UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE CIÊNCIAS

DEPARTAMENTO DE BIOQUÍMICA E BIOLOGIA MOLECULAR

CURSO DE PÓS-GRADUAÇÃO EM BIOQUIMICA

ANÁLISES TRANSCRIPTOMICA E PROTEÔMICA DE SEMENTES DE CAJUEIRO (*Anacardium occidentale* L.) VISANDO APLICAÇÕES BIOTECNOLÓGICAS

JOÃO GARCIA ALVES FILHO

FORTALEZA-CE

2013

JOÃO GARCIA ALVES FILHO

ANÁLISES TRANSCRIPTOMICA E PROTEÔMICA DE SEMENTES DE CAJUEIRO (*Anacardium occidentale* L.) VISANDO APLICAÇÕES BIOTECNOLÓGICAS

Tese apresentada como parte dos requisitos para a obtenção do grau de Doutor em Bioquímica pela Universidade Federal do Ceará.

FORTALEZA-CE

2013

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

A479a Alves Filho, João Garcia.
Análise transcriptomica e proteômica de sementes de cajueiro (Anacardium occidentale L.) visando aplicações biotecnológicas / João Garcia Alves Filho. – 2013.
171 f. : il. color.

Tese (doutorado) – Universidade Federal do Ceará, Centro de Ciências, Programa de Pós-Graduação em Bioquímica , Fortaleza, 2013.
Orientação: Prof. Dr. Benildo Sousa Cavada.

1. Castanha de caju. 2. Proteoma. 3. Transcriptoma. I. Título.

CDD 572

Esta tese foi apresentada como parte dos requisitos necessários para a obtenção do grau de Doutor em Bioquímica, outorgado pela Universidade Federal do Ceará.

João Garcia Alves Filho

João Garcia Alves Filho

Tese aprovada em: 21 de fevereiro de 2013.

Benildo Sousa Cavada, *Dr.*
Orientador
Departamento de Bioquímica e Biologia Molecular - UFC

Kyria Santiago Nascimento, Dr.
Departamento de Bioquímica e Biologia Molecular - UFC

Celso Shiniti Nagano, Dr.
Departamento de Engenharia de Pesca – UFC

Rodrigo Maranguape Silva da Cunha, Dr.
Universidade Estadual Vale do Acaraú – UVA

João Paulo Matos Santos Lima, Dr.
Universidade Federal do Rio Grande do Norte – UFRN

À Ana Eduarda

AGRADECIMENTOS

À Deus por ter me dado vida e consciência.

Aos meus pais João Garcia e Rosália de Sousa pelo incentivo aos estudos desde minha infância.

Ao meu orientador, prof. Benildo Sousa Cavada, por ter me orientado desde a época do mestrado.

Ao prof. Rodrigo Maranguape Silva da Cunha pela orientação, incentivo e amizade cultivados ao longo de quase uma década.

A toda equipe da Embrapa Caprinos, em especial a Dr. Angela Eloy e João Ricardo, pelo auxílio prestado à parte de eletroforese bidimensional.

Ao Lemap, em nome do prof. Celso Nagano pelo auxílio prestado à parte de espectrometria de massa.

Aos professores João Paulo Matos e Kyria Santiago pela disponibilidade em compor a banca e auxílios prestados na parte de bioinformática e proteômica.

A Vitória Virgínia pelo grande auxílio prestado nas análises dos géis bidimensionais além da amizade, parceria e cumplicidade ao longo desses anos.

Ao Raulzito Fernandes pelo apoio na elaboração das análises com microssatélites.

À toda equipe do Laboratório de Biologia Molecular do NUBIS: Maria Amélia Soares, Cleane Moreira, Áurea Morgana, Aurilene Gomes, Mônica Valéria, Crislay Fontenele, Flavia Muniz, Jedson Aragão, Nayanne Hardy, Bruno Bezerra, Antonio Francisco de Sousa, Mariana da Silva, Dauana Mesquita e Marcos Silvino.

Aos amigos Ricardo Basto, Maria Auxiliadora Oliveira, Daniel de Brito, Tatiana Farias, Nagila Carneiro, Gleiciane Martins pela amizade ao longo desses anos.

A todos aqueles que de forma direta ou indireta fizeram com que a execução deste trabalho fosse possível.

Este trabalho foi realizado graças ao apoio das seguintes instituições:

Fundação Cearense de Apoio ao Desenvolvimento Científico e Tecnológico (FUNCAP), através da bolsa de Doutorado concedida, nos anos de 2009 a 2011.

Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), através da Bolsa de Doutorado concedida, no ano de 2012.

Universidade Federal do Ceará (UFC).

Departamento de Bioquímica e Biologia Molecular (DBBM) pertencente à UFC.

Laboratório de Moléculas Biologicamente Ativas (Biomol-Lab) do Departamento de Bioquímica e Biologia Molecular/UFC.

Laboratório de Espectrometria de Massa de Proteínas (LEMAP) pertencente à UFC.

Faculdade de Medicina de Sobral (FAMED) pertencente à UFC.

Núcleo de Biotecnologia de Sobral (NUBIS) cuja sede fica localizada na Faculdade de Medicina de Sobral.

“Eu gosto do impossível porque lá a concorrência é menor”

Walt Disney

RESUMO

O cajueiro (*Anacardium occidentale* L.) é uma importante espécie vegetal brasileira, ocorrendo especialmente na região Nordeste. Na natureza, existem basicamente dois tipos de cajueiros: o tipo comum, também conhecido como gigante, e o cajueiro anão-precoce, o qual apresenta baixo porte e precocidade na produção de frutos. O fruto do cajueiro (castanha-de-caju) é fonte de lipídeos e proteínas, sendo utilizado como alimento no mundo todo. No entanto, a comercialização de subprodutos da castanha-de-caju tem mostrado atenção especial devido à presença de proteínas alergênicas. Além disso, as potencialidades de uso do cajueiro são limitadas devida a carência de dados moleculares. Assim, o presente trabalho tem como objetivo estabelecer um panorama geral sobre transcritos e proteínas de castanhas de cajueiro por meio de análise transcriptômica e proteômica. Com esta finalidade, sementes em maturação de cajueiro comum e anão CCP 76 foram utilizadas para obtenção de RNA mensageiro. Posterior sequenciamento por plataforma Illumina com cobertura média de 4x, obtiveram-se sequências com qualidade Phred acima de 30. A montagem do transcriptoma revelou 37.422 e 77.371 sequências contíguas, relativa aos fragmentos de DNA complementares de sementes de cajueiro comum e anão CCP 76, respectivamente. O percentual de sequências do transcriptoma identificadas pelo BLAST variou de 15,2 a 45,9% para o cajueiro CCP 76 e comum, respectivamente. Análise comparativa entre os dois genótipos permitiu a identificação de três novos microssatélites polimórficos. Adicionalmente, a análise do proteoma de sementes quiescentes do cajueiro comum, juntamente com os genótipos anão-precoce CCP 76, CCP 09, BRS 226 e BRS 275 mostrou a presença de seis proteínas diferencialmente expressas. Em suma, as análises transcriptômica e proteômica realizadas no presente estudo contribuíram para a adição de novos dados moleculares para *A. occidentale* L., assim como permitiram a identificação de possíveis marcadores com grande potencial biotecnológico.

PALAVRAS-CHAVE: Castanha de caju. Proteoma. Transcriptoma.

ABSTRACT

The cashew (*Anacardium occidentale* L.) is an important Brazilian plant species occurring especially in the northeast region. In nature, there are basically two types of cashew, the common also known as giant, and dwarf cashew which show low size and precocity for fruit production. The fruit of the cashew (cashew-nut) is a source of lipids and proteins, it has been utilized to food in the world. However, the sub-product commercialization of cashew-nut had shown special attention by presence of allergenic proteins. Futhermore, the potentialities of the cashew use are limited due to lack of molecular data. Thus, the present study aimed to establish an overview about transcripts and proteins from cashew-nut through transcriptome and proteome analysis. For this purpose, seeds of the common and dwarf CCP 76 cashew in maturation were used to obtain messenger RNA. After sequencing by Illumina platform with 4x average coverage, we obtained sequences with Phred score above 30. The transcriptome assembly revealed 37,422 and 77,371 contigs, relative to complementary DNA fragment of seeds of the common and dwarf CCP 76 cashew, respectively. The percentage of transcriptome sequences identified by BLAST varied of 15.2 to 45.9% for CCP 76 and common cashew, respectively. The comparative analysis between the two genotypes allowed the identification of tree new polymorphic microsatellites. Additionally, the proteome analysis of seed quiescent of both common cashew and dwarf CCP 76, CCP 09, BRS 226 e BRS 275 genotypes showed presence of six different expressed proteins. In summary, transcriptome and proteome analysis contributed to addition of new molecular data for *A. occidentale* L., as well as allowed the identification of possible marks with high biotechnological potential.

KEYWORDS: Cashew nut. Proteome. Transcriptome.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE GRÁFICOS

LISTA DE ABREVIATURAS E SIGLAS

ACC – Amêndoa da Castanha de Caju

AMP – Adenosina Monofosfato

ATP – Adenosina Trifosfato

BLAST - Basic Local Alignment Search Tool

CCP 06 - Clone Cajueiro Pacajus 06

CCP 09 – Clone Cajueiro Pacajus 09

CCP 1001 - Clone Cajueiro Pacajus 1001

CCP 76 - Clone Cajueiro Pacajus 76

cDNA – DNA complementar

CTE – Cadeia Transportadora de Elétrons

DNA – Ácido Desoxirribonucléico

ESTs - Etiquetas de Sequência Expressa

FTP – File Transfer Protocol

GC – Guanina e Citosina

GO - Gene Ontology

IBGE – Instituto Brasileiro de Geografia e Estatística

KEGG - Kyoto Encyclopedia of Genes and Genomes

KO - KEGG Orthology

LCC – Líquido da castanha de caju

LSPA - Levantamento Sistemático da Produção Agrícola

mRNA - RNA mensageiro

MySQL – My Structured Query Language

MS – Espectrometria de massa

NCBI - National Center for Biotechnology Information

NGS - Nova Geração de Sequenciamento

nsLTPs - Proteínas de transferência de lipídeos não específicos

ORFs - open reading frames

pb – pares de base

PCR – Reação em cadeia da polimerase

PR-proteínas - Proteínas relacionadas a patogênese

qRT-PCR – PCR em tempo real

RNA – Ácido ribonucléico

RNA-Seq - sequenciamento de RNA

RPKM - Reads per kilobase per milion

rRNA – RNA ribossômico

SAGE – Análise Serial da Expressão Gênica

SOLiD – Sequenciamento por ligação e detecção de oligonucleotídeo

tRNA – RNA de transferência

SUMÁRIO

INTRODUÇÃO

O Estado do Ceará é o maior exportador do país de castanha e líquido da castanha de caju. De acordo com o último levantamento Sistemático da Produção Agrícola (LSPA), do IBGE, realizado em outubro de 2011, a agricultura cearense se recuperou após os problemas climáticos de 2010. Um dos destaques foi a castanha de caju cuja safra aumentou de 39,6 mil toneladas em 2010 para 111,7 mil toneladas em 2011 (um aumento de 182,2% sobre a safra anterior). Vale ressaltar que a baixa produtividade de castanhas de caju em 2010 obrigou as indústrias cearenses a importar cerca de 84 mil toneladas da África (ADECE, 2012).

A amêndoa da castanha do caju contém proteínas de alta qualidade nutricional, e é considerada uma fonte de proteínas de baixo custo (OGUNWOLU et al., 2009). Contudo, a amêndoa da castanha contém proteínas que são alergênicas para pessoas que tenham hipersensibilidade. Cerca de 20% dos casos registrados de alergia a alimentos nos Estados Unidos devem-se a presença de compostos alergênicos encontrados na castanha do caju (TEUBER et al., 2002). Desta forma, é de grande interesse a caracterização das proteínas de reserva de sementes de cajueiro para viabilizar uma melhor caracterização dos alérgenos.

Na natureza existem dois grupos de cajueiro, o anão-precoce e o comum. Particularmente o tipo anão-precoce tem sido alvo de esforços concentrados de melhoramento, devido às suas características de interesse, especialmente para exploração dentro dos modernos sistemas de cultivo. Várias características intrínsecas do cajueiro anão, como precocidade e porte, têm sido alvo dos programas de melhoramento. Atualmente, vários genótipos de cajueiro anão-precoce estão disponíveis no mercado. Cada genótipo parece exibir diferenças fenotípicas, mas a literatura é escassa no tocante a caracterização molecular dos genótipos de cajueiro.

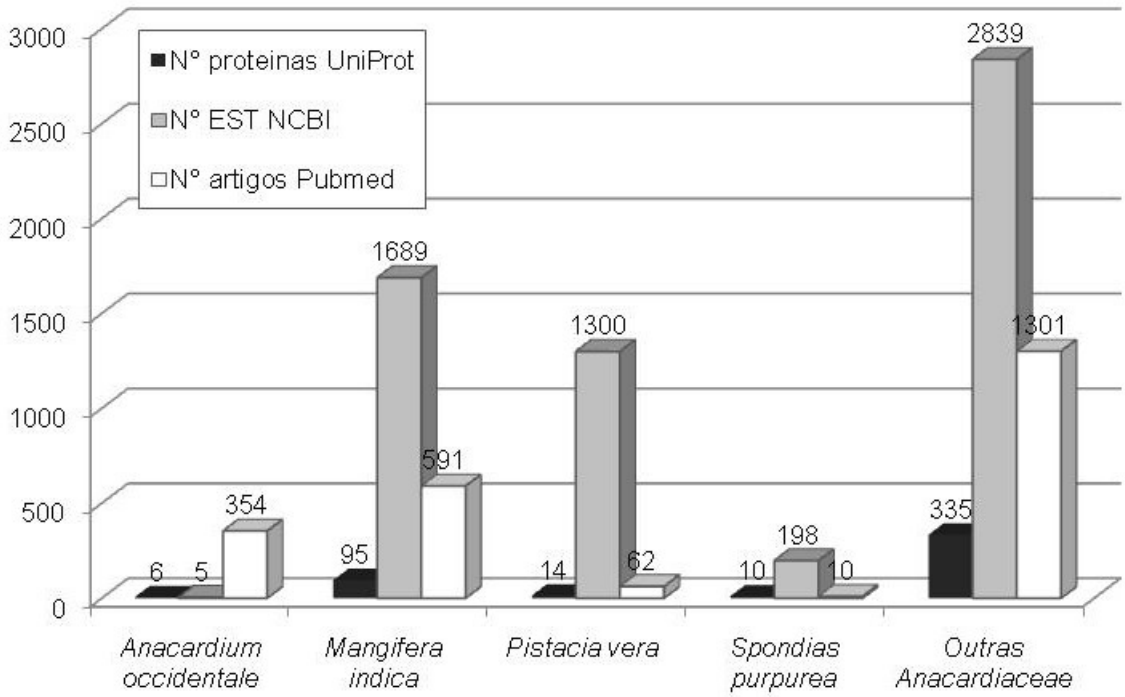
Recentes avanços na área de genômica, transcriptômica e proteômica podem contribuir com a formação de um banco de dados de sequências de cajueiro, o qual irá auxiliar no entendimento mais aprofundado dos mecanismos

moleculares que envolvem toda a fisiologia do desta espécie.

O banco de dados *Swiss-Prot* possui duas seções nos quais as sequências podem ser anotadas manualmente (*Swiss-Prot*) ou automaticamente (TrEMBL). A família Anacardiaceae possui 460 proteínas depositadas (*Swiss-Prot*/TrEMBL), entre elas apenas 12 (3%) pertence ao gênero *Anacardium* e destas 6 sequências pertencem a espécie *A. occidentale* (Gráfico 1). Além disso, *A. occidentale* possui apenas 39 sequências de nucleotídeos depositadas no *GenBank*.

Em síntese, a proposta do presente trabalho é realizar estudos transcriptomicos e proteômicos de sementes de cajueiro comum e anão CCP 76 visando à identificação de genes ou proteínas que possam ser usados como marcadores moleculares.

Gráfico 1 - Número de proteínas, ESTs e artigos relacionados à família Anacardiaceae encontrados em bancos de dados públicos.



Três espécies importantes de Anacardiaceae foram destacadas: o cajueiro (*Anacardium occidentale*), manga (*Mangifera indica*) e o pistache (*Pistacia vera*). O número de proteínas foi obtido no bando de dados UniProt; o número de ESTs foi obtido pelo NCBI e o número de artigos pelo PubMed, também pertencente ao NCBI. **Fonte**: UniProt (www.uniprot.org), e NCBI (http://www.ncbi.nlm.nih.gov). Acessado em 01/01/2013.

CAPÍTULO 1:
REVISÃO DE LITERATURA

1 REVISÃO DE LITERATURA

1.1 A biologia do cajueiro

O cajueiro (*Anacardium occidentale* L.) pertence à família Anacardiaceae, a qual possui 60 a 74 gêneros e 400 a 600 espécies de árvores, arbustos, subarbustos e trepadeiras ocorrendo principalmente em climas tropicais e subtropicais. Além do cajueiro, várias outras Anacardiaceae são exploradas economicamente, destacando-se a manga (*Mangifera indica* L.), o pistache (*Pistacia vera* L.), o umbu (*Spondias tuberosa* Arr. Câm.), a cajá (*Spondias mombim* L.), a seriguela (*Spondias purpurea* L.) e a cajá-manga (*Spondias cytherea*), encontradas na América tropical (PAIVA; CRISÓSTOMO; BARROS, 2003).

O gênero *Anacardium* possui 21 espécies, mas *A. occidentale* é a única espécie cultivada. O cajueiro é uma árvore ou arbusto (anão-precoce) apresentando folhas simples, alternas, subcoriáceas, glabras, ovadas, obtusas, onduladas, pecioladas, roxo-avermelhadas quando jovem, de cor verde-amareladas quando maduras podendo cair após maturação (Figura 1). As flores são pequenas, curto-pediceladas, pálidas, avermelhadas, dispostas em panículas terminais, pedunculadas e ramificadas. O sistema reprodutivo é predominantemente alogâmico (fecundação cruzada). O fruto (castanha) é um aquênio reniforme, formado por epicarpo, mesocarpo e endocarpo. O epicarpo é liso, coriáceo e cinzento. O mesocarpo é espesso, alveolado e cheio de LCC (líquido da castanha do caju). A amêndoa (parte comestível) é reniforme, composta por dois cotilédones brancos, carnosos, oleosos e revestida por uma película pergaminácea. O pedúnculo floral ou pseudofruto (caju) é hipertrofiado, carnoso, suculento e variável em tamanho, peso e cor (BARROS et al., 1993).

O processo de maturação de sementes é o resultado de alterações morfológicas, fisiológicas e bioquímicas, como aumento de tamanho, variações no teor de água e vigor, que iniciam desde a fertilização até o momento em que as sementes estão maduras e se desprendem da planta-mãe (CARVALHO; NAKAGAWA, 2000). Sendo assim, o estudo da maturação é importante para que se tenha um conhecimento mais amplo da reprodução vegetal, possibilitando prever a época adequada de colheita.

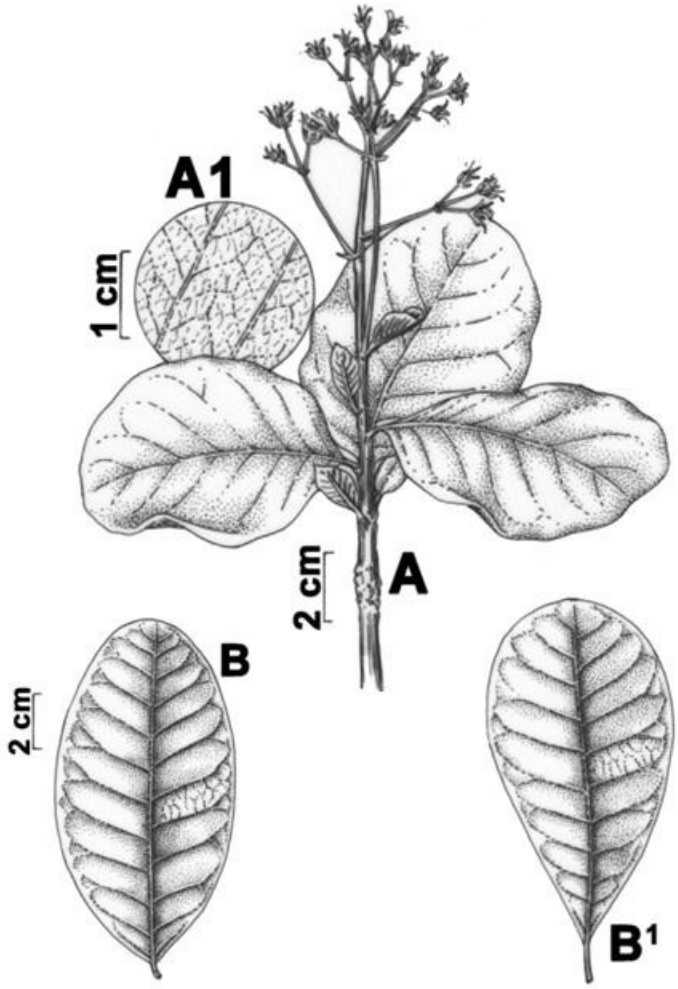
O desenvolvimento da castanha-de-caju pode ser avaliado de acordo com a cor da castanha. Castanhas de cor avermelhada, também chamada de maturi, são as mais jovens. Até a castanha atingir o tamanho máximo, ela irá crescer e mudar de cor de vermelho a verde. A castanha totalmente verde está em crescimento máximo. Por fim, a castanha de cor cinza já está no final do processo de maturação (RAO; HASSAN, 1957).

É durante a maturação das castanhas onde ocorrem os primeiros ataques de insetos herbívoros. De acordo com Bleicher e colaboradores (1995), a traça-da-castanha, *Anacampis sp.* (Lepidoptera: Gelechiidae), inicia o ataque em castanhas de cor vermelha e aumenta sua intensidade à medida que o fruto se desenvolve (BLEICHER; ABREU; MELO, 1995).

Na natureza, existem basicamente dois tipos de cajueiro, o cajueiro comum (*A. occidentale* L.) e o cajueiro anão-precoce (*A. occidentale* var. *nanum*) (Figura 1).

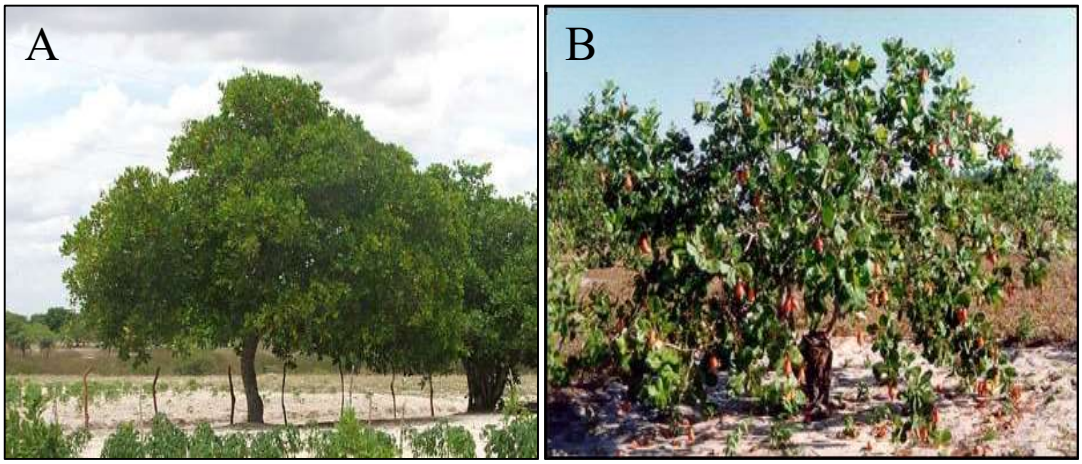
O cajueiro comum é o mais predominante no Nordeste ocorrendo de forma natural sem a necessidadede plantio. Ele apresenta porte elevado, com altura variando entre 8 e 15 m e extensão da copa podendo atingir até 20 m, dependendo das condições de clima, solo e sanidade (CRISÓSTOMO et al., 2001). O peso do fruto varia de 3 - 33 g, e o do pedúnculo de 20 – 500 g. O segundo tipo de cajueiro é conhecido como anão-precoce (cajueiro de 6 meses), e é caracterizado, principalmente pelo seu porte baixo, com altura entre 3 e 4 m e copa atingindo até 9 m (LIMA, 1988). O tipo anão-precoce também é caracterizado pela precocidade, iniciando o florescimento entre 6 e 18 meses contra os cinco a sete meses do tipo comum. O peso do fruto, nas populações naturais, varia de 3 – 10 g e o do pedúnculo de 20 – 160 g (LIMA, 1988). As principais características dos cajueiros dos tipos comum e anão-precoce são ilustradas na Tabela 1 (CAVALCANTI et al., 2009).

Figura 1 – Detalhe da folha e inflorescencia do cajueiro (*Anacardium occidentale*).



(A) Ramo florífero; **(A1)** detalhe das nervuras da face abaxial; **(B-B1)** Variação foliar.
Fonte: (LUZ, 2011)

Figura 2 - Tipos de cajueiro existentes na natureza.



(A). Cajueiro comum (*A. occidentale* L.); **(B)** cajueiro anão-precoce (*A. occidentale var. nanum*). **Fonte**: (PAIVA; BARROS, 2004).

Tabela 1 - Principais características dos cajueiros dos tipos comum e anão-precoce.

Características	Comum	Anão-precoce
Porte (m)	Alto (8-15)	Baixo (<5)
Tamanho da copa (m)	>7	5 a 7
Primeira Florada	2 a 5 anos	6 a 18 meses
Variação do peso da castanha (g)	3 a 33	3 a 13
Variação do peso do pedúnculo (g)	20 a 500	20 a 160
Produção: castanha/planta/safra (Kg)	<1 a >100	Até 43

Fonte: (CAVALCANTI et al., 2009).

As proteínas de semente podem ser classificadas em proteínas de reserva, proteínas estruturais e metabólicas e proteínas de proteção (FERREIRA; BORGHETTI, 2004). As proteínas de reserva de semente, por sua vez geralmente são classificadas de acordo com a sua solubilidade em: água (albuminas), solução salina diluída (globulinas), solução alcoólica diluída (prolaminas) e soluções diluídas ácidas ou alcalinas (glutelinas) (OSBORNE, 1924; SHEWRY; NAPIER; TATHAN, 1995).

Entre as proteínas da amêndoa da castanha de interesse nutricional,

destaca-se a anacardeína, uma globulina 13S (do tipo legumina) que pode conter até 50% do conteúdo total de nitrogênio da amêndoa. Ela é composta de pelo menos dois polipeptídeos com massas de 18-24 KDa e 30-37 KDa, com ponto isoelétrico entre 6,2 e 7,2 não apresentando carboidratos ligados covalentemente (SATHE et al., 1997). Outras proteínas encontradas na amêndoa da castanha do caju como as vicilinas são conhecidas por serem alergênicas.

A maioria dos alérgenos pode ser agrupada em poucas famílias baseando-se nas suas propriedades estruturais e funcionais como a superfamília das cupinas, superfamília das prolaminas e proteínas do sistema de defesa. A superfamília das cupinas se subdividem em vicilinas (globulinas 7S triméricas) e leguminas (globulinas 11S hexaméricas). A subfamília das prolaminas são formadas pelas albuminas 2S, proteínas de transferência de lipídeos não específicos (nsLTPs), a α-amilase de cereais e inibidores de proteases. As proteínas relacionadas à defesa são bastante diversas como as proteínas relacionadas a patogênese (PR-proteínas) (BREITENEDER; RADAUER, 2004).

Alguns dos alérgenos encontrados na amêndoa da castanha do caju têm sido caracterizados como o Ana o 1 (WANG et al., 2002), que é a principal proteína alergênica sendo classificada como uma albumina 2S (tipo vicilina). O segundo maior alérgeno da ACC é designada como Ana o 2 (WANG et al., 2003). Ela é classificada como uma globulina 11S (da família das leguminas). O terceiro maior alérgeno é uma albumina 2S designado como Ana o 3 (ROBOTHAM et al., 2005). As proteínas do cajueiro já identificadas estão listadas na tabela 2.

.

Tabela 2 - Lista de proteínas de *Anacardium occidentale* depositadas no banco de dados *SwissProt/TrEMBL.*

Número de acesso	Nome da proteína	pI/ MW	Número de aminoácidos
Q8L5L5	Ana o 1.0101 (Globulina: Vicilina)	5,64/ 61.840	538
Q8L5L6	Ana o 1.0102 (Globulina: Vicilina)	5,64/ 61.638	536
Q8GZP6	Ana o 2 (Globulina: Legumina)	6,18/ 5.1996	457
Q8H2B8	Ana o 3 (Albumina)	5,68/ 16.335	138
Q5WPQ1	Cadeia maior da ribulose bifosfato carboxilase	7,2/ 51.636	466
I3RXT2	Gliceraldeído-3-fosfato desidrogenase	7,6/21.234	201
H2KNA3	Maturase K	9,5/61.249	515

Fonte: UNIPROT acesso em 14/02/2013

1.2 A agroindústria do caju

O Brasil ocupa o sétimo lugar na produção de castanhas no mundo (Gráfico 2). Essa produção ocorre principalmente na região Nordeste.

O Estado do Ceará é o maior exportador do país de castanha no Brasil. De acordo com dados do IBGE, a safra da castanha de caju aumentou de 39,6 mil toneladas em 2010 para 111,7 mil toneladas em 2011 (um aumento de 182,2% sobre a safra anterior). Vale ressaltar que a baixa produtividade de castanhas de caju em 2010 obrigou as indústrias cearenses a importar cerca de 84 mil toneladas da África (ADECE, 2012).

O pedúnculo ou falso-fruto, que representa cerca de 90% do peso do fruto completo, vem se tornando, importante segmento da agroindústria do caju. Entre

os diversos produtos destacam-se o suco concentrado, doces, refrigerante gaseificado e cajuína (PAIVA; BARROS, 2004). Também foi observado que o ácido anacárdico presente no pedúnculo tem efeito contra a bactéria Gram-negativa *Helicobacter pylori* (KUBO; LEE, 1999) além de efeito antitumor (KUBO et al., 1993).

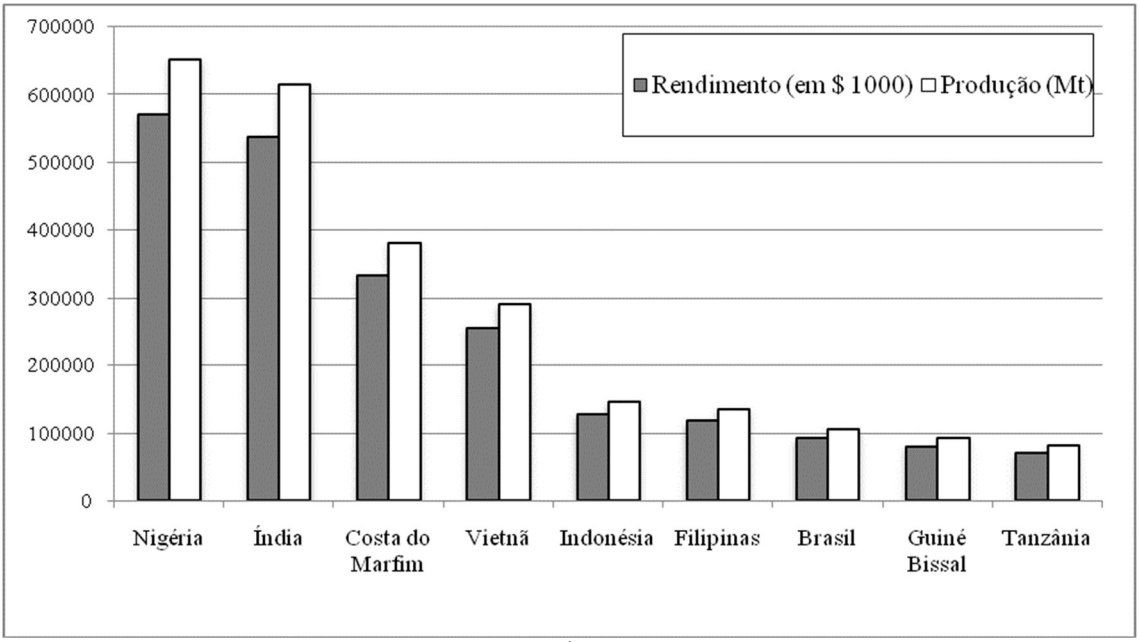
O líquido da castanha do caju (LCC) tem muitas aplicações como, por exemplo, no uso como suplemento de combustível para geração de calor (MOHOD; KHANDETOD; POWAR, 2008). Além disso, o óleo da castanha do caju contém vários compostos fenólicos como o ácido anacárdico, cardol entre inúmeros outros com diversas atividades microbianas contra bactérias Gram-positivas *Bacillus subtilis*, *Bacillus ammoniagenes*, *Staphilococcus aureus*, *Streptococcus mutans* e *Propionibacterium acnes* (HIMEJIMA; KUBO, 1991). Além do efeito antimicrobiano foi observado também efeito moluscicida (KUBO; KOMATSU; OCHI, 1986), atividade inibitória da síntese das prostaglandinas (KUBO et al., 1987) e efeito inibitório da gliceraldeido-3-fostato desidrogenase de *Trypanosoma Cruzi* (PEREIRA et al., 2008). Mais recentemente, várias atividades biológicas do LCC têm sido mais elucidadas do ponto de vista químico como a interação com membranas biológicas, efeito antioxidante, propriedade antimutagênica, efeito citotóxico, indução de apoptose pela ativação de caspases, efeito antimicrobiano e efeito na estrutura e função de proteínas. Esses fatores certamente irão contribuir para sua aplicação na farmacologia e na medicina (STASIUK; KOZUBEK, 2010).

A amêndoa da castanha do caju tem alto teor nutricional contendo altos níveis de proteínas (21%) e lipídios (48%), além de minerais como fósforo, potássio, magnésio, ferro (FETUGA; BABATUNDE; OYENUGA, 1974) e selênio (KANNAMKUMARATH et al., 2002), o qual desempenha um importante papel como antioxidante.

O baixo rendimento do cajueiro é atribuído, principalmente, aos fatores ambientais como estresse hídrico e alta salinidade dos solos bem como à ocorrência de pragas e doenças. As pragas mais importantes do cajueiro são causadas por insetos da ordem *Lepidoptera*, *Hemiptera*, *Thysanoptera* e *Coleoptera*. Com relação às doenças, destacam-se a antracnose, o mofo-preto

e a resinose, ambas causadas por fungos (PAIVA; BARROS, 2004). Uma lista de alguns fatores físicos e biológicos que diminuem o rendimento do cajueiro pode ser visualizada na tabela 3.

Gráfico 2 – Produção de castanha de caju em 2010, no mundo.



A receita (rendimento) é dada em dólares ($) e a produção em milhões de toneladas (Mt). **Fonte**: (FAO, 2008)

Tabela 3 - Principais fatores responsáveis pelo baixo rendimento do cajueiro.

Fatores físicos		
Estresse		Referência
Estresse hídrico		(BLAIKIE; CHACKO, 1998)
Estresse salino		(VOIGT et al., 2009)
Pragas		
Pragas	Espécie (Ordem: Família)	Referencia
Broca-das-pontas	*Anthistarcha binocularis* (Lepidoptera: Gelechiidae)	(BLEICHER et al., 2007)
Traça-da-castanha	*Anacampsis sp*. (Lepidoptera: Gelechiidae)	(MESQUITA; BECKER; SOBRINHO, 1998)
Pulgão	*Aphis gossypii* (Hemiptera: Aphididae)	(BLEICHER et al., 1997)
Mosca-branca	*Aleurodicus cocois* (Hemiptera: Aleyrodidae)	(MARACAJA et al., 2008)
Tripes-da-cinta-vermelha	*Selenothrips rubrocinctus* (Thysanoptera: Thripidae)	(PENG; CHRISTIAN, 2004)
Broca-da-raiz	*Marshallius spp* (Coleoptera: Curculionidae)	(BLEICHER et al., 2010)
Coleobrocas	*Apate spp* (Coleoptera: Bostrychidae)	(DWOMOH; ACKONOR; AFUN, 2008)
Doenças causadas por fungos		
Doença	Espécie	Referência
Antracnose	*Colletotrichum gloeosporioides*	(FIGUEIRÊDO et al., 2012)
Mofo-preto	*Pilgeriella anacardii*	(VIANA et al., 2012)
Resinose	*Lasiodiplodia theobromae*	(MUNIZ et al., 2011)

Fonte: O Autor.

1.3 Transcriptômica

Entende-se por transcriptômica, a área científica responsável pelo estudo do transcriptoma, ou seja, o conjunto completo de transcritos em uma célula, e suas quantidades em um estágio específico do desenvolvimento ou condição fisiológica (Wang; Gerstein; Snyder, 2009). Entre as metodologias disponíveis para a identificação de genes transcritos, a análise de sequências expressas (ESTs) tem provado ser uma estratégia bastante informativa, uma vez que permite identificar os genes expressos em uma linhagem celular ou tecido em um estágio de desenvolvimento específico (JORGE, 2002). As ESTs são sequências parciais de uma das extremidades da molécula de DNA complementar (cDNA) a um dado mRNA, resutante do sequenciamento sistemático dos clones de uma biblioteca de cDNA. Segundo Adams e colaboradores (1991), há informação suficiente nas 200-400 bases de nucleotídeos sequenciados para a identificação desses genes expressos (ADAMS et al., 1991).

A concentração relativa dos transcritos é geralmente diretamente proporcional ao seu nível de expressão. Sendo assim, através da quantificação dos transcritos, é possível inferir sobre os níveis de expressão desses genes em uma determinada condição. Entre os principais estudos envolvendo expressão gênica, destacam-se a PCR em tempo real (qRT-PCR) e *Northern blot* (AMARAL et al., 2006). No entanto, essas técnicas possibilitam o estudo de um ou poucos genes por vez.

A análise da expressão gênica em larga escala tem se desenvolvido graças à incorporação de técnicas como os arranjos de DNA (DNA array) (EISEN; BROWN, 1999) e SAGE (*Serial Analysis of Gene Expression*) (VELCULESCU et al., 1995). Entretanto, algumas limitações do uso dessas tecnologias têm surgido (HINTON et al., 2004). Devido a isso, a aplicação de sequenciamento de cDNA surgiu como uma alternativa eficiente suprir dados transcriptômicos independentemente da necessidade de uma sequência genômica de referência previamente descrita (ANDREOTE, 2011).

O sequenciamento de cDNA pelo método de Sanger, mesmo em se tratando de ESTs, é muito dispendioso e economicamente inviável devido ao grande número de sequências. Graças ao surgimento da Nova Geração de Sequenciamento (NGS) que tem se tornado possível o sequenciamento em larga escala do transcriptoma (RNA-Seq).

1.4 Nova geração de sequenciamento (NGS)

Nas abordagens de sequenciamento de DNA (sequenciamento *De novo*) em larga escala, longas cadeias de DNA são fragmentadas seja por enzimas de restrição ou quebra mecânica. Em seguida, cada fragmento de DNA passa por uma etapa de seleção clonal e amplificação. No método clássico, o DNA é clonado em um vetor e amplificado em *Escherichia coli,* os fragmentos são individualmente sequenciados e agrupados eletronicamente formando sequências contíguas. Essa abordagem vem atualmente sendo substituída por métodos menos laboriosos e totalmente automatizada.

A *Emulsion PCR* isola moléculas de DNA individuais em *beads* magnéticos (1 a 28 μm) revestidos por *primers*. Cada *bead* fica em gotas aquosas dentro de uma fase apolar. O enriquecimento dos *beads* consiste na amplificação do *template* em cada *bead* resultando em *beads* com várias cópias do mesmo fragmento sendo então imobilizadas em placas ou chips para posterior leitura. Essa abordagem é usada nas plataformas 454 (pirosequenciamento) e SOLiD, por exemplo (SHENDURE; JI, 2008) (Figura 3).

Outra estratégia de amplificação clonal é a *Bridge PCR,* que consiste na amplificação de uma sequência de DNA (imobilizado em placa na extremidade 5’ ou 3’) contendo um adaptador flexível o qual se dobra para se ligar em um *primer* que também está imobilizado nesta mesma placa. O resultado é a amplificação do *template* na ordem de aproximadamente 1000 cópias próximas ao ponto de origem em uma determinada região da placa. Essa estratégia é usada na plataforma *Illumina*, comumente referida como *Solexa* (SHENDURE; JI, 2008) (Figura 3)

. O sistema 454 (pirosequenciamento) é considerado a primeira plataforma de sequenciamento do tipo *next-generation*. A seleção clonal é feita por *Emulsion PCR,* com *amplicons* capturados na superfície de *beads* de 28 μm.

Depois da quebra da emulsão, os *beads* são sujeitos a enriquecimento baseado em amplificação e são então depositados em placas contendo micropoços (com capacidade para apenas um *bead* por poço). A estratégia para leitura da sequência é baseada na adição de nucleotídeos com um acompanhamento de emissão de luz. Resumidamente, um desoxinucleotídeo é adicionado na reação, caso encontre sua base complementar na sequência ela é incorporada havendo liberação de pirofosfato o qual é incorporado ao AMP pela sulfurilase produzindo ATP. Esse ATP produzido é substrato para a luciferase que produz luz sendo então captada pelo sistema óptico do equipamento (ANSORGE, 2009). Caso a base adicionada não for complementar ao *template* não há sinal de luz e rapidamente inicia-se um novo ciclo. Caso o *template* contenha um trecho com o mesmo nucleotídeo repetido algumas vezes, nota-se um aumento na intensidade de luz naquela região proporcional ao número de nucleotídeos adicionados.

A plataforma *Solexa*, agora parte da *Illumina*, desenvolveram uma tecnologia baseada em *dye-terminators* reversíveis. Inicialmente, a seleção clonal é feita por *Bridge PCR* no qual os fragmentos de DNA gerados por quebra mecânica (ondas acústicas) são ligados a adaptadores e ligam-se aleatoriamente na superfície de uma placa. O adaptador presente na placa é elongado com base na sequência de DNA que acabou de se ligar e essa nova fita se dobra tocando a superfície da placa fazendo com que ela encontra um novo adaptador havendo uma amplificação conhecida como *Bridge amplification*. A *Bridge PCR* ocorre em vários ciclos gerando milhares de cópias do DNA de interesse em regiões específicas. Para que ocorra a leitura, o *primer* de sequenciamento é anelado seguido pela adição de base (uma por vez) cada base é marcada com uma fluorescência e a detecção dessa fluorescência é feita pela parte óptica do equipamento em quatro canais (SHENDURE; JI, 2008).

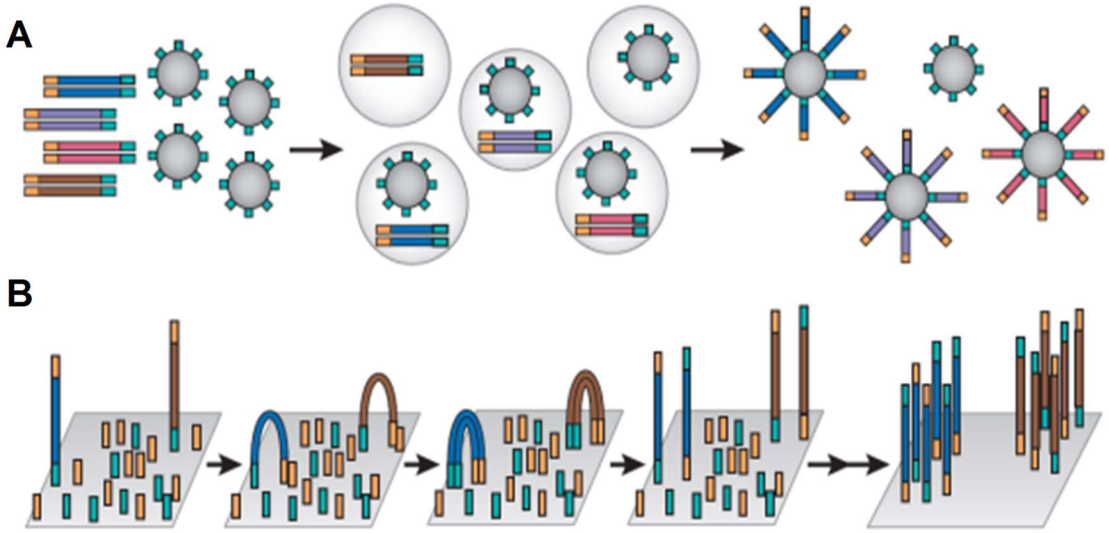
A tecnologia SOLiD de sequenciamento (da *Applied Biosystems*) emprega a DNA ligase ao invés da DNA polimerase na reação. A seleção clonal é feita por *Emulsion PCR* com *beads* paramagnéticos de 1 μm sendo então imobilizados em uma placa. O sequenciamento é dirigido por uma DNA ligase onde *primers* universais de sequenciamento são adicionados seguidos pela adição de uma mistura de oligonucleotídeos (um *pool* de octâmeros que possui

em uma extremidade um dinucleotídeo complementar à sequência alvo e a outra extremidade é marcada com fluorescência). A media que um oligonucleotídeo específico hibridiza com a sequência alvo, a DNA ligase liga os fragmentos e o sinal de fluorescência é captado, em seguida, ocorre a clivagem da extremidade 5’ do oligonucleotídeo havendo perda da fluorescência, porém liberando essa extremidade para a ligação do próximo oligonucleotídeo complementar a sequência. Sucessivos ciclos e *rounds* de hibridização e coleta de dados ocorrem com a mesma sequência alvo e o computador decifra o código de dois nucleotídeos (*two-base encoding*) (SHENDURE; JI, 2008).

