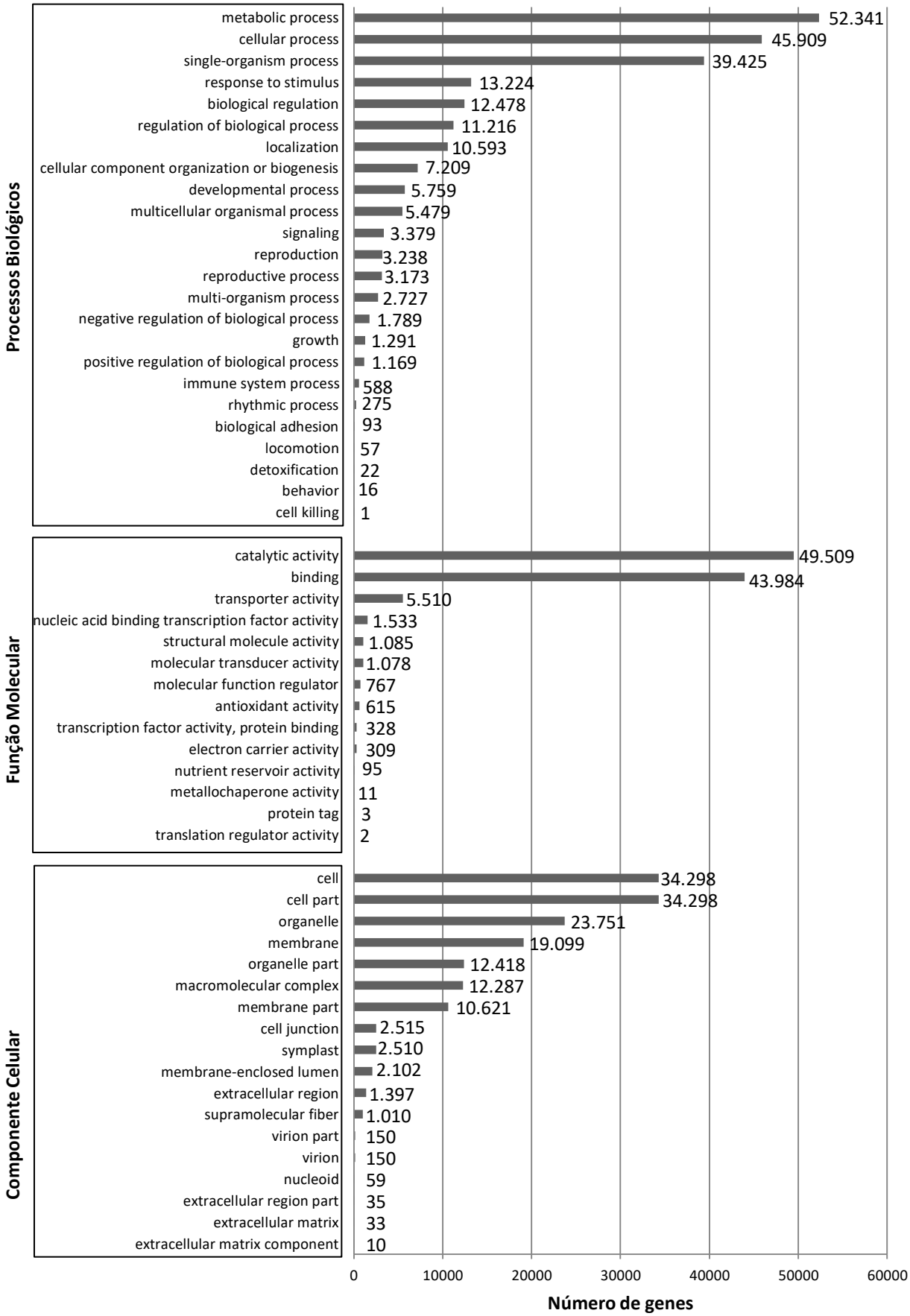

Figura 2 – Anotação funcional de sequências gênicas nucleares codificantes de proteínas de *Dipteryx alata.*

DISCUSSÃO

A quantidade de dados gerada no sequenciamento de DNA de *Dipteryx alata* usando o plataforma MiSeq Illumina foi suficiente para montar uma amostra de sequências genômicas nucleares (355 MB) para a espécie (Capítulo 2). As sequências de DNA montadas representaram 44% do genoma de *D. alata*, visto que o tamanho do genoma desta espécie foi estimado em 807,2 MB usando citometria de fluxo (Capítulo 1). Essa representação parcial do genoma permitiu a abrangente caracterização gênica estrutural e funcional e gerou informações genômica até então não disponível para esta espécie do Cerrado brasileiro.

Nas sequências nucleares de *D. alata* foram preditos 1.431 genes codificantes de RNAs não traduzidos (ncRNAs). Com exceção dos rRNAs e tRNAs, os ncRNAs foram identificados por pesquisas de similaridade de sequência, metodologia eficiente devido ao alto nível de conservação das sequências nas regiões centrais dos genes. A anotação baseada em similaridade depende do acesso a sequências genômicas conhecidas, o que não é um problema visto o crescimento exponencial dos bancos de dados nas duas últimas décadas (Lagesen et al 2007). Os rRNA e tRNAs foram anotados com programas que fazem uso de modelos específicos para essas famílias de RNAs, capazes de gerar resultados ainda mais precisos que o Blast. Os resultados da anotação de ncRNAs em *D. alata* são semelhantes ao encontrado na espécies *Trifolium pratense*, também pertencente a família Leguminosae. O tamanho médio dos tRNAs, snRNAs e miRNAs anotados em *T. pratense* foram 73, 105,9 e 130,6 bases, respectivamente (Ištvánek et al 2014). Os ncRNAs desempenham funções importantes na tradução e na regulação da expressão gênica e investigações desses genes, assim como a anotação dos genes codificadores de proteínas, são importantes para a exploração global do genoma e a compreensão da biologia molecular fundamental das espécies.

Nas sequências de *D. alata* foram anotados 62.200 genes putativos codificadores de proteínas, e a maior parte deles são homólogos a genes de *Glycine max*, *Phaseolus vulgaris*, *Cicer arietinum* e *Medicago truncatula*. Essas espécies são representantes da família Leguminosae, e, portanto são filogeneticamente próximas de *D. alata*. Espécies pertencentes a grupos taxonômicos próximos geralmente apresentam similaridade genômica em grande escala, com sequências gênicas conservadas e alta sintenia (Young et al 2012). Nesse sentido, era esperado a alta similaridade das sequências de *D. alata* com genes de espécies da mesma família que possuem genomas parcial ou totalmente estudados. A maior similaridade das regiões codificantes de *D. alata* com genes de *G. max* pode ser explicada pelo grande volume de sequências depositadas em bancos de dados para essa espécie (Tang et al 2014). Segundo Young et al (2012), *G. max* é alvo de extensa pesquisa genômica e fornece a base para o conhecimento atual sobre o genoma de leguminosas.

A quantidade de genes nucleares identificados para *D. alata* é alta em comparação com o conjunto de genes anotados para outras espécies da família Leguminosae. O genoma de *Phaseolus vulgaris* foi montado em 473 MB e possui 27.197 genes (Schmutz et al 2014). O genoma de *Cicer arietinum* foi montado em 738 MB e possui 28.269 genes (Varshney et al 2013). O genoma de *Glycine max* foi montado em 950 MB e possui 46.430 genes (Schmutz et al 2010). O genoma de *Medicago truncatula* foi montado em 360 MB e possui 50.894 genes (Tang et al 2014). O número diferente de genes entre as espécies da mesma família resulta de eventos de perdas ou ganhos (duplicações) de genes que ocorreram ao longo da evolução

(Young et al 2012). Dessa forma, a variação no número de genes representa a diversidade genética resultante da evolução de famílias gênicas (Denton et al 2014). Porém, em alguns casos, o número elevado de genes pode ser resultado da predição incorreta a partir de montagens parciais do genoma. A super estimativa do número de genes ocorre devido à fragmentação de um único gene em múltiplos scaffolds e da predição *ab initio* de um gene único como vários genes separados (Denton et al 2014).

Como o número de genes identificados (62.200) em 44% do genoma de *D. alata* é superior ao encontrado em genomas completos de outras espécies de Leguminosae é possível inferir que a anotação gênica esteja super estimada. Resultados semelhantes foram encontrados para a espécie *Trifolium pratense* em que foram preditos 64.761 genes em 314,6 MB de sequências genômicas (Ištvánek et al 2014). As montagens de *D. alata* e de *Trifolium pratense* possuem números altos de sequências, 275.709 e 176.760 scaffolds respectivamente, o que corrobora a hipótese de super estimativa do número de genes por fragmentação gênica. A anotação de pseudogenes também pode contribuir para a super estimativa do número de genes nos genomas. Futuros esforços de sequenciamento, com a geração de mais dados de sequências, incluindo o sequenciamento de fragmentos maiores como os de Pacific BioSciences, poderão aumentar a contiguidade das sequências e melhorar a qualidade da montagem e das anotações para *D. alata* (Schatz et al 2012; Denton et al 2014). Dados adicionais como, por exemplo, de RNA-seq também poderão contribuir para a identificação do número correto dos genes de *D. alata* (Denton et al 2014).

Apesar da super estimativa no número de genes, é possível inferir que as sequências gênicas de *D. alata* foram preditas corretamente, uma vez que 82% das sequências de aminoácidos codificadas pelos transcritos tiveram similaridade significativa com sequências conhecidas armazenadas em bancos de dados. Além disso, o tamanho médio de éxons (246 pb) e íntrons (255 pb) de *D. alata* foram semelhantes aos anotados para outras espécies de Leguminosae. Os genes de *Phaseolus vulgaris* possuem em média 5,5 éxons, os éxons possuem em média 160 pb e os íntrons 200 pb (Schmutz et al 2014). Os genes de *Cicer arietinum* possuem em média 4,93 éxons, os éxons possuem em média 236 pb e os íntrons 480pb (Varshney et al 2013). O número médio de éxons por gene (3,03) em *D. alata* foi menor que o de *Phaseolus vulgaris* e *Cicer arietinum*, que pode ser justificado pela fragmentação dos genes. Outra evidência da fragmentação é o tamanho médio dos genes, em *D. alata* o tamanho médio é 1.156pb, valor inferior ao da espécie *Medicago truncatula*, que possui genes com tamanho médio de 3.280pb (Tang et al 2014).

Do total de sequências codificadoras de proteínas identificadas para *D. alata*, 53% foram anotados em categorias funcionais. O estudo funcional das sequências gênicas é importante para descrever a identidade biológica de um produto de gene (Berardini et al 2004). Genes envolvidos em inúmeras vias bioquímicas e processos biológicos foram anotados nas sequências nucleares de *D. alata*. As sequências identificadas como genes, mas sem anotação funcional podem representar regiões gênicas que provavelmente não dispõem dos domínios funcionais conservados (Hou et al 2011; Liu et al 2012). Nesse sentido, a ausência de sequências homólogas nas bases de dados públicas pode indicar papéis específicos para esses genes (Shi et al 2011). De acordo com a literatura existem genes que são comuns a todas as espécies de plantas e genes que são específicos para um subconjunto de

espécies. Esses genes específicos podem desempenhar papeis na adaptação das espécies a nichos ecológicos particulares (Delsenya et al 2010).

Nesse contexto, um exemplo é a análise comparativa realizada entre o repertório gênico das espécies monocotiledôneas *Sorghum bicolor* e *Oryza sativa* e as espécie dicotiledôneas *Arabidopsis thaliana* e *Populus trichocarpa* (Delsenya et al 2010). O genoma de *S. bicolor* foi montado em 730 MB de sequências, no qual foram anotados um total de 34.496 genes. Do conjunto de famílias gênicas do genoma de *S. bicolor* cerca de 58% (9.503) foram compartilhadas entre todas as quatro espécies estudadas, 24% (3.983) foram específicas para as gramíneas (*S. bicolor e O. sativa*) e 7% (1.153) foram únicas para *S. bicolor*. Esse resultado mostra que um catálogo específico de genes foi desenvolvidos pelas gramíneas ao longo da evolução, provavelmente devido a características particulares das espécies desse grupo. A diferença na composição gênica é responsável pela variação morfológica e fisiológica entre as espécies de plantas e contribuem para a adaptação a ambientes distintos. No estudo supracitado os genes identificados exclusivamente em *S. bicolor* são duplicações recentes de genes de microRNA que contribuem para a tolerância de *S. bicolor* à seca (Paterson et al 2009; Delsenya et al 2010). Considerando as investigações supracitadas realizadas por Paterson et al (2009) e os resultados do presente estudo é possível sugerir que as espécies nativas do Cerrado ou mesmo *D. alata* possua conjuntos específicos de genes em sua arquitetura genômica.

O conjunto de dados gerados para *D. alata* nesse estudo e disponibilizado em bancos de dados públicos é uma adição valiosa a informação genômica dessa espécie do Cerrado. A quantidade de informações biológicas obtidas nesse estudo confirma que o uso de tecnologias de sequenciamento de nova geração oferece grande oportunidade para o rápido acesso a informação genética em plantas não modelo (Gordo et al 2012). Futuros esforços de sequenciamento poderão aumentar ainda mais o conhecimento de sequências gênicas de *D. alata* e permitir inúmeras outras investigações, tais como a elucidação de redes de regulação de genes e vias metabólicas, eventos de duplicação de todo o genoma e análise de sintenia baseada na comparação com o genoma de outras espécies da família Leguminosae que já possuem genoma sequenciado. O aumento na quantidade de genomas sequenciados e anotados tem gerado uma série de oportunidades para a genômica comparativa. Nesse sentido, esse estudo inicial abre caminho para novas pesquisas genômicas de *D. alata*.

CONCLUSÃO

Os dados genômicos obtidos neste estudo permitiram conhecer parte do repertório gênico de *D. alata* e representou uma contribuição substancial para as informações até então disponíveis para essa espécie. Os resultados geraram conhecimentos sobre a estrutura e a organização do genoma de *D. alata* e ainda abriram caminho para futuras investigações, como a análise de sintenia baseada na comparação com o genoma de outras espécies da família Leguminosae tais como *Glycine max*, *Phaseolus vulgaris*, *Cicer arietinum* e *Medicago truncatula*.

MATERIAL SUPLEMENTAR

https://www.dropbox.com/sh/i6u8roglrx5dusw/AACVy6CoglMGYSyDZj8J9NEPa?dl=0

REFERÊNCIAS

Altschul SF; Gish W; Miller W; Myers EW; Lipman DJ. (1990) Basic Local Alignment Seash Tool. L. Mol. Biol, 215:403-410.

Amar D; Frades I; Danek A; Goldberg T; Sharma SK; Hedley PE; Proux-Wera E; Andreasson E; Shamir R; Tzfadia O; Alexandersson E. (2014) Evaluation and integration of functional annotation pipelines for newly sequenced organisms: the potato genome as a test case. BMC Plant Biology, 14:329.

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Berardini TZ; Mundodi Z; Reiser L; Huala E; Garcia-Hernandez M; Zhang P; Mueller LA; Yoon J; Doyle A; Lander G; Moseyko N; Yoo D; Xu I; Zoeckler B; Montoya M; Miller N; Weems D; Rhee SY. (2004) Functional Annotation of the *Arabidopsis* Genome Using Controlled Vocabularies. Plant Physiology, 135:745–755.

Bolger AM; Lohse M; Usadel B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 1.

Caicedo AL; Purugganan MD. (2005) Comparative Plant Genomics. Frontiers and Prospects. Plant Physiology, 138:545–547.

Conesa A; Götz S; García-Gómez JM; Terol J; Talón M; Robles M. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics Applications Note, 21(18).

Conesa A; Gotz S. Blast2GO: A Comprehensive Suite for Functional Analysis in Plant Genomics. International Journal of Plant Genomics, 2008.

Delsenya M; Hanb B; Hsingc YI. (2010) High throughput DNA sequencing: The new sequencing revolution. Plant Science, 179:407–422.

Denton D; Lugo-Martinez J; Tucker AE; Schrider DR; Warren WC; Hahn MW. (2014) Extensive Error in the Number of Genes Inferred from Draft Genome Assemblies. PLOS Computational Biology, 10.

Doyle JJ; Doyle JL. (1987) Isolation of plant DNA from fresh tissue. Focus (San Francisco, Calif.), 12:13-15.

Gordo SMC; Pinheiro DG; Moreira ECO; Rodrigues SM; Poltronieri MC; Lemos OF; Silva IT; Ramos RTJ; Silva A; Schneider H; Jr WAS; Sampaio I; Darnet S. (2012) High-throughput sequencing of black pepper root transcriptome. BMC Plant Biology, 12:168.

Gotz S; Garcıa-Gomez JM; Terol J; Williams TD; Nagaraj SH; Nueda MJ; Robles M; Talon M; Dopazo D; Conesa A. (2008) High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Research, 36(10):3420–3435.

Hou R; Bao Z; Wang S; Su H; Li Y; Du H; Hu J; Wang S; Hu X. (2011) Transcriptome Sequencing and De Novo Analysis for Yesso Scallop (*Patinopecten yessoensis*) Using 454 GS FLX. PLoS ONE, 6.

Ištvánek J; Jaroš M; Křenek A; Řepková J. (2014) Genome Assembly and Annotation for Red Clover (*Trifolium Pratense*; Fabaceae). American Journal of Botany, 101(2):327–337.

Lagesen K; Hallin P; Rødland EA; Stærfeldt HH; Rognes T; Ussery DW. (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Research, 35(9).

Langmead B; Trapnell C; Pop M; Salzberg SL.(2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology, 10.

Liu M.; Qiao G; Jiang J; Yang H; Xie L; Xie J; Zhuo R. (2012) Transcriptome Sequencing and De Novo Analysis for Ma Bamboo (*Dendrocalamus latiflorus* Munro) Using the Illumina Platform. PLoS ONE, 7(10).

Lohse M; Nagel A; Herter T; May P; Schroda M; Zrenner R; Tohge T; Fernie AR; Stitt M; Usadel B. (2014) Mercator: a fast and simple web server for genome scale functional annotation of plant sequence data. Plant, Cell and Environment, 37:1250–1258.

Lowe TM; Eddy SR. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Research, 25(5).

Mardis ER. (2008) Next-Generation DNA Sequencing Methods. Annu. Rev. Genomics Hum. Genet., 9:387–402.

Mathé C; Sagot MF; Schiex T; Rouzé P. (2002) Current methods of gene prediction, their strengths and weaknesses. Nucl. Acids Res. 30(19):4103-4117.

Metzker M. Sequencing technologies — the next generation. (2010) Nature Reviews Genetics, 11.

Mittler R; Shulaev V. (2013) Functional genomics, challenges and perspectives for the future. Physiologia Plantarum, 148:317–321.

Paterson et al (2009) The *Sorghum bicolor* genome and the diversification of grasses. Nature, 457.

Safonova Y; Bankevich A; Pevzner PA. (2015) dipSPAdes: Assembler for Highly Polymorphic Diploid Genomes. Journal Of Computational Biology, 22(6):528-545.

Sano SM; Ribeiro JF; Brito MA. (2004) Baru: biologia e uso. Planaltina, DF: Embrapa Cerrados, 52p.

Schatz MC; Witkowski J; McCombie WR. (2012) Current challenges in de novo plant genome sequencing and assembly. Genome Biology, 13.

Schmutz J et al (2010) Genome sequence of the palaeopolyploid soybean. Nature, 463:178-183.

Schmutz J et al (2014) A reference genome for common bean and genome-wide analysis of dual domestications. Nature Gnetics, 46(7).

Shi CY; Yang H; Wei1 CL; Yu O; Zhang ZZ; Jiang CJ; Sun J; Li YY; Chen Q; Xia T; Wan XC. (2011) Deep sequencing of the *Camellia sinensis* transcriptome revealed candidate genes for major metabolic pathways of tea-specific compounds. BMC Genomics, 12:131.

Soares TN; Melo DB; Resende LV; Vianello RP; Chaves LJ; Collevatti RG; Telles MPC. (2012) Development of microsatellite markers for the neotropical tree species *Dipteryx alata* (Fabaceae). American Journal of Botany.

Stanke M; Morgenstern B. (2005) AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Research, 33.

Stein L. (2001) Genome Annotation: From Sequence to Biology. Nature Reviews Genetics, 2.

Tang H; Krishnakumar V; Bidwell S; Rosen B; Chan A; Zhou S; Gentzbittel L; Childs KL; Yandell M; Gundlach H; Mayer KFX; Schwartz DC; Town CD. (2014) An improved genome release (version Mt4.0) for the model legume *Medicago truncatula*. BMC Genomics, 15:312.

Varshney RK. (2013) Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. Nature Biotechnology, 31:240–246.

Yi X; Zhang Z; Ling Y; Xu W; Su Z. (2015) PNRD: a plant non-coding RNA database. Nucleic Acids Research, 43.

Young ND; Bharti AK. Genome-Enabled Insights into Legume Biology. (2012) Annu. Rev. Plant Biol, 63:283–305.

Zieleziński A; Potarzycki P; Książkiewicz M; Karłowski WM. (2012) Annotating a non-model plant genome – a study on the narrow-leafed lupin. Journal of Biotechnology, Computational Biology and Bionanotechnology. 93(3):318-332.

