Universidade de Brasília

Instituto de Química

Programa de Pós-Graduação em Tecnologias Química e Biológica

Vinicius Nattan Silva Lemos

Exploração *in silico* de dados de RNA-Seq de sementes de *Elaeis guineensis, Jatropha curcas* e *Ricinus communis* para a reconstrução de vias metabólicas de ácidos graxos e anotação de genes de interesse biotecnológico

Brasília

2019

Vinicius Nattan Silva Lemos

Exploração *in silico* de dados de RNA-Seq proveniente de sementes de *Elaeis guineensis, Jatropha curcas* e *Ricinus communis* para a reconstrução de vias metabólicas de ácidos graxos e anotação de genes de interesse biotecnológico

Dissertação apresentada ao Instituto de Química da Universidade de Brasília como requisito parcial para a obtenção do título de Mestre em Tecnologia Química e Biológica.

Área de Concentração:
Tecnologias Química e Biológica

Orientador:
Prof. Dr. Elíbio Leopoldo Rech Filho

Co-orientadora:
Dra. Priscila Grynberg

Brasília

2019

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoa de Nível Superior – Brasil (CAPES) – Código de Financiamento 001

FOLHA DE APROVAÇÃO

Comunicamos a aprovação da Defesa de Dissertação do (a) aluno (a) **Vinicius Nattan Silva Lemos**, matrícula no **17/0090337**, intitulada "**Exploração in silico de dados de RNA-Seq proveniente de sementes de Elaeis guineensis, Jatropha curcas e Ricinus communis para anotação de vias metabólicas de ácidos graxos**", apresentada no (a) da Universidade de Brasília (UnB) em 10 de julho de 2019.

Prof. Dr. Elibio Leopoldo Rechi Filho

Presidente de Banca

Prof.a Dra. Talita Souza Carmo

Membro Titular

Prof. Dr. Georgios Joannis Pappas Junior

Membro Titular IB/UnB

Prof.a Dra. Nádia Skorupa Parachin

Membro Suplente

Em 10 de julho de 2019.

Dedico este trabalho aos meus familiares e amigos que sempre me apoiaram, acreditaram no meu potencial e estiveram comigo nos melhores e piores momentos.

Agradecimentos

Gostaria de agradecer primeiramente a minha mãe Nilcea, pois essa vitória não é apenas minha, mas também dela que compartilha desse sonho comigo. Ao José Lemos, meu pai, por todo apoio durante os anos acadêmicos. A todos meus familiares que torcem pela minha vitória em especial a minha avó Rosa, que sempre sonhou em ter um neto com boas formações acadêmicas e estou feliz em trilhar esse caminho.

Agradeço a Universidade de Brasília onde iniciei em 2011 meu processo de formação e sou imensamente grato por todas as experiências boas e ruins neste ambiente acadêmico. E aos professores que passaram pela minha vida acadêmica, pois entendo a importância que cada um teve no desenvolvimento de uma visão crítica e de mundo. Ao Programa de Pós-graduação em Tecnologias Química e Biológica pela oportunidade de fazer parte da congregação dos estudos sobre processos químicos e biológicos de transformação de matéria prima.

À EMBRAPA pela oportunidade de participar de seu corpo institucional neste período. A todos os integrantes do Laboratório de Bioinformática da EMBRAPA Recursos Genéticos e Biotecnologia, em especial ao Gabriel, Anna Zotta e Marcos Costa pelas conversas acadêmicas e não acadêmicas, pela troca de experiências e ensinamento.

Ao meu orientador Dr. Elibio Leopoldo Rech Filho pela confiança no meu trabalho e por todo apoio, e a co-orientadora Dra. Priscila Grynberg pela confiança, ensinamentos e oportunidades. Ao Dr. Roberto Coiti Togawa por toda ajuda em toda e qualquer dificuldade encontrada e nos momentos de “pouca" fé.

Resumo

Jatropha curcas (pinhão-manso), *Ricinus communis* (mamona) e *Elaeis guineensis* (dendê) produzem ácidos graxos que podem ser utilizados como fonte renovável na matriz energética, apresentando um grande potencial biotecnológico para as indústrias da área. Um melhor conhecimento das vias metabólicas associadas com a síntese de ácidos graxos contribuirá para a maximização da produção utilizando métodos de engenharia metabólica e biologia sintética. O objetivo deste trabalho foi identificar transcritos relacionados com a síntese de ácidos graxos e inferir a sua presença em vias metabólicas. Para isso, o RNA total de sementes destas três espécies foi extraído e sequenciado utilizando a tecnologia *Illumina HiSeq 2500*. Após a exclusão de *reads* de baixa qualidade, os transcritomas foram montados utilizando seis programas diferentes. Os resultados obtidos foram filtrados com o programa *Evidential Gene* para a obtenção de resultados robustos. Um banco de dados local com 250 proteínas associadas à 12 vias metabólicas de ácidos graxos de 10 plantas foi criado para a identificação dos transcritos de interesse. Um total de 110, 110 e 109 proteínas únicas relacionadas com a produção de óleo foram identificados, com 146, 146 e 148 ocorrências em *J. curcas*, *R. communis* e *E. guineensis* respectivamente nas 12 vias metabólicas associadas a ácidos graxos. O número esperado de ocorrências, de acordo com o KEGG, seria de 158, 162 e 159, respectivamente. A estratégia desenvolvida foi capaz de identificar cinco transcritos que não haviam sido anotados nos genomas e na base de dados KEGG. A informação obtida neste trabalho pode ser utilizada como material de referência para estudos de engenharia metabólica e melhoramento da produção de ácidos graxos para propósitos industriais, principalmente voltado a biocombustíveis e produção de fármacos.

Palavras-chave: Transcritoma, RNA-Seq, ácidos graxos, *Jatropha curcas*, *Elaeis guineensis*, *Ricinus communis*, *J. curcas*, *E. guineensis*, *R. communis*, vias metabólicas, vias metabólicas de ácidos graxos.

Abstract

Jatropha curcas, *Ricinus communis* and *Elaeis guineensis* produce fatty acids that can be used as a renewable source in the energy matrix, presenting a great biotechnological potential for the industry. A better understanding of the metabolic pathways associated with fatty acid synthesis will contribute to the maximization of production using metabolic engineering and synthetic biology. The objective of this work was to identify transcripts related to the synthesis of fatty acids and to infer their presence in metabolic pathways. For this, the total seed RNA of these three species was extracted and sequenced using Illumina HiSeq 2500 technology. After exclusion of low-quality reads, the transcripts were assembled using six different programs. The results obtained were filtered with the Evidential Gene program to obtain robust results. A local database of 250 proteins associated with 12 metabolic pathways of 10 plant fatty acids was created for the identification of the transcripts of interest. A total of 110, 110 and 109 unique proteins related to the production of oil were identified, with 146, 146 and 148 occurrences in *J. curcas*, *R. communis* and *E. guineensis* respectively in the 12 metabolic pathways associated with fatty acids. The expected number of occurrences, according to the KEGG, would be 158, 162 and 159, respectively. The strategy developed was able to identify five transcripts that had not been annotated in the genomes and in the KEGG database. The information obtained in this work can be used as reference material for studies of metabolic engineering and improvement of the production of fatty acids for industrial purposes, mainly focused on biofuels and drug production.

Keywords: Transcriptome, RNA-Seq, fatty acid, *Jatropha curcas*, *Elaeis guineensis*, *Ricinus communis*, *J. curcas*, *E. guineensis*, *R. communis*, metabolic pathways, fatty acid pathways.

Lista de Figuras

Lista de Tabelas

Lista de Quadros

Lista de Abreviações e Glossário

BioCyc – Banco de dados de vias e genomas, contendo cerca de 14.600 vias.

BLAST - *Basic Local Alignment Search Tool*

BUSCO - *Benchmarking Universal Single-copy Orthologs* – Programa de computador que procura em um banco de dados contendo os chamados "cópias únicas ortólogas" que mostra o quão completo está determinado grupo de sequências (seja ela genômica ou transcritômica)

Contig – Sequência genômica contígua da qual a ordem das bases possui um alto valor de confiabilidade

DNA - ácido desoxirribonucleico

Fasta – formato de arquivo de texto utilizado em geral para representar sequências de nucleotídeos ou de aminoácidos. Um arquivo .fasta pode conter uma ou mais sequências. O nome da sequência é representado pelo sinal de maior '>', sendo que na próxima linha após o nome é a sequência propriamente dita.