Um sistema relacionado ao SOLiD é o *Polonator*, que também é baseado em *Emulsion PCR* e sequenciamento por ligação. Porém, o custo do *Polonator* é bastante reduzido quando comparado aos outros sequenciadores da segunda geração. Outro sistema diferente dos demais é o *HeliScope* que é capaz de fazer a leitura sem haver a necessidade de amplificação clonal. Em vez disso, um sistema de detecção de fluorescência altamente sensitivo é utilizado para dirigir o sequenciamento por síntese. As bibliotecas contendo caudas poli-A são hibridizadas em placa rica em oligo dT. Em seguida, uma DNA polimerase adiciona nucleotídeos marcados com fluorescência que são captados pelo equipamento (SHENDURE; JI, 2008).

O *Ion Torrent Systems Inc* desenvolveu um sistema baseado em uma química de sequenciamento padrão, mas com um novo sistema de detecção baseado em semicondutores. Esse método é baseado na detecção de íons hidrogênio que são liberados durante a polimerização do DNA. Em um micropoço contendo o DNA *template* é adicionado um único tipo de nucleotídeo. Se o nucleotídeo adicionado for complementar ao da sequência, ele é incorporado causando a liberação de um íon hidrogênio que ativa o sensor hipersensitivo, indicando que a reação ocorreu. Em casos de homopolímeros, uma maior quantidade de íons hidrogênio é liberada e o sinal eletrônico é aumentado na mesma proporção (RUSK, 2011).

Figura 3 – Estratégias de amplificação clonal utilizando nova geração de sequenciamento.



(A) as plataforms 454, Polonator e SOLiD utilizam amplificação clonal baseado em PCR em emulsão. **(B)** A plataforma Solexa (*Illumina*) utiliza a amplificação clonal baseado em bridge PCR. **Fonte**: (SHENDURE; JI, 2008).

1.5 Proteômica

O proteoma é o conjunto de todas as proteínas presentes em um material biológico e que, ao contrário do genoma, está em contínua mudança em resposta aos estímulos internos e externos (WILKINS, 1995; PATERSON; AEBERSOLD, 2003). A proteômica, o estudo do proteoma, é uma área recente tendo surgido em meados dos anos 90. Contudo, algumas das principais técnicas já eram bastante difundidas. A eletroforese bidimensional, por exemplo, teve início nos anos 70 (O´FARREL, 1975), e o primeiro espectrômetro de massa data de 1918. No entanto, a invenção das fontes de ionização MALDI (*Matrix-Assisted Laser Desorption/Ionization*) e ESI (*EletroSpray Ionization*) nos anos 80 possibilitou com que as proteínas pudessem ser levadas à fase gasosa sem a necessidade de altas temperaturas. De uma maneira geral, várias áreas da pesquisa convergiram para tornar possível o estudo de proteínas em larga escala.

Em genômica, o problema de medir a expressão de genes tem sido resolvido pela introdução de métodos como a PCR, hibridização e microarranjos. Por outro lado, proteínas não podem ser amplificadas por uma técnica semelhante à PCR. Uma pequena quantidade de polipeptídeo deve ser

detectada em pequenas quantidades sem nenhum tipo de amplificação ou hibridização (LIEBLER, 2002). Sendo assim, a análise do proteoma requer uma série de metodologias para que esse estudo se torne possível. Entre as ferramentas para proteômica destacam-se a eletroforese bidimensional, a espectrometria de massa e ferramentas de bioinformática que, além de analisarem os dados, criam e gerenciam banco de dados.

A eletroforese bidimensional passou por melhorias com a criação do IPG (gradiente de pH imobilizado) (GÖRK et al., 2007) e é atualmente usado em estudos de proteômica comparativa, no qual o objetivo é identificar diferenças qualitativas e quantitativas entre amostras de proteínas, pois mesmo com as limitações da técnica, gera dados em um formato que permite fácil avaliação visual (CÁNOVAS et al., 2004).

A separação de proteínas por ferramentas analíticas como o SDS-PAGE, focalização isoelétrica, 2D-PAGE e HPLC têm ajudado a simplificar uma mistura complexa de proteínas para proteínas individuais ou pequenos grupos de proteínas facilitando sua identificação via MS. Como uma alternativa para a 2D-PAGE, a tecnologia de identificação multidimensional de proteínas (MudPIT) é um método que tem provado ser bastante eficiente para a identificação de proteínas em misturas complexas (MATTHIESEN, 2007).

Em contraste com outras técnicas de identificação de proteínas como a degradação de Edman, a MS fornece uma abordagem massiva para a identificação de proteínas em larga escala. Existem basicamente dois métodos de identificação de proteínas por MS. A impressão digital da massa do peptídeo ou PMF (do inglês *Peptide Mass Fingerprint*) possibilita a determinação precisa da massa dos fragmentos obtidos pela digestão com tripsina. Paralelamente, proteínas sequenciadas previamente podem ser clivadas *in silico* com tripsina a fim de se obter a massa teórica dos fragmentos. A comparação das massas dos fragmentos calculados com os dados experimentais permite uma identificação precisa da proteína. O segundo método de identificação de proteínas por MS é através da espectrometria de massa sequencial (MS/MS ou MS^2) que se baseia na fragmentação dos íons peptídicos em câmara de colisão contendo gás nobre. Esses fragmentos irão compor o novo espectro que será tomado como base para fazer o sequenciamento *de novo* dos aminoácidos constituintes. A utilização de computadores tem sido bastante útil para a identificação de proteínas por MS em

larga escala (LIEBLER, 2002).

Inicialmente, durante os primeiros estudos de 2DE, a análise dos *spots* de proteínas era feita manualmente, sem a ajuda de softwares. A presença ou ausência de *spots* era convertida em uma matriz binária que era então agrupada e a analisada. Devido ao fato de a 2-DE ser imperfeita, devido a distorções no padrão de proteínas causadas pelos procedimentos de polimerização e corrida dos géis, a necessidade por softwares logo se tornou aparente para ser possível um melhor alinhamento e comparação dos géis (BIRON et al., 2006).

Os primeiros softwares foram feitos com base naqueles usados por astrônomos para mapeamento de estrelas, um deles se chama “Tycho” em homenagem ao famoso astrônomo Tycho Brahe (1546-1601). No final do século 20, outros softwares pioneiros como o MELANIE da Swiss Institute of Bioinformatics foram desenvolvidos. Na última década, um importante número de *softwares* comerciais, envolvendo algoritmos mais poderosos e ferramentas estatísticas do que as gerações prévias de tais programas, foram desenhados para auxiliar os pesquisadores a lidar com a enorme quantidade de dados produzidos (BIRON et al., 2006).

Atualmente, vários *softwares* têm sido desenvolvidos para a análise de eletroforeses bidimensionais. Um exemplo é o SWISS-2DPAGE (http://au.expasy.org/ch2d/) que pode localizar proteínas em mapas bidimensionais a partir do banco de dados SWISS-PROT (www.expasy.ch). O banco de dados UniProt possui duas seções nos quais as sequências podem ser anotadas manualmente (Swiss-Prot) ou automaticamente (TrEMBL). Análises nos espectros de massas podem ser feitos em vários *softwares* disponíveis online como o MOWSE, ProFound, PeptIdent, MASCOT, etc.

CAPÍTULO 2:
MONTAGEM *De novo* DO TRANSCRIPTOMA DO CAJUEIRO UTILIZANDO SEQUENCIAMENTO POR SÍNTESE E A MONTAGEM POR DOIS CONJUNTOS DE ALGORITMOS

2 MONTAGEM *DE NOVO* DO TRANSCRIPTOMA DO CAJUEIRO UTILIZANDO SEQUENCIAMENTO POR SÍNTESE E A MONTAGEM POR DOIS CONJUNTOS DE ALGORITMOS

2.1 INTRODUÇÃO

O transcriptoma é o conjunto de todos os genes transcritos em uma célula e sua abundância relativa varia conforme estágio de desenvolvimento ou condição fisiológica (WANG; GERSTEIN; SNYDER, 2009). Assim como o genoma, o transcriptoma pode ser obtido pelas novas tecnologias de sequenciamento e essa tecnologia passou a ser chamada de RNA-Seq. Embora o transcriptoma seja menor que o genoma, o sequenciamento por RNA-Seq é bastante desafiador por vários motivos. O primeiro é que, ao contrário do genoma, o transcriptoma contém genes cuja expressão (número de cópias de RNA) varia em várias ordens de magnitude. Além disso, o processamento de RNA torna a tarefa de montagem bem mais complexa quando comparada a montagem de genomas.

Mesmo que um organismo já possua seu genoma sequenciado, o estudo do transcriptoma por RNA-Seq é importante, pois permite a análise da expressão de milhares de genes simultaneamente, mesmo que eles sejam pouco expressos. Outra vantagem é a possibilidade de estudos envolvendo processamento de mRNA, RNA não codificante, micro RNA, etc. A técnica de RNA-Seq é mais versátil que estudos envolvendo microarranjos e PCR em tempo real e está sendo bastante utilizada no estudo de transcriptoma em plantas não modelo.

Um experimento de RNA-Seq, moléculas de RNA total ou mRNA são fragmentados e convertidos em uma biblioteca de cDNA. A biblioteca é então sequenciada em sequenciadores de nova geração para produzir milhões a bilhões de sequências curtas chamadas de *reads*. Os *reads* podem ser lidos a partir de uma extremidade (*Single End*) ou em ambas as extremidades (*Paired End*) dependendo dos adaptadores utilizados. Por fim, as sequências dos *reads* são montadas utilizando programas específicos para a reconstrução do transcriptoma (MARTIN; WANG, 2011). Programas como *Phrap* podem consumir muita memória RAM e processador em um computador e a montagem

de um transcriptoma poderia levar meses para se concluir. Novos programas como o *Velvet* utilizam novos algoritmos para construção de um genoma ou transcriptoma baseado no grafo de Bruijn (Figura 4). Essa nova metodologia tem reduzido o tempo de processamento, mas ainda requer bastante memória RAM e processamento, especialmente em transcriptomas eucariontes.

A montagem do transcriptoma pode ser feita com o auxílio do genoma do organismo como referência (Figura 4). No entanto, em organismos que não possuem genomas sequenciados, é possível fazer uma montagem usando a estratégia *De novo*. Na maioria das espécies, especialmente as plantas poliplóides, a falta genomas de referência ocorre devido ao tamanho e a complexidade de seus genomas (MARTIN; WANG, 2011).

Para uma montagem *De novo* de um transcriptoma de um organismo eucarionte é necessário milhões ou bilhões de *reads* e a construção do grafo de Bruijn pode levar dias ou semanas mesmo em um computador que tenha centenas de gigabytes de memória RAM até produzir um transcriptoma que tenha cobertura acima de 30 x (MARTIN; WANG, 2011).

Em geral, a montagem de transcriptomas de organismos complexos requerem sequenciamentos poderosos como as plataformas *Illumina* e *SOLiD*. Mesmo assim, a montagem *De novo* de organismo eucariontes é muito mais desafiadora não apenas pelo tamanho do transcriptoma, mas também pelas dificuldades envolvidas na identificação de *splicing* alternativos (MARTIN; WANG, 2011).

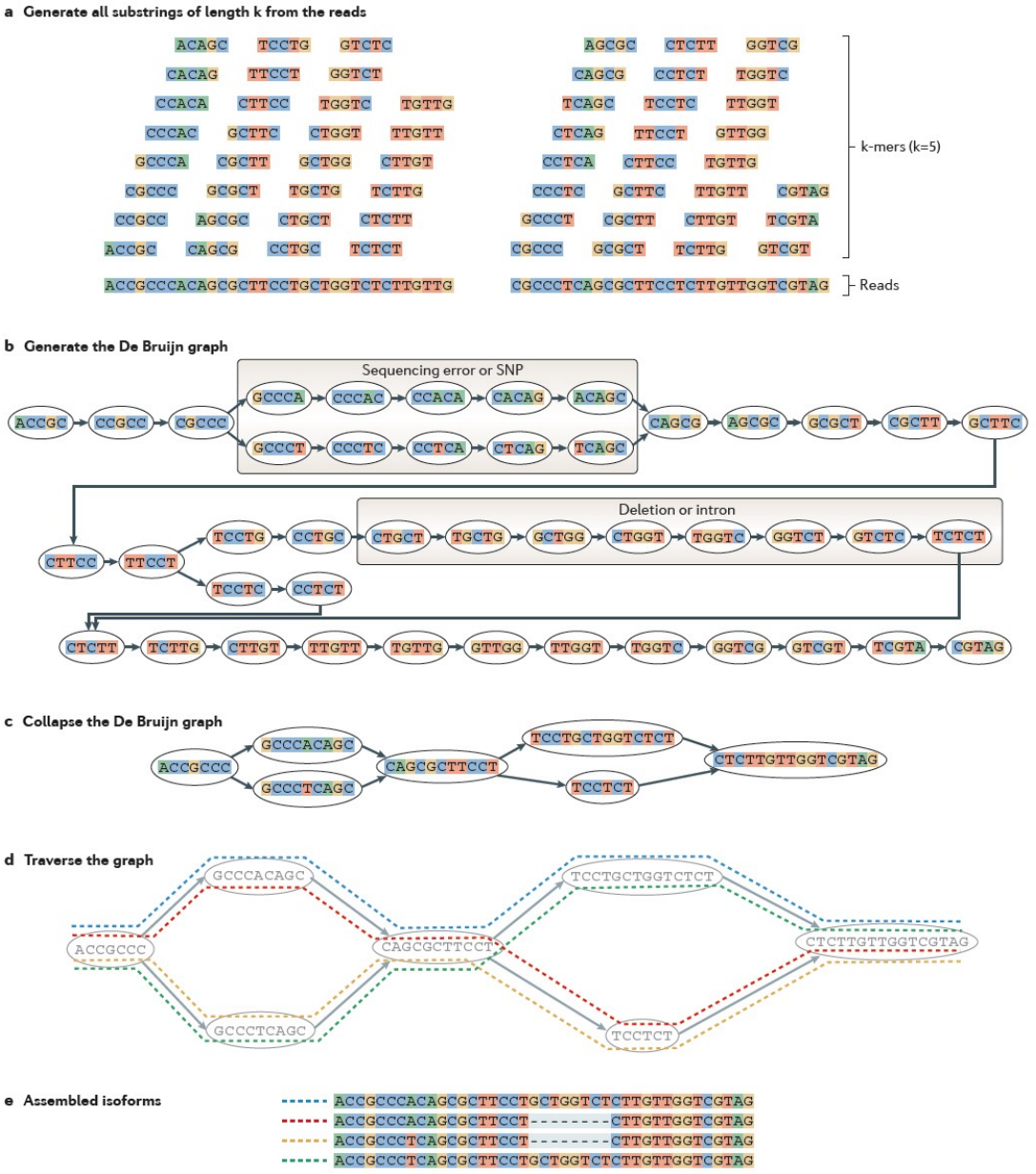
Entre os inúmeros programas montadores de genoma e transcriptoma, um dos mais populares para a criação de um transcriptoma sem um genoma de referência é o *Velvet*. O programa *Velvet* é um conjunto de algoritmos criados para manipulação de grafo de Bruijn (uma representação baseada em palavras curtas denominadas de k-*mers*), utilizado na reconstrução de sequências genômicas a partir de um conjunto de dados de leituras curtas (Figura 4) como é o caso do Illumina/Solexa (ZERBINO; BIRNEY, 2008).

O *Velvet* é formado por duas partes: *Velveth* e *Velvetg*. O *Velveth* tem como função principal a criação de uma tabela de índices (*hash*) a partir de um conjunto de sequências de leituras, computando sobreposições entre k-mers, e

gerar no final do processo dois arquivos principais: um arquivo de 36 sequências indexadas (sequence) e um arquivo contendo a representação das sobreposições entre os k-mers chamado de mapa de vias (Roadmaps) (ZERBINO; BIRNEY, 2008).

O *Velvetg* é responsável pela construção do genoma através da manipulação do grafo de Bruijn, correção de erros e resolução de repetições, gerando um arquivo de *contigs* ou *scaffolds* (conjunto de *contigs*) (ZERBINO; BIRNEY, 2008).

Figura 4 – Visão geral da estratégia de montagem *De novo* do transcriptoma.



(A) todos os possíveis *k-mers* são gerados a partir de cada *read*. (**B)** cada *k-mer* é utilizado para a construção de um nó no grafo de Bruijn e os pares de nós são conectados se deslocando um *k-mer* por um caractere criando uma sobreposição de *k-1* entre dois *k-mers*. **(C,D)** Cadeias de nós adjacentes no grafo são colapsados em um único nó. As isoformas são então unidas. **Fonte:** (MARTIN; WANG, 2011):

2.2 OBJETIVOS

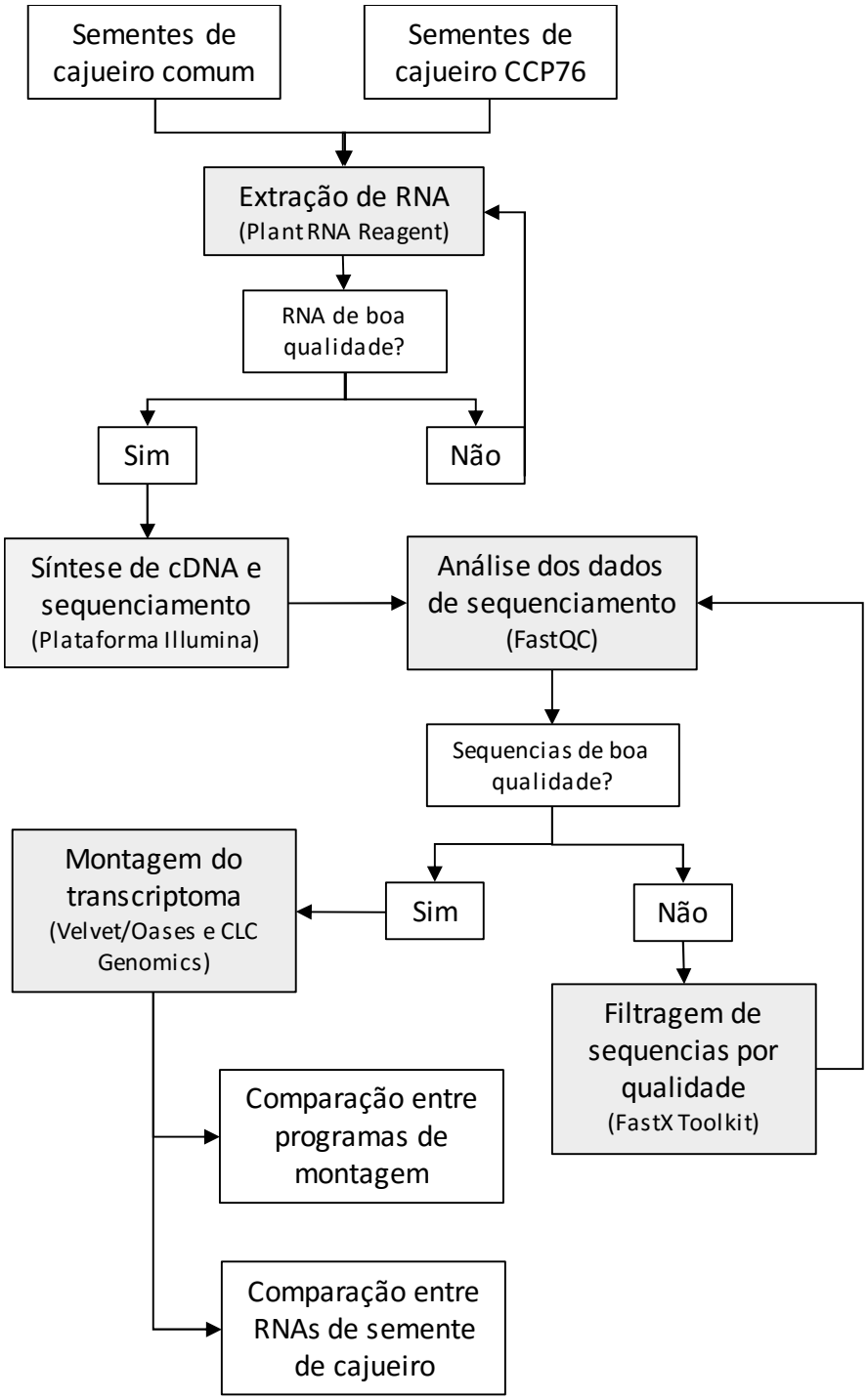
2.2.1 Geral

Obter uma montagem *De novo* do transcriptoma de cajueiro comum e cajueiro anão CCP 76 utilizando dados obtidos por sequenciamento de próxima geração.

2.2.2 Específicos

- Extrair e purificar RNA de sementes em formação de cajueiro comum e cajueiro anão (CCP 76);
- Construir uma biblioteca para cada genótipo compatível com o sequenciamento por síntese;
- Obter sequências com qualidade para montagem *De novo* do transcriptoma de sementes em formação de cada genótipo;
- Comparar a montagem do transcriptoma obtidas pelos programas de *Velvet* e *CLC Genomics WorkBench*.
- Comparar montagem do cajueiro comum com o cajueiro CCP 76;

2.3 ESTRATÉGIA EXPERIMENTAL



2.4 METODOLOGIA

2.4.1 Coleta de material biológico

Sementes de cajueiro comum e anão-precoce CCP 76 foram coletadas no sítio Recanto dos Cajueiros localizado no município de Itapipoca-CE. Castanhas jovens em diferentes estágios de maturação foram lavadas com água destilada, coletadas e armazenadas em nitrogênio líquido até a etapa de extração de RNA.

2.4.2 Isolamento de RNA total

O RNA total foi obtido pelo *Plant RNA Reagent* (*Invitrogen*™). Brevemente, sementes jovens de cajueiro foram coletadas e imediatamente congeladas em nitrogênio líquido. As amostras foram então maceradas e incubadas com 1 mL de tampão de extração. Os tubos foram levemente invertidos e incubados a temperatura ambiente por 15 minutos. Em seguida, a solução foi centrifugada por 2 minutos a 12.000 x *g* a temperatura ambiente e o sobrenadante, transferido para um novo tubo. Os tubos foram misturados por inversão após adição de 0,1 mL de NaCl 5 M e 0,3 mL de clorofórmio. As amostras foram novamente centrifugadas a 10.000 x *g* por 10 minutos a 4 °C para separar as fases. A fase superior aquosa foi transferida para um novo tubo livre de RNAse seguido pela adição de um volume igual de isopropanol. As amostras foram misturadas por inversão e deixadas à temperatura ambiente por 10 minutos. Os tubos foram centrifugados a 12.000 x *g* por 10 minutos. O sobrenadante foi descartado e ao precipitado foi adicionado 1 ml de etanol 75%. Os tubos foram centrifugados a temperatura ambiente por 1 minuto a 12.000 x g. O líquido foi descartado e ao precipitado será adicionado 30 µL de água livre de RNAse misturando levemente com pipeta e deixando em repouso até o precipitado se dissolver por completo. As amostras foram armazenadas em freezer -80 °C para uso posterior.

2.4.3 RNA-Seq utilizando a plataforma Illumina

Após a extração de RNA, foi feito um *pool* de amostra para o cajueiro comum e outro para o cajueiro anão-precoce. Cerca de 10 µg de RNA total extraído foi levado ao Laboratório de Biotecnologia Animal (ESALQ-USP) para a construção das bibliotecas de cDNA do tipo *Paired End* foi utilizado a plataforma de sequenciamento *Illumina HiSeq2000*.

2.4.4 Montagem do transcriptoma

O processamento dos dados foi feito em um computador *HP Proliant* com 8 núcleos de processamento e 16,7 Gb de memória RAM. Os dados brutos (arquivos no formato *fastq*) contendo dados de sequência e qualidade de base foram avaliados pelo programa *FastQC*. Os *reads* de baixa qualidade serão removidos (*trimados*) utilizando a ferramenta *FastX Toolkit*. A montagem *De novo* foi feita utilizando o programa *Velvet*. O programa *VelvetOptimiser* foi utilizado para escolher os melhores parâmetros do *Velvet* para cada biblioteca. Após o processamento no *Velvet*, o arquivo de saída foi processado pelo programa *Oases* para a análise dos transcritos. O resultado da montagem do *Velvet* foi avaliado usando o programa estatístico *R* com o módulo *Plotrix* para criação de gráficos.

Alternativamente foi utilizado o programa *CLC Genomics Workbench* para montagem. A montagem foi feita utilizando os parâmetros recomendados pelo programa (*default*). Em ambos os programas foram avaliados dados como o tamanho do maior *contig*, N50, número de *contigs*, etc.

2.5 RESULTADOS

2.5.1 Extração de RNA

A extração de RNA total de sementes de cajueiro do tipo anão CCP 76 (Figura 5) e cajueiro comum (Figura 6) foi bem-sucedida, pois ambas mostraram as bandas de RNA ribossômico que são indicadores de que não houve degradação da amostra. A quantificação do RNA mostrou valores de concentração entre 680 a 1454 µg/µL e uma relação das absorbâncias $A_{260/280}$ entre 1,8 e 2,0. O preparo das bibliotecas de cDNAs e sequenciamento pela plataforma *Illumina* também ocorreram conforme o esperado (dados não mostrados).

2.5.2 Análise de dados brutos

Os dados brutos obtidos contendo as sequências e valores de qualidade para cada base (formato *fastq*) foram analisados pelo software *FastQC*. Foi observado que metade da biblioteca continha *reads* com baixa qualidade de sequência, o que contribuía com a diminuição da média de qualidade do total de *reads* (dados não mostrados). Sendo assim, os *reads* curtos (menos de 15 nucleotídeos) e de baixa qualidade (valor *Phred* abaixo de 30) foram removidos utilizando o *software FastX Toolkit*. Após este processamento, as sequências foram novamente analisadas pelo software *FastQC*.

O gráfico de qualidade da sequência por base apresentou altos valores de qualidade tanto nas sequências do cajueiro CCP 76 (Gráfico 3) quanto no cajueiro comum (Gráfico 4) com valor *Phred* médio acima de 30 em todas as bases. O desvio padrão foi relativamente pequeno exceto na base de número 29.

Mesmo após a remoção das bases com qualidade abaixo de 30, ainda pode-se observar que existem dois grupos de *reads* com diferentes valores de qualidade. Isso pode ser observado no gráfico de qualidade por sequência do cajueiro CCP 76 (Gráfico 5) e do cajueiro comum (Gráfico 6).

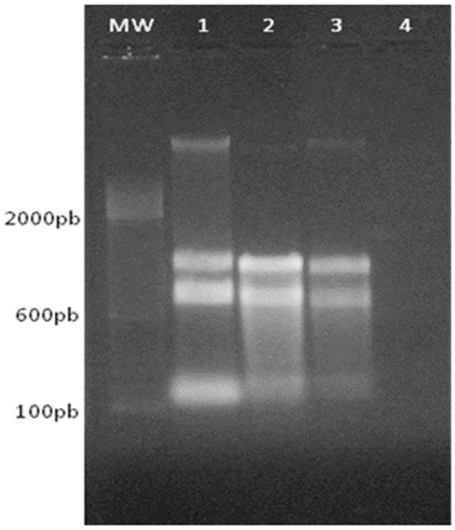
O conteúdo de GC dos *reads* também foi avaliado para saber se eles estão em uma distribuição normal teórica. Tanto o cajueiro anão CCP 76 (Gráfico

7) quanto o cajueiro comum (Gráfico 8) mostraram valores de GC aceitáveis.

Geralmente, uma base de baixa qualidade é representada pela letra “N” que representa qualquer base nitrogenada. O gráfico que indica o conteúdo de N também é útil para avaliar a qualidade dos *reads*. Pode-se observar que praticamente não existem “N” no cajueiro anão CCP 76 (Gráfico 9) e no cajueiro comum (Gráfico 10).

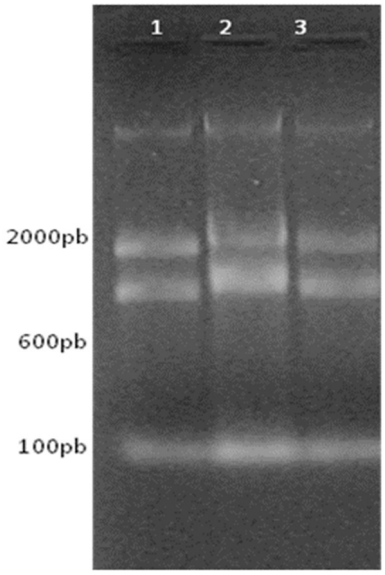
Alguns sequenciadores produzem fragmentos de tamanho não uniforme e mesmo produzindo sequências uniformes, um *pipeline* pode remover (*trimar*) as sequências de baixa qualidade. O gráfico de distribuição de sequência mostra que a maioria dos *reads* possui tamanho de 50 pb em ambas as bibliotecas do cajueiro CCP 76 (Gráfico 11) e cajueiro comum (Gráfico 12). Vale ressaltar que antes de *trimar* as bases, todas as sequências tinham o mesmo tamanho (dados não mostrados).

Figura 5 - Eletroforese em gel de agarose mostrando RNA total de cajueiro CCP 76 em quatro estádios de maturação.



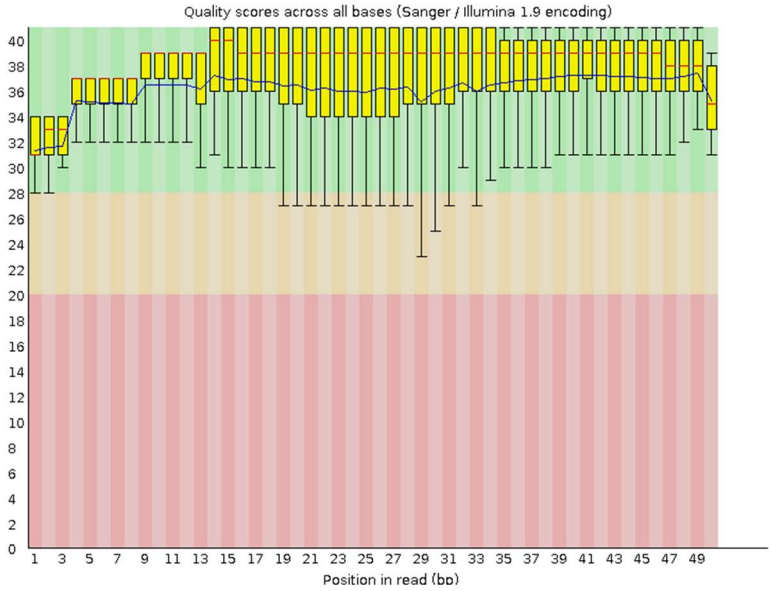
MW: marcador de peso molecular *Ladder* 100pb; **raia 1**: estadio de maturação 1 (roxo); **raia 2**: estadio de maturação 2 (roxo+verde); **raia 3**: estádio de maturação 3 (verde); **raia 4**: estádio de maturação 4 (cinza). **Fonte**: O Autor.

Figura 6 - Eletroforese em gel de agarose mostrando RNA total de cajueiro comum em quatro estádios de maturação.



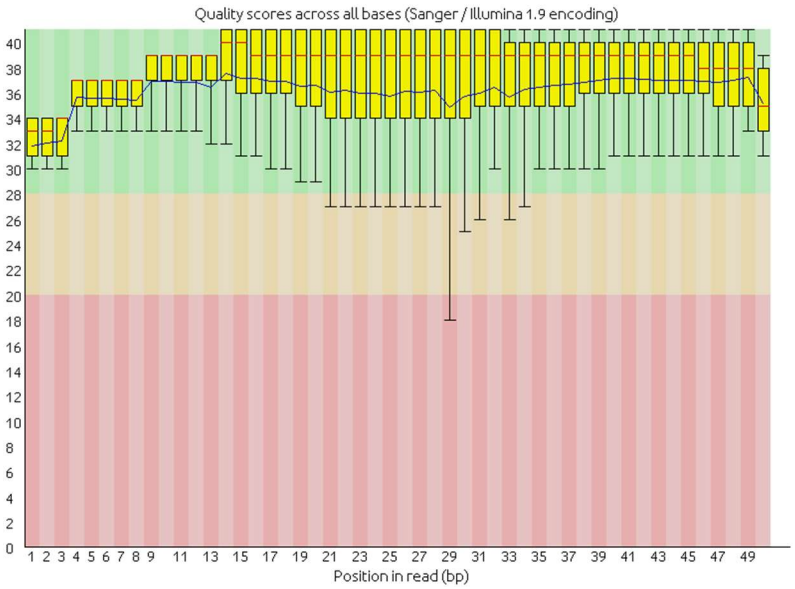
Raia 1: estadio de maturação 1 (roxo); **raia 2**: estadio de maturação 2 (roxo+verde); **raia 3:** estádio de maturação 3 (verde). **Fonte**: O Autor.

Gráfico 3 - Qualidade de sequência dos reads transcriptoma do cajueiro CCP 76 por base.



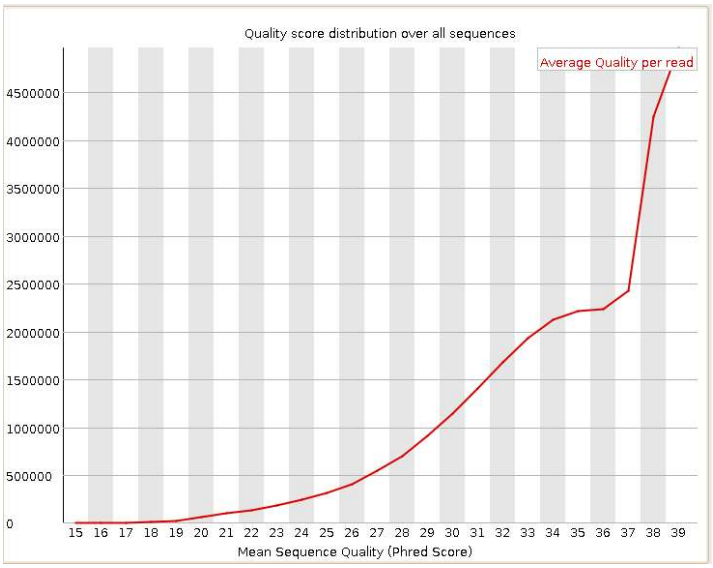
A linha vermelha central indica o valor de mediana; A caixa amarela representa a variação inter-quartil (25-75%); a linha azul representa a qualidade média. O eixo y indica a qualidade Phred (boa acima de 28). O gráfico foi feito utilizando o programa FastQC. **Fonte**: O Autor.

Gráfico 4 - Qualidade de sequência dos reads do transcriptoma do cajueiro comum de acordo por base.



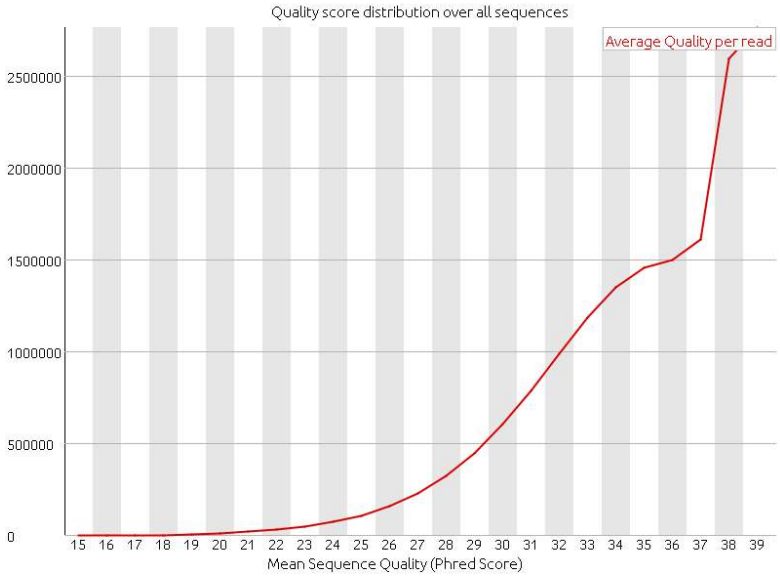
A linha vermelha central indica o valor de mediana; A caixa amarela representa a variação inter-quartil (25-75%); a linha azul representa a qualidade média. O eixo y indica a qualidade Phred (boa acima de 28). O gráfico foi feito utilizando o programa FastQC. **Fonte**: O Autor.

Gráfico 5 - Qualidade dos reads do transcriptoma do cajueiro CCP 76 por sequência.



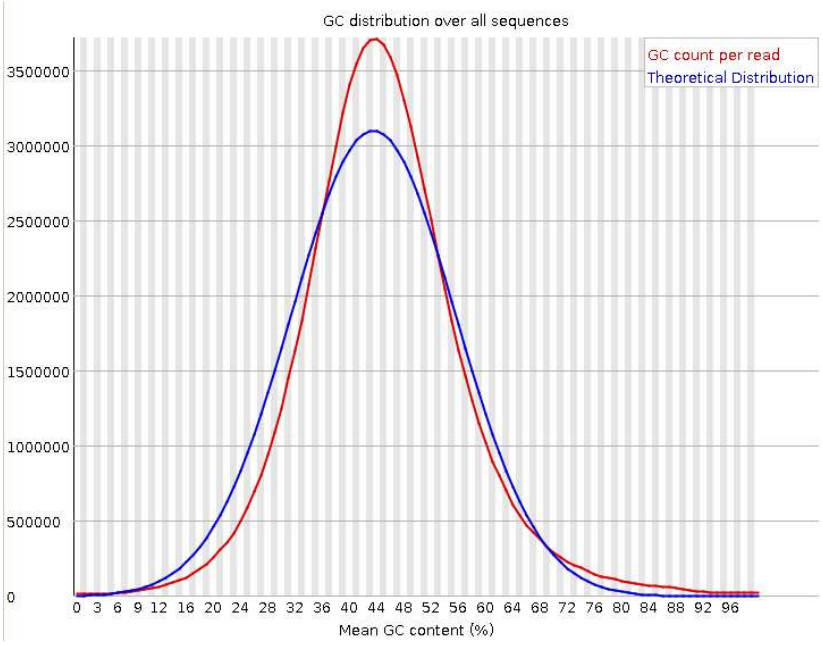
O gráfico foi feito utilizando o programa *FastQC*. **Fonte**: O Autor.

Gráfico 6 - Qualidade dos reads do transcriptoma do cajueiro comum por sequência.



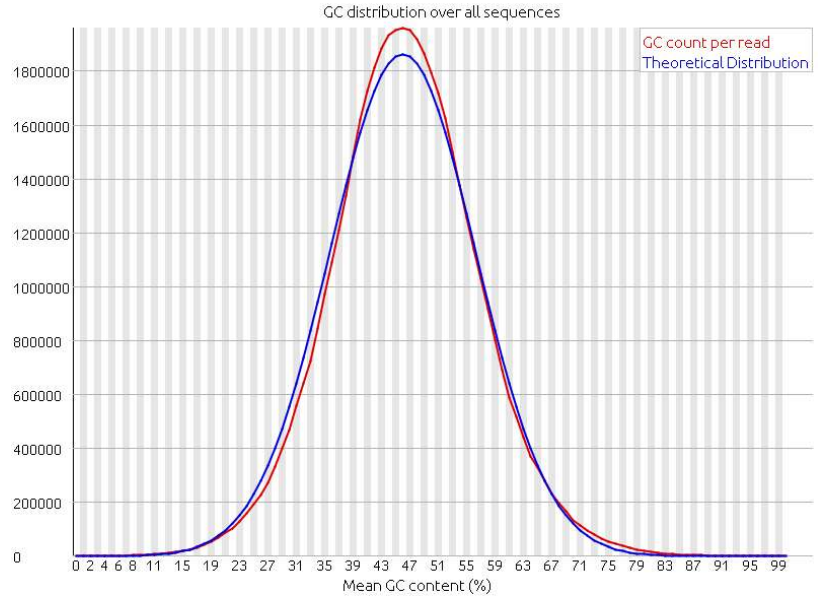
O gráfico foi feito utilizando o programa *FastQC*. **Fonte**: O Autor.

Gráfico 7 - Conteúdo de GC dos *reads* do transcriptoma do cajueiro CCP 76.



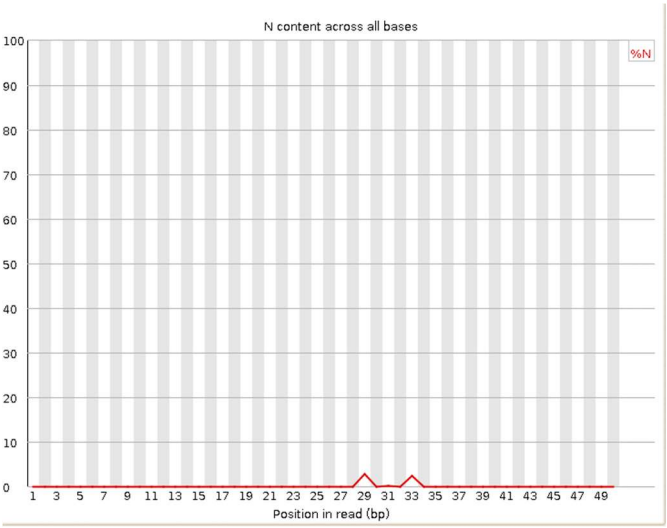
O gráfico foi feito utilizando o programa *FastQC*. **Fonte**: O Autor.

Gráfico 8 - Conteúdo de GC dos *reads* do transcriptoma do cajueiro comum.



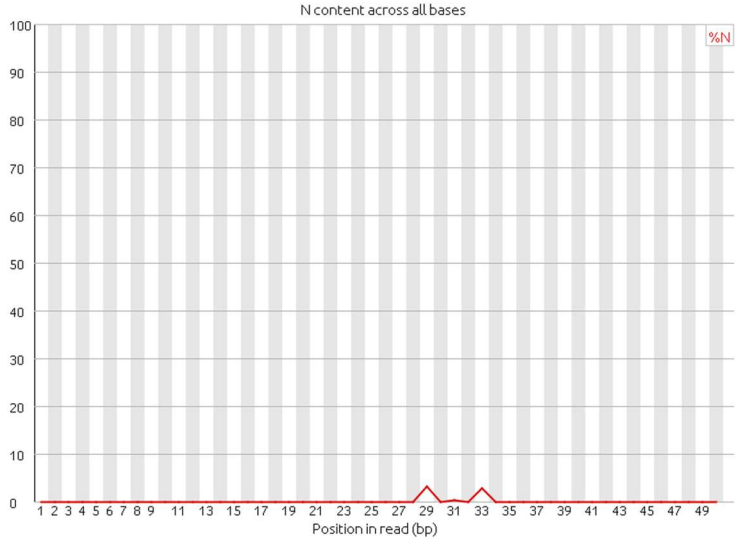
O gráfico foi feito utilizando o programa *FastQC*. **Fonte**: O Autor.