CAPÍTULO 4

DESCOBERTA E CARACTERIZAÇÃO DE POLIMORFISMOS DE NUCLEOTÍDEO ÚNICO NO GENOMA NUCLEAR DE *Dipteryx alata* (LEGUMINOSAE)

RESUMO

Os polimorfismos de nucleotídeos únicos (SNPs) têm sido caracterizados como marcadores moleculares de preferência em vários estudos, uma vez que permitem uma ampla detecção da variação genômica. Avanços nas tecnologias de seqüenciamento e nas ferramentas bioinformáticas permitem a descoberta de SNPs com alto rendimento, mesmo em organismos não-modelo como o *Dipteryx alata*, espécie arbórea nativa do Bioma Cerrado Brasileiro. O presente estudo teve como objetivos descobrir e caracterizar SNPs em uma amostra do genoma de *D. alata*. Para tanto, amostras de DNA foram extraídas de folhas de quatro indivíduos coletados em diferentes localidades do Cerrado. As bibliotecas genômicas foram preparadas seguindo o protocolo Nextera e sequenciadas na plataforma Illumina MiSeq. A montagem dos reads foi realizada com o software Dipspades. Para a descoberta de variantes os reads de cada indivíduo foram alinhados as sequências montadas usando o software BWA e SNPs e InDels foram descobertos usando as ferramentas do pacote GATK (*Genome Analysis Toolkit*). Após a obtenção de SNPs e InDels putativos para a espécie, foram realizadas duas rodadas de recalibração e novas chamadas de variantes. VCFtools e predição de genes do software Augustus foram utilizados para estimar quantos SNPs estavam no contexto dos genes. O software SnpEff foi usado para predição dos efeitos dos SNPs sobre a função dos genes. Um total de 736.787 SNPs e 90.803 InDels foram descobertos em 355Mb de sequências (N50 de 1598). Foi identificada uma média de 1 SNP a cada 189 pb do genoma e a razão entre as mutações de transição (Ts) e transversão (Tv) foi de 1,58. Em relação às mudanças geradas pelos SNPs nas sequências, o número de transições C/T e A/G foi semelhante. As transversões A/C e G/T tiveram frequências semelhantes, enquanto as transversões A/T e C/G foram as mais e menos frequentes, respectivamente. Uma porcentagem de 46,5% dos SNPs ocorre no contexto gênico. Os efeitos dos SNPs foram anotados principalmente em éxons e regiões intergênicas. Os resultados deste estudo geraram novos conhecimentos sobre a estrutura e organização genômica de *D. alata*. Os SNPs identificados são consistentes para o desenvolvimento futuro de marcadores moleculares para aplicações em estudos de genômica populacional.

Palavras – chaves: Baru, Cerrado, SNPs, variação genômica.

INTRODUÇÃO

Os polimorfismos de nucleotídeo único (SNPs) são variações pontuais onde ocorre a variação de apenas uma base nitrogenada entre duas ou mais sequências de nucleotídeos (Kwok e Gu, 1999; Nadeau, 2002). As variações genéticas dos SNPs surgem por mutações nucleotídicas que geram substituições ou InDels (Brumfield et al 2003). As substituições consistem na troca de um nucleotídeo por outro e dão origem a diferentes alelos contendo bases alternativas (Brumfield et al 2003). As substituições do tipo transição ocorrem geralmente com maior frequência nos genomas do que as substituições do tipo transversão (Caetano, 2009). As mutações InDels, por sua vez, consistem em inserções ou deleções de nucleotídeo na sequência de DNA. As variações nas sequências de DNA dos SNPs podem constituir marcadores genéticos, ou seja, polimorfismos hereditários medidos e monitorados em uma ou mais populações de indivíduos (Davey et al 2011).

Os SNPs têm sido caracterizados como marcadores genéticos de preferência em vários estudos devido a sua abundância, uma vez que cobrem todo o genoma permitindo uma ampla detecção da variação genômica (Syvanen, 2001; Kerstens et al 2009; Ganal et al 2009). Por exemplo, foram descobertos 4.902.039 SNPs e 810.467 InDels no genoma da planta modelo *Arabidopsis thaliana* a partir da análise do DNA de 80 indivíduos provenientes de 8 regiões geográficas distintas (Cao et al 2011). Assim, os SNPs possuem ocorrência frequente nas sequências de DNA e podem ocorrer em regiões codantes ou não codantes do genoma. Os SNPs presentes nas regiões gênicas e regulatórias podem alterar a estrutura e a função das proteínas codificadas e modificar a expressão gênica (Caetano, 2009).

A descoberta e análise dos SNPs possibilitam realizar várias investigações biológicas importantes. Os SNPs possuem potencial de uso em investigações de diversidade genética, estrutura populacional, análise de cruzamentos consanguíneos, confirmação de paternidade, evolução de espécies, análise de adaptação em ecologia molecular, reconstrução filogenética, análise de características de produção visando o melhoramento animal e vegetal, estudos forenses, diagnóstico de doenças, desenvolvimento de novos fármacos terapeuticos, entre outras aplicações (Chakravarti, 2001; Syvanen, 2001; Nachman, 2001; Schlotterer, 2004; Caetano, 2009; Ganal et al 2009; Ma et al 2011; Ekblom e Galindo, 2011; Masoodi et al 2013).

Estudos baseados em marcadores SNPs dependem de análises em larga escala visto que um único SNP é pouco informativo (Syvanen, 2001; Brumfield et al 2003). Várias abordagens estão disponíveis para a descoberta de SNPs com alto rendimento (Ganal et al 2009). Uma dessas abordagens é o sequenciamento e a montagem do genoma da espécie seguido do re-sequenciamento de vários indivíduos visando comparar as sequências e identificar variantes. Essa abordagem é facilitada pelos avanços nas tecnologias de sequenciamento que são capazes de gerar vários bilhões de bases por corrida, em um curto espaço de tempo e com custos reduzidos. A grande quantidade de dados gerada pelo sequenciamento de nova geração e o uso das ferramentas de bioinformática permitem descobrir e analisar milhares de SNPs em centenas de indivíduos, mesmo em organismos não modelos com pouca ou nenhuma informação genética disponível (Kerstens et al 2009; Davey et al 2011; Jiang et al 2016).

O presente estudo teve como objetivo descobrir e caracterizar SNPs em uma amostra do genoma de *Dipteryx alata* usando dados de sequenciamento de nova geração. *D. alata* é uma espécie arbórea não modelo nativa do Bioma Cerrado Brasileiro que pertence à família Leguminosae. Possui potencial econômico uma vez que a poupa e as castanhas são comestíveis e processadas por pequenas indústrias para fabricação de sorvetes, licores, castanha torrada, entre outros produtos. Além disso, *D. alata* possui utilidade medicinal, forrageira, no paisagismo, na recuperação de áreas degradadas e na extração de madeira (Sano et al 2004). Os estudos realizados até agora para *D. alata* fizeram uso de pequeno número de marcadores moleculares, principalmente marcadores microssatélites (Collevatti et al 2013; Soares et al 2015). A descoberta e a seleção de SNPs com potencial para marcadores moleculares permitirá a análise da diversidade alélica em *D. alata* usando a genotipagem em larga escala e abrirá caminho para os estudos de genômica de populações.

MATERIAL E METÓDOS

Extração de DNA e sequenciamento Illumina

Conforme descrito no capítulo2, para a extração de DNA foram coletadas folhas de quatro indivíduos adultos de *Dipteryx alata* provenientes de diferentes regiões de Cerrado, sendo elas localizadas nas cidades de Natividade-TO, Camapuã-MS, Aquidauana-MS e Cáceres-MT. O método brometo de cetiltrimetilamônio (CTAB) descrito por Doyle e Doyle (1987) foi usado para extrair o DNA de cada indivíduo. Quatro bibliotecas de DNA, uma para cada amostra, foram preparadas para sequenciamento utilizando o kit *Nextera DNA Sample Preparation* (Illumina). As quatro bibliotecas foram agrupadas e sequenciadas em plataforma MiSeq da Illumina usando o kit de sequenciamento MiSeq v3 de 600 ciclos (2x250bp, reads paired-end). Foram realizadas duas corridas de sequenciamento, a primeira com as quatro bibliotecas agrupadas e a segunda somente com a biblioteca de Camapuã-MS, biblioteca que apresentou maior rendimento no primeiro sequenciamento.

Análise de qualidade e montagem de sequências genômicas nucleares

A qualidade dos reads foi avaliada usando o software FastQC (Andrews, 2010) e bases de baixa qualidade foram removidas usando o software Trimmomatic (Bolger et al 2014) (Capítulo 2). A montagem das sequências foi realizada usando o software dipSPAdes, um montador *de novo* que usa como algoritmo a construção de grafos de bruijn e foi desenvolvido para genomas diploides altamente polimórficos (Safonova et al 2015). Sequências menores que 500pb e sequências organelares foram excluídas da montagem (Capítulo 2). Sequências repetitivas foram identificadas e mascaradas usando os softwares RepeatMasker e RepeatModeler (Tarailo-Graovac e Chen, 2009; Smit e Hubley, 2015). O RepeatMasker identificou repetições nas sequências de DNA de *D. alata* por meio da análise de similaridade com a biblioteca de eudicotyledoneas armazenada em banco de dados. Além disso, o RepeatMasker identificou famílias de repetições de novo usando o banco de dados criado pelo RepeatModeler para as sequências de *D. alata*.

A qualidade da montagem foi avaliada usando o script assemblathon desenvolvido por Keith Bradnam (2011). A cobertura da montagem foi estimada através do alinhamento dos reads contra o fasta da montagem usando o software bowtie2 (Langmead et al 2009) e a ferramenta genomeCoverageBed da suíte bedtools (Quinlan e Hall, 2010).

Descoberta e caracterização de SNPs

Para a descoberta de polimorfismos de nucleotídeo único foram usadas as sequências montadas, livres de sequências organelares e com regiões repetitivas mascaradas, e os reads de cada biblioteca. A biblioteca sequenciada na segunda corrida Illumina, mesmo sendo repetição de uma das quatro bibliotecas do primeiro sequenciamento, foi considerada como uma biblioteca diferente. Dessa forma, foram estabelecidas cinco bibliotecas, a partir de quatro indivíduos, para a identificação de SNPs. A descoberta de variantes seguiu o pipeline desenvolvido por Nunes (dados não publicados), descrito a seguir.

Os reads de cada biblioteca de *D. alata* foram mapeados nas sequências genômicas montadas para a espécie usando como alinhador o software BWA (Li e Durbin, 2009). Várias ferramentas do pacote *Picard-Tools* (http://broadinstitute.github.io/picard/) foram usadas para preparar o arquivo de alinhamento para a chamada de variantes. A ferramenta *SortSam* do picard foi usada para converter o arquivo resultado do alinhamento do formato SAM (*Sequence Alignment/Map*) para o formato BAM (*Binary Alignment/Map*) e para ordenar os reads por coordenada de alinhamento. A ferramenta *MarkDuplicates* do picard foi utilizada para marcar possíveis duplicatas de PCR. Moléculas duplamente sequenciadas não foram usadas na chamada de variantes. A ferramenta *BuildBamIndex* do picard foi usada para criar os arquivos de índices para o arquivo BAM. A ferramenta *AddOrReplaceReadGroups* do picard foi usado para inserir dados de identificação nos arquivos de alinhamento.

A identificação de variantes foi realizada com o pacote GATK (*Genome Analysis Toolkit*) (McKenna et al 2010), que possui ferramentas para detectar variantes com alta confiança. O algoritmo *HaplotypeCaller* foi usado para chamar simultaneamente SNPs e InDels e armazená-los em um arquivo em formato VCF. Informações de SNPs e indels foram separadas em dois arquivos usando a ferramenta *SelectVariants*. A ferramenta *VariantFiltration* foi usada para filtrar SNPs e InDels de alta qualidade usando critérios determinados. Para SNPs foram usados os filtros QD < 1.0, FS > 5.0, MQ < 50.0, MQRankSum < -2.0, MQRankSum > 2.0, ReadPosRankSum < -2.5, ReadPosRankSum > 2.5 e para InDels os filtros QD < 0.6, FS > 8.0, ReadPosRankSum < -3.0, ReadPosRankSum > 3.0. As variantes que passaram pelos filtros foram marcadas com PASS enquanto as variantes consideradas ruins e filtradas foram marcadas com FILTER no arquivo VCF de saída. As variantes que apresentaram filtros diferentes de PASS foram excluídas das análises. *RealignerTargetCreator* e *IndelRealigner* do GATK foram usados para definir o intervalo de realinhamento e realinhar os indels, respectivamente. É comum a ocorrência de maior proporção de erros de alinhamento nas regiões próximas aos InDels. Dessa forma, o realinhamento é importante para minimizar a identificação de falsas variantes.

Após a obtenção de SNPs e InDels putativos para a espécie, foram realizadas duas rodadas de recalibração e novas chamadas de variantes. Em cada recalibração foi utilizado o

VCF com as variantes da etapa anterior com o intuito de observar a estabilização do número de polimorfismos identificados. A ferramenta *Base Quality Score Recalibrator* (BQSR) do GATK foi usada para realizar as recalibrações que visam melhorar as estimativas das notas de qualidade das bases. A chamada de variantes depende dos índices de qualidade de bases e as pontuações atribuídas pelos sequenciadores estão sujeitas a erros podendo estar super ou subestimadas. A recalibração da pontuação de qualidade de base consiste em modelar os erros empíricos e ajustar os resultados. A recalibração foi realizada em duas etapas: primeiro foi feita a construção de um modelo de covariação baseado nos dados e em seguida o ajuste dos scores de qualidade das bases baseado no modelo. A recalibração é importante porque permite obter bases mais precisas o que melhora a precisão da chamada de variantes. Após cada recalibração foi realizada uma nova chamada de variantes com o *HaplotypeCaller* usando os arquivos BAM recalibrados e aplicados os filtros supracitados para SNPs e InDels.

O programa VCFtools (Danecek et al 2011) foi usado para obter informações sobre os arquivos VCF de várias etapas do pipeline do GATK tais como o número de SNPs e InDels que passaram pelos filtros e a taxa de transição/transversão. VCFtools foi usado também para estimar quantos SNPs estão no contexto de genes. A análise dos SNPs em contexto gênico foi realizada usando o arquivo VCF com os resultados do GATK e o arquivo de predição de genes de *D. alata*. A predição de fragmentos gênicos de *D. alata* foi obtida usando o software Augustus (Stanke et al 2004). A predição dos efeitos dos SNPs sobre a função dos genes de *D. alata*, considerando alterações de aminonoácidos, foi obtida com o software SnpEff (Cingolani et al 2012). Scripts em R foram usados para separar apenas um SNP por contig, o de maior qualidade, e em distância de 200pb de outros SNPs na mesma sequência de *D. alata*, seleção realizada com o intuito de fornecer SNPs para o desenvolvimento de marcadores moleculares.

RESULTADOS

Montagem das sequências

Conforme descrito no capítulo 2, foram gerados com o sequenciamento Illumina 22 GB de dados que corresponde a 61.6 milhões de reads paired-end. Após o corte de qualidade foram mantidos 14.6 GB e 60.5 milhões de reads. Com esse esforço de sequenciamento foram montados 355 Mb scaffolds do genoma nuclear, com N50 igual a 1598. A figura 1 mostra a distribuição de tamanho das sequências na montagem. A cobertura da montagem foi estimada em 12.78X.

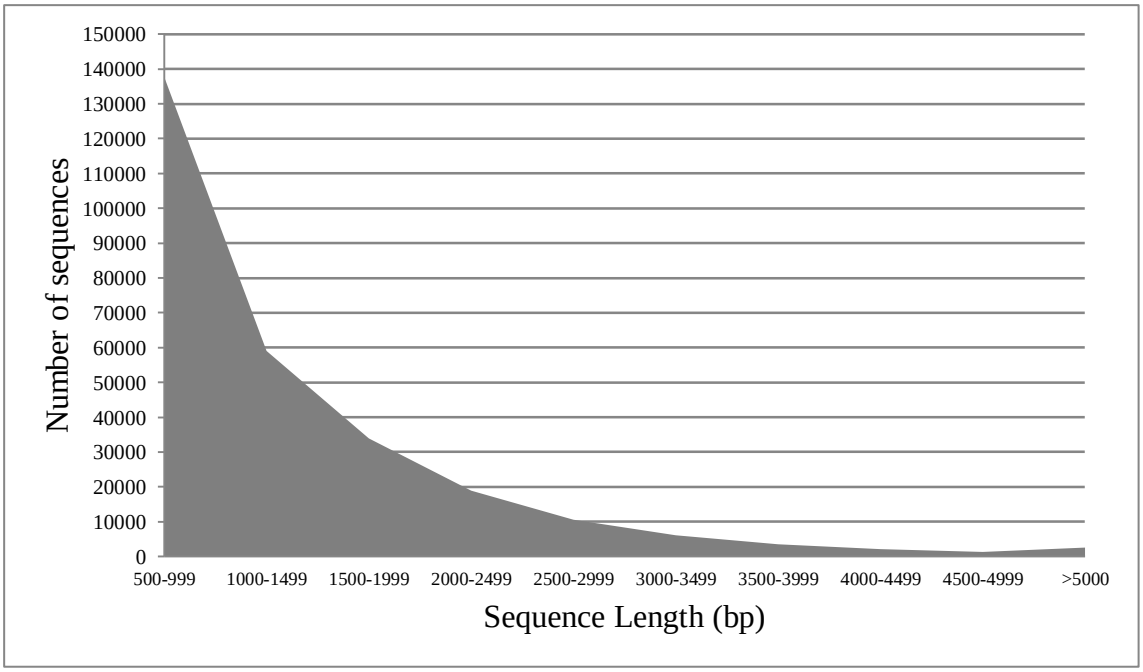

Figura 1 - Distribuição dos tamanhos das sequências nucleares na montagem de *D. alata*.

Descoberta de SNPs

Após a primeira chamada de variantes e primeira filtragem foram identificados 958.373 SNPs e 114.447 InDels, que foram considerados como referência para as posteriores etapas do pipeline do GATK (Tabela 1). Após a primeira recalibração e segunda filtragem foram identificados 765.701 SNPs e 94.504 InDels. Após a segunda recalibração e terceira filtragem foram identificados 736.787 SNPs e 90.803 InDels, variantes consideradas válidas para as análises futuras. Grandes quantidades de SNPs descobertos na primeira chamada de variantes foram mantidos após as recalibrações (593.954 SNPs) e a comparação dos resultados das duas recalibrações indica uma tendência de estabilização no número de variantes identificadas (Figura 2).

Após a segunda recalibração foram descobertos em média 1 SNP a cada 189 pb de sequência de *D. alata* e apenas 1.229 possuiam mais que dois alelos, ou seja, eram multi-alelicos. Em relação às mudanças geradas pelos SNPs nas sequências, o número de transições C/T e A/G foi semelhante. As transversões A/C e G/T tiveram frequências similares, enquanto as transversões A/T e C/G foram as mais e menos frequentes, respectivamente (Tabela 2). A razão entre mutações de transição (Ts) e transversão (Tv) foi de 1,58. Após a aplicação do filtro para selecionar 1 SNP por contig e em distância de 200pb foram mantidos 82.504 SNPs com Ts/Tv de 1,27.

Tabela 1 - Número de polimorfismos de nucleotídeo único (SNPs) e indels nas sequências de *D. alata* identificados em cada etapa do pipeline do GATK.

	Número total de indels	Número total de SNP	Número de transições (Ts)	Número de transversões (Tv)	Ts/Tv
Dados Brutos: Antes do filtro	116.608	1.071.768	661.148	410.620	1,61
Dados Brutos: Após filtro	114.447	958.373	587.572	370.801	1,58
Primeira recalibração: Antes do filtro	96.275	855.009	527.672	327.337	1,61
Primeira recalibração: Após filtro	94.504	765.701	469.711	295.990	1,58
Segunda recalibração: Antes do filtro	92.411	820.622	506.691	313.931	1,61
Segunda recalibração: Após filtro	90.803	736.787	452.272	284.515	1,58
1 SNP por contig e em distância de 200pb	_	82.504	46.265	36.239	1,27

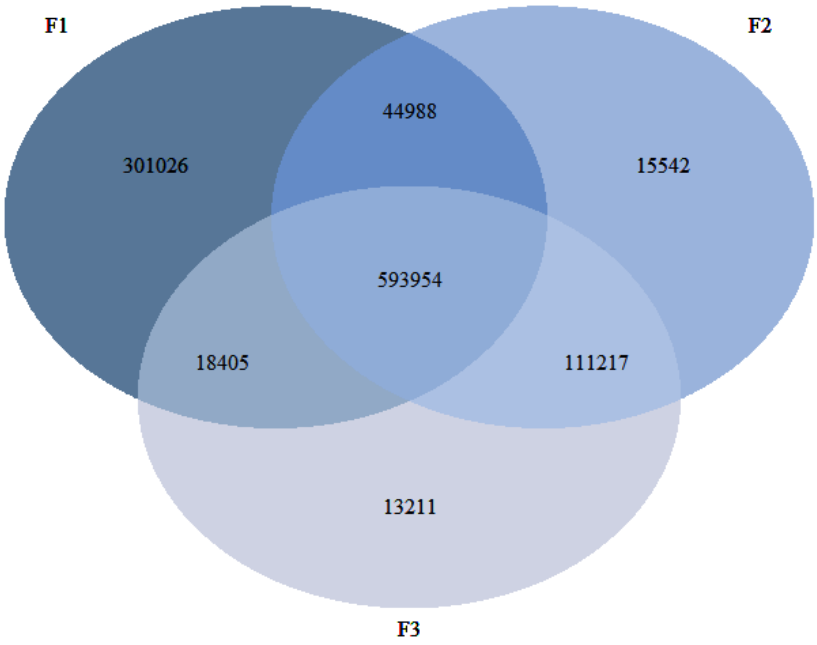

Figura 2 - Diagrama de Venn apresentando o número de SNPs após cada etapa de filtragem. F1 - Filtro após a primeira chamada de variantes. F2 - Filtro após a primeira recalibração. F3 - Filtro após a segunda recalibração.

