UFAM

UNIVERSIDADE FEDERAL DO AMAZONAS

PROGRAMA DE PÓS GRADUAÇÃO EM BIOTECNOLOGIA

TRANSCRIPTOMA DE *Copaifera multijuga* HAYNE: MONTAGEM E ANOTAÇÃO

ELIANE CARVALHO DOS SANTOS

MANAUS, AM

2018

UFAM

UNIVERSIDADE FEDERAL DO AMAZONAS
PROGRAMA DE PÓS GRADUAÇÃO EM BIOTECNOLOGIA

TRANSCRIPTOMA DE *Copaifera multijuga* HAYNE: MONTAGEM E ANOTAÇÃO

ELIANE CARVALHO DOS SANTOS

Tese apresentada ao Programa de Pós-Graduação em Biotecnologia da Universidade Federal do Amazonas como parte dos requisitos exigidos para obtenção do título de Doutora em BIOTECNOLOGIA.

Orientador: Spartaco Astolfi Filho
Co-Orientador: Edmar Vaz de Andrade

MANAUS, AM
2018

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

Carvalho, Eliane
C331t Transcriptoma de Copaifera multijuga Hayne: Montagem e Anotação / Eliane Carvalho. 2018
97 f.: il. color; 31 cm.

Orientadora: Spartaco Astolfi Filho
Coorientador: Edmar Vaz de Andrade
Tese (Doutorado em Biotecnologia) - Universidade Federal do Amazonas.

1. copaíba. 2. óleo-resina. 3. Amazônia. 4. metabolismo. 5. terpeno. I. Astolfi Filho, Spartaco II. Universidade Federal do Amazonas III. Título

DEDICATÓRIA

Dedico este trabalho ao meu filho Ethan Gabriel Carvalho dos Santos meu principal motivo para seguir sempre em frente e minha querida avó Raimunda Helena Carvalho (*in memorian*) que me ensinou a sempre buscar meus objetivos e não desistir nunca.

AGRADECIMENTOS

Agradeço ao meu Deus por ter me permitido chegar até aqui e pela força que me deu para seguir sempre em frente, ultrapassando todos os obstáculos que surgiram ao longo do meu caminho.

Ao meu amado filho Ethan Gabriel Carvalho dos Santos, meu principal motivador nesta trajetória e que me só me dá felicidades todos os dias e com seu amor me fez ter mais coragem de seguir essa caminhada.

Ao meu marido e melhor amigo Cleyson Alex Maia dos Santos por ter sido minha rocha, segurado minha mão e me dado forças para não desistir, a força que Deus me deu veio a mim por meio de suas mãos e palavras em todo o meu caminho. Te amo!

Minha avó Helena Carvalho (*in memorian*) obrigada por seus conselhos e por ter me incentivado a nunca desistir dos meus sonhos, quando segurava minha mão e me encorajava a seguir este caminho e que certamente estaria muito orgulhosa de mim. Infelizmente o Senhor a levou exatamente no meio desta minha trajetória, meu coração ficou em pedaços, mas sei que você me deu forças para não desistir.

À minha mãe que seguiu junto comigo toda essa caminhada, muitas vezes sendo mãe no meu lugar, quando minhas tormentas diárias me deixavam ausentes de casa. Obrigada mãe era você que estava lá me apoiando e segurando minha mão!

A toda minha família, meus irmãos, Andreza, Alberto e Aline. Minhas tias Joveci, Leila e Dora que certamente me deram todo apoio emocional que precisei ao longo deste projeto. Muitas vezes viajando a Manaus só para me apoiar. Obrigada por tudo!

A grande professora Andréa Ghelfi, por ter me ensinado muito sobre a bioinformática e me ajudado da melhor maneira possível neste projeto, me ajudando com as análises iniciais. Ensinando-me tudo metodicamente e com muito amor a ciência. Obrigada porque você que me ajudou a desenvolver todo

este trabalho, me oferecendo todo o aparato necessário para o meu aprendizado. Andréa você é “perfect” obrigada por tudo!

A grande família UFAM por toda a caminhada e descontração do laboratório em muitos momentos quando aos finais de semana e feriados que estávamos na correria. As festinhas maravilhosas, as discussões de projetos, e todas as horas que passamos juntos. Obrigada Diego Moreira, Kerolen, Edson, Elen, Enedina, Júlio, Elson, Maria, Marcelo, Pamela, Anita, Isabelle, Lorena, Filipe, Genilton e Dona Elza. E a todos os demais colegas que estavam sempre juntos e misturados nos melhores e piores momentos.

Ao grande amigo Diego Moreira pela parceria, que esteve junto comigo em muitos momentos de domingo a domingo no laboratório, me ajudando, incentivando, comemorando os resultados positivos, chorando junto nos resultados negativos e discutindo ciência. Obrigada Dieguinho!

A minha grande amiga Suelen Dias, que mais uma vez esteve ao meu lado, nas conversas, nas discussões sobre meu trabalho, na companhia, na motivação. Obrigada amiga/irmã por sua presença constante em minha vida.

Ao meu maravilhoso orientador Professor Spartaco Astolfi Filho, por ter me aceitado neste projeto e ter acreditado e confiado em mim. Por ter estado ao meu lado me ensinando como o melhor orientador que alguém pode querer. Orientador que vai para bancada, que segura a mão, que apoia, que ensina o amor a ciência, que dá puxões de orelha necessários, que ouve que estimula que faz surgir a luz no fim do túnel, que enfim é um grande pai de todos. Professor é um privilégio e orgulho imenso fazer parte da sua equipe e por ter aprendido com você e me faltam palavras para agradecer. Levarei por toda minha vida esse aprendizado. Obrigada!

Ao professor Edmar Andrade por ter estado junto comigo desde o início deste trabalho, muitas vezes apoiando, dando ideias, estimulando meu aprendizado, trocando experiências e conhecimentos. Como orientador também buscou dar o seu melhor para me apoiar e mostrar o quão importante é seguir metodicamente todos os passos e critérios da pesquisa. Que me fez

crescer intelectualmente por incentivar a minha busca às suas muitas perguntas. Obrigada professor Edmar por todo o seu apoio nesta trajetória.

A Dra Tainá Raiol, por ter me ajudado nas análises deste projeto e dado todo o seu apoio em todos os momentos que precisei.

Ao Dr. Jorge Luis Lopes Lozano, amigo que por muitas vezes que me ouviu, deu conselhos e me ofereceu sua ajuda desde o início até a conclusão deste trabalho, dando dicas sábias e importantes para minha caminhada. Obrigada professor!

Ao Dr. Adolfo Mota, por toda a sua solicitude, tirando-me muitas dúvidas, obrigada por suas ideias, por me ensinar parte do conhecimento que adquiri ao longo deste projeto.

Ao programa do PPG-BIOTEC.

À Capes pela bolsa cedida.

“A menos que modifiquemos a nossa maneira de pensar, não seremos capazes de resolver os problemas causados pela forma como nos acostumamos a ver o mundo”.

(Albert Einstein)

RESUMO

Copaifera multijuga Hayne é uma espécie de planta de grande porte, popularmente conhecida como copaíba, nativa da América Latina e da África. São amplamente utilizadas na medicina popular amazônica, devido as propriedades do óleo-resina extraído do tronco de suas árvores, utilizado principalmente como: diurético, laxativo, antitetânico, anti-inflamatório, antitussígeno, cicatrizante e anti-tumoral. Sendo, portanto, de grande importância para pesquisas que visem identificar nas plantas substâncias com potenciais finalidades médicas e biotecnológicas. Por isto, este trabalho visou pesquisar o transcriptoma de *C. multijuga* Hayne, o qual foi sequênciado utilizando a plataforma 454 Roche, sendo obtido um total de **638.576** *reads*, montados através da metodologia *de novo* com auxílio das plataformas MIRA e TRINITY. Sendo gerados, a partir da montagem por TRINITY **53.319** *contigs* e **62.839** pela montagem MIRA. A anotação do transcriptoma foi realizada utilizando BLASTx (NCBI) e a anotação funcional através do banco Gene Ontology (GO). Os resultados obtidos foram categorizados de acordo com GO, onde os *unigene*s foram agrupados nas categorias: “Componente Celular” com **6.249** *contigs* envolvidos nesta categoria, “Função Molecular” com total **17.208** *contigs* e “Processo Biológico” com **9.499** *contigs.* Nas três categorias os *contigs* evidenciados estão envolvidos no metabolismo primário vegetal. Já do metabolismo secundário foram detectados **184** *unigenes*, sendo organizados em 22 *clusters*, envolvidos principalmente em respostas ao estresse oxidativo vegetal, compostos terpenos, importantes metabólitos envolvidos na formação do óleo-resina de copaíba, diterpenos (porção resinosa do óleo-resina) e sesquiterpenos (componentes voláteis do óleo) e *unigenes* relacionados com a pigmentação vegetal tais como: flavonóides, carotenóides e antocianinas. Na análise filogenética foram feitas comparações evolutivas de enzimas de terpenos sintases, componentes de formação do óleo de copaíba, com enzimas do banco de dados NCBI. O *unigene* de TPS4-2 de *C. multijuga* mostrou-se mais próximo a outras sequências de terpenos TPS4-2 de *Copaifera* disponíveis no banco. Com isto, neste trabalho realizou-se o transcriptoma de *C. multijuga* com a montagem *de novo* e anotação o que leva a perspectiva de estudos com os *unigenes* codificadores de enzimas

componentes do óleo-resina evidenciados neste trabalho, visando com isto futuras pesquisas para fins biotecnológicos.

Palavras-chave: copaíba, óleo-resina, Amazônia, metabolismo, terpeno, RNA-seq.

ABSTRACT

Copaifera multijuga Hayne is a large plant species, popularly known as copaiba, native to Latin America and Africa. They are widely used in Amazonian folk medicine, due to the properties of oleoresin extracted from the trunk of their trees, mainly used as: diuretic, laxative, anti-tetanic, anti-inflammatory, antitussive, healing and anti-tumor. Therefore, it is of great importance for research aimed at identifying in plants substances with potential medical and biotechnological purposes. Therefore, this work aimed to search for the transcriptome of *C. multijuga* Hayne, which was sequenced using the Roche 454 platform, obtaining a total of 638,576 reads, assembled using the *de novo* methodology using the MIRA and TRINITY platforms. Being generated, from the assembly by TRINITY **53.319 contigs** and **62.839** by assembly MIRA. The transcriptome annotation was performed using BLASTx (NCBI) and the functional annotation through the Gene Ontology (GO) bank. The results were categorized according to GO, where the *contig*s were grouped in the categories: "Cell Component" with **6.249 *contigs*** involved in this category, "Molecular Function" with a total of **17.208 *contigs*** and "Biological Process" with **9,499 *contigs***. In the three categories the *contig*s evidenced are involved in the primary plant metabolism. A total of **184** *unigenes* were detected in 22 clusters, mainly involved in responses to plant oxidative stress, terpenes, important metabolites involved in the formation of copaiba oleoresin, diterpenes (resinous portion of oleoresin) and sesquiterpenes (volatile components of oil) and *unigene*s related to vegetal pigmentation such as: flavonoids, carotenoids and anthocyanins. In phylogenetic analysis, evolutionary comparisons of terpene synthase enzymes, copaiba oil formation components, with enzymes from the NCBI database were made. *C. multijuga* TPS4-2 *unigene* had similarity to other Copaifera TPS4-2 terpene sequences available from the bank. Thus, in this work the transcriptome of *C. multijuga* was carried out with new assembly and annotation, which leads to the perspective of studies with the *unigene*s encoding enzymes components of the oleoresin evidenced in this work, aiming with this future research for biotechnological purposes.

Keywords: copaiba, oil-resin, Amazon, metabolism, terpene, RNA-seq.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

BLAST= Basic Local Alignment Search Tool

ESTs= *Expressed Sequence Tags*

GO= *Gene Ontology*

GPP= geranil difosfato

IPP= isopentenil pirofosfato

KEGG= Kyoto Encyclopedia of Genes and Genomes

MEP= metileritritol 4- phosphate

MEV= *mevalonic acid* ou mevalonato,

NCBI- National Center for Biotechnology Information

NGS= *New Generation Sequence* (sequenciamento de nova geração)

TPS- Terpeno sintase

TPS-4-2= Terpeno sintase 4-2.

Introdução

1. INTRODUÇÃO

As espécies de *Copaifera,* denominadas popularmente de copaíbas, são árvores de grande porte, podendo alcançar 40 metros de altura e 4 metros de diâmetro. As espécies deste gênero de plantas chegam a viver cerca de 400 anos, sendo nativas da região tropical da América Latina e também da África Ocidental. Na América Latina são encontradas espécies de copaíbas na região que se estende do México ao norte da Argentina (VEIGA-JÚNIOR e PINTO, 2002; PIERI *et al.,* 2009; HECK *et al.,* 2012).

As copaíbas são amplamente utilizadas na medicina popular, devido às propriedades etnofarmacológicas do óleo-resina, extraído do tronco de suas espécies, bem como pela utilização da madeira de boa qualidade indicada para construção civil (LEANDRO *et al.,* 2012; GONÇALVES *et al.,* 2014).

A espécie *C. multijuga* Hayne destaca-se como uma das espécies frequentemente utilizadas na produção do óleo de copaíba (VEIGA-JUNIOR e PINTO, 2002). O óleo de copaíba é formado a partir do metabolismo secundário vegetal, sendo um exudato constituído por ácidos resinosos e compostos voláteis, por isto é designado óleo-resina, caracterizado principalmente por ser um líquido transparente de forte odor, cuja coloração varia do amarelo para marrom claro, com exceção do óleo da espécie *C. langsdorfii*, que possui coloração avermelhada (VEIGA JUNIOR *et al.,* 2002, HECK *et al.,* 2012).

Devido as importantes propriedades reconhecidas do óleo de copaíba, no ano de 2011 foi realizado um trabalho de caracterização do óleo extraído de folhas de plantas jovens de *C. multijuga* Hayne e de sequenciamento do transcriptoma desse mesmo tipo de material (BASTOS, 2011), os dados gerados serviram para um trabalho inicial de anotação e também para o trabalho que será apresentado nessa tese.

Por isso a partir dos dados brutos obtidos a partir deste trabalho, o principal objetivo desta tese foi fazer a montagem *de novo* e anotação do transcriptoma de *C. multijuga*, visando buscar vias tanto do metabolismo primário como também do metabolismo secundário da espécie, com ênfase principal ao metabolismo de terpenos formadores do óleo-resina de copaíba. Sabe-se que a compreensão destas rotas é de grande importância não só para o conhecimento da biologia da espécie *C. multijuga* Hayne, como também para o rastreio e conhecimento de genes da família terpeno sintase, os quais são formadores de compostos secundários essenciais para utilização biotecnológica.

Revisão Bibliográfica

2. REVISÃO BIBLIOGRÁFICA

2.1. Características e classificação do gênero *Copaifera*

As copaíbas pertencem ao gênero *Copaifera* da Família Fabaceae e sub-família *Caesalpinioideae*. Inclui 72 espécies descritas muito semelhantes entre si, principalmente pelo grande porte das árvores, com 16 diferentes espécies que são encontradas somente no Brasil, nas regiões Amazônica e Centro-Oeste (VEIGA-JUNIOR e PINTO, 2002; RIGAMONTE *et al.,* 2004). Entre as espécies mais abundantes, destacam-se: *C. officinalis* L. (norte do Amazonas, Roraima, Colômbia, Venezuela e San Salvador), *C. guianensis* Desf. (Guianas), *C. reticulata* Ducke, *C. multijuga* Hayne (Amazônia), popularmente conhecida como copaíba branca, *C. confertiflora* Bth (Piauí), *C. langsdorffii* Desf. (Brasil, Argentina e Paraguay) copaíba vermelha, *C. coriacea* Mart. (Bahia) e *C. cearensis* Huber ex Ducke (Ceará) (VEIGA-JÚNIOR *et al*., 1997; VEIGA-JÚNIOR e PINTO, 2002; VEIGA-JÚNIOR *et al*, 2007; PIERI *et al.,* 2012).

As espécies de copaíbas são adaptadas a diversos ambientes, podendo ocorrer em florestas de terra firme, terras alagadas, margens de lagos e igarapés da bacia Amazônica bem como nas matas do cerrado (RIGAMONTE *et al.,* 2004)

O nome copaíba possivelmente tem origem tupi “cupa-yba” que significa “árvore de depósito” (VEIGA-JUNIOR e PINTO, 2002; PIERI *et al.,* 2009), em referência ao óleo depositado no tronco das árvores das espécies. O óleo de copaíba e suas propriedades medicinais eram muito conhecidos pelos índios latino-americanos, supõe-se que tal conhecimento advém da observação do comportamento de alguns animais feridos, que atritavam seu couro no tronco das árvores de copaíba, na busca da cicatrização de suas feridas (VEIGA-JUNIOR e PINTO, 2002; PIERI *et al*., 2009).

No Brasil, os óleos de copaíba são amplamente utilizados na medicina popular, sendo administrados oralmente e por aplicação tópica do óleo *in*

natura ou através do uso tópico de pomadas. Nos estados da região norte é comum o uso principalmente para tratar infecções na garganta (VEIGA-JUNIOR *et al.,* 1997).

Espécies de *Copaifera,* consequentemente, ganharam o nome comum de "árvore diesel", pois a copaíba é uma fonte atraente de biodiesel, entretanto seu uso é atualmente limitado pela sua distribuição geográfica, uma vez que as árvores desse gênero só crescem nos trópicos. Para utilizar o óleo-resina como biodiesel, os programas de reprodução podem ser úteis para adaptar as *Copaiferas* em regiões temperadas. Alternativamente, a engenharia genética poderia ser usada para transferir todas as vias bioquímicas para a produção de óleo-resina de *Copaifera* em outras plantas para produzir novas culturas de bioenergia adequadas para regiões geográficas mais amplas (CHEN *et al*., 2009)

2.2 Características do óleo de *Copaifera*

O óleo-resina de *Copaifera* é essencialmente encontrado em canais secretores localizados em todas as partes da árvore, e possivelmente, o óleo é produto da desintoxicação do organismo vegetal e funciona como defesa da planta contra animais, fungos e bactérias (VEIGA JUNIOR *et al.,* 2002, HECK *et al.,* 2012). Devido ao seu valor comercial são exportados para as indústrias de cosméticos europeias (VEIJA-JUNIOR *et al.,* 1997).

A produção do óleo-resina por espécies de copaíbas é extremamente variável e os fatores que determinam a produção são pouco estudados. As condições ambientais do local de crescimento da árvore, época do ano e suas características genéticas são tidas como fatores determinantes para a variação na produção (RIGAMONTE *et al.,* 2004).

O óleo-resina pode ser obtido por perfuração no tronco das arvores de copaíba, sendo vários os métodos de coleta, alguns extremamente invasivos que podem levar o indivíduo à morte. Porém, atualmente utiliza-se uma técnica que tem sido considerada a única prática não agressiva, a qual consiste na

perfuração do tronco com um trado de aproximadamente 2 metros de diâmetro em dois furos, e após insere-se um cano de PVC de ¾ de polegada nos orifícios, por onde o óleo ocorre o escoamento do óleo. Após a finalização da coleta o orifício é fechado, utilizando argila ou tampa plástica vedante, para proteção contra fungos e cupins (VEIGA JUNIOR *et al.,* 2002; RIGAMONTE *et al.,* 2004; PIERI *et al.,* 2012).

A composição química do óleo de copaíba foi determinada a partir de vários trabalhos, nos quais foram utilizadas técnicas clássicas e metodologias modernas de isolamento e de identificação. Dentre estas, pode-se citar a cromatografia líquida de alta eficiência (HPLC), cromatografia com fluído super-crítico com detector de infravermelho (SFC-FT-IR) e cromatografia gasosa acoplada à espectrometria de massas com colunas cromatográficas de fase estacionária quiral (β-ciclodextrina permetilada) (VEIGA JUNIOR *et al.,* 2002, LEANDRO *et al.,* 2012).

Por meio dessas técnicas constatou-se que os óleos de *Copaifera* são óleo-resinas, pois são constituídos por uma parte sólida formado por ácidos diterpênicos, e cerca de 55% a 60%, diluídos em óleo essencial, com princípio volátil, composto principalmente pelos sesquiterpenos (HECK *et al.,* 2012), que são responsáveis pelo aroma e também pela atividade anti-inflamatória do óleo-resina (BARRETO JUNIOR *et al.,* 2005). A porção resinosa do óleo é formado principalmente por diterpenos, há registros na literatura de 28 diterpenos diferentes (BARRETO JUNIOR *et al.,* 2005).

Apesar das copaibeiras apresentarem alterações nas quantidades e tipos de diterpenos e sesquiterpenos, apenas essas classes de substâncias podem estar presentes na composição do óleo puro. Contudo, apenas 5 espécies de *Copaifera* tiveram a composição descrita na literatura (*C. multijuga* Hayne, *C. langsdorfii*, *C. cearensis*, *C. officinalis* L. e *C. reticulada* Ducke), *C. guianensis* totalizando apenas 18 relatos químicos no total (LEANDRO *et al.,* 2012; CASCON e GILBERT *et al*., 2000).

Alguns trabalhos foram realizados usando cromatografia gasosa para determinar a composição química dos compostos mais abundantes presentes nos óleos-resina de diferentes espécies de *Copaiferas*, as (**Tabelas 1 e 2**) destacam esta composição, tanto de compostos sesquisterpenos componentes do óleo, quando de componentes diterpenos presentes na composição resinosa.

VEIGA-JUNIOR e colaboradores (1997) ao verificarem a autenticidade de 16 óleos de copaíba comerciais de diferentes espécies de *Copaiferas*, constataram que o ácido copálico foi o único composto diterpeno detectado em todos os óleos estudados, indicando, portanto, ser este um biomarcador para o gênero *Copaifera* (VEIGA JUNIOR *et al.,* 2007; HECK *et al.,* 2012).

BASTOS, 2011 em um estudo feito utilizando óleos extraídos de folhas de plantas jovem de *C. multijuga* Hayne, observou que o composto sesquiterpeno mais abundante na composição do óleo essencial foi D-germacreno, seguido de cis-cariofileno. BARDAJÍ *et al.*, 2016 também por cromatografia gasosa descreveram que os principais constituintes do óleo da espécie *C. reticulata* são compostos principalmente por β-bisaboleno, trans-α-bergamoteno, β-selineno e α-selineno.