Gráfico 9 - Conteúdo de bases indeterminadas (N) nos *reads* do transcriptoma do cajueiro CCP 76.



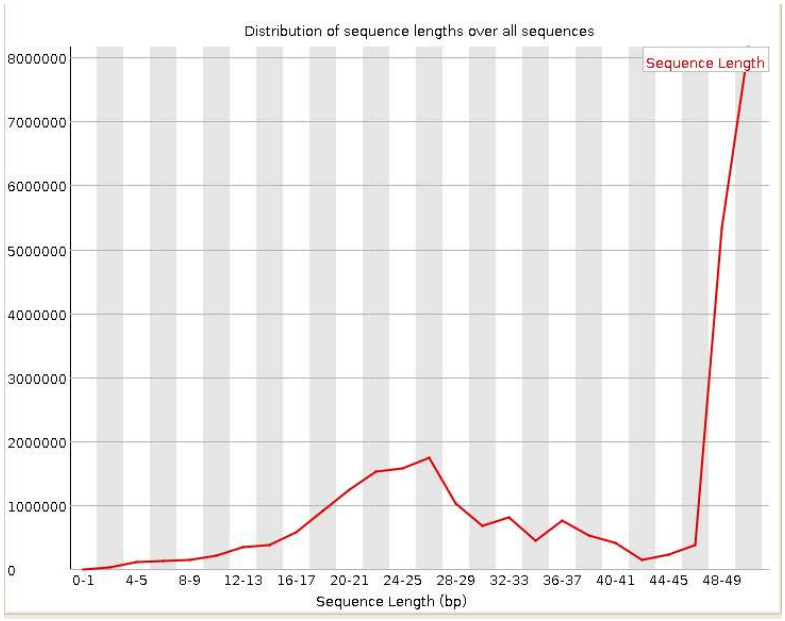
O gráfico foi feito utilizando o programa *FastQC*. **Fonte**: O Autor.

Gráfico 10 - Conteúdo de bases indeterminadas (N) nos *reads* do transcriptoma do cajueiro comum.



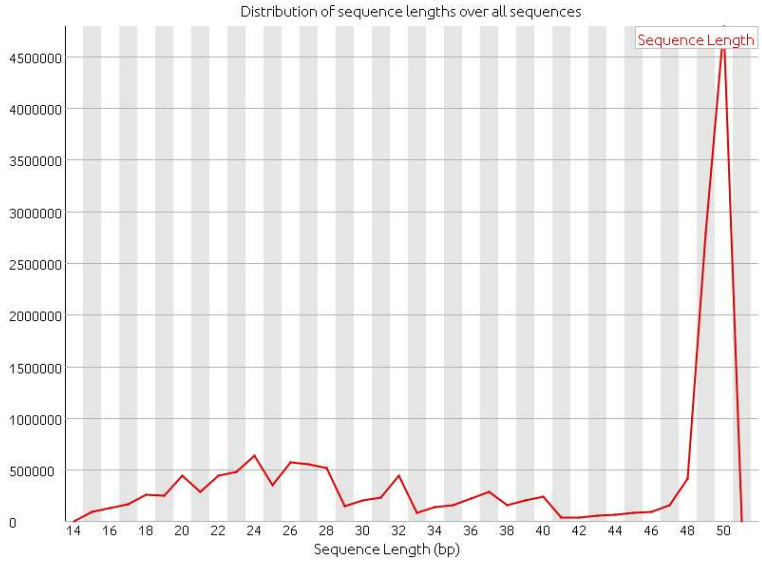
O gráfico foi feito utilizando o programa *FastQC*. **Fonte**: O Autor.

Gráfico 11 - Distribuição do comprimento da sequência dos *reads* do transcriptoma do cajueiro CCP 76.



O gráfico foi feito utilizando o programa *FastQC*. **Fonte**: O Autor.

Gráfico 12 - Distribuição do comprimento da sequência dos *reads* do transcriptoma do cajueiro comum.



O gráfico foi feito utilizando o programa *FastQC*. **Fonte**: O Autor.

2.5.3 Montagem do transcriptoma utilizando o Velvet e Oases

Com base no programa *VelvetOptimiser*, o melhor valor de *k-mer* para as bibliotecas de cajueiro comum e anão CCP 76 foi de 31 e a cobertura esperada (Exp_cov) foi 4. Ainda com base no mesmo programa, os melhores valores de corte de cobertura (Cov_cutoff) devem ser de 1,04 e 0,28 para o cajueiro CCP 76 e cajueiro comum, respectivamente.

Para a montagem do cajueiro anão CCP 76, foi utilizado um total de 28.085.118 *reads* e a montagem revelaram a presença de 117.667 *contigs* sendo que o maior deles possui 3.773 nucleotídeos de tamanho. O valor de N50 é 228 e a montagem foi feita utilizando 47,8% dos *reads*. O programa *Oases* utilizou os dados processados pelo *Velvet* adicionando como parâmetro o menor transcrito possuindo 100 pb resultando em 77.371 transcritos (Tabela 4).

O cajueiro comum possui um total de 16.347.083 *reads* e através da montagem, foram produzidos 54.457 *contigs* sendo que o maior *contig* possui 3.042 pb. O valor de N50 é 171 e a montagem foi feita com 33,4% dos *reads*. De acordo com o programa *Oases*, o número de transcritos é de 37.422 (Tabela 4).

Tanto o cajueiro CCP 76 (Gráfico 13) quanto o cajueiro comum (Gráfico 14) apresentaram 4 x de cobertura de acordo com as estatísticas do programa *Velvet*.

2.5.4 Montagem do transcriptoma utilizando o CLC Genomics Workbench

Os dados brutos também foram montados utilizando o programa montador *CLC Genomis Workbench*. Todos os parâmetros utilizados foram automáticos conforme sugerido pela interface gráfica do programa.

O cajueiro comum e o anão CCP 76 apresentaram aproximadamente 28% de AT e 21% de GC. O conteúdo de N, ou seja, bases não identificadas corretamente, foi de 0,3% para o cajueiro CCP 76 e 0,4% para o cajueiro comum (Tabela 5).

A montagem do cajueiro CCP 76 mostrou a presença de 36.746 *contigs* e o maior *contig* possui 6.582 pb e o valor de N50 foi de 672. O cajueiro comum por sua vez possui 16.899 *contigs* sendo que o maior *contig* possui 5.386 pb e o

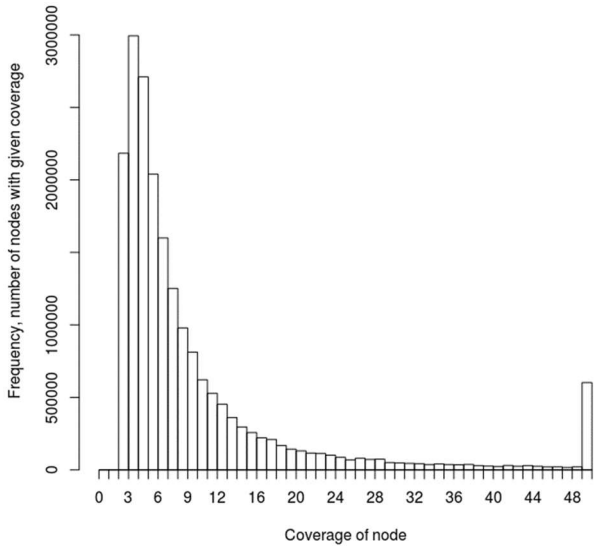
N50 é 534 (Tabela 6). A maioria dos contigs possuem tamanho menor que 1000 pb tanto no cajueiro CCP 76 (Gráfico 15) como no cajueiro comum (Gráfico 16).

Tabela 4 – Estatísticas de montagem do transcriptoma de cajueiro CCP 76 e cajueiro comum usando os programas *Velvet* e *Oases*.

Parâmetro	**CCP 76**	**Cajueiro comum**
Tamanho do K-mer	31	31
Cobertura esperada	4	4
Corte de cobertura	1,04	0,28
Tamanho mínimo de transcrito	100	100
N° *reads*	28.085.118	16.347.083
N° nós	132.193	64.201
N50	228	171
Tamanho do maior *contig*	3.773	3.042
Total *contigs*	117.667	54.457
Total transcritos	77.371	37.422
% *reads* utilizados	13.432.379 (47,8 %)	5.455.068 (33,4 %)

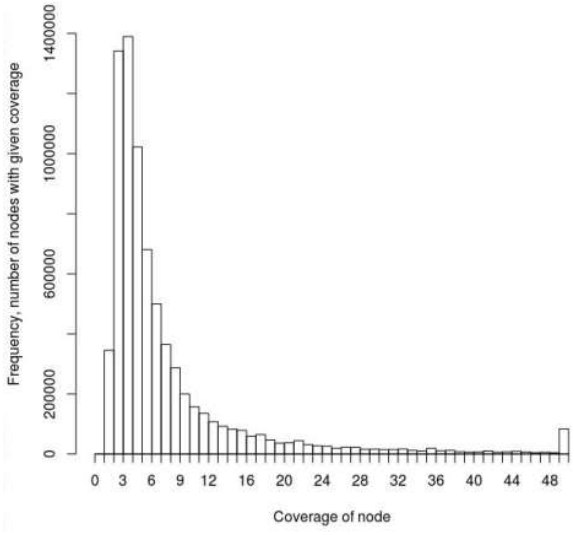
Fonte: O Autor.

Gráfico 13 - Resultado da montagem do transcriptoma do cajueiro CCP 76 utilizando o programa de montagem Velvet.



O gráfico foi feito utilizando o programa R *graphics* (*plotrix*) com base no arquivo *stats.txt* obtido na montagem do *Velvet*. **Fonte**: O Autor.

Gráfico 14 - Resultado da montagem do transcriptoma do cajueiro comum utilizando o programa de montagem Velvet.



O gráfico foi feito utilizando o programa R *graphics* (*plotrix*) com base no arquivo *stats.txt* obtido na montagem do *Velvet*. **Fonte**: O Autor.

Tabela 5 – Distribuição de nucleotídeos do transcriptoma do cajueiro anão e comum mapeados pelo programa CLC Genomics Workbench.

Nucleotídeo	CCP 76	Cajueiro comum
	Contagem (Frequência)	Contagem (Frequência)
Adenina (A)	5.800.208 (28.8%)	2.262.664 (28.3%)
Citosina (C)	4.265.589 (21.2%)	1.729.423 (21.6%)
Guanina (G)	4.226.901 (21.0%)	1.711.531 (21.4%)
Timina (T)	5.789.137 (28.7%)	2.253.995 (28.2%)
Qualquer nucleotídeo (N)	56.333 (0.3%)	32.373 (0.4%)

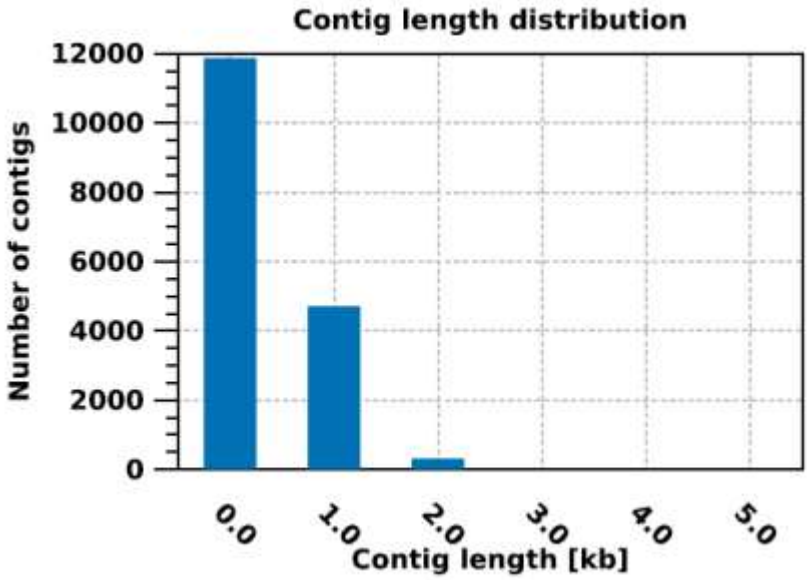
Fonte: O Autor.

Tabela 6 – Medidas de contigs do transcriptoma do cajueiro anão e comum mapeados pelo programa CLC Genomics Workbench.

Parâmetro	CCP 76	Comum
N75	382	329
N50	672	534
N25	1.184	908
Mínimo	200	200
Máximo	6.582	5.386
Médio	548	473
Contagem	36.746	16.899
Total	20.138.168	7.989.989

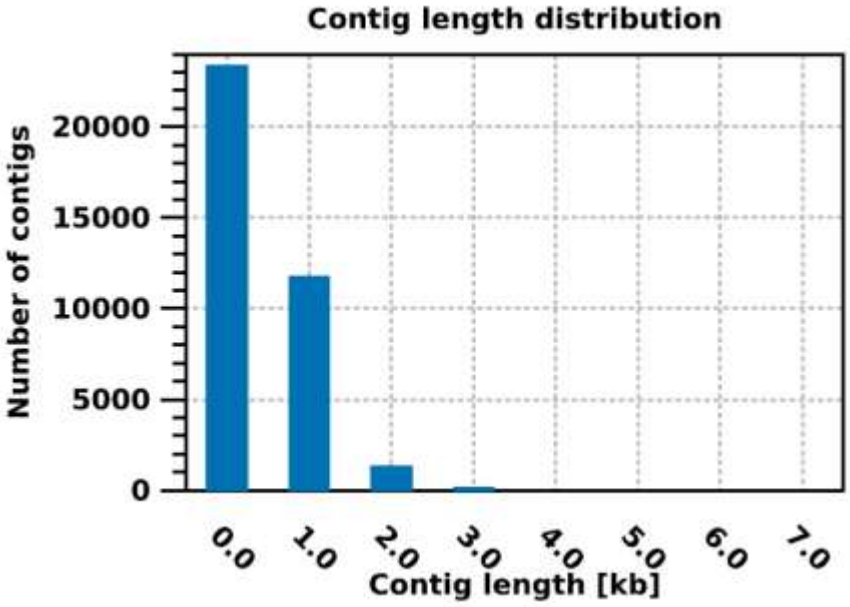
Fonte: O Autor.

Gráfico 15 – Distribuição do tamanho dos contigs no transcriptoma do cajueiro CCP 76 usando o programa CLC Genomics Workbench.



Fonte: O Autor.

Gráfico 16 – Distribuição do tamanho dos *contigs* no transcriptoma do cajueiro comum usando o programa *CLC Genomics Workbench*.



Fonte: O Autor.

2.6 DISCUSSÃO

A comparação dos transcritos entre organismos ou de um mesmo organismo sobre diferentes condições pode nos ajudar a compreender os diversos mecanismos moleculares relacionados às diferenças fenotípicas e á plasticidade metabólica. Ao longo dos anos, diferentes abordagens para estudos de transcritos foram desenvolvidas como a técnica de PCR em tempo real, microarranjo de DNA, análises de ESTs, sequenciamento de biblioteca de cDNA, análise serial da expressão gênica e, mais recentemente, o RNA-Seq.

O surgimento das novas tecnologias de sequenciamento massivo paralelo de ácidos nucléicos revolucionou os estudos genomicos e transcriptômicos, pois a velocidade de aquisição dos dados aumentou exponencialmente. Assim, em estudos transcriptomicos, as análises de ESTs foram substituídas pela técnica de RNA-Seq, onde toda a população de RNA é sequenciada através de uma biblioteca de cDNA. Deste modo, é notória a necessidade da obtenção de RNAs de boa qualidade para o sucesso das análises de transcritos. A qualidade de RNAs pode ser avaliada através de dois parâmetros, a presença de contaminantes e a integridade do RNA.

A relação entre as absorbâncias a 260 nm e 280 nm é um forte indicativo da presença de contaminantes e deve ser maior que 1,75 para que uma amostra seja considerada de boa qualidade e livre de contaminação de proteínas (SAMBROOK; FRITSCH; MANIATIS, 1989). Contudo, outros pesquisadores defendem que amostras que apresentem relação entre 1,6 e 2,0 já possuem qualidade satisfatória sendo que valores maiores que 2,0 indicam contaminação por fenol ou com outros tipos de álcoois (ROMANO, 1998). A quantificação do RNA de castanha de caju mostrou valores de concentração entre 680 a 1.454 µg/µL e uma relação das absorbâncias $A_{260/280}$ entre 1,8 e 2,0 sendo, portanto aceitáveis para análise transcriptômica.

A análise direta de qualidade dos dados brutos utilizando o programa *FastQC* demonstra que as sequências obtidas possuem perfil com escore *PHRED* superior a 30. O perfil gráfico do *FastQC*, observando-se o valor de *PHRED* para cada posição do *read* obtido, assemelha-se aos dados relatados por (SILVEIRA, 2012).

Após a constatação da qualidade dos *reads* obtidos, procedemos a montagem dos *Contigs* utilizando o programa *Velvet*. O *Velvet* é um *software* de montagem que sequências que utiliza o grafo de Bruijn e, ainda, oferece dados estatísticos (ZERBINO; BIRNEY, 2008). Um dos aspectos mais importantes na montagem de *Contigs* é a escolha da sobreposição ou *k-mer*. O valor de *k-mer* está relacionado com o tamanho e o número de *contigs* assim como o valor de N50. Altos valores de *k-mer* fazem com que os genes mais expressos apresentem *contigs* maiores e altos valores de N50, porém o número de *contigs* diminui, pois, os genes menos expressos não conseguem fechar os *gaps* virtuais. Por outro lado, baixos valores de *k-mer* aumentam o número de *contigs,* mas diminuem o tamanho médio dos *contigs* e o valor de N50.

Os parâmetros de montagem podem ser configurados para serem mais exigentes resultando em dados mais confiáveis, mas com uma quantidade menor de informação. Programas como o *VelvetOptimiser* foi desenvolvido para automatizar o processo de escolha de parâmetros, como o k-mer (ZERBINO, 2010). Com base no programa *VelvetOptimiser*, o melhor valor de *k-mer* para as bibliotecas de cajueiro comum e anão CCP 76 foi de 31 para uma cobertura esperada (Exp_cov) de 4 vezes e o maior valor de N50 possível.

Diante deste dilema que é bastante comum em montagem de transcriptoma *De novo*, novas estratégias têm sido desenvolvidas para a otimização do processo de montagem, uma delas é o método do múltiplo *k-mer*. Este método consiste em uma montagem inicial com um valor de *k-mer* alto e os *reads* não utilizados são montados separadamente com um valor de *k-mer* menor. Por fim, é feito um *pool* das sequências seguido pela remoção de *contigs* redundantes (SURGET-GROBA; MONTOYA-BURGOS, 2010).

Com base nos resultados obtidos, o cajueiro anão CCP 76 obteve uma melhor montagem quando comparada ao cajueiro comum, mas ambos tiveram apenas 4 x de cobertura. Isso se deve, entre outros fatores, ao fato de que o cajueiro anão possui cerca de 28 milhões de *reads* (quase 12 milhões a mais do que o cajueiro comum). Um grande número de *reads* auxilia a montagem reduzindo o número de *contigs* e aumentando os valores de N50 e cobertura média. Um exemplo é o transcriptoma *De novo* do eucalipto (*Eucalyptus grandis*)

utilizando a plataforma *Illumina*, onde 62 milhões de *reads* conseguiram montar aproximadamente 19 mil *contigs* com uma cobertura média de 37 x (MIZRACHI et al., 2010).

Numa tentativa de encontrar uma melhor montagem, os programas *Velvet* e *CLC Genomics WorkBench* foram comparados sendo que o último foi melhor em todas as bibliotecas testadas de acordo com os resultados obtidos. Um trabalho semelhante foi feito com o grão-de-bico (*Cicer arietinum* L.), onde se observou que, embora o N50 tenha sido melhor no *CLC Genomics Workbench*, a média do comprimento do *contig* e o número total de *contigs* foi muito maior que o esperado (GARG et al., 2011).

O resultado da montagem do *Velvet* foi melhorado após a utilização do programa *Oases*, o qual reduziu o número de *contigs*. O programa *Oases* foi desenvolvido especificamente para a montagem de transcriptomas *De novo* usando *reads* curtos, o qual leva em consideração a montagem feita pelo *Velvet* e explora o pareamento das sequências para produzir isoformas transcritas. Tem sido sugerido que a montagem pelo *Velvet* seguido pelo *Oases* produz os melhores *contigs*/transcritos (GARG et al., 2011).

Os critérios para avaliar as montagens de genomas estão em desenvolvimento, no entanto os padrões de avaliação sistemática da qualidade das montagens do transcriptoma ainda não foram estabelecidos (MARTIN; WANG, 2011). Sendo assim, a montagem do transcriptoma do cajueiro pode ser considerada como bem-sucedida uma vez que outras montagens de baixa cobertura terem sido encontradas na literatura.

2.7 CONCLUSÃO

A análise do transcriptoma do cajueiro comum e anão-precoce CCP 76 por RNA-Seq mostrou sequências de alta qualidade Phred e uma montagem satisfatória com cobertura média de 4x utilizando os programas de montagem *Velvet* e *CLC Genomics*. As sequências do cajueiro anão CCP 76 mostraram melhor qualidade em relação ao cajueiro comum, principalmente devido a um maior número de sequências obtidas. O programa de montagem *Velvet*, apesar de não ter se mostrado melhor que o *CLC Genomics* mostrou uma montagem de elevado nível com a vantagem de ser uma ferramenta com código fonte gratuito.

CAPÍTULO 3:
IDENTIFICAÇÃO DE MARCADORES SSR (*In silico*) NO TRANSCRIPTOMA DE SEMENTES DE CAJUEIRO COMUM E ANÃO CCP 76

3 IDENTIFICAÇÃO DE MARCADORES SSR (*IN SILICO*) NO TRANSCRIPTOMA DE SEMENTES DE CAJUEIRO COMUM E ANÃO CCP 76

3.1 INTRODUÇÃO

A moderna agropecuária tem exigido maior eficiência na produção e a redução dos custos de produção. Deste modo, técnicas de biologia molecular que auxiliem nos programas de melhoramento genético das espécies são cada vez mais desejadas. Segundo Ferreira e Gattaplagia (1998) qualquer fenótipo molecular oriundo de um gene expresso ou segmento específicos do DNA correspondentes às regiões expressas, ou não, do genoma pode ser definido como marcador molecular (FERREIRA; GATTAPLAGIA, 1998). Marcadores moleculares que apresentam comportamento mendeliano simples podem ser empregados como marcadores genéticos.

Os marcadores moleculares podem ser divididos em dois grandes grupos, marcadores baseados em proteínas e marcadores de DNA. As aloenzimas e isoenzimas são os principais representantes dos marcadores moleculares de proteínas (PATERSON; TANKSLEY; SORRELLS, 1991). Já com relação aos marcadores de DNA, existem diversos tipos, dentre os quais destacamos: *Amplified Fragment Length Polymorphism*– AFLP (ZABEAU; VOS, 1993; VOS et al., 1995), *Random Amplified Polymorphic DNA* – RAPD (WILLIAMS et al., 1990), *Restriction Fragment Length Polymorphism DNA* – RAPD (GRODZICKER et al., 1974), *Single Nucleotide Polymorphism* – SNP, Minissatélites ou *Variable Number of Tandem Repeats* (VNTR) (JEFFREYS; WILSON; THEIN, 1985) e Microssatélites (HAMADA; PETRINO; KAKUNAGA, 1982; WEBER, 1990). Estes marcadores moleculares são classificados como baseados em hibridização (RFLP), em amplificação por PCR (RAPD, AFLP, SSR) e os polimorfismo de sequencia, como os SNPs (VARSHNEY et al., 2007; SEHGAL; RAINA, 2008).

Segundo Sansaloni (2008), os marcadores moleculares internacionalmente adotados para a identificação individual de plantas e animais são baseados na amplificação de segmentos curtos de DNA de 1 a 6 pares de bases repetidos em tandem, denominados de microssatélites ou STR ou, ainda, SSR (SANSALONI, 2008). Contudo, é importante ressaltar a crescente

importância dos marcadores de polimorfismo de nucleotídeo simples. São marcadores codominantes, altamente multialélicos, com resultados reprodutivos, extensiva cobertura no genoma, apresentando maior conteúdo informativo por loco gênico entre todas as classes de marcadores moleculares e viáveis de automação (GOLDSTEIN; SCHLÖTTERER, 1999; PARIDA et al., 2009). Os microssatélites são altamente frequentes no genoma humano, sendo que ocorrem a cada 1.000 a 2.000 pares de bases. Tais características permitem analisar desde indivíduos até espécies proximamente relacionadas.

Microssatélites são classificados de diversas maneiras, dentre as quais destacamos a classificação por tamanho e tipo de unidade de repetição, e pela localização no genoma. A maior parte dos microssatélites identificados é nuclear, contudo há relatos da distribuição dos SSRs nos genomas mitocondriais e cloroplastiais. Baseados na sua localização genômica, os SSRs podem ser classificados em nuSSR (nuclares), mtSSR (mitocondrial) ou cpSSR (cloroplastial) (KALIA et al., 2011). Dependendo da quantidade de nucleotídeos por unidade de repetição, os SSRs podem ser classificados em mono, di, tri, tetra, penta ou hexanucleotídicos. Wang e colaboradores (2009) como simples perfeito ou simples imperfeito (WANG; BARKLEY; JENKINS, 2009). Os arranjos perfeitos são compostos pelas unidades de repetição do motivo, enquanto que os imperfeitos são interrompidos por motivos não participantes da unidade de repetição.

Wang e colaboradores (2009) apontam quatro possibilidades para o polimorfismo dos *loci* contendo os SSRs, como o deslizamento da DNA polimerase III durante o processo de replicação de regiões contendo os motivos, erros do sistema de reparo pós-replicacional, a recombinação da dupla fita (como nos casos do *crossing-over*) ou retrotransposição.

A análise da distribuição dos SSRs ao longo do genoma de eucariontes tem revelado que a maior parte dos SSRs encontram em região não codificante (KALIA, RAI, et al., 2011). Nos cereais, como milho, trigo, sorgo, arroz, somente cerca de 1,5 a 7,5% dos SSRs estão localizados em etiquetas de sequencias transcritas – EST (KANTETY et al., 2002; THIEL et al., 2003). Li e colaboradores (2004) demonstraram que nos RNA transcritos, as regiões não traduzidas, como

os introns, 5´-UTR e 3´-UTR, possuem mais regiões repetidas que os éxons (LI et al., 2004).

Os dinucleotídeos são mais comuns em muitas espécies e são mais frequentes em regiões não codificantes que em regiões codificadoras (LI et al., 2004); (WANG; BARKLEY; JENKINS, 2009). Os motivos trinucleotídicos são mais frequentes nos éxons, devido ao frame de leitura dos códons (LI et al., 2004). Nas plantas, o triplete AAG é mais comum, contudo nos cereais o CCG é mais frequente (VARSHNEY et al., 2007; THIEL et al., 2003), uma característica das monocotiledôneas (KALIA et al., 2011). Apesar da frequência destes microssatélites, Kalia e colaboradores (2011) relatam que em plantas o papel funcional dos SSRs é pouco estudado e pobremente compreendido.

Os SSRs têm sido escolhidos como o marcador molecular para diversas aplicações em estudos de plantas devido à hipervariabilidade, codominância e extensiva cobertura no genoma (KALIA et al., 2011). As abordagens em plantas estão relacionadas ao acesso à variabilidade genética de coleções de germoplasma, e desta forma orientando a escolha apropriada de genótipos para cruzamento, mapeamento e etiquetagem de genes, análises de *loci* com informações quantitativas para características agronômicas e seleção assistida por marcadores moleculares (KALIA et al., 2011).

Após a identificação de uma região com motivos de microssatélite, duas outras condições tornam um *loci* propício a ser utilizado nas abordagens supracitadas. Incialmente é importante que as regiões flanqueadoras do loci repetitivo sejam conservadas o bastante para poderem ser utilizadas para a construção de primers para a reação de PCR. Além disso, é importante haver polimorfismo para o *loci*. Vale salientar que o conteúdo informativo do SSRs não está na diferença entre o número de unidade de repetição de um alelo, mas sim na frequência deste alelo em uma população. Deste modo, após a identificação, o SSR deverá ser validado, os tipos alélicos determinados e, por fim, a frequência destes alelos deve ser identificada.

Apesar dos dados moleculares de cajueiro serem extremamente escassos, já foram identificados 11 SSRs para cajueiro (CAVALCANTI; WILKINSON, 2007). Estes marcadores têm sido utilizados para análises de diversidade genética e QTL (Locus de Caracteres Quantitativos) de cajueiro e outras análises em espécies relacionadas.

3.2 OBJETIVOS

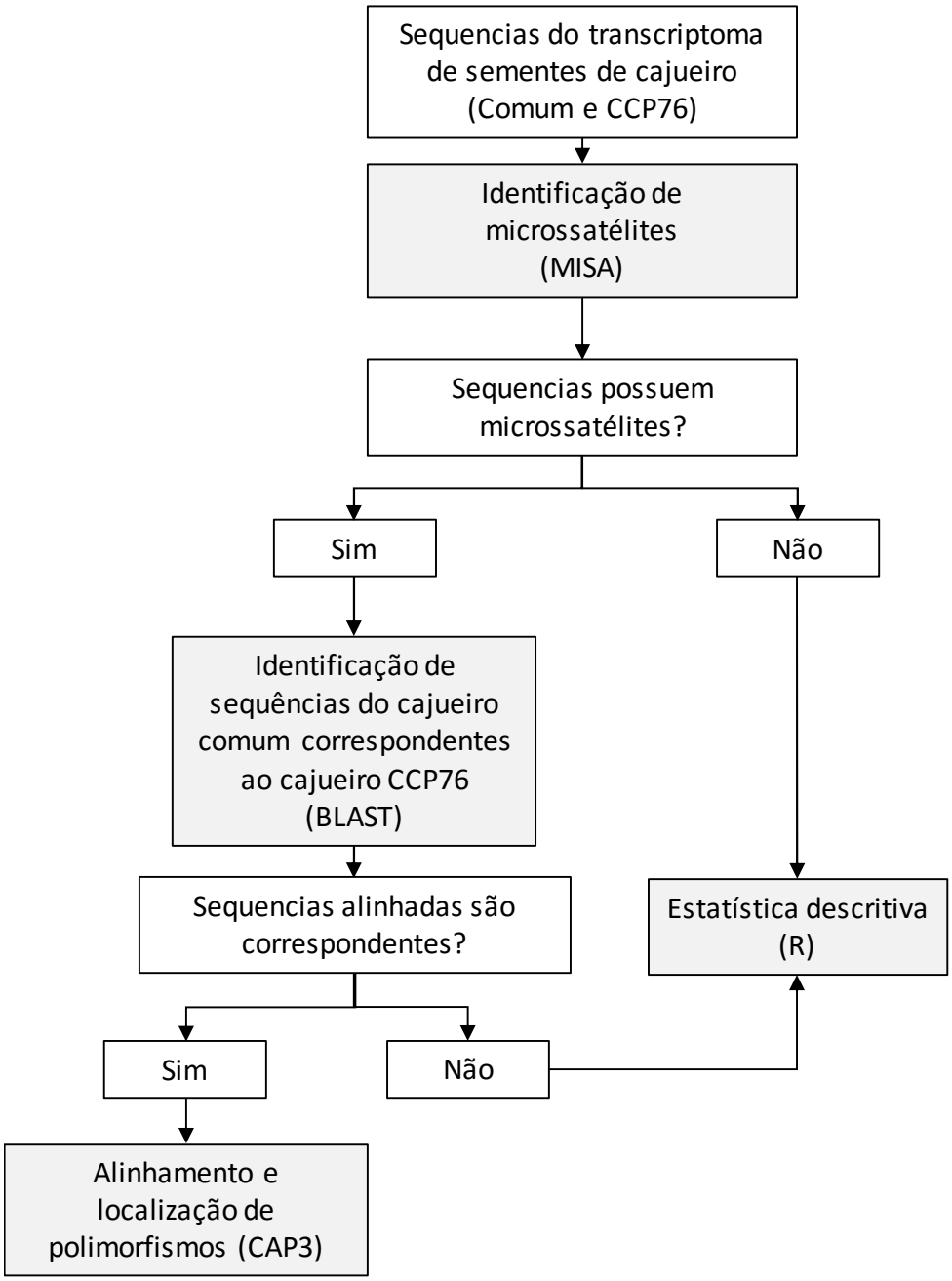
3.2.1 Geral

Identificar a presença de microssatélites no transcriptoma de cajueiro comum e CCP 76 utilizando ferramentas de bioinformática.

3.2.2 Específicos

- Observar a frequência dos microssatélites do tipo di, tri ou tetranucleotídeos no cajueiro comum e CCP 76;
- Utilizar a ferramenta BLAST para investigar os microssatélites que se apresentam nos *contigs* compartilhados entre o cajueiro comum e no CCP 76;
- Usar a ferramenta CAP3 para alinhar os *contigs* e fazer uma inspeção visual do local onde se encontram os microssatélites.

3.3 ESTRATÉGIA EXPERIMENTAL



3.4 METODOLOGIA

As sequências dos transcritos de cajueiro comum e CCP 76 foram obtidas utilizando o programa *Velvet* seguido pelo *Oases*, conforme descrito no capítulo 2. Um total de 77.371 transcritos do cajueiro CCP 76 e 37.422 transcritos do cajueiro comum foi utilizado como arquivo de entrada para a busca por microssatélites.

O mapeamento dos microssatélites se deu pela utilização da ferramenta MISA (MIcroSAtellite) (THIEL et al., 2003). O script MISA, desenvolvido em Perl, encontra-se depositado no seguinte endereço eletrônico: http://pgrc.ipk-gatersleben.de/misa/.

Para saber quais microssatélites estavam em ambos os genótipos de cajueiro, todos os *contigs* que continham microssatélites foram alinhados utilizando o algoritmo BLAST (ALTSCHUL et al., 1990), utilizando as sequências do transcriptoma de castanhas de CCP 76 como banco de dados.

A detecção de polimorfismos nos SSR presentes em ambas as sequencias de cajueiro foi feita pelo alinhamento dos *contigs* contendo microssatélites foi feito utilizando o programa de alinhamento e montagem CAP3 (HUANG; MADAN, 1999). As análises estatísticas descritivas foram feitas utilizando comandos Shell, LibreOffice Calc e Microsoft Excel™.

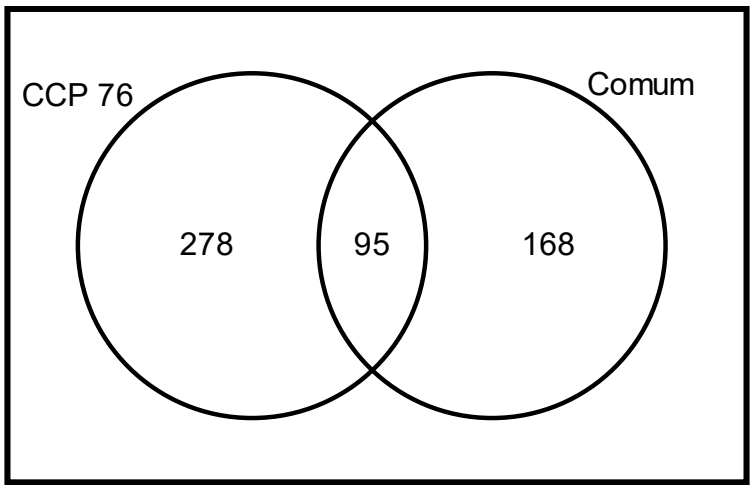
3.5 RESULTADOS

A busca por regiões SSR pelo MISA identificou a presença de 263 SSR para o cajueiro comum e 373 para CCP 76. O alinhamento local dos *contigs* contendo os SSR identificados revelou que 95 SSRs são compartilhados por ambos os genótipos (Figura 7).

Foram identificados nas bibliotecas transcritos de sementes em formação de cajueiro anão e comum a presença de SSRs com motivos di, tri e tetranucleotidicos. Os motivos mais abundantes foram do tipo trinucleotideos, sendo 200 para o comum e 237 para o CCP 76 com variação de 6 a 11 unidades de repetição (Gráfico 17). Os motivos dinucleotídicos foram os segundo mais abundante, sendo possível detectar 53 e 98 SSRs para os genótipos de cajueiro comum e CCP 76, respectivamente. Os tamanhos destes microssatélites variaram de 9 a 16 unidades de repetição. Os poucos microssatélites do tipo tetranucleotidicos identificados, apenas 10 para o comum e 8 para CCP 76, variam de 5 a 7 unidades de repetição.

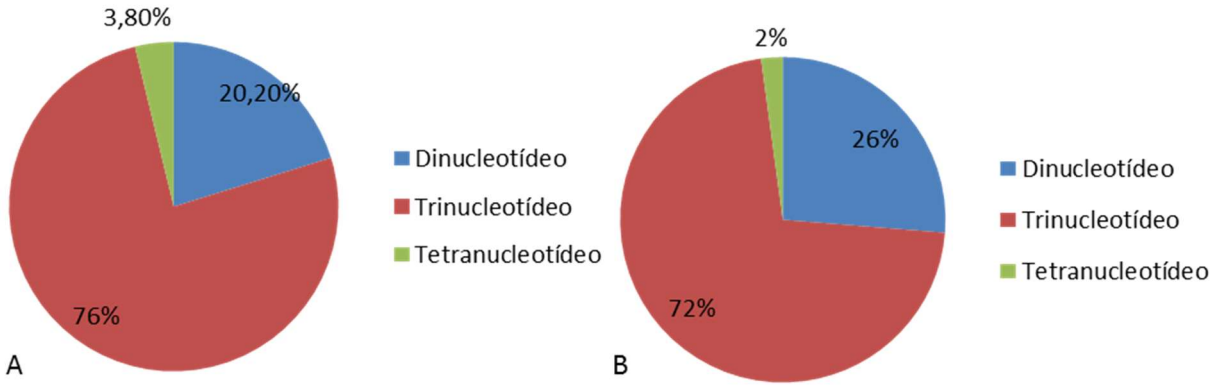
Os motivos encontrados foram separados em grupos de acordo com dados de Yu e colaboradores (2011) (YU et al*.,* 2011; ZHAO et al*.,* 2012). Os dinucleotideos foram alocados em quatro grupos, sendo observada uma maior distribuição de motivos para grupo 3 (CT/TC/GA/AG) em ambos os genótipos (Gráfico 18). Vale ressaltar que não houve a presença de motivos para o grupo 1 (GC/CG) nos ecótipos em estudo. A análise da distribuição do número de unidades de repetição no grupo 3 revela que o cajueiro anão possui mais SSR com 9 ou 12 unidades de repetição, enquanto identificamos no cajueiro comum a maior frequência de 9 ou 15 (Gráfico 19).

Figura 7 - Diagrama de Venn demonstrando o numero de SSRs totais e compartilhadas para o cajueiro anão CCP 76 e cajueiro comum.



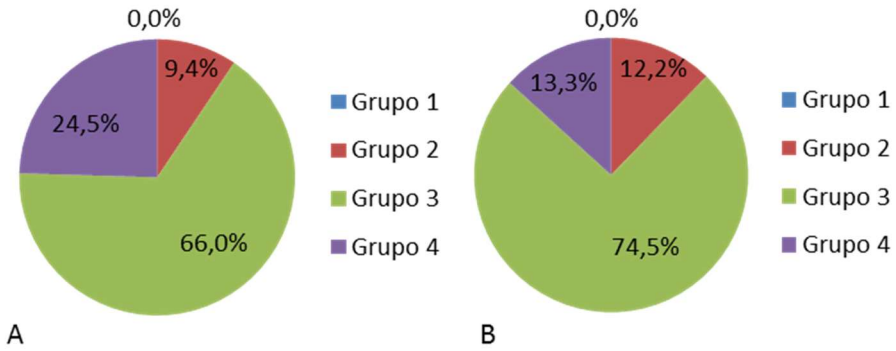
Fonte: O Autor.

Gráfico 17 - Distribuição dos microssatelites correspondendo aos motivos di, tri e tetranucleotídicos no transcriptoma do cajueiro.



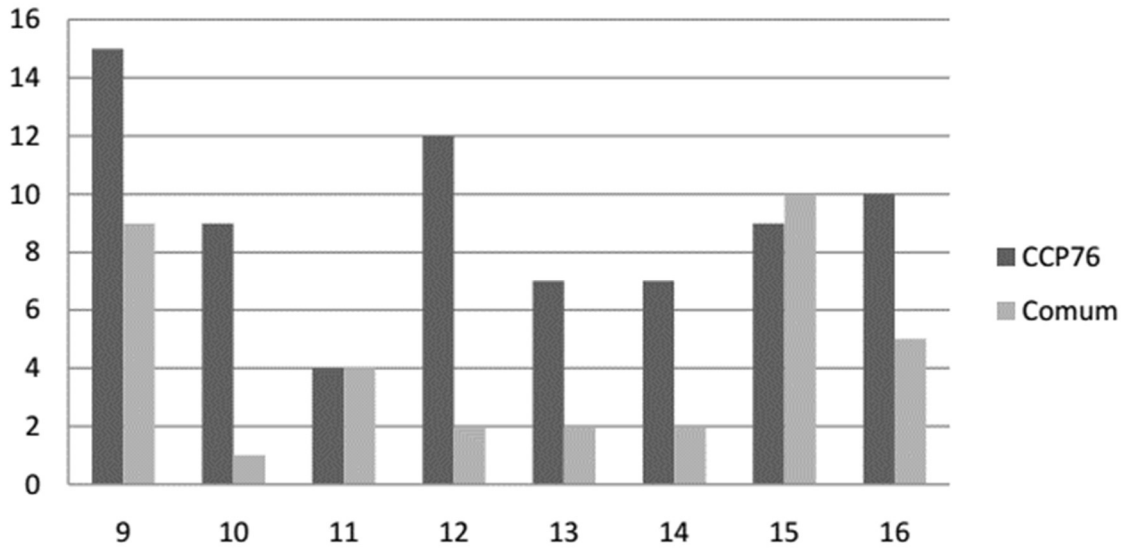
(A) distribuição dos microssatelites no cajueiro comum; **(B)** distribuição dos microssatelites no cajueiro CCP 76. **Fonte**: O Autor.

Gráfico 18 - Distribuição dos microssatelites correspondendo aos motivos do tipo dinucleotídicos encontrados no transcriptoma do cajueiro.



(A) distribuição dos microssatelites do tipo dinucleotideos no cajueiro comum; **(B)** distribuição dos microssatelites do tipo dinucleotideos no cajueiro CCP 76; Grupo 1 indica os motivos GC/CG, grupo 2 CA/AC/GT/TG, grupo 3 CT/TC/GA/AG e grupo 4 AT/TA. **Fonte**: O Autor.

Gráfico 19 - Distribuição das unidades de repetição para o grupo 3 dos motivos do tipo dinucleotidícos.

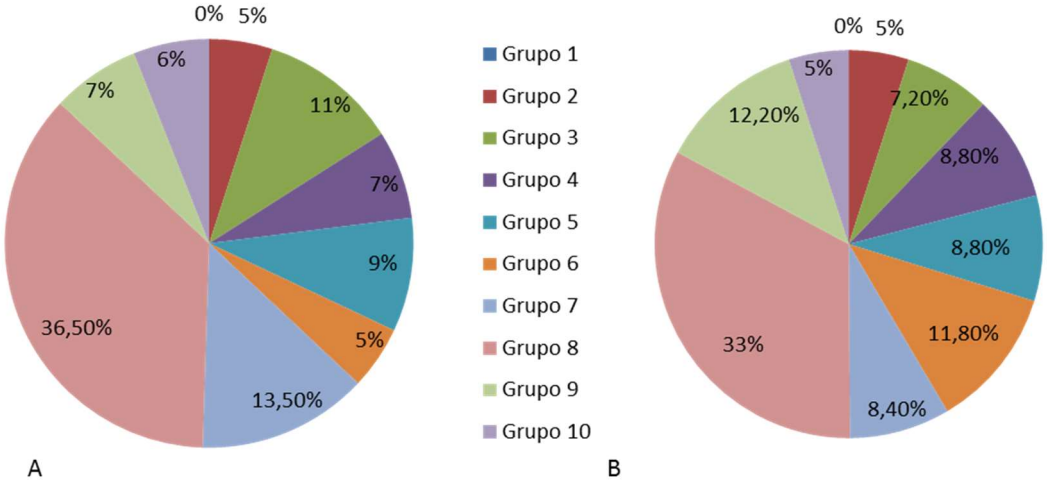


Grupo 3 (CT/TC/GA/AG); As unidades variaram de 9 a 16 repetições. **Fonte**: O Autor.