Tabela 2 - Quantidades de transição e transversão para os SNPs identificados em *D. alata*.

Substituições	Número de SNPs
Transição (Ts)	451.918
C/T	226.193
A/G	225.725
Transversão (Tv)	284.869
C/G	50.413
A/T	93.196
A/C	70.791
G/T	70.469
Ts/Tv taxa	1,5864

Anotação das variantes e seus efeitos

Após a segunda recalibração 342.326 SNPs estavam em contexto de gene (Ts/Tv: 1,56) e 394.461 SNPs não estavam em contexto de gene (Ts/Tv: 1,61). Dos SNPs únicos por contig e em distância de 200 pb um total de 26.704 estava em contexto de genes (Ts/Tv:1,35) e 55.800 não estava em contexto de genes (Ts/Tv: 1,24). Os efeitos dos SNPs estavam principalmente em éxons (30,82%) seguidos de regiões intergênicas (20,09%), íntrons (19,51%), regiões upstream (10,78% - 5 KB do sitio de início da transcrição) e downstream (10,3% - 5KB a jusante da poliA) de genes (Figura 3).

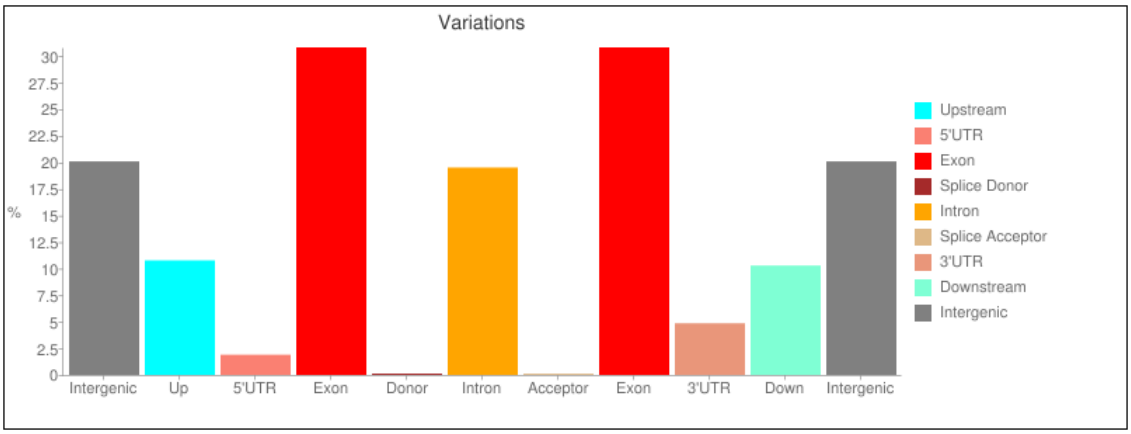

Figura 3 - Distribuição do percentual de efeitos dos SNPs baseada na região genômica onde ocorre em *D. alata*.

Os principais tipos de efeitos dos SNPs foram a variação na sequência de íntrons (20,59%), em regiões intergênicas (19,70%), variação missense (16,54%), variação sinônima (13,39%) e variação em regiões upstream (10,57%) e downstream (10,10%) de genes (Tabela 3).

Tabela 3 – Número de efeitos por tipos identificados na anotação funcional de SNPs em *D. alata*.

Tipos de efeitos	**Quantidade de efeitos**	**Percentual**
intron_variant	432.155	20.59%
intergenic_region	413.492	19.70%
missense_variant	347.200	16.54%
synonymous_variant	281.120	13.39%
upstream_gene_variant	221.758	10.57%
downstream_gene_variant	211.943	10.10%
3_prime_UTR_variant	100.005	4.77%
splice_region_variant	36.992	1.76%
5_prime_UTR_variant	33.610	1.60%
stop_gained	7.604	0.36%
5_prime_UTR_premature_start_codon_gain_variant	5.550	0.26%
splice_donor_variant	2.606	0.12%
splice_acceptor_variant	2.441	0.12%
stop_lost	1.102	0.05%
start_lost	774	0.04%
stop_retained_variant	441	0.02%
initiator_codon_variant	150	0.01%
non_canonical_start_codon	7	0%

Os efeitos dos SNPs foram classificados quanto ao impacto em alto (0,7%), baixo (15,2%), moderado (16,8%) e modificador (67,1%). Quanto às classes funcionais, os efeitos dos SNPs foram classificados em missense (54,7%), nonsense (1,1%) e silencioso (44,1%).

DISCUSSÃO

O esforço de sequenciamento de DNA para *Dipteryx alata* usando a plataforma MiSeq Illumina permitiu montar 44% do genoma da espécie, que foi estimado em 807.2MB usando Citometria de Fluxo (Capítulo 1). A montagem obtida, apesar de não amostrar o genoma completo de *D. alata*, foi considerada satisfatória para a identificação e caracterização de variantes. O número de sequências obtidas para *D. alata* permitiu anotar um grande número de SNPs e InDels e contribuiu para aumentar as informações genômicas disponíveis para esta espécie do Cerrado. Até o desenvolvimento desse trabalho, não foi encontrado na literatura estudos de descoberta e caracterização de SNPs em *D. alata*.

A quantidade de SNPs e InDels descobertos (736.787 e 90.803, respectivamente) sugere alto nível de polimorfismos entre indivíduos de *D. alata* oriundos de diferentes populações do Cerrado. As variantes putativas identificadas podem ser consideradas consistentes uma vez que sobreviveram as duas etapas de recalibração e aos filtros de alta qualidade. Dos SNPs identificados 61,3% representavam mutações de transição e 38,7% mutações de transversão, o que permitiu estimar a razão transição/transversão (Ts/Tv) em 1,58. A razão Ts/Tv está associada a erros de sequenciamento, quanto menor o Ts/Tv mais a

chance de identificação de falsos SNPs. O valor de Ts/Tv estimado para *D. alata* é semelhante ao encontrado na literatura para *Phaseolus vulgaris* (Ts/Tv igual a 1,56), espécie que também pertence à família Leguminosae (Ariani et al 2016).

O total de 342.326 SNPs, que corresponde a 46,5% dos SNPs putativos descobertos, ocorre em contexto de genes. Quantidade significativa de SNPs estavam localizados em éxons, íntrons e regiões intergênicas. Os SNPs presentes em contexto gênico podem mudar a forma alélica do gene e alterar a estrutura e a função das proteínas codificadas (Caetano, 2009). Em *D. alata* 67,1% dos SNPs exercem impacto modificador no genoma, tipo de impacto que não é classificado nas outras categorias (alto, baixo e moderado) e consiste principalmente de mudanças nas sequências de DNA. São exemplos de impactos modificadores a variação em íntrons e a variação em regiões intergênicas. SNPs com impacto moderado e baixo tiveram frequências semelhantes em *D. alata* (16,8% e 15,2%, respectivamente). Mutações missenses, ou seja, mutações que levam a codificação de um aminoácido diferente do normal possuem impacto moderado no genoma. Mutações missenses foram a classe funcional mais frequente de efeitos de SNPs em *D. alata*. Outros exemplos de SNPs com impacto moderado são variação em regiões de splice e em regiões regulatórias. Mutações classificadas como de alto impacto, tais como perda de éxons, perda de stop ou start códon e ganho de stop códon (substituição nonsense) não são frequentes no genoma de *D. alata*. Esse resultado era esperado visto que variantes com alto impacto estão sob forte pressão seletiva.

Os SNPs identificados para *D. alata* fornecem subsídios para o futuro desenvolvimento de marcadores moleculares para aplicação em estudos de genômica de populações para a espécie. O desenvolvimento de marcadores moleculares permitirá a análise da diversidade alélica em indivíduos e populações de *D. alata* usando a genotipagem em larga escala. Investigações usando grande número de marcadores elevará o poder para fazer inferência em diversos estudos já que aumenta a precisão e a exatidão da estimativa de parâmetros genéticos populacionais. Além disso, o uso de marcadores em larga escala permitirá que novas questões sejam abordadas como, por exemplo, estudos de processos adaptativos que dependem de regiões genômicas em resposta à seleção (Luikart et al 2003).

CONCLUSÃO

O total de 736.787 SNPs e 90.803 InDels putativos foram descobertos e caracterizados no genoma de *D. alata*. As variantes identificadas são consistentes para o futuro desenvolvimento de marcadores moleculares para *D. alata* que tornarão viáveis estudos de genômica de populações. Muitos SNPs foram identificados em contexto gênico e serão importantes para avaliar o efeito da seleção natural.

MATERIAL SUPLEMENTAR

https://www.dropbox.com/sh/i6u8roglrx5dusw/AACVy6CoglMGYSyDZj8J9NEPa?dl=0

REFERÊNCIAS

Altschul SF; Gish W; Miller W; Myers EW; Lipman D J. (1990). Basic local alignment search tool. Journal of molecular biology, 215(3):403-410.

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Ariani A; Mier JCB; Gepts P (2016) Genome-wide identification of SNPs and copy number variation in common bean (*Phaseolus vulgaris* L.) using genotyping-by-sequencing (GBS). Mol Breeding, 36:87.

Bolger AM; Lohse M; Usadel B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 1.

Brumfield RT; Beerli P; Nickerson DA; Edwards SV. (2003) The utility of single nucleotide polymorphisms in inferences of population history. Trends in Ecology and Evolution, 18(5):249-256.

Caetano AR. (2009) Marcadores SNP: conceitos básicos, aplicações no manejo e no melhoramento animal e perspectivas para o futuro. R. Bras. Zootec., 38:64-71.

Cao J; Schneeberger K; Ossowski S; Günther T; Bender S; Fitz J; Koenig D; Lanz C; Stegle O; Lippert C; Wang X; Ott F; Müller J; Alonso-Blanco C; Borgwardt K; Schmid KJ; Weigel D. (2011) Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nature Genetics, 43:956-966.

Chakravarti A. (2001) Single nucleotide polymorphisms …to a future of genetic medicine. Nature, 409:822-823.

Cingolani R; Platts A; Wang L; Coon M; Nguyen T; Wang L; Land SJ; Lu X; Ruden DM. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly, 6(2):80-92.

Collevatti RG; Telles MPC; Nabout JC; Chaves LJ; Soares TN (2013) Demographic history and the low genetic diversity in *Dipteryx alata* (Fabaceae) from Brazilian Neotropical savannas. Heredity, 111:97–105.

Danecek D; Auton A; Abecasis G; Albers CA; Banks E; DePristo MA; Handsaker RE; Lunter G; Marth GT; Sherry ST; McVean G; Durbin R; 1000 Genomes Project Analysis Group (2011) The variant call format and VCFtools. Bioinformatics Applications note, 27:2156-2158.

Davey JW; Hohenlohe PA; Etter PD; Boone JQ; Catchen JM; Blaxter ML. (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Genetics, 12.

Ekblom R; Galindo J. (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity, 107:1–15.

Ganal MW; Altmann T; Roder MS (2009) SNP identification in crop plants, Current Opinion in Plant Biology 12:211–217.

Jiang Z; Wang H; Michal JJ; Zhou X; Liu B; Woods LCS; Fuchs RA. (2016) Genome Wide Sampling Sequencing for SNP Genotyping: Methods, Challenges and Future Development. International Journal of Biological Sciences, 12(1):100-108.

Kerstens HHD; Crooijmans RPMA; Veenendaal A; Dibbits BW; Chin-A-Woeng TFC; Dunnen JTD; Groenen MAM. (2009) Large scale single nucleotide polymorphism discovery in unsequenced genomes using second generation high throughput sequencing technology: applied to turkey. BMC Genomics, 10.

Kwok PY; Gu Z. (1999) Single nucleotide polymorphism libraries: why and how are we building them? Molecular Medicine Today, 1999.

Langmead B; Trapnell C; Pop M; Salzberg SL. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology, 10.

Li H; Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics, 25:1754-60.

Luikart G; England PR; Tallmon D; Jordan S; Taberlet P. (2003) The Power and Promise of Population Genomics: From Genotyping to Genome Typing. Nature Reviews Genetics. 4:981-994.

Ma H; Ma Q; Ma C; Ma L. (2011) Isolation and characterization of gene-derived single nucleotide polymorphism (SNP) markers in *Scylla paramamosain*. Biochemical Systematics and Ecology, 39:419–424.

Masoodi TA; Shammari SAA; Muammar MNA; Alhamdan AA; Talluri VR. (2013) Exploration of deleterious single nucleotide polymorphisms in late-onset Alzheimer disease susceptibility genes. Gene 512:429–437.

McKenna A; Hanna M; Banks E; Sivachenko A; Cibulskis K; Kernytsky A. (2010) The genome analysis toolkit: a MapReduce frame work for analyzing next-generation DNA sequencing data. Genome Res. 20:1297–1303.

Nachman WN. (2001) Single nucleotide polymorphisms and recombination rate in humans. TRENDS in Genetics, 9(17):481-485.

Nadeau JH. (2002) Tackling complexity. Nature, 420:517-518.

Quinlan AR; Hall IM. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, *26*(6):841-842.

Safonova Y; Bankevich A; Pevzner PA. (2015) dipSPAdes: Assembler for Highly Polymorphic Diploid Genomes. Journal Of Computational Biology, 22(6):528-545.

Sano SM; Ribeiro JF; Brito MA. (2004) Baru: biologia e uso. Planaltina, DF: Embrapa Cerrados, 52p.

Schlotterer C. (2004) The evolution of molecular markers — just a matter of fashion? Nature Genetics, 5:73-69.

Smit AFA; Hubley R. (2015) RepeatModeler Open-1.0 http://www.repeatmasker.org

Soares TN; Chaves LJ; Telles MPC; Diniz-Filho JAF; Resende LV. (2008a) Distribuição espacial da variabilidade genética intrapopulacional de *Dipteryx alata*. Pesq. agropec. bras., 43(9):1151-1158.

Soares TN; Diniz-Filho JAF; Nabout JC; Telles MPC; Terribile RC; Chaves LJ. (2015) Patterns of genetic variability in central and peripheral populations of *Dipteryx alata* (Fabaceae) in the Brazilian Cerrado. Plant Syst Evol, 301:1315–1324.

Stanke M; Morgenstern B. (2005) AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints.Nucleic Acids Research, 33.

Syvanen AC. (2001) Accessing genetic variation: genotyping Single Nucleotide Polymorphisms. Nature Genetics, 2:930-942.

Tarailo-Graovac M; Chen N. (2009) Using RepeatMasker to Identify UNIT 4.10. Repetitive Elements in Genomic Sequences. Current Protocols in Bioinformatics.

CAPÍTULO 5

MONTAGEM *DE NOVO* E CARACTERIZAÇÃO DE SEQUÊNCIAS GENÔMICAS CLOROPLASTIDIAIS E MITOCONDRIAIS DE *Dipteryx alata* (LEGUMINOSAE) USANDO DADOS DE SEQUENCIAMENTO DE NOVA GERAÇÃO

RESUMO

Os avanços nas plataformas de sequenciamento e nas ferramentas de bioinformática permitiram um aumento significativo no número de genomas organelares completos ou parciais depositados em bancos de dados. O objetivo do presente estudo foi sequenciar, montar e anotar sequências genômicas cloroplastidiais e mitocondriais de *Dipteryx alata*, espécie de árvore endêmica do Cerrado Brasileiro, usando dados de sequenciamento Illumina. Bibliotecas genômicas foram preparadas a partir de amostras de DNA total foliar e sequenciadas usando a plataforma Illumina MiSeq. Os reads organelares foram isolados por meio da similaridade com sequências conhecidas de cloroplastos e mitocôndrias depositadas em banco de dados para outras espécies de plantas. Reads organelares foram montados com software Spades. A anotação gênica foi obtida com os softwares Augustus, Blast2GO, RNAmmer, tRNAscan-SE e Blastn usando o banco de dados PNRD (Plant Non-coding RNA Database). Anotação de elementos repetitivos foi realizada com os softwares QDD e RepeatMasker. Foram montados 110 KB de sequências cloroplastidiais com N50 de 2.384pb e 327 KB de sequências mitocondriais com N50 de 1.784pb. Foram anotados genes de 3 rRNA, 13 tRNA, 6 miRNA e 20 lncRNA para o cloroplasto e genes de 4 rRNA, 26 tRNA, 7 miRNA e 54 lncRNA para a mitocôndria. Para o cloroplasto foram preditos 20 genes codificantes de proteínas com tamanho médio de 2.374pb e para a mitocôndria foram preditos 176 genes com tamanho médio de 1.279pb. O número estimado de mRNAs transcritos por esse conjunto de genes é igual a 63 e 525 para cloroplasto e mitocôndria, respectivamente. Foram anotados também 39 regiões microssatélites e 4 elementos transponíveis nas sequências cloroplastidiais e 158 regiões microssatélites e 26 elementos transponíveis nas sequências mitocondriais. Os microssatélites mais abundantes foram os dinucleotídeos AT e os elementos transponíveis mais comuns foram os retroelementos. As sequências de genes e elementos repetitivos organelares obtidas para *D. alata* representam uma adição valiosa a informação genômica dessa espécie do Cerrado. Essas sequências poderão ser usadas para o futuro desenvolvimento de marcadores moleculares úteis em estudos de diversidade genética, filogenia, sistemática molecular e evolução.

Palavras-chaves: Cerrado, sequenciamento Illumina, genes, microssatélites, elementos transponíveis.

INTRODUÇÃO

As células vegetais contem, além do genoma nuclear, os genomas das organelas cloroplasto e mitocôndria. É amplamente aceito que mitocôndrias e cloroplastos surgiram pela simbiose entre cianobactérias e células eucarióticas, o que justifica estas organelas possuírem genoma próprio (Xiong et al 2008; Qian et al 2013; Kim et al 2016). Os genomas organelares estão presentes em múltiplas cópias na célula, por exemplo, na espécie modelo *Arabidopsis thaliana* uma célula foliar apresenta em média 100 cópias de DNA mitocondrial e 1000 cópias de DNA cloroplastidial (Wang et al 2012). Em geral, os genomas cloroplastidiais e mitocondriais possuem herança materna e o genoma mitocondrial é maior e mais complexo que o genoma cloroplastidial (Xiong et al 2008; Zhang et al 2012; Shetty et al 2016).

O genoma do cloroplasto é constituído por moléculas circulares de DNA dupla fita com tamanho entre 120 e 160 KB. É organizado em quatro partes conservadas, que inclui duas cópias invertidas com tamanho entre 20 e 28KB que divide uma região grande de cópia única (80-90 KB) de uma região pequena de cópia única (16-27KB) (Zhang et al 2011; Qian et al 2013; Zhang et al 2014; Ni et al 2016). O conteúdo e a ordem dos genes são altamente conservados na maioria dos genomas cloroplastidiais (Zhang et al 2012; Yi et al 2016). Por outro lado, o genoma da mitocôndria é maior, constituído por DNA circular de cadeia dupla com tamanho entre 200 e 2400 KB, e possui estrutura bastante variável entre as espécies de plantas. O genoma mitocondrial de plantas é de 10-100 vezes maior que os genomas mitocondriais animais (Xiong et al 2008). O conteúdo de genes é conservado, mas a ordem dos genes vária no genoma mitocondrial. Além disso, o tamanho do genoma da mitocôndria é mais variável que o do cloroplasto devido ao maior percentual de elementos repetitivos (Zhang et al 2011; Zhang et al 2012). O genoma do cloroplasto apresenta pouca ou nenhuma recombinação e possui taxa de mutação lenta, o que mantem a estrutura genômica relativamente conservada ao longo da evolução, enquanto a taxa de mutação no genoma da mitocôndria é comparativamente maior (Jeong et al 2014; Hoang et al 2015; Yi et al 2016; Prabhudas et al 2016).

Cloroplastos e mitocôndrias possuem a maquinaria transcricional e traducional completa para a expressão da informação genética (Xiong et al 2008; Asif et al 2013). Possuem genes codificantes de RNAs não traduzidos tais como RNAs transportadores (tRNA) e RNAs ribossômicos (rRNA) e genes que codificam RNAs mensageiros (mRNA) que são traduzidos em proteínas. O genoma do cloroplasto possui os genes essenciais para a fotossíntese e para a biossíntese de amido, ácidos graxos, pigmentos e aminoácidos (Zhang et al 2011; Qian et al 2013; Ni et al 2016; Prabhudas et al 2016). Os genes mitocondriais codificam uma serie de componentes necessários a fosforilação oxidativa e aos outros processos em que as mitocôndrias estão envolvidas tais como sinalização celular, diferenciação celular e morte celular (Xiong et al 2008; Wang etl al 2012). Nos últimos anos houve um aumento nas informações acerca da estrutura e função dos genomas organelares devido à ampliação do número de genomas sequenciados e anotados.