Em *C. multijuga* estudos fitoquímicos demonstram que o óleo contém ácido linoléico e palmítico, bem como, misturas de sesquisterpenos e metabólitos secundários que auxiliam na proteção vegetal (GONÇALVES *et al.,* 2014). Sabe-se que a composição química do óleo, cor e viscosidade podem ser variadas entre as espécies de *Copaiferas* e as regiões onde estas espécies se encontram (VEIGA JUNIOR *et al.,* 2002; PLOWDEN, 2003).

Dentre as propriedades farmacológicas do óleo destacam-se as anti-inflamatórias, (VEIGA JUNIOR *et al.,* 2007; BARBOSA *et al.,* 2012; LUCCA *et al.,* 2015), atividade larvicida em *Aedes aegypti* (GERIS *et al.,* 2008), antitumorais (LEANDRO *et al.,* 2012) e antimicrobianas, que pode ser justificada pela presença de ácido caurenóico (VEIGA JUNIOR *et al.,* 2007; PIERI *et al.,* 2012), além das propriedades antinociceptiva e anti-leshmania

(GOMES *et al.,* 2007; LEANDRO *et al.,* 2012), onde o óleo resina produzido pela espécie *C. multijuga* se destaca.

Tabela 1- Compostos sesquiterpenos mais abundantes presentes em óleos de diferentes espécies de *Copaiferas*.

Espécie	Componente Sesquiterpeno	Concentração (%)	Material	Idade da planta	referência
C. multijuga	D-germacreno guaiol cis-cariofileno δ-cadineno α-cubebeno α-eudesmol α-copaeno	31,46 7,53 5,74 5,71 5,06 3,75 1,79	Folha	jovem	Bastos, 2011
C. langsdorffii	γ-muuroleno β-cariofileno δ-elemeno D-germacreno	25,2 16,6 2,1 1,4	Folha	adulta	Gramosa e Silveira, 2005
C. multijuga	9-epi(E) cariofileno D-germacreno- α-copaeno	46,92 17,39 10,90	Caule	jovem	Bastos, 2011
C. langsdorffi	Óxido cariofileno Caureno β-caryophyllene	31.0 30.2 8,0	madeira do tronco	adulta	Gramosa e Silveira, 2005
C. langsdorffii	β-cariofileno germacreno B β-selineno α-humuleno γ-elemene	53,3 8,7 6.5 6,1 4,8	óleo puro	Adulta	Gramosa e Silveira, 2005
C. multijuga	β-cariofileno α-humuleno α-bergamoteno	57,5 8,3 2,6	óleo puro	Adulta	Veiga Junior *et al.* 2007
C. cearensis	β-cariofileno α-copaeno β-bisaboleno	19,7 8,2 8,2	óleo puro	adulta	Veiga Junior *et al.* 2007
C. reticulata	b-Bisaboleno trans-a-ergamoteneo b-Selineno a-Selineno	24.91 21.99 12.17 11.43	óleo puro	adulta	Veiga Junior *et al.* 2007
C. multijuga	α-copaeno β-cariofileno δ-cadineno	5.2 60,3 2,9	óleo puro	adulta	Cascon e Gilbert, 2000
C. guianensis	β-cariofileno trans-a- ergamotene	4.7 7,2	óleo puro	adulta	Cascon e Gilbert, 2000

Tabela 2- Compostos diterpenos mais abundantes presentes em óleos-resina de diferentes espécies de *Copaiferas*.

Espécie	Componente Diterpeno	Concentração (%)	material	Idade da planta	referência
C. multijuga	ácido copálico 3-acetoxicopalico	11,0 6,2	óleo puro	adulta	Cascon e Gilbert, 2000
C. guianensis	caur-16-en-19-oico ácido polialtico ácido copálico	17,5 10,6 1,4	óleo puro	adulta	Cascon e Gilbert, 2000
C. multijuga	ácido copálico	6,2	óleo puro	adulta	Veiga Junior *et al.* 2007
C. cearensis	ácido copálico metil clorecinato	2,1 11,3	óleo puro	adulta	Veiga Junior *et al.* 2007
C. reticulata	ácido copálico methyl kaurenoate	2,4 3,9	óleo puro	adulta	Veiga Junior *et al.* 2007

2.3 Metabólitos Secundários em Vegetais

Os fitoquímicos, muitas vezes também chamados de metabolitos secundários, são compostos químicos produzido por plantas através de várias vias químicas. Geralmente, a maioria dos fitoquímicos ajuda as plantas na proteção contra patógenos, herbívoros ou estresses abióticos, como: altos níveis de radiação UV (ADAMCZYK *et al*., 2018; HOLOPAINEN *et al*., 2018). Os fitoquímicos defensivos podem ser detectados visualmente, provados ou até mesmo seus aromas sentidos pelos seres humano e outros organismos. Vários fitoquímicos dão cor (por exemplo, antocianinas) ou sabor e aroma (por exemplo, monoterpenos voláteis, sesquiterpenos e alguns fenilpropanóides) (HOLOPAINEN *et al*., 2018). Essas características permitem o uso destes compostos em grandes aplicações, como produtos farmacêuticos, inseticidas, corantes, aromas e fragrâncias (GOOSSENS *et al.*, 2003).

A infinidade de substâncias fitoquímicas que possuem papéis altamente especializados em interações ecológicas, muitas vezes ocorre para atração de inimigos dos herbívoros, predadores ou parasitóides, que utilizam os compostos voláteis como pistas para localizar suas presas ou hospedeiros;

este mecanismo é denominado "defesa indireta" (SCHNEE *et al.,* 2002; THOLL e LEE, 2011).

O conjunto de transformações de moléculas orgânicas catalisadas por enzimas que ocorrem em uma célula são denominadas de metabolismo. Em espécies vegetais ocorrem duas classificações para o tipo de metabolismo realizado, universalmente categorizado como metabolismo primário e secundário. O metabolismo primário é realizado por um conjunto de micromoléculas comuns nos vegetais que realizam todos os processos vitais celulares, tais como os fitosteróis, nucleotídeos, aminoácidos e ácidos orgânicos. Estes são encontrados em todas as plantas e executam funções metabólicas que são essenciais e, normalmente, evidentes (CROTEAU *et al.,* 2000). O metabolismo primário é reconfigurado para suportar o aumento da demanda da resposta de resistência (BOLTON, 2009). Ocorrem nas células uma grande produção e distribuição destas micromoléculas para que as funções essenciais do organismo ocorram, tais como: a divisão celular, crescimento, respiração, armazenamento e reprodução.

O metabolismo secundário, em contraste com metabolismo primário, envolve compostos de baixo peso molecular que incluem uma vasta gama de compostos que somam-se mais de 200.000 estruturas definidas, abrange todas as facetas fisiológicas e bioquímicas de "produtos secundários", incluindo aspectos funcionais e evolutivos, de diversidade e complexidade estrutural, que são pouco abundantes no organismo, com uma frequência inferior a 1 % do carbono total, ou pelo fato de sua estocagem ocorrer em órgãos ou células específicas (BOURGAUD *et al.,* 2001; HARTMANN, 2007; WINK, 2010).

Os produtos secundários têm um papel importante na adaptação das plantas aos seus ambientes. Geralmente são considerados não essenciais para o crescimento e desenvolvimento vegetal, ao contrário dos produtos do metabolismo primário. No entanto, estes são indispensáveis para a sobrevivência de uma espécie. O papel funcional destes fitoquímicos vai desde a ecologia à defesa, melhoria da proteção contra ambos os estresses bióticos e

abióticos, além de estarem envolvidos em papéis ecológicos como atraentes ou repelentes, para polinizadores e fitofagia respectivamente, cores e aromas de órgãos reprodutores (flores e frutos) (IRITI e FAORO, 2009). Além disso, alguns destes metabólitos constituem importantes compostos que absorvem a luz ultravioleta evitando que as folhas sejam danificadas (FUMAGALI *et al.,* 2008). Uma característica de plantas e outros organismos sésseis, que não podem fugir em caso de perigo ou que não têm um sistema imunológico de combate a agentes patogênicos, é a sua grande capacidade para sintetizar uma enorme variedade destes compostos (STONE e WILLIAMS, 1992; WINK, 2010).

Devido às importantes funções desempenhadas nas células, os metabólitos secundários despertam grande interesse, não só pelas atividades biológicas exercidas pelas plantas em resposta aos estímulos do meio ambiente, mas também pela imensa atividade farmacológica que possuem. Muitos são de importância comercial não apenas na área farmacêutica, mas também nas áreas alimentar, agronômica, perfumaria entre outras (PEREIRA e CARDOSO, 2012). Os metabólitos presentes no óleo de copaíba, por exemplo, são de grande importância para o ramo farmacêutico, devido às inúmeras propriedades observadas (VEIGA JUNIOR *et al.,* 2007; BARBOSA *et al.,* 2012; LUCCA *et al.,* 2015).

As distinções entre o metabolismo secundário e metabolismo básico ou primário já eram reconhecidas desde a segunda metade do século 19. Mas o termo foi provavelmente introduzido pela primeira vez somente em 1891 por Albrecht Kossel (BOURGAUD *et al.,* 2001; HARTMANN, 2007).

As pesquisas com metabólitos secundários iniciaram em meados do século 20. No início eram feitas apenas compilações de várias classes de constituintes químicos (por exemplo, terpenóides, fenilpropanóides, poliquetídeos e alcalóides) e a descrição de sua distribuição dentro do reino vegetal bem como sua ocorrência e acumulação dentro dos tecidos de plantas. O interesse em produtos naturais não foi puramente acadêmico, mas sim motivado por sua

grande utilidade como corantes, polímeros, fibras, colas, ceras, óleos, agentes aromatizantes, perfumes e drogas (CROTEAU *et al.,* 2000).

O estudo com compostos secundários de plantas pode ser considerado como tendo iniciado em 1806, quando Friedrich Wilhelm Serturner isolou a morfina (Principium Somniferum) a partir do ópio da papoula. Logo após, surgiram pesquisas em velocidade recorde para o isolamento de metabólitos secundários de plantas, bem como as primeiras bases avançadas na indústria farmacêutica para o descobrimento de novas drogas. A primeira síntese de um produto secundário, índigo, por Von Baeyer em 1886, forneceu um marco na química orgânica sintética. A estrutura química da morfina foi elucidada em 1823, ao passo que a primeira síntese total de sua complexa estrutura só foi concluída até 1950, quase 150 anos após seu isolamento (HARTMANN, 2007).

De um modo geral, os precursores de vias metabólicas secundárias são produtos do metabolismo primário. Portanto, um fator de estresse grave ou de longa duração poderia induzir uma mudança excessiva entre o metabolismo primário e secundário e, consequentemente, um desvio de recursos essenciais disponíveis do crescimento para a defesa vegetal (IRITI e FAORO, 2009). Os metabólitos secundários de plantas são usualmente classificados de acordo com a sua rota biossintética ou rota metabólica. Geralmente são consideradas três principais famílias de moléculas: os compostos fenólicos, terpenóides e os alcalóides (HARBONE, 1999; IRITI e FAORO, 2009; PEREIRA e CARDOSO, 2012).

Os compostos fenólicos são uma classe de derivados de fenilalanina com um grupo C6-C3 básico (fenil-propano). Representando cerca de 40% do carbono orgânico que circula na biosfera os 8.000 ou mais compostos fenólicos encontrados são formados por meio de qualquer rota do ácido chiquímico ou o malonato (CROTEAU *et al.,* 2000; IRITI e FAORO, 2009). Há teorias que sugerem que a adaptação evolutiva bem-sucedida das plantas para a terra foi conseguida em grande parte pela formação maciça de compostos fenólicos.

Embora a maior parte destas substâncias tenha assumido papéis estruturais da parede celular, uma vasta gama de componentes não estruturais também foi formada, tendo vários papéis importantes como: defesa das plantas, determinando certas características distintivas de diferentes madeiras e cascas (por exemplo, durabilidade), coloração das flores, e contribuindo substancialmente para certos sabores e odores. Estas funções desempenhadas por compostos fenólicos são essenciais para a sobrevivência de todos os tipos de plantas vasculares. Há inclusive certas espécies vegetais que desenvolveram compostos fenólicos para inibir o crescimento de outras plantas competidoras (ação alelopática) (CROTEAU *et al.,* 2000).

Dentre os metabólitos secundários vegetais estão os da classe terpeno (isoprenóides). Os terpenos são compostos chave presentes na composição de óleos essenciais, terebintinas e resinas e formam a base de uma gama de produtos comercialmente úteis (TRAPP e CROTEAU, 2001). Muitos terpenóides possuem potentes atividades farmacêuticas. Por exemplo, taxol e artemisinina são medicamentos eficazes para tratar câncer e malária, respectivamente (CHEN *et al.,* 2009).

Os terpenóides constituem a maior classe de metabólitos secundários de plantas. Mais de 22.000 compostos já foram descritos e só na década de 70 o número de estruturas definidas dobrou (MCGARVEY e CROTEAU, 1995). Possuem baixo peso molecular, tais como monoterpenos com 10 carbonos e sesquisterpenos de 15 carbonos. São voláteis à temperatura ambiente e são encontrados como componentes principais de aromas florais e óleos essenciais de ervas, vegetais, frutas. Todos os terpenos de plantas são derivados de precursores isopreno. O verdadeiro precursor dos terpenos foi caracterizado como ácido mevalônico (MVA: *mevalonic acid*) ou mevalonato, via sintética que ocorre no citosol, proveniente da união de unidades de acetil coenzima A (YAHYAA *et al.,* 2015).

Os alcalóides podem ser definidos como compostos farmacologicamente ativos, contendo um nitrogênio e derivados de aminoácidos. Estes são um

grupo extremamente diversificado de cíclicos, compostos com uma grande variedade de alvos e atividades biológicas, incluindo interferência de neurotransmissores, interrupção da replicação do DNA e inibição da síntese protéica. Os alcalóides são produzidos por 20-30% de todas as espécies de plantas mais altas, com frequentemente impactos significativos na alimentação de herbívoros (ZÜST e AGRAWAL, 2016).

Uma vez que que o papel dos alcalóides nas plantas ainda é uma questão difícil de ser respondida, mas de acordo com CROTEAU *et al.,* 2000, algumas respostas estão surgindo amparadas nas funções eco-químicas destes compostos. O papel dos alcalóides nas defesas químicas das plantas é sustentado pela grande variedade de efeitos fisiológicos que estes exercem sobre os animais e também por suas atividades antimicrobianas. Vários alcalóides são tóxicos aos insetos e atuam como repelente para herbívoros, o que os tornam de grande interesse industrial (BOURGAUD *et al.,* 2001).

Devido ao importante papel desempenhado pelos metabólitos secundários há um grande interesse no detalhamento das rotas biossintéticas destas micromoléculas. Atualmente técnicas como transcriptomas são empregadas para determinação precisa da complexidade destas rotas.

2.4 Síntese e características de terpenos em vegetais

Os metabolitos do terpeno não só são essenciais para o crescimento e desenvolvimento das plantas (por exemplo, fitohormonas de giberelina), mas também representam ferramentas importantes nas várias interações de plantas com o meio ambiente (THOLL, 2006).

Os genomas de plantas possuem conjuntos de genes relacionados que codificam enzimas para formação de terpenos. Essa família gênica definida como: terpeno sintases (TPS) (CHEN *et al*. 2011). Os conjuntos de dados da sequência genômica e de etiqueta de sequência expressa (EST) de plantas modelo tais como: *Arabidopsis*, milho, arroz, tomate, *Medicago* e *Picea*

mostraram a grande variedade de genes da família terpeno sintase (KULHEIM *et al.,* 2015).

A partir da análise de vários genomas CHEN *et al.* (2011) descreveram a família de genes terpeno sintases *TPS*, como uma família de tamanho médio, com um número de genes que variam de cerca de 20 a 150. Os produtos destas enzimas multi-produto podem ser ainda mais modificada a partir de uma oxigenação por citocromo P450 monooxigenases, ou metilação por metil transferases para formar os compostos adicionais.

Esta família está dividida em três classes e sete sub-famílias: **Classe I-** Consiste em TPS-c (difosfato copalil e ent-caureno), TPS-E / F (ent-caureno e outros diterpenos, bem como alguns mono e sesquiterpenos) e TPS-h (Selaginella específico). **Classe II** consiste em TPS-d (gimnosperma específica) e **Classe III** do TPS-a (sesquiterpenos), TPS-b (monoterpenos cíclicos e diterpenos) e TPS-g (acíclico monoterpenos). De acordo com estes subtipos foram identificados através de sequenciamento e estudos funcionais de uma ampla variedade de plantas tais como: *Arabidopsis thaliana*, *Vitis vinifera*, *Solanum lycopersicum*, *Selaginella moellendorfii* e *Populus trichocarpa* (KULHEIM *et al.,* 2015).

As enzimas formadas a partir de genes de TPS em vegetais são provenientes de duas vias metabólicas, a via do ácido mevalônico (MEV: *mevalonic acid*) ou mevalonato, no citosol, e a rota metabólica do metileritritol (MEP) nos plastídios das células (YAHYAA *et al.,* 2015; KULHEM *et al.*, 2015; WANG *et al.,* 2015; HATTAN *et al.,* 2016, LU *et al*. 2016) como pode ser observado (**Figura 1**). A via MEV faz parte do processo metabólico no citosol de eucariotos e em algumas das actinobactérias e archaea, enquanto que a via do MEP está presente em procariotos e em cloroplasto de plantas (HATTAN *et al.,* 2016, LU *et al*. 2016).

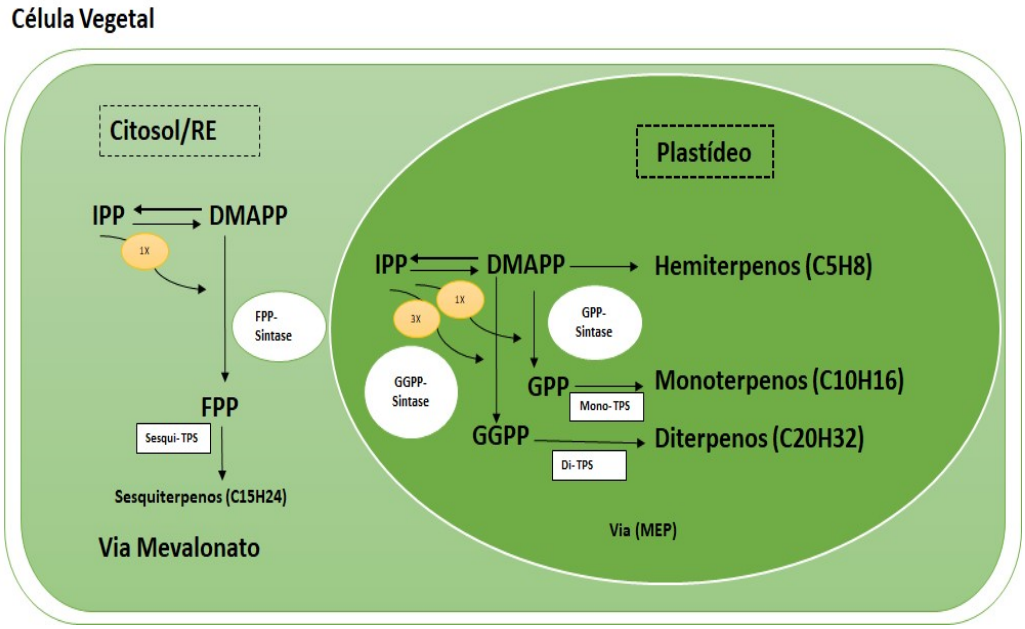


Figura 1- Vias metabólicas de síntese dos terpenos. O esquema suamariza duas vias de síntese de terpenos, a via do mevalonato e a via MEP. (IPP) isoprenil C5; Dimetilalil diposfato (DMAPP); geranildifosfato sintase (GPP); geranilgeranil difosfato sinatase (GGPP). Baseado em AUBOURG et al., 2002.

Os precursores dos terpenos formados nas vias MEV e MEP são dois isômeros, o pirofosfato de isopentenil (IPP) e o pirofosfato de dimetilalil (DMAPP) (AUBOURG *et al.,* 2002; CHEN *et al.,* 2011; KULHEIM *et al.,* 2015; SINGH e SHARMA, 2015). Em plantas, as enzimas terpeno-sintases (TPSs) são responsáveis pela síntese das várias moléculas de terpeno a partir destes dois precursores (HATTAN *et al.,* 2016). Comparado com a via MEVA, o MEP é considerada como uma via "deficiente em energia", onde a redução adicional é necessário poder para produzir precursores de terpeno (WANG *et al*., 2015).

Na via MEV (citosol) o IPP e o DMAPP são sintetizados a partir da união de três unidades de acetil coenzima A (YAHYAA *et al.,* 2015; WANG *et al.,* 2015). Enquanto que nos plastídios, a síntese é derivada a partir do piruvato e do gliceraldeido-3-fosfato através da metileritritol-fosfato. Nos plastídios o pirofosfato de dimetilalila (DMAPP) gerado a partir da via MEP é utilizado pelas enzimas isopreno sintases para formar os isoprenos. Nos plastídios, a junção

de uma molécula de IPP e uma molécula de DMAPP catalisada pela enzima geranil pirofosfato sintase (EC 2.5.1.1) (GPPS) forma o geranil pirofosfato (GPP), uma molécula com 10 carbonos, precursor universal de todos os monoterpenos. No citosol, a condensação de duas moléculas de IPP e uma molécula de DMAPP, catalisada pela enzima farnesil pirofosfato sintase (EC 2.5.1.92) (FPPS) resulta na formação da farnesil pirofosfato FPP, uma molécula com 15 carbonos, o precursor natural dos sesquiterpenos (PICHERSKY *et al.,* 2006; DUDAREVA *et al.,* 2006).

Os compostos terpênicos podem ser classificados em monoterpenos (C10), sesquiterpenos (C15), diterpenos (C20), triterpenos (C30) e tetraterpenos (C40) de acordo com o número de estruturas de isopreno presente (WANG et al., 2015) (**Tabela 3**). Os monoterpenos tem como principal característica serem compostos altamente voláteis e os sesquiterpenos são compostos semivoláteis, que fornecem aromas característicos de várias espécies de plantas e também são utilizados em perfumes (HOLOPAINEN *et al*., 2018).