Gaps – Espaços introduzidos no alinhamento para compensar as inserções e deleções de uma sequência em relação a outra

kb - kilo bases, 10^3

KEGG - *Kyoto encyclopedia genes and genomes*

KEGG API - *KEGG Application Programming Interface*

UniProt – *Universal Protein Resource* – Principal repositório de sequencias proteicas. As sequências disponibilizadas são altamente curadas

Mb - milhão de bases, 10^6

N50 – Define o tamanho mínimo que o *contig* precisa para cobrir 50% do genoma. Demonstra que metade da sequência genômica está em *contigs* maiores ou iguais ao valor de N50.

NCBI - *National Center for Biotechnology Information* – Principal repositório de sequências biológicas

PANTHER – *Protein ANalysis THrough Evolutionary Relationships*, banco de dados de classificação de proteínas.

pb - pares de bases

Phred - representação numérica da qualidade de identificação dos nucleotídeos gerados à partir do sequenciamento de DNA automatizado. Utilizado para avaliação da qualidade, reconhecimento e remoção de sequências de baixa qualidade.

RBNHs - *Reciprocal Best length - Normalised hit*, método de alta precisão para encontrar ortólogos normalizado pelo BLAST

Reactome – Banco de dados de vias metabólicas. Curada manualmente e de livre acesso.

RNA - *Ribonucleic acid,* ácido ribonucleico

RNA-Seq - RNA *sequencing*

Scaffold – uma porção da sequência genômica reconstruído a partir da junção de *contigs* e *gaps*

STAR – *Spliced Transcripts Alignment to a Reference*

Sumário

1 Introdução

O sequenciamento de RNA (ácido ribonucleico) engloba um conjunto de técnicas experimentais e computacionais que possibilitam identificar a quantidade de sequências de RNA em amostras biológicas em um determinado estágio de desenvolvimento (KORPELAINEN et al., 2015; WANG; GERSTEIN; SNYDER, 2009). Entender o transcritoma de uma planta é essencial para entender e interpretar os elementos genômicos, como, por exemplo, RNAs condificantes e não codificantes, e como estes influenciam nos processos biológicos (WANG; GERSTEIN; SNYDER, 2009).

A técnica de RNA-Seq também viabiliza a identificação de transcritos não anotados e miRNAs, polimorfismos de um único nucleotídeo e consequentemente a identificação de possíveis alvos moleculares para o melhoramento genético (POPLAWSKI et al., 2015; VAN VERK et al., 2013).

A anotação de vias metabólicas de ácidos graxos tem um importante papel na obtenção de informações a respeito das vias e suas possíveis utilizações industriais, voltados principalmente de biocombustíveis e lubrificantes, a partir dos óleos extraídos de plantas oleaginosas.

Planta oleaginosa tem como característica principal a presença de altas concentrações de ácidos graxos em sua composição. Essas plantas possuem diversas aplicações nos setores industriais, alimentícios, na produção de sabão, como combustíveis alternativos ao diesel e, na indústria de medicamentos e cosméticos, como a capacidade de serem utilizados na confecção de lubrificantes, óleos de cozinha, processamento de alimentos e produção de biocombustíveis (KANTAR et al., 2016; SAVADI et al., 2016). E ainda, podem ser utilizadas para a consolidação de programas de energia renovável (VILLELA et al. 2014) para a produção de biodiesel, pois proporcionam apoio à agricultura familiar, criando melhores

condições de vida (infraestrutura) em regiões carentes e valorizando potencialidades regionais.

Processos industriais relacionados as suas utilizações vêm sendo desenvolvidos para a melhor utilização dos óleos obtidos dessas plantas, onde os ácidos graxos são extraídos e por meio de processos químicos e biológicos são transformados em um produto comercial (BILAL et al., 2018; ZHU et al., 2016).

A variedade de plantas utilizadas pela indústria é relativamente baixa, sendo restrita principalmente à soja, milho, girassol e algodão, pois a domesticação/melhoramento de novas plantas é um processo que demora em média 10 anos e demanda um alto investimento, isso devido pelo crescimento do agronegócio, elevando a importância do melhoramento genético, biotecnologia e incorporação de novas tecnologias ao processo de produção de sementes (SANTOS et al., 2014). Essas plantas oleaginosas não convencionais, podem ser utilizadas para a produção de biodiesel e diminuição do impacto ambiental causado pela emissão de gases à partir da queima de combustíveis fosseis, assim a obtenção de informações genômicas são importantes para o estabelecimento de estratégias e delineamento de experimentos para a utilização dessas plantas na matriz energética (BACENETTI et al., 2017).

O pinhão-manso (*Jatropha curcas),* dendê (*Elaeis guineensis)* e a mamona (*Ricinus communis)* estão entre as plantas não convencionais capazes de se extrair óleos de interesse para as indústrias alimentícias, de biocombustíveis, cosmética e farmacêutica e apresentam um grande potencial biotecnológico, devido as informações que podem ser extraídas delas. Estas plantas apresentam diversas dificuldades, seja por seu modo de cultivo e crescimento ao longo do ano ou pela falta de informações disponíveis e corroboradas experimentalmente contidas no seu genoma. Estas espécies possuem genomas disponíveis, mas não estão completos, o que dificulta as análises e inferências quanto à suas propriedades físicas, químicas e genéticas. Na figura 1, é possível observar as separações filogenéticas dessas plantas criada à partir de informações taxonômicas dispostas no *Taxonomy browser*, do

National Center for Biotechnology Information (NCBI), mostrando assim que *Jatropha curcas* e *Ricinus communis* são mais próximas entre si, e *Elaeis guineensis* está mais distante delas. Segundo Loureiro et al (2013), as informações morfoanatômicas sobre as sementes de *Jatropha curcas* L. são ainda escassas, no entanto na literatura são encontrados estudos sobre a morfologia e anatomia de frutos e sementes de espécies da mesma família.

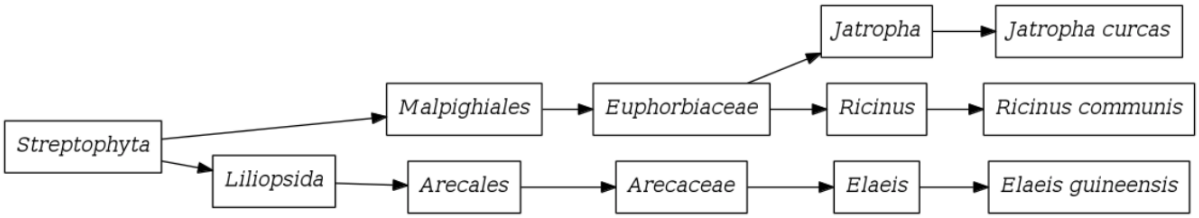


Figura 1 - Árvore filogenética das três plantas, das quais foram extraídos os RNAs totais e sequenciados para a realização das analises.
Fonte: NCBI, 2019.

Estas espécies apresentam vantagens de crescerem e se desenvolverem bem em habitats que possuem solos com nutrientes e capacidade hídrica limitados gerando assim uma oportunidade de desenvolvimento econômico e a diversidade do meio ambiente (CANVIN, 1965; SAVADI et al., 2017).

1.1 *Jatropha curcas*

J. curcas (Figura 2) pertence à família *Euphorbiaceae*, nativa da América Central, e tem como nome comum pinhão-manso. É uma pequena árvore semi-perene com resistência a ambientes áridos. É bastante utilizada para produção de biocombustíveis e na medicina natural por possuir altas taxas de óleo em suas sementes (cerca de 27-40% da composição) (ACHTEN et al., 2007, 2008). No entanto, presença de ésteres de forbol e curcina são as principais toxinas no óleo extraído, que inviabiliza seu uso na indústria alimentícia (DEVAPPA; MAKKAR; BECKER, 2010; KUMAR; SHARMA, 2008). As dificuldades encontradas são: alto custo para o plantio, falta de conhecimentos avançados sobre a biologia, incidência e manejo de pragas e doenças relacionadas, sobre cultivo e produtividade e

ausência de programas governamentais de incentivo (EDRISI et al., 2015). Industrialmente, esta planta pode ser utilizada de diversas formas como para ração animal, produção de biocombustíveis, utilização na indústria química na fabricação de pesticidas (MONTES; MELCHINGER, 2016).