Os motivos do tipo trinucleotideos foram distribuídos em 10 grupos, ainda conforme proposto por Yu e colaboradores (2011) e Zhao e colaboradores (2012). Houve uma maior abundancia de motivos para o grupo 8 (AAG/AGA/GAA/CTT/TTC/TCT), sendo este grupo correspondente a 36,5% e 33% do total de SSR trinucleotídicos mapeados para o genótipo comum e CCP 76 (Gráfico 20). Ainda, pode-se notar que não houve a presença do grupo 1 (GGC/GCG/CGG/CCG/CGC/GCC). Analises a respeito da distribuição das unidades de repetição para o grupo 8 demonstram, para ambos os genótipos, uma maior frequência dos microssatélites com 6 e 7 unidades de repetição (Gráfico 20 e Gráfico 21). Além disso, os resultados revelam que não houve a presença de SSR com 10 e 11 unidades de repetição para o cajueiro anão.

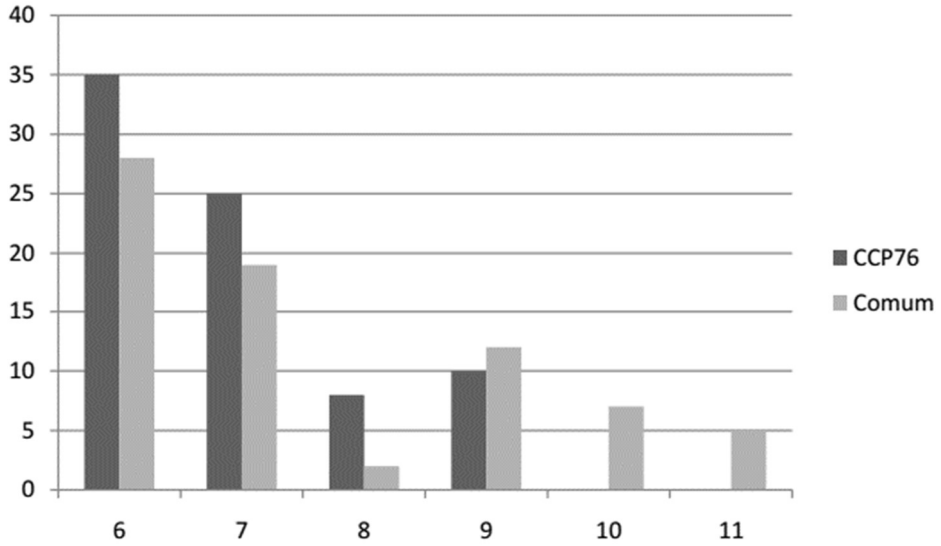
Por fim, vale ressaltar que dos 95 SSRs compartilhados entre os genótipos anão e comum (Apêndice A), três demonstraram diferenças entre os genótipos avaliados, sendo este um forte indício da identificação de *loci* com polimorfismos. A identificação de SSR polimórficos viabiliza o estudo de diversidade genética e possui uma forte aplicação nos programas de melhoramento genético. É importante registrar que os *loci* polimórficos foram identificados em motivos trinucleotidicos (Figura 8 e Gráfico 22).

Gráfico 20 - Distribuição dos microssatelites com motivos do tipo trinucleotídicos encontrados no transcriptoma do cajueiro.



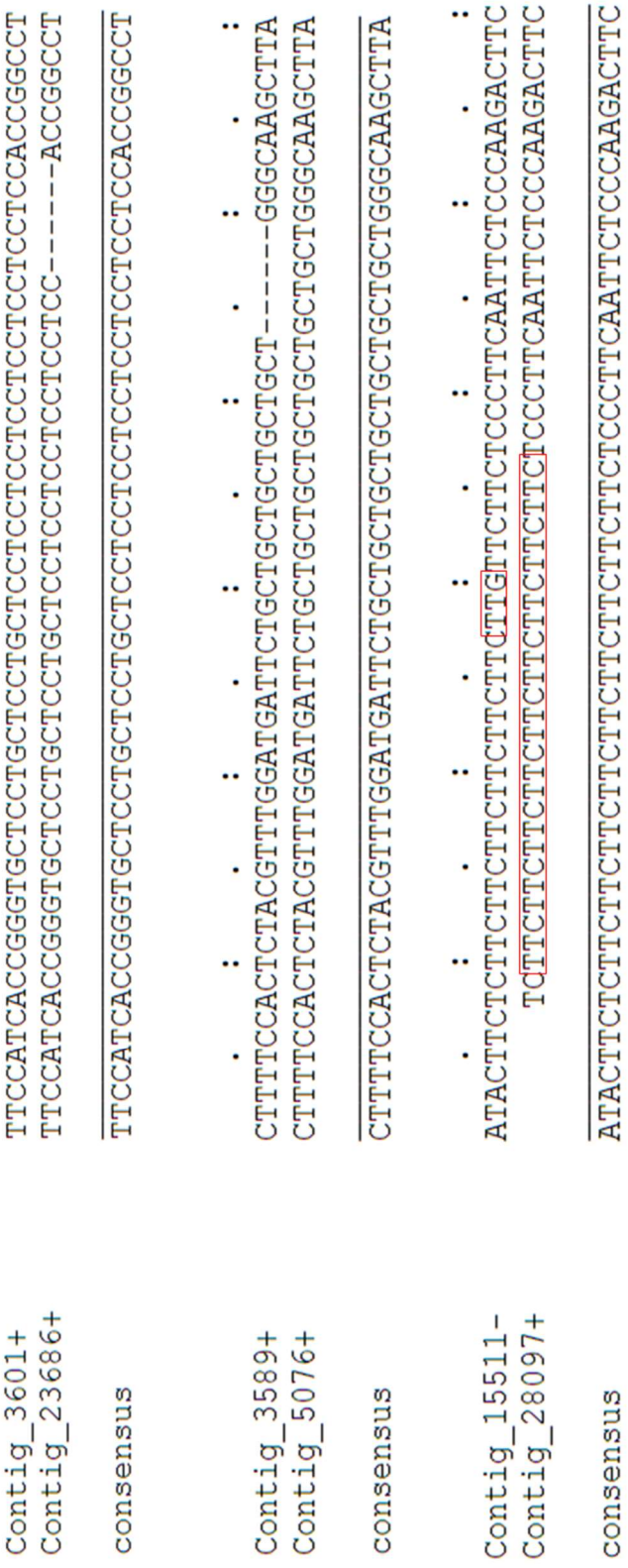
(A) microssatelites com motivos do tipo trinucleotideos encontrados no transcriptoma do cajueiro comum e **(B)** microssatelites com motivos do tipo trinucleotideos encontrados no transcriptoma do cajueiro CCP 76. Grupo 1 indica os motivos GGC/GCG/CGG/CCG/CGC/GCC, grupo 2 ACG/CGA/GAC/TGC/GCT/CTG, grupo 3 AGC/GCA/CAG/TCG/CGT/GTC e grupo 4 ACC/CAC/CCA/TGG/GTG/GGT, grupo 5 AGG/GGA/GAG/TCC/CCT/CTC, grupo 6 AGT/GTA/TAG/ACT/CAT/ATC, grupo 7 ATG/TGA/GAT/TAC/ACT/CTA, grupo 8 AAG/AGA/GAA/CTT/TTC/TCT, grupo 9 AAC/ACA/CAA/TTG/TGT/GTT e grupo 10 AAT/ATA/TAA/ATT/TTA/TAT. **Fonte**: O Autor.

Gráfico 21 - Distribuição das unidades de repetição para o grupo 8 dos motivos do tipo triucleotídicos.



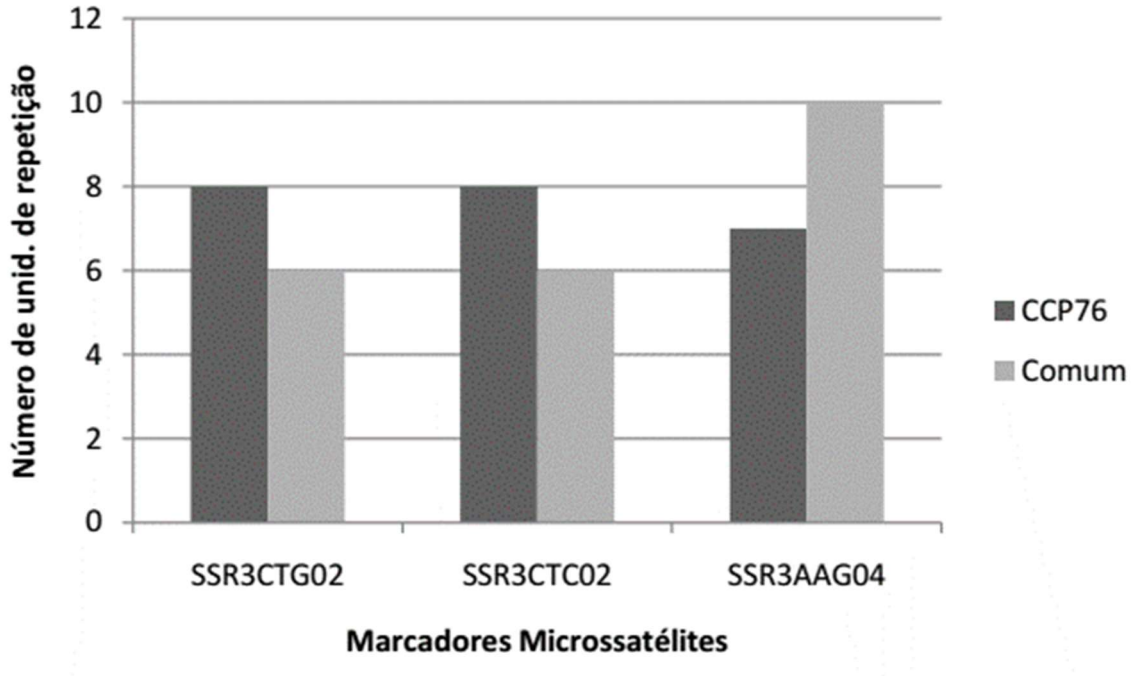
Grupo 8 (AAG/AGA/GAA/CTT/TTC/TCT). As unidades variaram de 6 a 11 repetições. **Fonte**: O Autor.

Figura 8 – Alinhamento das sequencias contendo microssatélites polimórficos utilizando o programa CAP3.



Fonte: O Autor.

Gráfico 22 - Microssatélites polimórficos dos genótipos CCP 76 e cajueiro comum.



Fonte: O Autor.

3.6 DISCUSSÃO

Os microssatélites possuem diversas características que os tornam marcadores moleculares de grande relevância científica e tecnológica. Contudo, alguns fatores como o tempo necessário para criação de bibliotecas e o alto custo envolvido nesse desenvolvimento são entraves para a maior utilização dos marcadores SSR (CHO et al., 2010; ZALAPA et al., 2012). Amiruddin e colaboradores (2012) relatam que SSRs representam uma rica fonte de marcadores moleculares devido à alta taxa de mutação causada ao longo da fita de DNA nessas regiões (AMIRUDDIN et al., 2012). Ainda, acredita-se que os microssatélites são importantes para evolução dos genomas devido a estímulos de variabilidade genética e influencia na expressão de genes (LI et al., 2002; KALIA et al., 2011). Os microssatélites são importantes para o desenvolvimento de estratégias para conservação da espécie (KALIA et al., 2011). Assim, a identificação de SSRs para a espécie estudada representa importante fonte de dados para o desenvolvimento de estratégias de identificação, conservação e relações evolutivas entre os genótipos.

A estratégia tradicional para criação de mapas de SSRs seguia o fluxo geral a seguir: digestão do material, hibridização, clonagem, PCR e sequenciamento (EDWARDS et al., 1996; ZANE; BARGELLONI; PATARNELLO, 2002; KALIA et al., 2011). Com o advento da nova geração de sequênciamento (NGS) para o estudo do genoma e transcriptoma, bem como os avanços de bioinformática, têm permitido uma rápida varredura de SSRs em diferentes espécies (YAN et al., 2011; OUYANG et al., 2011). Isso devido ao menor tempo e ao reduzido número de etapas envolvidas para geração destes dados (ZALAPA et al., 2012). Estudo recente relata a utilização de NGS para o mapeamento de microssatélites em *Vaccinium macrocarpon* Ait. (GEORGI et al., 2012), *Mikania micranta* (YAN et al., 2011), *Sonncratia caseolaris* (OUYANG et al., 2011), *Ipomoea batatas* (WANG et al., 2010) e *Cicer arietinum* (GARG et al., 2011).

De modo geral, microssatélites em vegetais são repetições dinucleotidicas com motivos AG/CT (KALIA et al., 2011). Contudo, quando se trabalha com regiões codificadoras os motivos de repetição tendem a ser do tipo trinucleotidicos (LI et al., 2002). Tal afirmação corrobora com resultados observados no presente estudo, onde se pôde observar uma maior distribuição deste tipo de motivos para ambos os

genótipos, 72% dos SSRs para anão e 76% para cajueiro comum. Além disso, a presença de motivos do tipo di e tetranucleotideos podem estar presentes em íntrons ou das regiões não traduzidas á montante e à jusante da região codificante (LIU et al., 1999). Outro aspecto importante que fora observado no presente estudo é o fato de que a maioria dos SSRs dinucleotídicos identificados, as unidades de repetição são múltiplas de 3. Postulamos que esse fato está relacionado aos *frames* de leitura das regiões codificantes. Ainda, a ausência ou baixa frequência de motivos do tipo dinucleotideos e trinucleotideos para grupo 1 em ambos os genótipos, parece ser uma tendência para alguns vegetais (ZHAO; PRAKASH; HE, 2012).

Trabalho realizado Cavalcanti e Wilkinson (2007) relata a identificação de 20 microssatélites, dos quais 11 são polimóficos, para *Anacardium occidentale* seguindo a estratégia tradicional de utilização de sequenciamento de Sanger (CAVALCANTI; WILKINSON, 2007). No presente trabalho foram identificados mais de 600 SSRs no transcriptoma de sementes em formação de cajueiro. A utilização de abordagens diferentes, Sanger contra NGS, justifica a maior identificação de SSRs do presente trabalho.

O pequeno número de regiões polimórficas encontradas no presente estudo deve-se ao fato dos SSRs mapeados são oriundos de transcriptoma, portanto pertencentes às regiões codificantes do genoma do cajueiro, logo apresentam baixo grau de polimorfismo (LI et al., 2002; KALIA et al., 2011). Além disso, os dados encontrados são detecções de variação na porção expressa do genoma, assim, representando uma importante informação sobre os genótipos estudados.

Por fim, os microssatélites aqui encontrados podem ser usados para outros genótipos da espécie em estudo ou para espécies relacionadas por serem oriundos de regiões codificadoras e, portanto, possuírem regiões flanqueadoras de alta conservação (LI et al., 2002).

3.7 CONCLUSÃO

Microssatélites do tipo di, tri e tetranucleotídeos foram encontrados no transcriptoma do cajueiro comum e CCP 76 utilizando ferramenta *in silico*. Embora cerca de 300 SSR tenham sido encontrados, apenas 95 estão presentes em sequências compartilhados entre o cajueiro comum e CCP 76. Destas, encontramos três novos microssatélites polimórficos, podendo ser utilizados como marcadores moleculares.

CAPÍTULO 4:

ANOTAÇÃO FUNCIONAL DO TRANSCRIPTOMA DE SEMENTES DO CAJUEIRO

4 ANOTAÇÃO FUNCIONAL DO TRANSCRIPTOMA DE SEMENTES DO CAJUEIRO

4.1 INTRODUÇÃO

O processo de anotação de um genoma consiste na análise e interpretação dos dados de sequenciamento com o intuito de colocá-lo no contexto de nossa compreensão dos processos biológicos (STEIN, 2001). Desta forma, fica claro que sem a anotação funcional a sequência de um genoma ou transcriptoma não passa de um enorme conjunto de letras armazenadas em um disco rígido.

A anotação de um genoma, segundo Stein (2001), pode ser dividida em três níveis. A anotação a nível de nucleotídeo consiste na localização de ORFs, genes, microssatélites, tRNAs, rRNAs, etc. A identificação de genes, proteínas, e sua função é considerada uma anotação a nível de proteína. Por fim, o processo de anotação mais complexa de um genoma envolve a construção de blocos de genes e proteínas relacionados ao ciclo celular, morte celular, metabolismo, entre outros e faz parte da anotação a nível de processos.

Por outro lado, em um transcriptoma, temos a possibilidade de estudar níveis de expressão gênica com base na cobertura de um determinado gene. Esse índice de expressão é chamado de *reads per kilobase per milion* (RPKM) (MORTAZAVI et al., 2008).

Uma das ferramentas mais importantes para comparação rápida de sequencias é o algoritmo BLAST (*Basic Local Alignment Search Tool*) (ALTSCHUL et al., 1990). Essa ferramenta simples e robusta é capaz de comparar sequências de proteína/proteína (BLASTp), DNA/DNA (BLASTn), DNA/proteína (BLASTx) e proteína/DNA (tBLASTn). O programa BLAST pode ser executado online acessando a página do *National Center for Biotechnology Information* – NCBI (www.ncbi.nlm.nih.gov). Obviamente não convém fazer buscas online de 11 mil genes de um transcriptoma um por um. Sendo assim, o programa BLAST pode ser executado na plataforma *Unix* utilizando linhas de comando no *Shell*. Para isso é necessário fazer download do banco de dados para que o BLAST seja feito localmente de acordo com

os parâmetros ajustados pelo usuário.

O objetivo do BLAST é simples, encontrar uma sequência homóloga em outro organismo que possua uma função conhecida. Torna-se então necessário a criação de um banco de dados que associe proteínas e suas respectivas funções biológicas. Esse foi o principal objetivo do *Gene Ontology Consortium* (GO), que pode ser acessado pelo site (http://www.geneontology.org). Os termos do *Gene Ontology* são agrupados em três grandes categorias: Função molecular, processo biológico e componente celular (ASHBURNER et al., 2000).

Para visualização de vias metabólicas, uma das ferramentas mais utilizadas é o KEGG (*Kyoto Encyclopedia of Genes and Genomes*) *Pathways* (KANEHISA et al., 2006). Para a visualização de mapas KEGG é necessário ter os termos de *KEGG Othology* (KO) que podem ser obtidos pelo *IDMapping* do *UniProt* o qual converte IDs do *UniProt* em termos de KO. A anotação funcional pode, no entanto, ser realizada através do programa *Blast2go* que possui uma interface gráfica sendo capaz de realizar BLAST, busca de termos de GO e visualização de mapas KEGG (CONESA; GÖTZ, 2008).

4.2 OBJETIVOS

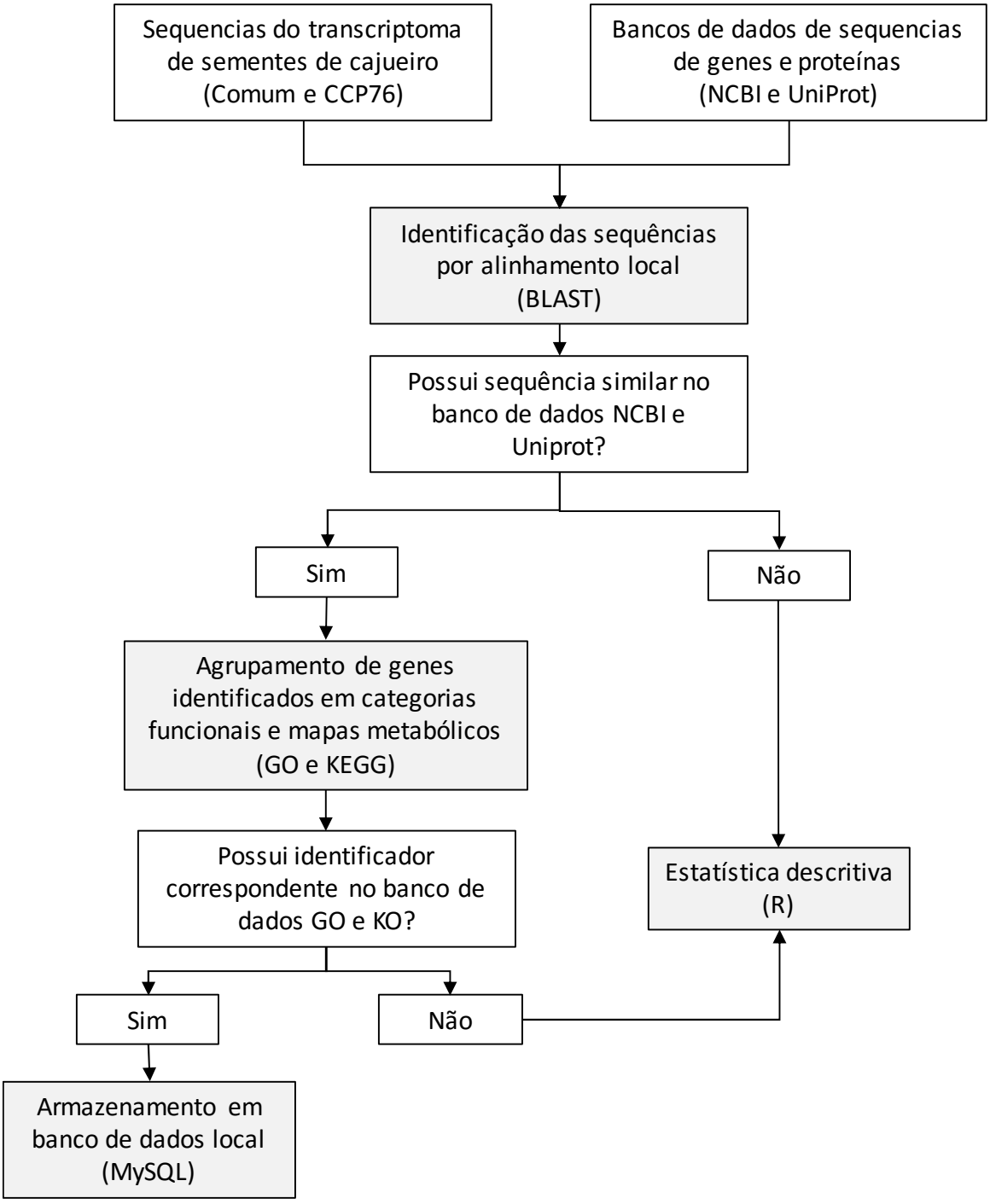
4.2.1 Geral

Identificar os principais genes expressos no transcriptoma de sementes em maturação de cajueiro comum e anão CCP 76 com base em alinhamento com genes ortólogos presentes em bancos de dados.

4.2.2 Específicos

- Alinhar as sequencias do transcriptoma do cajueiro com os bancos de dados do NCBI e UniProt utilizando o programa BLAST.
- Agrupar as sequências identificadas do transcriptoma de sementes de cajueiro em categorias funcionais do Gene Ontology.
- Mapear as sequências identificadas em mapas metabólicos utilizando o KEGG *Pathways*.

4.3 ESTRATÉGIA EXPERIMENTAL



4.4 METODOLOGIA

4.4.1 Identificação de genes pelo BLAST

Para a detecção de genes baseados em sequências ortólogas, foi realizado o programa BLAST. O banco de dados do NCBInr e do *SwissProt* foram baixados pelo site FTP do NCBI e *Uniprot*, respectivamente, acessados em outubro de 2012. Os parâmetros do BLAST incluem um BLASTX, um *cutoff* de *E-value* para $1x10^{-5}$ e $1x10^{-30}$, 8 núcleos de processamento e saída no formato tabular (.txt). Os comandos de Shell AWK, SORT e UNIQ foram utilizados para selecionar uma lista de valores de *Score* e *Identity* os quais resultaram em gráficos criados pelo módulo *Plotrix* do programa R.

O resultado do BLAST no formato tabular (.txt) foi exportado para um banco de dados criado no MySQL para facilitar buscas avançadas.

4.4.2 Anotação funcional pelo Gene Ontology

Inicialmente foi obtida uma lista de IDs do resultado do BLAST. Para obter a lista de IDs, foram utilizados os comandos do Shell SED, AWK, SORT e UNIQ.

Para a obtenção dos IDs de GO, foi utilizado o algoritmo *AmiGO* disponível online no site (www.geneontology.org) com base nos IDs contidos no resultado do BLAST.

A lista contendo os IDs e seus respectivos termos de GO foram adicionados ao MySQL e mesclados com o banco de dados contendo o resultado do BLAST utilizando a função "*Join table*". Em seguida foi feita uma busca no MySQL pelo número de registros contendo as principais categorias de GO (*Molecular Function, Cellular Component* e *Biological Process*) e exportados para criação de um gráfico de pizza utilizando o Microsoft Excel.

4.4.3 Obtenção de vias metabólicas no KEGG Pathways

Inicialmente foi obtida uma lista de IDs do resultado do BLAST usando os comandos do Shell SED, AWK, SORT e UNIQ. Para obter os números de KO, foi utilizada a opção *IDMapping* disponível online no site do *UniProt* (www.uniprot.org) com base nos IDs contidos no resultado do BLAST.

A identificação de enzimas em mapas de vias metabólicas foi feita utilizando o módulo *KEGG Orthology Database* (www.genome.jp/kegg/ko.html). Dezenas de vias metabólicas foram obtidas e destas quatro vias do metabolismo central (Glicólise, Ciclo de Krebs, CTE e Biossíntese de Ácidos Graxos) foram mostrados nos resultados como exemplo.

4.5 RESULTADOS

Para identificar a possível função dos transcritos, eles foram comparados contra o banco de dados não redundantes do *Swiss-Prot* disponível na base de dados do *UniProt*. Um total de 11.795 (15,2%) dos transcritos de cajueiro anão CCP 76 e 17.205 (45,9%) do cajueiro comum mostraram *hits* significantes com o banco de dados do *UniProt*. O banco de dados do *Swiss-Prot* acessado em outubro de 2012 contém 536.789 sequências (256,5 Mb no formato fasta) e a busca no *Blastx* contra os transcritos levou aproximadamente 5 horas de processamento (Tabela 7).

Para avaliar o perfil geral da identificação pelo BLAST, foi feito um gráfico de frequência com base nos valores de identidade utilizando a ferramenta *Plotrix* do programa R. O gráfico 23 mostra que o cajueiro anão CCP 76 possui uma distribuição de identidade melhor em relação ao cajueiro comum (Gráfico 24). É possível notar no gráfico 23 que a maioria das identificações possui valores de identidade em cerca de 88% em relação ao seu respectivo *subject*. Já o gráfico 24 mostra que a maioria das identificações possuem valores de identidade próximos de 39%.

Também foram avaliadas a frequência do *score* em ambas as amostras mostrando resultados semelhantes no cajueiro anão CCP 76 (Gráfico 25) e no cajueiro comum (Gráfico 26). Percebe-se que o *score* se inicia a partir de 50, mas sua frequência cai com *score* acima de 500.

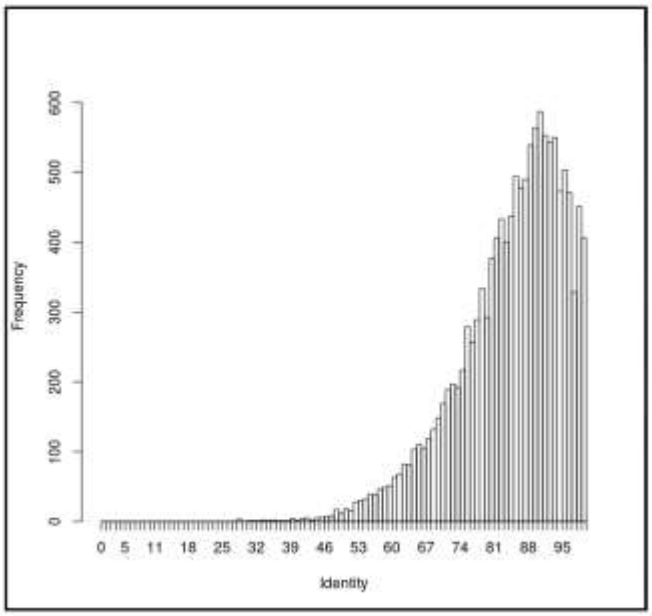
Com base na identificação pelo BLAST, foi feita uma anotação funcional pelo GO. Um total de 5.186 (44%) e 7.915 (46,0%) de termos de GO foi encontrado para o cajueiro anão CCP 76 e cajueiro comum, respectivamente. Os principais termos de GO encontrados no transcriptoma do cajueiro estão mostrados na Tabela 8.

Tabela 7 – Estatísticas do BLAST do cajueiro anão CCP 76 e cajueiro comum.

Parâmetro	Cajueiro CCP 76	Cajueiro comum
Número de sequências do transcriptoma	77.371	37.422
Tamanho do arquivo do transcriptoma	15 Mb	12 Mb
Banco de dados	*Swiss-Prot*	*Swiss-Prot*
Número de sequências da database	536.789	536.789
Tamanho do arquivo da database	256,5 Mb	256,5 Mb
Tipo de BLAST	*Blastx*	*Blastx*
Número de processadores utilizados	8	8
Corte do E-value	1.10^{-5}	1.10^{-5}
Tempo de processamento	5 horas	5 horas
Formato de saída	.txt	.txt
Data de acesso a database	10/10/2012	10/10/2012

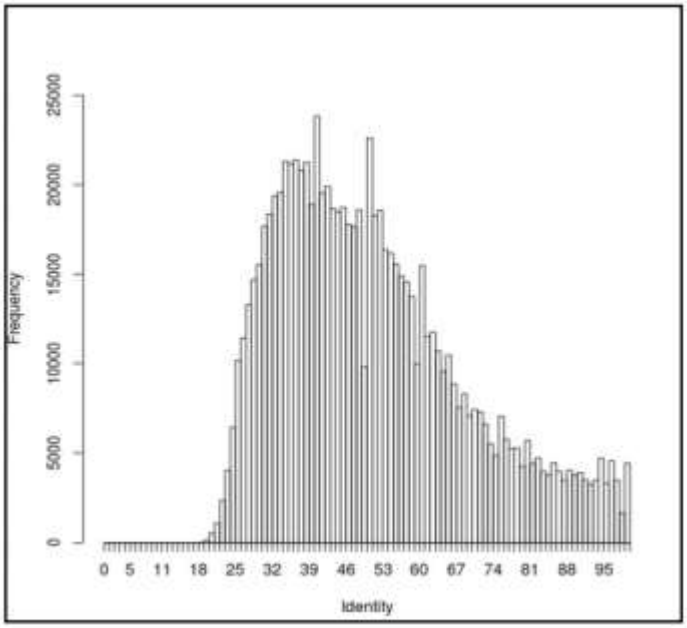
Fonte: O Autor

Gráfico 23 - Frequência da identidade do BLAST dos transcritos do cajueiro CCP 76 contra o *Swiss-Prot.*



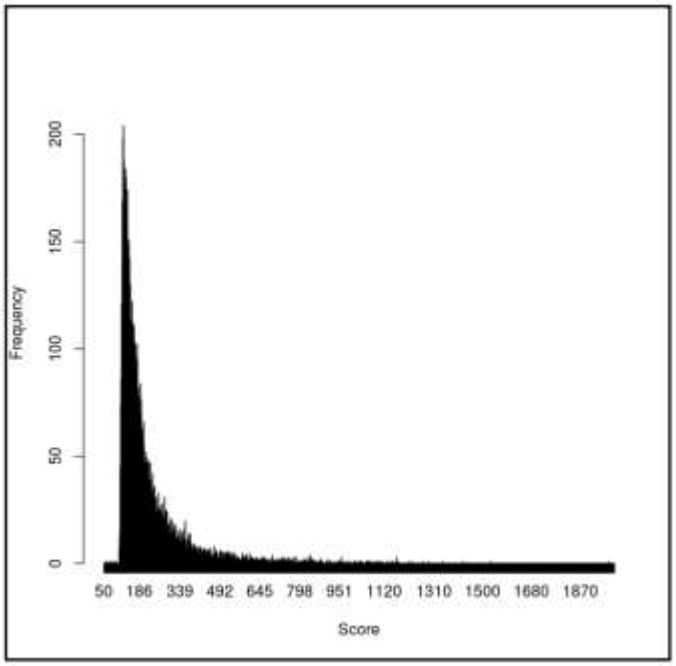
O gráfico foi feito utilizando o pacote *plotrix* do programa estatístico R. **Fonte**: O Autor.

Gráfico 24 - Frequência da identidade do BLAST dos transcritos do cajueiro comum contra o *Swiss-Prot*.



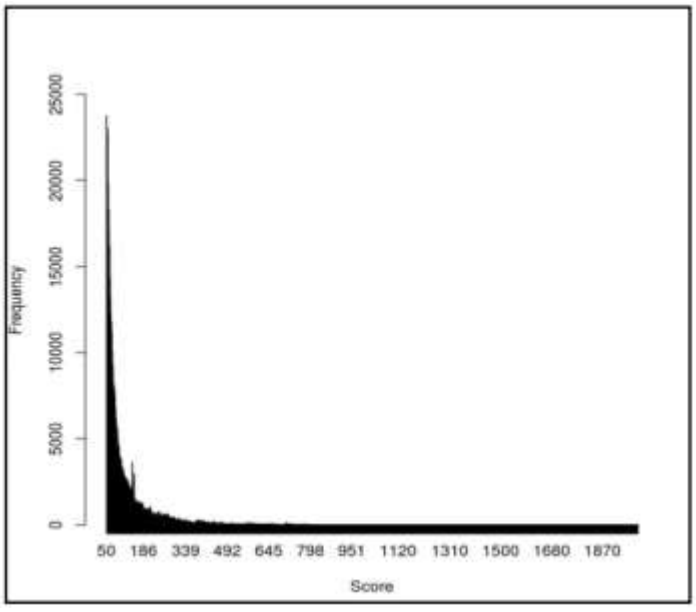
O gráfico foi feito utilizando o pacote *plotrix* do programa estatístico R. **Fonte**: O Autor.

Gráfico 25 - Frequência do score do BLAST dos transcritos do cajueiro CCP 76 contra o banco de dados do *Swiss-Prot*.



O gráfico foi feito utilizando o pacote *plotrix* do programa estatístico R (*plotrix*). **Fonte**: O Autor.

Gráfico 26 - Frequência do score do BLAST dos transcritos do cajueiro comum contra o banco de dados do *Swiss-Prot*.



O gráfico foi feito utilizando o pacote *plotrix* do programa estatístico R (plotrix). **Fonte**: O Autor.

Tabela 8 - Classificação funcional do transcriptoma do cajueiro comum e anão CCP 76 com base nos termos de *Gene ontology* mais representativos.

Termos de GO	Nº de Unigenes (CCP 76)	Nº de Unigenes (Comum)
Processo Biológico	4058	6225
Processo Celular	3331	5138
Processo Metabólico	2856	4501
Resposta ao estímulo	1209	1829
Regulação Biológica	1181	1653
Single-organism process	895	1233
Componente Celular	4328	6524
Parte Celular	3978	6052
Organela	3018	4607
Membrana	1745	2501
Parte da Organela	1577	2376
Parte da Membrana	1215	1750
Função Molecular	4066	6342
Ligação	2822	4314
Atividade Catalítica	2532	4023
Atividade Transportadora	332	487
Atividade Fator de transcrição ligante a sequência específica de DNA	233	310
Atividade Molécula Estrtural	199	412

Fonte: O Autor.

Ainda com base nas identificações do BLAST, foi feita uma busca por termos de KO utilizando a ferramenta *IDMapping* do *Uniprot*. Um total de 1.594 (13,5%) e 1.989 (11,5%) de termos de KO foi encontrado no transcriptoma do cajueiro anão CCP 76 e comum, respectivamente.

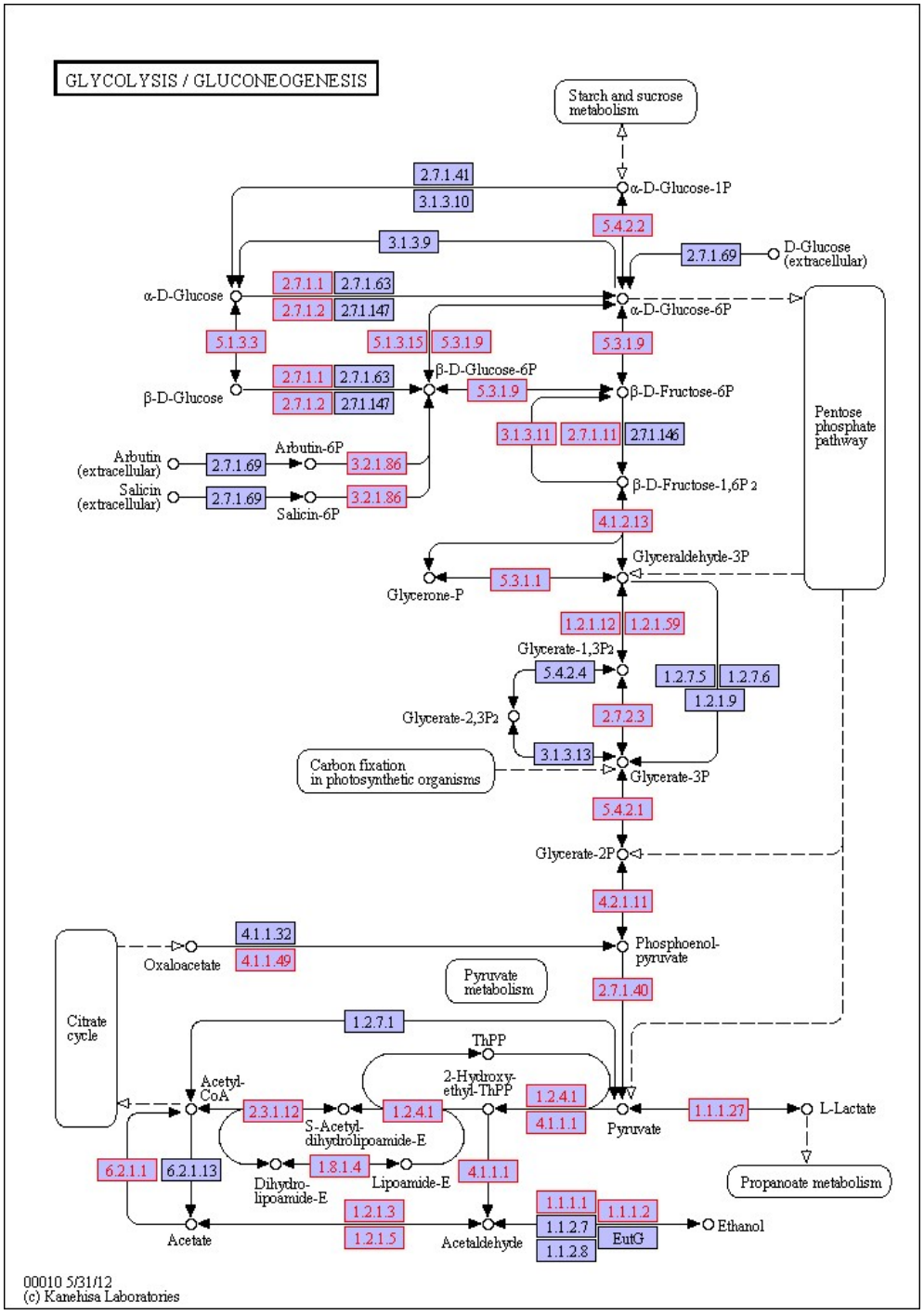
Os mapas KEGG de dezenas de vias metabólicas foram gerados para ambos transcriptomas do cajueiro comum e CCP 76. Quatro exemplos de mapas KEGG para a glicólise/gliconeogenese, ciclo do ácido tricarboxílico, fosforilação oxidativa e biossíntese de ácidos graxos estão mostrados nas figuras 9 a 16.

Praticamente todas as enzimas da glicólise e gliconeogênese estão compartilhadas no cajueiro anão (Figura 9) e cajueiro comum (Figura 10) sendo que apenas a gliceraldeído-3-fosfato desidrogenase (NADP) [EC:1.2.1.9] e a álcool desidrogenase (EutG) foram encontradas apenas no cajueiro comum. Todas as enzimas do ciclo do ácido tricarboxílico estão compartilhadas no cajueiro anão (Figura 11) e cajueiro comum (Figura 12).

Das enzimas identificadas, as quais estavam relacionadas à fosforilação oxidativa, 27 estavam presentes no cajueiro comum e ausentes no CCP 76, enquanto que apenas uma foi identificada no CCP 76 e não no cajueiro comum (Figura 13 e 14). Enzimas presentes no cajueiro comum: protoheme IX farnesiltransferase (COX10), Subunidade D da NADH:quinona oxidoredutase (nuoD) e ubiquinona citocromo-c-redutase (QCR6). Enzima presente no CCP 76: Subunidade C da ATPase transportadora de prótons tipo V (C).

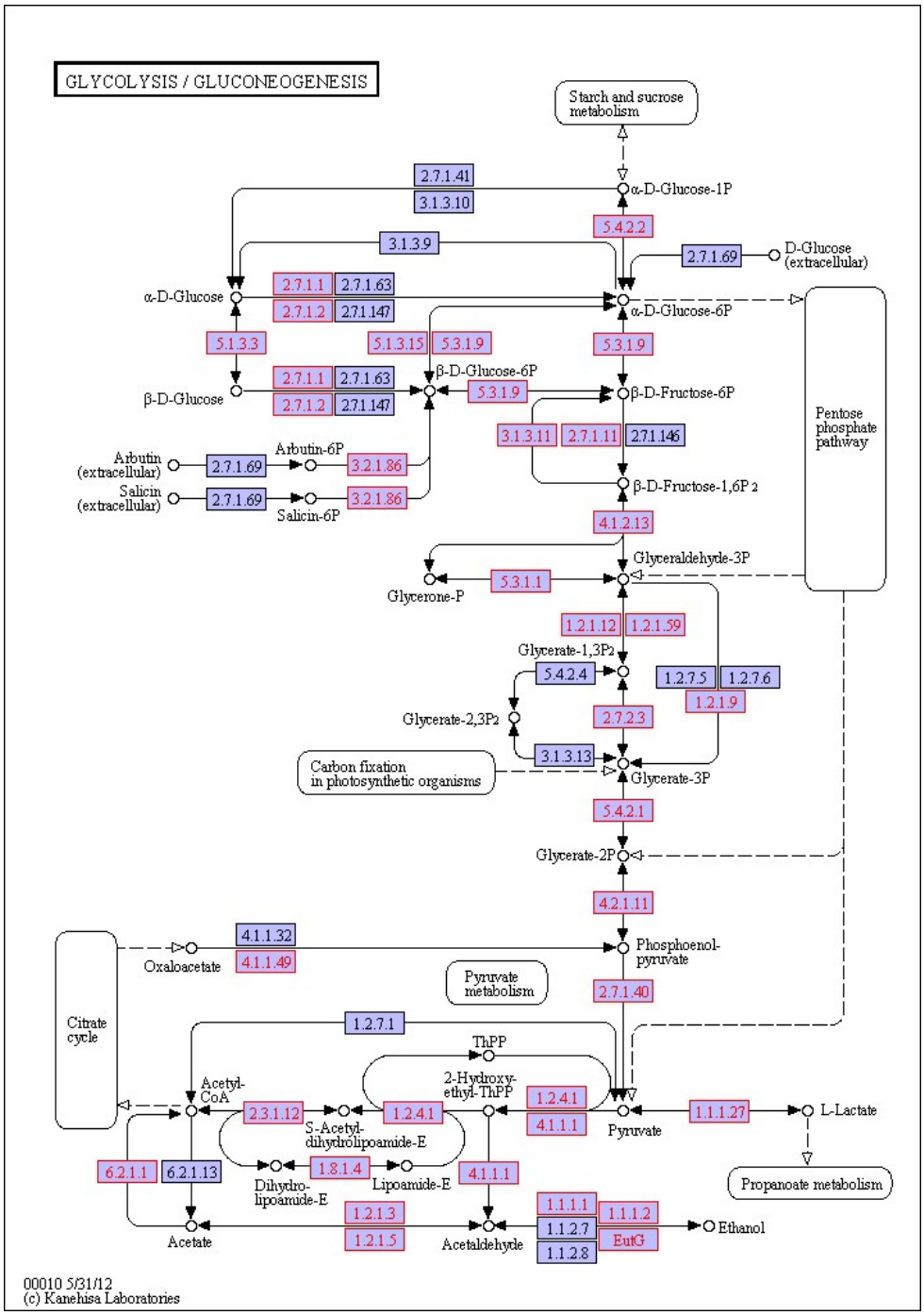
Com relação a biossíntese de ácidos graxos, percebe-se que as enzimas enoil-ACP redutase III (fabL), Acil graxo-ACP tioesterase B (FatB), oleoil-ACP hidrolase (3.1.2.14) e a Acil graxo-ACP tioesterase A (FatA) foram encontradas no cajueiro comum (Figura 16) mas não estão presentes no anão CCP 76. Enquanto que não há enzimas que estão presentes no CCP 76 e ausentes no cajueiro comum (Figura 15).