Tabela 3- Classificação de terpenóides baseada no número de unidades isopreno.

Terpenóides		
Classe	**Unidades Isopreno**	**Fórmula Molecular**
Hemiterpenóides	1	C5H8
Monoterpenóides	2	C10H16
Sesquiterpenóides	3	C15H24
Diterpenóides	4	C20H32
Sesterterpenóides	5	C25H40
Triterpenóides	6	C30H48
Tetraterpenóides	8	C40H64
Politerpenóides	>8	(C5H8)n

Os terpenos têm diversas aplicações industriais como químicos, nutracêuticos, antioxidantes e drogas (AJIKUMAR *et al.*, 2010; SINGH *e* SHARMA, 2015). Muitos têm atividades farmacológicas e biológicas e,

portanto, são interessantes para a medicina e a biotecnologia. A indústria de sabor e fragrância tem, por si só, mais de US $ 1 bilhão no mercado de terpenos. (WU *et* al., 2006; SINGH *et al*., 2015). Além disso, propriedades termoquimicas e termofísica dos terpenos os tornam ideais para uso como combustíveis. Por exemplo, a estrutura em anel de C10 limoneno permite maior densidade de energia e pode servir precursor de combustíveis (WANG *et al*, 2015).

Sesquiterpenos C15 como b-cariofileno é um componente importante de oleorresina *Copaifera* que pode ser utilizado diretamente como diesel (CHEN *et al*., 2009; WANG *et al.* 2015).

No entanto, a produção natural de terpenos geralmente tem baixo rendimento e gera misturas de complexas terpenos e, portanto, não atende à crescente demanda na indústria de terpenos. Este dilema permite o papel da produção de terpeno fotossintético como uma rota potencial para produzir terpenos em vez de usar os recursos naturais (WANG 2015), o que leva gera um grande interesse nas industrias biotecnológicas. Desenvolvimentos em engenharia metabólica e biologia sintética oferecem novas possibilidades para a superprodução de produtos naturais complexos ao otimizar mais hospedeiros microbianos tecnicamente aceitáveis (AJIKUMAR *et al*., 2010).

2.4 Projetos Transcriptoma

A técnica de sequenciamento de DNA mudou completamente a visão sobre a biologia e, particularmente, biologia vegetal. Com as plataformas de sequenciamento foi possível caracterizar um grande número de genes por meio de suas sequências de nucleotídeos, proporcionando assim um atalho para as sequências proteicas correspondentes, bem como as funções por elas desempenhadas. As informações sobre as sequências gênicas e seus polimorfismos facilitaram o mapeamento genético, clonagem de genes e a compreensão das relações evolucionárias e permitiu o início de estudos sobre a biodiversidade (DELSENY *et al.,* 2010).

O método de sequenciamento mais popular desenvolvido foi o de Sanger e Colson descrito pela primeira vez em 1977, o qual tem sido continuamente melhorado. Permitiu o sequenciamento de fragmentos de DNA cada vez maiores e, finalmente, genomas completos de diversas espécies de referência como: *Haemophilus influenzae*, *Saccharomyces cerevisiae*, *Escherichia coli*, *Caenorabditis elegans*, *Drosophila melanogaster*, *Arabidopsis thaliana* e finalmente, o complexo genoma de *Homo sapiens* e de *Oriza sativa*. O sequenciamento desses genomas levou a era da genômica funcional e investigação biológica totalmente modificada (DELSENY *et al.,* 2010, KIRCHER e KELSO, 2010).

Hoje existem basicamente dois tipos de projetos genoma. Um chamado estrutural que é o sequenciamento total do genoma, e outro chamado de sequenciamento funcional do genoma ou transcriptoma, que se baseia no sequenciamento apenas dos genes transcritos e tem a vantagem de poder caracterizar a expressão temporal e local dos genes (CARNEIRO *et al.,* 2000).

Por analogia com o termo genoma, que representa o conjunto completo dos genes (DNA) de um organismo, criou-se o termo transcriptoma, que representa o conjunto completo dos transcritos (RNAs), o que gera um grande desafio para análise global desse sistema (PASSOS e JORDAN, 2000). O transcriptoma é um conjunto completo de moléculas de RNA (transcritos) que estão num dado momento, presentes num dado organismo/órgão/tecido/célula.

Transcriptômica é uma maneira eficiente de descobrir genes ou famílias de genes que codificam enzimas envolvidas em várias vias metabólicas em vegetais. As tecnologias de sequenciamento de nova geração (NGS) de alto rendimento revolucionaram a transcriptômica, especialmente com o advento do sequenciamento de RNA (RNA-seq). Esta tecnologia pode ser usada para obter sequências de RNA em uma grande escala com enorme capacidade de seqüenciamento. Apesar dessas vantagens, as leituras de sequência obtidas a partir de plataformas NGS, como Illumina, SOLiD e Roche-454, são muitas

vezes curtas (35-500 pb) em comparação com o sequenciamento Sanger tradicional (> 700 pb) (METZKER, 2010; XIAO *et al.*, 2013).

A quantidade de cada molécula de RNA é uma resultante do equilíbrio entre a transcrição, o processamento do RNA e os eventos de degradação de RNA. O transcriptoma sofre constantes alterações qualitativas e quantitativas que refletem processos fisiológicos naturais ou são desencadeados por estímulos externos. A compreensão do transcriptoma é essencial para interpretar os elementos funcionais do genoma e revelando os constituintes moleculares de células e tecidos, e também para a compreensão do desenvolvimento de doenças. Os principais objetivos de um transcriptoma são: catalogar todos tipos de transcrição, incluindo mRNAs, RNAs não-codificantes e pequenos RNAs; determinar as estruturas dos transcritos gênicos, em termos dos seus locais de iniciação e poliadenilação (5' e 3'), padrões de *splicing* e de outras modificações pós-transcricionais; e para quantificar alteração dos níveis de expressão de cada transcrito durante o desenvolvimento e sob diferentes condições (WANG *et al.,* 2009; ŻMIEŃKO *et al.,* 2011).

O sequenciamento tradicional de transcriptoma é feito em cerca de 800 pares de bases (pb) de apenas uma das fitas de cDNA. Considerando que as bibliotecas de cDNA são montadas direcionalmente, a maior parte do sequenciamento é feito a partir da extremidade 5’ do cDNA, devido ser a mais informativa e conservada. As análises de etiqueta dos genes expressos (ESTs-*Expressed Sequence Tags*) são feitas a partir de bancos de dados de ESTs tem mostrado ser uma grande fonte de identificação, principalmente da função bioquímica de muitos genes e de funções biológicas que podem ser inferidas com base na frequência de certos genes, que são identificados em bibliotecas de cDNAs construídas sob diferentes condições (CARNEIRO *et al.,* 2000).

O sequenciamento tradicional de ESTs**,** através da construção de uma biblioteca de EST é demorado, bem como oneroso. Felizmente, as novas metodologias de sequenciamento por RNA-seq, desenvolvidas recentemente, geram perfis de transcriptomas de forma a obter um alto rendimento utilizando

tecnologias de sequenciamento NGS, tais como: pirosequenciamento 454 e Illumina. Além do menor custo essas técnicas de RNA-seq também permitem: (1) catalogar todos os tipos de transcrições incluindo mRNAs, RNAs não codificantes, e pequenos RNAs; (2) investigar a estrutura da transcrição de genes, os padrões de *splicing*, isoformas gênicas; (3) estudar a modificação pós-transcricionais e mutações e (4) a expressão do gene precisamente quantificados em grande escala, ao mesmo tempo de sequenciamento. Além disso, o RNA-seq é independente do genoma de uma espécie sendo útil para analisar o transcriptoma de uma espécie sem informações completas do genoma. Nos últimos anos, o RNA-seq aumentou a compreensão da complexidade dos padrões de transcrição do gene, estrutura e variações do gene, bem como as redes de regulação de genes (XIE *et al.,* 2012).

Contudo, devido ao grande volume de dados gerados pelas tecnologias de sequenciamento de alto desempenho, os projetos de genômica e transcritômica necessitam fortemente do auxílio de metodologias computacionais de bioinformática.

2.5 Genoma e Transcriptoma de espécies vegetais

O sequenciamento de nova geração desencadeou uma explosão de recursos genômicos e transcriptômicos disponíveis nas ciências de plantas. Após a publicação do sequenciamento do genoma da planta modelo *Arabidopsis thaliana* (119 Mb) no ano de 2000 (THE *ARABIDOPSIS* GENOME INITIATIVE, 2000), foram publicadas cerca de 180 sequenciamentos de genomas da plantas. E este número é amplamente aprimorado, incluindo montagens de transcriptomas vegetais.

Usando a mesma abordagem tomada por Arabidopsis Genome Iniativa, o genoma do arroz (*Oryza sativa*) (INTERNATIONAL RICE GENOME SEQUENCING PROJECT 2005), milho (SCHNABLE *et al.*,2009) e soja (KIM *et al*., 2010) foram sequênciados. O genoma de *O. sativa* está quase completo, assim como genoma de *A. thaliana*, ainda possui lacunas em regiões repetitivas do genoma, mas ainda assim, é considerado o padrão-ouro para

genomas de plantas devido à alta qualidade e natureza final do sequenciamento (HAMILTON e BUELL, 2012).

A partir de 2016, o banco de dados de montagem *shotgun* de transcriptoma do (NCBI) enumerou mais de 450 montagens de genomas de plantas, enquanto o projeto Planta 1KP (http://www.onekp.com/samples/list.php) atualmente inclui pouco mais de 1300 transcriptomas de espécies vegetais. Este aumento notável é o resultado da revolução genômica que forneceu as ferramentas para sequênciar mais rápido e barato transcriptomas inteiros, que consequentemente podem produzir dados suficientes para um inventário transcriptômico de qualidade operacional ("transcriptoma") (BOLGER *et al*, 2017).

Sabe-se que o sequenciamento do genoma de uma planta fornece meios para entender a base genética das diferenças entre plantas e outros eucariotos, e fornece as bases para caracterização funcional detalhada de genes de espécies vegetais (THE *ARABIDOPSIS* GENOME INITIATIVE, 2000).

Sequenciamentos NGS de genomas de plantas forneceram uma nova visão da variação genômica dentro e entre espécies como: polimorfismos de nucleotídeo único (SNPs), variações de número de cópias (CNVs) e outros tipos de variações estruturais (SVs), e *status* da epigenética vegetal. Da mesma forma, o seqüenciamento de transcriptomas de plantas (RNA-seq) forneceu atlas da expressão gênica de várias espécies, caracterizando pequenos RNAs e identificando genes e caminhos envolvidos em aclimatação a estresses bióticos e abióticos (O'ROURKE *et al.*, 2014).

Para espécies vegetais com genomas desconhecidos, a utilização das novas plataformas de sequenciamento ainda é limitada. O tamanho das leituras produzidas é incompatível com a montagem dos genomas nucleares gigantescos e altamente repetitivos das plantas. Os poucos trabalhos realizados têm sido destinados ao sequenciamento de transcritos, ressequenciamento e sequenciamento *De novo* de genomas plastidiais, os

quais são menores (~150Kb) e contêm pouca quantidade de DNA repetitivo (CARVALHO e SILVA 2010).

Sabe-se que as abordagens transcriptômicas em plantas são relativamente novas, os transcriptomas, avaliado por microarrays ou sequenciamento de nova geração, produziram uma grande quantidade de dados até então inéditos em relação à identidade e níveis de transcrição em sistemas de plantas (USADEL e FERNIE, 2013).

Aos poucos o sequenciamento de nova geração está substituindo plataformas baseadas em chips (microarrays) para estudos genômicos. Esta mudança está colocando uma pressão extraordinária sobre o acesso à bioinformática, a velocidade computacional e ao armazenamento e interpretação de dados. O gargalo para resolver problemas de longa data na biologia das plantas será montar e peneirar os terabytes de dados desses estudos (O'ROURKE *et al.*, 2014).

2.6 Montagem *De novo* e Anotação dos transcritos

Montagem é o processo de reconstrução da sequência original obtida a partir de um genoma ou transcriptoma. Para montar um genoma, os programas computacionais geralmente usam dados que consistem em leituras únicas e emparelhadas. As leituras únicas são simplesmente os pequenos fragmentos sequênciados; que podem ser unidos através de sobreposições de regiões em sequências conhecidas como *'contigs'* (BAKER 2012), com o propósito de formar *scaffolds*/supercontigs (regiões maiores) (LEÃO, 2016).

Atualmente existem duas diferentes abordagens para montagem de genomas. A montagem baseada em um genoma de referência e a montagem *de novo* (**Figura 2)**. A primeira consiste no alinhamento dos *reads* com sequências de organismo com genoma já finalizado, assim a nova sequência é

construída com base na ordem dos *scaffolds* do organismo modelo (LEÃO, 2016).

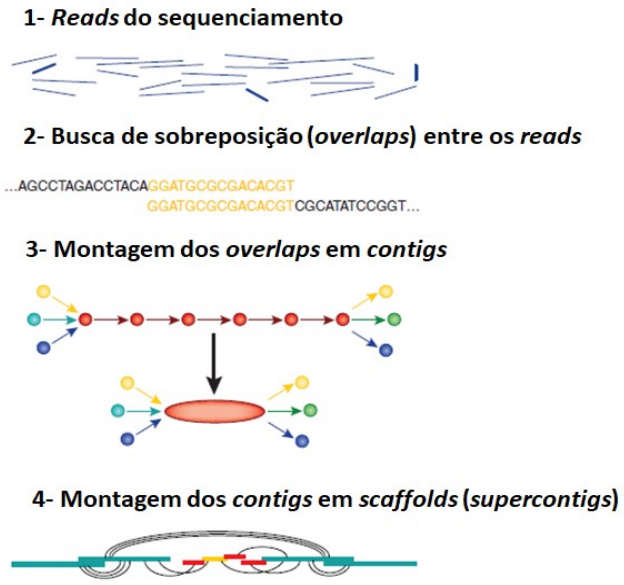


Figura 2- Esquema mostrando a montagem *de novo*, baseado em Baker, 2012.

A montagem de sequências *de novo* é o processo pelo qual sequências individuais (*reads*) resultantes do sequenciamento são organizadas e dispostas umas as outras para formar sequências longas completas (*contigs*). O objetivo é obter *in silico* uma sequência equivalente ao DNA molde original. Quando não se tem o genoma de referência do organismo estudado é preciso gerar, ou montar, esse genoma *de novo*, ou seja, pela primeira vez (PASZKIEWICZ e STUDHOLME, 2010). bootApós a montagem, os *contigs* resultantes são ordenados baseados em um genoma já pronto e que apresente proximidade ao organismo em estudo. Ao final da ordenação, os *contigs* estarão organizados e passarão para o fechamento das lacunas, ou *gaps* como também são chamados, presentes entre os *contigs*, com o objetivo de reduzir a quantidade de pequenos fragmentos e aumentar o número de *scaffolds* da montagem (LEÃO, 2016).

As tecnologias de sequenciamento de alto desempenho foram usadas na montagem *de novo* de muitos genomas de plantas e animais na última

década. No entanto, os genomas montados a partir de leituras curtas são geralmente genomas de rascunho que consistem em até dezenas de milhares de *contigs*, especialmente em plantas, devido ao alto conteúdo de repetição ou baixa cobertura de sequenciamento em regiões complexas GC ou AT. Os conjuntos de genoma geralmente contêm genes fragmentados, repetições colapsadas ou redundantes e *contigs* quiméricos que confundem a função do gene e a detecção de variações estruturais (SV). Portanto, melhorar e validar as montagens do genoma através da aplicação de métodos computacionais e experimentais atualizados é uma importante tarefa de construção de fundação para estudos baseados em genômica (DU *et al.*, 2017).

RNA-Seq é uma poderosa ferramenta para análise de transcriptoma e utiliza tecnologias de sequenciamento de alto desempenho para produzir milhões de *reads* (sequências) de cDNA com curtos fragmentos entre 30-400pb (pares de bases) (METZKER, 2010). As *reads* resultantes são alinhadas com um genoma ou transcriptoma de referência. Alternativamente são montados pela metodologia *de novo* para produzir um mapa do transcriptoma que consiste tanto em estrutura da transcrição como em nível de expressão para cada gene, em um dado estágio de desenvolvimento em particular (GARCIA-SECO *et al.,* 2015).

Atualmente, existem dois principais tipos de algoritmos de montagem para sequências do genoma: gráfico de sobreposição para leituras longas e gráfico de Bruijin para leituras curtas produzidas pela segunda geração de tecnologias de sequenciamento. Existem várias plataformas de montagem de genomas tais como: Mira, Phusion, ABySS , Velvet e Trinity, que utiliza três algoritmos diferentes: Inchworm, Chrysalis e Butterfly (CLARKE *et al*., 2013).

A plataforma TRINITY fornece ainda estatísticas ao final da montagem tais como: quantidade de transcritos montados, quantidade de genes, conteúdo GC das sequências, o tamanho do maior e do menor *scaffold* (*supercontigs*) e o N50, o qual corresponde a estatística média ponderada dos contigs totais, de

tal modo que 50% de todo o conjunto de sequências montadas está contido em *contigs* iguais ou maiores do que este valor (FERNANDES, 2015).

As montagens de genomas continuam sendo um desafio para a comunidade de bioinformática principalmente para equilibrar a eficiência de computação e a precisão da montagem. CLARKE *et al*, 2013, testou cinco modelos computacionais de montagem utilizando várias leituras curtas de RNA partir de dados do Consórcio de Controle de RNA Externo (ERCC) e cromossomo humano 22. Observou que a plataforma TRINITY teve um bom desempenho relativo tanto para ERCC quanto para dados humanos, mas pode não gerar transcrições consistentemente completas. O ABySS foi o método mais rápido, mas a qualidade da montagem foi baixa. Mira gerou uma boa taxa para mapear seus *contigs* no cromossoma humano 22, mas é a velocidade computacional não foi satisfatória.

Após a etapa de montagem os transcritos passam pelo processo de anotação, sendo que para novas tecnologias de sequenciamento (NGS), a anotação automática é a usual, pela gama de resultados obtidos com o sequenciamento.

O objetivo de uma anotação de alta qualidade é identificar as principais características do genoma, em particular os genes e seus produtos. A anotação pode ser dividida em três classes podendo responder a três importantes perguntas: anotação a nível de nucleotídeos, ou seja, busca-se a localização das sequências e genes (Onde os genes estão presentes?); anotação a nível de proteínas, para averiguar as funções dos genes presentes (O que os genes fazem?); e o nível de processo, que busca saber quais as vias metabólicas que os genes estão presentes (Como os genes atuam?) (STEIN, 2001; SOUZA, 2012).

Um dos aspectos mais importantes em triagem de dados genômicos é associar sequências individuais e informações de expressão relacionada com a função biológica. A anotação funcional automática é uma forma eficaz para resolver este problema. Na anotação automática, todas as sequências (*contigs*)

são anotadas, por comparação com sequências disponíveis em bancos de dados públicos como o NCBI. Os *contigs* inicialmente são classificados de acordo com as categorias do COG (*Clusters of Orthologous Groups*) e GO (*Gene Ontology*), para determinação funcional gênica (CONESA *et al.,* 2005).

Após a identificação e a anotação dos genes, a prioridade usual passa a ser o estabelecimento dos padrões de expressão para fins de confirmação de suas funções. Apesar de serem bem definidos, sensíveis e robustos, os métodos tradicionais de análise de expressão gênica como *Northern blot*, hibridização *in situ*, reação em cadeia da DNA polimerase (PCR, do inglês, *Polymerase Chain Reaction*) ou PCR quantitativa (em tempo real) precedida de transcrição reversa (RT-qPCR, do inglês, *Reverse Transcription - Quantitative, real-time PCR*), são adequadas para a análise de um número pequeno de genes. O estabelecimento de perfis transcricionais proporciona um enorme avanço, pois permite uma análise eficiente de um grande número de genes ao mesmo tempo.

A análise da expressão gênica é cada vez mais importante no campo de pesquisas biológicas. O entendimento dos padrões de expressão gênica gera informações sobre a rede complexa de regulação e contribui para a identificação de genes relevantes para novos processos biológicos. Dois métodos foram desenvolvidos para medir a abundância da transcrição ganharam muita popularidade e são frequentemente aplicadas. A técnica de *Microarrays* permite a análise paralela de milhares de genes em duas populações de RNA marcados diferencialmente, enquanto a técnica de RT-PCR proporciona a medição simultânea da expressão de genes em muitas amostras diferentes para um número limitado de genes, e é especialmente adequada quando apenas um número pequeno de células está disponível (VANDESOMPELE *et al.,* 2002).

O uso das novas tecnologias de sequenciamento (NGS) permitiu a determinação robusta do transcriptoma de inúmeras espécies vegetais, principalmente espécie sem genoma de referência como: o gênero *Copaifera,* sendo este com grande potencial econômico, devido principalmente a comercialização do óleo, extraído do tronco das árvores, ao qual são atribuídas

inúmeras propriedades farmacológicas como: antimicrobiana, anti-leshmania, anti-tumoral e antinoceptiva, sendo na Amazônia a espécie *C. multijuga* uma das mais promissoras, tanto para comercialização do óleo como para estudos biotecnológicos.

Objetivos

3. OBJETIVOS

3.1 GERAL:

- Montagem *de novo* e análise do transcriptoma de *Copaifera multijuga* Hayne.

3.2 ESPECÍFICOS:

- Realizar a montagem *de novo* do transcriptoma de folhas de *C. multijuga*;
- Proceder a anotação primária do transcriptoma por meio do programa BLASTx;
- Efetuar a anotação funcional do transcriptoma utilizando o GO.
- Anotar os processos metabólicos relacionados à biossíntese de terpenóides;
- Analisar a filogenia de *unigenes* responsáveis pela síntese de sesquiterpenos *C. multijuga*.

Material e Métodos

4- MATERIAL E MÉTODOS

4.1 Material Vegetal

Para análise do transcriptoma de *C. multijuga* o material obtido neste estudo foi coletado, de acordo com BASTOS, 2011, no período chuvoso, a partir de plantas jovens da espécie *C. multijuga* Hayne com tamanho entre 1 e 1,5 m de altura, medidos da raiz ao caule e com 0,5 cm a 3,0 cm de espessura de caule, obtidos no Viveiro de Plantas Medicinais da Universidade Federal do Amazonas-UFAM/ AM-Brasil (Latitude- 4º 55' 58"; Longitude= 52º 19' 58").