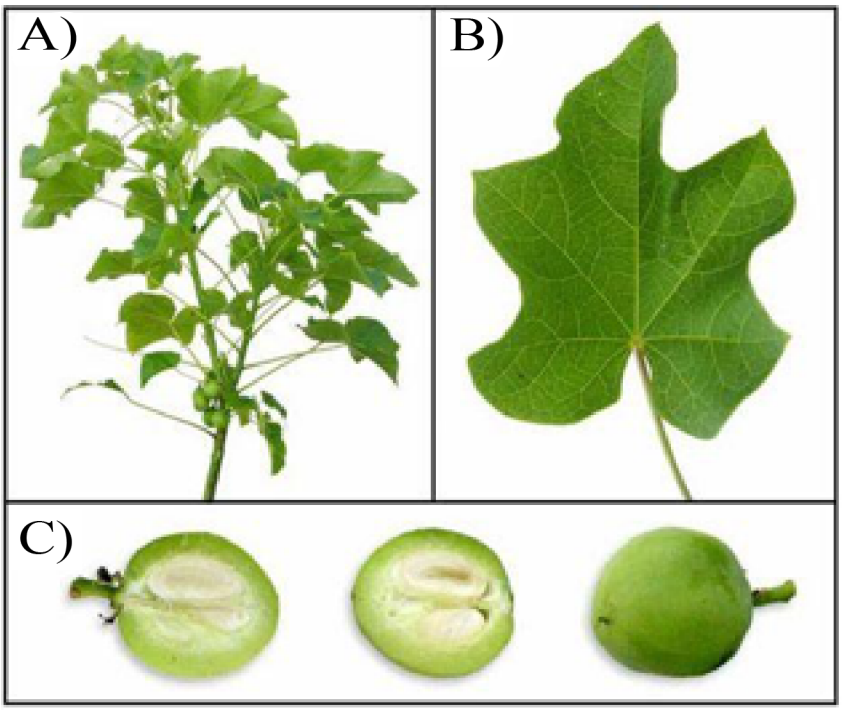


Figura 2 - Representação da árvore (A), folha (B) e fruto inteiro e cortado transversalmente (C) de *J. curcas*.

Fonte: Disponível em: sustainabledesignupdate.com, acesso em: Maio/2019.

Estudos sobre a utilização do biodiesel a partir do óleo de *J. curcas* estão sendo realizado (BRITTAINE; LUTALADIO, 2010; CASTRO GONZÁLES, 2016; YANG et al., 2012). Este contém uma viscosidade maior (0,93292 gm/cc), enquanto no diesel fóssil é de 0,850 gm/cc. Essa diferença na viscosidade impede a sua utilização completa na indústria automotiva devido às modificações que devem ser feitas no motor para a queima eficiente do combustível, modificações essas que não foram publicadas na literatura. A melhor alternativa até agora para sua utilização é fazer uma mistura entre o biodiesel produzido pelo óleo de *J. curcas* e o diesel para manter estabilidade de viscosidade na maioria das temperaturas (PRAMANIK, 2003). A torta (parte sólida resultante da moagem industrial) proveniente da

produção de biodiesel também pode ser aproveitada como substrato para digestões anaeróbicas, adubos e outros processos que visam a produção de outros compostos químicos de interesse (NAVARRO-PINEDA et al., 2016).

O genoma da *J. curcas* (JatCur_1.0) foi montado pela *Chinese Academy of Science* em 2014 e está disponível no NCBI (WANG et al., 2014). Está organizado em nível de suportes com um N50 (tamanho mínimo para cobrir metade do genoma) de 746 kb. Ele foi sequenciado com tecnologia *Illumina HiSeq*, que permite o sequenciamento via a síntese das sequencias, e montado com o programa *SOAPdenovo* (LUO et al., 2012), conferindo uma cobertura de 189x.

1.2 *Elaeis guineensis*

E. guineensis (Figura 3), pertencente à família *Arecaceae,* é conhecida no Brasil como dendê, é uma árvore produtora de óleo de palma (azeite de dendê). Seu local de origem não é bem definido, mas é de conhecimento que é no continente africano (SAMBANTHAMURTHI et al., 2009). Completa seu ciclo em 1 ano, passando pelas germinação e florescimento neste período. É amplamente cultivada no continente africano (OBAHIAGBON, 2012), com uma produção de 196 milhões de toneladas em 2017. No entanto, a Indonésia e a Malásia, juntas, foram responsáveis por 260 milhões de toneladas neste ano (FAO, 2019). O óleo de palma é composto principalmente por ácido palmítico, esteárico e mirístico (saturados) e ácido oleico e linoleico (insaturados), totalizando aproximadamente 95% (Tabela 1). (AHSAN; AHAD; SIDDIQUI, 2015; EDEM, 2002; MBA; DUMONT; NGADI, 2015).

Tabela 1 - Composição do óleo de *Elaeis guineensis*

	Nome sistemático	Símbolos	Porcentagem do peso total
Láurico	n-Dodecanoico	C 12:0	<1
Miristico	n-Tertadecanoico	C 14:0	1-6
Palmítico	n-Hexadecanoico	C 16:0	32-47
Esteárico	n-Octadecanoico	C 18:0	1-6
Arachide	n-Eicosanoico	C 20:0	<1
Palmitoleico	n-Hexadec-9-enoico	C 16:1	<1
Oleico	n-Octadec-9-enoico	C 18:1	40-52
Gadoleico	n-Eicos-9-enoic	C 20:1	<1
Linoleico	n-Octadec-a,12-denoic	C 18:2	5-7

Fonte: OBAHIAGBON, 2012.

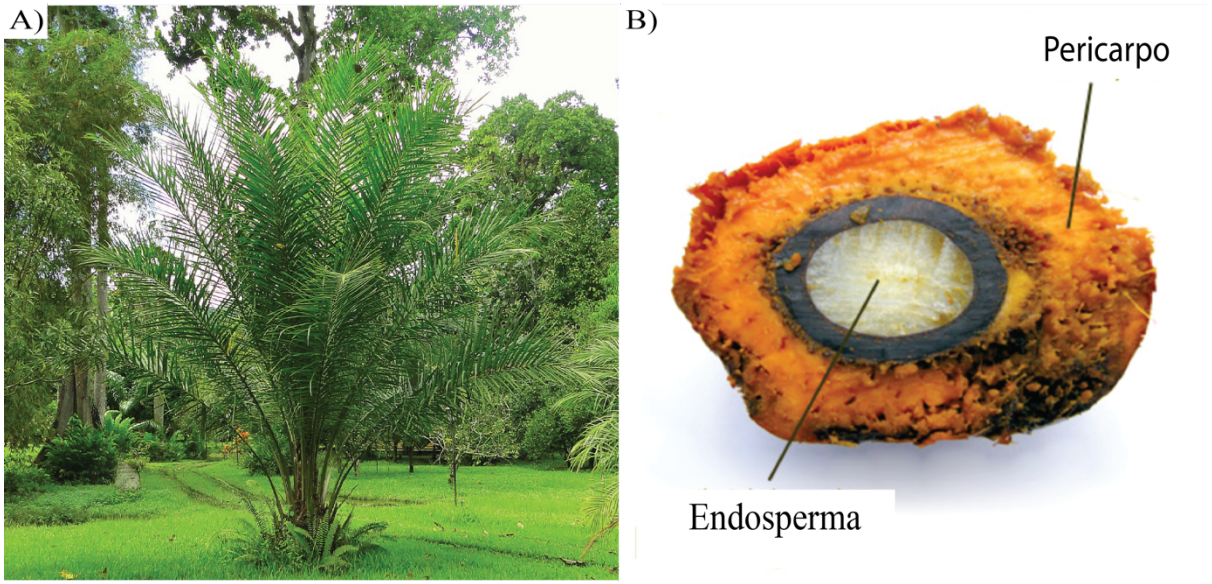


Figura 3 – Árvore de *E. guineensis* (A) e corte transversal da semente apresentando o pericarpo e endosperma (B).