As primeiras sequências organelares foram obtidas pelos sequenciadores Sanger, porém, o surgimento do sequenciamento de nova geração e os avanços nas ferramentas de bioinformática revolucionaram a capacidade de produzir e anotar dados de sequências de

nucleotídeos. As plataformas de sequenciamento de nova geração permitem gerar grandes quantidades de sequências de DNA e um curto período de tempo e com custo reduzido. As ferramentas de bioinformática possibilitam a montagem e anotação de sequências genômicas complexas de forma eficiente e sem a existência de genomas de referência (Zhang et al 2011; Mardis 2008; Metzker 2010; Yi et al 2016). Com as novas abordagens o número de genomas organelares completos ou parciais depositadas no banco de dados GenBank aumentou significativamente (http://www.ncbi.nlm.nih.gov/genome/organelle/) (Shetty et al 2016). O sequenciamento e a anotação dos genomas organelares fornecem informações valiosas que abrem caminhos para diversos estudos.

A obtenção de sequências genômicas possibilita, além da anotação de genes, a caracterização de elementos repetitivos organelares tais como microssatélites e elementos transponíveis (TEs). Os microssatélites, também chamados de repetições de sequências simples (SSRs), são repetições curtas em tandem com motivos de 1 a 6 nucleotídeos (Foulongne-Oriol et al 2013). Os TEs são sequências de DNA capazes de se transpor no genoma, ou seja, de serem inseridos em novas localizações cromossômicas (Feschotte et al 2002). SSRs e TEs organelares geram polimorfismos genéticos que podem ser usados para o desenvolvimento de marcadores moleculares (Zhang et al 2012). Marcadores moleculares baseados em elementos repetitivos são ferramentas úteis em estudos de diversidade genética, estudos filogenéticos e evolutivos (Zhang et al 2012; Ni et al 2016; Shetty et al 2016). Genes organelares também servem como marcadores em estudos de sistemática molecular e evolução (Xu et al 2015). A natureza conservada, mas com taxa de mutação suficiente para captar eventos evolutivos, fazem das sequências organelares, principalmente das cloroplastidiais, ferramentas adequadas e de valor inestimável para diversos estudos (Zhang et al 2014).

Nesse sentido, o objetivo do presente estudo foi sequenciar, montar e anotar sequências genômicas cloroplastidiais e mitocondriais de *Dipteryx alata* usando dados de sequenciamento de nova geração. *D. alata*, conhecido popularmente como Baru, é uma espécie de árvore endêmica do Cerrado Brasileiro pertencente à família Leguminosae (Sano et al 2004). A família Leguminosae é uma das maiores famílias de angiospermas (Guo et al 2007). Várias espécies da família Leguminosae pertencentes ao Cerrado e este pode ser considerado um dos primeiros estudos genômicos para as espécies nativas deste bioma. Os dados obtidos podem contribuir para uma melhor compreensão da estrutura e organização dos genomas organelares de *D. alata* e ainda poderão ser usados para o desenvolvimento de marcadores moleculares para a espécie.

MATERIAL E METÓDOS

Material vegetal e extração de DNA

Foram coletadas folhas de quatro indivíduos adultos de *Dipteryx alata* provenientes das cidades Natividade-TO, Camapuã-MS, Aquidauana-MS e Cáceres-MT, todas regiões de bioma Cerrado. O DNA total de cada indivíduo foi extraído separadamente usando o método brometo de cetiltrimetilamônio (CTAB) descrito por Doyle e Doyle (1987).

As análises qualitativas e quantitativas das amostras de DNA foram realizadas por eletroforese horizontal em gel de agarose e ensaio Qubit® DNA Fluorometer (Invitrogen), respectivamente.

Preparo das bibliotecas de DNA e sequenciamento Illumina

As amostras de DNA total foram usadas para preparar as bibliotecas seguindo o protocolo do kit Nextera da Illumina. Foram preparadas quatro bibliotecas, uma para cada amostra, e a validação foi realizada com Bioanalyzer (Agilent Technologies 2100) e PCR em tempo real (Applied Biosystems StepOne). Os kits usados foram o High Sensitivity DNA kit para o Bioanalyzer e o kit Kappa (Kapa Biosystems) para PCR em tempo real. As bibliotecas foram agrupadas e sequenciadas usando o kit de sequenciamento Miseq v3 de 600 ciclos (2x250bp, reads paired-end) e a plataforma MiSeq da Illumina. Foram realizadas duas corridas de sequenciamento, a primeira com as quatro bibliotecas agrupadas e a segunda somente com a biblioteca de Camapuã-MS.

Análise de qualidade e montagem de sequências genômicas organelares

A qualidade dos reads foi avaliada usando o software fastQC e sequências de baixa qualidade foram removidas usando trimmomatic (Andrews, 2010; Bolger et al 2014). Phred Q>30 foi considerado como o limiar mínimo de qualidade. Reads menores do que 20 pares de bases foram eliminados. Para a separação dos reads de sequências organelares e nucleares foram usados o software bowtie e bancos de dados organelares (Langmead et al 2009). Foram criados dois bancos de dados como um subconjunto do banco de dados GenBank do NCBI. Sequências de cloroplasto e mitocôndrias de todas as espécies de angiospermas depositada no RefSeq do GenBank ate janeiro de 2015 foram amostradas. O conjunto total de reads foi alinhado contra o banco de sequências de DNA de cloroplasto e os reads mapeados contra esse banco foram separados para a posterior montagem das sequências cloroplastidiais de *Dipteryx alata*. O mesmo procedimento foi repetido usando um banco de dados de genomas mitocondriais e reads mapeados foram separados para a montagem das sequencias de DNA mitocondriais de *Dipteryx alata*.

Para a montagem das sequências de cloroplasto e mitocôndria foram testados três montadores que usam como algoritmo a construção de grafos de bruijn: ABySS (Simpson et al 2009), SOAPdenovo2 (Luo et al 2012) e Spades (Bankevich et al 2012). A qualidade das montagens foi avaliada usando o script em perl disponibilizado pelo grupo do Assemblathon. A montagem usada para as anotações foi selecionada com base na comparação dos seguintes parâmetros: volume de dados montados, número de scaffolds, tamanho do maior scaffold e N50.

Anotação de genes de RNAs não codificantes (ncRNAs)

RNAs ribossomais (rRNAs) foram preditos usando o programa RNAmmer (Lagesen et al 2007). RNAs transportadores (tRNAs) foram preditos usando o programa

tRNAscan-SE (Lowe e Eddy, 1997). Outros RNAs não codificantes de proteínas (ncRNAs), tais como pequenos RNA nucleares (snRNAs), pequenos RNA nucleolares (snoRNAs), microRNAs (miRNAs) e longos RNAs não codificante (lncRNAs), foram preditos usando o Blastn (Basic Local Alignment Search Tool) com E-valor de 10^{-6} e o banco de dados PNRD (Plant Non-coding RNA Database). O banco PNRD (http://structuralbiology.cau.edu.cn/PNRD/) diponibiliza sequências de ncRNAs de 150 espécies de plantas (Yi et al 2015).

Anotação de genes codificantes de proteínas

Genes completos e parciais (sequências codantes) codificantes de proteínas foram preditos com o software Augustus usando o genoma de Arabidopsis como matriz de treinamento (Stanke et al 2004). Arquivos com as sequências dos transcritos foram obtidos a partir dos resultados do Augustus para cloroplasto e mitocôndria. O software BLASTx foi usado para comparar as sequências transcritas contra um banco de dados de nucleotídeos traduzidos com E-valor menor que 10^{-6} (Altschul at al 1990). Os melhores resultados do alinhamento foram analisados com o software Blast2GO v3.3 para anotação gênica funcional de *D. alata* (Conesa et al 2005). As sequências foram anotadas nas três categorias funcionais definidas pelo Gene Ontology (http://www.geneontology.org): componentes celulares, função molecular e processos biológicos.

Anotação de regiões microssatélites

O software QDD versão 3.1.2 instalado em VirtualBox 5.0.14 para hospedeiro Windows foi usado para identificação de regiões microssatélites perfeitas nas sequências organelares (Meglécz et al 2010). Para a busca dos microssatélites foram usados como número mínimo de repetições para cada tamanho de motivo: quatro para di e tri-nucleotídeos e três para tetra, penta e hexa-nucleotídeos.

Anotação de TEs

Os elementos transponíveis nas sequências dos genomas organelares de *D. alata* foram detectadas e anotadas usando o programa RepeatMasker (Tarailo-Graovac e Chen, 2009). O RepeatMasker rastreia elementos repetitivo em sequências de DNA por meio da comparação contra bibliotecas de espécies com genomas conhecidos armazenados em banco de dados. Nesse sentido, as sequências de *D. alata* foram analisadas contra o banco de Eudicotiledôneas.

RESULTADOS

Montagem das sequências

As duas rodadas de sequenciamento com a plataforma MiSeq da Illumina geraram 22 GB de dados que corresponde a 61.62 milhões de reads paired-end. Depois do corte de qualidade, os reads foram separados 175.174 mil cloroplastidiais (33 MB) e 184.064 mil mitocondriais (34 MB) e estes dados foram usados para a montagem das sequências genômicas de *D. alata* (Tabela 1). As montagens com o software Spades atingiram os melhores resultados, maior N50, e foram selecionadas para as análises de anotações (Tabela 1; Figura 1 e 2).

Tabela 1 – Qualidade das montagens de sequências genômicas organelares de *Dipteryx alata*.

Parâmetros	**Cloroplasto**			**Mitocôndria**		
Montadores	ABySS	SOAP	Spades	ABySS	SOAP	Spades
Volume de dados montados (KB)	92	121	110	274	305	327
Número total de scaffolds	76	151	69	466	483	251
Tamanho total dos scaffold (pb)	92099	120633	107457	272805	303376	318878
Tamanho do maior scaffold (pb)	16449	10260	23116	13320	5169	16466
Tamanho do menor scaffold (pb)	63	103	130	83	117	362
N50 dos scaffolds	2246	2142	2384	822	841	1764
Scaffolds > 500 nucleotídeos	40	47	46	173	195	188
Scaffolds > 1k nucleotídeos	26	30	27	65	76	85
Scaffolds > 10k nucleotídeos	1	1	1	1	0	2

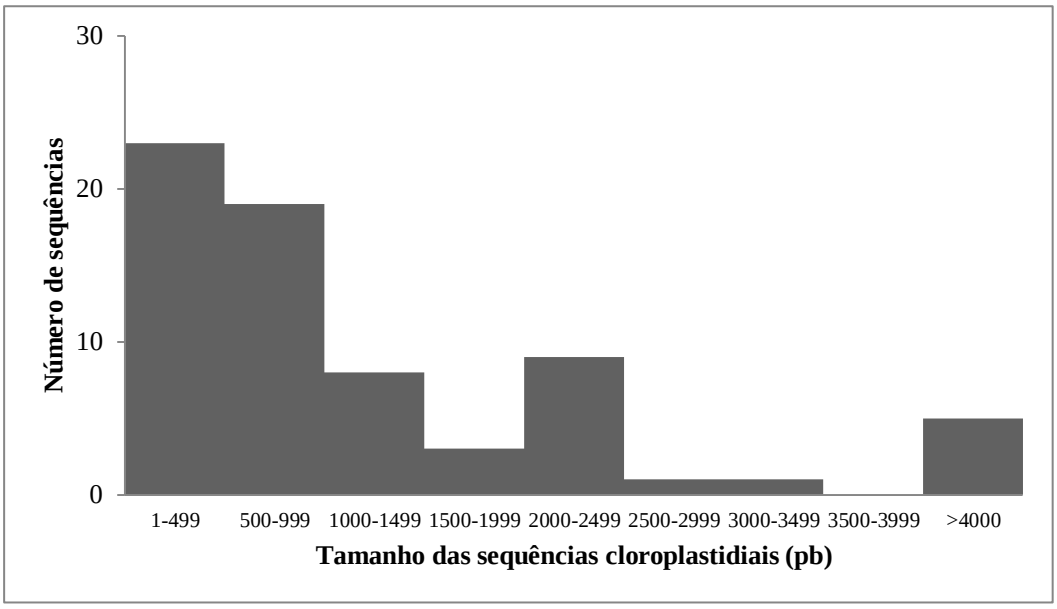

Figura 1 - Distribuição dos tamanhos das sequências organelares cloroplastidiais na montagem de *D. alata* com o software Spades.

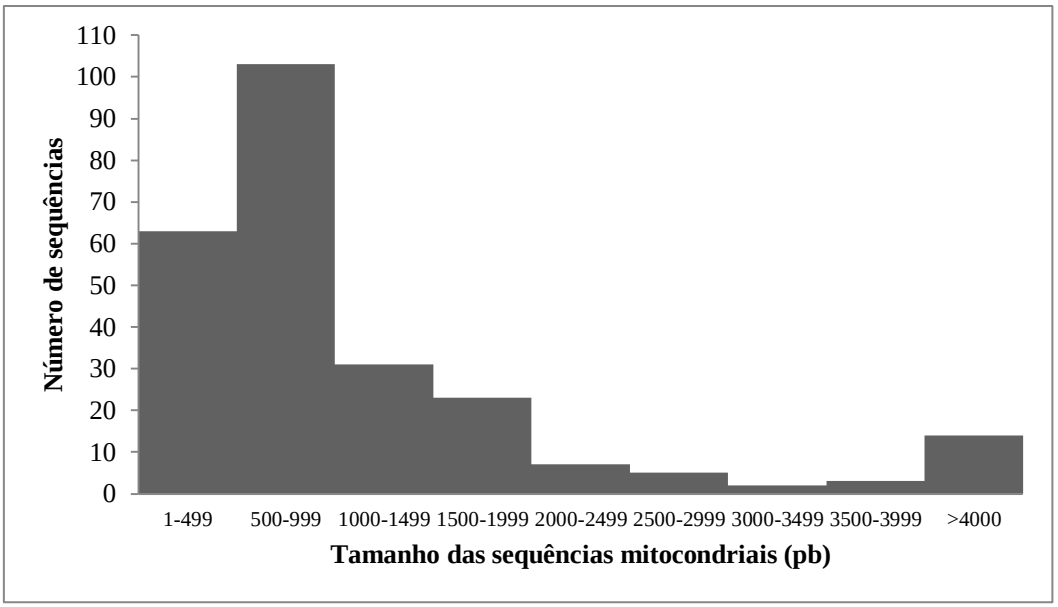

Figura 2 - Distribuição dos tamanhos das sequências organelares mitocondriais na montagem de *D. alata* com o software Spades.

Anotação gênica de RNAs não traduzidos

Foram identificados, nas sequências organelares montadas para *D. alata*, 133 genes codificantes de RNAs não traduzidos, sendo 42 de cloroplasto e 91 de mitocôndria (Tabela 2). Em ambas as organelas os RNAs transportadores e os longos RNAs não codificantes foram identificados em maior quantidade. Foram anotados também RNAs ribossômicos e microRNA.

Tabela 2 - Anotação estrutural de genes codificantes de RNAs não traduzidos em sequências organelares de *D. alata*.

Tipo de RNA	Número de RNAs	Tamanho (pb)		
		Médio	Mínimo	Máximo
Cloroplasto				
RNA ribossômico (rRNA)	3	1626	114	3288
RNA Transportador (tRNA)	13	75	71	88
microRNA (miRNA)	6	100	75	168
Longo RNA não codificante (lncRNA)	20	2884	292	14802
Mitocôndria				
RNA ribossômico (rRNA)	4	1246	104	1478
RNA Transportador (tRNA)	26	73	60	87
microRNA (miRNA)	7	109	75	168
Longo RNA não codificante (lncRNA)	54	1055	213	4532

Anotação gênica de RNAs codificantes de proteínas

Um total de 20 genes completos e parciais, compostos por 313 éxons e 271 íntrons, foi predito para o cloroplasto de *D. alata*. O número máximo de genes por sequência foi 4 e o tamanho médio dos genes foi 2374 pb. O número médio de éxons por gene foi 4.96 e o tamanho médio dos éxons foi 355 pb. O número médio de íntrons por gene foi 5.21 e o tamanho médio dos íntrons foi 241 pb. O número estimado de mRNAs transcritos pelo conjunto de genes cloroplastidiais anotados foi igual a 63. O número médio de transcritos por gene foi 3.15 e o tamanho médio dos transcritos foi 2810 pb. Nas sequências de mitocôndrias de *D. alata* foram preditos 176 genes, compostos por 2053 éxons e 1706 íntrons. O número máximo de genes por sequência foi 4 e o tamanho médio dos genes foi 1279 pb. O número médio de éxons por gene foi 3.91 e o tamanho médio dos éxons foi 288 pb. O número médio de íntrons por gene foi 3.74 e o tamanho médio dos íntrons foi 200 pb. O número estimado de mRNAs transcritos pelo conjunto de genes mitocondriais anotados foi igual a 525. O número médio de transcritos por gene foi 2.98 e o tamanho médio dos transcritos foi 1801 pb (Tabela 2; Material suplementar online Figura1 e Figura2).

Tabela 3 – Anotação estrutural de genes organelares codificantes de proteínas em *D. alata*.

Parâmetros	**Cloroplasto**			**Mitocôndria**		
	Valores estimados			Valores estimados		
	Médio	Mínimo	Máximo	Médio	Mínimo	Máximo
Número de genes por scaffold	1.25	1	4	1.22	1	4
Tamanho dos genes (pb)	2374	129	7590	1279	150	7807
Número de éxons por gene	4.96	1	21	3.91	1	22
Tamanho dos éxons (pb)	355	42	2990	288	8	2570
Número de íntrons por gene	5.21	1	21	3.74	1	20
Tamanho dos íntrons (pb)	241	67	2087	200	5	1192
Número de transcritos por gene	3.15	1	14	2.98	1	17
Tamanho dos transcritos (pb)	2810	129	7590	1801	150	7807

Anotação funcional de genes codificantes de proteínas

A comparação das sequências organelares transcritas com sequências já conhecidas armazenadas em base de dados permitiu identificar 58 mRNA de cloroplastos e 323 mRNA de mitocôndrias. Entre as espécies que apareceram como Top-Hits do Blast estão *Medicago truncatula*, *Vicia faba* (feijão-fava), *Glycine max* (soja), e *Medicago sativa*, espécies da família Leguminosae (Figura 3 e 4). Entre as sequências cloroplastidiais identificadas pelo blast, 53 foram classificadas nas categorias funcionais do Gene Ontology, enquanto 5 tiveram similaridade com sequências cujas funções ainda não estão descritas (Figura 5). Entre os transcritos mitocondriais identificados pelo Blast, 281 foram anotados funcionalmente, enquanto 42 não tiveram similaridade com sequências com funções conhecidas (Figura 5).

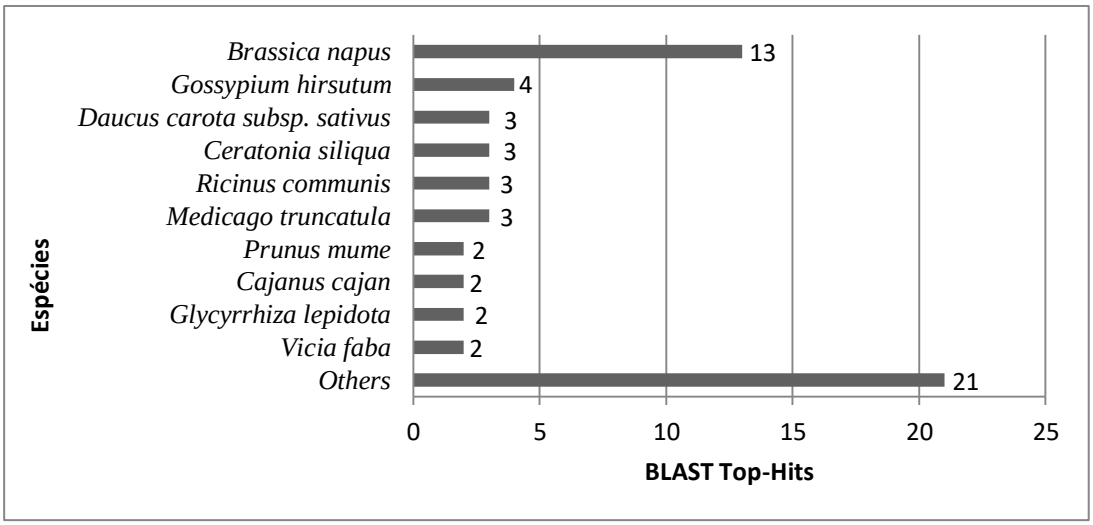

Figura 3 – Distribuição das espécies Top-Hits do Blast das sequências cloroplastidiais.