4.2 Extração de RNA e construção da biblioteca de cDNA.

O RNA total foi obtido a partir de folhas previamente maceradas em nitrogênio líquido. A extração seguiu o protocolo de Kiefer e colaboradores com modificação, suprimindo o uso da resina *NucleonTM PhytoPureTM DNA extration resin.* O isolamento da fração mRNA poli (A)+ foi realizado com *Protocol for small-Scale mRNA isolation PolyATract systens III kit* – Promega (Part#TM021). Realizou-se a síntese do cDNA com o kit *Universal Riboclone® cDNA Synthesis System* – Promega (Part#TM038). A biblioteca de cDNA foi quantificada e analisada quanto a faixa de tamanho de seus fragmentos por meio de eletroforese em gel de agarose (1%).

4.3 Sequenciamento do transcriptoma de *C. multijuga* Hayne

A preparação da reação de sequenciamento foi feita utilizando-se 5µL de fragmentos da biblioteca de cDNA (400ng/µL) por meio da metodologia de *shotgun,* em sequênciador 454 FLX/ROCHE. Até esta etapa o trabalho experimental foi realizado por BASTOS 2011 no Centro de Apoio Multidisciplinar (CAM) da UFAM com apoio do Laboratório Nacional de Computação Científica (LNCC) – Petrópolis –RJ.

4.4 Anotação funcional do transcriptoma de *C. multijuga* Hayne

A partir dos dados do sequenciamento de *C. multijuga*, foi feita a montagem *De novo* das sequências utilizando os programas MIRA (Mimicking Intelligent Read Assembly) e TRINITY, onde as sequências foram agrupadas em *contigs* e *singlets*.

A anotação foi realizada por comparação com bancos de dados de sequências de aminoácidos de plantas, VIRIDAEPLANTAE, que possuem genomas completos anotados (exemplo: *Arabidopsis thaliana* e *Oryza sativa*) utilizando a ferramenta BLASTx local (ALTSCHUL *et al.,* 1997). Os *contigs* também foram comparados utilizando os bancos curados de sequências de aminoácidos Refseq, Swiss-Prot (http: //www. expasy.ch/sprot/); TREMBL e contra banco de dados NCBI (http://www.ncbi.nlm.nih.gov/), banco de dados KOG (ftp://ftp.ncbi.nih.gov/pub/COG/KOG/kyva) com o programa BLAST (*E-value* <1E-5). O melhor alinhamento resultante foi selecionado para anotar os *unigenes*. A anotação funcional por termos de ontologia (GO, http://www.geneontology.org) foi analisada pelo *software* BLAST2Go.

As vias metabólicas foram obtidas usando o padrão KEGG (http://www.genome.jp/KEGG/) e banco de dados KOG (ftp://ftp.ncbi.nih.gov/pub/COG/KOG/kyva) com o programa BLAST (*E-value* <1E-5). O melhor alinhamento resultante foi selecionado para anotar os *unigenes*. A anotação funcional por termos de ontologia (GO, http://www.geneontology.org) foi analisada pelo *software* BLAST2Go.

Para a avaliação dos dados brutos do sequenciamento, foi necessário inicialmente fazer a checagem em todos os *reads* gerados, os quais foram agrupados separadamente em arquivos (5´ e 3´). Na etapa de trimagem de sequências foi utilizado o pacote de programas FASTX-Toolkit, versão 0.0.14 (http://hannonlab.cshl.edu/fastx_toolkit/commandline.html). As sequências menores que 50 bases foram *trimadas* utilizando *fastq_quality_trimmer*, em seguida foram selecionadas as sequências que possuíam pelo menos 80 % das bases com *score* 20 de qualidade, com o comando fastq_quality_filter. A

análise estatística como: tamanho e média de *contigs* foi obtida utilizando linguagem de programação R e o N50 foi calculado com o pacote Biostrings.

4.5 Análise e predição de funções gênicas e construção de vias metabólicas e filogenia de terpenos

A anotação foi ratificada com o uso do BLAST e os *contigs* foram classificados de acordo com as categorias do COG (*Clusters of Orthologous Groups*) e GO (*Gene Ontology*). A classificação foi realizada por comparação dos *contigs* de copaíba com todas as sequências disponíveis no repositório COG e GO.

As vias metabólicas foram montadas utilizando o banco de dados KEGG (Kyoto *Encyclopedia of Genes and Genomes*, 2013). Os códigos das enzimas (*Enzyme Commission numbers*, ECs) foram obtidos a partir dos dados em *flatfiles* disponíveis no banco Swiss-Prot para cada proteína identificada entre os *contigs* anotados. Em seguida, os ECs foram submetidos ao banco de dados KEGG PATHWAY http://www.genome.jp/kegg/pathway.html para identificação das vias metabólicas de *Copaifera*. Este é um recurso de banco de dados para a compreensão das funções de sistemas biológicos, especialmente em conjuntos de dados moleculares gerados pelo sequenciamento do genoma e outras tecnologias experimentais de elevado rendimento.

O agrupamento filogenético de terpenos sintases foi feito utilizando 15 sequências parciais nucleotídeos de *unigenes* obtidos a partir do banco de *C. multijuga* Hayne e sequências de outras espécies de *Copaifera* disponíveis no banco de dados (NCBI), foi feita a construção de uma árvore filogenética de Neighbor-Joining (SAITOU e NEI, 1987), com *bootstrap* calculado a partir de 500 réplicas de árvores geradas, utilizando o programa MEGA7 7.0.21 (KUMAR % TAMURA, 2016).

O cálculo de *bootstrap* permite a obtenção de um parâmetro que indique a “confiança” para cada clado de uma árvore filogenética produzida (EFRON et al, 1996), o qual baseia-se na construção de sub-amostras a partir de uma

amostra inicial da população calculando-se as estatísticas, o programa de inicialização gera inúmeras cópias da amostra original para criar uma pseudopopulação, o que gera várias árvores-réplicas, após gerar todas as réplicas, a árvore consenso final é obtida, sendo considerado valores maiores ou iguais que 90% com alto grau de confiabilidade. Os valores presentes na árvore indicam a topologia de consenso das árvores-réplicas (HILLIS e BULL, 2003, CARVALHO, 2012). As distâncias evolutivas foram calculadas usando o método Jukes-Cantor, o qual gera uma matriz de distância com o número de diferenças entre os pares de sequências alinhadas (JUKES e CANTOR, 1969; CARVALHO, 2012).

Resultados e Discussão

5. RESULTADOS E DISCUSSÃO

5.1 Sequenciamento e Montagem *De novo* das sequências de *C. multijuga* Hayne

Sabe-se que transcriptômica é uma ferramenta poderosa e conveniente para analisar a expressão de genes e gerar valiosas informações sobre as sequências gênicas de diversos organismos. Tem sido usada para analisar perfis de expressão gênica entre espécies estreitamente relacionadas ou diferentes tecidos em insetos e plantas. Especialmente para plantas medicinais, os números de transcriptomas sequênciados dispararam nos últimos anos e continua a crescer em ritmo intenso.

Os esforços no sequenciamento de transcriptoma de plantas devem-se ao grande interesse em explorar a biossíntese especializada de metabolitos em plantas medicinais não-modelos e facilitam a descoberta *de novo*s genes biossintéticos responsáveis pela produção de compostos únicos de várias espécies (XIAO *et al.* 2013; SUI *et al.,* 2015), como no caso da espécie *C. multijuga* Hayne.

Os resultados obtidos pelo sequenciamento a partir de folhas da espécie *C. multijuga* Hayne estão resumidos na **Tabela 4**. Foram obtidos um total de **638.576** *reads* (dados brutos de sequências, **Figura 2-A**), dos quais aproximadamente **344.868** sequências (54 %), foram recuperadas após o pré-processamento (**Figura 2-B**). O tamanho médio das sequências variou entre **50** e **496** bp (pares de bases). Estas foram agrupadas a partir da montagem *de novo* pela plataforma TRINITY em **53.319** *contigs* (sequências) e pela montagem por MIRA obteve-se **62.839** *contigs*. Ambas montagens geraram **39.807** sequências *singlets*.

Um total de **11.050** *unigenes* foi determinado após o alinhamento com BLASTx (**Tabela 2**). Na montagem feita a partir da plataforma TRINITY, **5.177** *unigenes* foram exclusivos a partir dessa construção, enquanto que por MIRA foram **5.873** *unigenes* identificados. Do total de *unigenes* anotados (11.050), **2.366** foram redundantes a partir das duas montagens **(Figura 3)**. Com isto

pode-se observar que em montagens de transcriptomas, é indispensável a utilização de mais de uma plataforma de montagem, para assim maximizar os resultados obtidos na etapa de anotação, pois não há um consenso no uso de um único *software* para montagens transcriptômicas gerados a partir de sequenciamento de nova geração, sendo ainda um desafio para a bioinformática (CLARKE *et al.,* 2013).

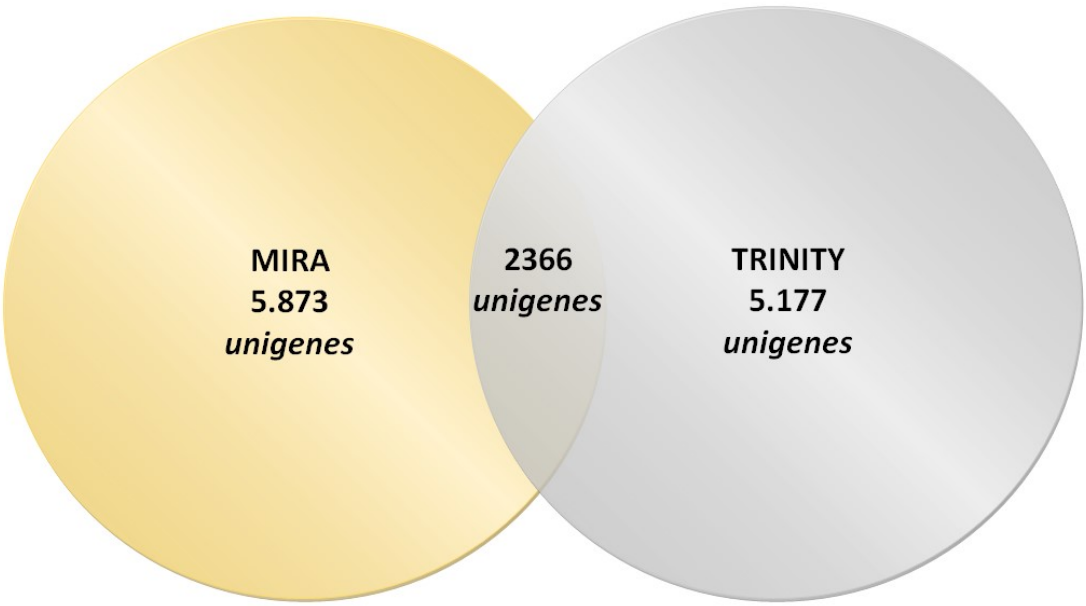
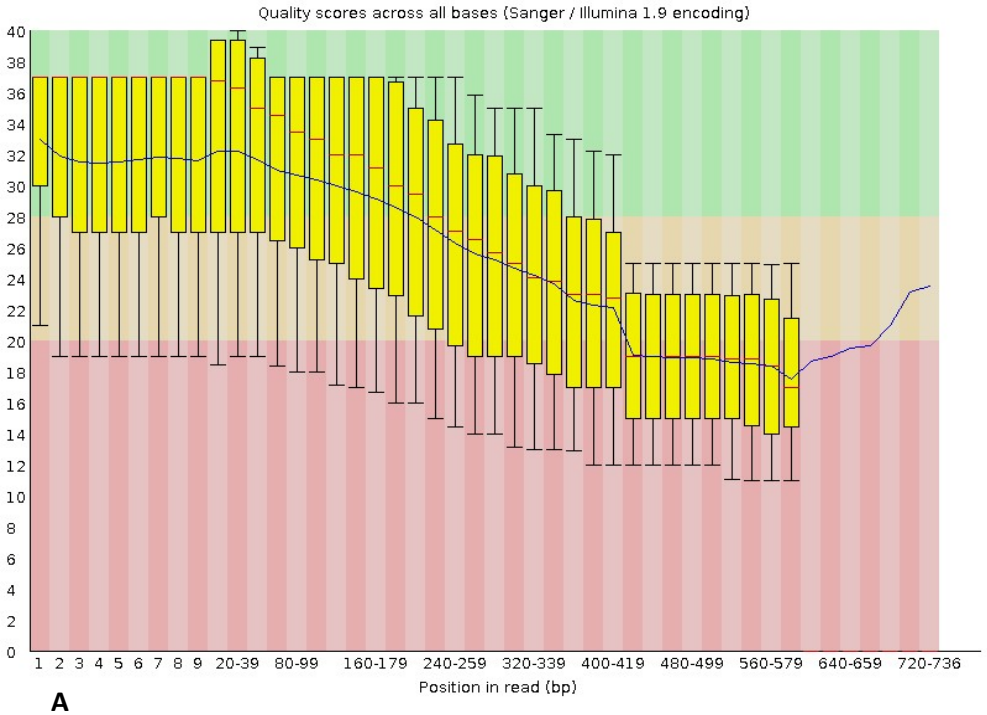
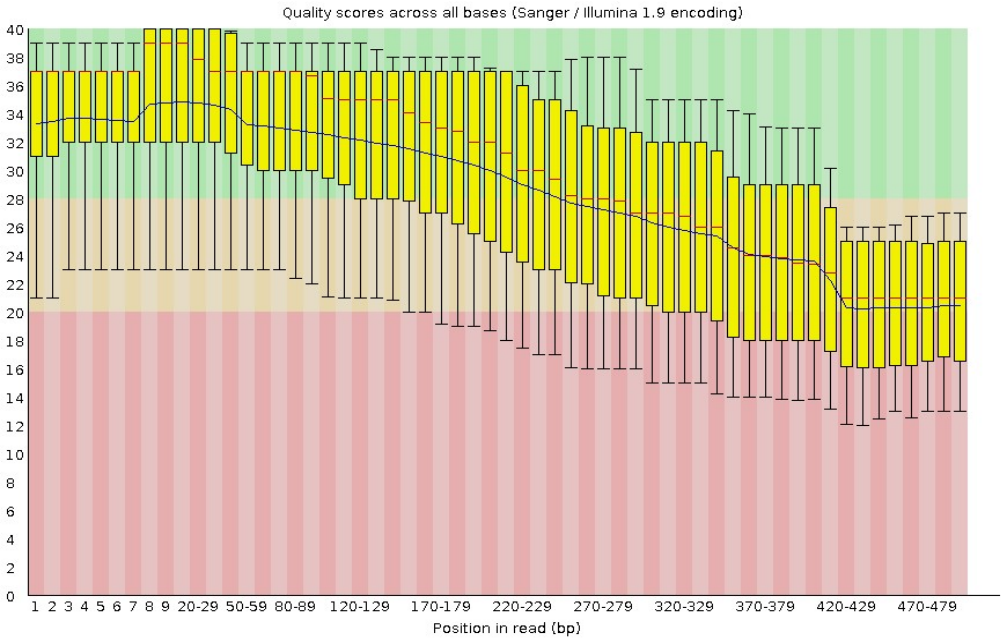


Figura 3- Diagramas de Venn mostrando os resultados obtidos nas montagens MIRA e TRINITY. O número de *unigene*s foi determinado por alinhamento das sequências com BLASTx. A região de interseção indica a redundância de genes anotados utilizando as duas plataformas de montagem.

As sequências selecionadas com melhor qualidade possuíram pelo menos 80% das bases com *score* 20 de qualidade de acordo com o comando *fastq_quality_filter*. versão 0.51 (**Figura 4-A e 4-B**). Dentre os diversos parâmetros avaliados por *fastq_quality_filter* para determinar a qualidade das sequências, a medida por base se mostrou fundamental para avaliar a eficiência dos processos de triagem e limpeza que se seguiram. Assim, na figura 1 é mostrada a comparação na qualidade das sequências por base entre os *reads* brutos (sem qualquer pré-processamento; 4-A) e os *reads* finais (aqueles que foram submetidos à montagem dos *contigs*; 4-B).



A



B

Figura 4- Gráfico que mostra a qualidade por base das sequências dos *reads* brutos pré-processamento (A) e pós-processamento (B), em fastq_quality_filter. As barras amarelas correspondem à amplitude interquartílica (25%-75%) dos valores de qualidade, enquanto que os limites em preto acima e abaixo se atribuem aos valores 10% e 90%; dentro das barras de qualidade, a linha em vermelho é o valor mediano da qualidade; por fim, entre as barras, a linha em azul representa a variação do valor médio da qualidade.

A partir dos *contigs* gerados, a partir das montagens, foram analisados parâmetros de eficiência como: o número de sequências geradas, tamanho mínimo e máximo dos *contigs* e quantidade (N50) de *contigs* utilizados como demonstrado na (**Tabela 4**). As montagens genômicas são medidas pelo tamanho e precisão de seus *contigs* e *scaffolds*, estes às vezes chamados de *supercontigs* ou *metacontigs*, definem a ordem e a orientação do *contig* e os tamanhos das lacunas que ocorrem entre os *contigs*. O tamanho das montagens geralmente é fornecido por estatísticas, incluindo comprimento máximo, comprimento médio, comprimento total combinado e N50. O *contig* N50 é o comprimento dos *contigs*, calculado a partir de um conjunto que contém todos os *contigs* ordenados, cujo comprimento combinado destes representa pelo menos 50% do total da montagem (MILLER *et al.,* 2010).

Tabela 4- Resumo dos dados referentes a montagem dos contigs pelos programas TRINITY e MIRA

Dados de Montagem	TRINITY	MIRA
Número de *reads*	638.576	638.576
tamanho dos reads	224.9 (46- 496 pb)	224.9 (46- 496 pb)
Número de *contigs*	53.319	62.839
Número de *unigenes*	5.177	5.873
Media reads/ *unigenes*	24.677	2.88
Tamanho do menor *contig**	224	40
Tamanho do maior *contig**	5.379	2.519
Tamanho mediano dos *contigs**	371	371
N50	402	402

***Tamanho dos *contigs* em pares de bases (pb)**

5.2 Busca por homologia de sequências por BLAST

A busca por homologias das sequências de *C. multijuga* realizada pelo algoritmo BLASTx, utilizando como base banco de proteínas não redundantes (nr) do NCBI, o *e-value* médio das sequências foi de 10^{-1}, enquanto que a

similaridade média variou em torno de 80% nas sequências. Ao analisar a variedade das espécies presentes nos alinhamentos de maior *e-value* do BLAST, foi observado que para todas as sequências as principais espécies vegetais comparadas com *C. multijuga* (**Figura 5**) foram: *Vitis vinifera*, *Glycine max*, *Medicago truncatula, Glycine soja* e *Phaseolus vulgaris*. O valor absoluto de *hits* anotados no BLAST foi **11.356**. O maior percentual de *hits* foi observado com *V. vinifera* (13,3%; 1.505 hits), seguido por com *G. max* (11,5%; 1.305 hits). Estes resultados corroboram dados descritos na literatura para outras espécies, como por ZWENGER *et al.,* 2010, os quais realizaram a anotação funcional de *C. officinalis*, obtiveram 143 *hits* comparados com *V. vinífera* e 137 com *G. max*. De acordo com ZWENGER *et al.,* 2010 com estes resultados é difícil determinar se a distribuição das espécies para as sequências de copaíba foi verdadeiramente semelhante às sequências das espécies comparadas ou se depende em grande parte da número de sequências para essa espécie dentro do NCBI banco de dados.

Destas, *G. max*, *G. soja* e *P. vulgaris* são espécies pertencentes a família Fabaceae, compartilhando assim maior proximidade filogenética com *C. multijuga* a qual também faz parte desta família.

A família Fabaceae (Leguminosae) é a terceira maior família de angiospermas, contendo 18.000 espécies distribuídas em 650 gêneros. Em comparação com cereais, para os quais uma ampla gama de recursos genéticos e genômicos estão disponíveis os bancos de dados genômicos para as leguminosas são muito escassos. Com isso, grandes esforços estão sendo direcionados para o desenvolvimento de ferramentas e conjuntos de dados genômicos específicos de espécies desta família (KAUR *et al.,* 2012).