Fonte: A – Flickr. Disponível em: https://www.flickr.com/photos/barloventomagico/5332238650, acesso em: Maio/2019. B – Stuartxchange. Disponível em http://www.stuartxchange.com/AfricanOilPalm.html, acesso em: Maio/2019.

Estudos relatam o uso do óleo de palma há mais de 5.000 anos (OBAHIAGBON, 2012). Atualmente, na indústria alimentícia, o óleo de palma é usado como óleo para cozimento, na produção de margarinas, sorvetes, cremes vegetais, molhos de saladas, queijos vegetais e em suplementos e vitaminas. O óleo de palma também é muito usado na indústria química. A partir dele são produzidos surfactantes, cosméticos, graxas e lubrificantes, sabão,

tinta de impressora, dentre outros produtos. E, por fim, esta planta é usada para a fabricação de biodiesel, móveis, fertilizantes, produção de papel, além de outros produtos. Há ainda algumas aplicações, que são iminentes, na indústria farmacêutica. Estudos recentes mostraram que o óleo de palma é capaz de inibir a proliferação de células de câncer de mama, possui efeitos antioxidantes, inibe a síntese do colesterol total e previne o dano ao DNA (AHSAN; AHAD; SIDDIQUI, 2015; EDEM, 2002; MBA; DUMONT; NGADI, 2015; OBAHIAGBON, 2012).

Atualmente há quatro cultivares de *E. guineensis*: Macrocaria, Dura, Pisifera e Tenera. A classificação da cultivar é feita com base na estrutura da fruta e na produtividade. Para todos os cultivares, o alcance máximo na produtividade se dá em 30 anos, mesmo a planta podendo viver até 200 anos. Apenas a Pisifera e Tenera são cobiçadas comercialmente por industrias de biocombustíveis, cosméticos e químicos, quando comparadas com os outros cultivares que são mais utilizadas em ornamentação (VERHEYE, 2010).

O melhoramento convencional, apesar do período necessário de estudos (cerca de 10 anos para a escolha das melhores progênies) aumentou consideravelmente a qualidade e produtividade do óleo extraído. No entanto, a possibilidade de explorar as informações genômicas pode encurtar este tempo e melhorar a precisão de escolha de progênies baseados em fenótipos de acordo com o interesse econômico (SAMBANTHAMURTHI et al., 2009).

O genoma de *E. guineensis* disponível é o EG5 depositado no NCBI pela *Orion Genomics* em 2013 com uma cobertura de 16x (SINGH et al., 2013; UTHAIPAISANWONG et al., 2012). O sequenciamento foi feito com a Roche 454 e montado com o programa Newbler v. 2.6 (MARGULIES et al., 2006). Possui 16 cromossomos com um *scaffold* de mais de 1,1 MB, e sequências com comprimento total de 1,5 GB e 288 gaps entre *scaffolds*. O comprimento dos gaps chega a 4,8 MB, correspondendo a espaçamentos pequenos quando comparados ao tamanho total do genoma.

1.3 *Ricinus communis*

R. communis (Figura 4) pertence também à família *Euphorbiaceae*. No Brasil é popularmente chamada de mamona. É uma planta perene, produzida principalmente no Brasil, China e Índia, países em desenvolvimento e com vulnerabilidade econômica. Juntos, esses países são responsáveis por 93% da produção mundial. A utilização deste cultivar possibilita o crescimento da região, além de diminuir os impactos ambientais, por ser uma fonte renovável de energia e possibilita o aquecimento de empresas de biotecnologia e agricultura regional (BARAJAS FORERO, 2005; OGUNNIYI, 2006). A produção do óleo de mamona (*Castor oil*) corresponde a apenas 0,15% da produção mundial de óleos vegetais. No entanto, a demanda pela indústria aumentou consideravelmente, saltando de 400 mil toneladas em 1985 para 610 mil toneladas em 2010, isto devido ao aumento de químicos, cosméticos e biocombustíveis formados a partir desses óleos (SEVERINO et al., 2012).

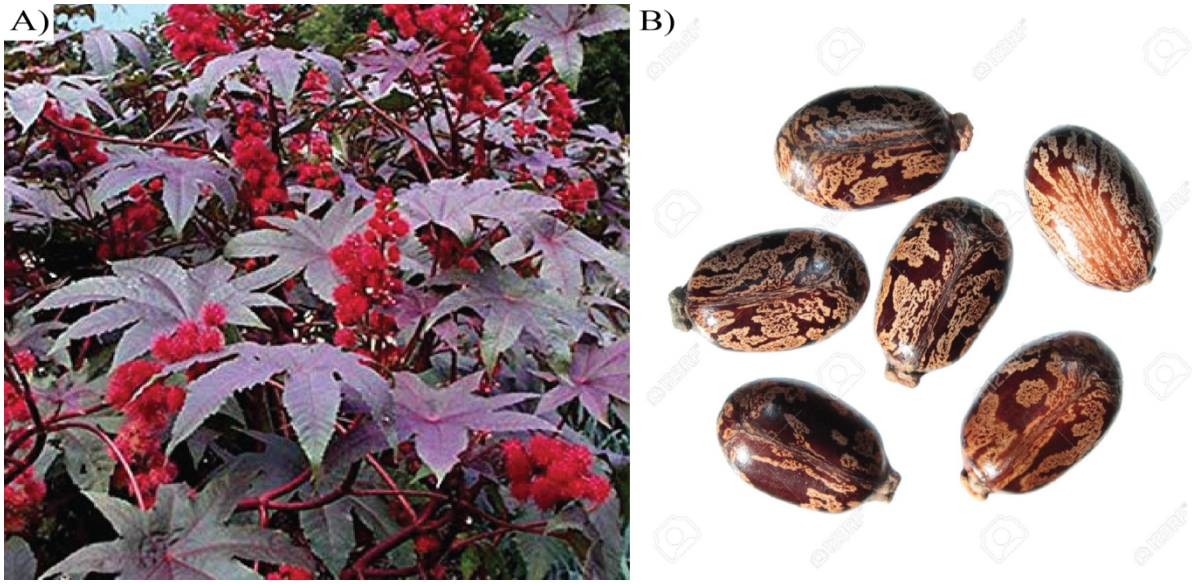


Figura 4 - Árvore de *R. communis* (A) e sementes (B) extraídas do fruto.

Fonte: A – Plant world seeds (Disponível em: https://www.plant-world-seeds.com/store/view_seed_item/2480, acesso em: Abril/2019). B – KAZAKOV, M. Seeds of Castor Bean Plant (Ricinus communis). Disponível em: https://www.123rf.com/photo_73690678_seeds-of-castor-bean-plant-ricinus-communis-on-white-background.html, acesso em: Abril/2019.

Seu uso remonta a aproximadamente 1.400 anos, quando foi domesticado no nordeste africano e introduzido na China, sendo assim utilizada por povos regionais, mas não é utilizado amplamente por indústrias de químicos, biocombustíveis e cosmética (HONG; BLACKMORE, 2015).

O óleo de mamona é amplamente empregado na indústria química. Cerca de 40-60% da composição da semente é ácido graxo (BARAJAS FORERO, 2005; TAN et al., 2009). Cerca de 90% do óleo é formado por ácido ricinoleico, que possibilita a produção de derivados de pureza elevada.

A mamona é considerada uma das plantas mais promissoras para uso em combustíveis e biodiesel devido à sua alta produção de sementes anual e pelo fato de crescer em terras marginais e clima semi-árido. O biodiesel feito a partir do óleo de *R. communis* foi 18% mais eficiente quanto ao retorno de temperatura quando comparado com o diesel de petróleo o que ajuda a conservar os equipamentos. No entanto, o custo de produção limita a adoção deste óleo pelas indústrias farmacêuticas, de biocombustíveis e química. Mas esse problema pode ser eliminado se houver o uso com uma mistura de outros óleos vegetais (soja, milho e canola) na produção, barateando parte do processo de obtenção e extração dos óleos e ainda obtendo um produto final de qualidade (OGUNNIYI, 2006; SEVERINO et al., 2012).