Figura 9 - Mapa KEGG para a via da glicólise e gliconeogênese mostrando enzimas encontradas no transcriptoma do cajueiro anão CCP 76.



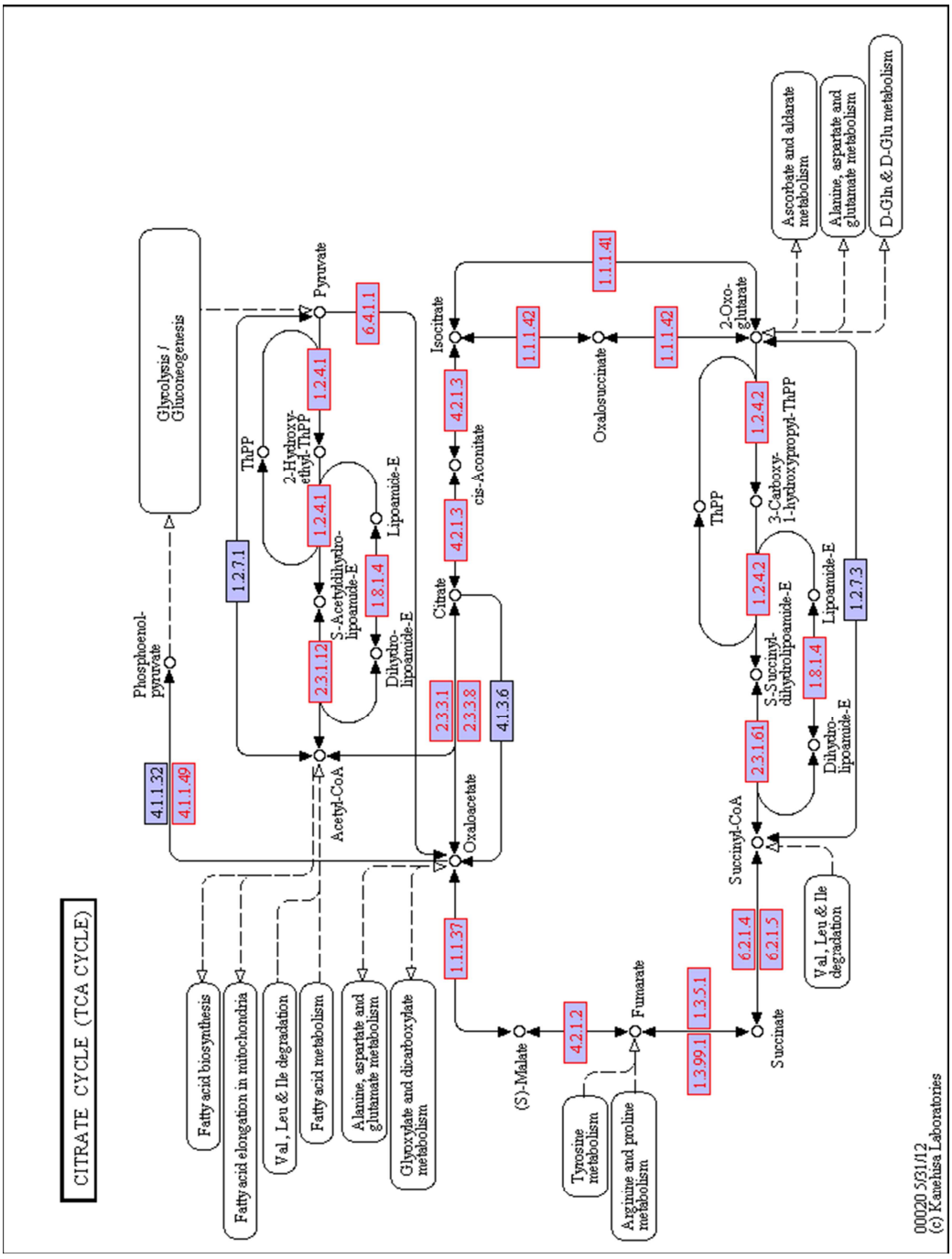
Os números dentro das caixas representam o código da enzima (EC) envolvida na via sendo que cada caixa destacada de vermelho corresponde às enzimas identificadas no transcriptoma de cajueiro. **Fonte**: O Autor.

Figura 10 - Mapa KEGG para a via da glicólise e gliconeogênese mostrando enzimas encontradas no transcriptoma do cajueiro comum.



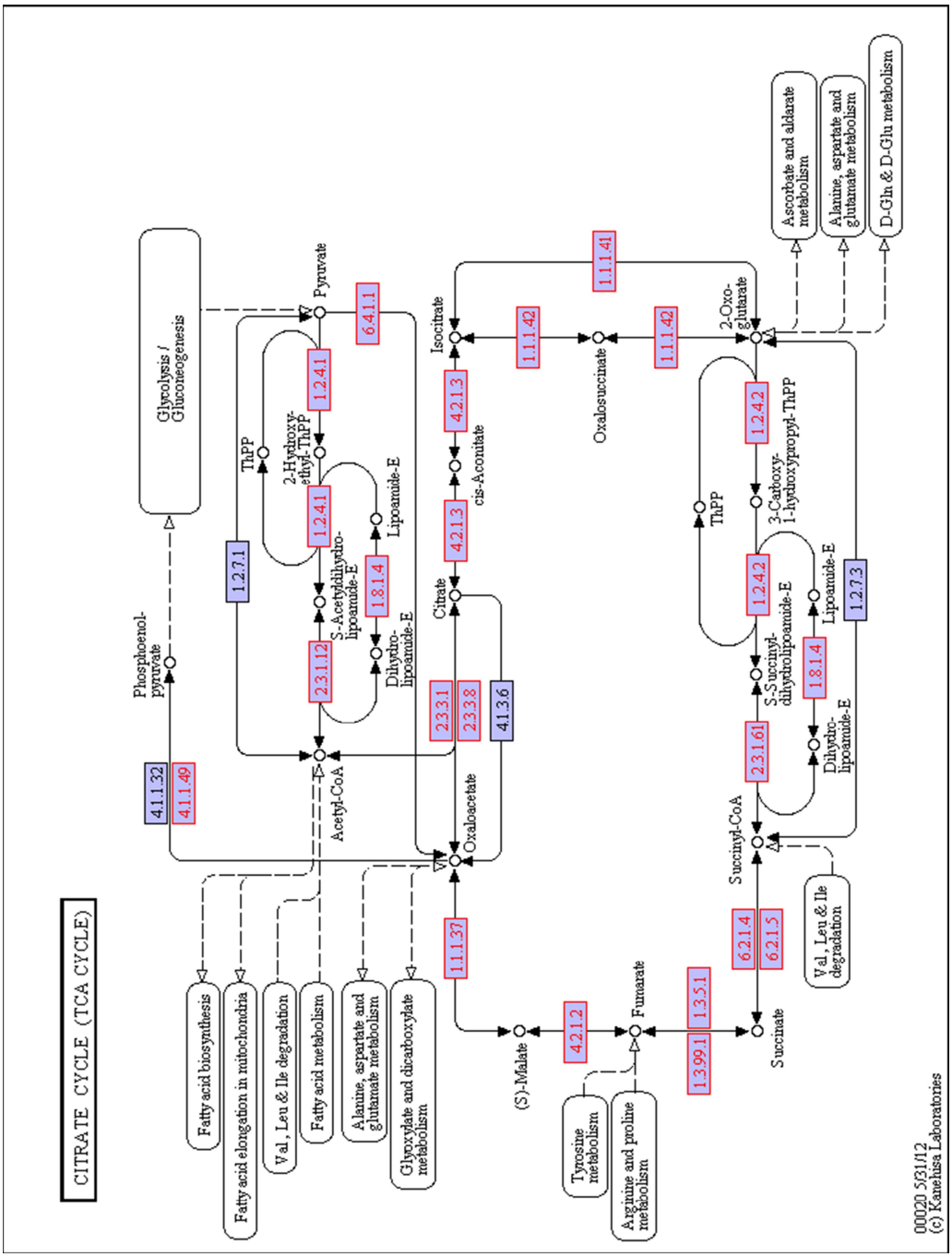
Os números dentro das caixas representam o código da enzima (EC) envolvida na via sendo que cada caixa destacada de vermelho corresponde às enzimas identificadas no transcriptoma de cajueiro. **Fonte:** O Autor

Figura 11 - Mapa KEGG para a via do ciclo do ácido tricarboxílico mostrando enzimas encontradas no transcriptoma do cajueiro anão CCP 76.



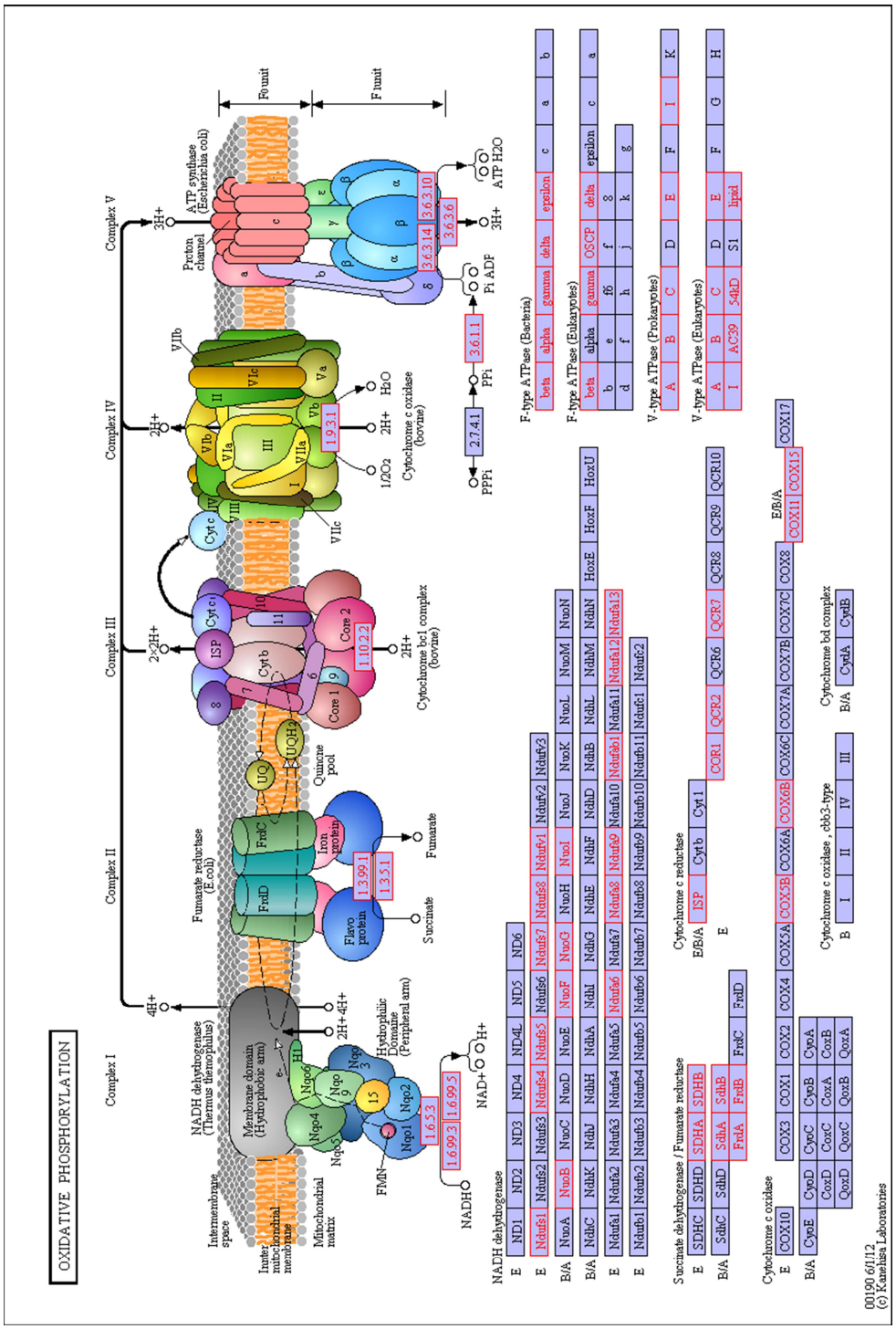
Os números dentro das caixas representam o código da enzima (EC) envolvida na via sendo que cada caixa destacada de vermelho corresponde às enzimas identificadas no transcriptoma de cajueiro. **Fonte**: O Autor.

Figura 12 - Mapa KEGG para a via do ciclo do ácido tricarboxílico mostrando enzimas encontradas no transcriptoma do cajueiro comum.



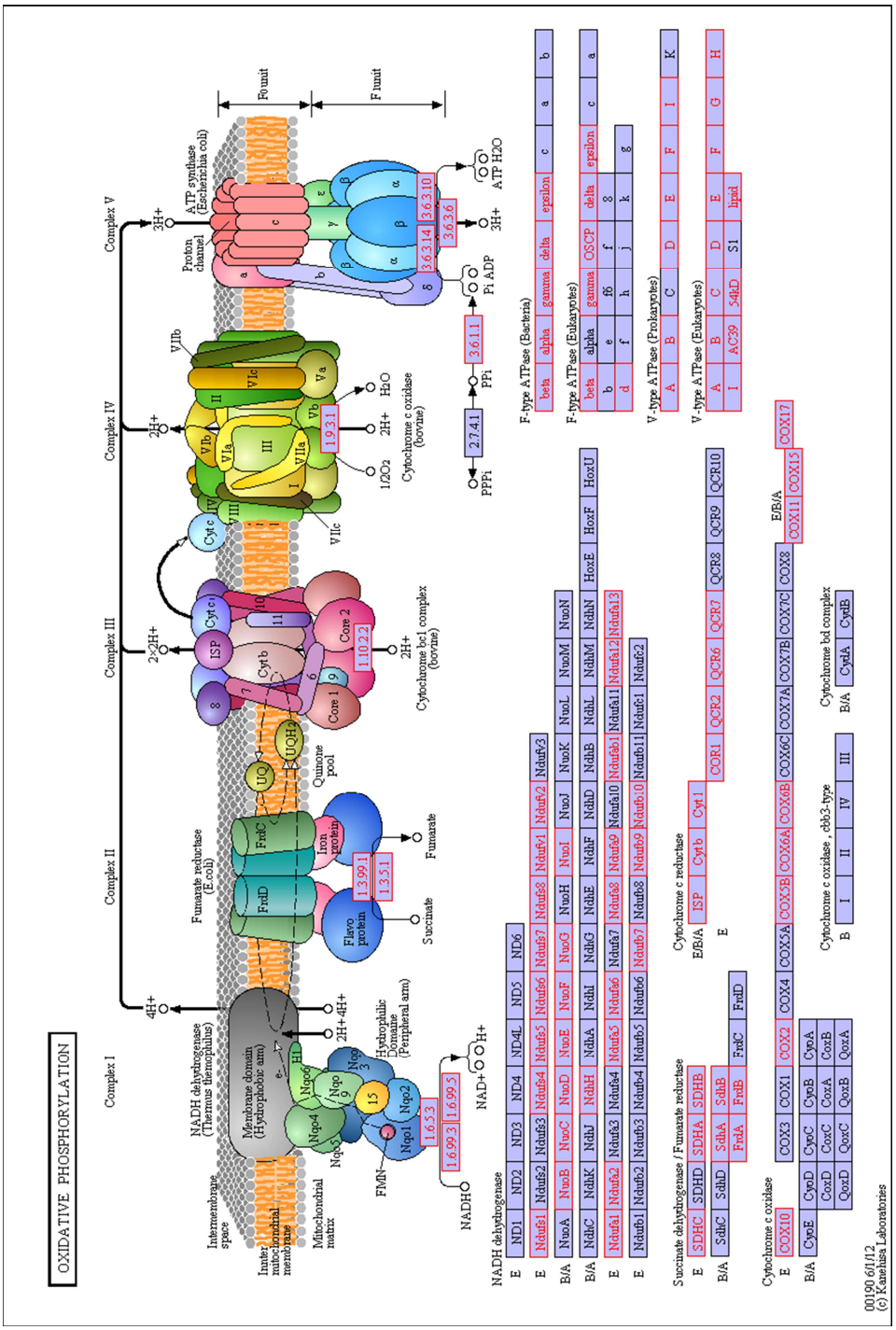
Os números dentro das caixas representam o código da enzima (EC) envolvida na via sendo que cada caixa destacada de vermelho corresponde às enzimas identificadas no transcriptoma de cajueiro. **Fonte**: O Autor.

Figura 13 - Mapa KEGG para a via da fosforilação oxidativa mostrando enzimas encontradas no transcriptoma do cajueiro anão CCP 76.



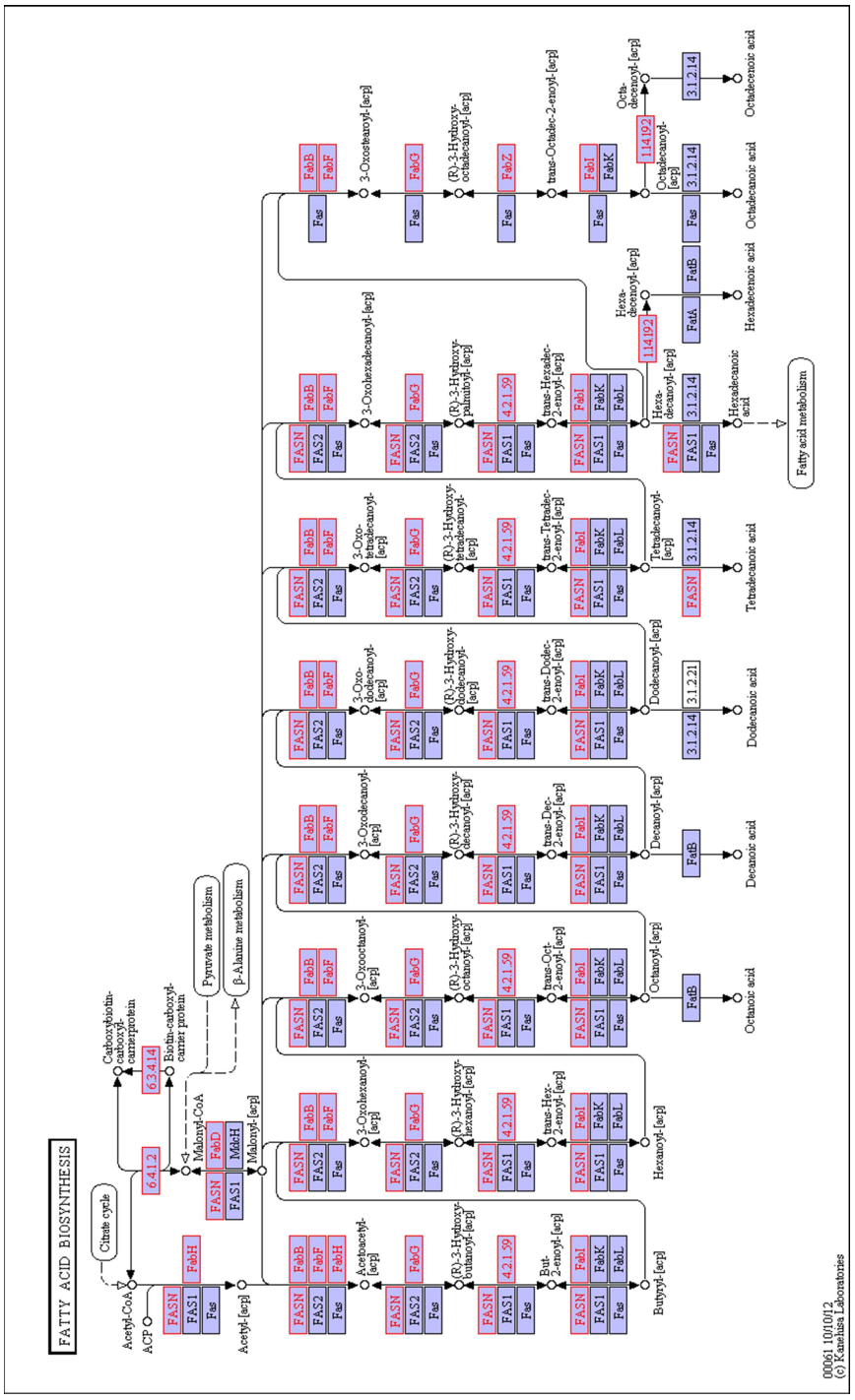
Os números dentro das caixas representam o código da enzima (EC) envolvida na via sendo que cada caixa destacada de vermelho corresponde às enzimas identificadas no transcriptoma de cajueiro.
Fonte: O Autor

Figura 14 - Mapa KEGG para a via da fosforilação oxidativa mostrando enzimas encontradas no transcriptoma do cajueiro comum.



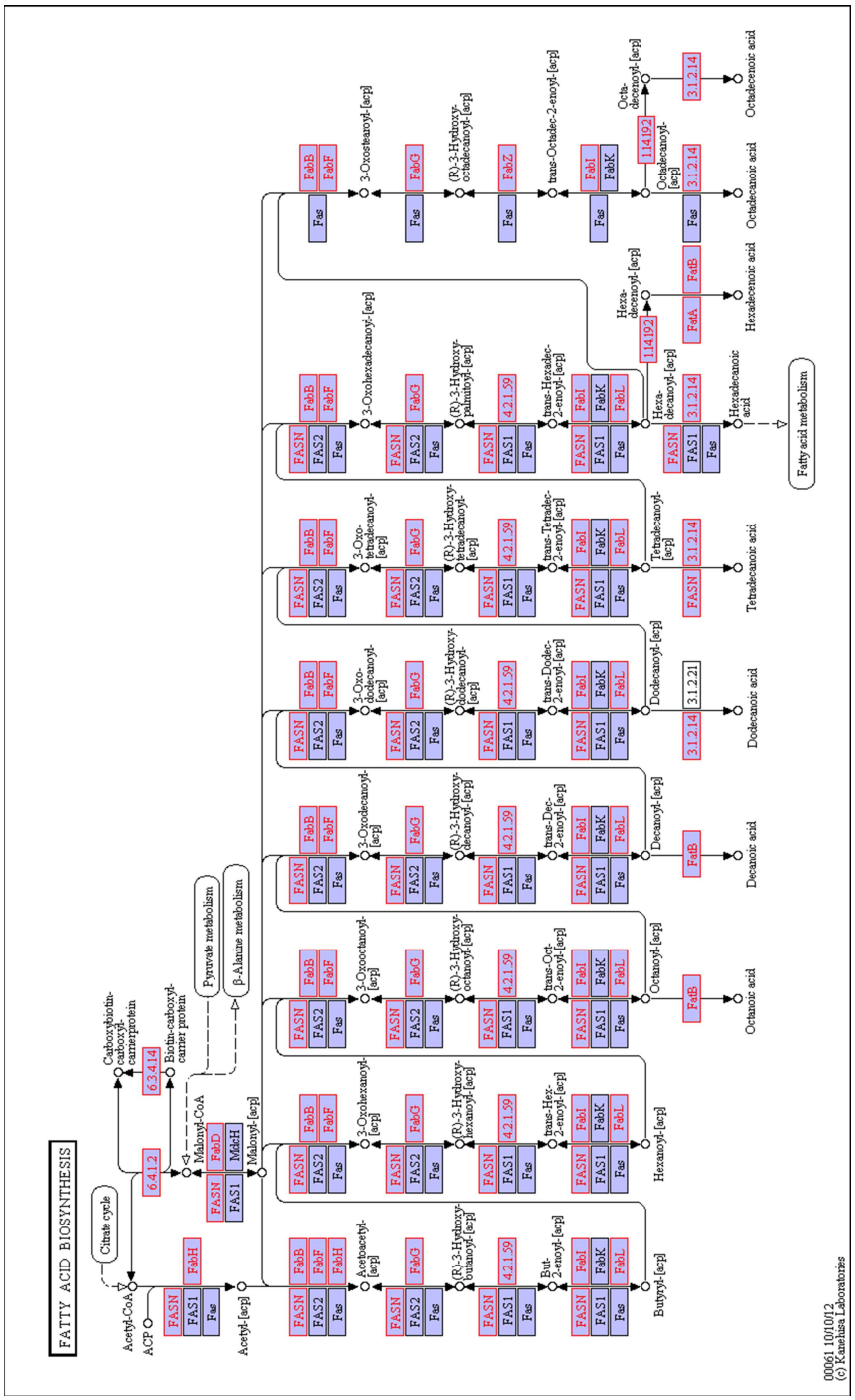
Os números dentro das caixas representam o código da enzima (EC) envolvida na via sendo que cada caixa destacada de vermelho corresponde às enzimas identificadas no transcriptoma de cajueiro. **Fonte**: O Autor.

Figura 15 - Mapa KEGG para a via da biossíntese de ácidos graxos mostrando enzimas encontradas no transcriptoma do cajueiro anão CCP 76.



Os números dentro das caixas representam o código da enzima (EC) envolvida na via sendo que cada caixa destacada de vermelho corresponde às enzimas identificadas no transcriptoma de cajueiro. **Fonte**: O Autor.

Figura 16 - Mapa KEGG para a via da biossíntese de ácidos graxos mostrando enzimas encontradas no transcriptoma do cajueiro comum.



Os números dentro das caixas representam o código da enzima (EC) envolvida na via sendo que cada caixa destacada de vermelho corresponde às enzimas identificadas no transcriptoma de cajueiro. **Fonte**: O Autor.

4.6 DISCUSSÃO

O baixo percentual de proteínas identificadas pelo BLAST no cajueiro comum (45,9%) e no cajueiro anão CCP 76 (15,2%), é devido talvez à escassez de dados moleculares para a família *Anacardiaceae*. Comparando nossos resultados com aqueles obtidos pelo transcriptoma do grão-de-bico (*Cicer arietinum* L.), uma leguminosa cultivada em regiões áridas, nota-se que 44,7% dos transcritos mostraram identificação pelo BLAST contra o banco de dados do *UniProt* (GARG et al., 2011).

Também se observa baixo percentual de identificações do BLAST na microalga *Dunaliella tertiolecta* onde se observou que apenas 25% dos transcritos foram identificados contra o banco de dados não redundante do NCBI e com valor e abaixo de 10^{-6} (RISMANI-YAZDI; HAZNEDAROGLU; PECCIA, 2011). Parece ser razoável supor que a falta de organismos modelo é o principal fator limitante no percentual de identificações pelo BLAST uma vez que todos os trabalhos citados mostraram *reads* de boa qualidade.

O transcriptoma da semente de cajueiro apresentou um alto percentual de transcritos que não foram identificados pelo BLAST utilizando o banco de dados do *Swiss-Prot*. Uma vez que os resultados de montagem revelaram *reads* de boa qualidade, o estudo dessas sequências no futuro poderá revelar resultados de grande significância biológica.

Com relação à anotação funcional pelo *Gene Ontology*, quase metade dos transcritos identificados pelo BLAST do cajueiro comum (46%) e CCP 76 (44%) foram anotados com termos de GO. Um percentual semelhante foi encontrado no transcriptoma do grão de bico, onde 65% dos transcritos foram identificados com termos de GO (GARG et al., 2011).

A análise do transcriptoma da oliva (*Olea europaea*) revelou que os principais termos de GO pertententes à categoria de processos biológicos são de processo celular, processo metabólico e regulação biológica. Os principais termos relacionados à função molecular são ligação, atividade catalítica e atividade de transporte. Com relação à categoria componente celular, os termos de GO que se destacam são: célula, organela e membrana (MUÑOZ-MÉRIDA et al., 2013). Os mesmos termos foram encontrados em maior abundância também no transcriptoma de sementes de

cajueiro.

Além da anotação pelo *Gene Ontology*, descrevemos algumas das principais vias metabólicas em mapas obtidos pelo *KEGG Orthology* (KO). A conversão dos IDs de proteínas identificadas pelo BLAST feita pelo *IDMapping* do *Uniprot* mostrou que o número de termos de KO foi de apenas 13,5% para o cajueiro CCP 76 e 11,5% para o cajueiro comum. Em comparação, a anotação do transcriptoma de raiz do ginseng americano (*Penax quinquefolius*) mostrou que 53% das sequências tiveram identificação pelo banco de dados do KEGG (SUN et al., 2010).

4.7 CONCLUSÃO

Com base nos resultados obtidos, apenas uma pequena parte do transcriptoma das sementes de cajueiro pôde ser identificada pelo BLAST e destes uma pequena fração foi anotada com termos de GO e visualizada em mapas KEGG. O grande número de transcritos de alta qualidade não identificados reflete a falta de dados moleculares do cajueiro e espécies próximas e constitui um desafio aos estudos posteriores. O acesso aos mapas metabólicos mostrou um grande número de enzimas, mas algumas estiveram ausentes, talvez devido a falhas no processo de montagem do transcriptoma.

CAPÍTULO 5

PROTEÔMICA COMPARATIVA DO CAJUEIRO COMUM E ANÃO CCP 76 UTILIZANDO ELETROFORESE BIDIMENSIONAL

5 PROTEÔMICA COMPARATIVA DO CAJUEIRO COMUM E ANÃO CCP 76 UTILIZANDO ELETROFORESE BIDIMENSIONAL.

5.1 INTRODUÇÃO

Nos últimos anos vários trabalhos têm enfocado a caracterização da dinâmica de proteínas ao longo do desenvolvimento vegetal (CANOVAS et al., 2004). A dinâmica das proteínas em um sistema vivo é influenciada por diversos fatores internos e externos, que determinam modificações estruturais e a conformação das proteínas. Nesse sentido, o estudo e a caracterização de mapas proteômicos apresentam-se como importantes ferramentas complementares aos estudos anatômicos, fisiológicos e de genômica (BALBUENA, 2009). A análise do proteoma permite examinar simultaneamente alterações e classificar padrões temporais de acúmulo de proteínas que ocorrem durante o desenvolvimento da semente, permitindo a identificação de proteínas marcadoras estágio específicas (DIAS et al., 2007).

A proteômica vegetal encontra-se menos avançada em relação à pesquisa proteômica para os organismos unicelulares e leveduras (PARK, 2004). Os primeiros estudos envolvendo variações no proteoma durante o desenvolvimento e germinação de sementes foi realizada em *Arabidopsis thaliana* (GALLARDO et al., 2001) e *Medicago truncatula* (GALLARDO et al., 2003). Trabalhos semelhantes têm sido realizados em outras espécies modelos e naquelas que apresentam valor econômico como a soja (*Glycine max*), o arroz (*Oryza sativa*), o tomate (*Lycopersicum esculentum*), o trigo (*Triticum aestrivum*), e a cevada (*Hordeum vulgare*).

Uma das maiores dificuldades em proteômica vegetal é a capacidade de identificação de proteínas em genomas não sequenciados (MORAES, 2006). A identificação e caracterização de proteínas são aceleradas pela disponibilidade de sequencias genômicas e de ESTs. Para contornar os problemas da falta de sequências genômicas, duas alternativas são possíveis. Se sequências EST estão disponíveis, proteínas podem ser identificadas por sequências *tags* obtidas por ESI-MS/MS. No caso do cajueiro, existem apenas 5 ESTs disponíveis no GenBank até a presente data. Outro caminho seria realizar buscas baseadas na homologia, preferencialmente usando sequências obtidas de MS/MS ou sequenciamento de Edman. Proteínas vegetais, em geral, compartilham uma quantidade significativa de

identidade/homologia entre elas.

A preparação de amostra de material vegetal é reconhecidamente mais difícil devido à rigidez das paredes celulares e ao acúmulo de uma grande variedade de metabólitos secundários no vacúolo central. Inclusive, o próprio vacúolo que ocupa grande parte do volume da célula faz com que a quantidade de proteína intracelular seja menor quando comparado a outras células que não apresentam vacúolo. Métodos envolvendo precipitação de proteínas vegetais utilizando acetona/TCA têm sido eficientes na proteômica vegetal. Contudo, a simples preparação de amostras de qualidade seguida pela boa resolução de géis bidimensionais altamente reprodutíveis ainda parece ser um desafio no estudo proteômico de plantas.

5.2 OBJETIVOS

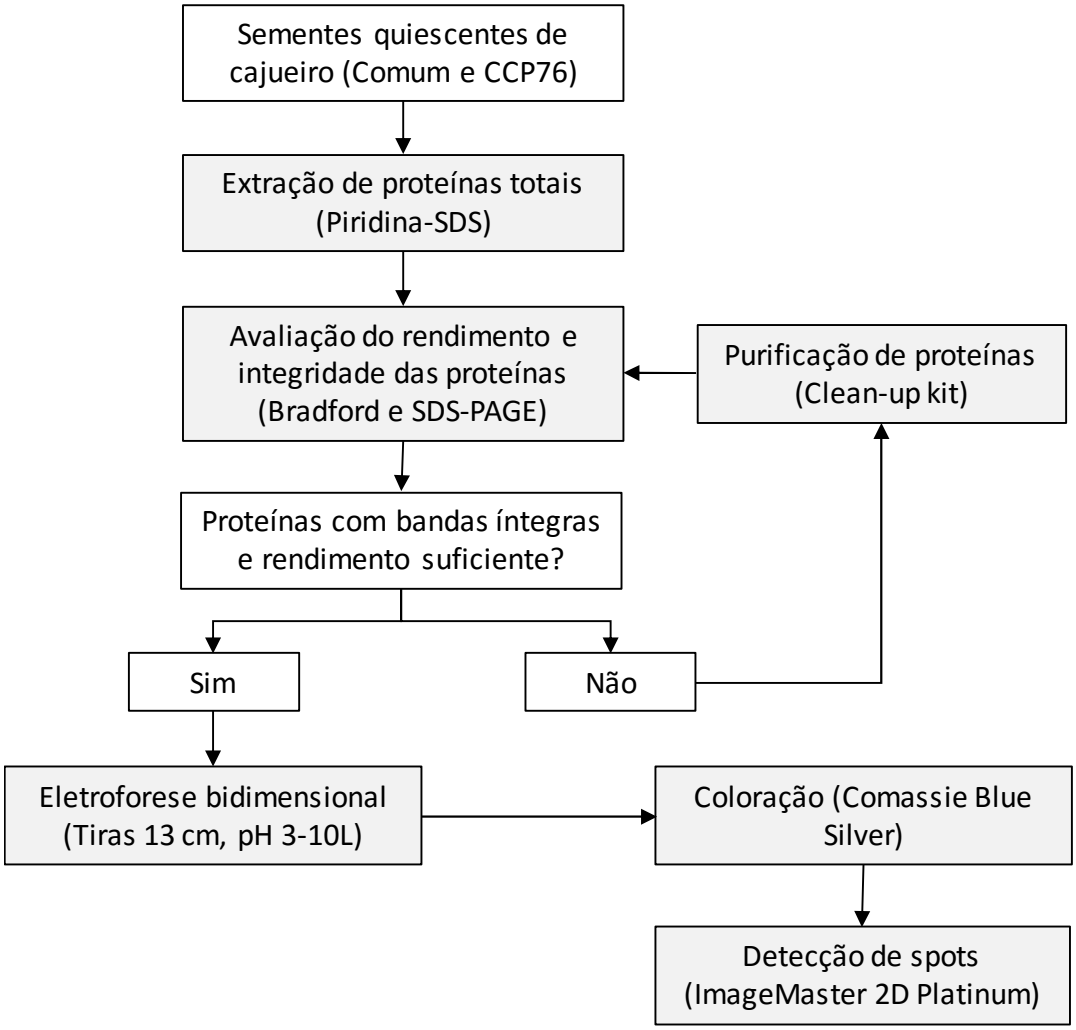
5.2.1 Geral

Comparar o perfil de proteínas de semente de cajueiro comum e CCP 76 utilizando eletroforese bidimensional.

5.2.2 Específicos

- Estabelecer um protocolo experimental reprodutível para extração de proteínas da amêndoa da castanha do caju compatíveis com 2D-PAGE e MS.
- Criar um banco de dados com as informações obtidas no presente projeto (imagens dos géis 2D, dados sobre a massa e pI das proteínas, sequências dos peptídeos e etc).
- Determinar a presença de proteínas específicas para os genótipos estudados.

5.3 ESTRATÉGIA EXPERIMENTAL



5.4 METODOLOGIA

5.4.1 Coleta de material biológico

As sementes de cajueiro comum e do genótipo anão-precoce CCP 76 foram adquiridas do Sítio Recanto dos Cajueiros, localizado na cidade de Itapipoca, Ceará, distando cerca de 130 Km de Fortaleza. As sementes foram utilizadas para a extração de proteínas em condições normais (quiescentes).

5.4.2 Extração de proteínas totais de semente

Frutos (castanhas) do cajueiro (Comum e CCP 76) foram cuidadosamente quebrados para a liberação das amêndoas que posteriormente foram cortadas em cubos de aproximadamente 1 mm. Em seguida, as amêndoas cortadas (aproximadamente 15 g) foram delipidadas em 100 mL de acetona sob agitação constante durante uma noite. A acetona foi trocada quatro vezes. Logo após, foram maceradas em pistilo e almofariz na presença de nitrogênio líquido até a obtenção do pó (farinha). As proteínas totais das sementes foram extraídas usando 20 mg de farinha da castanha acrescido de 40 mg de PVPP (polivinilpolipirrolidona) e 800 µL de tampão piridina-SDS (piridina 50 mM, tiouréia 10 mM, SDS 1%, pH 5,0) (proporção de 1:2:40). A mistura foi agitada por 2 h e 30 min a 4ºC seguido de centrifugação a 10.000 x g por 40 min. Ao sobrenadante foram adicionados quatro volumes de acetona gelada contendo 10% de TCA (ácido tricloroacético) *overnight* a -20 ºC (DAMERVAL, C; et al., 1986). Em seguida, a amostra foi centrifugada a 10.000 x g por 30 min e o sobrenadante descartado. O precipitado contendo as proteínas totais foi lavado com 1 mL de acetona gelada 100% (3 vezes) e seco a temperatura ambiente. Posteriormente foram solubilizadas em 200 µL de uréia/tiouréia9 M.

5.4.3 Dosagem de proteínas

As proteínas totais foram quantificadas pelo o método de Bradford (1976) e a integridade protéica foi analisada por SDS-PAGE (LAEMMLI, 1970).

5.4.4 Eletroforese bidimensional (2 DE)

As proteínas da amêndoa da castanha de caju de cada genótipo (250 µg), foram diluídas em solução de reidratação (uréia 7 M, tiouréia 2 M, DTT 65 mM, CHAPS 1% p/v, *IPGBuffer* 0,5% v/v e azul de Bromofenol 0,002% p/v) para um volume final de 250 µl (GORK et al., 2007). As amostras diluídas na solução de reidratação foram aplicadas no *IPGBox* (*GE Healthcare*) e posteriormente incubadas com tiras com gradiente de pH imobilizado (*IPGStrip*) de 13 cm e faixa de pH linear de 3-10 e de 4-7, por um período de 16 h. A focalização isoelétrica (IEF) Foi realizada no equipamento *Ettan™ IPGPhor III™* (*GEHealthcare*) utilizando as seguintes condições: etapa 1 (500 V por 0:30 h); etapa 2 (4000 V por 2:30 h) e etapa 3 (10000 V até atingir 18.000 Vh totais). Após a IEF, as tiras foram armazenadas em freezer -20 °C para posterior utilização.

Após a focalização, as tiras de IPG foram equilibradas, sob agitação, em solução de equilíbrio (tris 50 mM, glicerol 30%, uréia 6 M, SDS 2% e azul de bromofenol 0,002%) com ditiotreitol (DTT) a 1% por 15 min para a redução das proteínas e alquiladas com iodoacetamida (IAA) a 3%, também em solução de equilíbrio por 15 min. Terminado o equilíbrio as tiras foram mergulhadas em tampão de corrida por 10 s para retirar o excesso da solução de equilíbrio. Em seguida, as tiras foram postas no topo dos géis da segunda dimensão e cobertos com 2 mL de agarose morna (agarose 0,5 %, SDS 1% e azul de bromofenol 0,002%) adicionando um pente para a formação do poço do marcador e deixados até solidificar. A corrida foi realizada em uma unidade de eletroforese vertical (*HoeferSE 600 Ruby, AmershamBiosciences*®) utilizando os seguintes parâmetros: 15 mA/gel por 15 min e em seguida 25 mA/gel a uma temperatura de 18 °C e permanecendo assim até que o azul de bromofenol atinja o limite inferior dos géis. Após a corrida, os géis foram colocados em solução de fixação, composta por etanol, ácido acético e água (4:1:5 v/v/v), durante 15 min, corados com solução de Comassie G-250 (*Blue Silver)* (CANDIANO el al., 2004) por 24 h e armazenados em solução de ácido acético 5%. Os géis foram escaneados utilizando-se o equipamento *ImageScannerIII* e gerenciados pelo programa *LabScan 6.0* (ambos da *GE Healthcare*). As imagens obtidas (com 300 dpi de resolução) foram analisadas e editadas no programa *Image Master 2D Platinum* 6.0 (GE *Healthcare*), onde os *spots* foram detectados

automaticamente pelo programa passando por uma revisão manual para eliminação de *spots* artefactuais.

5.4.5 Pesquisa no banco de dados

Os *spots* analisados nos géis bidimensionais pelo *ImageMaster*™ foram inicialmente identificados utilizando a ferramenta *TagIdent* do ExPASy (Expert Protein Analysis System), com banco de dados do Swiss-Prot.

5.5 RESULTADOS

5.5.1 Concentração e eletroforese unidimensional

O valor da concentração das proteínas totais, obtidas da amêndoa da castanha de caju, foi de 12,5 mg/ml e 9,88 mg/ml, para os genótipos de cajueiro comum e CCP 76, respectivamente. A porcentagem de proteína presente na farinha para esses genótipos (comum e CCP 76) foi de 12,5% e 9,88%, respectivamente (Tabela 9).

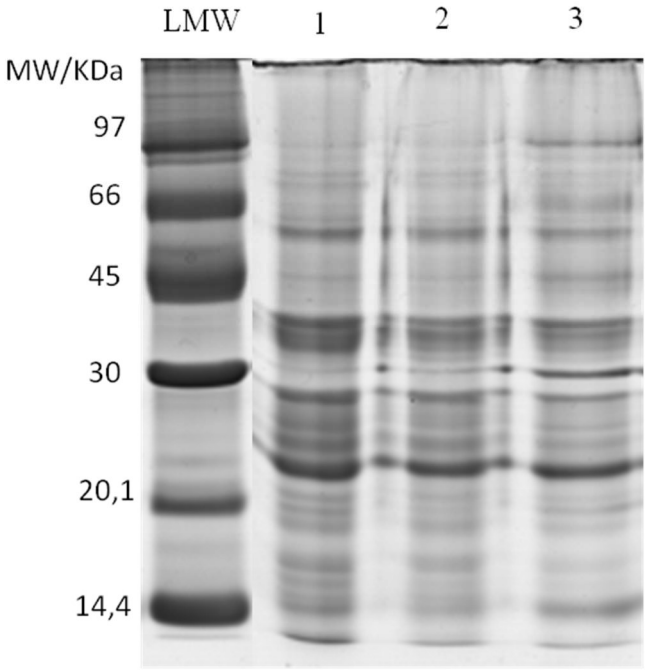
Nos resultados obtidos com SDS-PAGE (1 DE), pôde-se perceber bandas protéicas livres de contaminantes, como carboidratos e lipídeos (Figura 17). Nas raias 1 e 2, estão presentes amostras dos genótipos comum e CCP 76, respectivamente. Observando os resultados da quantificação e da eletroforese, percebe-se que as amostras possuem quantidade e qualidade aceitáveis para realização da eletroforese bidimensional (2 DE).