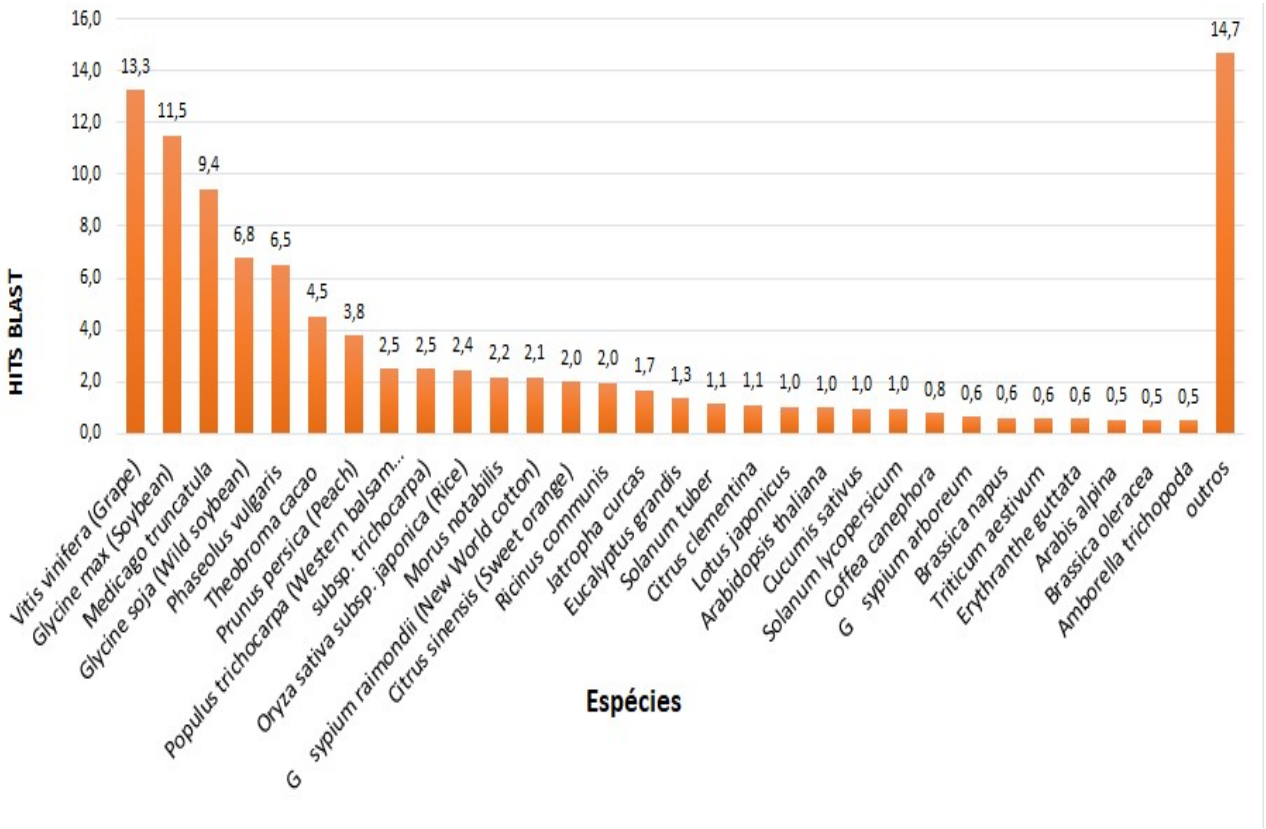


Figura 5- Gráfico das sequências de *C. multijuga* Hayne organizadas de acordo com a frequência dos hits BLASTx. As maiores relações comparativas foram para *V. vinifera, G. max, M. truncatula*. O valor absoluto dos hits comparados foi de 11.356. O eixo (x) corresponde as espécies comparadas com *C. multijuga*.

5.3 Ontologia gênica e anotação funcional de sequências de *C. multijuga* Hayne

No mapeamento gênico feito pelo *Gene Ontology* (GO) para anotação funcional dos *unigenes* presentes em *C. multijuga*, foi possível determinar que 54% dos *contigs* obtidos foram anotados totalizando, **32.956** termos GO assinalados. A categorização dos genes identificados de acordo com os termos do GO foi: **9.499** *contigs* para "Processo Biológico" (**Figura 6**), **6.249** *contigs* correspondendo a categorização "Componente Celular" **(Figura 7)** e **17.208** *contigs* envolvidos na "Função Molecular" **(Figura 8)**. Os *contigs* mais abundantes dos três processos ontológicos analisados por GO foram organizados de acordo com a (**Figura 13)**.

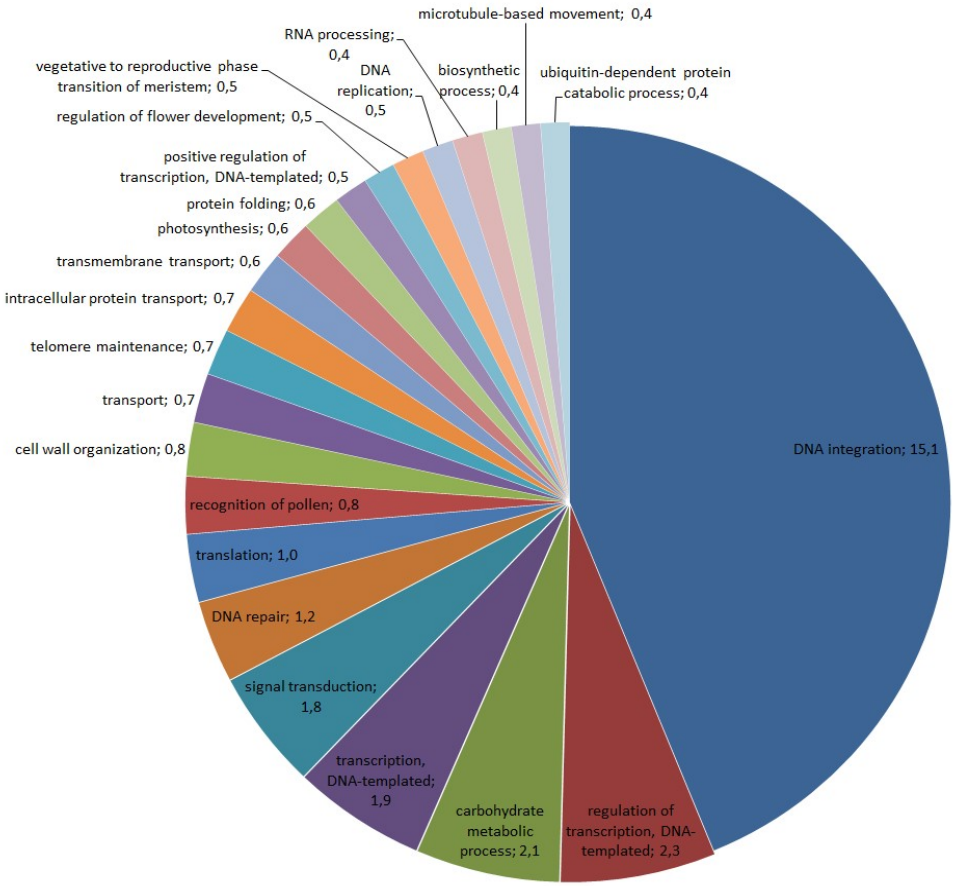


Figura 6- Gráfico dos *contigs* mais abundantes descritos na categoria "Processo Biológico" de *C. multijuga* Hayne. Os valores representam o percentual a partir do número absoluto de 9.499 *contigs* nesta categoria por GO.

A Ontologia gênica define o universo dos conceitos para funções de genes ("termos GO") e como essas funções estão relacionadas entre si ("relações"). A base de dados é constantemente revisada e expandida à medida que o conhecimento biológico se acumula. O GO descreve atividade biológica em relação a três aspectos: função molecular (atividades de nível molecular realizadas por produtos de genes), componente celular (os locais relativos às estruturas celulares em que um produto de gene desempenha uma função), e processo biológico (processos maiores ou "programas biológicos" realizado por múltiplas atividades moleculares) (The Gene Ontology Consortium, 2016).

Na categoria GO “Processo Biológico”, do total de *contigs* analisados, os mais abundantes fazem parte de processos de integração de DNA (15,1%) e na regulação de transcrição com 2,3%. Para os produtos gênicos, um nível mais alto de similaridade indica uma descrição mais específica dos termos da ontologia gênica. Sabe-se que a maior parte dos *contigs* categorizados pelo GO são genes mantenedores da célula que participam do metabolismo primário celular. Isto pode ser explicado pelo fato de que as células especializam-se em funções fisiológicas distintas, determinadas pelo repertório de proteínas associado a cada tipo de metabolismo. Este repertório de proteínas é determinado, por sua vez, pela regulação da expressão dos genes, especialmente em nível de transcrição, por genes mantenedores da célula e geralmente constitutivos (*housekeeping genes*). Estes são necessários para a manutenção de funções celulares basais que são essenciais para a existência de uma célula, os quais são transcritos em todos os tipos celulares, (EISENBERG e LEVANON, 2013).

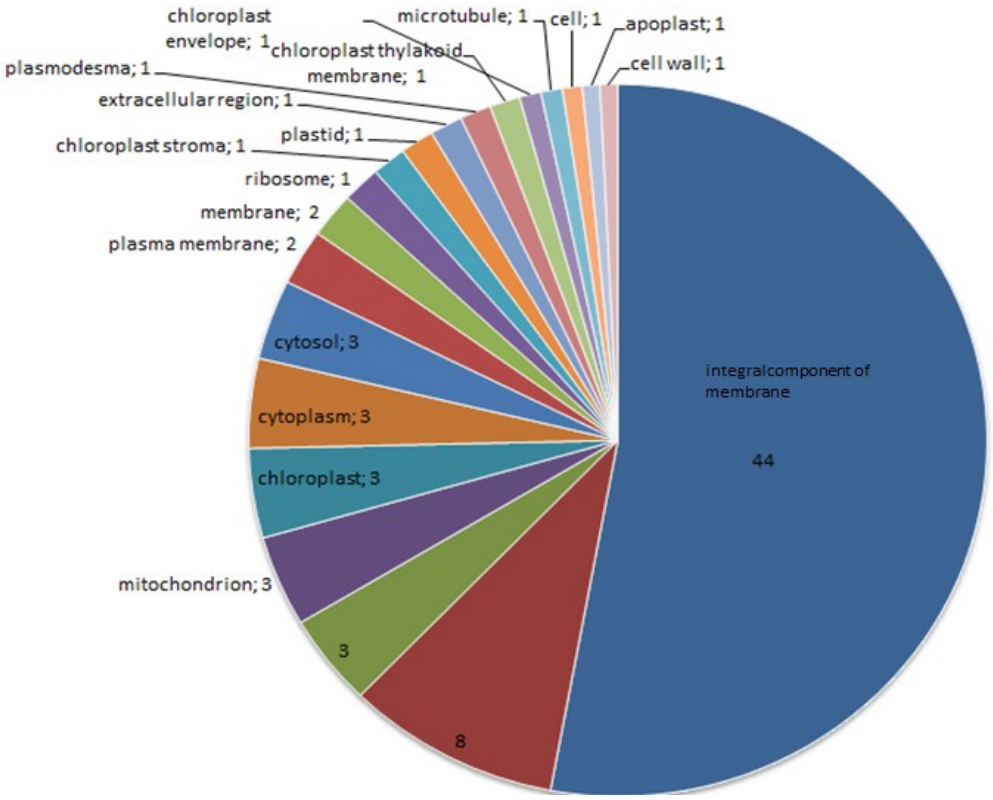


Figura 7- Gráfico dos *contig*s mais abundantes na categoria "Componente Celular" de *C. multijuga* Há|yne. Os valores representam o percentual a partir do número absoluto de 6.249 *contig*s nesta categoria por GO.

Como mostrado na (**Figura 7)**, na categoria "Componente celular" os *contigs* mais abundantes foram relacionados com componentes integrais de membrana (44%) e proteínas associadas ao núcleo celular (7,8%). Esses resultados são similares aos obtidos por KAUR *et al.,* 2012, com outras espécies da família Fabaceae, principalmente os relacionados com genes dos componentes citoplasmáticos, componentes integrais de membrana e de cloroplasto, cruciais para a manutenção da homeostase vegetal.

Na categorização "Função molecular" (**Figura 8**) é possível observar que os *contigs* mais abundantes codificam moléculas ligantes de ácidos nucleicos, correspondendo a 14,6% das sequências anotadas nesta categoria, ligantes de íon zinco (9,9%) e ligantes de ATP com (9,8%). Resultados similares foram descritos por ZWENGER *et al.,* 2010, em trabalhos com transcriptoma em *C. officinalis*, onde demonstrou que 46% dos 535 *contigs* anotados codificam

ligantes (ex. ligantes de nucleotídeos), isto demonstra que em espécies do gênero *Copaifera* estes metabólitos primários são fundamentais na formação de macromoléculas importantes nestas espécies.

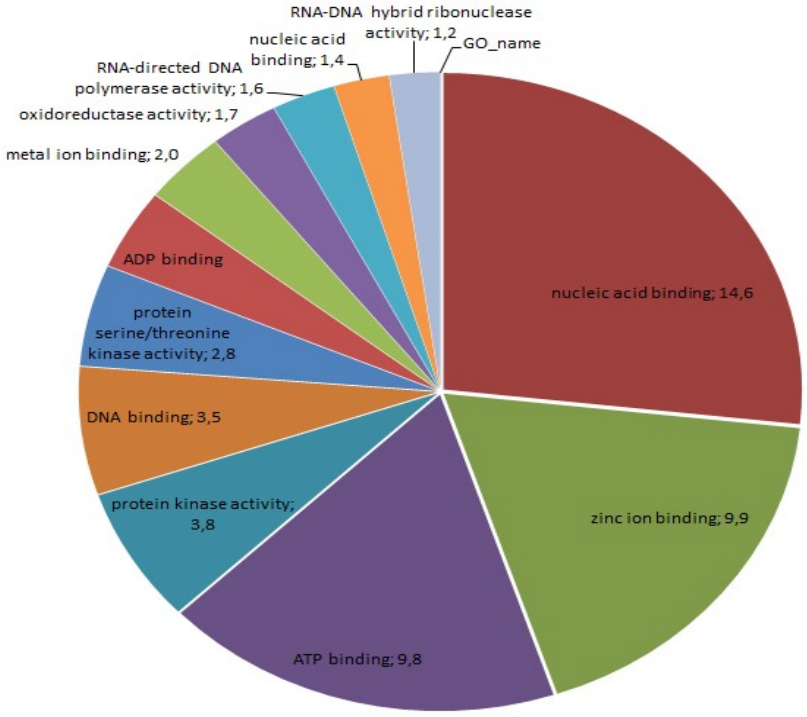


Figura 8- Gráfico mostrando os *contigs* mais abundantes envolvidos na "Função Molecular" de *C. multijuga* Hayne, os valores representam o percentual a partir do número absoluto de 17.208 *contig*s nesta categoria por GO.

Os resultados obtidos por HAO *et al.,* 2012; GALLA *et al.,* 2015; PATEL *et al.,* 2015 e ZHANG *et al.* 2015, utilizando diferentes gêneros de espécies vegetais, também demonstraram resultados similares aos obtidos em *C. multijuga* na "Função Molecular". Vários genes como os do cassete ligante de ATP formam uma classe de proteínas integrantes de uma superfamília que é uma das maiores e mais antigas famílias de proteínas com representantes em todos os filos existentes tanto em procariotos quanto em eucariotos. A capacidade transportadora do cassete de proteínas de ligação ao ATP-(ABC) é surpreendente e vão desde o transporte de íons minerais, lipídios e peptídeos para a regulação dos canais até outras bombas primárias (REA, 2007).

5.4 Metabolismo Secundário de *C. multijuga* Hayne

As plantas são capazes de sintetizar uma variedade enorme de compostos orgânicos de baixo peso molecular, os metabólitos secundários, geralmente com estruturas originais e complexas. Grupos de proteínas que se acumulam nas células em resposta a sinais fisiológicos específicos, determinados por reguladores do crescimento, estresse abiótico ou biótico ou pela disponibilidade de nutrientes, fazem parte do metabolismo secundário (CROTEAU *et al.,* 2000). Os principais metabólitos secundários em plantas são distribuídos em três grupos de acordo com sua rota biossintética: terpenos, compostos fenólicos e compostos contendo nitrogênio (TAYZ e ZEIGER, 2004).

Nas vias do metabolismo secundário de *C. multijuga*, foi mapeado um total de **184** *contigs*, que por terem menor abrangência comparado ao total de genes identificados no metabolismo primário de *Copaifera* foram agrupados separadamente. Na **Figura 9** são mostrados os percentuais mais representativos dos *contigs* evidenciados em *Copaifera*.

Os genes que atuam na síntese de compostos do metabolismo secundário mais representativos foram *contigs* de: carotenóides e tetraterpenos com 8,3%; metabólitos de resposta a UV com 4,8%, flavonóides 3,6%, diterpenos 2,4% e antocianina com 1,2%.

Estes grupos de genes codificadores de metabólitos observados, estão diretamente relacionados com a biossíntese do processo de pigmentação em *Copaifera*, tais como: carotenóide, o qual forma um grupo de pigmentos tetraterpenoides naturais, distribuídos amplamente em plantas, algas, fungos e bactérias, estes desempenham papéis essenciais na fotossíntese e fotoproteção vegetal e estão onipresentes nas membranas de todos organismos fotossintéticos, sendo sua ocorrência já destacada na literatura (DOMONKOS *et al.,* 2013; SUN *et al.*, 2017).

Os resultados obtidos mostram ainda, que os *contigs* observados, em sua grande maioria, são genes de metabólitos secundários diretamente relacionados a respostas ao estresse oxidativo, como os flavonóides que já foram descritos no gênero *Copaifera* (BATISTA *et al.,* 2015). Sabe-se, no

entanto que as plantas produzem mais de 200.000 tipos diferentes de compostos (FIEHN, 2002; TANAKA *et al.,* 2008), incluindo muitos pigmentos que auxiliam na proteção ao estresse, como respostas aos raios UV, sendo as três principais classes de pigmentos para coloração em plantas: flavonóides/antocianinas, betalaina e carotenóides, dos quais somente a classe betalaina, não foi detectada no transcriptoma de *C. multijuga* Hayne.

Os flavonóides, são um grupo de metabólitos secundários que pertencem à classe de fenilpropanilpropanóides, tem a mais larga gama de cores, de amarelo claro ao azul. Em particular, antocianinas, uma classe dos flavonóides que são responsáveis diretos pelas cores laranja ao azul encontradas em muitas flores, folhas, frutos, sementes e outros tecidos. São largamente distribuídos nas plantas de sementes, são solúveis em água, e são armazenadas em vacúolos (TANAKA *et al.,* 2008). Sendo alguns *contigs* (1,2%) formadores de antocianina identificados neste trabalho em *Copaifera*.

Sabe-se que após a clorofila, a antocianina é o grupo mais importante de pigmentos de origem vegetal. Compõem o maior grupo de pigmentos solúveis em água do reino vegetal e são encontradas em maior quantidade nas angiospermas. As antocianinas desempenham funções variadas como ação antioxidante, proteção contra a ação da luz e mecanismos de defesa. São pigmentos que variam do azul ao vermelho e têm papel importante em vários mecanismos reprodutores das plantas, tais como a polinização e a dispersão de sementes (LOPES *et al.,* 2007, TANAKA *et al.,* 2008).

Os metabólitos secundários, amplamente distribuídos em plantas, são classificados em seis grandes subgrupos: chalconas, flavonas, flavonóis, flavandiols, antocianinas e proantocianidinas ou taninos condensados e ainda um sétimo grupo é encontrado em algumas espécies, como auronas. Os flavonóides são uma grande classe estruturalmente diversificada de compostos fenólicos encontrados em todas as plantas superiores (FERRER *et al.,* 2008). Os diferentes flavonóides têm diversas funções biológicas, incluindo a proteção contra a radiação ultravioleta (UV) e fitopatogenos, sinalização durante a nodulação, a fertilidade masculina, do transporte de auxina, bem como a

coloração das flores como um sinal visual para atração a polinizadores (FERREYRA *et al.,* 2012).

De acordo com análise observada dos *contigs* relacionados com o metabolismo secundário em *C. multijuga*, observa-se que *contigs* de uma mesma via podem atuar sinergicamente para exercerem as suas funções biológicas de seu metabolismo secundário, como proteção vegetal em resposta ao estresse oxidativo.

A exposição ao estresse ambiental, muitas vezes resulta na planta um aumento da produção de espécies oxidativas, tais como superóxido (O_2^-), peróxido de hidrogênio (H_2O_2) e oxido nítrico (NO). A habilidade para sobreviver a essas toxinas celulares depende da capacidade de resposta metabólica de mecanismos de desintoxicação. Espécies reativas de oxigênio (ROS) e de nitrogênio tem efeitos diretos e indiretos sobre o estado redox celular e a expressão de diversos genes relacionados ao estresse, incluindo aqueles envolvidos na defesa antioxidante e metabólitos secundários fenólicos (GRACE e LOGAN, 2000).

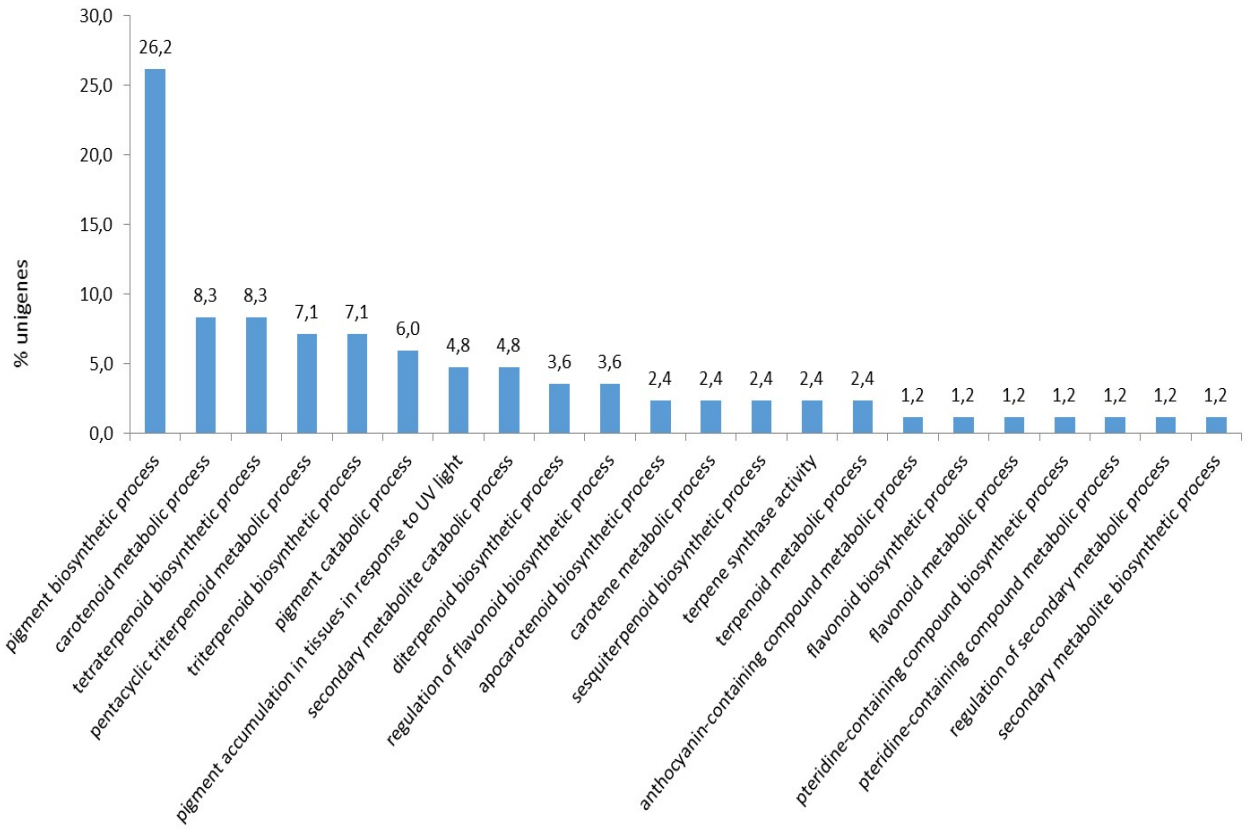


Figura 9- Gráfico que indica a porcentagem de *contig*s relacionados ao metabolismo secundário em *C. multijuga*. O eixo "y" indica o percentual de tipos de *contig*s envolvidos nos processos para formação de metabólitos secundários em *C. multijuga*, a partir do total de (184 metabólitos). O eixo "x" indica os nomes dos *contigs* identificados.