Além disso, o óleo de mamona e seus derivados podem ser usados na síntese de monômeros e polímeros renováveis e biodegradáveis, na fabricação de ceras, graxas e sabão, lubrificantes automotivos, que não necessita de aditivos por ter alta viscosidade em baixas temperaturas devido à hidroxila incomum encontrada nesse ácido, como pode ser observado na estrutura química na Figura 5, fertilizantes, tintas e uso médico e farmacológico (PATEL et al., 2016).

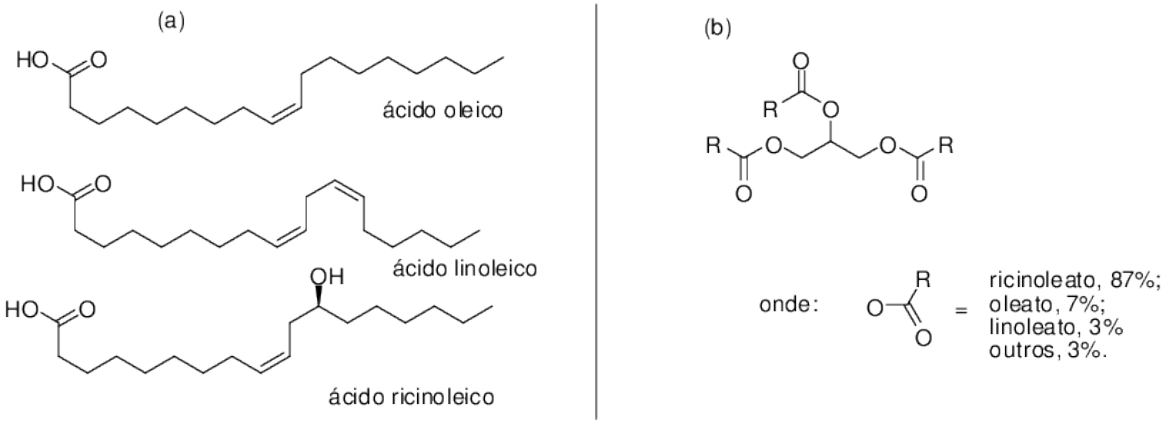


Figura 5 – a) Estrutura molecular geral dos principais ácidos graxos presentes no óleo de mamona. b) estrutura molecular típica mostrando os percentuais desses ácidos graxos presentes na mistura na forma de ésteres de glicerol e triglicerídeos.

Fonte: (MORAIS et al., 2013)

A mamona possui alta toxicidade devido à presença da ricina. Ela corresponde entre 1%-5% do peso total da semente, e permanece após a extração do óleo. Ela pode ser utilizada como pesticida devido ao baixo risco de contaminação em humanos e pelo fato de que animais ruminantes conseguem metabolizar esta toxina, possui um LD50 (dose letal) por via oral de 20 mg/kg, sendo classificada assim como muito toxica (SEVERINO et al., 2012).

O sequenciamento do genoma de R. *communis (*JCVI_RCG_1.1) foi iniciado em 2011 pelo *J. Craig Venter Institute*, no entanto, até o presente momento não foi publicado. A montagem, em nível de *scaffolds*, está disponível no website do NCBI (https://www.ncbi.nlm.nih.gov/assembly/GCF_000151685.1). A tecnologia Sanger foi usada para o sequenciamento e a montagem foi feita com o programa Celera Assembler. A cobertura é de 4.5x, possui 350 Mb, N50 de 496 kb e gaps com 13 Mb.

1.4 **Produção de óleo vegetal**

No ano de 2017, a produção anual de culturas oleaginosas, segundo dados oficiais, estimados ou inferidos da FAO, foi de 994 milhões de toneladas, sendo que o Brasil produziu 123 milhões de toneladas, ou seja, 12,4% da produção mundial. A soja e o óleo de palma

foram responsáveis por aproximadamente dois terços da produção mundial. A Indonésia foi, neste ano, o maior produtor de óleo de palma, responsável por 158 milhões de toneladas (aproximadamente 16% da produção total mundial e 50% da produção de óleo de palma). O Brasil produziu 1,7 milhões de toneladas no mesmo período (FAO, 2019).

A produção de óleo vegetal mundial vem crescendo ano após ano (Figura 6), pois é utilizado em muitas áreas diferentes como a alimentação, produção de combustível e de fármacos. A matéria prima para a geração de óleo vegetal são as plantações, que utilizam em geral de monoculturas em larga escala. A necessidade de produção aumenta a cada ano, mas a eficiência do plantio não acompanha essa realidade.

Assim o melhoramento genético das plantas, principalmente por meio de cruzamentos e técnicas do DNA recombinante para minimizar o estresse hídrico, herbivoria e competitividade com outras plantas locais é necessário, para que se consiga aumentar a produção sem a necessidade de aumentar das terras de plantio, a bioinformática permite a observação de alvos para o melhoramento, identificação do melhor alvo e obtenção de informações não existentes previamente para esses espécies (DURAND-GASSELIN et al., 2011).

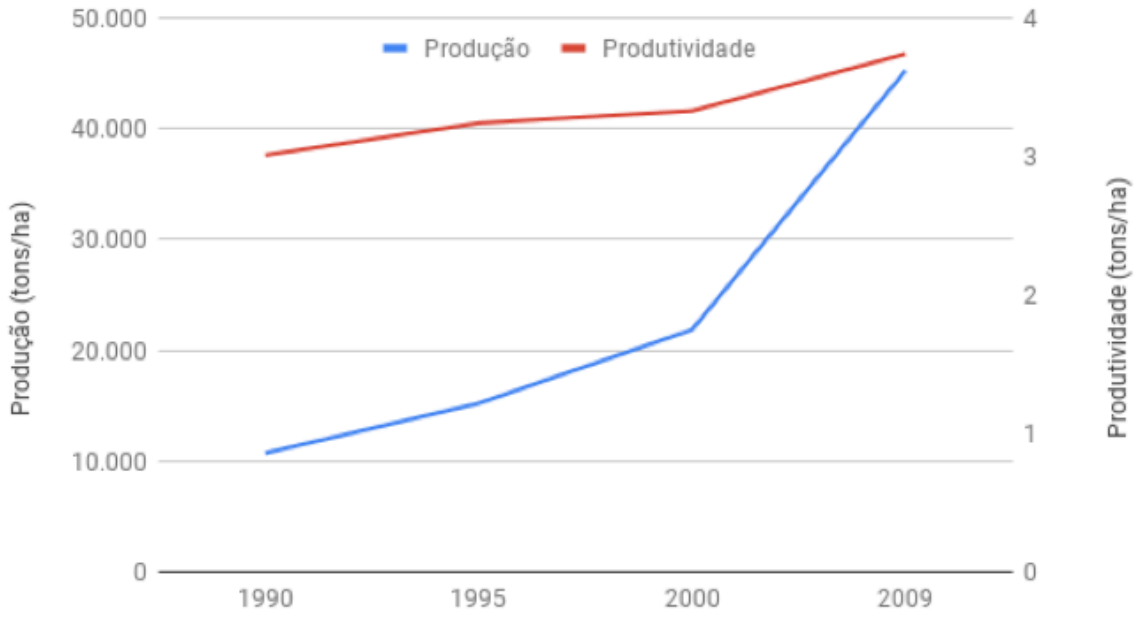


Figura 6 - Produção e Produtividade média de plantas oleaginosas entre 1990 e 2009.

Fonte: DURAND-GASSELIN et al. (2011).