Tabela 9 - Quantificação de proteínas totais de amêndoas de diferentes genótipos de cajueiro utilizando o método de Bradford.

Amostra	Concentração de proteínas (mg/mL)	Rendimento (Relação proteína/semente)
Comum	12,5	12,5%
CCP 76	9,88	9,88%

Fonte: O Autor.

Figura 17 - Eletroforese em gel de poliacrilamida (SDS-PAGE) a 12,5%, de proteínas de amêndoas de cajueiro coradas com comassie R-350.



LMW: Marcador de baixo peso molecular com bandas variando de 14,4 a 97 KDa; **Raia 1** – proteínas de sementes de cajueiro comum; **Raia 2** – proteínas de semente de cajueiro CCP 76. **Fonte**: O Autor.

5.5.2 Eletroforese bidimensional em ampla faixa de pH

Os mapas proteômicos obtidos confirmaram a similaridade geral entre os perfis protéicos dos genótipos testados, contudo podemos observar a existência de algumas diferenças importantes no perfil de 2 DE obtidos. Utilizando tiras com faixa de pH de 3-10, foram detectados 157 e 196 *spots* no cajueiro comum (comum) e CCP 76, respectivamente.

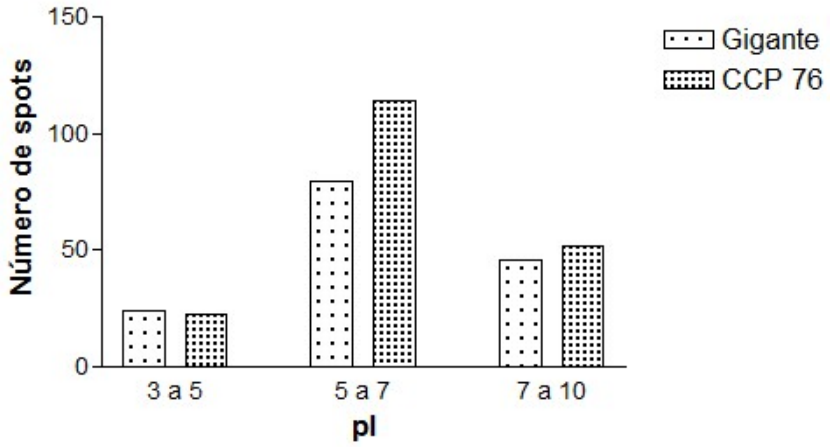
Com relação ao ponto isoelétrico, as proteínas de sementes de cajueiro se distribuem de forma mais abundante na faixa de pH 5 entre e 7 tanto no cajueiro CCP 76 como no cajueiro comum (Gráfico 27). Com relação à massa molecular, as proteínas estão mais distribuidas no intervalo de 20 KDa a 50KDa conforme mostrado no (Gráfico 28).

Os géis bidimensionais das proteínas de semente de cajueiro comum (Figura 18) e cajueiro CCP 76 (Figura 19) foram bem resolvidos usando tiras com faixa de pH

de 3-10. As áreas delimitadas com retângulos nos géis indicam regiões onde os *spots* altamente expressos se acumularam (R1), ou regiões que contenham proteínas alcalinas de difícil focalização (R3), ou ainda regiões com proteínas com diferentes níveis de intensidade (R2).

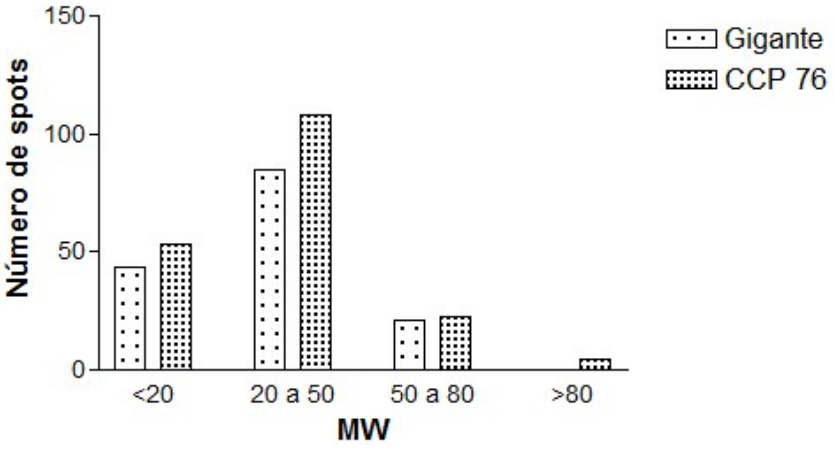
A identificação dos *spots* com base nos dados de ponto isoelétrico e massa molecular, feita pelo *TagIdent* contra o banco de dados do swiss-prot, revelou que a maioria das proteínas identificadas pertencem a classe das globulinas, tanto no cajueiro comum (Gráfico 29) como no cajueiro CCP 76 (Gráfico 30).

Gráfico 27 - Distribuição dos spots protéicos de cajueiro comum e cajueiro CCP 76 de acordo com o ponto isoelétrico (pI).



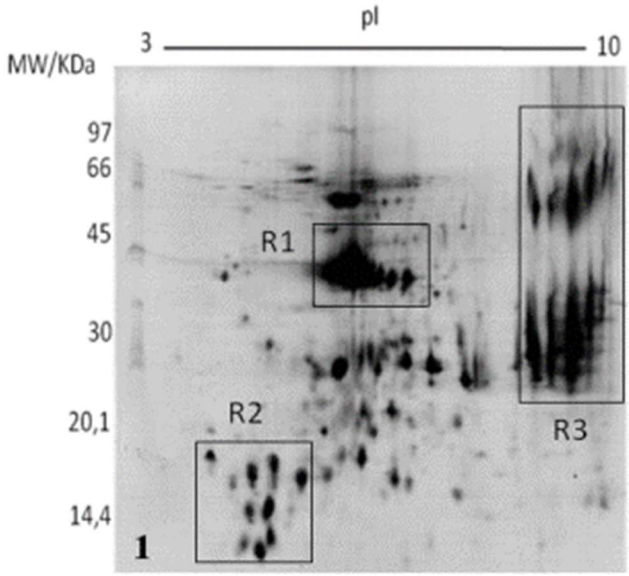
Fonte: O Autor.

Gráfico 28 - Distribuição dos spots protéicos de cajueiro comum e cajueiro CCP 76 de acordo com a massa molecular (MW).



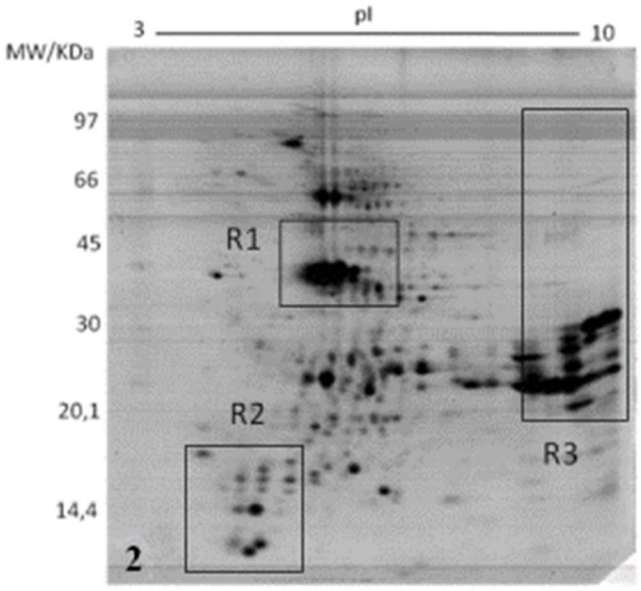
Fonte: O Autor.

Figura 18 – Gel bidimensional das proteínas totais da amêndoa da castanha de cajueiro comum usando tiras de pH 3-10.



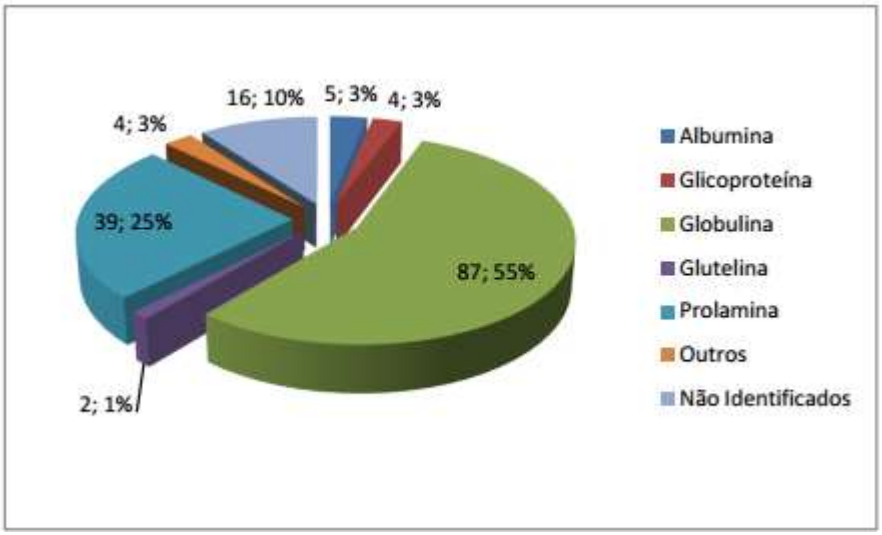
As proteínas estão distribuídas em tiras de 13 cm em uma faixa de pH 3 a 10. A esquerda marcador de peso molecular variando de 14 a 90 KDa; R1 – região de difícil resolução devido a presença de proteínas muito expressas; R2 – região de difícil resolução por conter proteínas em pH alcalino; R3 – região onde a intensidade dos *spots* variou nas variedades de cajueiro estudadas. **Fonte**: O Autor.

Figura 19 – Gel bidimensional das proteínas totais da amêndoa da castanha de cajueiro CCP 76 usando tiras de pH 3-10.



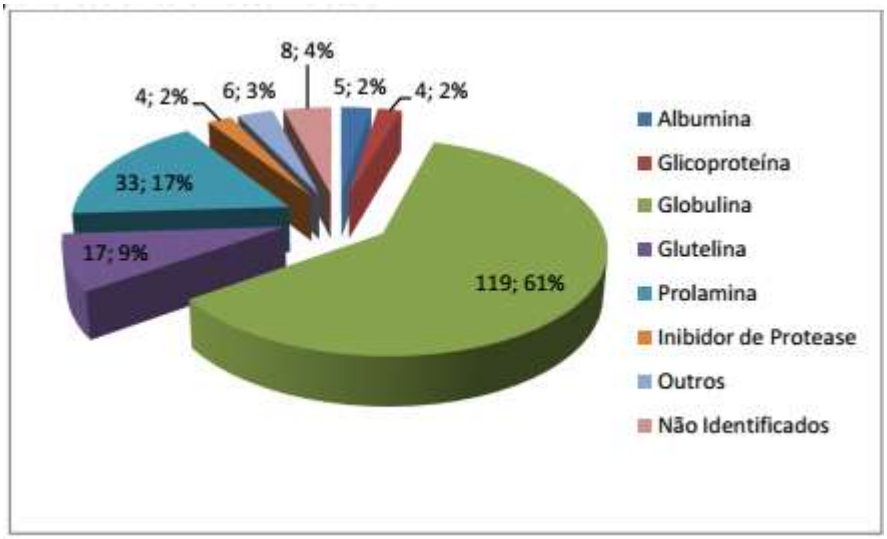
As proteínas estão distribuídas em tiras de 13 cm em uma faixa de pH 3 a 10. A esquerda marcador de peso molecular variando de 14 a 90 KDa; R1 – região de difícil resolução devido a presença de proteínas muito expressas; R2 – região de difícil resolução por conter proteínas em pH alcalino; R3 – região onde a intensidade dos *spots* variou nas variedades de cajueiro estudadas. **Fonte**: O Autor.

Gráfico 29 – Identificação dos spots proteicos presentes nos géis do cajueiro CCP 76 com base em dados de ponto isoelétrico e massa molecular.



As identificações foram feitas utilizando o programa *TagIdent*, do *ExPASy* contra o banco de dados do Swiss-Prot. **Fonte**: O Autor.

Gráfico 30 – Identificação dos spots proteicos presentes nos géis do cajueiro CCP 76 com base em dados de ponto isoelétrico e massa molecular.



As identificações foram feitas utilizando o programa *TagIdent*, do *ExPASy* contra o banco de dados do Swiss-Prot. **Fonte**: O Autor.

5.5.3 Eletroforese Bidimensional em estreita faixa de pH

Com os resultados descritos anteriormente, pode-se perceber que a maioria dos *spots* protéicos estão distribuídos em uma faixa de pH de 5-7, e pelo fato de haver uma grande concentração destas proteínas em uma faixa estreita, algumas delas não aparecem bem definidas. Assim, foram utilizados novos géis com tiras de pH 4-7 para as mesmas amostras.

Os géis bidimensionais de proteínas de semente de cajueiro comum (Figura 20) e cajueiro CCP 76 (Figura 21) apresentaram 263 e 390 *spots*, respectivamente. O número de *matches* presente nos géis de referência do cajueiro comum e CCP 76 são 343 e 216, respectivamente. Os coeficientes de correlação linear para os mesmos (cajueiro comum e CCP 76) são de 0,991 e 0,968, respectivamente (Gráfico 31).

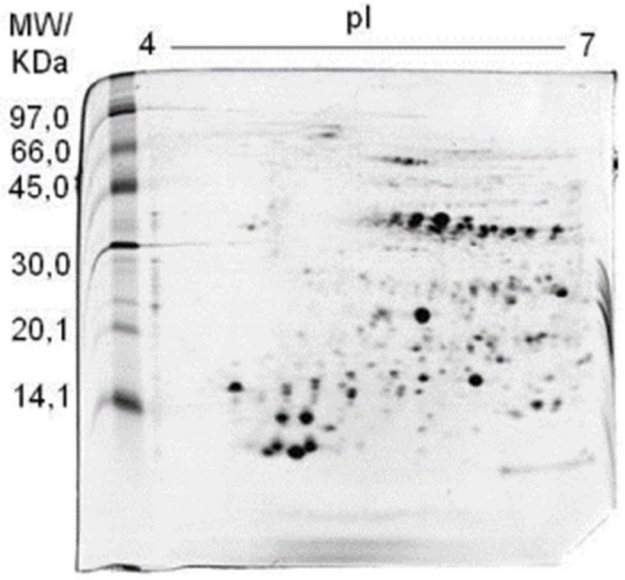
A distribuição das proteínas do cajueiro por ponto isoelétrico (Gráfico 32) mostra que o cajueiro comum possui uma maior diversidade de *spots* nas faixas de pH próximos de 6. A distribuição das proteínas de acordo com sua massa molecular (Gráfico 33) mostra que o cajueiro comum possui uma maior variedade de *spots* com massa entre 20 e 50 KDa.

A identificação das proteínas diferencialmente expressas entre os genótipos de cajueiro é essencial para a escolha de biomarcadores e a compreensão das diferenças fenotípicas. Algumas proteínas diferencialmente expressas são mostradas na (Figura 22). Uma identificação baseada nos dados de ponto isoelétrico e massa molecular, semelhante a outras proteínas de sementes, foi feita utilizando o software *TagIdent* contra o banco de dados do *SwissProt*.

Os *spots* 288, 459 e 461 identificados pelo *TagIdent* como glicinina (P02858), glicoproteína (P10743), e conglutina (Q6PSU2-3) respectivamente, mostraram-se mais expressos no cajueiro CCP 76 quando comparada com o cajueiro comum.

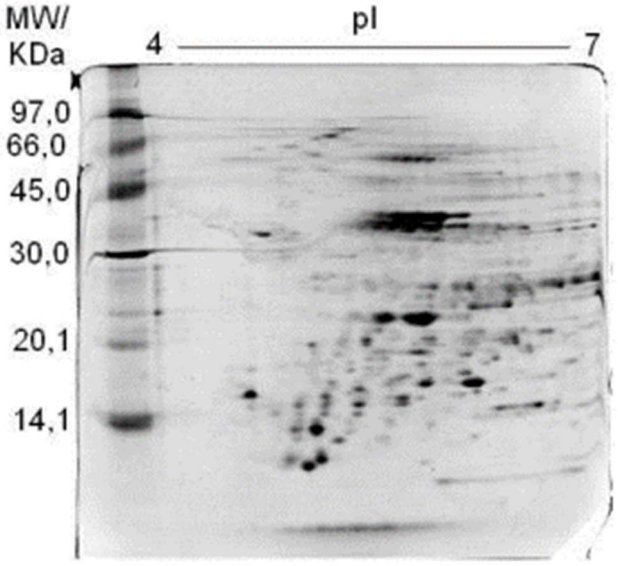
Os *spots* 561, 562, 583, 594 e 597 não foram identificados pelo *TagIdent*, e foram mais expressos no cajueiro comum. Por outro lado, os *spots* 288, 459, 461 e 481 mostraram-se mais expressos no cajueiro CCP 76 (Figura 22).

Figura 20 – Gel bidimensional das proteínas totais da amêndoa da castanha de cajueiro comum usando tiras de pH 4-7.



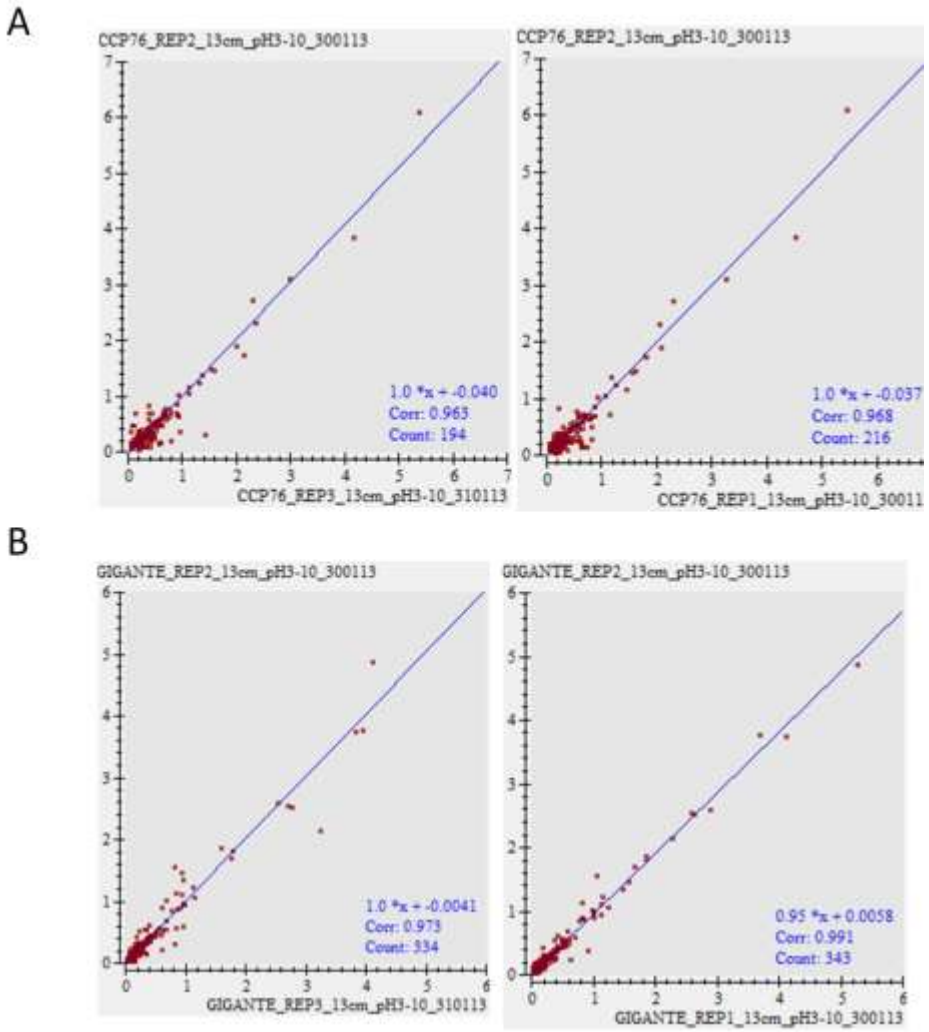
As proteínas estão distribuídas em tiras de 13 cm em uma faixa de pH 3 a 10. A esquerda marcador de peso molecular (MW) variando de 14 a 90 KDa. **Fonte**: O Autor.

Figura 21 – Gel bidimensional das proteínas totais da amêndoa da castanha de cajueiro CCP 76 usando tiras de pH 4-7.



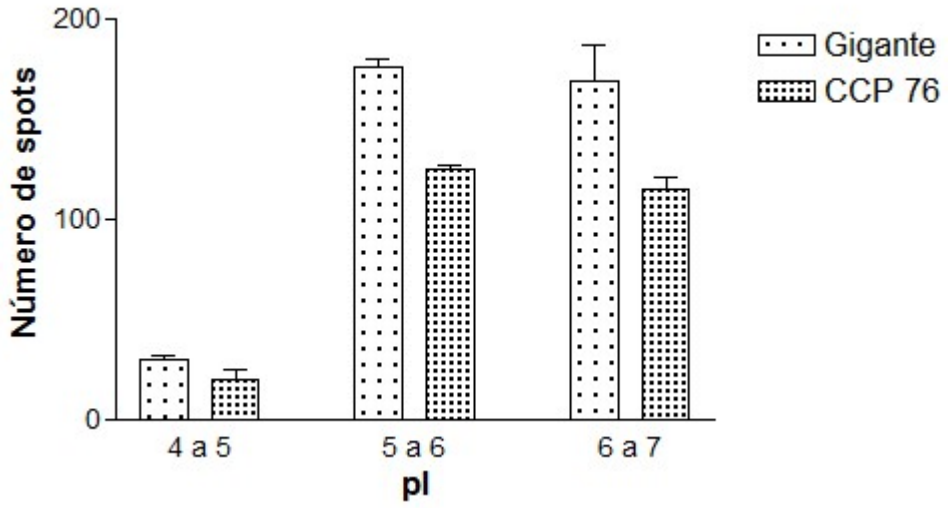
As proteínas estão distribuídas em tiras de 13 cm em uma faixa de pH 3 a 10. A esquerda marcador de peso molecular (MW) variando de 14 a 90 KDa. **Fonte**: O Autor.

Gráfico 31 – Dispersão dos *spots* protéicos das replicatas de géis bidimensionais de sementes do cajueiro comum e CCP 76.



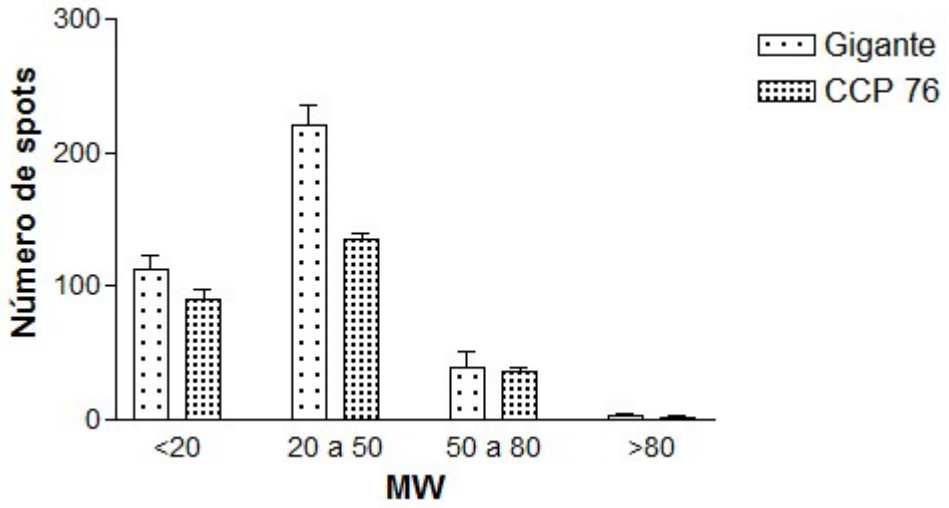
(A) gráfico de dispersão das replicatas 1 e 2 contra o gel de referencia do cajueiro CCP 76. **(B)** gráficos de dispersão das replicatas 1 e 2 contra o gel de referencia do cajueiro comum. O gráfico foi feito de acordo com a porcentagem do volume dos *spots*. **Corr.** - correlação linear entre as replicatas; **Count.** – Número de *spots* compartilhados (*matches*) entre as replicatas. Os gráficos foram obtidos pelo programa *ImageMaster Platinum* versão 6.0. **Fonte**: O Autor.

Gráfico 32 – Distribuição dos spots protéicos de cajueiro comum e cajueiro CCP 76 de acordo com o ponto isoelétrico (pI).



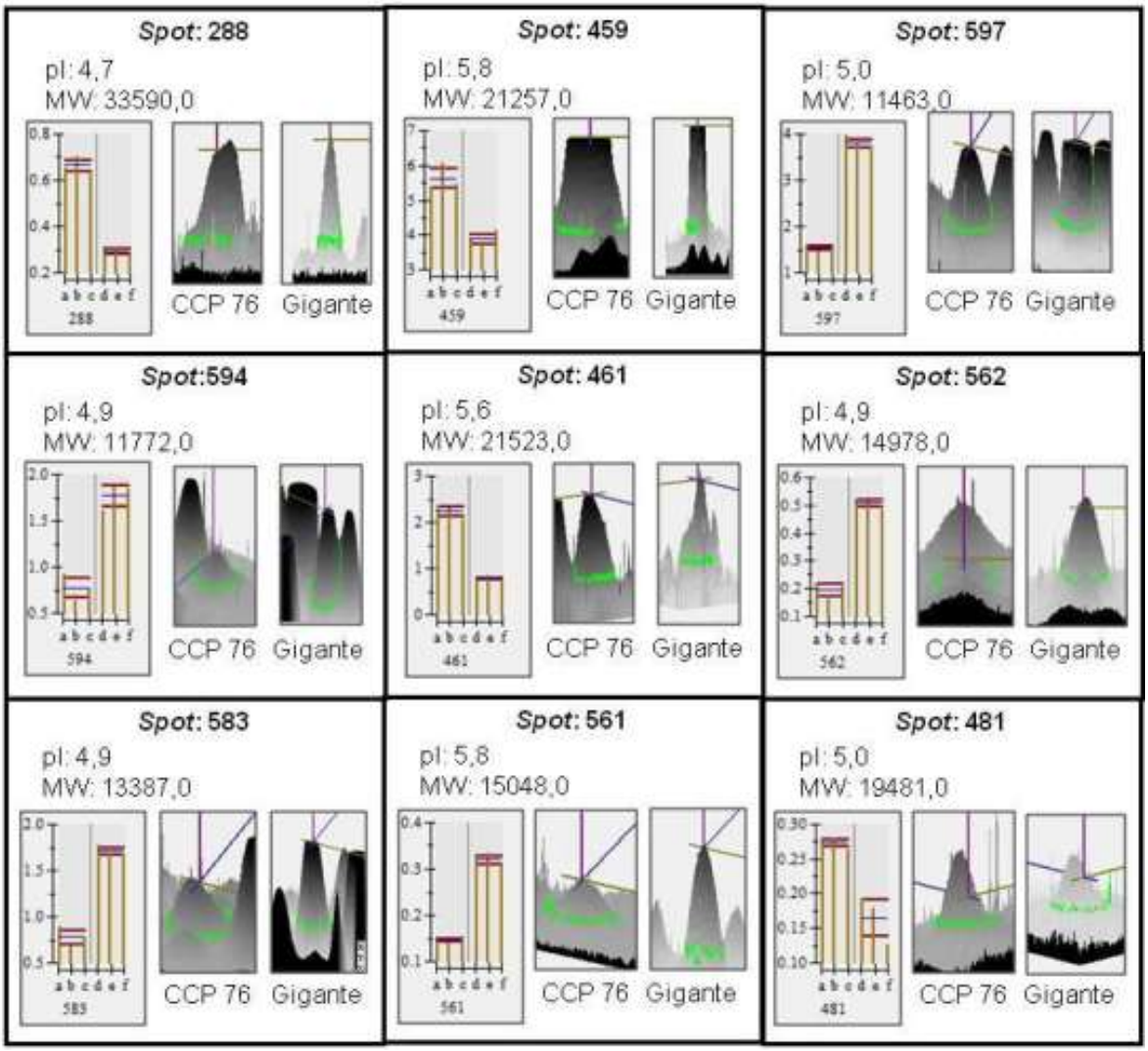
Fonte: O Autor.

Gráfico 33 - Distribuição dos spots protéicos de cajueiro comum e cajueiro CCP 76 de acordo com a massa molecular (MW).



Fonte: O Autor.

Figura 22 – Proteínas diferencialmente expressas nos géis bidimensionais de proteínas de sementes de cajueiro comum e CCP 76 separados na faixa de pH de 4-7.



O painel está dividido em nove quadros contendo informações sobre os *spots* diferencialmente expressos. Acima o número do *spot* com os valores de ponto isoelétrico (pI) e massa molecular (MW); o histograma à esquerda mostra os valores de porcentagem do volume dos *spots* nas replicatas dos géis do cajueiro CCP 76 (a, b) e do cajueiro comum (c, d) onde a linha amarela vertical representa a porcentagem do volume do *spot*, a linha horizontal azul é a média e as linhas horizontais vermelhas representam o desvio padrão; as imagens à direita representam visualizações tridimensionais dos *spots* presentes nos géis de referencia do cajueiro CCP 76 e cajueiro comum. **Fonte**: O Autor.

5.6 DISCUSSÃO

A eletroforese bidimensional é um dos métodos mais poderosos para analisar o proteoma completo de uma célula, tecido ou órgão em estudos proteômicos (WITTMANN-LIEBOLD; GRAACK; POHL, 2006). A importância de estudos proteômicos, juntamente com a transcriptômica, ajuda a compreender os reais níveis de proteínas em uma célula, pois os níveis de RNA nem sempre são consistentes com a abundância relativa de proteínas (ANDERSON; ANDERSON, 1998); (GYGI et al., 1999).

A análise proteômica de semente de amendoim (*Arachis hypogea*) revelou a presença de 169 *spots* protéicos usando uma tira de pH 3-11 NL (SCHMIDT et al., 2009). De acordo com nossos resultados, obtivemos 160 a 200 *spots* protéicos em sementes de cajueiro usando tiras de pH 3-10 L. Para cada gel foi feito uma repetição onde foram obtidos números semelhantes (dados não mostrados).

Estudos realizados com sementes de *Medicago truncatula* mostraram que as proteínas mais abundantes identificadas por espectrometria de massa foram globulinas 7S (vicilinas e convicilinas) e globulinas 11S (leguminas) (GALLARDO et al., 2003). Com relação ao cajueiro, utilizando a ferramenta *TagIdent* que identifica os *spots* baseado nos valores de pI e massa molecular, a maioria dos *spots* identificados pertencem a classe das globulinas.

De acordo com os gráficos de distribuição dos *spots* por ponto isoelétrico e massa molecular, observamos que as proteínas de semente de cajueiro ocorrem com maior frequência no intervalo de pI entre 5 e 7. Estudos eletroforéticos de *Araucaria angustifolia* demonstraram também que a disposição preferencial das proteínas, quando submetidas ao fracionamento em gradiente amplo de pH (3-10), é em torno do pH 6 (BALBUENA, 2009).

Para aumentar a resolução dos géis e consequentemente o número de *spots* protéicos próximos do pH 6, foram feitos novos géis com faixa de pH de 4-7L onde número de *spots* aumentou para cerca de 260 a 300 utilizando 240 µg de proteínas. Um experimento semelhante foi feito com sementes de cevada (*Hordeum vulgare*) onde 200 µg de proteínas foram aplicadas em tiras de pH 4-7 e os géis bidimensionais foram corados com azul de comassie coloidal resultando na revelação de 600 *spots*

protéicos (OSTERGAARD et al., 2004).

A análise intra-classe dos géis bidimensionais de cajueiro comum revelou que mais de 50 % dos *spots* são compartilhados entre as triplicatas e que todos os coeficientes de correlação linear estão acima de 0,9. O coeficiente de correlação linear mede o grau de correlação entre duas variáveis e deve assumir um valor entre zero e um. De acordo com a correlação de Pearson, valores de correlação linear acima de 0,85 são aceitáveis para indicar a reprodutibilidade entre os géis (VIEIRA, 1980).

A proteômica funcional tem como objetivo a identificação de proteínas e suas respectivas funções biológicas em uma célula ou tecido, enquanto que a proteômica quantitativa visa às alterações nos níveis de expressão de proteínas, que podem ser úteis como marcadores moleculares (LIEBLER, 2002). A análise dos histogramas dos géis bidimensionais de cajueiro comum e comum revelou proteínas com diferentes valores de porcentagem do volume, sugerindo alterações nos níveis de expressão.

A comparação entre dois géis de referência pode revelar sobre as alterações nos níveis de expressão de proteínas desde que as alterações na intensidade ou volume do spot sejam superiores a duas vezes para ser considerado como diferença estatística (RIGHETTI et al., 2004). Encontramos seis *spots* protéicos diferencialmente expressos nos géis do cajueiro comum e CCP 76, que podem ser utilizados como marcadores moleculares.

5.7 CONCLUSÃO

Com base nos resultados obtidos, foi possível estabelecer um protocolo experimental reprodutível para eletroforese bidimensional do cajueiro. Foram encontradas seis proteínas com níveis de expressão diferenciados no cajueiro comum e CCP 76. A maioria das proteínas identificadas com base em dados de pI e massa pertence à classe das globulinas.

CAPÍTULO 6:
PROTEÔMICA COMPARATIVA DAS PROTEÍNAS DE SEMENTE DE QUATRO VARIEDADES DE CAJUEIRO

6 PROTEÔMICA COMPARATIVA DAS PROTEÍNAS DE SEMENTE DE QUATRO VARIEDADES DE CAJUEIRO.

6.1 INTRODUÇÃO

6.1.1 Clones de cajueiro anão-precoce

Uma das mais importantes contribuições do melhoramento genético de plantas tem sido o desenvolvimento de variedades para diferentes ambientes. O uso de clones resistentes representa uma forma de manejo econômico, ecológico e seguro, impedindo a invasão de pragas e doenças, além de proporcionar uma melhor utilização da variabilidade genética da espécie. (PAIVA; BARROS, 2004).

O cajueiro tipo anão-precoce, também conhecido por cajueiro de seis meses, caracteriza-se pelo porte baixo, com altura de planta em torno de 3 a 4 m, copa compacta e homogênea, e envergadura de copa média em torno de 7 a 9 m. Inicia o florescimento entre 6 e 18 meses. O peso do fruto varia de 3-10 g e do pedúnculo de 20-160 g (LIMA, 1988).

Os primeiros clones comerciais de cajueiro do tipo anão-precoce, o CCP 06, CCP 76, CCP 09 e CCP 1001 foram lançados nos anos 80, visando uma melhoria da qualidade de castanhas. Em 1996, foram implantados dois outros clones de cajueiro do tipo anão-precoce, com a denominação de Embrapa 50 e Embrapa 51. Para o plantio irrigado foi lançado o clone BRS 189, tendo aproveitamento da castanha e do pedúnculo para o mercado de mesa, devido ao alto teor de vitamina C e baixa quantidade de taninos. Para plantio comercial na região dos Baixões Piauienses, foi lançado o clone BRS 226. Estudos posteriores mostram que o clone BRS 226 apresenta resistência a resinose (PAIVA et al., 2008; PAIVA; BARROS, 2004). Uma comparação entre as diferentes variedades de cajueiro pode ser vista na tabela 10.

6.1.2 Proteômica visando caracterização de genótipos

Com relação à caracterização de genótipos, pode-se afirmar que alguns trabalhos têm sido bem-sucedidos na identificação de proteínas exclusivas de cada genótipo, podendo estas serem utilizadas como marcadores. Um exemplo é o estudo comparativo de dois cultivares de arroz (*Oryza sativa* L.) que, embora pertencendo a

mesma subespécie *indica*, têm diferenças em respeito a sua morfologia, fisiologia e qualidade do grão. Variações na composição de proteínas no endosperma foram estudadas por comparação dos mapas 2-DE para esses dois cultivares de arroz. Depois da análise por MALDI-TOF, algumas proteínas foram encontradas como sendo exclusivamente de um dos genótipos e vice-versa mostrando que é possível realizar estudos comparativos dos dois cultivares de arroz via proteômica (YANG; SHEN; KUANG, 2006).

Quatro genótipos de girassol (*Helianthus annus*) com diferentes níveis de óleo e ácido oléico foram comparados por 2-DE e LC-MS/MS revelando alterações nas enzimas envolvidas na glicólise e na síntese de aminoácidos. Isso sugere que o conteúdo de óleo na semente dos genótipos é fortemente ligado ao metabolismo de carboidratos e síntese de proteínas de uma maneira complexa (HAJDUCH et al., 2007).

A composição de proteínas pode ser afetada pelo melhoramento e pelas condições ambientais. Sendo assim, estudos envolvendo a comparação das principais proteínas de reserva da soja têm mostrado variações no conteúdo dessas proteínas. Diferentes *spots* de proteína foram encontrados pela comparação das proteínas de reserva da soja selvagem (*Glycine soja*) e da soja cultivada (*Glycine max*) (NATARAJAN et al., 2006). Posteriormente, estudos envolvendo apenas as 5 subunidades da glicinina têm sido eficientes para a caracterização de 16 diferentes genótipos de soja (NATARAJAN et al., 2007).

Além da caracterização de genótipos, a eletroforese bidimensional, quando combinada com Western Blot é eficiente na identificação de alérgenos. Tal metodologia foi feita utilizando anticorpo IgG anti Ara h 3 de amendoim (*Arachis hipogea*) que pôde detectar a presença de polipeptídeos homólogos ao Ara h 3 no tremoço-branco (*Lupinus albus*) e na soja (*Glycine max*) levando a descoberta de novos alérgenos putativos (MAGNI et al., 2005).

Tabela 10 - Comparação entre os diferentes clones de cajueiro anão-precoce.

	Peso da castanha	**Amêndoa despeliculada**	**Peso médio do pedúnculo**	**Coloração do pedúnculo**
CCP 09	7,7g	2,1 g	87 g	Laranja
CCP 76	8,6g	1,8 g	135 g	Laranja
BRS 226	9,75g	2,72 g	102,6 g	Laranja
BRS 275	11,4g	3,13 g	108,0 g	Laranja

Fonte: Adaptado de: (BARROS et al., 1993; PAIVA; BARROS, 2004).

6.2 OBJETIVOS

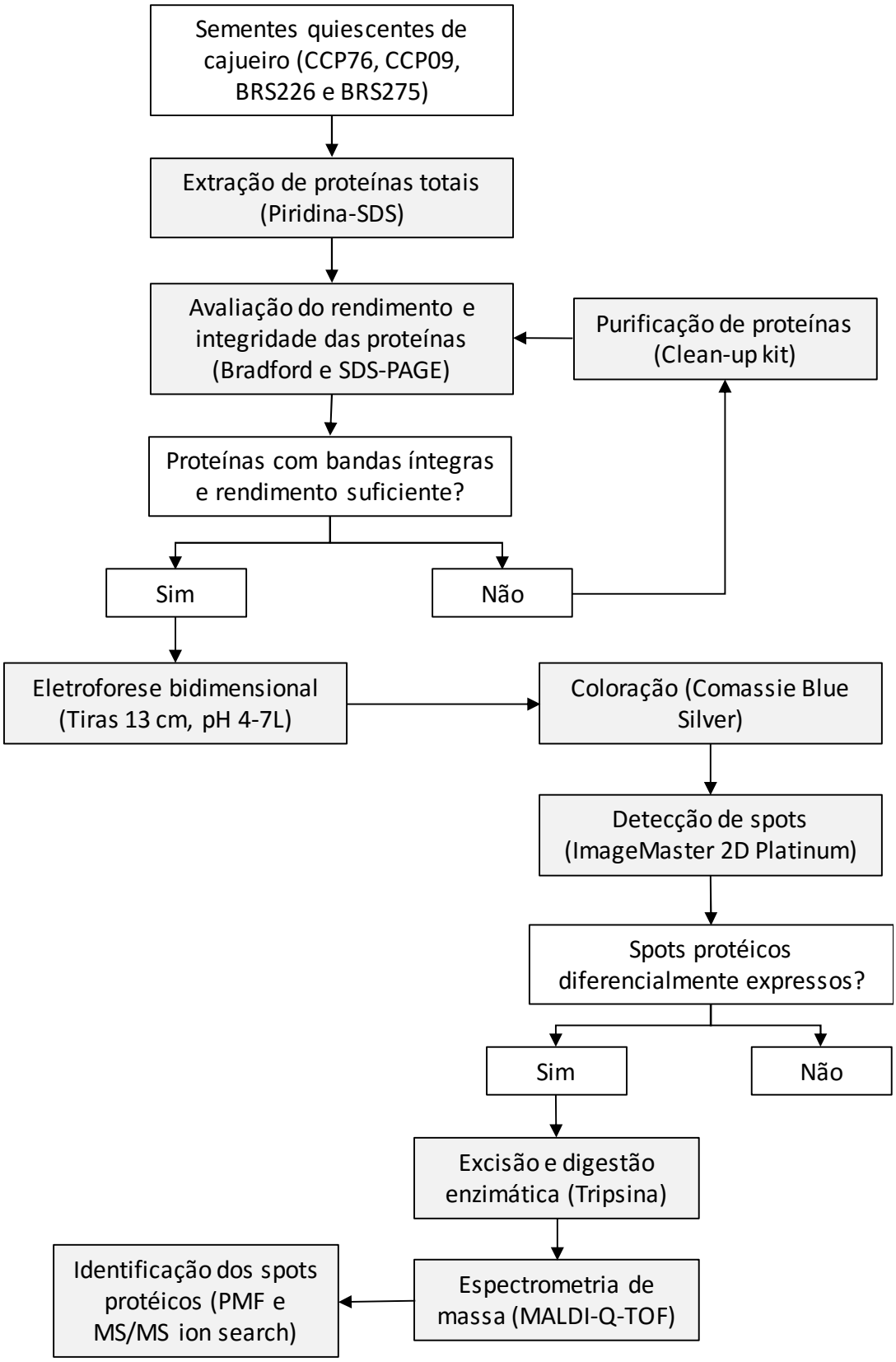
6.2.1 Geral

Comparar o perfil protéico de sementes de quatro variedades de cajueiro (CCP 76, CCP 09, BRS 226 e BRS 275).

6.2.2 Específicos

- Obter géis bidimensionais de qualidade para as quatro variedades de cajueiro.
- Buscar proteínas diferencialmente expressas nas quatro variedades de cajueiro que possam ser usadas como marcadores moleculares.
- Identificar as proteínas mais expressas no cajueiro CCP 76 por espectrometria de massa.

6.3 ESTRATÉGIA EXPERIMENTAL



6.4 METODOLOGIA

6.4.1 Coleta de material biológico

As sementes de cajueiro (CCP 76, CCP 09, BRS 226 e BRS 275) foram adquiridas da Embrapa Agroindústria Tropical (Fortaleza, Ceará). As sementes foram utilizadas para a extração de proteínas em condições normais (quiescentes).

6.4.2 Extração de proteínas totais de semente

Frutos (castanhas) do cajueiro (comum e CCP 76) foram cuidadosamente quebrados para a liberação das amêndoas que posteriormente foram cortadas em cubos de aproximadamente 1 mm. Em seguida, as amêndoas cortadas (aproximadamente 15 g) foram delipidadas em 100 mL de acetona sob agitação constante durante uma noite. A acetona foi trocada quatro vezes. Logo após, foram maceradas em pistilo e almofariz na presença de nitrogênio líquido até a obtenção do pó (farinha). As proteínas totais das sementes foram extraídas usando 20 mg de farinha da castanha acrescido de 40 mg de PVPP (polivinilpolipirrolidona) e 800 µL de tampão piridina-SDS (piridina 50 mM, tiouréia 10 mM, SDS 1%, pH 5,0) (proporção de 1:2:40). A mistura foi agitada por 2 h e 30 min a 4ºC seguido de centrifugação a 10.000 x g por 40 min. Ao sobrenadante foram adicionados quatro volumes de acetona gelada contendo 10% de TCA (ácido tricloroacético) *overnight* a -20 ºC (DAMERVAL et al., 1986). Em seguida, a amostra foi centrifugada a 10.000 x g por 30 min e o sobrenadante descartado. O precipitado contendo as proteínas totais foi lavado com 1mL de acetona gelada 100% (3 vezes) e seco a temperatura ambiente. Posteriormente foram solubilizadas em 200 µL de uréia/tiouréia 9 M.