5.5 Terpenos e vias metabólicas em *C. multijuga* Hayne

O maior grupo de *contigs* relacionado ao metabolismo secundário de *C. multijuga*, enquadram-se na classe de terpenos (**Figura 9**). A partir dos dados obtidos de *Copaifera* foram obtidos mapas metabólicos de terpenos e das enzimas identificadas através do banco de dados KEGG. Participantes da via do metabolismo de terpenos estão destacadas na **Tabela 4** e as suas rotas metabólicas estão demostradas nas **Figuras 10**, **11 e 12.** O banco de dados KEGG é amplamente utilizado como uma base de dados de referência das vias metabólicas de um organismo para a integração e interpretação de conjuntos

de dados em grande escala gerados por tecnologia de sequenciamento de alto rendimento (XIE *et al.,*2012).

As vias metabólicas de compostos essenciais na formação do óleo-resina foram desenhadas a fim de detalhar a composição de diterpenos e sesquiterpenos que foram identificados no transcriptoma de *C. multijuga*, corroborando estudos já descritos para composição do óleo-resina do gênero desta espécie. De acordo com HECK *et al.,* 2012 e ALMEIDA *et al.,* 2016, afirmam que óleo essencial de *Copaifera* é constituído por uma parte sólida, os ácidos diterpenos, cerca de 55% a 60%, diluídos em óleo essencial, contendo compostos voláteis, principalmente sesquiterpenos, que são encontrados nas cavidades de *Copaifera* e nos canais resiníferos onde são produzidos. CASCON e GILBERT, 2000 analisaram a composição química do óleo-resina em diferentes espécies de *Copaifera* e observaram que em *C. multijuga*, aproximadamente 80% da oleorresina é composta de sesquiterpenos, enquanto que na *C. guianensis* é de apenas cerca de 44%.

Neste trabalho foram observados *contigs* de enzimas responsáveis pela síntese de compostos terpenóides em *C. multijuga*, destacados na (**Tabela 5**), destacados de acordo com a composição química já descrita para o óleo-resina de acordo com HECK *et al.* (2012) e ALMEIDA *et al.* (2016).

A partir da anotação funcional foi possível identificar *contigs* de enzimas precursoras relacionadas à formação de compostos terpênicos das duas vias de formação. A via padrão do mavalonato (MEV) está presente no citosol de animais, fungos e plantas, e gera os precursores da produção de triterpenos e sequiterpenos. A via alternativa MEP, presente no plastídeo somente em plantas e bactérias, e é a fonte dos precursores dos mono, di e dos tetraterpenóides (ROBERTS, 2007).

Na **Figura 10**, destacam-se os *contigs* identificados nas duas vias como: acetiltransferase (EC. 2.2.3.19 com 10 *reads*) da via MEV; 2-C-metil-D-eritritol-4-fosfato citidililtransferase (EC. 2.2.1.7/ 39 *reads*) precursor da via MEP;

difosfatomevalonato decarboxilase (EC 4.1.1.33/12 *reads*); 4-hidroxi-3-meteilbut-2-em-1-il-difosfato redutase (EC 1.17.1.2/37 *reads*).

Estes resultados possibilitam inferir que o processo para à formação de compostos associados a formação do óleo-resina de *C. multijuga* estão presentes também em plantas jovens e ambas vias de formação estão ativas.

Na porção formada por sesquiterpenos em *C. multijuga* Hayne foram destacados os principais *contigs* expressos na formação do metabolismo secundário bem com o número de *reads* obtidos na montagem transcriptomica (**Tabela 5**). Sendo obtido quase por completa a via metabólica (**Figura 12**) destes *contigs*, os quais destacam-se: δ-cadineno, que apresentou maior número de *reads* observados (281); β-farneseno (188 *reads*); β-cariofileno sintase (75 *reads*), D-germacreno sintase (45 *reads*) e γ-humulene synthase (65 *reads*). Estes resultados são compatíveis com os resultados obtidos por cromatografia gasosa, nos quais destacam que os principais sesquiterpenos encontrados nos óleos de copaíbas adultas e jovens são: β-cariofileno, β-farneseno, óxido de cariofileno, α-humuleno e δ-cadineno (VEIGA-JUNIOR e PINTO, 2002; BASTOS, 2011; LEANDRO, 2012). ALMEIDA *et al.,* 2016 também analisou por cromatografia o óleo de folhas de *C. langsdorfii* e destacou estes compostos, ressaltando principalmente a composição dos sesquiterpenos, D-germacreno e β-cariofileno.

BASTOS (2011) já havia identificado por cromatografia gasosa a presença do composto δ-cadineno, em 5,71%, no óleo extraído a partir de folhas de *C. multijuga*, assim como CASCON e GILBERT (2000) que também em estudos com *C. multijuga* identificaram a presença em 2,9% deste composto no óleo puro. Estes resultados sugerem que os *contigs* para formação deste composto possivelmente são comumente expressos em *C. multijuga*, uma vez que enzimas δ-Cadineno sintases (DCS) são sesquiterpeno-ciclases que catalisam a ciclização do farnesil difosfato para formação de fitolaxinas importantes para proteção vegetal contra bactérias e fungos, dados já comprovado em algodão (CHEN *et* al., 1995; GENNADIOS *et a*l., 2010).

CHEN *et al.*, 2009, a partir de estudos com *C. officinallis*, para averiguar composição de sesquiterpenos do óleo em plantas jovens e adultas, identificou, por cromatografia gasosa, um total de 26 sesquiterpenos a partir de tecido foliar, 25 sesquiterpenos de tecido de caule e somente 2 sesquiterpenos detectados a partir tecidos radiculares. O composto D-germacreno foi evidenciado somente em oléo extraído a partir de folhas e caule, com maior abundancia em folhas de plantas jovem representando 81,4% da composição do óleo. Já o composto β-cariofileno foi o único composto detectado a partir das três composições vegetais analisadas.

Dos *contigs* formadores de compostos encontrados em *C. multijuga* Hayne, é importante destacar a atividade do composto β-cariofileno, sabe-se que a este é atribuída a atividade inti-inflamatória descrita para o óleo-resina (VEIGA JUNIOR *et al.,* 2007; BARBOSA *et al.,* 2012; LUCCA *et al.,* 2015).

No que se refere a *contigs* de enzimas de compostos da porção resinosa de *C. multijuga*, foram identificados genes de quatro enzimas precursoras responsáveis pela síntese de diterpenos (**Tabela 5**), tais como: ent-copalil difosfato sintase (89 *reads*), casbano sintase (45 *reads*), ácido ent-caurenóico hidroxilase (35 *reads*) e 3-beta-dioxigenase giberelina (15 *reads*), a via metabólica desta composição é destacada na (**Figura 12**).

Destes *contigs* destaca-se o que codifica a *ent*-copalil difosfato sintase, pois sabe-se que esta enzima catalisa o primeiro passo na biossíntese de giberelina (IRMISCH *et al.,* 2015), um importante hormônio vegetal que regula muitos aspectos do crescimento e desenvolvimento durante o ciclo de vida das plantas (ITO *et al*, 2017). É importante destacar também, a presença do *contig* de formação da enzima casbano sintase, uma vez que esta enzima catalisa a ciclização do geranilgeranil difosfato para o casbano, uma fitoalexina diterpênica com atividade antibacteriana e antifúngica (HILL *et al*., 1996), possivelmente é devido a presença desse composto que o óleo-resina da copaíba possui atividade antimicrobiana e antifúngica (HECK *et al*., 2012).

Tabela 5- Principais *contig*s de enzimas componentes da formação do óleo-resina evidenciadas em *C. multijuga*.

Terpenos em *C. multijuga* Hayne				
EC	**Nome**	***Reads/contig***	**Identidade**	***E-value***
1.1.1.34	Hydroxymethylglutaryl CoA reductase	40	41.43-93.59	3e-34-0.097
1.17.1.2/1.17.7.4*	4-hydroxy-3-methylbut-2-en-1-yl diphosphate reductase	37	40.54-97.32	1e-72-0.100
2.2.1.7	2-C-methyl-D-erytriotol 4-phosphate cytidylyltransferase	39	43.94-96.72	2e-33-0.098
2.3.1.9	Acetyltransferase	10	46.51-84.00	5e-17-0.062
2.5.1.1	Heptaprenyl synthase diphosphate	8	55.88-68.75	0.009-0.094
2.5.1.30	heptaprenyl diphosphate synthase	16	42.86-78.57	0.010-0.091
2.5.1.31	undecaprenyl diphosphate synthase	12	42.31-85.71	0.007-0.084
2.5.1.84	Undecaprenyl diphosphate synthase	28	39.58-81.25	0.004-0.099
2.7.1.148	Isoprene synthase	21	41.86-85.71	9e-06-0.096
2.7.7.60	4-diphosphocytidyl-2-C-methyl-D erythriotol kinase	24	44.00-90.91	1e-06-0.095
4.1.1.33	Diphosphomevalonate decarboxylase	12	46.15-73.68	0.031-0.095
4.2.3.27	α-trans-nonaprenyl-diphosphate synthase	102	38.33-92.31	1e-25-0.097
1.14.13.77	13-α-hidroxylase	4	55.88-70.59	0.040-0.081
Diterpenos em *C. multijuga* Hayne				
EC	**Nome**	***Reads/contig***	**Identidade**	**E-value**
5.5.1.13	Ent-copalyl diphosphate synthase	89	40.23- 6.84	1e-39-0.099
4.2.3.8	Casbene synthase	45	42.47-7.50	1e-13-0.099
1.14.13.79	ent-kaurenoic acid hydroxylase	35	48.48-2.31	1e-33-0.098
1.14.11.15	Gibberelin 3-beta-dioxygenase	15	52.00-1.43	5e-08-0.095
Sesquiterpenos em *C. multijuga* Hayne				
EC	**Nome**	***Reads/contig***	**Identidade**	**E-value**
4.2.3.46	α-farnesene synthase	65	39.18-92.31	2e-19-0.099
4.2.3.47	β-farneseno	188	41.51-92.86	3e-08-0.098
4.2.3.22	Germacrene-D synthase	45	42.99-83.33	0.009-0.096
4.2.3.23	germacrene-A synthase	34	37.80-80.00	0.003-0.097
4.2.3.13	δ-Cadinene	281	44.07 -81.58	1e-36-0.0097

4.2.3.57/ TPS21	β-caryophyllene synthase	75	41.51-92.86	2e-06-0.100
4.2.3.56	γ-humulene synthase	65	42.86-77.78	0.002-0.099
4.2.3.39	epi-cedrol synthase	35	41.67-75.00	0.012-0.098
4.2.3.9	aristolochene synthase	17	43.10-71.43	0.006-0.085
TPS1	Valencene (EC:4.2.3.73)	34	43.14-75.61	4e-27-0.093
4.2.3.21	vetispiradiene synthase	7	45.16-56.25	8e-04-0.095

*** EC. 1.17.1.2 obsoleto no banco KEGG modificado para 1.17.7.4***

Na literatura os compostos mais citados na composição da fração resinosa a partir do óleo puro de copaiferas são: os ácidos copálico, polialtico, caurenóico e ent-caurenóico, juntamente com os seus derivados, os ácidos: 3-hidróxi-copálico, 3-acetóxi-copálico, e ent-agático (CASCON e GILBERT, 2000; VEIGA-JUNIOR e PINTO, 2002; VEIGA-JUNIOR, 2007). No transcriptoma de *C. multijuga* não foi possível identificar *contigs* relacionados com a formação destes compostos, possivelmente devido o fato da análise transcriptomica não ter sido completa o suficiente para detectar estas sequências, ou as enzimas responsáveis pela síntese destes compostos resinosos são produzidas em outras partes da planta, como por exemplo casca do tronco ou em outra fase de desenvolvimento da planta. Sabe-se também, que a composição química do óleo, cor e viscosidade podem ser variadas entre as espécies de *Copaiferas* e as regiões onde estas espécies se encontram (VEIGA JUNIOR *et al.,* 2002; PLOWDEN, 2003; ALMEIDA *et al.,* 2016). MEDEIROS e VIEIRA (2008) a partir de análises de *C. multijuga*, demonstraram que fatores abióticos e bióticos contribuíram para produção de óleo-resina (por exemplo, térmicos, idade e tamanho da árvore).

Genomas de plantas possuem famílias gênicas relacionados à síntese destes compostos secundários, as quais codificam enzimas que podem diferir para cada linhagem de plantas, como a família de genes terpeno sintases TPS (CHEN *et al.,* 2011). Em *C. multijuga* Hayne *contigs* codificadores destas enzimas foram identificados os quais possibilitaram a construção de vias metabólicas relacionadas à composição do óleo-resina desta espécie.

Sabe-se que o grupo de terpenóides constitui a maior família de produtos naturais, dos quais mais de 22.000 compostos individuais já foram descritos, e o número de estruturas definidas duplicou a cada década desde os anos 1970 (MCGARVEY e CROTEAU, 1995). Apesar da variação na composição química dos óleos-resinas de *Copaifera* em geral, o β-cariofileno, é o principal constituinte da fração volatil e o ácido copálico o principal constituinte da fração resinosa. Estes compostos são considerados os marcadores químicos dos óleos-resinas da copaiba (VEIGA-JUNIOR *et al.,* 1997; GILBERT, 2000). PIERI *et al.,* (2009) sugerem que o ácido copálico, encontrado em todos os óleos-resinas de copaíba já estudados, possa vir a ser utilizado como um marcador da autenticidade dos óleos vegetais.

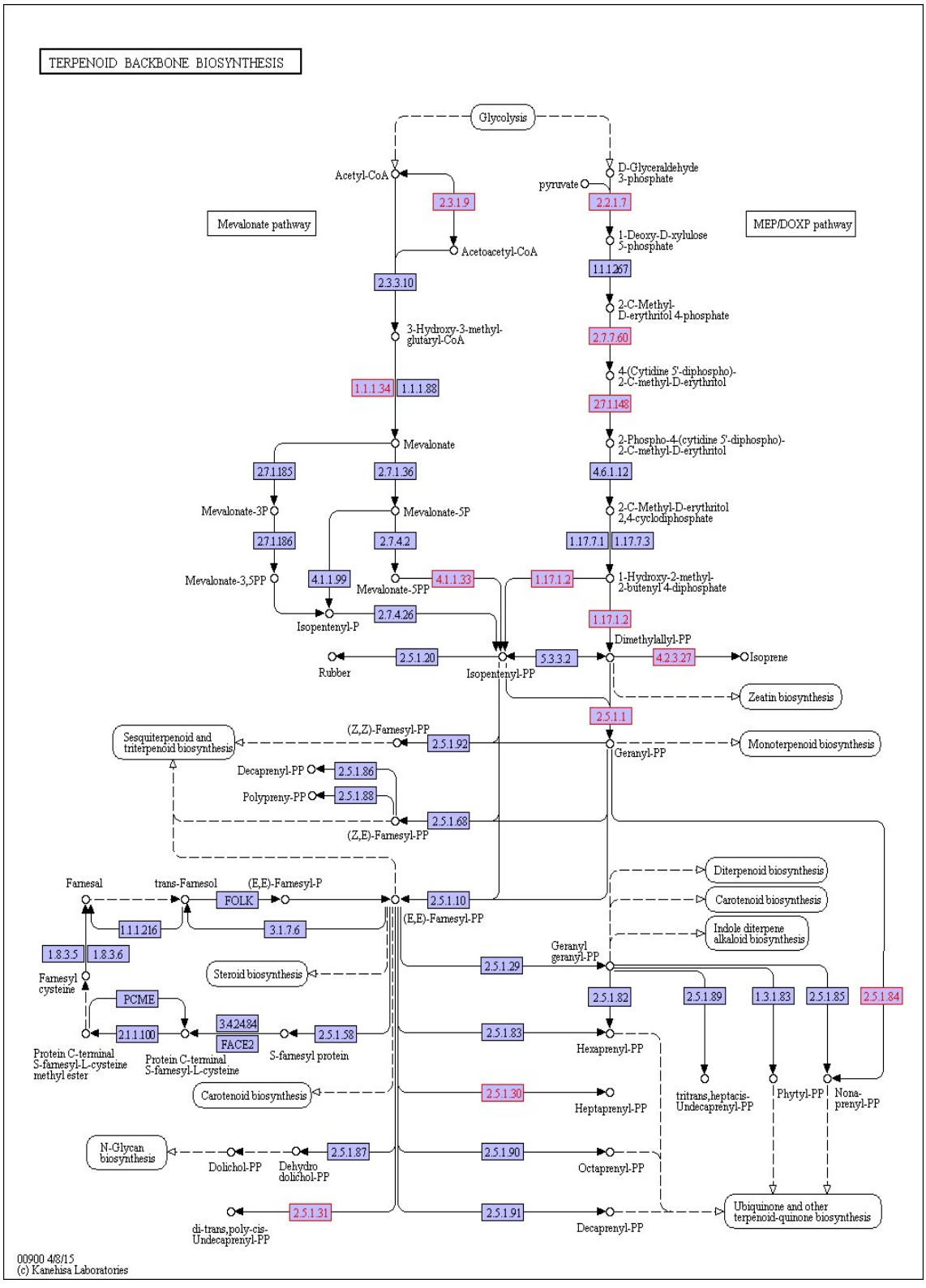


Figura 10- Esquema representando a via metabólica de terpenos geradas em programa KEGG. Os códigos enzimáticos (ECs) componentes da via e identificados em *C. multijuga*, estão destacados em rosa.

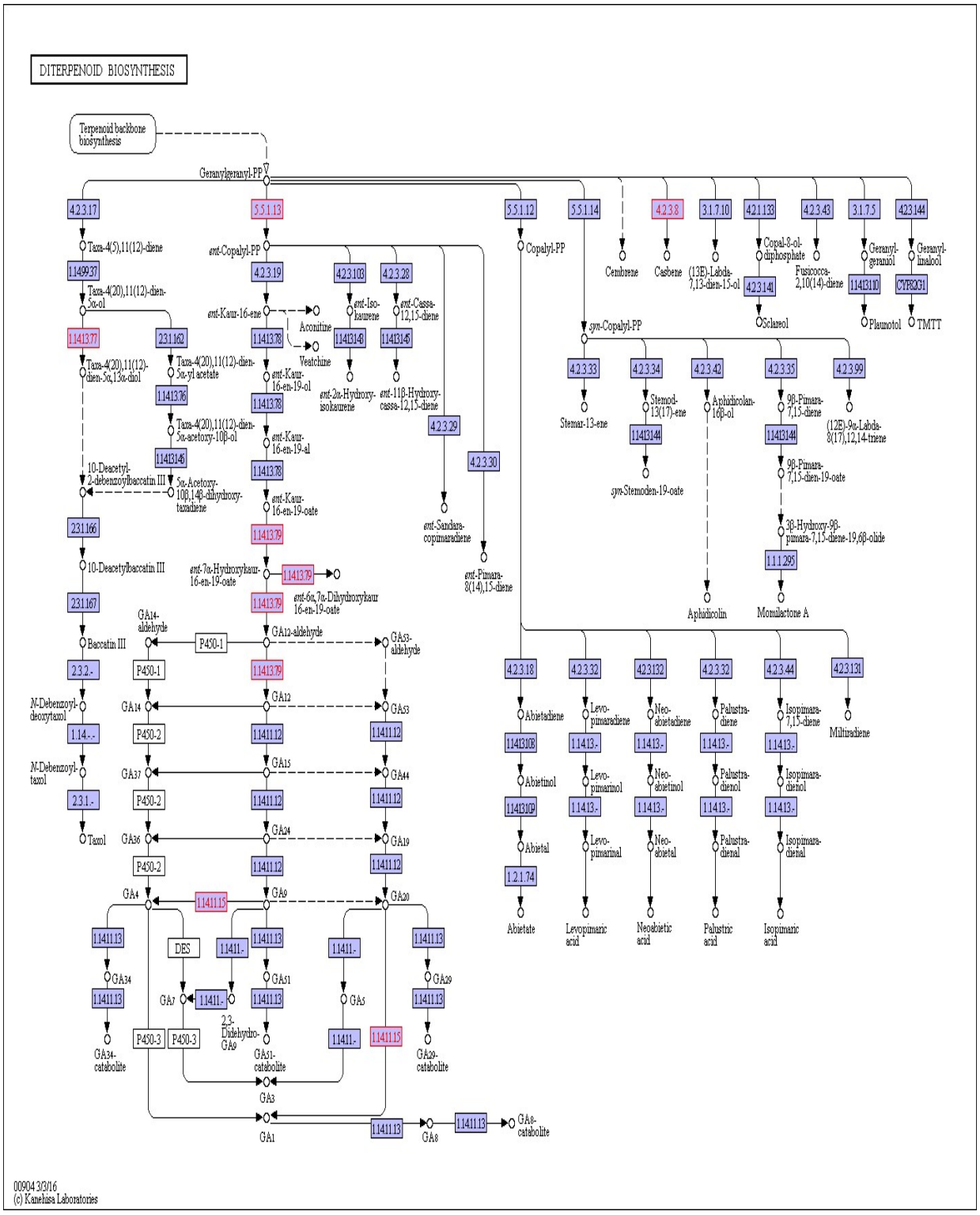


Figura 11-Esquema representando a via metabólica de diterpenos (20 carbonos) geradas em programa KEGG. Os códigos enzimáticos (ECs) componentes da via e identificados em *C. multijuga*, estão destacados em rosa.