No contexto atual, em que estudos biotecnológicos auxiliam na obtenção de informações quanto a anotação das vias metabólicas há a possibilidade: de aumento da concentração de ácidos graxos, a velocidade de crescimento e a resistência das plantas para terem lugar de destaque na economia e na sociedade (SAVADI et al, 2016). Há apenas duas maneiras de aumentar o coeficiente de óleos vegetais: a. através do aumento as áreas de plantação, e b. aumentar a produtividade das áreas atualmente plantadas. O aumento da área de plantio não é um método sustentável, pois aumenta automaticamente o desmatamento e reduz as condições ideais de plantio para aquele cultivar. O aumento da produtividade é mais apropriado comercialmente e para o meio ambiente. Novos métodos de melhoramento devem ser empregados para que haja um aumento na produtividade, já que o método tradicional (cruzamento) tem seus limites e dificuldades de implementação em função da baixa variabilidade genética de algumas espécies, tais como *E. guineensis*. Os estudos sobre os mecanismos de resposta planta-praga, aos estresses e como as vias metabólicas se relacionam com o meio ambiente e as necessidades do organismo (fixação de nitrogênio, resistência a

seca, entre outros) são necessários para melhorar a produtividade dos cultivos no futuro, por meio de novas tecnologias e protocolos (biologia sintética e engenharia metabólica) (TEH et al., 2017).

A biologia sintética é definida como digital e padronizada, pois, os cientistas utilizam e/ou modificam técnicas de outras áreas biológicas, tais como: engenharia genética, microbiologia e bioinformática, para transformarem microrganismos naturais em sintéticos de modo sistemático (Centro Ecológico, 2009-2010). E a engenharia metabólica, segundo Silva e Paulillo (2015) é a utilização dos *biobricks* (cassetes padronizados para expressão gênica) para sintetizar e transplantar genomas contendo apenas genes específicos, além da eliminação de genes próprios, visando a melhoria das atividades metabólicas dos microrganismos pela manipulação de suas funções enzimáticas.

1.5 Indústrias e ácidos graxos

Com a necessidade de investimento em energias renováveis, devido à preocupação com questões ambientais, envolvendo o aquecimento global, poluição de rios e mares provenientes da utilização de combustíveis fosseis, as indústrias começaram a utilizar óleos vegetais para produção, em especial, de biocombustíveis (BILAL et al., 2018), o que gerou um aumento do preço do óleo vegetal nos últimos anos, chegando a valores de 2.500,00 US$ por tonelada onde anteriormente era de 800 US$ por tonelada (oilworld.biz). A indústria de produção de biocombustíveis busca utilizar plantas que possam ser amplamente cultivadas e que não onerem a demanda nutricional do país, e que possuam preços competitivos (ALVIM, 2011; ZHOU et al., 2016).

Trabalhos visando o uso de plantas não utilizadas no consumo alimentício e que possuam boa capacidade energética começaram a surgir e a obter financiamentos. No entanto, há pouca informação genômica quanto a essas espécies, pois como são espécies que entraram recentemente no interesse industrial, os estudos quanto às suas propriedades ainda estão em

processo de elucidação e validação. Assim, toda informação gerada é relevante e importante para o processo de domesticação, por estarem contidas no genoma da planta (DEMIRBAS; BALAT, 2006; HANSEN et al., 2017).

Muitas dessas plantas possuem gargalos na utilização das mesmas como matéria prima, seja por dificuldades no plantio ou no processamento na cadeia industrial, pela ocorrência de substâncias indesejadas, principalmente de toxinas. A descoberta de novas enzimas com base nos estudos de vias metabólicas dessas plantas pode contribuir com o melhoramento de espécies já utilizadas na agricultura, bem como o melhor manejo dessas plantas não convencionais na agricultura industrial (MOGHE et al., 2017; NÜTZMANN; HUANG; OSBOURN, 2016; SHU, 1998). Estudos com diferentes cultivares possibilitam novos usos da engenharia metabólica, podendo auxiliar no melhoramento genético de novas oleaginosas.

A biotecnologia de plantas mostra ter um papel importante na promoção de maneiras mais sustentáveis da utilização de óleos vegetais (LU et al., 2011; NAPIER; GRAHAM, 2010). A indústria farmacêutica também vem investindo cada vez mais em pesquisas sobre a utilização de produtos vegetais com funções para melhora da saúde humana, na descoberta de novas drogas (MOGHE et al., 2017; SHU, 1998).

As moléculas formadas pelas vias metabólicas de ácidos graxos também são objeto de interesse industrial devido sua capacidade de conversão em produtos de maior valor agregado, tais como: biocombustíveis e lubrificantes. Os compostos que possuem propriedades únicas, geram produtos com maior valor agregado, como o N-Acetil-D-esfingosina, que é produzido a partir da via de metabolismo de esfingolipídeo e vendido pelo preço de R$169/g (Sigma-Aldrich). Outras aplicações farmacêuticas incluem o acoplamento de esfingosina com fenetil isotiocianato que possui ação inibitória ao crescimento de células com leucemia entre outras aplicações (MIAZEK et al., 2016).

Além da formação de produtos, essas vias nos trazem informações quanto à sua utilização para o aumento da formação de biomassa, não só para as espécies deste projeto, a

partir da inibição de partes das vias de biossíntese de esteroides, como também pelo aumento da captação de luz e fixação de nitrogênio (TAKATSUTO et al., 2005; VIGEOLAS et al., 2007). Outra utilização seria a produção de hidrolases capazes de facilitar o processamento de material lignocelulósico e facilitar o acesso de microrganismos ao substrato para produções industriais (JOHNSON et al., 2007; STICKLEN, 2008).

1.6 **Vias metabólicas**

O estudo das vias metabólicas é de extrema importância para os organismos, pois por meio delas é possível explorar a síntese de metabólitos primários e secundários que são essenciais para o crescimento, adaptação, resistência, resposta imune e reprodução dos seres vivos. A elucidação de como esses mecanismos ocorre para que se possa controlá-los ou incentivar o acontecimento deles é a base dos estudos de biologia sintética (BENNER; SISMOUR, 2005; KHALIL; COLLINS, 2010).

Para a biotecnologia, as vias metabólicas se apresentam como uma oportunidade de negócio, seja para aumentar a produção de determinado produto com valor agregado, seja para diminuir quantidade de metabólitos secundários oriundos da produção ou melhorar o consumo de substrato (NIELSEN; KEASLING, 2016). O conhecimento das vias metabólicas vem sendo utilizado ao longo do tempo para o melhoramento genético na agricultura, objetivando conferir características à planta como: maior resistência à seca e pragas, melhores e maiores frutos e outras características que sejam vantajosas para seu cultivo ou utilização (DERISI; IYER; BROWN, 1997). Essas vias formam uma malha complexa que é totalmente integrada. Desvendar o conhecimento sobre como uma via influencia as outras vias é necessário, pois mudanças podem ocasionar até mesmo a morte do organismo, caso se alterem processos essenciais, como a glicólise e síntese de ácidos graxos (STEPHANOPOULOS, 1999).

Bancos de dados especializados em vias metabólicas foram desenvolvidos ao longo das últimas três décadas. Um dos principais é o KEGG (Kyoto Encyclopedia of Genes and Genomes). Ele possibilita a procura, visualização e *download* de sequências de genes relacionados às vias metabólicas. O KEGG tem em sua base de dados mais de 29 milhões de sequências de 6.243 organismos, sendo 102 plantas, dos quais em geral possuem muitas informações, sendo classificados assim como bem anotados e diversas conexões com outros bancos de dados curados (KANEHISA et al., 2008).

1.6.1. **Vias de ácidos graxos**

Os organismos geralmente utilizam as vias de ácidos graxos para sintetizar reservas de energia em forma de amido e ácidos graxos, sendo assim de grande importância. A partir delas podemos sintetizar componentes essenciais (glicose e ATP) e armazenar energia (amido e ácidos graxos). O armazenamento dessa energia é feito na forma de amido para ser utilizado na fase de germinação. Quando essa fonte é esgotada, entram em ação os lipídeos (ácidos graxos) presentes na semente (HILDEBRAND, 2011; XU; SHANKLIN, 2016). Essas vias também são precursoras de hormônios em humanos e terpenos em plantas. Nos dois casos são responsáveis pela sinalização para a ativação ou inibição de outras vias metabólicas (glicólise ou gliconeogênese). São utilizados também em plantas como forma de proteção, muitas vezes contra herbivoria. Em geral, essas vias são acionadas quando há abundância de nutrientes (glicose) ou há uma pressão ecológica (seca ou herbivoria) (MOGHE et al., 2017).