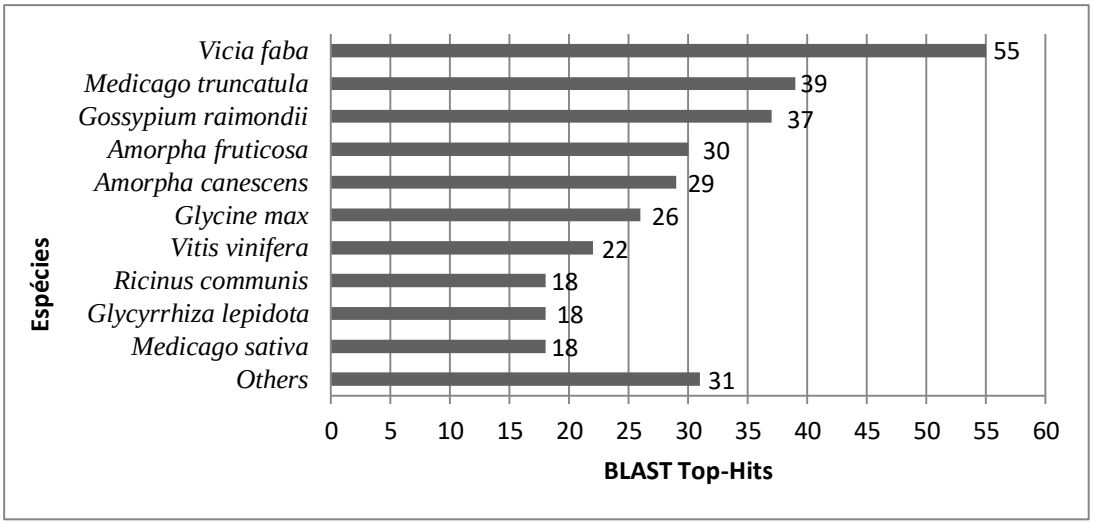

Figura 4 – Distribuição das espécies Top-Hits do Blast das sequências mitocondriais.

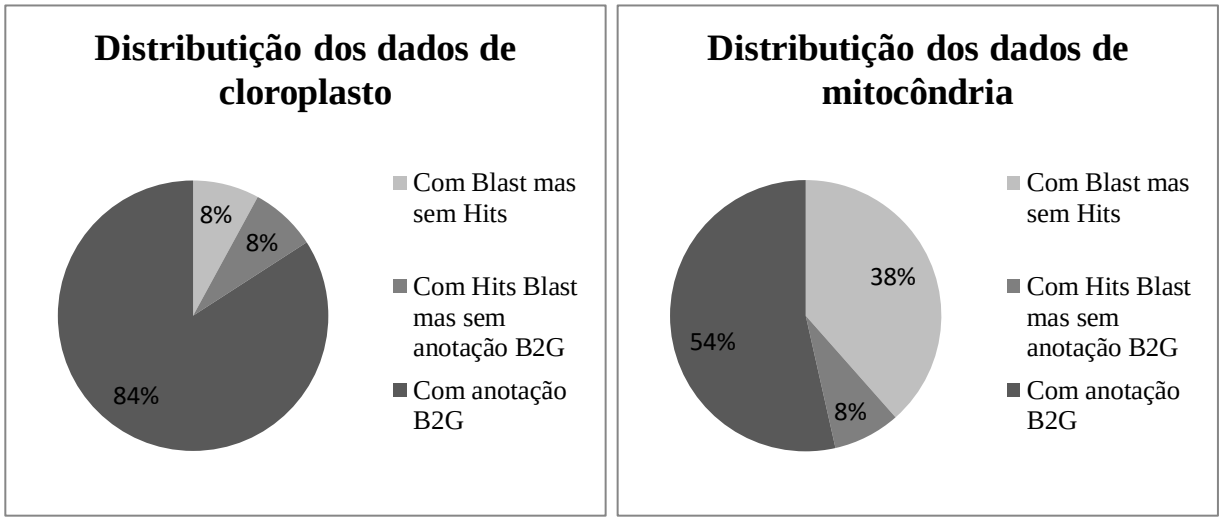

Figura 5 – Distribuição dos dados referentes à análise funcional de 63 sequências cloroplastidiais e 525 sequências mitocondriais. B2G:Blast2GO

Para as sequências de cloroplasto foram obtidas 169 anotações para a categoria processos biológicos, 71 anotações para função molecular e 96 anotações para componentes celulares (Figura 6). Os processos biológicos mais frequentes foram síntese de ATP, processo de oxido-redução e fotossíntese. As funções moleculares mais frequentes foram ligação ao ATP, atividade NADH desidrogenase e transporte de elétrons. Na categoria de componentes celulares a maioria dos produtos de genes foi anotada como componentes de membrana, componentes dos fotossistemas I e II e componentes dos tilacóides dos cloroplastos. Alguns produtos de genes cloroplastidiais foram anotados com localização mitocondrial.

Para as sequências mitocondriais foram obtidas 759 anotações para a categoria processos biológicos, 327 anotações para função molecular e 537 anotações para componentes celulares (Figura 7). Os processos biológicos mais frequentes foram síntese de

ATP, processo de oxido-redução e respiração aeróbica. As funções moleculares mais frequentes foram NADH desidrogenase, atividade citocromo c oxidase e ligação a ATP. Na categoria componentes celulares a maioria dos produtos de genes foi anotada com localização mitocondrial (cadeia respiratória), ribossomal e como componente de membrana. Alguns produtos de genes mitocondriais foram anotados com localização cloroplastidial.

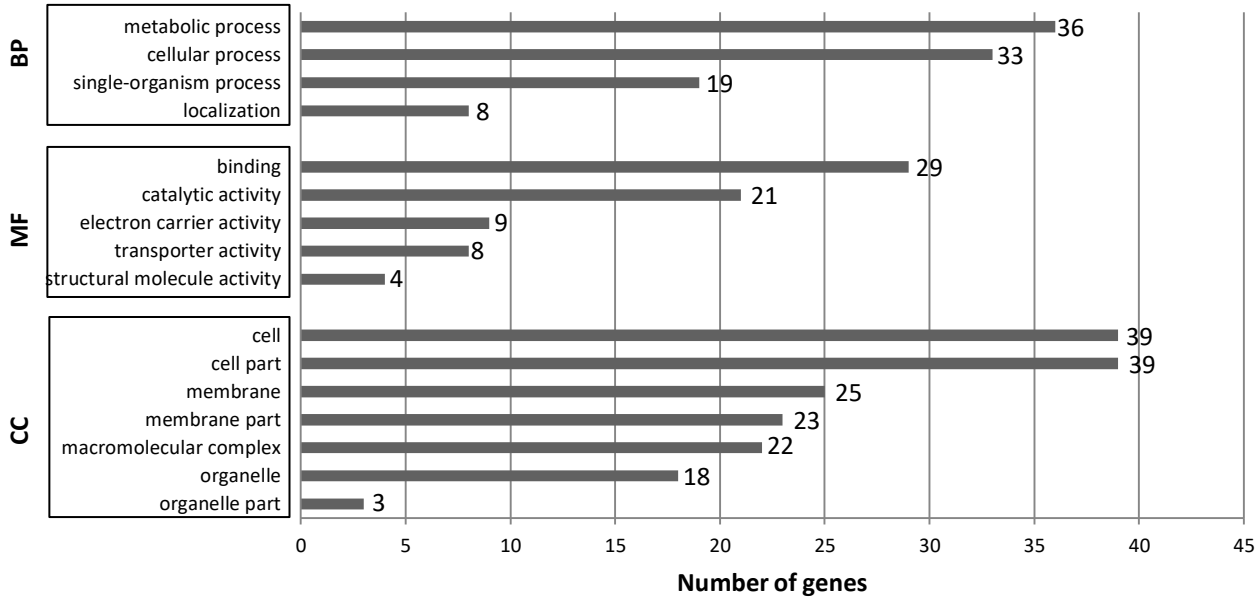

Figura 6 – Anotação funcional de sequências gênicas cloroplastidiais codificantes de proteínas de *Dipteryx alata.* BP: processos biológicos, MF: função molecular, CC: componentes celulares.

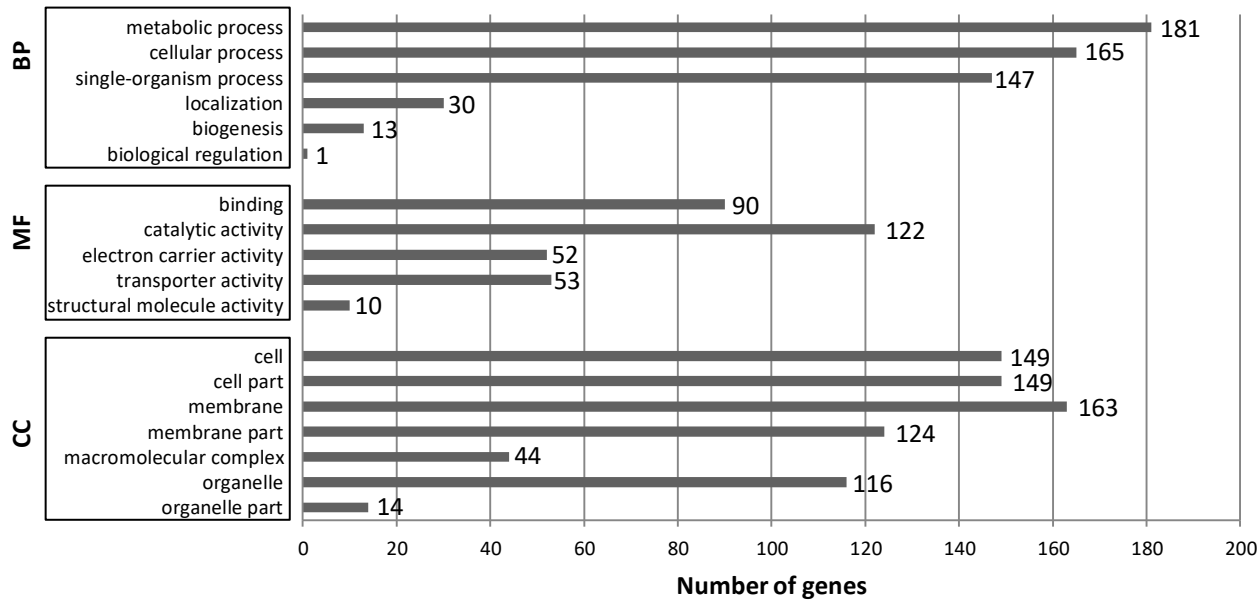

Figura 7 – Anotação funcional de sequências gênicas mitocondriais codificantes de proteínas de *Dipteryx alata.* BP: processos biológicos, MF: função molecular, CC: componentes celulares.

Anotação de microssatélites

Um total de 20 sequências de cloroplastos e 86 sequências de mitocôndrias possuem regiões microssatélites. Em cloroplasto o número mínimo de microssatélites por sequência é 1 e o número máximo é 13. Em mitocôndria o número mínimo de microssatélites por sequência foi 1 e o número máximo foi 9. Nas sequências de cloroplasto foram identificadas 39 e nas sequências de mitocôndrias foram identificadas 158 regiões microssatélites. Em ambas as organelas o tamanho de motivo de repetição mais comum foi o dinucleotídeos e o motivo de repetição mais frequente foi a AT (Tabela 4 e Tabela 5). Os motivos microssatélites foram repetidos em tandem entre 3 e 43 vezes. O total de microssatélites identificados ocuparam 420 pb e 1358 pb nas sequências de cloroplasto e mitocôndria, respectivamente.

Tabela 4 - Quantidade de microssatélites encontrados em sequências organelares de *Dipteryx alata* por tamanho dos motivos de repetição.

Tamanho do motivo repetição	**Cloroplasto**		**Mitocôndria**	
	Frequência Absoluta	**Frequência Relativa (%)**	**Frequência Absoluta**	**Frequência Relativa (%)**
Dinucleotídeos	35	90	111	71
Trinucleotídeos	3	8	10	6
Tetranucleotídeos	1	2	35	22
Pentanucleotídeos	0	0	2	1
Hexanucleotídeos	0	0	0	0
Total	39	100	158	100

Tabela 5 - Motivos de repetição mais frequentes em sequências organelares de *D. alata*.

Cloroplasto			**Mitocôndria**		
Motivo	**Frequência absoluta**	**Frequência relativa (%)**	**Motivo**	**Frequência absoluta**	**Frequência relativa (%)**
AT	21	53.84	AT	49	31.01
CT	9	23.07	AG	27	17.08
AG	4	10.25	CT	22	13.92
GAA	2	5.12	AC	6	3.79
TG	1	2.56	AATA	4	2.53
CTT	1	2.56	CG	3	1.89
TTTA	1	2.56	GT	3	1.89
Total	39	100	AAAG	3	1.89
			AAG	2	1.26
			GGTT	2	1.26
			TCTT	2	1.26
			Outros	35	22.15
			Total	158	100

Anotação de elementos transponíveis

Um total de 0.28% e 1.09% das sequências de cloroplasto e mitocôndria foram mascaradas pelo RepeatMasker, o que correspondeu a 297 pb e 3464 pb, respectivamente (Tabela 6). Os retrotransposons LINE foram os TEs mais abundantes nas sequências de cloroplasto, enquanto os retrotransposons LTR (*gypsy*/DIRS1 e Ty1/*copia*) foram os TEs mais abundantes nas sequências de mitocôndrias.

Tabela 6 – Anotação de TEs em sequências organelares de *D. alata* usando RepeatMasker.

	Ordem	Superfamília	Número de elementos	Porcentagem de elementos (%)	Sequências ocupadas (bp)	Percentual de sequências no genoma (%)
Cloroplasto						
Retroelements	LTR elementos	*Copia*	1	25	144	0.134
	LINE	LINE/L1	2	50	96	0.089
Rolling-circles		Helitron	1	25	57	0.053
Total			4	100	297	0.276
Mitocôndria						
Retroelements	LTR elementos	Copia	7	26.923	1206	0.378
		Gypsy	11	42.307	1777	0.557
	LINE	LINE/L1	2	7.692	94	0.029
DNA transposons	TIR	CMC-EnSpm	3	11.538	193	0.060
		TcMar-Stowaway	1	3.846	87	0.027
Rolling-circles		Helitron	2	7.692	107	0.033
Total			26	100	3464	1.086

DISCUSSÃO

O sequenciamento usando a plataforma Illumina MiSeq permitiu gerar quantidade significativa de reads de DNA total (22 GB de dados) a partir da qual foram separados os reads organelares, por meio de similaridade com sequências de outras espécies depositadas em banco de dados de cloroplasto e mitocôndria. Esse procedimento simples e rápido, aplicado com sucesso no presente estudo, foi usado eficientemente também em outros trabalhos disponíveis na literatura (Zhang et al 2011; Nock et al 2011). A melhor montagem dos reads organelares isolados foi obtida com o software Spades. Os três softwares testados usaram como estratégia de montagem os grafos de bruijn, no entanto cada montador possui

métodos diferentes, por exemplo, para lidar com erros de sequenciamento e com sequências repetitivas (Miller et al 2010). O montador Spades possui particularidades que permitiram gerar o menor número de scaffolds e o maior N50, e por isso foi considerado o montador mais eficiente para o conjunto de dados organelares de *D. alata*.

A quantidade de sequências montadas para o cloroplasto (110 KB) e a mitocôndria (327 KB) de *D. alata* é semelhante a genomas publicados para outras espécies da família Leguminosae. O genoma cloroplastidial de *Medicago truncatula* possui 124 KB (Shaver et al 2008), de *Cicer arietinum* possui 125 KB (Jansen et al 2008), de *Phaseolus vulgaris* possui 150 KB (Guo et al 2007), de *Lupinus luteus* possui 151 KB (Martin et al 2014) e de *Glycine max* possui 152 KB de sequências (Saski et al 2005). O genoma mitocondrial de *Glycine max* possui 402 KB (Chang et al 2013), de *Vigna radiata* possui 401 KB (Alverson et al 2011) e de *Lotus japonicus* possui 380 KB de sequências (Kazakoff et al 2012). Dessa forma é possível sugerir que grande percentual dos genomas organelares de *D. alata* foram amostrados nesse estudo. Muitas investigações biológicas importantes podem ser realizadas sem o sequenciamento da versão completa dos genomas. Para *D. alata* as sequências organelares permitiram a anotação de genes e elementos repetitivos.

A quantidade de genes não traduzidos em proteínas anotados nas sequências cloroplastidiais (3 genes de rRNA e 13 genes de tRNA) e mitocondriais (4 genes de rRNA e 26 genes de tRNA) de *D. alata* é semelhante ao encontrado na literatura para outras espécies da mesma família. O genoma do cloroplasto de *Lupinus luteus* possui 4 genes de rRNA e 30 genes de tRNA (Martin et al 2014). O genoma cloroplastidial de *Cicer arietinum* possui 4 genes de rRNA e 29 genes de tRNA (Jansen et al 2008), enquanto o genoma de *Phaseolus vulgaris* possui 4 genes de rRNA e 30 genes de tRNA (Guo et al 2007). No genoma da mitocôndria de *Glicine max* foram anotados 3 genes para rRNA e 19 genes para tRNA e no genoma de *Vigna radiata* foram identificados 3 genes para rRNA e 16 genes para tRNA (Alverson et al 2011; Chang et al 2013). A quantidade semelhante de genes não traduzidos pode ser justificada pelo elevado grau de conservação das sequências não traduzidas nos genomas organelares.

A quantidade de genes cloroplastidiais codificantes de proteínas (20 genes) anotados em *D. alata*, por outro lado, é menor que o encontrado em outras espécies de Leguminosae. O cloroplasto de *Glycine max* possui 111 genes codificadores de proteínas (Chang et al 2013), de *Medicago truncatula* possui 109 genes (Young et al 2011, Tang et al 2014), de *Lupinus luteus* possui 77 genes (Martin et al 2014), de *Cicer arietinum* e *Phaseolus vulgaris* possuem 75 genes (Guo et al 2007; Jansen et al 2008). O número reduzido de genes cloroplastidiais pode estar relacionado a perdas gênicas durante a história evolutiva de uma espécie, o que é relatado em angiospermas (Guo et al 2007; Qian et al 2013; Asif et al 2013; Xu et al 2015; Yi et al 2016). No entanto, é mais provável que em *D. alata* o número reduzido de genes esteja associado à amostragem parcial do genoma cloroplastidial.

Ao contrário do observado no genoma do cloroplasto de *D. alata*, o número de genes traduzidos em proteínas (176 genes) preditos nas sequências mitocondriais é significativamente maior que o encontrado em outras espécies da mesma família. Foram anotados no genoma de *Glicine max* 110 genes codificadores de proteínas (Chang et al 2013), no genoma de *Lotus japonicus* 65 genes (Kazakoff et al 2012) e no genoma de *Vigna radiata*

31 genes (Alverson et al 2011). O número elevado de genes no genoma mitocondrial pode estar relacionado à predição incorreta devido à montagem parcial do genoma. A super estimativa ocorre porque um único gene é fragmentado em múltiplos scaffolds e por isso identificado como genes separados (Denton et al 2014). No entanto, a duplicação de genes e a transferência de material genético entre os genomas organelares também podem contribuir para o elevado número de genes identificado no genoma mitocondrial de *D. alata* (Asif et al 2013; Hoang et al 2015).

Os resultados obtidos no presente estudo permitem sugerir a transferência de material genético do cloroplasto para a mitocôndria de *D. alata*, visto que nas sequências cloroplastidiais foram identificados baixo número de genes traduzidos e nas sequências mitocondriais esse número foi elevado. Além disso, alguns produtos de genes cloroplastidiais foram anotados com localização mitocondrial e alguns produtos de genes mitocondriais foram anotados com localização cloroplastidial, resultado que contribui para corroborar a hipótese de transferência de material genético entre as organelas de *D. alata*. Vários estudos relatam eventos de transferência de DNA entre organelas e sugerem que na maioria dos casos a transferência é unidirecional do cloroplasto para a mitocôndria (Chang et al 2013; Hoang et al 2015). Pode ocorrer também a transferência de material genético do núcleo para as organelas. Dessa forma, o genoma mitocondrial de plantas pode ser expandido através da integração de DNA cloroplastidial e/ou nuclerar. A transferência de genes ocorre por meio de RNAs intermediários e os genes integrados podem ser expressos pela maquinaria de transcrição mitocondrial (Xiong et al 2008). Esforços futuros de sequenciamento permitirão a montagem do genoma cloroplastidial e mitocondrial completo de *D. alata* e a confirmação do número de genes dessas organelas.

A maioria dos mRNAs codificados pelos genes cloroplastidiais montados para *D. alata* (92% dos mRNAs) foram semelhantes a sequências de outras espécies da família Leguminosae, resultado esperado visto que grupos taxonômicos próximos geralmente apresentam alta similaridade genômica (Young et al 2012). Para a mitocôndria 38% dos mRNAs identificados para *D. alata* não tiveram similaridade com sequências de proteínas anotadas para outras espécies, o que pode ser justificado pela natureza menos conservada do genoma da mitocôndria. Para ambas as organelas de *D. alata,* cloroplasto e mitocôndria, foram anotados funcionalmente genes envolvidos em inúmeras vias bioquímicas e processos biológicos. Os genes cloroplastidiais anotados estão diretamente relacionados com as funções desempenhadas por essa organela, que segundo a literatura possui toda a maquinaria enzimática para a fotossíntese e para a biossíntese de ácidos graxos, aminoácidos, pigmentos e vitaminas (Saski et al, 2005; Shetty et al 2016). O mesmo foi verificado para a mitocôndria de *D. alata*, onde os genes anotados estão relacionados à produção de ATP através da fosforilação oxidativa, um papel executado pelas mitocôndrias que é fundamental no metabolismo dos seres vivos (Wang et al 2012). A anotação em categorias funcionais é importante, pois permite descrever a identidade biológica de um produto de gene (Berardini et al 2004). As sequências codificadoras de proteínas identificadas pelo Blast para *D. alata* e não anotadas em categorias funcionais (8.6% dos mRNAs do cloroplasto e 13% dos mRNAs mitocondriais) provavelmente não possuem domínios funcionais conservados ((Delsenya et al 2010; Shi et al 2011; Hou et al 2011; Liu et al 2012).