6.4.3 Dosagem de proteínas

As proteínas totais foram quantificadas pelo o método de Bradford (1976) e a integridade protéica foi analisada por SDS-PAGE (LAEMMLI, 1970).

6.4.4 Eletroforese bidimensional

As proteínas da amêndoa da castanha de caju de cada genótipo (250 µg), foram diluídas em solução de reidratação (uréia 7 M, tiouréia 2 M, DTT 65 mM, CHAPS 1% p/v, *IPGBuffer* 0,5% v/v e azul de Bromofenol 0,002% p/v) para um volume final

de 250 µl (GÖRK et al., 2007). As amostras diluídas na solução de reidratação foram aplicadas no *IPGBox* (*GE Healthcare*) e posteriormente incubadas com tiras com gradiente de pH imobilizado (*IPGStrip*) de 13 cm e faixa de pH linear de 3-10, por um período de 16 h. A focalização isoelétrica (IEF) foi realizada no equipamento *Ettan*™ *IPGPhor III*™ (*GE Healthcare*) utilizando as seguintes condições: etapa 1 (500 V por 0:30 h); etapa 2 (4000 V por 2:30 h) e etapa 3 (10000 V até atingir 18.000 Vh totais). Após a IEF, as tiras foram armazenadas em freezer -20 °C para posterior utilização.

Após a focalização, as tiras de IPG foram equilibradas, sob agitação, em solução de equilíbrio (tris 50 mM, glicerol 30%, uréia 6 M, SDS 2% e azul de bromofenol 0,002 %) com ditiotreitol (DTT) a 1% por 15 min para a redução das proteínas e alquiladas com iodoacetamida (IAA) a 3%, também em solução de equilíbrio por 15 min. Terminado o equilíbrio as tiras foram mergulhadas em tampão de corrida por 10 s para retirar o excesso da solução de equilíbrio. Em seguida, as tiras foram postas no topo dos géis da segunda dimensão e cobertos com 2 mL de agarose morna (agarose 0,5%, SDS 1% e azul de bromofenol 0,002%) adicionando um pente para a formação do poço do marcador e deixados até solidificar. A corrida Foi realizada em uma unidade de eletroforese vertical (*Hoefer SE 600 Ruby, Amersham Biosciences*®) utilizando os seguintes parâmetros: 15 mA/gel por 15 min e em seguida 25 mA/gel a uma temperatura de 18 °C e permanecendo assim até que o azul de bromofenol atinja o limite inferior dos géis. Após a corrida, os géis foram colocados em solução de fixação, composta por etanol, ácido acético e água (4:1:5 v/v/v),durante 15 min, corados com solução de Comassie G-250 (B*lue Silver)* (CANDIANO et al., 2004) por 24 h e armazenados em solução de ácido acético 5%. Os géis foram escaneados utilizando-se o equipamento *Image Scanner III* e gerenciados pelo programa *LabScan 6.0* (ambos da *GE Healthcare*). As imagens obtidas foram analisadas e editadas no programa *Image Master 2D Platinum* 6.0 (GE *Healthcare*), onde os *spots* foram detectados automaticamente pelo programa passando por uma revisão manual para eliminação de *spots* artefactuais.

6.4.5 Espectrometria de massa

Os *spots* de interesse foram excisados dos géis bidimensionais com o auxílio de um bisturi e transferidas para tubos *eppendorf*, onde foi realizada a digestão com tripsina, de acordo com o método de (HELLMAN et al., 1995) com algumas alterações

(NOGUEIRA, 2007). Os *spots* selecionados foram descorados em solução de bicarbonato de amônio 25 mM/acetonitrila (1:1) em pelo menos 3 lavagens de 30 min, e desidratados, duas vezes, com acetonitrila 100% por 5 minutos. O solvente remanescente foi removido dos pedaços de gel em concentrador de amostra *Speed Vac* (LABCONCO). Os géis foram reidratados a 4 °C na proporção de 1:50 (proteínas/tripsina) e a digestão realizada a 37 °C por 16 horas.

Os peptídeos obtidos pela digestão tríptica (1 µL) foram aplicados na placa de MALDI e seco a temperatura ambiente. Em seguida, foi coberto por uma solução saturada de ácido α-ciano-4-hidroxicinâmico em acetonitrila 50%/ácido trifluoroacético 0,1% (1 µL). Após secos a temperatura ambiente, a amostra foi analisada por um espectrômetro de massa híbrido quadrupolo com mobilidade de íon Tempo de Vôo com aceleração ortogonal de alta definição Synapt HDMS (*Waters*). O espectrômetro será operado em modo positivo, com fonte de ionização MALDI.

O espectrômetro de massa possui duas fontes de ionização (MALDI e ESI), analisadores quadrupolo (com camara de colisão) e tempo de vôo além do sistema de *íon mobility* (que permite determinar a mobilidade do íon) (PRINGLE; et al., 2007). O instrumento foi calibrado com íon glucofibrinopeptidio-B (M+H) = 1570,68 Da. Foi selecionada a função DDA (análise direta de dados) para a seleção dos íons que serão fragmentados por decomposição induzida por colisão (CID). Os espectros de MS foram coletados entre 300 m/z a 3000 m/z e os espectros de MS/MS entre 50 m/z a 3000 m/z. A coleta e analise de dados foi feita utilizando o software MassLynx®.

6.4.6 Pesquisa no banco de dados

Os *spots* identificados nos géis bidimensionais pelo *ImageMaster*™ foram inicialmente identificadas utilizando a ferramenta *TagIdent* do *ExPASy* (*Expert Protein Analysis System*). Os espectros de MS e MS/MS gerados a partir da fragmentação dos íons precusores serão processados e analisados utilizando o programa MASCOT (PERKINS et al., 1999) para MS/MS e o *ProFound* (http://prowl.rockefeller.edu) para identificações baseadas em PMF.

6.5 RESULTADOS

6.5.1 Dosagem de proteínas e eletroforese unidimensional

Para verificar a qualidade da amostra utilizou-se a técnica de SDS-PAGE, na qual se observou a presença de bandas íntegras e bem definidas, mostrando que a extração de proteínas de castanha de caju foi eficiente para a obtenção de amostras de boa qualidade (Figura 23). As bandas mais intensas têm massa molecular de 20 e 30 KDa. O número de bandas protéicas é igual em todos os genótipos analisados havendo apenas uma diferença sutil nas intensidades dessas bandas.

Na quantificação por Bradford obtivemos as concentrações das proteínas totais, calculadas a partir das médias das triplicatas de absorbâncias, onde houve pouca variação entre estas medias nos quatro genótipos. As concentrações variaram de 2,81 a 5,34 (Tabela 11).

6.5.2 Análise dos géis bidimensionais

A focalização isoelétrica não mostrou alterações na voltagem nem amperagem de modo que a corrida ocorreu conforme o esperado (Figura 24). Para cada genótipo de cajueiro anão-precoce estudado, foram feitos géis bidimensionais em triplicata a fim de analisar a reprodutibilidade dos géis.

Os géis bidimensionais das quatro variedades de cajueiro mostraram-se muito semelhantes (Figura 25) a ponto de serem confundidas com replicatas. Para avaliar se as diferenças encontradas nas variedades de cajueiro são maiores do que aquelas diferenças metodológicas encontradas nas replicatas de géis, foi feita uma análise estatística de agrupamento heurístico (*heuristic clustering*). As triplicatas de todos os géis foram analisadas e agrupadas automaticamente pelo programa *ImageMaster* criando um dendograma conforme mostrado na figura 26. O dendograma mostra claramente quatro classes contendo tres ramificações em cada classe. O agrupamento feito automaticamente pelo programa é consistente com os dados experimentais mostrando que a variavel biológica é maior que a variavel metodológica.

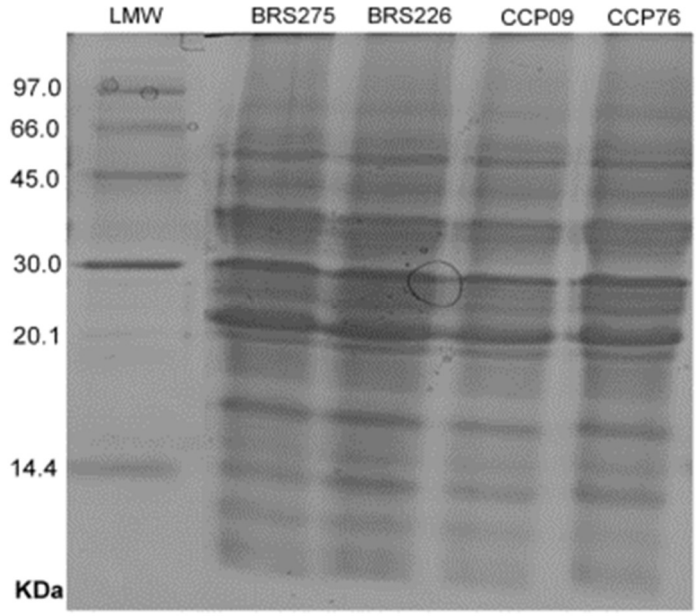
O número médio de *spots* presentes nos géis bidimensionais variou de 298 (CCP 09) a 371 (BRS 275) (Tabela 12). O gráfico de dispersão dos *spots* mostra que

o coeficiente de correlação linear de todos os géis estão acima de 0,9 (Gráfico 34).

O gráfico de distribuição dos *spots* de acordo com o ponto isoelétrico mostrou que a maioria dos *spots* ficaram distribuídos na faixa de pH entre 5-6, sendo que o cajueiro BRS 275 apresentou o maior número de *spots* em relação as outras variedades de cajueiro (Gráfico 35). O gráfico de distribuição dos *spots* de acordo com a massa molecular mostrou que os *spots* estão mais distribuídos na faixa de massa entre 20-60. O cajueiro BRS 275 teve uma maior quantidade de *spots* com tamanhos acima de 60 KDa em relação as outras variedades de cajueiro (Gráfico 36).

Para avaliar as semelhanças globais entre as variedades de cajueiro utilizando os dados de eletroforese bidimensional, foi feita uma análise fatorial (*factor analysis*), o qual mostrou que o cajueiro BRS 275 está mais próximo do cajueiro CCP 76 (Gráfico 37).

Figura 23 - Eletroforese em gel de poliacrilamida (SDS-PAGE) a 12,5%, de proteínas de amêndoas de quatro variedades de cajueiro coradas com comassie R-350.



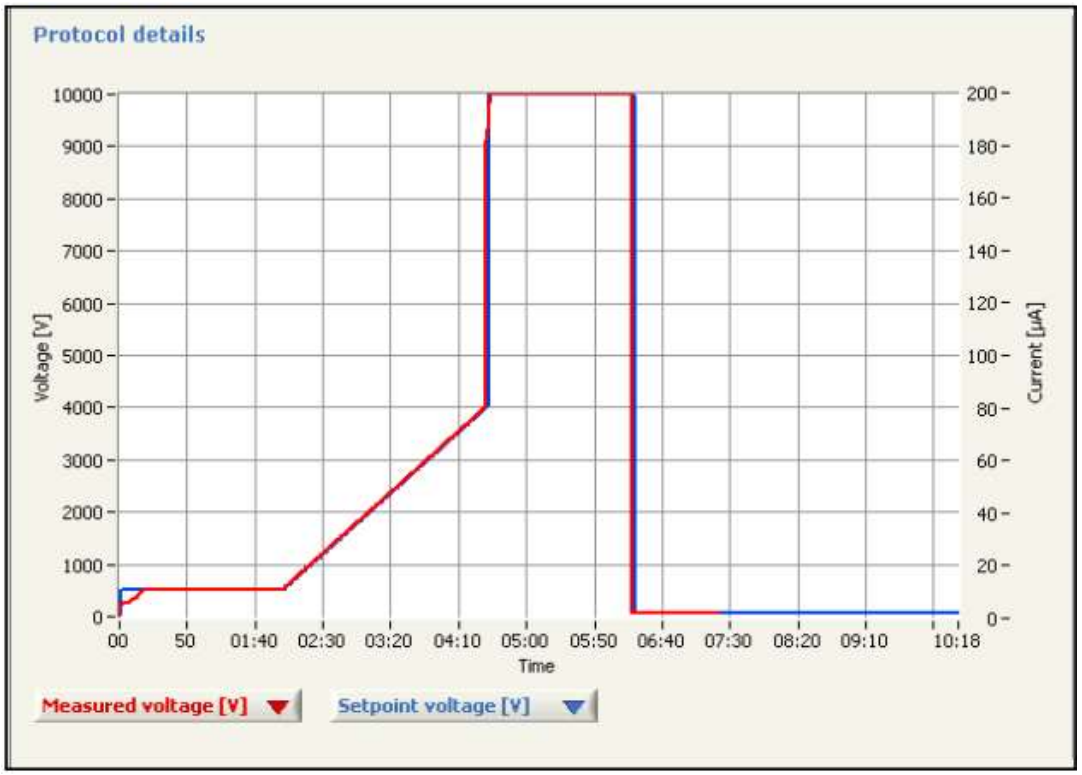
LMW marcador de baixo peso molecular com bandas variando de 14,4 a 97 KDa; Raia 1 – proteínas de sementes de cajueiro comum; raia 2 – proteínas de semente de cajueiro CCP 76. **Fonte**: O Autor.

Tabela 11 - Dosagem de proteínas de castanha dos genótipos do cajueiro anão-precoce (*A. occidentale* var. *nanum* L.).

Genótipos	Concentração de Proteína (µg/µL)	Desvio Padrão
CCP 76	2,81	0,05
BRS 226	3,66	0,06
CCP 09	3,56	0,04
BRS 275	5,34	0,08

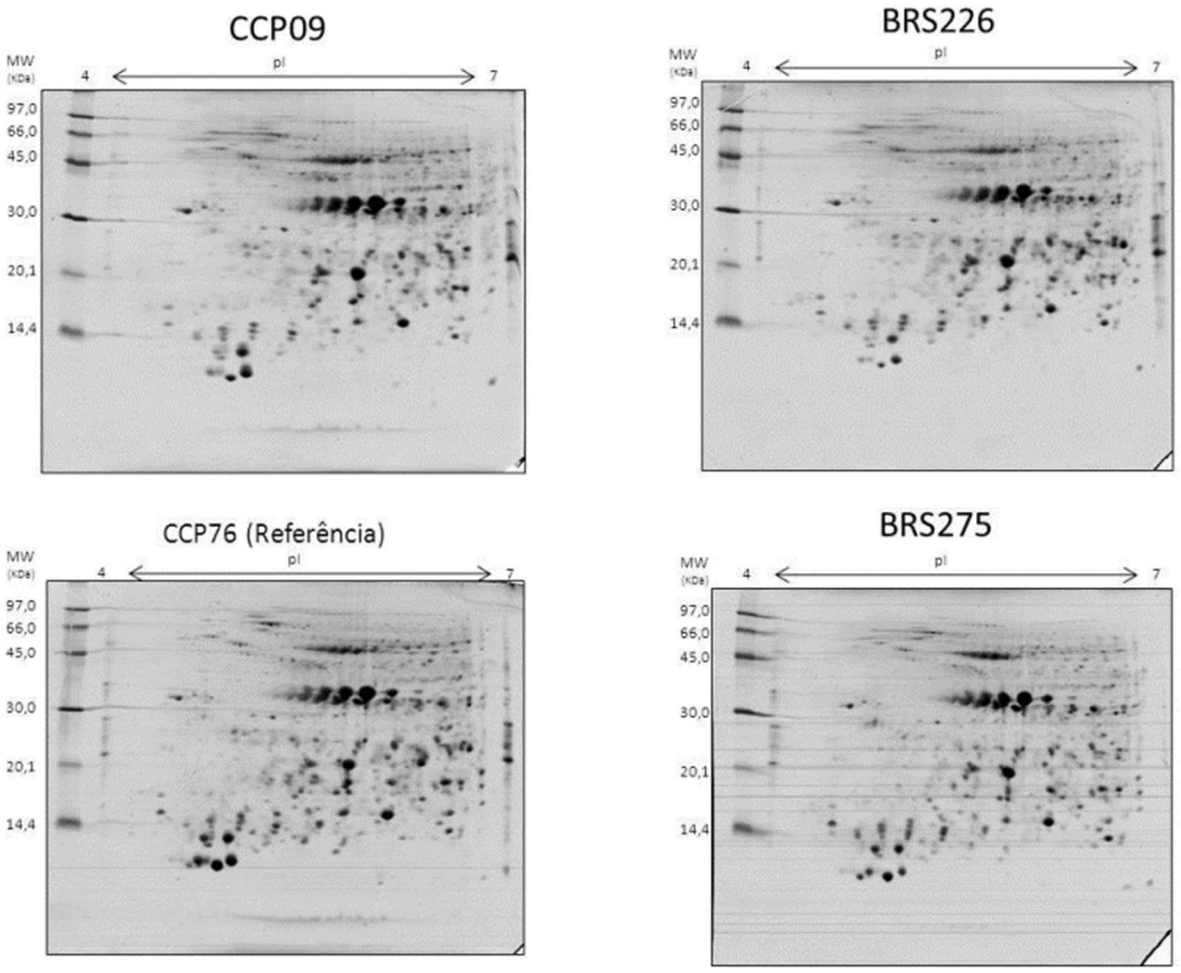
Fonte: O Autor.

Figura 24 – Focalização isoelétrica das proteínas de semente das quatro variedades de cajueiro.



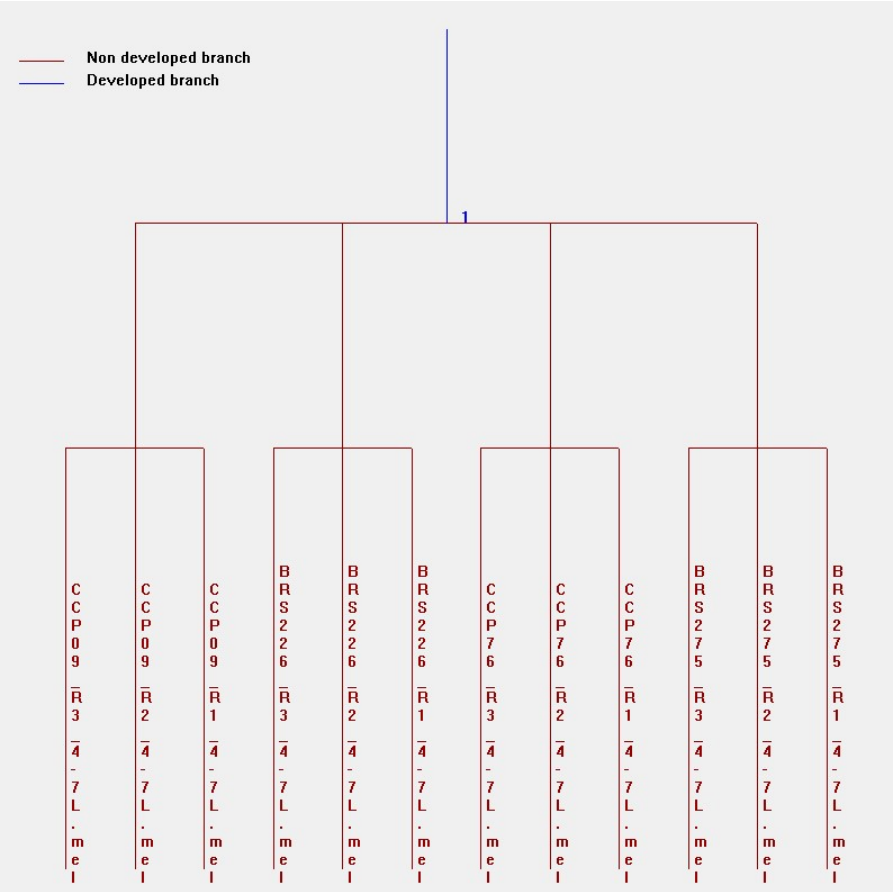
A linha azul indica as condições teóricas de focalização isoelétrica e a linha vermelha indica as condições experimentais. O gráfico foi obtido pelo programa *IPGPhor III*. **Fonte**: O Autor.

Figura 25 - Géis bidimensionais de proteínas de sementes quiescentes de cajueiro CCP 76, CCP 09, BRS 226, BRS 275 usando tiras de pH 4-7.



As proteínas estão distribuídas em tiras de 13 cm em uma faixa de pH 4 a 7. A esquerda marcador de peso molecular (MW) variando de 14 a 90 KDa. **Fonte**: O Autor.

Figura 26 – Agrupamento heurístico (*Heuristic clustering*) dos géis bidimensionais de proteínas de sementes de cajueiro.



Triplicatas dos géis de proteínas de sementes de cajueiro: CCP 09 (CCP 09_R3_4-7L.mel, CCP 09_R2_4-7L.mel, CCP 09_R3_4-7L.mel); BRS 226 (BRS 226_R3_4-7L.mel, BRS 226_R2_4-7L.mel, BRS 226_R1_4-7L.mel); CCP 76 (CCP 76_R3_4-7L.mel, CCP 76_R2_4-7L.mel, CCP 76_R1_4-7L.mel); BRS 275 (BRS 275_R3_4-7L.mel, BRS 275_R2_4-7L.mel, BRS 275_R1_4-7L.mel). O gráfico foi obtido pelo programa *ImageMaster Platinum* versão 6.0. **Fonte**: O Autor.

Tabela 12 - Número de *spots* presentes nos géis bidimensionais das quatro variedades de cajueiro.

Genótipos	Rep 1	Rep 2	Rep 3	Média	Desvio Padrão
CCP 09	284	322	289	298,33	20,65
BRS 226	330	319	289	312,67	21,22
CCP 76	349	270	269	296,00	45,90
BRS 275	377	367	371	371,67	5,03

Fonte: O Autor.

Gráfico 34 - Dispersão dos *spots* protéicos das replicatas de géis bidimensionais de sementes.

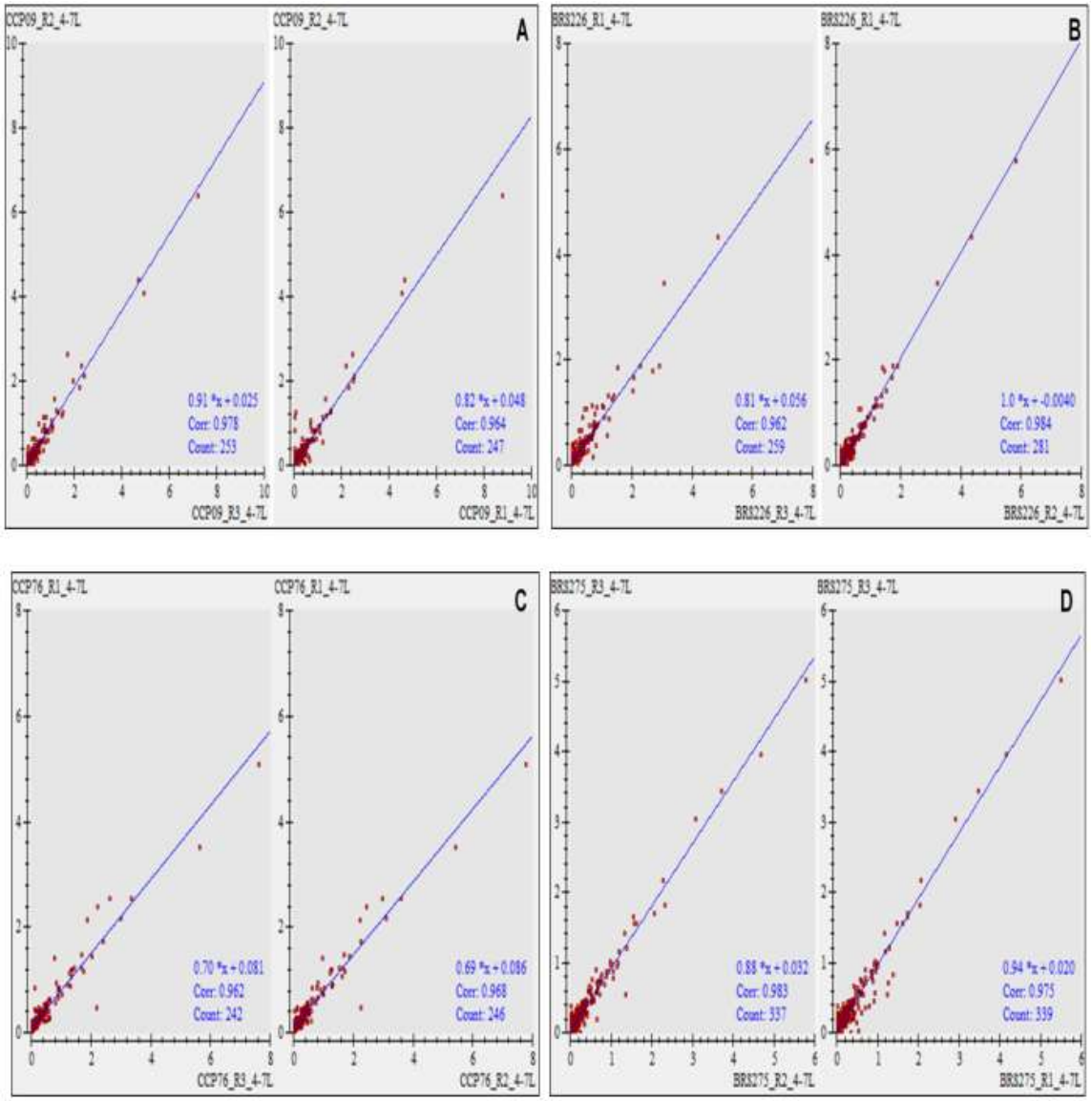
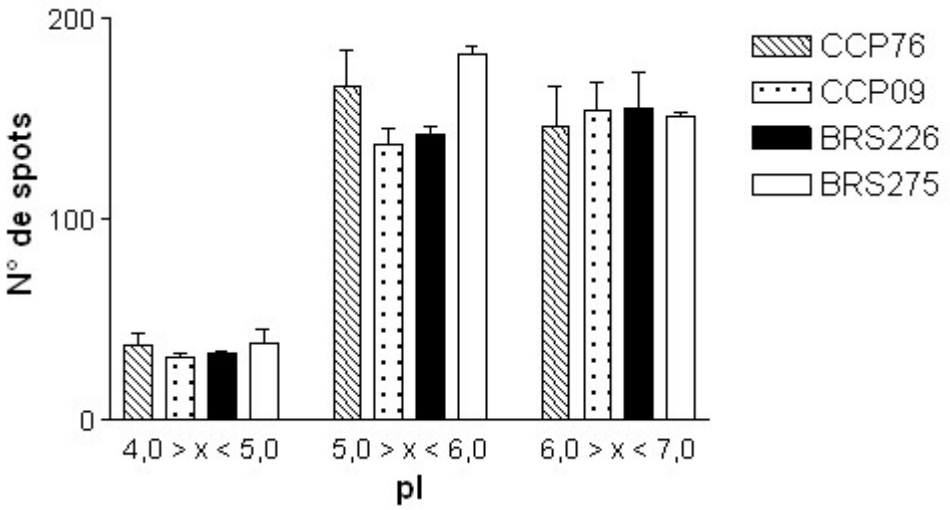


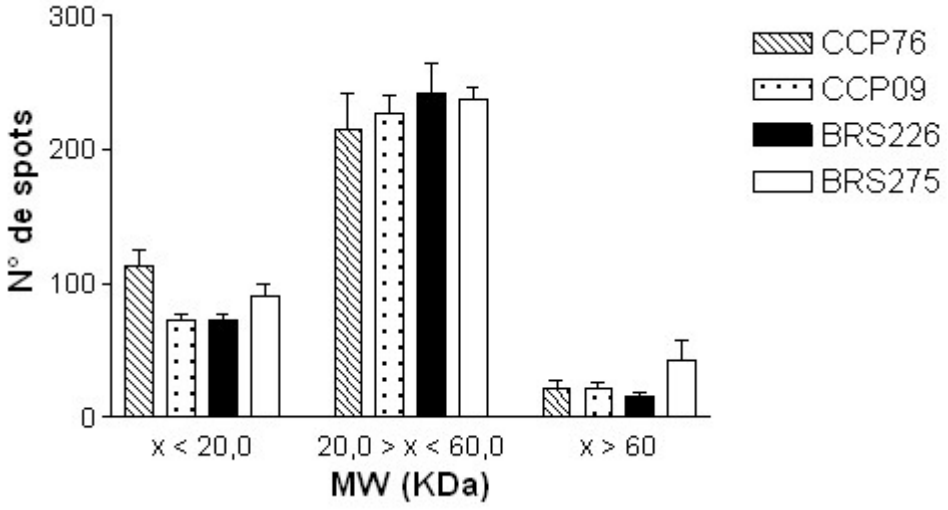
Gráfico de dispersão das replicatas 1 e 2 contra o gel de referencia do cajueiro: (A) CCP 09; (B) BRS 226; (C) CCP 76; (D) BRS 275. O gráfico foi feito de acordo com a porcentagem do volume dos *spots*. Corr. - correlação linear entre as replicatas; Count. – Número de *spots* compartilhados (*matches*) entre as replicatas. Os gráficos foram obtidos pelo programa *ImageMaster Platinum* versão 6.0. **Fonte**: O Autor.

Gráfico 35 - Distribuição dos spots dos géis bidimensionais das quatro variedades de cajueiro de acordo com os valores de ponto isoelétrico (pI).



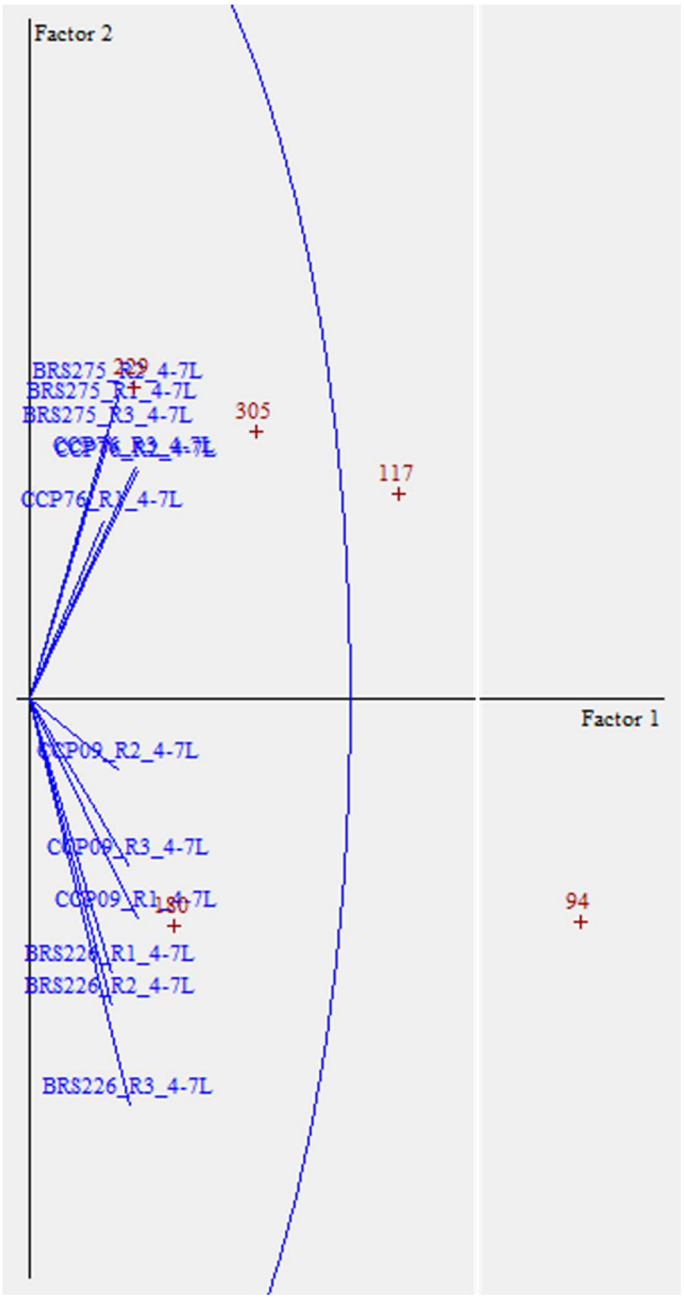
Fonte: O Autor.

Gráfico 36 - Distribuição dos spots dos géis bidimensionais dos genótipos de cajueiro de acordo com os valores de massa molecular (MW).



Fonte: O Autor.

Gráfico 37 – Análise fatorial (*factor analysis*) dos géis bidimensionais de proteínas de semente de quatro variedades de cajueiro.



Triplicatas dos géis de proteínas de sementes de cajueiro: CCP 09 (CCP 09_R3_4-7L, CCP 09_R2_4-7L, CCP 09_R3_4-7L); BRS 226 (BRS 226_R3_4-7L, BRS 226_R2_4-7L, BRS 226_R1_4-7L); CCP 76 (CCP 76_R3_4-7L, CCP 76_R2_4-7L, CCP 76_R1_4-7L); BRS 275 (BRS 275_R3_4-7L, BRS 275_R2_4-7L, BRS 275_R1_4-7L). O gráfico foi obtido pelo programa *ImageMaster Platinum* versão 6.0. **Fonte**: O Autor.

6.5.3 Identificação de proteínas por espectrometria de massa

O gel de referência contém 349 *spots*. A maioria dessas proteínas foram excisadas do gel e levadas ao espectrômetro de massa. O critério de escolha foram *spots* diferencialmente expressos e os *spots* mais expressos. Um total de 96 *spots* ionizaram gerando arquivos (.pkl). Desse número, 65 proteínas foram identificadas por ferramentas online. Apenas seis proteínas foram identificadas por MS/MS utilizando a ferramenta MASCOT (Tabela 13). Através dos dados de saída do MASCOT contendo valores de massa, foi possível fazer identificação por PMF usando a ferramenta ProFound resultando num total de 59 identificações, algumas estão mostradas na tabela 14. Vale ressaltar a busca foi filtrada para plantas (*Viridiplantae*).

A maioria das proteínas (48 *spots*) está identificada como hipotéticas preditas ou desconhecidas (Gráfico 38). As enzimas identificadas somam um total de seis: fosfoenolpiruvato carboxilase (gi | 81158942), clorofila a oxigenase (gi|238767608), pectinesterase (gi|357510873), maturase (gi|21388458), N-acetiltransferase (gi|357139437), celulose sintase (gi|114509162).

Entre as demais proteínas destacam-se a metalotioneína (gi|110559312), provavelmente relacionada à ligação com selênio, e uma proteína de choque térmico classe II de 17,3 KDa (HSP21_SOLPE).

No entanto, mesmo tentando identificar as proteínas por PMF, 37 não foram identificadas pelo ProFound. Por outro lado, usando o MASCOT, o número de proteínas não identificadas foi 105 *spots*.

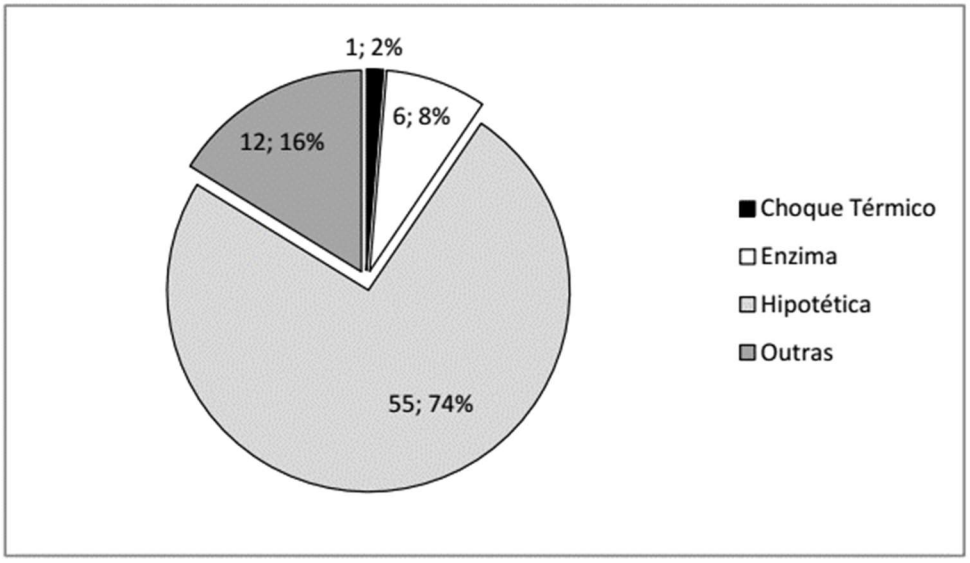
Algumas das proteínas identificadas por espectrometria de massa localizadas no mapa de referência do cajueiro CCP 76 estão mostradas na Figura 27.

Tabela 13 - Lista das proteínas de semente quiescente de cajueiro CCP 76 analisadas por MALDI-QUAD-TOF e identificadas por MS/MS.

ID	Num. Acesso	Nome da proteína	*Score*	*Expect*	*Matches*	Massa (KDa)
25	UP09_LACSN	Protein desconhecida 9 de 2D-PAGE	23	2.7e+03	1	1099
55	HSP21_SOLPE	Proteína de choque térmico 17.3 kDa classe II	61	0.0012	1	17312
80	TL33_SPIOL	Proteína do lúmen do tilacóide 33.6 kDa (Fragmento)	17	6.5e+02	1	1274
30	RR15_DIOEL	Proteína 30S ribossomal S15, do cloroplasto	20	3.1e+02	2	10827
4	RR4_ELEIN	Proteína ribossomal 30S S4, do cloroplasto (Fragmento)	51	0.26	4	22900
00	CWP20_SOLLC	Proteína de parede celular 66 kDa (Fragmento)	26	83	2	2175

Foi utilizado o programa MASCOT para as identificações por MS/MS contra o banco do Swiss-Prot. **Fonte**: O Autor.

Gráfico 38 – Visão geral das proteínas de semente de cajueiro CCP 76 quiescente identificadas por espectrometria de massa.



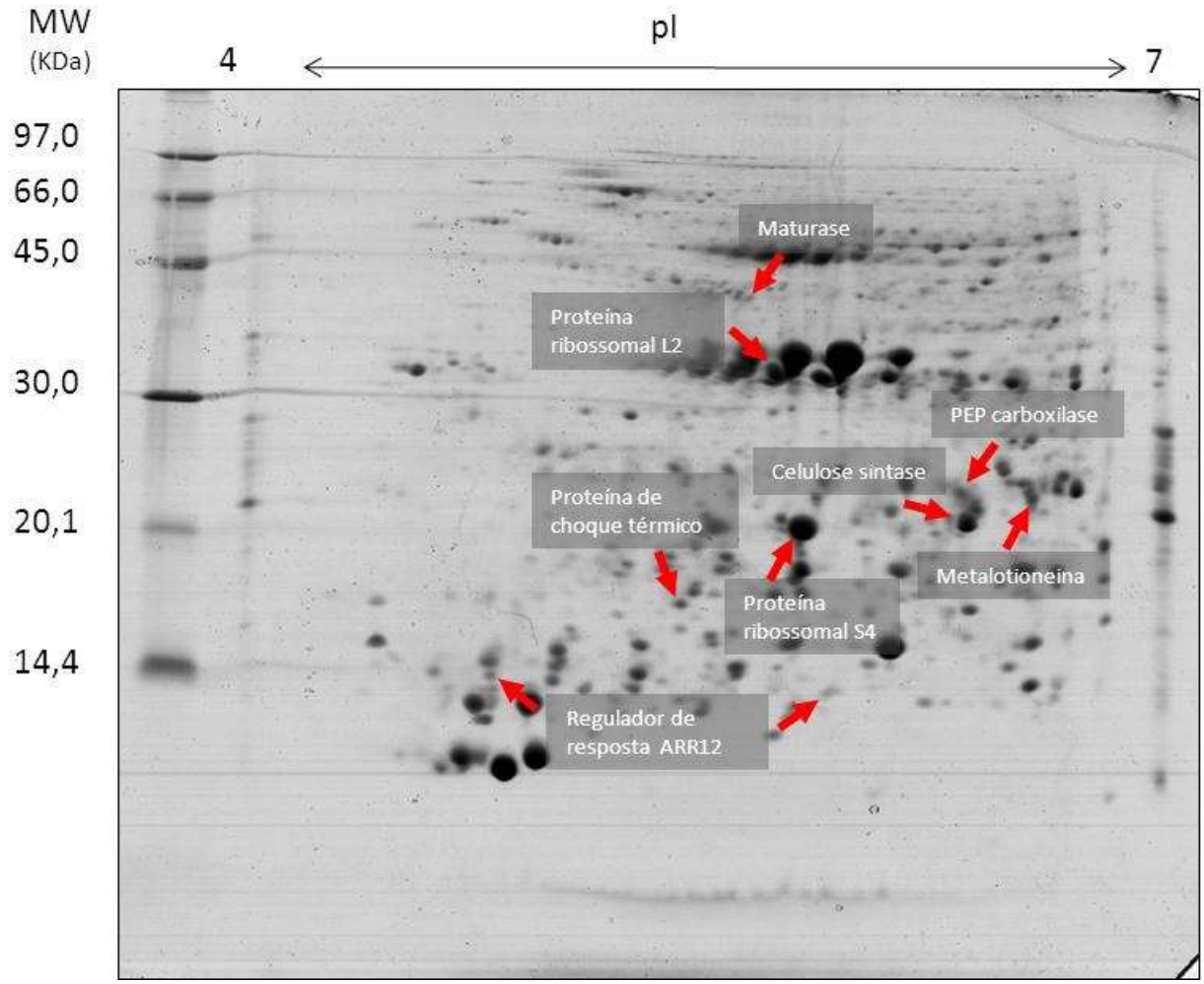
Fonte: O Autor.

Tabela 14 - Lista das proteínas de semente quiescente de cajueiro CCP 76 analisadas por MALDI-QUAD-TOF e identificadas por PMF.

ID	Num. acesso	Protein information	Valor E	Cobertura (%)	pI	Massa (KDa)
53	gi|357139437	N-acetiltransferase não caracterizada tipo ycf52	0.45	12	10.6	22.43
78	gi|21388458	Maturase	0.50	28	9.7	9.44
183	gi | 81158942	Carboxilase fosfoenolpiruvato	0.47	36	5.6	6.81
201	gi|114509162	Celulose sintase- semelhante a D4	0.84	5	5.8	131.30
54	gi | 350539597	Proteína EIL2	0.76	6	5.3	69.69
109	gi|239786301	Proteína ribossomal L2	0.36	18	11.7	11.81
184	gi | 110559312	Metalotioneína	0.15	55	4.7	7.09
202	gi|372450304	Produto do gene rps4 (mitocôndria)	0.13	15	11.3	43.71
267	gi|321439673	CYCLOIDEA 2	0.76	10	10.1	21.93
292	gi|357462821	Regulador de resposta duplo componente ARR12	0.76	9	4.5	9.34
301	gi|357462821	Regulador de resposta duplo componente ARR12	0.48	9	4.5	9.34
315	gi|224140593	Provável citocromo P450 campestanol to 6-Deoxocatasterona ou 6-oxocampestanol catasterona	0.29	13	7.0	10.10

Foi utilizado o programa ProFound para as identificações por PMF contra o banco do NCBInr.
Fonte: O Autor.

Figura 27 – Localização das proteínas identificadas no gel bidimensional de referência (CCP 76) utilizando espectrometria de massa.



MW – Marcador de peso molecular. **Fonte**: O Autor.