Os ácidos graxos formados por essas vias possuem inúmeras utilidades industriais como produção de sabões, lubrificantes, surfactantes, detergentes, estabilizantes, tratamento de efluentes, conservantes de alimentos e até mesmo em produtos farmacêuticos e cosméticos (ANNEKEN et al., 2006). Um dos maiores usos dos ácidos graxos são para a produção de biocombustível, por conter uma alta concentração de energia armazenada, diminuindo assim o uso de combustíveis fósseis (ZHOU et al., 2016).

1.6.2. **Ciências ômicas aplicada à engenharia metabólica**

As tecnologias "ômicas" trazem possibilidades de novas descobertas e maneiras de abordagens à questões do passado, presente e futuro. O sequenciamento de DNA vêm sendo parte fundamental da biologia, pois possibilita a obtenção de informações de organismos com base no seu genoma (GOODWIN; MCPHERSON; MCCOMBIE, 2016). Estudos genômicos trazem conhecimentos e possibilidades para a biotecnologia através do melhoramento de organismos e gerando novas linhagens comerciais (VARSHNEY et al., 2009). O melhoramento genético de organismos colabora com a indústria para otimizar e incluir processos para otimizar produtos e subprodutos, além de facilitar as cadeias de produção intermediárias e substratos utilizados na produção. A utilização de sequenciamento de alto desempenho possibilitou, por exemplo, o acesso à informações quanto à estrutura genética e possíveis candidatos para síntese, degradação e elongação de ácidos graxos (TEH et al., 2017).

A utilização de diferentes abordagens ômicas traz melhores resultados e mais confiabilidade dos dados. Um exemplo é a junção de múltiplas tecnologias de sequenciamento (Illumina, PacBio entre outras) que podem se complementar, e o uso da genômica, transcritômica e proteômica pode ser utilizado para validação e delineamento da parte experimental (MANZONI et al., 2018; MUTZ et al., 2013). Apesar das tecnologias de sequenciamento estarem mais acessíveis, ainda é cara a utilização de mais de uma abordagem, além da necessidade de um bom poder computacional e de pessoas capacitadas a executar estas análises multidisciplinares (HORNER et al., 2010). O principal gargalo nos estudos das “ômicas” está em relacioná-los com o sistema completo, onde a função do gene atua em vias e regulações metabólicas, não sendo possível isola-lo de todo o processo sistêmico (WAY et al., 2014).

1.7 **RNA-Seq: do sequenciamento à análise dos dados**

Segundo Harvey et al (2015) a técnica de RNA-Seq utiliza o sequenciamento de próxima geração e possibilita identificar e quantificar os RNAs mensageiros (mRNAs) presentes em um determinado momento no organismo. Isso revela o que foi transcrito de DNA para RNA para possível tradução posterior, a partir do dogma central da biologia molecular, onde DNA é transcrito para RNA e posteriormente traduzido em aminoácidos, que formam as proteínas. A partir dessas informações, podemos iniciar o processo de conhecimento sobre o que está acontecendo com o organismo quando há alguma alteração do seu estado normal, como por exemplo, um estresse hídrico. As informações sobre a expressão diferencial de genes, se há a expressão de determinado gene em determinadas condições ou não, associados com informações genômicas e fenotípicas podem ajudar com o melhoramento de organismos, uma vez que se tem informações sobre como alguns genes agem e se organizam para superar dificuldades ou produzir compostos (HARVEY; EDRADA-EBEL; QUINN, 2015). Esta técnica não permite saber se há de fato o produto proteico ou não, pois o RNA-Seq disponibiliza apenas informações sobre esses transcritos. Quando o genoma está disponível é possível determinar, através dos dados transcritômicos, os genes presentes após o RNA-Seq, os produtos gênicos alternativos (*splicing* alternativo) e até mesmo identificar novos genes, e assim associá-los com vias metabólicas e avaliar suas influências no metabolismo (MAMANOVA et al., 2010).

As análises pós-sequenciamento são fundamentais para a robustez dos resultados. O pré-processamento é caracterizado pela avaliação da qualidade do sequenciamento, retirado de adaptadores e exclusão de sequências de baixa qualidade. As sequências consideradas de alta qualidade precisam passar por algoritmos que as juntem formando transcritos, no caso do RNA-Seq ou *contigs* (Sequência genômica contígua da qual a ordem das bases possui um alto valor de confiabilidade) e *scaffolds* (uma porção da sequência genômica reconstruído a partir da junção de *contigs* e *gaps*), no caso de sequenciamento de DNA. Esses algoritmos são

chamados de montadores, que podem utilizar um genoma de referência, se houver, ou utilizar um método *de novo* caso nenhum genoma de referência esteja disponível. O alinhamento dos segmentos de sequências são definidos pelo *K-mers*, que significa o tamanho em que as *reads* serão divididas para construção do grafo de *brujin*, que representa a sobreposição das sequências de DNA utilizado nos algoritmos mais populares para a montagem de genomas e transcritomas (IQBAL et al., 2012). Ex: Com um *k-mer* = 3 e sequência "ATGGCGT" temos como possibilidades as seguintes sequências ATG, TGG, GGC, GCG e CGT, das quais seriam alinhadas umas contras as outras para a formação dos vértices do gráfico. Se dois *k-mers* se sobrepõem com exceção de um nucleotídeo, eles podem se ligar. A repetição desse processo culmina na montagem do genoma ou trascritoma (pelo gráfico *de bruijn)* (Figura 7) (COMPEAU; PEVZNER; TESLER, 2011). Esses valores necessitam seguir 2 regras básicas: o valor precisa ser ímpar, para que não ocorra palíndromos, e deverá sempre ser menor que o tamanho das sequências.