Além dos genes foram anotados também elementos repetitivos nas sequências de *D. alata*. O número de elementos repetitivos é maior no genoma mitocondrial (158 regiões microssatélites e 26 TEs) do que no genoma cloroplastidial (39 regiões microssatélites e 4 TEs). Os genomas mitocondriais são variáveis em tamanho e estrutura e acumulam muitas sequências repetitivas (Xiong el al 2008; Zhang et al 2011; Chang et al 2013). Em contrapartida os genomas cloroplastidiais são semelhantes entre as espécies de plantas e possuem baixo conteúdo repetitivo (Saski et al 2005). A maior quantidade de elementos repetitivos justifica o genoma mitocondrial ser maior que o cloroplastidial, visto que as sequências repetitivas são as principais responsáveis pelo aumento no tamanho dos genomas (Mehrotra et al 2014). Em ambas as organelas de *D. alata* o tamanho de motivo de repetição microssatélite mais comum foi o dinucleotídeos e o motivo de repetição mais frequente foi a AT. Segunda a literatura os microssatélites com motivos de repetição mais curtos são mais frequentes e o motivo AT é comum em muitas espécies de plantas (Yu et al, 2011; Zalapa et al 2012). Os retroelementos foram os TEs mais abundantes nos genomas de cloroplasto e mitocôndria de *D. alata* e são os mais frequentes também nos genomas de outras espécies da família Leguminosae (Schmutz et al 2010; Young et al 2011; Schmutz et al 2014).

As sequências de genes e elementos repetitivos organelares obtidas para *D. alata* representam uma adição valiosa a informação genômica dessa espécie do Cerrado. Essas sequências podem ser usadas para o futuro desenvolvimento de marcadores moleculares úteis em estudos de diversidade genética, filogenia, sistemática molecular e evolução (Guo et al 2007; Xu et al 2015; Shetty et al 2016; Yi et al 2016). Nesse sentido, esse estudo inicial abre caminho para novas pesquisas genômicas de *D. alata*.

CONCLUSÃO

A estratégia de sequenciar o DNA total de *D. alata* usando a plataforma Illumina e isolar os reads organelares usando similaridade com sequências de outras espécies depositadas em banco de dados de cloroplasto e mitocôndria foi considerada eficiente. Essa abordagem permitiu montar quantidade de sequências suficiente para identificar e anotar um grande número de genes e elementos repetitivos para essa espécie do Cerrado. As sequências de genes, microssatélites e elementos transponíveis organelares obtidas para *D. alata* representam uma adição valiosa a informação genômica dessa espécie do Cerrado. Essas sequências poderão ser usadas para o futuro desenvolvimento de marcadores moleculares úteis em estudos de diversidade genética, filogenia, sistemática molecular e evolução. A natureza conservada, mas com taxa de mutação suficiente para captar eventos evolutivos, fazem das sequências organelares, principalmente das cloroplastidiais, ferramentas adequadas e de valor inestimável para diversos estudos.

MATERIAL SUPLEMENTAR

https://www.dropbox.com/sh/i6u8roglrx5dusw/AACVy6CoglMGYSyDZj8J9NEPa?dl=0

REFERÊNCIAS

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Asif H; Khan A; Iqbal A; Khan IA; Heinze B; Azim MK. (2013) The chloroplast genome sequence of *Syzygium cumini* (L.) and its relationship with other angiosperms. Tree Genetics & Genomes, 9:867–877.

Altschul SF; Gish W; Miller W; Myers EW; Lipman DJ. (1990) Basic Local Alignment Seash Tool. L. Mol. Biol, 215:403-410.

Alverson AJ; Zhuo S; Rice DW; Sloan DB; Palmer JD. (2011) The Mitochondrial Genome of the Legume *Vigna radiata* and the Analysis of Recombination across Short Mitochondrial Repeats. Plos One, 6.

Bankevich A; Nurk S; Antipov D; Gurevich AA; Dvorkin M; Kulikov AS; Lesin VM; Nikolenko SI; Pham S; Prjibelski AD; Pyshkin AV; Sirotkin AV; Vyahhi N; Tesler G; Alekseyev MA; Pevzner PA. (2012) SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol. 19(5):455–477.

Berardini TZ; Mundodi Z; Reiser L; Huala E; Garcia-Hernandez M; Zhang P; Mueller LA; Yoon J; Doyle A; Lander G; Moseyko N; Yoo D; Xu I; Zoeckler B; Montoya M; Miller N; Weems D; Rhee SY. (2004) Functional Annotation of the *Arabidopsis* Genome Using Controlled Vocabularies. Plant Physiology, 135:745–755.

Bolger AM; Lohse M; Usadel B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 1.

Chang S; Wang Y; Lu J; Gai J; Li J; Chu P; Guan R; Zhao T. (2013) The Mitochondrial Genome of *Soybean* Reveals Complex Genome Structures and Gene Evolution at Intercellular and Phylogenetic Levels. PLOS ONE, 8.

Conesa A; Götz S; García-Gómez JM; Terol J; Talón M; Robles M. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics Applications Note, 21(18).

Delsenya M; Hanb B; Hsingc YI. (2010) High throughput DNA sequencing: The new sequencing revolution. Plant Science, 179:407–422.

Denton D; Lugo-Martinez J; Tucker AE; Schrider DR; Warren WC; Hahn MW. (2014) Extensive Error in the Number of Genes Inferred from Draft Genome Assemblies. PLOS Computational Biology, 10.

Doyle JJ; Doyle JL. (1987). Isolation of plant DNA from fresh tissue. Focus (San Francisco, Calif.), 12:13-15.

Feschotte C; Jiang N; Wessler SR. (2002) Plant transposable elements: where genetics meets genomics. Nature Reviews Genetics, 3.

Foulongne-Oriol M; Murat M; Castanera R; Ramírez L; Sonnenberg ASM. (2013) Genome-wide survey of repetitive DNA elements in the button mushroom *Agaricus bisporus* Fungal Genetics and Biology, 55:6–21.

Guo X; Castillo-Ramírez S; González V; Bustos P; Fernández-Vázquez JL; Santamaría RI; Arellano J; Cevallos MA; Dávila G. (2007) Rapid evolutionary change of common bean (*Phaseolus vulgaris* L) plastome, and the genomic diversification of legume chloroplasts. BMC Genomics, 8:228.

Hoang NV; Furtado A; McQualter RB; Henry RJ. (2015) Next generation sequencing of total DNA from sugarcane provides no evidence for chloroplast heteroplasmy. New Negatives in Plant Science 1(2):p.33–45.

Hou R; Bao Z; Wang S; Su H; Li Y; Du H; Hu J; Wang S; Hu X. (2011) Transcriptome Sequencing and De Novo Analysis for Yesso Scallop (*Patinopecten yessoensis*) Using 454 GS FLX. PLoS ONE, 6.

Jansen RK; Wojciechowski MF; Sanniyasic E; Leec SB; Daniell H. (2008) Complete plastid genome sequence of the chickpea (*Cicer arietinum*) and the phylogenetic distribution of rps12 and clpP intron losses among legumes (Leguminosae). Mol Phylogenet Evol. 48(3):1204–1217.

Jeong YM; Chung WH; Munc JH; Kimb N; Yu HJ. (2014) De novo assembly and characterization of the complete chloroplast genome of radish (*Raphanus sativus* L.). Gene, 551:.39–48.

Kazakoff SH; Imelfort M; Edwards D; Koehorst J; Biswas B; Batley J; Scott PT; Gresshoff PM. (2012) Capturing the biofuel wellhead and powerhouse: the chloroplast and mitochondrial genomes of the leguminous feedstock tree *Pongamia pinnata*. PLoS One, 7(12).

Kim TS; Lee JR; Raveendar S; Lee GA; Jeon YA; Lee HS; Ma KH; Lee SY; Chung JW. (2016) Complete chloroplast genome sequence of *Capsicum baccatum* var. baccatum. Mol Breeding, 36:110.

Lagesen K; Hallin P; Rødland EA; Stærfeldt HH; Rognes T; Ussery DW. (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Research, 35(9).

Langmead B; Trapnell C; Pop M; Salzberg SL. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology, 10.

Liu M; Qiao G; Jiang J; Yang H; Xie L; Xie J; Zhuo R. (2012) Transcriptome Sequencing and De Novo Analysis for Ma Bamboo (*Dendrocalamus latiflorus* Munro) Using the Illumina Platform. PLoS ONE, 7(10).

Lowe TM; Eddy SR. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Research, 25(5).

Luo et al (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience, 1:18.

Mardis ER. (2008) Next-Generation DNA Sequencing Methods. Annu. Rev. Genomics Hum. Genet., 9:387–402.

Martin GE; Rousseau-Gueutin M; Cordonnier S; Lima O; Michon-Coudouel S; Naquin D; Carvalho JF; Ainouche M; Salmon A; Ainouche A. (2014) The first complete chloroplast genome of the Genistoid legume *Lupinus luteus*: evidence for a novel major lineage-specific rearrangement and new insights regarding plastome evolution in the legume family. Annals of Botany, 113:1197–1210.

Meglécz E; Costedoat C; Dubut V; Gilles A; Malausa T; Pech N; Martin JF. (2010) QDD: a user-friendly program to select microsatellite markers and design primers from large sequencing projects. Bioinformatics Applications Note, 26(3):403-404.

Mehrotra S; Goyal V. (2014) Repetitive Sequences in Plant Nuclear DNA: Types, Distribution, Evolution and Function. Genomics Proteomics Bioinformatics, 12:164–171.

Metzker M. (2010) Sequencing technologies — the next generation. Nature Reviews Genetics, v.11.

Miller JR; Koren S; Sutton G. (2010) Assembly algorithms for next-generation sequencing data. Genomics 95:315–327.

Ni L; Zhao Z; Xua H; Chen S; Dorje G. (2016) The complete chloroplast genome of *Gentiana straminea* (Gentianaceae), an endemic species to the Sino-Himalayan subregion. Gene, 577:281–288.

Nock CJ; Waters DL; Edwards MA; Bowen SG; Rice N; Cordeiro GM; Henry RJ. (2011) Chloroplast genome sequences from total DNA for plant identification. Plant biotechnology journal, 9:328-333.

Prabhudas SK; Prayaga S; Madasamy P; Natarajan P. (2016) Shallow hole Genome Sequencing for the Assembly of Complete Chloroplast Genome Sequence of *Arachis hypogaea* L. Frontiers in Plant Science, 7.

Qian J; Song J; Gao H; Zhu Y; Xu J; Pang X; Yao H; Sun C; Li X; Li C; Liu J; Xu H; Chen S. (2013) The Complete Chloroplast Genome Sequence of the Medicinal Plant Salvia miltiorrhiza. Plos One, 8.

Sano SM; Ribeiro JF; Brito MA. (2004) Baru: biologia e uso. Planaltina, DF: Embrapa Cerrados, 52p.

Saski C; Lee SB; Daniell H; Wood TC; Tomkins J; Kim HG; Jansen RK. (2005) Complete chloroplast genome sequence of *Glycine max* and comparative analyses with other legume genomes. Plant Molecular Biology, 59:309–322.

Schmutz J et al (2010) Genome sequence of the palaeopolyploid soybean. Nature, 463:178-183.

Schmutz J et al (2014) A reference genome for common bean and genome-wide analysis of dual domestications. Nature Gnetics, 46(7).

Shaver JM; Oldenburg DJ; Bendich AJ. (2008) The Structure of Chloroplast DNA Molecules and the Effects of Light on the Amount of Chloroplast DNA during Development in *Medicago truncatula*. Plant Physiology, 146:1064–107.

Shetty SM; Shah MU; Makale K; Mohd-Yusuf Y; Khalid N; Othman RY. (2016) Complete Chloroplast Genome Sequence of *Musa balbisiana* Corroborates Structural Heterogeneity of Inverted Repeats in Wild Progenitors of Cultivated Bananas and Plantains. The plant genome, 9(2).

Shi CY; Yang H; Wei1 CL; Yu O; Zhang ZZ; Jiang CJ; Sun J; Li YY; Chen Q; Xia T; Wan XC. (2011) Deep sequencing of the *Camellia sinensis* transcriptome revealed candidate genes for major metabolic pathways of tea-specific compounds. BMC Genomics, 12:131.

Simpson JT; Wong K; Jackman SD; Schein JE; Jones SJM; Birol I.(2009) ABySS: A parallel assembler for short read sequence data. Genome Res. 19(6):1117–1123.

Soares TN; Chaves LJ; Telles MPC; Diniz-Filho JAF; Resende LV. (2008) Distribuição espacial da variabilidade genética intrapopulacional de Dipteryx alata. Pesq. agropec. bras., 43(9):1151-1158.

Stanke M; Morgenstern B. (2005) AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Research, 33.

Tang H; Krishnakumar V; Bidwell S; Rosen B; Chan A; Zhou S; Gentzbittel L; Childs KL; Yandell M; Gundlach H; Mayer KFX; Schwartz DC; Town CD. (2014) An improved genome release (version Mt4.0) for the model legume *Medicago truncatula*. BMC Genomics, 15:312.

Tarailo-Graovac M; Chen N. (2009) Using RepeatMasker to Identify UNIT 4.10. Repetitive Elements in Genomic Sequences. Current Protocols in Bioinformatics.

Wang W; Wu Y; Messing J. (2012) The Mitochondrial Genome of an Aquatic Plant, *Spirodela polyrhiza*. Plos One, 7(10).

Xiong AS; Peng RH; Zhuang J; Gao F; Zhu B; Fu XY; Xue Y; Jin XF; Tian YS; Zhao W; Yao QH. (2008) Gene duplication and transfer events in plant mitochondria genome. Biochemical and Biophysical Research Communications, 376:1–4.

Xu JH; Liu Q; Hua W; Wanga T; Xue Q; Messing J. (2015) Dynamics of chloroplast genomes in green plants. Genomics, 106:221–231.

Yi X; Zhang Z; Ling Y; Xu W; Su Z. (2015) PNRD: a plant non-coding RNA database. Nucleic Acids Research, 43.

Yi DK; Choi K; Joo M; Yang JC; Mustafina FU; Han JS; Son DC; Chang KS; Shin CH; Lee YM. (2016) The complete chloroplast genome sequence of *Abies nephrolepis* (Pinaceae: Abietoideae). Journal of Asia-Pacific Biodiversity, (9):245-249.

Young ND et al (2011) The *Medicago* genome provides insight into the evolution of rhizobial symbioses. Nature, 480.

Young ND; Bharti AK. (2012) Genome-Enabled Insights into Legume Biology. Annu. Rev. Plant Biol, 63:283–305.

Yu JN; Won C; Jun J; Lim YW; Kwak M. (2011) Fast and Cost-Effective Mining of Microsatellite Markers Using NGS Technology: An Example of a Korean Water Deer Hydropotes inermis argyropus. Plos One, 6.

Zalapa JE; Cuevas H; Zhu H; Steffan S; Senalik D; Zeldin E; Mccown B; Harbut R; Simon P. (2012) Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. American Journal of Botany,.99(2):193–208.

Zhang T; Hu S; Zhang G; Pan L; Zhang X; Al-Mssallem IS; Yu J. (2012) The Organelle Genomes of Hassawi Rice (*Oryza sativa* L.) and Its Hybrid in Saudi Arabia: Genome Variation, Rearrangement, and Origins. Plos One, 7.

Zhang T; Zhang X; Hu S; Yu J. (2011) An efficient procedure for plant organellar genome assembly, based on whole genome data from the 454 GS FLX sequencing platform. Plant Methods, 7:38.

Zhang Y; Li L; Yan TL; Liu Q (2014) Complete chloroplast genome sequences of Praxelis (*Eupatorium catarium* Veldkamp), an important invasive species. Gene, 549:58–69.

CONSIDERAÇÕES FINAIS

O tamanho do genoma de *D. alata* foi estimado em 0.825pg e o número cromossômico em 2n=16. Esses valores definiram o genoma de *D. alata* como muito pequeno e permitiu indicar a espécie como potencial candidata dentre as espécies do Cerrado brasileiro para projetos de sequenciamento genômico. O sequenciamento de DNA usando a plataforma Illumina MiSeq permitiu a montagem de 355Mb de sequências nucleares, 110 KB de sequências cloroplastidiais e 327 KB de sequências mitocondriais. O conjunto de sequências montadas permitiu a identificação e caracterização abrangente de regiões microssatélites, elementos transponíveis, regiões gênicas traduzidas e não traduzidas e de polimorfismos de nucleotídeo único (SNPs) em *D. alata*. Os resultados indicam que informações biológicas significativas podem ser obtidas sem a necessidade de sequenciar o genoma completo. A montagem do genoma parcial, estimado em 44% do genoma completo, permitiu gerar novos conhecimentos sobre a estrutura e a organização do genoma dessa espécie do Cerrado. Esse pode ser considerado um dos primeiros estudos genômicos para espécies de árvores nativas do Cerrado brasileiro e os resultados obtidos abrem caminho para novas investigações, como por exemplo, o desenvolvimento de maior número de marcadores microssatélites para uso de genética de populações, desenvolvimento de marcadores baseados em elementos transponíveis para uso de estudos evolutivos e desenvolvimento de marcadores SNPs aplicados a genômica de populações.

APÊNDICE

Manuscript ready for submission:

Trends in genome size and chromosome number evolution in subfamily Papilionoideae (Leguminosae) with emphasis on basal clade ADA

Adriana Maria Antunes[1], Fernando Roa[1], Gustavo Souza[2], Thannya Nascimentos Soares[1], Mariana Pires de Campos Telles[1,3]

[1]Laboratório de Genética & Biodiversidade – Universidade Federal de Goiás (UFG)- Avenida Esperança s/n, Campus Samambaia, CEP 74690-900- Goiânia – GO, Brasil.

[2]Laboratório de Citogenética e Evolução Vegetal – Universidade Federal de Pernambuco - Rua Nelson Chaves S/N- Cidade Universitária, CEP 50670-420 - Recife, PE – Brasil

[3] Departamento de Genética da UFG e Departamento de Biologia da Pontifícia Universidade Católica de Goiás (PUCGO).

Corresponding author:
[1]Mariana Pires de Campos Telles
[1]Laboratório de Genética & Biodiversidade, Instituto de Ciências Biológicas, Universidade Federal de Goiás, Goiânia, Goiás, Brasil. CEP: 74001-970
e-mail: tellesmpc@gmail.com

ACKNOWLEDGMENTS

Our research was supported by the project "Núcleo de Excelência em Recursos Genéticos Vegetais do Cerrado (CERGEN)" - GECER (PRONEX/ Fundação de Amparo á Pesquisa do Estado de Goiás - FAPEG /Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq #CP07-2012), and by several grants and fellowships for the research network GENPAC ("Geographical Genetics and Regional Planning for Natural Resources in Brazilian Cerrado") from CNPq/MCT/ Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES/FAPEG. Research by M.P.C. Telles and T.N. Soares was also supported by productivity grants from CNPq.

ABSTRACT

Genome size (C-value) and chromosome number (2*n* or *n*) present a huge diversity in plants. Despite improvements in methods of study, 1C and 2*n* estimates are available for a limited number of species. This study aimed to investigate the evolution of those parameters in a phylogenetic framework for Leguminosae with focus on the basal Papilionoideae clade ADA (Angylocalyceae + Dipterygeae + Amburaneae) and contribute with new estimates for *Dipteryx alata* and *Amburana cearensis*. A phylogeny for 349 members of Leguminosae was built based on combined *mat*K and *rbc*L plastidial genes and the 5.8S rRNA nuclear gene family, followed by ancestral states character reconstruction for 2*n* and 2C from all these species. The present work estimated the genome size of *D. alata* as 1C = 0.825 pg, which corresponds to a haploid genome of 807.2 MB with 2*n* = 16 and from *A. cearensis* 1C = 0.483 pg (1C = 472.3 MB; 2*n* = 22). These data are similar to the observed in other species of the clade ADA, that present low chromosome numbers and small genomes. The genome size and chromosome number reconstructed on the phylogeny suggest that this is the plesiomorphic condition in Papilionoideae, although several clade specific shifts are observed, like a conspicuous giant genome size increase in Fabeae. The estimated 2n and 1C values for Dipteryx alata and Amburana cearensis characterize these species as family candidates to genome sequencing due to their small genomic sizes and to their basal positions in the clade Papilionoideae.

Keywords: *Amburana cearensis*, C-value, cytogenetics, *Dipteryx alata*, genome model, Leguminosae.

INTRODUCTION

Genome size (1C-value) and chromosome number (2*n* or *n*) exhibit a huge diversity in plants. Genome size, expressed as C-value in picograms (pg) for haploid cells is not strictly correlated with the complexity of organisms (Huang et al 2015; Hidalgo et al 2015). Among the angiosperms, the variation in genome size is about 2400-fold, from 0.06 (*Genlisea margaretae*) to 152.23 pg (*Paris japonica*) (Greilhuber et al 2006; Pellicer et al 2010). The genome size is generally stable within a species, but may differ strongly among even closely phylogenetic related taxa (Nag et al 2011; Pandit et al 2014; Hidalgo et al, 2015). The variation of the chromosome number in plants is also high, ranging from 2n = 4 found in species such as *Haplopappus gracilis* and *Brachyscome dichromosomatica,* up to 2n = 1400 found in the fern species *Ophioglossum reticulatum* (Castiglione & Cremonini 2012; Dyer et al 2013).