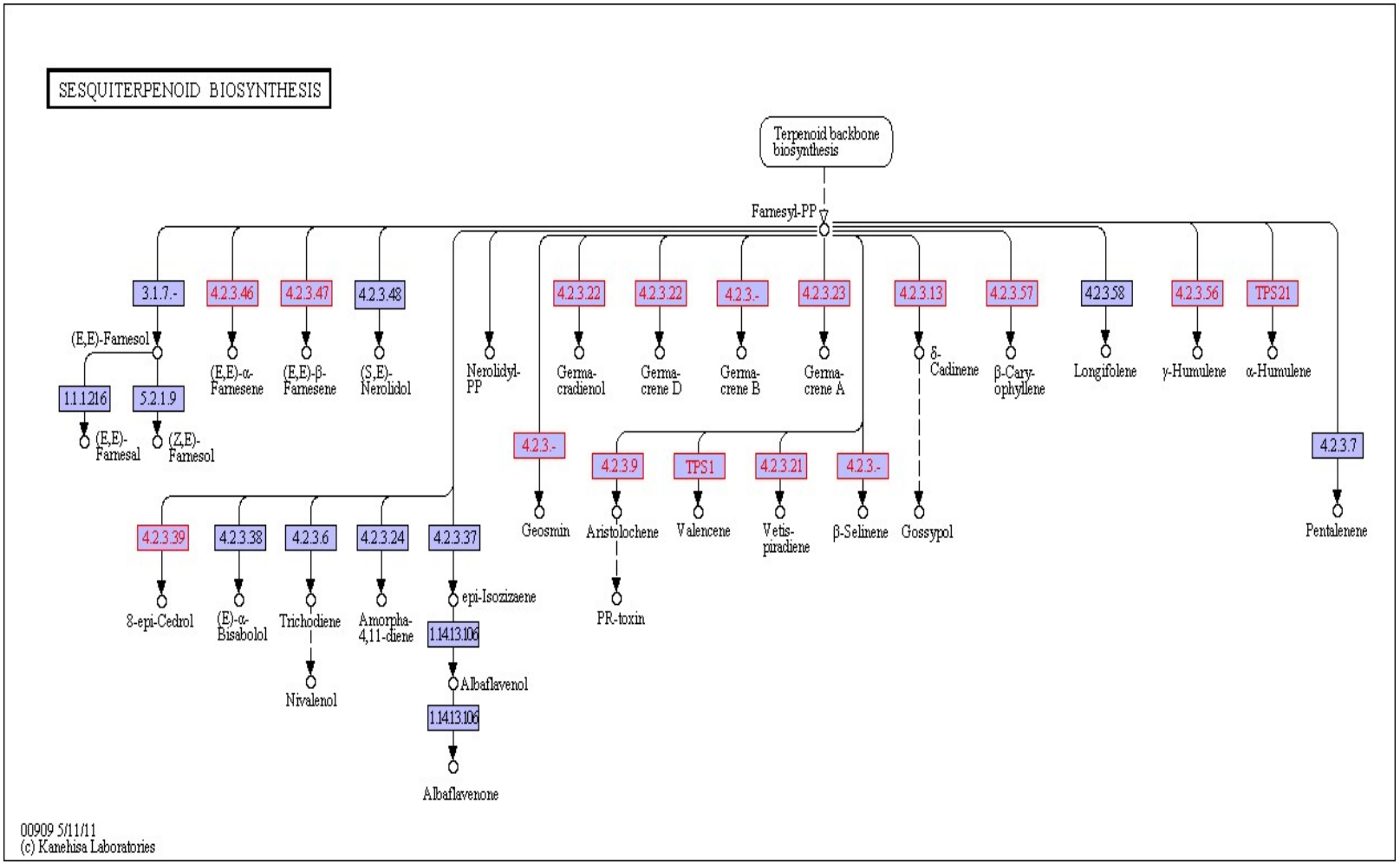


Figura 12- Esquema representando a via metabólica de sesquiterpenos (15 carbonos) geradas em programa KEGG. Os códigos enzimáticos (ECs) componentes da via e identificados em *C. multijuga*, estão destacados em rosa.

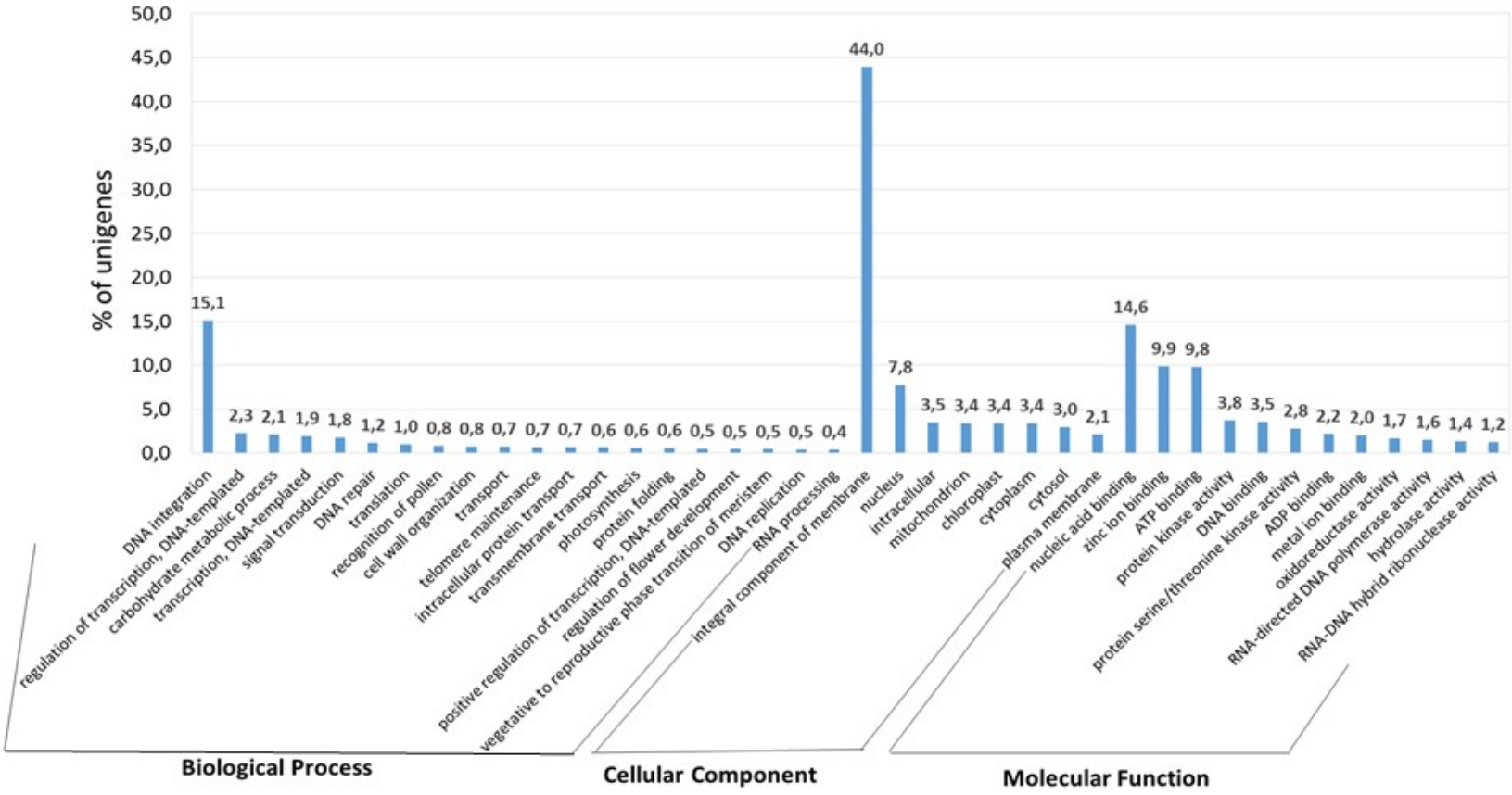


Figura 13- Resultados sumarizados dos metabólitos secundários mais abundantes *em C. multijuga* divididos em três categorias de acordo com o Gene Ontology: Biological Process (BP) Cellular Component (CC) e Molecular Function (MF). O eixo "y" do gráfico indica o percentual do número de genes na categoria G.O. O valor absoluto total dessa categoria para BP= 6.249 *contig*s; CC= 9.499 *contig*s; MF= 17.208 *contig*s. O eixo "x" indica os nomes das *contigs* mais abundantes.

5.6 Filogenia de terpenos sintases de *C. multijuga* Hayne

Para determinação do grau de proximidade genética dos *contigs* de terpenos sintases, encontradas no banco de *C. multijuga*, com outras espécies do gênero *Copaifera* disponíveis no NCBI, foi montada uma árvore filogenética baseada no método de Neighbor-Joining (SAITOU e NEI, 1987) mostrada na **Figura 14**, baseada na distância genética entre diferentes sequências de espécies do gênero *Copaifera: C. offinalis, C. langsdorfii* e *C. multijuga*.

Os resultados obtidos indicam características comuns da família de genes terpenos sintases entre as espécies do mesmo gênero, sendo o *contig* de TPS4-2 observado no dendograma filogenético com valor *bootstrap* do clado de 96%, indicando assim proximidade evolutiva (comparadas a nível de nucleotídeos) à outras sequências de terpenos TPS4-2 de *Copaifera* disponíveis no banco NCBI. Porém as sequências de terpenos sintases (TPS-1; TPS-2; TPS3 e TPS5), não foram agrupadas com as sequências de terpenos sintases disponíveis no banco NCBI, formando assim grupamentos separados dos ramos de TPS do banco. Esse resultado e os números baixos de bootstrap podem ser devido ao fato dessas sequências de terpenos sintases não serem completas.

O fato dos genes de terpenos sintases 4 terem agrupado em um só clado com alto grau de confiabilidade (*bootstrap* de 96%), evidencia o alto grau de similaridade genética. Isso nem sempre ocorre em famílias gênicas pois em alguns casos conjuntos de genes relacionados que codificam enzimas que utilizam substratos semelhantes e produzem produtos similares, têm suas sequências claramente divergentes (CHEN *et al*. 2011).

Genes de enzimas terpenos sintases já foram caracterizados em muitas espécies vegetais como milho (KOLLNER *et al.,* 2004); tomateiro (FALARA *et al.,* 2011*)*; soja (ZHANG *et al.,* 2013; LIU *et al.,* 2014) e em eucalipto (KÜLHEIM *et al.,* 2015), porém não há descrição de estudos envolvendo o gênero *Copaifera*. Deste gênero, ZWENGER *et al.,* 2010 por sequenciamento (ESTs) de folhas de *C. officinalis* jovem, anotaram um total de 613 *contigs*,

principalmente relacionados a respostas de calor, porém não relataram nenhuma seqüência correspondente ao metabolismo de terpenos sintases.

Sabe-se que pesquisas envolvendo a família das terpeno sintases são relativamente recentes. Alguns poucos trabalhos se destacam na caracterização desta família como o de LIU *et al*., 2014, no qual mostrou que a família TPS de soja (GmTPSs) é composta por mais de 20 membros.

Como estudos envolvendo abordagem transcriptômica de espécies amazônicas são escassos, especialmente as que envolvem pesquisas com genes de terpenos sintases, pode-se dizer que este é o primeiro trabalho voltado para caraterização desses genes a partir de planta uma do bioma amazônico.

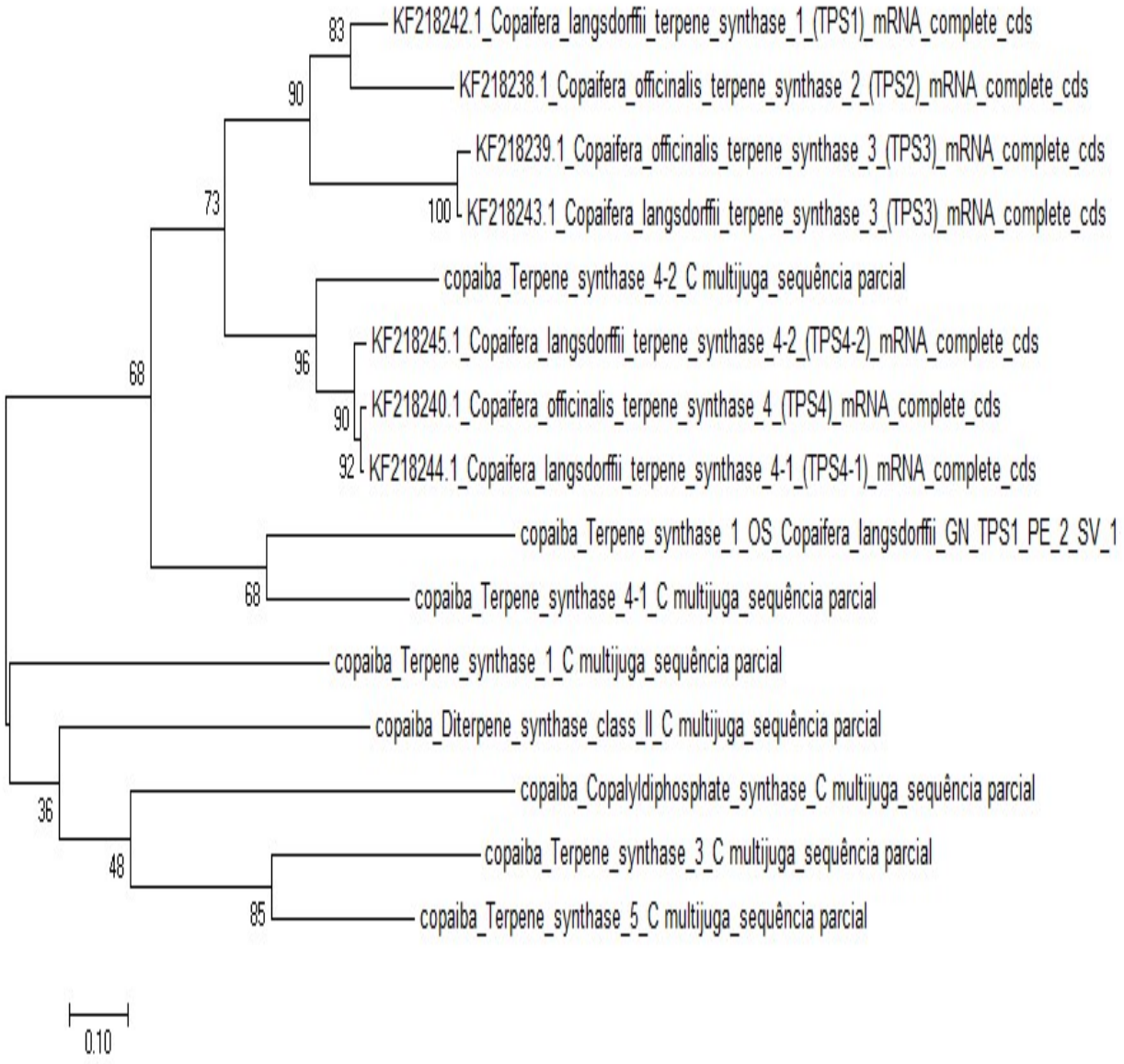


Figura 14- Árvore filogenética de genes de terpeno sintases de diferentes espécies de copaíba.Relações evolucionários dos táxons de *Copaifera.* A história evolutiva foi inferida usando o método de agrupamentos vizinhos Neighbor-Joining (SAITOU e NEI, 1987). A árvore ideal com a soma do comprimento do ramo = 5.25604012 é mostrada.

Conclusões

6- CONCLUSÕES

- Neste trabalho foi possível montar o transcriptoma de folhas de planta jovem *C. multijuga* a partir de 638.576 *reads*, gerando-se um total de 11.050 *contigs*.
- Foi possível identificar muitos *contigs* importantes envolvidos no metabolismo primário de *C. multijuga* Hayne, muitos dos quais já são conhecidos em outras espécies vegetais, como os genes *housekeeping.*
- Em relação ao metabolismo secundário foi possível evidenciar diversos *contigs* envolvidos na composição do óleo-resina, destancando genes associados ao estresse oxidativo vegetal (flavonóides e antocianinas) e genes envolvidos na formação dos compostos terpênicos, sesqui e diterpênicos, promissores para pesquisas biotecnológicas por serem potenciais antitumorais, anti-oxidadantes, anti-inflamatórios.
- No metabolismo secundário no que se refere à síntese de terpenos evidenciou-se que em plantas jovens tanto a vias do mevalonato (MEV) como a do Metileritritol-fosfato (MEP) estão ativas.
- Na comparação filogenética de *C. multijuga* com outras espécies de *Copaifera* o *contig* de TPS-4-2 destacou-se como mais conservado geneticamente, sendo alvo promissor para clonagem e expressão.

Referências

7- REFERÊNCIAS

AJIKUMAR, P. K; *et al.* Isoprenoid Pathway Optimization for Taxol Precursor Overproduction in *Escherichia coli. Science. 330 (6000): 70–74.* 2010.

AUBOURG, S.; LECHARNY A.; BOHLMANN J. Genomic analysis of the terpenoid synthase (AtTPS) gene family of Arabidopsis thaliana. *Molecular Genetics Genomics.* 267(6):730-45. 2002.

ALMEIDA, L.F.R. *et al.* Non-Oxygenated Sesquiterpenes in the Essential Oil of Copaifera langsdorffii Desf. Increase during the Day in the Dry Season, *Plos One*, 1-12, 2016.

ALTSCHUL, S.F; MADDEN, T.L; SCHÄFFER, A.A; ZHANG, J.; ZHANG, Z.; MILLER, W.; LIPMAN D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research.* 1; 25(17):3389-402, 1997.

ADAMCZYK, B.; ADAMCZYK, S; SMOLANDER, A.; KITUNEN, V.; SIMON, J. Plant Secondary Metabolites-Missing Pieces in the Soil Organic Matter Puzzle of Boreal Forests. *Soils*, 2-10, 2018.

BARBOSA, P.C.S.; MEDEIROS, R.S.; SAMPAIO, P.T.B.; VIEIRA, G.; WIEDEMANN, L.S.M.; VEIGA JUNIOR, V.F. Influence of Abiotic Factors on the Chemical Composition of Copaiba Oil (*Copaifera multijuga* Hayne): Soil Composition, Seasonality and Diameter at Breast Height. *Journal of the Brazilian Chemical Society*.Vol.23, No. 10, 1823-1833, 2012.

BARDAJÍ, D.K; et al. Copaifera reticulata oleoresin: Chemical characterization and antibacterial properties against oral pathogens. *Anaerobe*, 40:18-27, 2016.

BARRETO-JUNIOR, *et al.* Cromatografia de troca-iônica aplicada ao isolamento da fração ácida do óleo de copaíba (*Copaifera multijuga*) E DA SACACA (*Croton cajucara*). *Química Nova*, Vol. 28, No. 4, 719-722, 2005.

BASTOS, A.P.M. R. *Análise cromatográfica, morfológica e molecular da síntese do oleoresina em plantas jovens de Copaifera multijuga Hayne (Fabaceae – Caesalpinioideae).* Tese de doutorado do Programa de Pós-Graduação Multi-Institucional em Biotecnologia da Universidade Federal do Amazonas, 2011. 67p.

BAKER, M. *De novo* Genome Assembly: What Every Biologistshould Know. *Nature Methods*.Vol.9 N.4, 333-337, 2012.

BATISTA, A. G. *et al.* Polyphenols, antioxidants, and antimutagenic effects of Copaifera langsdorffii fruit. *Food Chemistry* XXX, 2015.

BIONDO, E.; MIOTTO, T.S.; SCHIFINO-WITTMANN, M.T. Citogenética de espécies arbóreas da subfamília Caesalpinioideae –Leguminosae do Sul do Brasil.*Ciência Florestal*, vol. 15, n. 3, p. 241-248 241. 2005.

BOLGER, M. E.; ARSOVA, B.; USADEL, B. Plant genome and transcriptome annotations: from misconceptions to simple solutions. *Briefings in Bioinformatics*, 1–13, 2017.

BOLTON, M. D. Primary Metabolism and Plant Defense—Fuel for the Fire. *Molecular Plant-Microbe Interactions,* 5: 487–497, 2009.

BOURGAUD, F.; MILESI, A.; GRAVOT, S.; GONTIER. E. Production of plant secondary metabolites: a historical perspective. *Plant Science* 161: 839–851, 2001.

CARNEIRO, N.P.; CARNEIRO, A.A.; GUIMARÃES, C.T.; PAIVA, E. Desvendando o Código Genético. *Biotecnologia Ciência e Desenvolvimento*, Vol.17, p.50-58, 2000.

CARVALHO, E.C. *Identificação fenotípica e molecular de bactérias patogênicas associadas à criação de peixes amazônicos.* Programa de Pós-graduação em Genética Conservação e Biologia Evolutiva, Instituto

Nacional de Pesquisas da Amazônia. Dissertação de mestrado. 120p, 2012.

CARVALHO, M. C. D. C. G. DE; SILVA, D. C. G. D. Sequenciamento de DNA de nova geração e suas aplicações na genômica de plantas. *Ciência Rural*, Santa Maria, v.40, n.3, p.735-744, 2010.

CLARKE, K.; YANG, Y.; MARSH, R.; XIE, L.; ZHANG, K. K. Comparative analysis of *de novo* transcriptome assembly. Vol.56 No.2: 156–162, 2013.

CONESA, A.; GÖTZ,S.; GARCÍA-GÓMEZ, J.M; TEROL, J.; TALÓN, M.; ROBLES, M. BLAST2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics Applications Note.* Vol. 21(18), 3674–3676, 2005.

CHEN, F.; AL-AHMAD, H.; JOYCE, B.; ZHAO, N.; KOLLNER, T.G.; DEGENHARDT, J.; JR. C.N. S. Within-plant distribution and emission of sesquiterpenes from *Copaifera officinalis*. *Plant Physiology and Biochemistry* 47, 1017–1023, 2009.

CHEN, F; THOLL, D.; BOHLMANN, J.; PICHERSKY, E. The Plant Genome: En evolutionary View on Structure and Function. The family of Terpene Synthases in plants: A mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. *The Plant Journal*, 66, 212–229, 2011.

CHOI, J.H.; KEUM, K.C.; LEE, S.Y. Production of recombinant proteins by high cell density culture of *Escherichia coli*. *Chem Eng Sci*, 61:876—885, 2006.