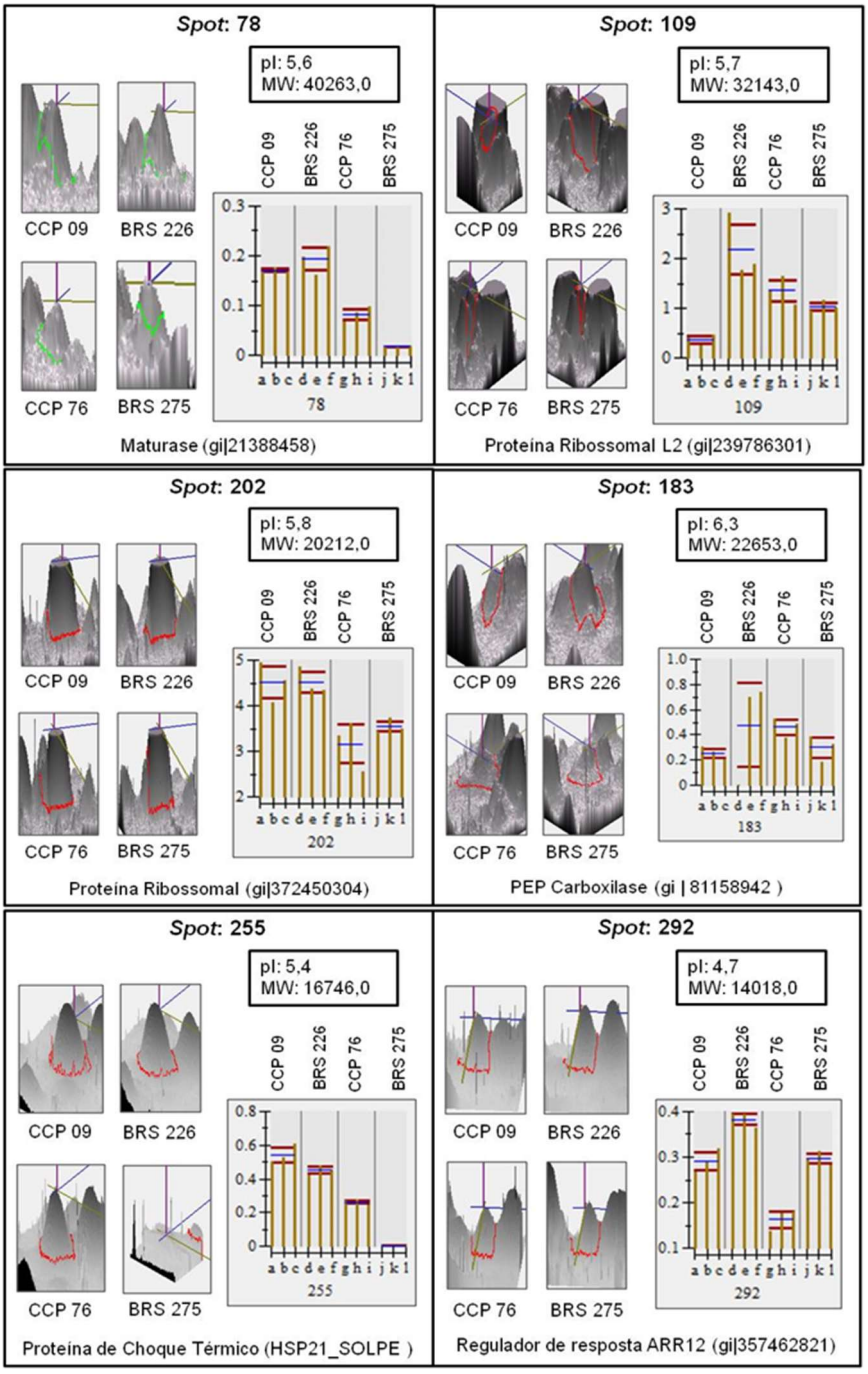
6.5.4 Proteínas diferencialmente expressas

Como explicado anteriormente, o objetivo da identificação de proteínas por MS foi avaliar o perfil protéico da amostra (*mining proteome*) e identificar as proteínas diferencialmente expressas (*protein expression profiling*). Foram encontradas seis proteínas candidatas a marcadores moleculares das variedades de cajueiro por terem valores de expressão diferenciadas.

A maturase (*spot* 78) e a proteína de choque térmico (*spot* 255) foram encontradas em baixos níveis no cajueiro BRS 275 em relação às outras variedades (Figura 28).

A proteína ribossomal L2 (*spot* 109) e a PEP carboxilase (*spot* 183) tiveram os menores níveis no cajueiro CCP 09. O regulador de resposta ARR12 (*spot* 292) teve os menores níveis de expressão no cajueiro CCP 76. Considerando diferenças de expressão valores de intensidade acima de duas vezes, os *spots* diferencialmente expressos reduzem para três (*spots* 78, 255 e 292) (Figura 28).

Figura 28 – Proteínas diferencialmente expressas nos géis bidimensionais de proteínas de sementes de quatro variedades de cajueiro separados na faixa de pH de 4-7 e identificados por espectrometria de massa.



O painel está dividido em seis quadros contendo informações sobre os *spots* diferencialmente expressos. Acima o número do *spot* com os valores de ponto isoelétrico (pI) e massa molecular (MW); o histograma à direita mostra os valores de porcentagem do volume dos *spots* nas replicatas dos géis do cajueiro: CCP 09 (a, b, c), BRS 226 (d, e, f), CCP 76 (g, h, i), BRS 275 (j, k, l). A linha amarela vertical representa a porcentagem do volume do *spot*, a linha horizontal azul é a média e as linhas horizontais vermelhas representam o desvio padrão; as imagens à esquerda representam visualizações tridimensionais dos *spots* presentes nos géis de referencia do cajueiro CCP 09, BRS 226, CCP 76 e BRS 275. Os dados foram obtidos pelo programa *ImageMaster Platinum* v. 6.0. **Fonte**: O Autor.

6.6 DISCUSSÃO

Vários trabalhos têm utilizado as ferramentas da proteômica com o objetivo de comparação entre genótipos e busca por marcadores moleculares. Trabalhos envolvendo proteômica de soja visando à comparação entre genótipos mostraram géis bidimensionais bastante semelhantes na faixa de pH de 3-10, mas as variações no número e intensidade dos *spots* foram melhor resolvidas na faixa de pH de 4-7 e 6-11 (NATARAJAN et al., 2006). Os géis bidimensionais das quatro variedades de cajueiro obtidas foram muito semelhantes mesmo utilizando uma estreita faixa de pH de 4-7.

Para uma avaliação global dos géis bidimensionais do cajueiro, incluindo as replicatas, foi feito o agrupamento heurístico (*heuristic clustering*) o qual mostrou que os géis foram agrupados em quatro classes automaticamente onde estas quatro classes realmente representam as quatro variedades de géis estudados. Também foi feito uma análise fatorial (*factor analysis*) para avaliar o grau de semelhança entre os géis, onde se observou que o BRS 275 é mais próximo do CCP 76 em relação aos outros genótipos.

O número de proteínas identificadas por espectrometria de massa geralmente é baixo (5 a 50 %), mesmo com espectros de qualidade (JOHNSON et al., 2005). O número de *spots* dos géis bidimensionais de sementes das quatro variedades de cajueiro variou de 284 a 377. Destas, apenas 6 % foram identificadas por MS/MS e 50 % foi identificada por PMF. O baixo número de identificações também pode ser resultado de pequena quantidade de dados moleculares em bancos de dados públicos. A identificação de proteínas de soja (*Glycine max*) por PMF, por exemplo, foi grandemente auxiliada por sequencias EST (MOONEY; THELEN, 2004).

A análise do proteoma de sementes de variedades de lentilha (*Lens culinaris*), por eletroforese bidimensional identificou centenas de proteínas, e somente 122 foram identificadas por MS, devido a pouca informação no genoma. Destas, após analise estatística multivariada mostrou que 24 proteínas foram diferencialmente expressas essenciais para a discriminação das populações. Entre as proteínas destacam-se a maturase-k e as globulinas (7S e 11S) (SCIPPA et al., 2010). Encontramos seis proteínas diferencialmente expressas nas quatro variedades de cajueiro e uma delas, foi identificada por espectrometria de massa como maturase.

6.7 CONCLUSÃO

Pode-se concluir que, embora os géis bidimensionais de sementes de quatro variedades de cajueiro sejam muito semelhantes, encontramos seis proteínas com níveis de expressão diferenciadas que podem ser utilizadas como marcadores moleculares.

SÍNTESE DE RESULTADOS E CONSIDERAÇÕES FINAIS

O sequenciamento do RNA (RNA-Seq) do transcriptoma de castanhas de cajueiro comum utilizando a plataforma *Illumina* revelou a presença de 16.347.083 *reads* de alta qualidade. A montagem *De* novo utilizando o *Velvet/Oases* mostrou a presença de 37.442 transcritos com uma cobertura média de 4 x e N50 igual a 171.

Com relação ao cajueiro anão CCP 76, o sequenciamento revelou a presença de 28.085.118 *reads* de alta qualidade. A montagem *De* novo utilizando o *Velvet/Oases* mostrou a presença de 77.371 transcritos com uma cobertura média de 4 x e N50 igual a 228.

A anotação funcional do transcriptoma do cajueiro comum mostrou que 17.205 (46%) dos transcritos foram identificados pelo BLAST contra o banco de dados do *Swiss-Prot*. Destas, 7.915 (46%) foram assinadas com termos de GO e 1.989 (11,5%) de termos de KO.

Com relação ao cajueiro anão CCP 76, a anotação funcional mostrou que 11.795 (15,2%) dos transcritos foram identificados pelo BLAST contra o banco de dados do *Swiss-Prot*. Destas, 5.186 (44%) foram assinadas com termos de GO e 1.594 (13,5%) de termos de KO.

Os mapas de géis bidimensionais constatam a presença de proteínas mais ou menos abundantes em sementes quiescentes de cajueiro comum e CCP 76.

Pode-se concluir que a análise do transcriptoma e proteômica do cajueiro comum e anão CCP 76 revelaram a presença de milhares de genes expressos durante a maturação, mas apenas algumas dezenas são altamente expressas em sementes quiescentes. Também foram encontradas pequenas diferenças quantitativas e qualitativas nas proteínas presentes nos diferentes tipos de cajueiro podendo ser candidatos a biomarcadores.

BIBLIOGRAFIA

ADAMS, M. D. et al. Complementary DNA sequencing: expressed sequence tags and human genome project. **Science**, v. 252, p. 1651-1656, 1991.

ADECE. Agronegócio. **Agencia de Desenvolvimento do Estado do Ceará**, 2012. Disponivel em: <http://www.adece.ce.gov.br/index.php/br/agronegocio>. Acesso em: 30 mar. 2012.

ALTSCHUL, S. F. et al. Basic local alignment search tool. **J. Mol. Biol.**, v. 215, p. 403-410, 1990.

AMARAL, A. M. et al. Plataformas tecnológicas no estudo da bactéria causadora do cancro cítrico: Genômica, transcriptômica e proteômica. **Melhoramento e Biotecnologia**, v. 27, p. 355-372, 2006.

AMIRUDDIN, N. et al. Characterisation of full-length cDNA sequences provides insights into the *Eimeria tenella* transcriptome. **BMC Genomics**, v. 13, p. 21-31, 2012.

ANDERSON, N. L.; ANDERSON, N. G. Proteome and proteomics: new technologies, new concepts, and new words. **Electrophoresis**, v. 19, p. 1853-1861, 1998.

ANDREOTE, F. D. **Análises genômica e transcriptômica de Methylobacterium mesophilicum SR1.6/6 em interação com a planta hospedeira**. Piracicaba: Escola Superior de Agricultura Luiz de Queiroz, 2011.

ANSORGE, W. J. Next-generation DNA sequencing techniques. **New biotechnology**, v. 25, p. 195-203, 2009.

ASHBURNER, M. et al. Gene Ontology: tool for the unification of biology. **Nat Genet**, v. 25, p. 25-29, 2000.

BALBUENA, T. S. **Proteômica do desenvolvimento da semente de Araucaria angustifolia**. São Paulo: Universidade de São Paulo, 2009. Tese de Doutorado.

BARROS, L. M. et al. **Recomendações técnicas para a cultura do cajueiro anão-precoce**. Fortaleza: Embrapa, 1993.

BIRON, D. G. et al. The pitfalls of proteomics experiments without the correct use of bioinformatics tools. **Proteomics**, v. 6, p. 5577–5596, 2006.

BLAIKIE, S. J.; CHACKO, E. K. Sap flow, leaf gas exchange and chlorophyll fluorescence of container-grown cashew (*Anacardium occidentale* L.) trees subjected to repeated cycles of soil drying. **Australian Journal of Experimental Agriculture,** v. 38, p. 305 - 311, 1998.

BLEICHER, E. et al. Minimal effective dosis of dimethoate and metamidophos for the control of the cashew inflorescence aphid. **Revista Brasileira de Fruticultura**, v. 19, p. 145-148, 1997.

BLEICHER, E. et al. Aspectos da biologia de *Anthistarcha binocularis meyrick* em inflorescência de cajueiro. **Pesquisa Agropecuária Tropical**, v. 37, 2007.

BLEICHER, E. et al. Minimal effective dose of phosphine to control the cashew root borer, *Marshallius bondari* Rosado-Neto (Coleoptera: Curculionidae). **Revista Ciência Agronômica**, v. 41, 2010.

BLEICHER, E.; ABREU, A. R. M.; MELO, Q. M. S. **Influência da fase de maturação da castanha na infestação da traça**. Fortaleza: Embrapa-CNPAT, 1995.

BRADFORD, M. M. Rapid and Sensitive Method for Quantitation of Microgram Quantities of Protein Utilizing Principle of Protein-dye Binding. **Analytical Biochemistry**, 1976. 248-254.

BREITENEDER, H.; RADAUER, C. A classification of plant food allergens. **J. Allergy Clin. Immunol.**, v. 113, p. 821-830, 2004.

CANDIANO, G. et al. Blue silver: a very sensitive colloidal comassie G-250 staining for proteome analysis. **Electrophoresis**, p. 1327-1333, 2004.

CANOVAS, F. M. et al. Plant proteome analysis. **Proteomics**, v. 4, p. 285-298, 2004.

CARVALHO, N. M.; NAKAGAWA, J. **Sementes:** Ciência, tecnologia e produção. Jaboticabal: FUNEP, 2000.

CAVALCANTI, J. J. V.; WILKINSON, M. J. The first genetic maps of cashew (*Anacardium occidentale* L.). **Euphytica**, v. 157, p. 131–143, 2007.

CHO, Y. et al. Development and characterization of twenty-five new polymorphic microsatellite markers in proso millet (*Panicum miliaceum* L.). **Genes & Genomics**, v. 32, p. 267-273, 2010.

CONESA, A.; GÖTZ, S. Blast2GO: A Comprehensive Suite for Functional Analysis in Plant Genomics. **International Journal of Plant Genomics**, p. 1-12, 2008.

DAMERVAL, C. et al. Technical improvements in two-dimensional electrophoresis increase the level of genetic variation detected in wheat-seedling protein. **Electrophoresis**, p. 53-54, 1986.

DIAS, L. L. C. et al. Proteômica comparativa aplicada à cultura de tecidos de plantas. **Revista Brasileira de Horticultura Ornamental**, v. 13, p. 2002-2008, 2007.

DWOMOH, E. A.; ACKONOR, J. B.; AFUN, J. V. K. Survey of insect species associated with cashew (*Anacardium occidentale* Linn.) and their distribution in Ghana. **African Journal of Agricultural Research**, v. 3, p. 205-214, 2008.

EDWARDS, K. J. et al. Microsatellite libraries enriched for several microsatellite sequences in plants. **BioTechniques**, v. 20, p. 758-760, 1996.

EISEN, M. B.; BROWN, P. O. DNA arrays for analysis of gene expression. **Methods of Enzymology**, v. 303, p. 179-205, 1999.

FERREIRA, A. G.; BORGHETTI, F. **Germinação:** do básico ao aplicado. Porto Alegre: Artmed, 2004.

FERREIRA, M. E.; GATTAPLAGIA, D. **Introdução ao uso de marcadores moleculares em análises genética**. [S.l.]: Embrapa Cenargen, 1998.

FETUGA, B.; BABATUNDE, G.; OYENUGA, V. Composition and nutritive value of cashew nut to the rat. **J. Agric. Food Chem.**, v. 22, p. 678-682, 1974.

FIGUEIRÊDO, L. C. et al. Detection of isometric, dsRNA-containing viral particles in *Colletotrichum gloeosporioides* isolated from cashew tree. **Tropical Plant Pathology**, v. 37, 2012.

GALLARDO, K. C. et al. Proteomic analysis of Arabidopsis seed germination and priming. **Plant Physiology**, v. 126, p. 835-848, 2001.

GALLARDO, K. et al. Proteomics of *Medicago truncatulla* seed development establishes the time frame of diverse metabolic precesses related to reserve accumulation. **Plant Physiology**, v. 133, p. 664-682, 2003.

GARG, R. et al. De novo assembly of chickpea transcriptome using short reads for gene discovery and marker identification. **DNA Research**, v. 18, p. 53-63, 2011.

GEORGI, L. et al. Cranberry microsatellite marker development from assembled next-generation genomic sequence. **Mol Breeding**, v. 30, p. 227–237, 2012.

GOLDSTEIN, D.; SCHLÖTTERER, C. **Microsatellites:** Evolution and applications. Oxford: Oxford University Press, 1999.

GÖRK, A. et al. **Two-Dimensional Electrophoresis with Immobilized pH Gradients for Proteome Analysis**. Munich: Technical University of Munich, 2007.

GRODZICKER, T. et al. Physical mapping of temperature sensitive mutations. **Cold Spring Harbor Symp.Quart.Biol**., v. 39, p. 439-446, 1974.

GYGI, S. P. et al. Correlation between Protein and mRNA Abundance in Yeast. **Mol. Cell. Biol.**, v. 3, p. 1720-1730, 1999.

HAJDUCH, M. et al. Proteomic analysis of near-isogenic sunflower varieties differing in seed oil traits. **Journal of Proteome Research**, v. 6, p. 3232-3241, 2007.

HAMADA, H. M.; PETRINO, M. G.; KAKUNAGA, T. A novel repeated element with Z-DNAforming potential is widely found in evolutionarily diverse eukaryotic genomes. **Natl. Acad. Sci**, v. 79, p. 6465–6469, 1982.

HELLMAN, U. et al. Improvement of an "In-Gel" digestion procedure for the micropreparation of internal protein fragments for amino acid sequencing. **Analytical Biochemistry**, v. 224, p. 451-455, 1995.

HIMEJIMA, M.; KUBO, I. Antibacterial agents from the cashew *Anacardium occidentale* (Anacardiaceae) nut shell oil. **J. Agric Food Chem**, v. 39, p. 418-421, 1991.

HINTON, J. C. et al. Benefits and pitfalls of using microarrays to monitor bacterial gene expression during infection. **Current Opinion in Microbiology**, London, v. 7, p. 277-

282, 2004.

HUANG, X.; MADAN, A. CAP3: A DNA Sequence Assembly Program. **Genome Research**, v. 9, p. 868-877, 1999.

JEFFREYS, A.; WILSON, V.; THEIN, S. Hypervariable 'minisatellite' regions in human DNA. **Nature**, v. 314, p. 67–73, 1985.

JOHNSON, R. S. et al. Informatics for protein identification by mass spectrometry. **Methods**, v. 35, p. 223-236, 2005.

JORGE, E. C. **Identificação de sequencias expressas (EST) de embriões de Gallus gallus**. Piracicaba: Escola Superior de Agronomia Luiz de Queiroz, 2002.

KALIA, R. K. et al. Microssatellite markers: no overview of the recent progress in plants. **Euphytica**, v. 177, p. 309-334, 2011.

KANEHISA, M. et al. From genomics to chemical genomics: new developments in KEGG. **Nucleic Acids Research**, v. 34, p. D354-D357, 2006.

KANNAMKUMARATH, S. S. et al. HPLC-ICP-MS determination of selenium distribution and speciation in different types of nut. **Anal Bioanal. Chem**, v. 373, p. 454-460, 2002.

KANTETY, R. V. et al. Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. **Plant Mol Biol.**, v. 48, p. 501–510, 2002.

KUBO, I. et al. Prostaglandin synthetase inhibitors from an african medicinal plant Ozoroa mucronata. **Chem. Lett.**, p. 1101-1104, 1987. ISSN 13.

KUBO, I. et al. Antitumor Agents from the Cashew (*Anacardium occidentale*) Apple Juice. **J. Agric. Food Chem.**, v. 41, p. 1012-1015, 1993.

KUBO, I.; KOMATSU, S.; OCHI, M. Molluscicides from the cashew *Anacardium occidentale* and their large-scale isolation. **J. Agric. Food Chem**, p. 970-973, 1986. ISSN 34.

KUBO, J.; LEE, R. J. Anti-Helicobacter pylori agents from the cashew apple. **J. Agric.**

Food Chem, v. 47, p. 533-537, 1999.

LAEMMLI, U. K. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. **Nature**, v. 227, p. 680-685, 1970.

LI, Y. C. et al. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. **Molecular Ecology**, v. 11, p. 2453-2465, 2002.

LI, Y. C. et al. Microsatellites within genes: structure, function, and evolution. **Mol Biol**, v. 21, p. 991–1007, 2004.

LIEBLER, D. C. **Introduction to proteomics:** tools for the new biology. Totowa: Humana Press, 2002.

LIMA, V. P. M. S. A cultura do cajueiro no nordeste do Brasil. In: BARROS, L. M. **Melhoramento**. Fortaleza: Etene, 1988. p. 321-356.

LIU, Z. et al. Transcribed dinucleotide microsatellites and their associated genes from channel catfish Ictalurus punctatus. **and Biophysical Research Communication**, v. 259, p. 190-194, 1999.

LUZ, C. L. S. **Anacardiaceae R. Br. na flora fanerogâmica do estado de São Paulo**. São Paulo: USP, 2011.

MAGNI, C. et al. Two-dimentional electrophoresis and western-blotting analysis with Ara h3 basic subunit IgG evidence the cross-reacting polypeptides of Arachis hypogaea, Glycine max, and Lupinus albus seed proteomes. **Journal of Agricultural and Food Chemistry**, v. 53, p. 2275-2281, 2005.

MARACAJA, P. B. et al. Revitalização de pomares de cajueiro e o controle da mosca branca na Serra de Santana – RN – Brasil: Uma intervenção da extensão rural. **Informativo técnico do semi-árido**, v. 2, 2008.

MARTIN, J. A.; WANG, Z. Next-generation transcriptome assembly. **Nature Reviews**, v. 12, p. 671-682, 2011.

MATTHIESEN, R. **Mass spectrometry data analysis in proteomics**. Totowa: Humana Press, 2007.

MESQUITA, A. L. M.; BECKER, V. O.; SOBRINHO, R. B. Taxonomic identification of lepidopterous species of cashew plant in Brazil. **Anais da Sociedade Entomológica do Brasil**, v. 27, 1998.

MIZRACHI, E. et al. De novo assembled expressed gene catalog of a fast-growing Eucalyptus tree produced by Illumina mRNA-Seq. **BMC Genomics**, v. 11, p. 1-12, 2010.

MOHOD, A. G.; KHANDETOD, Y. P.; POWAR, A. G. Processed cashew shell waste as fuel supplement for heat generation. **Energy for Sustainable Development**, v. 12, p. 73-76, 2008. ISSN 4.

MOONEY, B. P.; THELEN, J. J. High-throughput peptide mass fingerprinting of soybean seed proteins: automated workflow and utility of UniGene expressed sequence tag databases for protein identification. **Phytochemistry**, v. 65, p. 1733–1744, 2004.

MORAES, F. M. S. **Análise proteômica da embriogênese somática e da aquisição de competência embriogênica de *Ocotea catharinensis* Mez. (Lauraceae)**. Brasília: Universidade de Brasília, 2006. Dissertação de Mestrado.

MORTAZAVI, A. et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq. **Nature**, v. 5, p. 621-628, 2008.

MUNIZ, C. R. et al. Colonization of cashew plants by *Lasiodiplodia theobromae*: microscopical features. **Micron**, v. 42, p. 419–428, 2011.

MUÑOZ-MÉRIDA, A. et al. De novo assembly and functional annotation of the olive (*Olea europaea*) transcriptome. **DNA Research**, v. 7, p. 1-16, 2013.

NATARAJAN, S. et al. Proteomic and genetic analysis of glycinin subunits of sixteen soybeans genotypes. **Plant Physiol. and Biochem.**, v. 45, p. 436-444, 2007.

NATARAJAN, S. S. et al. Characterization of storage proteins in wild (*Glicine soja*) and cultivated (*Glycine max*) soybeans seeds using proteomic analysis. **J. Agric. Food Chem.**, v. 54, p. 3114-3120, 2006.

NOGUEIRA, F. C. S. **Análise Proteômica da Deposição de Proteínas em**

Sementes em Desenvolvimento e Suspensões Celulares Embriogênicas de Feijão-de-Corda [*Vigna unguiculata* (L.) Walp.]. Fortaleza: UFC - Dissertação de Mestrado, 2007.

O´FARREL, P. H. High resolution two-dimensional electrophoresis. **The Journal of Biological Chemistry**, v. 250, p. 4007-4021, 1975.

OGUNWOLU, S. O. et al. Functional properties of protein concentrates and isolates produced from cashew (*Anacardium occidentale* L.) nut. **Food Chemistry**, v. 115, p. 852-858, 2009.

OSBORNE, T. B. **The vegetable protein**. London: Longmans, Green and Co., 1924.

OSTERGAARD, O. et al. Proteome analysis of barley seeds: identification of major proteins from two-dimensional gels (pI 4-7). **Proteomics**, v. 4, p. 2437-2447, 2004.

OUYANG, J. et al. Development and characterization of 18 EST-SSR markers in *Sonneratia caseolaris*. **American Journal of Botany**, v. 98, p. 78-80, 2011.

PAIVA, J. R. D.; BARROS, L. D. M. **Clones de cajueiro:** obtenção, características e perspectivas. Fortaleza: Embrapa Agroindústria Tropical, 2004.

PAIVA, J. R. et al. Desempenho de clones de cajueiro-anão-precoce no semi-árido do Estado do Piauí. **Rev. Ciên. Agron.**, v. 39, p. 295-300, 2008.

PAIVA, J. R.; CRISÓSTOMO, J. R.; BARROS, L. M. **Recursos genéticos do cajueiro:** coleta, caracterização e utilização. Fortaleza: Embrapa, 2003.

PARIDA, S. K. et al. Informative genomic microsatellite markers for efficient genotyping applications in sugarcane. **Theor. Appl. Genet.** , v. 118, p. 327–338, 2009.

PARK, K. O. Proteomics studies in plants. **Journal of Biotechnology and Molecular Biology**, v. 37, p. 133-138, 2004.

PATERSON, A. H.; TANKSLEY, S. D.; SORRELLS, M. E. DNA markers in plant improvement. **Advances in Agronomy**, v. 46, p. 39-90, 1991.

PATERSON, S. D.; AEBERSOLD, R. H. Proteomics: the first decade and beyond. **Nature**, v. 33, p. 311-323, 2003.

PENG, R. K.; CHRISTIAN, K. The weaver ant, *Oecophylla smaragdina* (Hymenoptera: Formicidae), an effective biological control agent of the red-banded thrips, *Selenothrips rubrocinctus* (Thysanoptera: Thripidae) in mango crops in the Northern Territory of Australia. **International Journal of Pest Management**, v. 50, p. 107-114, 2004.

PEREIRA, J. M. et al. Anacardic acid derivatives as inhibitors of glyceraldehyde-3-phosphate dehydrogenase from *Trypanosoma Cruzi*. **Bioorganic & Medicinal Chemistry**, v. 16, p. 8889–8895, 2008.

PERKINS, D. N. et al. Probability-based protein identification by searching sequence databases using mass spectrometry data. **Electrophoresis**, p. 3551–3567, 1999. ISSN 20.

PRINGLE, S. D. et al. An investigation of the mobility separation of some peptide and protein ions using a new hybrid quadrupole/travelling wave IMS/oa-ToF instrument. **International Journal of Mass Spectrometry**, v. 261, p. 1–12, 2007.

RAO, V. N. M.; HASSAN, M. V. Preliminary studies on the floral biology of cashew (*Anacardium occidentale* L.). **Indian J. Agric. Sci.**, v. 27, p. 277-287, 1957.

RIGHETTI, P. G. et al. Quantitative proteomics: a review of different methodologies. **Eur. Mass Spectrom.**, v. 10, p. 335-348, 2004.

RISMANI-YAZDI, H.; HAZNEDAROGLU, B. Z.; PECCIA, J. Transcriptome sequencing and annotation of the microalgae *Dunaliella tertiolecta*: Pathway description and gene discovery for production of next-generation biofuels. **BioMed Central Genomics**, v. 12, p. 1-17, 2011.

ROBOTHAM, J. M. et al. Ana o 3, an important cashew nut (*Anacardium occidentale* L.) allergen of the 2S albumin family. **J. Allergy Clin. Immunol.**, v. 115, p. 1284-1290, 2005.

ROMANO, E. Extração de DNA de tecidos vegetais. In: BRASILEIRO, A. C. M.; CARNEIRO, V. T. C. **Manual de transformação genética de plantas**. Brasilia: Embrapa, 1998. p. 163-177.

RUSK, N. Torrents of sequence. **Nature Methods**, v. 8, p. 44-44, 2011.

SAMBROOK, J.; FRITSCH, E. F.; MANIATIS. **Molecular cloning. A laboratory manual**. Second edition. ed. [S.l.]: Cold Spring Harbor Laboratory Press, 1989.

SANSALONI, C. P. **Desenvolvimento, caracterização e mapeamento de microssatélites de Tetra e Pentanucleotídeos**. [S.l.]: [s.n.], 2008.

SATHE, S. K. et al. Biochemical characterization and in vitro digestibility of the major globulin in cashew nut (*Anacardium occidentale*). **J. Agric. Food Chem.**, v. 45, p. 2854-2860, 1997.

SCHMIDT, H. et al. 2-D DIGE analysis of the proteome of extracts from peanut variants reveals striking differences in major allergen contents. **Proteomics**, v. 9, p. 3507–3521, 2009.

SCIPPA, G. S. et al. The proteome of lentil (*Lens culinaris* Medik.) seeds: Discriminating between landraces. **Electrophoresis**, v. 31, p. 497–506, 2010.

SEHGAL, D.; RAINA, S. N. DNA markers and germplasm resource diagnostics: new perspectives in crop improvement and conservation strategies. In: ARYA, I. D.; ARYA, S. **Utilization of biotechnology in plant sciences**. Dehradun: [s.n.], 2008. p. 39–54.

SHENDURE, J.; JI, H. Next-generation DNA sequencing. **Nature biotechnology**, v. 26, p. 1135-45, 2008.

SHEWRY, P. R.; NAPIER, J. A.; TATHAN, A. S. Seed storage proteins: structures and biosynthesis. **The Plant Cell**, v. 7, p. 945-956, 1995.

SILVEIRA, L. **Montagem e anotação parcial da sequência genômica da bactéria diazotrófica *Azospirillum brasiliense* FP2**. Curitiba: Universidade Federal do Paraná, 2012.

STASIUK, M.; KOZUBEK, A. Biological activity of phenolic lipids. **Cell. Mol. Life Sci.**, v. 67, p. 841-860, 2010.

STEIN, L. Genome annotation: from sequence to biology. **Nature**, v. 4, p. 493-503, 2001.

SUN, C. et al. De novo sequencing and analysis of the american ginseng root transcriptome using a GS FLX titanium platform to discover putative genes involved in

ginsenoside biosynthesis. **BMC Genomics**, v. 11, p. 1-12, 2010.

SURGET-GROBA, Y.; MONTOYA-BURGOS, J. I. Optimization of de novo transcriptome assembly from next-generation sequencing data. **Genome Research**, v. 20, p. 1432–1440, 2010.

TEUBER, S. S. et al. Characterization of the soluble allergenic proteins of cashew nut (*Anacardium occidentale* L.). **J. Agric. Food Chem.**, v. 50, p. 6543-6549, 2002.

THIEL, T. et al. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). **Theoretical and Applied Genetics**, v. 106, p. 411-422, 2003.

THIEL, T. et al. Exploiting EST databases for the development of cDNA derived microsatellite markers in barley (*Hordeum vulgare* L.). **Theor. Appl. Genet.** , v. 106, p. 411–42, 2003.

VARSHNEY, R. K. et al. Genic molecular markers in plants: development and applications. In: VARSHNEY, R. K.; TUBEROSA, R. **Genomics assisted crop improvement:** genomics approaches and. [S.l.]: Springer, 2007. p. 13–29.

VELCULESCU, V. E. et al. Serial analysis of gene expression. **Science**, v. 270, p. 368-369, 1995.

VIANA, F. M. P. et al. Control of cashew black mould by acibenzolar-S-methyl. **Tropical Plant Pathology**, v. 37, 2012.

VIEIRA, S. **Introdução a bioestatística**. 3. ed. [S.l.]: Campus, 1980.

VOIGT, E. L. et al. Source–sink regulation of cotyledonary reserve mobilization during cashew (*Anacardium occidentale*) seedling establishment under NaCl salinity. **Journal of Plant Physiology**, v. 166, p. 80-89, 2009.

VOS, P. et al. AFLP: a new technique for DNA fingerprinting. **Nucleic Acids Res.**, v. 21, 1995.

WANG, F. et al. Ana o 1, a cashew (*Anacardium occidental*) allergen of the vicilin seed storage protein family. **J. Allergy Clin. Immunol.**, v. 110, p. 160-166, 2002.

WANG, F. et al. Ana o 2 a Major Cashew (*Anacardium occidentale* L.) Nut Allergen of the Legumin Family. **Int. Arch. Allergy Immunol.**, v. 132, p. 27-29, 2003.

WANG, M. L.; BARKLEY, N. A.; JENKINS, T. M. Microsatellite markers in plants and insects. Part I. Applications of biotechnology. **Genes Genomes Genomics**, v. 3, p. 54–67, 2009.

WANG, Z. et al. De novo assembly and characterization of root transcriptome using Illumina paired-end sequencing and development of cSSR markers in sweetpotato (*Ipomoea batatas*). **BMC Genomics**, v. 11, p. 726-739, 2010.

WANG, Z.; GERSTEIN, M.; SNYDER, M. RNA-Seq: a revolutionary tool for transcriptomics. **NATURE REVIEwS**, v. 10, p. 57-63, 2009.

WEBER, J. L. Informativeness os human (dC-dA)n (dG-dT)n polymorphism. **Genomics**, v. 7, p. 524-530, 1990.

WILKINS, M. R. Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. **Biotech. Gen. Eng. Rev.**, v. 13, p. 19-50, 1995.

WILLIAMS, J. G. et al. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. **Nucleic Acids Res**, v. 18, p. 6531–6535, 1990.

WITTMANN-LIEBOLD, B.; GRAACK, H. R.; POHL, T. Two-dimensional gel electrophoresis as tool for proteomics studies in combination with protein identification by mass spectrometry. **Proteomics**, v. 6, p. 4688-4703, 2006.

YAN, Y. et al. Development and characterization of EST-SSR markers in the invasive weed *Mikania micrantha* (Asteraceae). **American Journal of Botany**, v. 98, p. 1-3, 2011.

YANG, P. F.; SHEN, S. H.; KUANG, T. Y. Comparative analysis of the endosperm proteins separated by 2-D electrophoresis for two cultivars of hybrid rice (*Oryza sativa*). **J. Integr. Plant Biol.**, v. 48, p. 1028-1033, 2006.

YU, J. N. et al. Fast and Cost-Effective Mining of Microsatellite Markers Using NGS Technology: An Example of a Korean Water Deer Hydropotes inermis argyropus.

PLoS ONE, v. 6, 2011.

ZABEAU, M.; VOS, P. **Selective restriction fragment amplification:** a general method for DNA fingerprinting. 0 534 858 A1, 1993.

ZALAPA, J. E. et al. Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. **American Journal of Botany**, v. 99, p. 193-208, 2012.

ZANE, L.; BARGELLONI, L.; PATARNELLO, T. Strategies for microsatellite isolation: a review. **Mol Ecol**, v. 11, p. 1–16, 2002.

ZERBINO, D. R. Using the Velvet de novo assembler for short-read sequencing technologies. **Curr Protoc Bioinformatics**, p. 1-13, 2010.

ZERBINO, D. R.; BIRNEY, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. **Genome Research**, v. 18, p. 821-829, 2008.

ZHAO, Y.; PRAKASH, C. S.; HE, G. Characterization and compilation of polymorphic simple sequence repeat (SSR) markers of peanut from public database. **BMC Research Notes**, v. 5, p. 362, 2012.

APENDICE

Tabela contendo dados dos microssatélites do cajueiro comum e anão CCP 76.

Di-nucleotídeos

Grupos de Motivos	Nome do SSR	Locus do Cajueiro Anão	Locus do Cajueiro Comum	Núm. de unid. de repetição
Grupo I (GC/CG)	Não houve detecção destes motivos nas bibliotecas analisadas.			
Grupo II (AC/CA/TG/GT)	SSR2AC01	Contig 39044	Contig 70	9
	SSR2TG01	Contig 7112	Contig 4170	10
Grupo III (AG/GA/TC/CT)	SSR2AG01	Contig 41841 Contig 11600	Contig 1094 Contig 1332	9
	SSR2AG02	Contig 3207	Contig 2627	9
	SSR2AG03	Contig 35731	Contig 7795	9
	SSR2AG04	Contig 11290	Contig 217	12
	SSR2AG05	Contig 38697	Contig 4291	16
	SSR2AG06	Contig 18060	Contig 908	15
	SSR2AG07	Contig 15123	Contig 3426	12
	SSR2GA02	Contig 1605	Contig 6204	16
	SSR2GA03	Contig 15381	Contig 14515	9
	SSR2GA04	Contig 39317	Contig 22393	15
	SSR2TC01	Contig 3741	Contig 4305	13
	SSR2TC02	Contig 5770	Contig 604	14
	SSR2CT01	Contig 20951	Contig 2557	11
Grupo IV (AT/TA)	SSR2TA01	Contig 41308	Contig 11909	9
	SSR2AT01	Contig 4414	Contig 1253	11

Tri-nucleotídeos

Grupos de Motivos	Nome do SSR	Locus do Cajueiro Anão	Locus do Cajueiro Comum	Núm. de unid. de repetição (CCP 76/Comum)
Grupo I (GGC/GCG/CGG/CCG/CGC/GCC)	Não houve detecção			
Grupo II (ACG/CGA/GAC/TGC/GCT/CTG)	SSR3CTG01	Contig 18491	Contig 5077	7
	SSR3CTG02	Contig 5076	Contig 3589	8/6
	SSR3GCT01	Contig 47737	Contig 3981	6
	SSR3GCT	Contig 6777	Contig 591	6

	02			
	SSR3TGC 01	Contig 7751	Contig 1855	6
	SSR3TGC 02	Contig 3978	Contig 16688	6
Grupo III (AGC/GCA/CAG/TC G/CGT/GTC)	SSR3AGC 01	Contig 7357	Contig 8039	7
	SSR3CAG 01	Contig 3400 Contig 9677	Contig 10080	6
	SSR3CAG 02	Contig 2587	Contig 15952	7
	SSR3CAG 03	Contig 34044	Contig 6401	7
	SSR3CAG 04	Contig 25240	Contig 7943	6
	SSR3CAG 05	Contig 44825	Contig 38037	6
	SSR3CAG 06	Contig 19617	Contig 6495	6
	SSR3GCA 01	Contig 29970	Contig 3255	6
	SSR3TCG 01	Contig 5901	Contig 1295	6
Grupo IV (ACC/CAC/CCA/TG G/GTG/GGT)	SSR3CAC 01	Contig 622	Contig 12001	6
	SSR3CAC 02	Contig 19578	Contig 12982	6
	SSR3GGT 01	Contig 6263	Contig 16880	6
	SSR3TGG 01	Contig 35043	Contig 11722	6
	SSR3TGG 02	Contig 8207	Contig 35499	6
	SSR3TGG 03	Contig 3821	Contig 6631	6
Grupo V (AGG/GGA/GAG/TC C/CCT/CTC)	SSR3AGG 01	Contig 13673	Contig 28410	8
	SSR3AGG 02	Contig 31939	Contig 26410	7
	SSR3AGG 03	Contig 4653	Contig 25854	9
	SSR3CTC 01	Contig 28503	Contig 11944	10
	SSR3CTC 02	Contig 3601	Contig 23686	8/6
	SSR3GGA	Contig	Contig 34086	7

	01	30307		
	SSR3TCC 01	Contig 22986	Contig 11981	6
	SSR3TCC 02	Contig 17742	Contig 18464	9
Grupo VI (AGT/GTA/TAG/ACT /CAT/ATC)	SSR3AGT 01	Contig 14673	Contig 29830	6
	SSR3CAT 01	Contig 28663	Contig 31043	6
	SSR3TCA 01	Contig 2093	Contig 4022	9
	SSR3TCA 02	Contig 5758	Contig 8215	6
	SSR3TCA 03	Contig 20655	Contig 16455	6
Grupo VII (ATG/TGA/GAT/TAC /ACT/CTA)	SSR3GAT 01	Contig 42660	Contig 1292	6
	SSR3GAT 02	Contig 2082	Contig 15987	7
	SSR3GAT 03	Contig 31382	Contig 17417	7
	SSR3GAT 04	Contig 10990	Contig 22781	7
	SSR3GAT 01	Contig 42660	Contig 1292	6
	SSR3GAT 02	Contig 2082	Contig 15987	7
	SSR3GAT 03	Contig 31382	Contig 17417	7
	SSR3GAT 04	Contig 10990	Contig 22781	7
	SSR3TGA 01	Contig 14532 Contig 175	Contig 1001	8
	SSR3TGA 02	Contig 38625	Contig 29864	7
	SSR3TGA 01	Contig 14532 Contig 175	Contig 1001	8
	SSR3TGA 02	Contig 38625	Contig 29864	7
Grupo VIII (AAG/AGA/GAA/CTT /TTC/TCT)	SSR3AAG 01	Contig 30890	Contig 1667	10
	SSR3AAG 02	Contig 10027	Contig 2600	6
	SSR3AAG	Contig	Contig 10953	9

	03	15720		
	SSR3AAG 04	Contig 15511	Contig 28097	7/10
	SSR3AGA 01	Contig 3009	Contig 17446 Contig 1923 Contig 2065	6/6
	SSR3AGA 02	Contig 7960	Contig 2189	11
	SSR3AGA 03	Contig 289 Contig 2455	Contig 1996	6
	SSR3CTT 01	Contig 11600	Contig 1094 Contig 1332	7
	SSR3CTT 02	Contig 23824	Contig 20005	7
	SSR3CTT 03	Contig 16399	Contig 27479	6
	SSR3CTT 04	Contig 16654	Contig 34467	7
	SSR3CTT 05	Contig 26303 Contig 2928	Contig 902	11
	SSR3GAA 01	Contig 16517	Contig 23078	6
	SSR3GAA 02	Contig 22670	Contig 11859	9
	SSR3GAA 03	Contig 15754	Contig 15750	8
	SSR3GAA 05	Contig 4264	Contig 4991	9
	SSR3TTC 01	Contig 14152	Contig 16997	6
	SSR3TTC 02	Contig 4920	Contig 3335	6
	SSR3TTC 03	Contig 5612	Contig 35592	6
	SSR3TTC 04	Contig 18679	Contig 32714	7
	SSR3TCT 02	Contig 36420	Contig 7278	6
	SSR3TCT 03	Contig 35718	Contig 11953	6
Grupo IX (AAC/ACA/CAA/TTG /TGT/GTT)	SSR3AAC 01	Contig 14120	Contig 294	11
	SSR3ACA 01	Contig 18127	Contig 1718	8
	SSR3ACA 02	Contig 9304	Contig 5431	8

	SSR3TGT 01	Contig 47946	Contig 5431	8
	SSR3TTG 01	Contig 744	Contig 28594	6
Grupo X (AAT/ATA/TAA/ATT/ TTA/TAT)	SSR3ATT 01	Contig 18524 Contig 40405	Contig 1265	8
	SSR3TTA 01	Contig 163	Contig 32232	6

Tetra-nucleotídeos

Grupos de Motivos	**Nome do SSR**	**Locus do Cajueiro Anão**	**Locus do Cajueiro Comum**	**Núm. de unid. de repetição**
Grupo I (AAAG/TTTC)	MIC4AAAG01	Contig 35145	Contig 10402	6
	MIC4TTTC01	Contig 3884	Contig 2335	6