Genome size and chromosome number data have been incorporated in the fields of ecology, physiology, plant breeding, conservation, and others (Bennett et al 2000; Hiremath et al 2004; Pandit et al 2014). In practice, knowledge of the genome size is important to estimate the time and cost of whole genome sequencing projects (Garcia et al 2014; Wang et al 2015). In an evolutionary perspective, the measurement of nuclear DNA content and the chromosomes number can be used to elucidate evolutionary patterns and mechanisms of adaptation in plants (Wang et al 2011; Wang et al 2015; Huang et al 2015).

Genome size estimates are available for only ca. 7500 angiosperms, i.e. 2% of the known species (Wang et al 2015), The family Leguminosae is the third largest family of flowering plants, with 24,505 species of 946 genera. Leguminosae is classified, based on phylogenetic data, into six subfamilies: Caesalpinioideae DC., Cercidoideae Legume Phylogeny Working Group (stat. Nov.), Detarioideae and Burmeister., Dialioideae Legume Phylogeny Working Group (stat. Nov.), Duparquetioideae Legume Phylogeny Working Group (stat. Nov.), Papilionoideae DC (Azani et al 2017). The papilionoid legumes (Papilionoideae) with ca.13,800 species, are a legacy of the extraordinarily broad ecological and biogeographical range largely associated with their diversifications in varied ecosystems (Lavin et al 2005; Lewis et al 2005).

Papilionoids are also well known for their unprecedented economic importance to agriculture and the food industry, as they include pulse legume crops. Due to its economic interest, several Papilionoideae are the subject of genomic sequencing projects, for example

Glycine max, *Phaseolus vulgaris*, *Cicer arietinum* e *Medicago truncatula* (Schmutz et al 2010; Varshney et al 2013; Schmutz et al 2014; Tang et al 2014). The potentially earliest-branching papilionoids fall within clade ADA, which includes the recircumscribed monophyletic tribes Angylocalyceae, Dipterygeae, and Amburanae. Angylocalyceae has 21 species belonging to 5 genera, Dipterygeae has 25 species distributed into 4 genera and Amburaneae has 29 species in 7 genera. Angylocalyceae has genera that occur in Africa, Madagascar, Australia and in tropical forests of northern South America. Genera of Dipterygeae and Amburaneae are found in the Neotropical region (Cardoso et al 2013). Among them, there are *Dipteryx*, (which has 12 species) and *Amburana* (which has 2 species) to which the species *Dipteryx alata* Vogel and *Amburana cearensis* (Allemão) A. C. Sm, the focus of this study, belong.

The diversity of chromosome numbers and genome size is also large in subfamily Papilionoideae. Chromosome numbers range from 2n = 8 to 180 (Rice et al 2015), while genome size shows a 80-fold variation from 2C = 0.69 to 54.8 pg (Garcia et al 2014). The present study aimed to estimate the genome size of *Dipteryx alata* and *Amburana cearensis*, ADA clade species, and the number of chromosomes of *Dipteryx alata*, and to reconstruct the evolution of these characters in a phylogenetic context.

MATERIAL AND METHODS

Plant material

For Flow Cytometry, leaves of *Dipteryx alata* and *Amburana cearensis* were collected from three individuals. Leaves of germinated seeds of *Glycine max* cv. 'Polanka' (2C-Value = 2.5 pg), *Pisum sativum* cv. 'Ctirad' (9.76 pg), and *Solanum lycopersicum* cv. 'Stupicke' (1.96 pg), used as internal standards for Flow Cytometry, were obtained from the Institute of Experimental Botany at Olomouc, Czech Republic (Doležel and Greilhuber 2010). Roots of *D. alata* for the chromosome analyses were collected from germinated seeds of the collection.

Flow Cytometry analysis

Leaves were kept on wet paper at low temperature (4-6 °C) over a period of two days before procedure. The estimation of genome size was performed according to the protocol described in Dolezel & Gohde (1995). Small fragments of leaf tissue (40mg) the sample and

standard were co-minced with razor blade in a petri dish containing 1.5 ml of WPB buffer (Tris-HCL 0.02M, MgCl2·6H2O 4mM, Triton X 100 1%, EDTA Na2·2H2O 2mM, NaCl 86mM, sodium metabisulfite 10mM, 2 PVP-10 1%). The samples were fragmented quickly to minimize the release of cytosolic compounds, passed through a 30µm pore filter to remove cell debris and stained with 50µl propidium iodide (1mg/ml). This preparation was repeated in triplicate for each plant and three measurements were carried out on consecutive days.

Samples were analyzed on a Partec CyFlow Space (Partec, Muenster, Germany) equipped with a Diode-pumped solid-state green laser (532 nm). For each sample, peaks of fluorescence were read at least for 2,000 particles and analyzed by using Partec FloMax software (Version 2.4). The sample and standard peaks were manually delimited in the histograms. Samples with a half peak coefficient of variation (HPCV) of more than 5% were not considered for analysis. The nuclear DNA content (2C value) was determined according to the following formula for relative fluorescence: ([average peak of sample / average of standard] x standard peak 2C-value in pg). The C-value was multiplied by 978 to obtain the size of the genome in megabases (MB), since 1 pg is equal to 978 bp.

Determination of the chromosomes number of *D. alata*

Root tips of *D. alata* were pre-treated with 0.002M of 8-hydroxyquinoline at room temperature for 5 h, fixed overnight in ethanol/glacial acetic acid 3:1 (v/v) and stored at -6 °C. Root tips were digested using dissolved 18U/ml cellulase and 18U/ml pectinase for 120 min at 37 °C. The meristem was dissected in acetic acid (45%) and squashed under a coverslip (removed after freezing in liquid nitrogen). Staining was performed with 4',6-diamidino-2-phenylindole (DAPI - 2µg/ml)/glycerol 1:1. Capture was done in a Zeiss Axio Scope A1 equipped with a Cool Cube 1m CCD camera and the software Cometa. The 2n were determined by chromosome counting of cells from at least four individuals.

Frequencies and correlations between 2C and 2*n*

The Kew 2C-value database (Garcia et al 2014) was consulted for the genome size and chromosome number of Leguminosae species, while the Chromosome Counts Database (CCDB) was consulted to obtain the chromosome number of 6,071 species (Rice et al 2015) (Supp. Mat., table S1). Histograms of frequencies and Gaussian kernel estimations of density were performed in R with the functions: geom_histogram, geom_point,

geom_density, while Kendall correlation (non-parametric) was used to assess the correlation between 2C and 2n with the function cor.test. The interactive "online resources", cited along the text and available at http://cyto.shinyapps.io/fabaceae were developed with shiny-R. (Wickham 2009; Wang et al 2013; Chang et al 2016; R Core Team 2016).

Phylogenetic analysis

Phylogenetic relationships were based on combined *mat*K and *rbc*L plastidial genes and 5.8S rDNA nuclear gene family for 349 species of Leguminosae (Supp. Mat., table S2). All species were obtained from Genbank (Supp. Mat., Table S2) and all Leguminosae species having known chromosome number and/or DNA content data were selected. All clades identified were named according to Borges et al (2013). Members of clade Caesalpinoideae were used as outgroups (Borges et al 2013).

The Bayesian Inference (BI) search was performed with combined *mat*K + *rbc*L + ITS sequences using Mr. Bayes v3.1.2 (Ronquist & Huelsenbeck, 2003) under a partitioned model as implemented on CIPRES Science Gateway V.3.1 (www.phylo.org). The most appropriate model of sequence evolution for each region and combined matrix were determined using the Akaike information criterion (AIC) as implemented in the program jModelTest 0.1.1 (Posada, 2008). The analysis was conducted for two independent runs and for 10,000,000 generations, sampling every 1,000 trees. Tracer v 1.0.1 was used to ensure that the Markov chains had reached stationary state and to determine the appropriate number of 'burn-in' runs (Rambaut et al 2014). Posterior probabilities (PP) were estimated by constructing a 50% majority-rule consensus tree with the sampled trees. Trees were visualized using FigTree v1.3.1 (Rambaut, 2009).

Ancestral state reconstruction and inference of chromosome number/genome size changes

For species with more than one ploidy level or chromosome number, only the lowest was used for the reconstruction of the ancestral diploid number (discussed further in results). To reconstruct ancestral diploid chromosome number (2*n*) along the phylogeny we used the maximum likelihood Markov k-state one-parameter (Mk1) model in Mesquite (v.2.75; Maddison & Maddison, 2011). Genome size data were regarded as continuous characters and analyzed under maximum parsimony in Mesquite.

RESULTS

Genome size and chromosome number

The genome size of *D. alata* was calculated as 1C = 0.825 pg corresponding to a haploid genome of 807.2MB (Figure 1a), while *A. cearensis* showed 1C = 0.483 pg (1C = 472.3 MB). *Glycine max* and *Solanum lycopersicum* were selected as reference standards for having nearby genome sizes without overlapping peaks. In all prometaphases analyzed, *D. alata* showed 2n = 16 chromosomes (Figure 1b). *A. cearenses* has 2n = 22 chromosomes (Darlington & Wylie, 1956; Fedorov 1969).

Genome size and chromosome numbers variation in Leguminosae

Dividing the sample of known chromosome numbers of Leguminosae in terciles (each with 33.3%), three groups were defined: low 2*n* (8 – 16), intermediate 2*n* (16 - 22) and high 2*n* (22 - 180). Most clades of Leguminosae and Papilionoideae, even the clade ADA, had species in each category. The distribution of 2*n* is partially discontinuous (multimodal) since polyploidy events can be observed as peaks on a histogram 2*n* (Figure 2). The most frequent numbers (modes) indicate recent polyploidy in Baphieae, Dalbergioids, Genistoids, Hologalegina, Indigofereae and MPA from Papilionoideae. In the Papilionoideae clades, 2*n* = 16 (low-intermediate 2*n*), 2*n* = 14 (low) and 2*n* = 22 (intermediate-high 2*n*) are among the three most frequent numbers for each clade in that order, while for other Leguminosae clades the most common are 2*n* = 28 and 24 (high 2*n*).

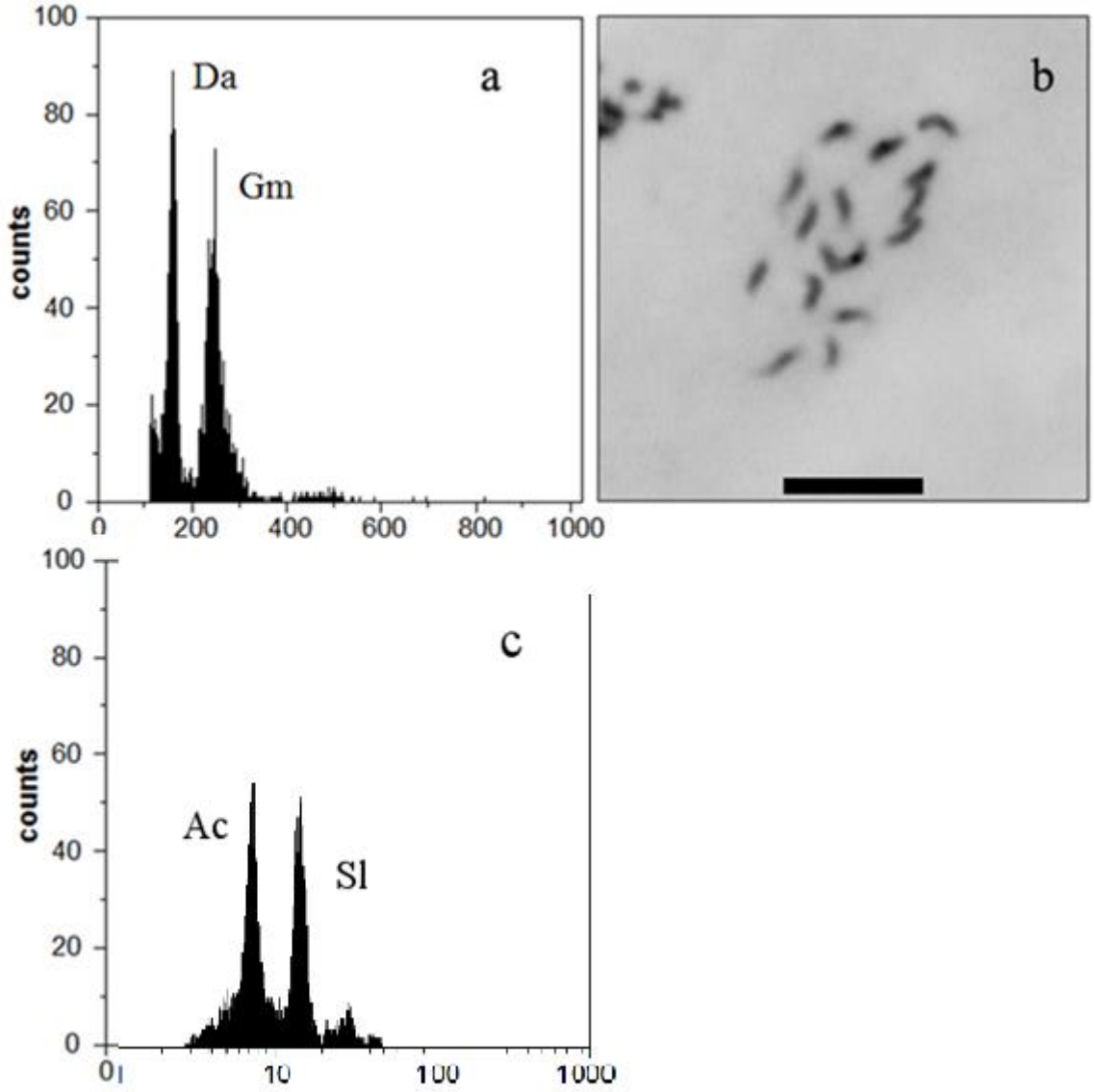

Fig 1 Histograms of relative fluorescence (arbitrary units) of propidium iodide stained nuclei by Flow Cytometry. **Fig1a** Da - *Dipteryx alata*. Gm - *Glycine max*, reference standard; **Fig1b** Prometaphase of *D. alata* 2n=16 stained with DAPI. Inverted color. Bar = 5µm; **Fig1c** Ac – *Amburana cearensis*, Sl – *Solanum lycopersicum*, reference standard

.

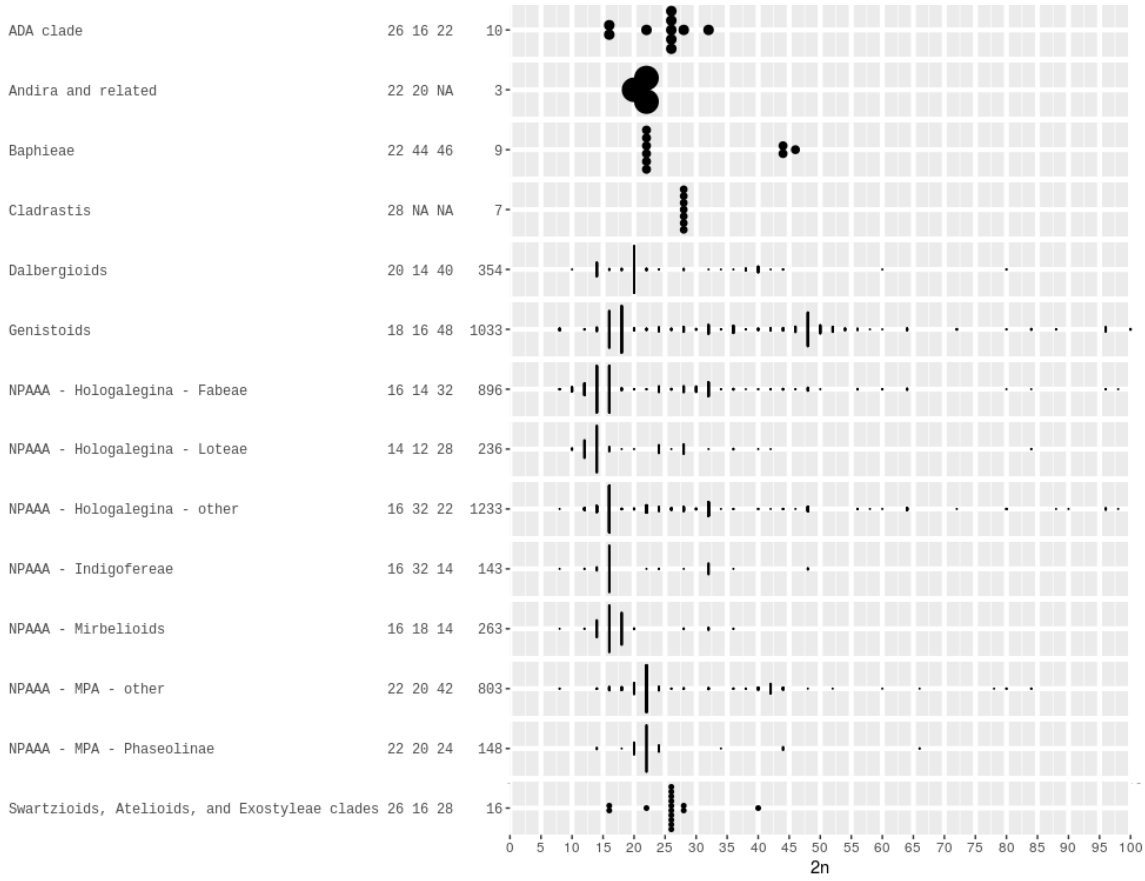

Fig 2 Dot plot distribution of 2n for Papilionoideae major clades; y axis contains name, three most frequent 2n and sample size. Dot size in different scales depending on the sample size.

Most clades of Leguminosae and Papilionoideae had species in all three categories established by dividing the sample of known genome sizes in terciles (low 2C 0.6 – 1.6 pg, intermediate 2C, 1.6 – 3.5 pg, and high 2C, 3.5 – 54.8 pg) Some exceptions are the clade ADA, with low 2C, and other poorly studied clades. The distribution of 2C is multimodal as shown by the Gaussian kernel density peaks for 2C in Figure 3, with the first peak corresponding to the low 2C category for most clades.

Correlation between 2n and 2C-values

At first sight it might seem that, in general, the Leguminosae, or at least the Papilionoideae, showed a restriction for high 2C-values in species with high 2n. This fact was confirmed for rich clades as Fabeae, which showed negative correlation between 2C and 2n. Accordingly, in Figure 4, clade Fabeae with low 2n shows large genomes sizes. However, for Hologalegina-Loteae and clade NPAAA-MPA a positive correlation could be established. In those groups, there is a restriction for low 2C-values in species with high 2n (online resources). According to that, in the phylogeny, the clade Erythrininae (NPAAA-MPA) with intermediate genome sizes (>2 pg) shows high chromosome number (Figure 3). Genistoids and Dalbergioids clades are the best sampled Papilionoideae outside of the NPAAA clade (which includes Hologalegina) and have chromosome numbers mainly in the intermediate and high 2n categories. However, is evident from the phylogeny and the histograms, that Genistoids have a lot more karyotypes with high 2n, but smaller genomes. In contrast, Dalbergiod genomes tend to be larger than genistoids genomes but with smaller 2n, showing opposite trends (Fig 3, online resources).

Ancestrality analysis of 2C-value and chromosome number

Maximum likelihood reconstruction 2C at ancient nodes showed a small genome size for the Papilionoideae ancestor less than ca. 2 pg (blue in figure 4). At some point in the evolution of Fabeae, it gained a huge genome of ca. 12 pg (green in figure 4), while other ancestral nodes showed ca. 4 pg, as in Dalbergioids (light blue). From the reconstruction of n, a process which discards any (recent) ploidy information, the ancestral 2n of Papilionoideae corresponds to an intermediate size karyotype of 2n = 21 to 33 (yellow). Two conspicuous trends are observed in the branches, a small chromosome number in the ancestor of clade Hologalegina, and higher chromosome numbers in the ancestor of other NPAAA clades and the less diversified basal clades, ADA, Genistoids and Dalbergioids.