CROTEAU, R.; KUTCHAN T.M, LEWIS. Natural Products (Secondary Metabolites). *In*: Buchanan B., Gruissem W., Jones R. (Eds.)*Biochemistry e Molecular Biology of Plants*, Rockville: *American Society of Plant Physiologists*, p.1250-1318, 2000.

DELSENY, M.; HAN, B.; HSING, Y.I. High throughput DNA sequencing: The new sequencing revolution. *Plant Science*. Vol. 179, p.407–422, 2010.

DOMONKOS, I.; KIS, M.; GOMBOS, Z.; UGHY, B. Carotenoids, versatile components of oxygenic photosynthesis. *Progress in Lipid Research*, p. 52 539–561, 2013.

DU, H. *et al*. Sequencing and *de novo* assembly of a near complete *indica* rice genome. *Nature Comunication*, 1-12, 2017.

DUDAREVA, N.; NEGRE, F.; NAGEGOWDA, D.A; ORLOVA, I. Critical Reviews in Plant Sciences, Plant Volatiles: Recent Advances and Future Perspectives. *Critical Reviews in Plant Sciences*, 25: 417–440, 2006.

EISENBERG, E; LEVANON, E.Y. Human housekeeping genes, *Trends in Genetics,* V. 29(10): 569- 574, 2013.

EFRON, B.; HALLORAN, E.; HOLMES, S. Bootstrap confidence levels for phylogenetic trees. *Proc. Natl. Acad. Sci.* USA 93, 13429-13434. 1996.

FALARA, V. *et al*. The tomato terpene synthase gene family. Plant Physiology.157(2):770-89, 2011.

FALCÃO, L. D. *et al.,* Copigmentação intra e intermolecular de antocianinas: uma revisão. Boletim CEPPA, Curitiba, v. 21, n.2, 351-366, 2003.

FERRER, J.L; AUSTIN, M.B., STEWART, C.J.; NOEL, J.P. Structure and function of enzymes involved in the biosynthesis of phenylpropanoids. *Plant Physiology and Biochemistry,* 46, 356e370, 2008.

FERREYRA, M.L.F.; RIUS, S.P.; CASATI, P. Flavonoids: biosynthesis, biological functions, and biotechnological applications. *Frontiers in Plant Science.* Vol.3. 1-15. 2012.

FERNANDES, L.A, *Montagem e anotação funcional de sequências gênicas de Handroanthus impetiginosus (Mart. ex DC.) Mattos*. Dissertação, Universidade Federal de Goiás. Instituto de Ciências Biológicas. Programa de Pós-Graduação em Genética e Biologia Molecular. 67 p. 2015.

FIEHN, O. Metabolomics--the link between genotypes and phenotypes. Plant Moleculas Biology. ;48(1-2):155-71. 2002.

FUMAGALI, E.; GONÇALVES, R.A.C.; SILVA, M.F.P; VIDOTI, G.J.; OLIVEIRA, A.J B de. Produção de metabólitos secundários em cultura de células e tecidos de plantas: O exemplo dos gêneros *Tabernaemontana* e *Aspidosperma*. *Revista Brasileira de Farmacognosia.* 18(4): 627-641, 2008.

GAETA, M.L. *Distribuição de DNA repetitivo nos cromossomos de Copaifera langsdorffii Desf. (Caesalpinioideae): Uma importante árvore produtora de óleos essenciais.* Dissertação de mestrado. Universidade Estadual de Londrina. 46p.2009.

GALLA, G.; VOGEL, H.; SHARBEL, T.F.; BARCACCIA, G. *De novo* sequencing of the Hypericum perforatum L. flower transcriptome to identify potential genes that are related to plant reproduction sensu lato. *BMC Genomics*, 16:254, 2015.

GARCIA-SECO, D.; ZHANG, Y. GUTIERREZ-MAÑERO, F.J; MARTIN, C.; RAMOS-SOLANO, B. RNA-Seq analysis and transcriptome assembly for blackberry (Rubus sp. Var. Lochness) fruit. *BMC Genomics* 16:5, 2015.

GERIS, R.; SILVA, I.G.; SILVA, H.H.G.; BARISON, A.; RODRIGUES-FILHO, E.; FERREIRA, G. DITERPENOIDS FROM *Copaifera reticulata* DUCKE WITH LARVICIDAL ACTIVITY AGAINST *Aedes aegypti* (L.) (DIPTERA, CULICIDAE). *Revista do Instituto de Medicina Tropical de São Paulo*. Vol. 50 (1):25-28,2008.

GOMES, N. M.; REZENDE, C.M.; FONTES, S.P.; MATHEUS, M.E.; FERNANDES, P.D. Antinociceptive activity of Amazonian Copaiba oils. *Journal of Ethnopharmacology*. Vol.109: 486–492, 2007.

GONÇALVES, E.; SILVA, J. R.; GOMES, C.L. NERY, M.B.L.; NAVARRO, D.M.A.F.; SANTOS, G. K.N.; SILVA-NETO, J.C.; COSTA-SILVA, J.H.; ARAÚJO, A.V.; WANDERLEY, A.G. Effects of the oral treatment with

Copaifera multijuga oil on reproductive performance of male Wistar rats. *Revista Brasileira de Farmagnosia.* Vol. 24.p. 355-362. 2014.

GOOSSENS, A. *et al.,* A functional genomics approach toward the understanding of secondary metabolism in plant cells. 2003. *PNAS, Proceedings of the National Academy of Sciences.* Vol. 100 (14): 8595–8600, 2003.

GRAMOSA, N. V; SILVEIRA, E.D. Volatile Constituents of *Copaifera langsdorffii* from the Brazilian Northeast. *Journal of Essential Oil Research*, 17, 130-132, 2005.

GRACE, S.C., LOGAN, B.A., Energy dissipation and radical scavenging by the plant phenylpronanoid pathway. *Phiosophical. Transactions Royal Society B*. Vol. 355, 1499–1510. 2000.

HAMILTON; JP, BUELL CR. Advances in plant genome sequencing. Plant J. 70(1):177-90, 2012.

HATTAN J., SHINDO K., ITO T., SHIBUYA Y., WATANABE A., TAGAKI C., *et al.* Identification of a novel hedycaryol synthase gene isolated from Camellia brevistyla flowers and floral scent of Camellia cultivars. Planta 243, 959–972, 2016.

HILL, A.; CANE, D.E.; MAU, C.J.D.; WEST, C.A. High Level Expression of Ricinus communis Casbene Synthase in Escherichia coli and Characterization of the Recombinant Enzyme. *Archives of Biochemistry and Biophysics*. Vol. 336, No. 2, 15, pp. 283–289, 1996.

HILLIS, DAVID M.; BULL, JAMES J. An Empirical Test of Bootstrapping as a Method for Assessing Confidence in Phylogenetic Analysis. *Systematic Biology*, 42 (2): 182-192. 1993.

HOLOPAINEN, J. K.; KIVIMÄENPÄÄ, M. JULKUNEN-TIITTO, R. New Light for Phytochemicals. *Trends in Biotechnology*, Vol. 36, No. 1, 2018.

HAO, D.C.; MA, P.; UM, J.; CHEN, S.L.; XIAO P.G.; PENG Y.; HUO, L., XU, L.J.; SUN, C. *De novo* characterization of the root transcriptome of a traditional Chinese medicinal plant *Polygonum cuspidatum*. *SCIENCE CHINA Life Sciences*. Vol.55 (5): 452–466, 2012.

HARBORNE, J. B.; GRAYER, R. J. The anthocyanins. London: Chapman and Hall, p. 1-20, 1988.

HARBORNE, JB. Classes and functions of secondary products, In: Walton NJ, Brown DE (Ed.). Chemicals from plants, perspectives on secondary plant products. *London: Imperial College*, p.1-25, 1999.

HARTMANN. THOMAS. From waste products to ecochemicals: Fifty years research of plant secondary metabolism. *Phytochemistry* 68: 2831–2846, 2007.

HECK, M.C.; VIANA, L.A.; VICENTINI, V.E.P. IMPORTÂNCIA DO ÓLEO DE Copaifera sp. (COPAÍBA). *SaBios: Revista Saúde e Biologia*, Vol.7, p.82-90, 2012.

INTERNATIONAL RICE GENOME SEQUENCING PROJECT. The map-based sequence of the rice genome. *Nature*, 436, 793–800, 2005.

IRITI, M.; FAORO, F. Chemical Diversity and Defence Metabolism: How Plants Cope with Pathogens and Ozone Pollution. *International Journal of Molecular Sciences*. 10: 3371-3399, 2009.

IRMISCH, S.; MULLER, A.T.; SCHMIDT, L.; GÜNTHER, J.; GERSHENZON, J.; KÖLLNER, T.G. One amino acid makes the difference: the formation of ent-kaurene and 16α-hydroxyent-kaurane by diterpene synthases in poplar. *BMC Plant* Biology. 15:262, 2015.

ITO, S. *et al.* Regulation of Strigolactone Biosynthesis by Gibberellin Signaling. *Plant Physiology*, Vol. 174, pp. 1250–12592017.

JUKES T.H. e CANTOR C.R. Evolution of protein molecules. In Munro HN, editor, Mammalian Protein Metabolism. *Academic Press*, New York. pp. 21-132, 1969.

KAUR, S.; PEMBLETON, L.W.; COGAN, N.O.; SAVIN, K.W.; LEONFORTE, T.; PAULL, J.; MATERNE, M.; FORSTER, J.W. Transcriptome sequencing of field pea and faba bean for discovery and validation of SSR genetic markers. *BMC Genomics*, 13:104, 2012.

KIRCHER, M.; KELSO, J. High-throughput DNA sequencing- concepts and limitations. *Bioessays Journal*. Vol. 32, p. 524–536, 2010.

KIM, M. Y. *et al.* Whole-genome sequencing and intensive analysis of the undomesticated soybean (Glycine soja Sieb. and Zucc.) genome. *Proc. Natl Acad. Sci*. USA 107, 22032–22037, 2010.

KÜLHEIM, A. P. *et al*. The Eucalyptus terpene synthase gene family Carsten. BMC Genomics, 16:450, 2015.

KÖLLNER, T.G; SCHNEE, C.; GERSHENZON, J. , DEGENHARDT, J. The variability of sesquiterpenes emitted from two Zea mays cultivars is controlled by allelic variation of two terpene synthase genes encoding stereoselective multiple product enzymes. Plant Cell. 16(5):1115-31, 2004.

LEÃO, A.C.R. *Montagem do Genoma DE Fonsecaea multimorphosa CBS 980.96T fungo isolado de abscesso cerebral felino.* Dissertação de mestrado. Programa de Pós-Graduação em Bioinformática. Universidade Federal do Paraná, 2016. 50p.

LEANDRO, L.M.; VARGAS, F.S.; BARBOSA, P.C.S.; NEVES, J.K.O.; SILVA, J.A.; VEIGA JUNIOR, V.F. Chemistry and Biological Activities of Terpenoids from Copaiba (*Copaifera* spp.) Oleoresins. *Molecules* vol.17, p.3866-3889.2012.

LIN J.; WANG, D.; CHEN, X.; KOLLNER, T. G.; GUO, H.; PANTALONE, V. R.; ARELLI, P.; STEWART, C. N. J.; WANG, N.; CHEN, F. An (E,E)-a-

farnesene synthase gene of soybean has a role in defence against nematodes and is involved in synthesizing insect-induced volatiles. *Plant Biotechnology Journal* 15, pp. 510–519, 2017.

LIU, J.; HUANG, F.; WANG, X.; ZHANG, M.; ZHENG, R.; WANG, J.;YU, D. Genome-wide analysis of terpene synthases in soybean: functional characterization of GmTPS3. *Gene*. 1;544(1):83-92, 2014.

LU, X.; TANG, K.; LI, P. Plant Metabolic Engineering Strategies for the Production of Pharmaceutical Terpenoids. *Front Plant Sci*; 7: 1647, 2016.

LUCCA, L.G.; MATOS, S.P.; BORILLE, B.T.; DIAS, D.O.; TEIXEIRA, H.F.; VEIGA JUNIOR, V.F.; LIMBERGER, R.P.; KOESTER, L. Determination of β-caryophyllene skin permeation/retention fromcrude copaiba oil (*Copaifera multijuga* Hayne) and respective oil-basednanoemulsion using a novel HS-GC/MS method. *Journal of Pharmaceutical and Biomedical Analysis.* Vol.104, 144–148, 2015.

LOPES, T. J.; XAVIER, M.F; QUADRI, M. G. N.; QUADRI, M. B. Antocianinas: Uma breve revisão das características estruturais e da estabilidade. *Revista Brasileira de Agrociência*, Pelotas, v.13, n.3, p. 291-297, 2007.

MCGARVEY, D.J.; CROTEAU. Terpenoid Metabolism. *The Plant Cell*. Vol.7, p.1015-1026. 1995.

METZKER, M. L. Sequencing technologies -the next generation. *Nature Review*, Vol.11: 32-40. 2010.

MILLER, J. R.; KOREN, S.; SUTTON, G. Assembly algorithms for next-generation sequencing data. *Genomics* 95, 315–327, 2010.

ONE KP, http://www.onekp.com/samples/list.php. Acesso: 14 de janeiro de 2018.

O'ROURKE, J.A; BOLON, Y-T.; BUCCIARELLI, B.; VANCE, C.P. Legume genomics: unders tanding biology through DNA and RNA sequencing. *Annals of Botany* 113: 1107–1120, 2014.

PASSOS, G.A.; JORDAN, C.N.B. Projeto Transcriptoma. Análise da expressão em larga escala usando DNA- Arrays. *Biotecnologia Ciência e Desenvolvimento. V*ol. 12. P.34-37, 2000.

PASZKIEWICZ, K.; STUDHOLME, D. J. *De novo* Assembly of Short Sequence reads. *Briefings in Bioinformatics*. Vol. 11(5):457- 472, 2010

PATEL, S.S.; SHAH, D.B.; PANCHAL, H.J. *De novo* Transcriptome Analysis of *Arachis Hypogaea* L. (*SRR1212866*). *OMICS Research*, V.5(1):1-6, 2015.

PEREIRA, R.J; CARDOSO, M.D.G. Metabólitos secundários vegetais e benefícios antioxidantes. *Journal of Biotechnology and Biodiversity* Vol. 3, N. 4: pp. 146-152, 2012.

PICHERSKY E.; NOEL J.P.; DUDAREVA, N. Biosynthesis of plant volatiles: nature's diversity and ingenuity. *Science.*10;311(5762):808-11, 2006.

PIERI, F.A.; MUSSI, M.C.; MOREIRA, M.A.S. Óleo de copaíba (*Copaifera* sp.): histórico, extração, aplicações industriais e propriedades medicinais. *Revista Brasileira Plantas Medicinais*, Botucatu, v.11, n.4, p.465-472, 2009.

PIERI, F.A.; SILVA, V.O.; SOUZA, C.F.; J.C.M., COSTA; L.F. SANTOS; MOREIRA, M.A.S. Antimicrobial profile screening of two oils of *Copaifera* genus. *Arquivo Brasileiro de Medicina Veterinária e Zootecnia. Vol.* 64, n.1, p.241-244, 2012.

PLOWDEN, C. Production ecology of copaiba *(Copaifera* spp.) Oleoresin in the Eastern Brazilian Amazon. *Economic Botany*. 57(4), 491-501. 2003.

REA, F. Plant ATP-Binding Cassette Transporters. *Annual Review of Plant Biology,*58:347-375, 2007.

RIGAMONTE-AZEVEDO, O.C.; WADT, P.G.S; WADT, L. H. O. Copaíba: ecologia e produção de óleo-resina. Rio Branco: EMBRAPA, MAPA, 2004. 28p.

ROBERTS, S.C. Production and engineering of terpenoids in plant cell culture. *Nature Chemical Biology*, 3, 387–395, 2007.

SAITOU N e NEI M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology Evolution*, 4(4):406-25, 1987.

SCHNABLE, P. et al. The B73 Maize Genome: Complexity, Diversity, and Dynamics. *Science* 326, 1112, 2009.

SCHNEE, C.; KÖLLNER, T. G; GERSHENZON, J.; DEGENHARDT, J. The Maize Gene *terpene synthase 1* Encodes a Sesquiterpene Synthase Catalyzing the Formation of (*E*)-β-Farnesene, (*E*)-Nerolidol, and (*E*,*E*)-Farnesol after Herbivore Damage. Plant Physiology, 130(4): 2049–2060. 2002.

SINGH, B e SHARMA, R.A. Plant terpenes: defense responses, phylogenetic analysis, regulation and clinical applications. *3 Biotech*, 5:129–151, 2015.

SOUZA, V. *Montagem do draft do genoma da bactéria Herbaspirillum huttiense subsp. Putei.* Dissertação, Universidade Federal do Paraná, 2012, 54p.

STONE, M.J.; WILLIAMS, D.H. M. On the evolution of functional secondary metabolites (natural products). *Molecular Microbiology*, 6(1): 29-34, 1992.

SUI, C.; CHEN,M.; XU,J.; WEI,J.; JIN,Y.; XU,Y.; SUN, J.;GAO,K.;YANGA,C.; ZHANGA, Z.; CHENA, S.; LUOA, H. Comparison of root transcriptomes and expressions of genes involved in main medicinal secondary metabolites from *Bupleurum chinense* and *Bupleurum scorzonerifolium*, the two Chinese official *Radix bupleuri* source species. *Physiologia Plantarum*. 153: 230–242. 2015.

SUN, T; YUAN, H, CAO, H,; YAZDANI, M.; TADMOR, Y.; LI, L. Carotenoid Metabolism in Plants: The Role of Plastids. *Molecular Plant*, 1- 42, 2017.

TAIZ, L.; ZEIGER, E. *Fisiologia vegetal*. Porto Alegre: Artmed, 2004. p.449-484.

TANAKA, Y.; SASAKI, N.; OHMIYA; A. Biosynthesis of plant pigments: anthocyanins, betalains and carotenoids. *The Plant Journal*, 54: 733–74, 2008.

THOLL, D. e LEE, S. Terpene Specialized Metabolism in Arabidopsis thaliana.

Arabidopsis Book. ;9:e0143. 2011.

TRAPP, S. e CROTEAU, R. Defensive Resin Biosynthesis in Conifers. *Plant Molecular Biology*, 52:689–724, 2001.

THE *ARABIDOPSIS* GENOME INITIATIVE. Analysis of the genome sequence of the owering plant *Arabidopsis thaliana*. *Nature*, 796- 815, 2000.

USADEL, B. e FERNIE, A. Theplanttranscriptome—from integrating observations to models*. Frontiers Plant Science*. Vol.4, 1- 3, 2013.

VANDESOMPELE, J.; PRETER, D.K.; PATTYN, F.; POPPE, B.; ROY, N.V.; PAEPE, A.D; SPELEMAN, F. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biology.* Vol. 3-7. 2002.

VEIGA-JR., PATITUCCI, M.L.; PINTO, A, C. Controle de autenticidade de óleos de copaíba comerciais por cromatografia gasosa de alta resolução. *Química Nova*, 20(6), 1997.

VEIGA-JÚNIOR, V. e PINTO, A.C. O Gênero *Copaifera* L. *Química Nova*. Vol. 25, no. 2, 273-286, 2002.

VEIGA JUNIOR, V.F. ROSAS, E.C.; CARVALHO, M.V.; HENRIQUES, M.G.M.O.; PINTO, A. C. Chemical composition and anti-inflammatory activity of copaiba oils from *Copaifera cearensis* Huber ex Ducke, *Copaifera*

reticulata Ducke and *Copaifera multijuga* Hayne—A comparative study. *Journal of Ethnopharmacology*. Vol.112, p, 248–254, 2007.

WANG, W.; WANG, Y.; ZHANG, Q.; QI, Y.; GUO, D. Global characterization of Artemisia annua glandular trichome transcriptome using 454 pyrosequencing. *BMC Genomics*, 10:465, 2009.

WINK, M. *Biochemistry of Plant Secondary*. Annual Review, 434p. 2010.

WU, X.; PRIOR, R. L. Standardized methods for the determination of antioxidant capacity and phenolics in foods and dietary supplements. *Journal of Agricultural and Food Chemistry*, v.53, p.3101-3113, 2005.

XIAO, M. *et al*. Transcriptome analysis based on next-generation sequencing of non-model plants producing specialized metabolites of biotechnological interest. *Journal of Biotechnology* 166, 122–134, 2013

XIE, F.; BURKLEW, C.E; YANG, B.; LIU, M.; XIAO, P.; ZHANG, B.; QIU, D. *De novo* sequencing and a comprehensive analysis of purple sweet potato (*Impomoea batatas* L.) transcriptome. *Planta.* 236:101–113, 2012.

YAHYAA, M.; MATSUBA, Y.; BRANDT, W.; DORON-FAIGENBOIM, A.; BAR, E.; MCCLAIN, A.; DAVIDOVICH-RIKANATI, R.; LEWINSOHN, E.; PICHERSKY, E.; IBDAH, M. Identification, Functional Characterization, and Evolution of Terpene Synthases from a Basal Dicot. *Plant Physiology*, Vol. 169:1683–1697, 2015.

ZHANG, M.; LIU, J.; LI, K.; YU, D. Identification and Characterization of a Novel

Monoterpene Synthase from Soybean estricted to Neryl Diphosphate Precursor. *PLOS ONE,* V.8, 1-10 2013.

ZHANG, S. *et al., De novo* characterization of Panax japonicus C. A. Mey transcriptome and genes related to triterpenoid saponin biosynthesis. *Biochemical and Biophysical Research Communications*. 466: 450-455, 2015.

ŻMIEŃKO, A; JACKOWIAK, P.; FIGLEROWICZ, M. Transcriptome sequencing: Next generation approach to RNA functional analysis. *Journal of Biotechnology, Computational Biology and Bionanotechnology.* Vol. 92(4) p. 311-319, 2011.

ZÜST, T e AGRAWAL, A.A. Mechanisms and evolution of plant resistance to aphids. *Nature Plants*, VOL 2, 1-9, 2016.

ZWENGER, S.; BASU, C. Plant terpenoids: applications and future potentials. *Biotechnology and Molecular Biology*. Reviews Vol. 3 (1), pp. 001-007, 2008.

ZWENGER, S.; REINSVOLD, R.E.; BASU, C. Analysis and functional annotation of expressed sequence tags from diesel tree (*Copaifera officinalis). Plant Biotechnology,* 27, 385-391. 2010.