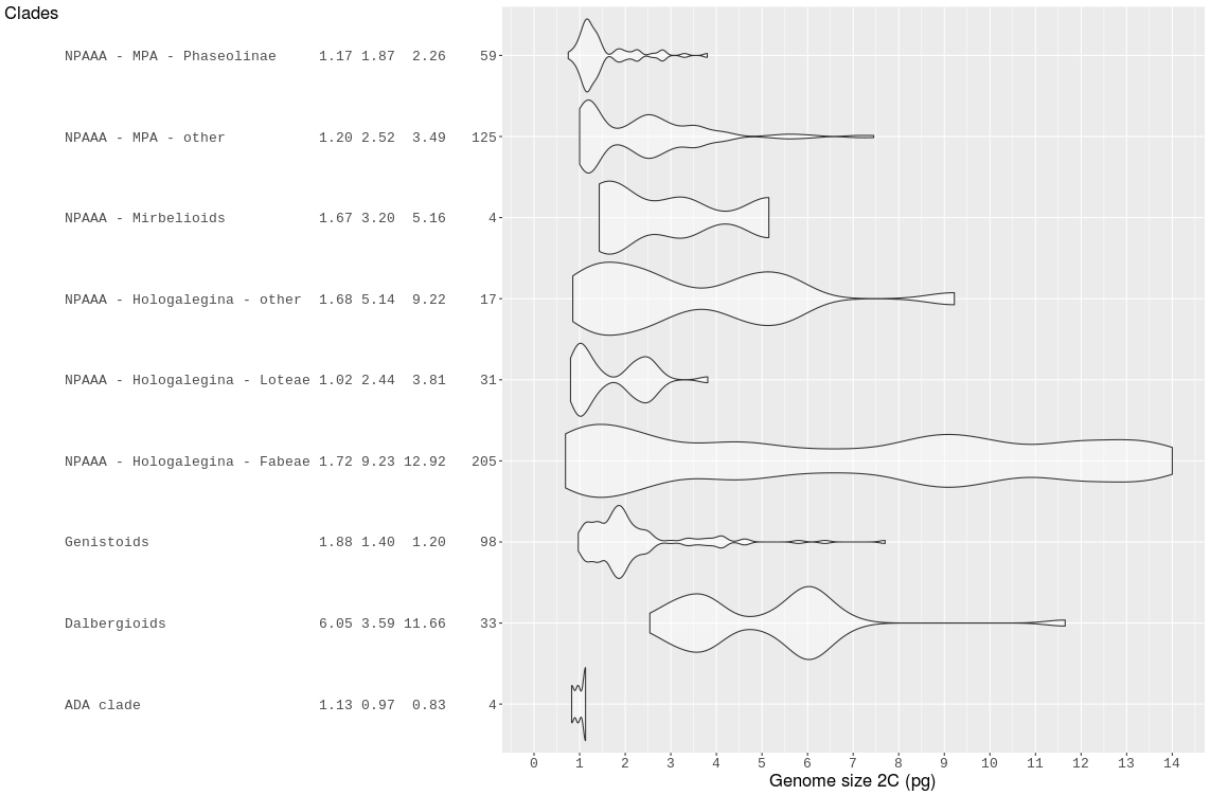

Figure 3. Violin plots showing the density of 2C in Papilionoideae clades. The y-axis contains the three highest peaks in pg and the sample size. Bandwidth of the density function customizable in the online resources.

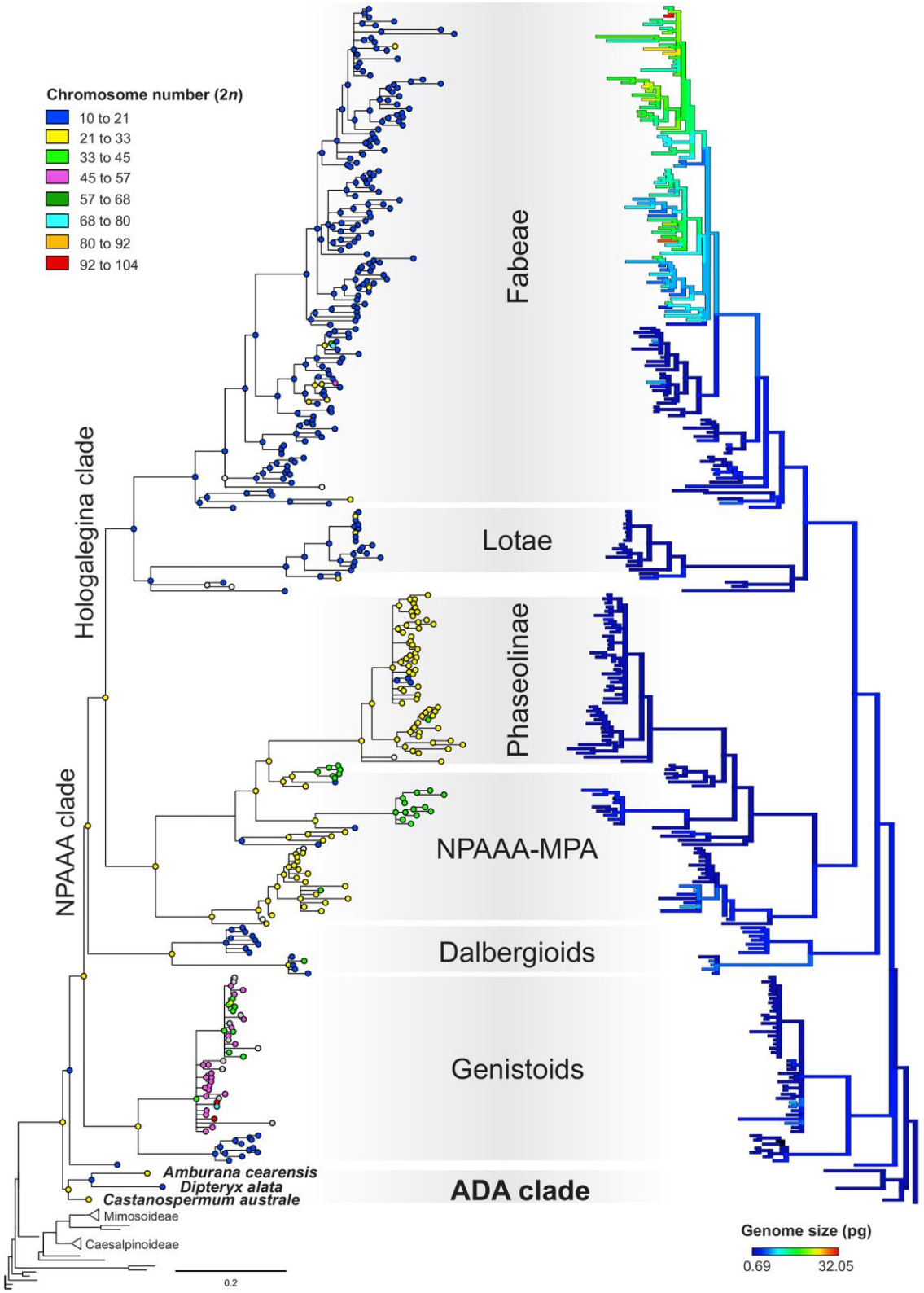

Fig 4 Evolution of genome size and chromosome number in Leguminosae species. Values of 2n and 2C are showed by colors in the ancient nodes and branches.

DISCUSSION

2C and 2n evolution in Leguminosae

Previous analysis of thousands of homolog genes of rosids showed that the ancestor of legumes and other core eudicots was a hexaploid, with probably *n* = 21 chromosomes, which rearranged to produce *n* = 6 Leguminosae ancestor through at least 16 fusions and 1 fission. Subsequently, the ancestor of all Papilionoideae suffered a whole genome duplication (WGD) event presenting *n* = 12 (Cannon et al 2015; Murat et al 2015). Doyle (2012) suggests an ancestor for Papilionoideae with *n* = 14. Historically, frequencies were used to propose the ancestral numbers for taxa, however, currently is more common to consider also the phylogenetic and genomic information available. Heuristic methods are used with more frequency to reconstruct the ancestral numbers, showing a range of ancestral values or a probability for one or more ancestral numbers at nodes (Mayrose et al 2010; Cusimano et al 2012; Peruzzi 2013; Sousa et al 2014; Macas et al 2015). The ancestral range (2n =21 - 33) for Papilionoideae obtained here corroborates the previously analyses, and fits well with the high 2n modes found in the basal clades ADA, Swarzioids, Atelioids and Cladrastis (modes 2n =16 - 28).

Among Leguminosae, duplication events occurred in ancestors of Cercidae and Detarieae, whereas inside Papilionoideae, ancestors of genistoids and Soybean (*Glycine*) a paleopolyploid, underwent WGD as shown by comparison of orthologs segments (Cannon et al 2015; Murat et al 2015). However, there is no evidence of those events in the present genome sizes in Papilionoideae. The reconstruction of ancestral C-values begins by excluding any possible (recent) polyploid value, because it models only continuous or gradual variations.

The dynamics of genome size evolution breaks up the drastic changes in karyotype after a WGD along long periods of time. Apart from WGD, several mechanisms are responsible for the expansion and contraction of genome size. Repetitive elements, especially transposable elements, in particular of retrotransposons are the key drivers of differences in genome size plants, since they can explain up to 85% differences in genome size, as studied in species of the tribe Fabeae (Macas et al 2015). Tandem repeats and segmental duplication of the genome are common in plants and are also positively correlated to the amount of DNA (Bennett et al 2000; Michael et al 2014). On the other hand, reducing the size of the genome may occur through deletions, such as errors in DNA replication and illegitimate or non-homologous recombination (Whitney et al 2010).

The present reconstruction of C-values shows two clades standing clearly for an ancestral increase of genome size, the Dalbergioids and NPAAA-Hologalegina-Fabeae, as explained, independently of WGD events. In addition, it is clear that several events of recent polyploidy are common in the family, especially in several clades of Papilionoideae, as evidenced by the multimodal distribution of 2n, but they have to be transformed or removed in the reconstruction of ancestral 2n as explained for the genome size reconstruction.

Contrasting evolutionary trends of genome size vs. chromosome number were found for clades of Leguminosae. Though it might be expected a priori to see a positive correlation between them, as observed in NPAAA-MPA and Hologalegina clades (excluding Fabeae), several ecological driven traits model those characteristics in special ways (Bennett 1987; Pandit et al 2014; Suda et al 2015). Depending on the habit and invasiveness of the species, its genome adapts not to expand, whereas the chromosome number may vary by polyploidy conferring new raw material for evolution (Pandit et al 2014). Genome size constraint mechanisms, aroused by natural selection, may allow for ecological strategies that might explain the absence large genomes in high 2n species. In the absence of such mechanisms an unlimited expansion of the genome may occur (Kelly et al 2015).

The ADA clade and the candidate species for analysis of the genome of Papilionoideae

The number of chromosomes and the genome size present variable values among the species of the ADA clade. *A. cearensis* has an intermemeary chromosome number (2n = 22) and a small genome size (1C = 0.483 pg). *D. alata* has a low chromosome number (2n = 16) and a small genome (1C = 0.825 pg). The genome sizes of D. alata and A. cearensis are among the smallest in the Leguminosae family, especially after the genus Trifolium (Fabeae), but unlike other genomes with a small genome, are part of a basal clade of Papilionoideae. A very small DNA content may help choosing candidate species for sequencing the genomes. The next generation sequencing (NGS) platforms have allowed the analysis of nucleotide sequences at high performance and reduced cost of pairs of sequenced bases. Despite advances in sequencing technology, genomes of smaller sizes have been prioritized due to lower costs for obtaining the total sequences set. In addition, small genomes facilitate assembly of the sequences, a complex and difficult computational job for large genomes (Kelly et al 2012). In this sense, there is support for placing *D. alata* and *A. cearensis* as Cerrado candidates for genome sequencing studies. The genomic study of these species will

allow comparative analysis with other Papilionoideae that already have a sequenced genome, such as *Glycine max* and *Phaseolus vulgaris*. Understanding the molecular differences between the genomes of the Papilionoideae species is important to elucidate the mechanisms that regulate C-value and chromosome number modifications.

CONCLUSION

This study contributed with new estimates of genome size and number of chromosomes for the Leguminosae family. The sizes of the haploid genomes of D. alata and A. cearensis, from the basal clade ADA, are among the smallest in Leguminosae or Papilionoideae. For those these characteristics, we propose the species as a candidates for genome sequencing projects. This will allow in depth evolutionary analysis by whole genome comparisons with other model species of the family as *Lotus* (Sato et al 2008), soybean (Schmutz et al 2010), chickpea (Varshney et al 2013) and bean (Schmutz et al 2014). Characterizing the evolution of cytogenetic traits of Leguminosae is the initial step before elucidating the mechanisms regulating C-value changes and rearranges that change the chromosome number.

DATA ARCHIVING STATEMENT

The data will be submitted to the Kew 2C-value and Chromosome Counts database.

REFERENCES

Azani, N. et al (2017) A new subfamily classification of the Leguminosae based on a taxonomically comprehensive phylogeny. The Legume Phylogeny Working Group (LPWG). Taxon, 66(1):44-77.

Bennett MD (1987) Variations in genome form in plants and its ecological implications. New Phytol 106:177–200.

Bennett MD, Bhandol P, Leitch, IJ (2000) Nuclear DNA Amounts in Angiosperms and their Modern Uses - 807 New Estimates. Annals of Botany 86:859-909.

Borges L, Bruneau A, Cardoso D, et al (2013) Towards a new classification system for legumes: Progress report from the 6th International Legume Conference. South African J Bot 89:3–9.

Cannon SB, McKain MR, Harkess A, et al (2015) Multiple polyploidy events in the early radiation of nodulating and nonnodulating legumes. Mol Biol Evol 32:193–210.

Cardoso D, Pennington RTT, de Queiroz LPP, et al (2013) Reconstructing the deep-branching relationships of the papilionoid legumes. South African J Bot 89:58–75.

Castiglione MR; Cremonini R. (2012) A fascinating island: 2n = 4. New Trends in Plant Cytogenetics and Cytoembryology, 146:711-726.

Chang W, Cheng J, Allaire JJ, Xie Y, McPherson J (2016) shiny: Web Application Framework for R.

Cusimano N, Sousa A, Renner SS (2012) Maximum likelihood inference implies a high, not a low, ancestral haploid chromosome number in Araceae, with a critique of the bias introduced by "x". Ann Bot 109:681–92.

Darlington CD, Wylie AP. (1956) Chromosome atlas of flowering plants. Chromosome atlas of flowering plants. George Allen and Unwin Ltd., London, UK.

Doležel J, Göhde W. (1995) Sex determination in dioecious plants Melandrium album and M. rubrum using high resolution flow cytometry. Cytometry, v.19, p.103–106.

Doležel J, Greilhuber J (2010) Nuclear genome size: Are we getting closer? Cytom Part A 77:635–642.

Doyle JJ (2012) Polyploidy in Legumes. Polyploidy Genome Evol. Springer Berlin Heidelberg, Berlin, Heidelberg, 47–180

Dyer RJ; Pellicer J; Savolainen V; Leitch IJ; Schneider H. (2013) Genome size expansion and the relationship between nuclear DNA content and spore size in the Asplenium monanthes fern complex (Aspleniaceae). BMC Plant Biology, 13:219.

Fedorov ANA. (1969) Chromosome numbers of flowering plants. Academy of science of theUSSR Komarov. Botanical Institute Leningrad (Reprint, 1974).

Garcia S, Gálvez F, Gras A, Kovařík A, Garnatje T (2014) Plant rDNA database: update and new features. Database (Oxford) 2014:1–7.

Greilhuber J, Borsch T, MAller K, Worberg A, Porembski S, Barthlott W (2006) Smallest angiosperm genomes found in Lentibulariaceae with chromosomes of bacterial size. Plant Biology 8:770-777.

Gunn CR, Service USAR (1983) A nomenclator of legume (Fabaceae) genera. U.S. Dept. of Agriculture, Agricultural Research Service

Hidalgo O, Garcia S, Garnatje T, Mumbru M, Patterson A, Vigo J, Valles J (2015) Genome size in aquatic and wetland plants: fitting with the large genome constraint hypothesis with a few relevant exceptions. Plant Syst Evol 31.

Hinchliff CE, Smith SA, Allman JF, et al (2015) Synthesis of phylogeny and taxonomy into a comprehensive tree of life. Proc Natl Acad Sci 112:12764–12769.

Hiremath SC, Nagasampige MH (2004) Genome size variation and evolution in some species of Dalbergia Linn.f. (Fabaceae). Caryologia 57:367-372.

Huang H, Tong Y, Zhang QJ, Gao LZ (2015) Genome Size Variation among and within Camellia Species by Using Flow Cytometric Analysis. Plos Onde 8.

Kelly LJ, Leitch AR, Fay MF, et al (2012) Why size really matters when sequencing plant genomes. Plant Ecol Divers 5:415–425.

Kelly LJ, Renny-Byfield S, Pellicer J, et al (2015) Analysis of the giant genomes of Fritillaria (Liliaceae) indicates that a lack of DNA removal characterizes extreme expansions in genome size. New Phytol 208:596–607.

Lavin M, Herendeen P, Wojciechowski M (2005) Evolutionary Rates Analysis of Leguminosae Implicates a Rapid Diversification of Lineages during the Tertiary. Syst Biol 54:575–594.

Lewis, G., Schrire, B., Mackinder, B., Lock, M. (eds.) (2005). Legumes of the World. Royal Botanic Gardens, Kew. 577 pp.

Macas JJ, Novak P, Pellicer J, et al (2015) In Depth Characterization of Repetitive DNA in 23 Plant Genomes Reveals Sources of Genome Size Variation in the Legume Tribe Fabeae. PLoS One 10:e0143424.

Maddison WP, Maddison DR. (2011) Mesquite: a modular system for evolutionary analysis. Version 2.75. http://mesquiteproject.org.

Mayrose I, Barker MS, Otto SP (2010) Probabilistic models of chromosome number evolution and the inference of polyploidy. Syst Biol 59:132–144.

Mendonça RC, Felfili JM, Walter BMT, Silva Junior MC, Rezende AV, Filgueiras TS, Nogueira PE, Fagg CW (2008) Flora vascular do Bioma Cerrado: checklist com 12.356 especies. In: Sano SM, Almeida SP, Ribeiro JF (eds) Cerrado: ecologia e flora. Embrapa Cerrados, Brasília, pp 423–1279

Michael TP (2014) Plant genome size variation: bloating and purging DNA. Briefings In Functional Genomics 13:308-317.

Murat F, Zhang R, Guizard S, et al (2015) Karyotype and gene order evolution from reconstructed extinct ancestors highlight contrasts in genome plasticity of modern rosid crops. Genome Biol Evol 7:735–749.

Nag A, Chanda S, Rajkumar S (2011) Estimation of nuclear genome size of important medicinal plant species from western Himalaya using flow cytometry. Journal of Cell & Plant Sciences 2:19-23.

Pandit MK, White SM, Pocock MJO (2014) The contrasting effects of genome size, chromosome number and ploidy level on plant invasiveness: A global analysis. New Phytol 203:697–703.

Pellicer J, Fay MF, Leitch IJ (2010) The largest eukaryotic genome of them all? Botanical Journal of the Linnean Society 164:10-15.

Peruzzi L (2013) “ x ” is not a bias, but a number with real biological significance. Plant Biosyst - An Int J Deal with all Asp Plant Biol 147:1238–1241.

Posada D. (2008) jModelTest: phylogenetic model averaging. Molecular Biology and Evolution, 25(7):1253.

R Core Team (2016) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria

Rambaut A, Suchard MA, Xie D, Drummond AJ. (2014) Tracer v1.6, Available from http://beast.bio.ed.ac.uk/Tracer

Rambaut A. (2009) FigTree version 1.3.1, Available from http://tree.bio.ed.ac.uk.

Rice A, Glick L, Abadi S, et al (2015) The Chromosome Counts Database (CCDB) - a community resource of plant chromosome numbers. New Phytol 206:19–26.

Ronquist F, Huelsenbeck JP. (2003) MRBAYES 3: Bayesian phylogenetic inference under mixed models. Bioinformatics, v19, p1572-1574.

Sato S, Nakamura Y, Kaneko T, et al (2008) Genome Structure of the Legume, Lotus japonicus. DNA Res 15:227–239.

Schmutz J, Cannon SB, Schlueter J, et al (2010) Genome sequence of the palaeopolyploid soybean. Nature 463:178–183.

Schmutz J, McClean PE, Mamidi S, et al (2014) A reference genome for common bean and genome-wide analysis of dual domestications. Nat Genet 46:707–713.

Sousa A, Cusimano N, Renner SS (2014) Combining FISH and model-based predictions to understand chromosome evolution in *Typhonium* (Araceae). Ann Bot 113:669–680.

Suda J, Meyerson LA, Leitch IJ, Pyšek P (2015) The hidden side of plant invasions: The role of genome size. New Phytol 205:994–1007.

Tang, H.; Krishnakumar, V.; Bidwell, S.; Rosen, B.; Chan, A.; Zhou, S.; Gentzbittel, L.; Childs, K.L.; Yandell, M.; Gundlach, H.; Mayer, K.F.X.; Schwartz, D.C.; Town, C.D. (2014) An improved genome release (version Mt4.0) for the model legume *Medicago truncatula*. BMC Genomics, v.15:312.

The Plant List (2013) Published on the Internet, Version 1.1. http://www.theplantlist.org/. Accessed 3 Sep 2016

Varshney RK, Song C, Saxena RK, et al (2013) Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. Nat Biotechnol 31:240–246.

Wang, X.; Zhang, Z.; Wen, Z.; Xiao, H.; Yang, Z.; Chen, G.; Zhao, X. (2011) The chromosome number, karyotype and genome size of the desert plant diploid Reaumuria soongorica (Pall.) Maxim. Plant Cell Rep 30:955–964.

Wang S, Deng Z, Chung F, Hu W (2013) From Gaussian kernel density estimation to kernel methods. Int J Mach Learn Cybern 4:119–137.

Wang J, Liu J, Kang M (2015) Quantitative testing of the methodology for genome size estimation in plants using flow cytometry: a case study of the *Primulina genus*. Frontiers in Plant Science 6.

Whitney KD, Baack EJ, Hamrick JL, Godt MJW, Barringer BC, Bennett MD, Eckert CG, Goodwillie C, Kalisz S, Leitch IJ, Ross-Ibarra J (2010) A role for nonadaptive processes in plant genome size evolution? Evolution 2097-2109.

Wickham H (2009) ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York