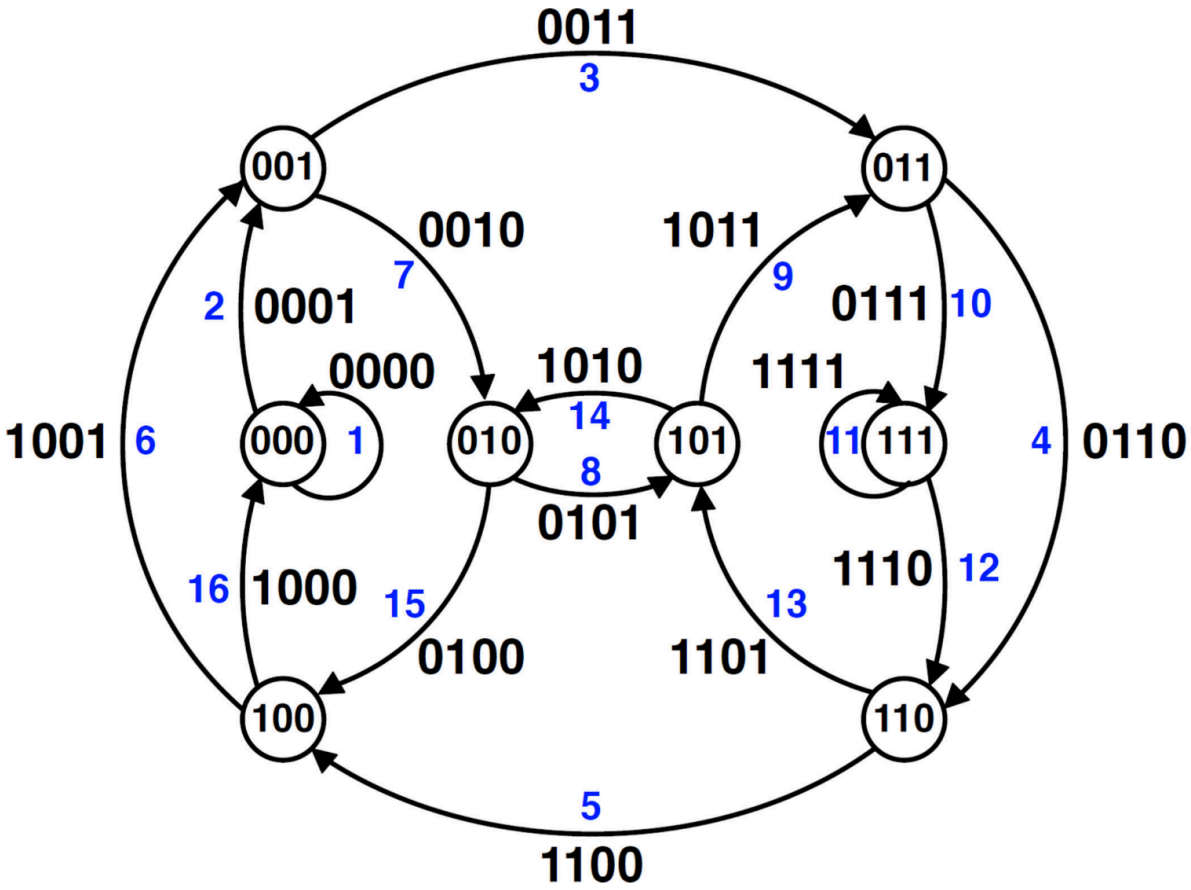


Figura 7 - Gráfico de Bruijn. Em um intervalo finito igual a quatro para cada palavra, sendo que o alfabeto é composto apenas de 2 letras (0 e 1). Este gráfico tem um ciclo Euleriano, sendo que cada nó pode interagir com 2 outros nós e recebe interação de 2 outros nós também. Seguindo a numeração em azul é possível observar na ordem de 1 ate 16 gera um ciclo Euleriano 0000, 0001, 0011, 0110, 1100, 1001, 0010, 0101, 1011, 0111, 1111, 1110, 1101, 1010, 0100, 1000, do qual gera ao final a sequencia 0000110010111101.

Fonte: Compeau, Pevzner e Tesler, 2011.

Valores altos de *k-mer* (acima de 45) podem gerar *contigs* menores por necessitarem de alinhamentos maiores (anelamento das sequencias entre si). Além disso, geralmente, precisam de maior poder de processamento e memória para armazenar as sequências. Porém, em certos casos, melhoram a montagem de partes repetitivas. Valores menores de *k-mers* possibilitam maiores chances de sobreposição, que aumentam as ocorrências de montagens ambíguas e reconstrução de genomas muito grandes e repetitivos, por exemplo, o genoma de eucalipto, podendo gerar assim erros na montagem. É importante que se ache um consenso entre o valor de *k-mer* para uma boa montagem, que pode ser realizado observando o

tamanho esperado do genoma e das sequencias obtidas através dos sequenciamentos, onde não haja muita redundância e que não se perca informações (ZERBINO, 2010).

Após gerar dados concisos e robustos, avaliando pela completude apresentada e com dados e informações da literatura, é necessário correlacioná-los a transcritos e/ou proteínas presentes em organismos, onde passamos pelo processo de anotação dos genes, que permite saber a localização desse gene ou proteína e quais são suas funções com base em informações previamente conhecidas em bancos de dados (Uniprot, NCBI e GO *consortium*) pela homologia das sequências. Há programas capazes de anotar esses genes automaticamente, alguns específicos a determinadas espécies, por terem a capacidade de identificar códons preferenciais aquela espécie (BLAST2GO, GO FEAT, *Companion,* DFAST).

1.8 **Analise geral do transcritoma**

As analises feitas com o transcritoma em geral buscam observar transcritos diferencialmente expressos em determinadas condições e a regulação de suas expressões, principalmente em estresse biótico (infecção) ou abiótico (hídrico), assim com estes estudos é possível obter informações de possíveis transcritos que se relacionam a mitigar os efeitos causados por esses estresses e ampliar as informações anteriormente obtidas, possibilitando assim o processo de melhoramento genético (COVINGTON et al., 2008; YAO et al., 2011). Esses dados obtidos devem ser corroborados posteriormente por analises experimentais, para que assim se possa confirmar a atuação dos transcritos encontrados.

Essas analises podem ser feitas de diferentes tecidos, que geram dados específicos as ocorrências dos transcritos sendo expressos em tecidos específicos da planta, Natarajan e Parani foram capazes de observar o aumento do nível de expressão de 28 transcritos relacionados a vias de metabólicas de ácidos graxos de *J. curcas*, utilizando analises do transcritoma de folhas maduras, flores, sementes em desenvolvimento, embrião e sementes maduras (NATARAJAN; PARANI, 2011).

Xia *et al* analisou o transcritoma de *Camellia oleifera* que possibilitou a anotação de transcritos a banco de dados que não estavam anotados anteriormente, dando a oportunidade de associa-los a vias de metabolismo de ácidos graxos, com as informações obtidas foi possível a reconstrução da via metabólica de síntese de ácidos graxos para essa espécie (Figura 8) (XIA et al., 2014).

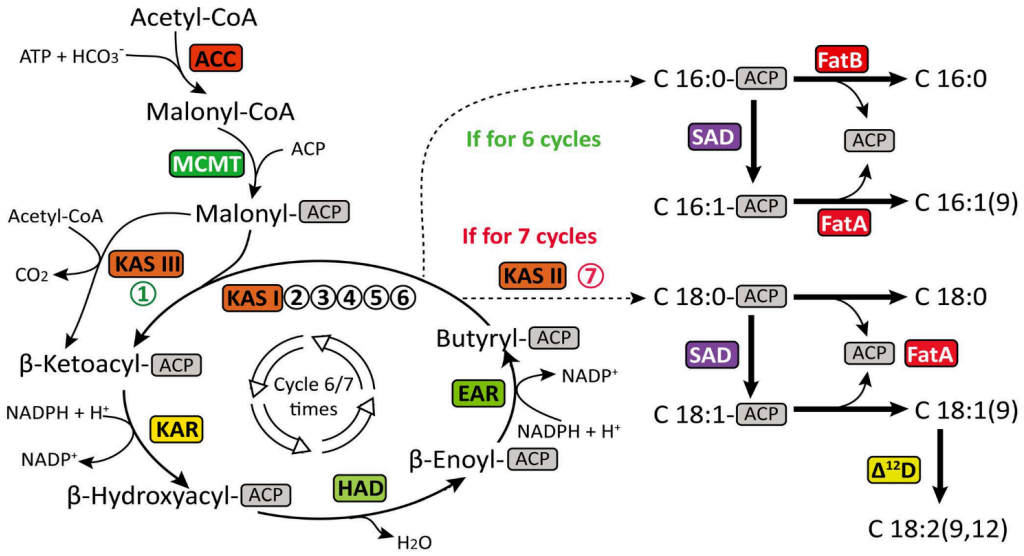


Figura 8 – Principais reações da biossíntese de ácidos graxos reconstruída a partir da informações obtidas pelo transcritoma de *Camellia oleifera*. Fonte: XIA *et al*, 2014.

2 Justificativa

Na agricultura, 94,6% da produção brasileira na safra de 2011/12 era composta por soja, enquanto outras plantas oleaginosas (algodão, mamona, dendê e girassol), que são utilizadas em linhas de produção (óleo de cozinha, ração animal, dentre outros), não possuíam espaço nas lavouras de acordo com dados do Ministério da Agricultura, Pecuária e Abastecimento (MAPA) (OSAKI, M. & BATALHA, 2011). Estas espécies pouco aproveitadas pela indústria podem ser usadas para a ração animal, produção de biocombustíveis e em setores da indústria de alimentos (MONTES; MELCHINGER, 2016). A importância destes cultivares se dá devido a sua versatilidade, como a possibilidade de aproveitar desde as sementes até suas flores e frutos.

J. curcas, *E. guineensis* e *R. communis* foram ótimos candidatos para a pesquisa deste estudo devido: capacidade de crescer em solo nacional; adaptação a condições contendo macronutrientes abundantes; tais como - como carbono, oxigênio, hidrogênio, fosforo, cálcio, magnésio e enxofre encontrados em condições naturais e limitações hídricas como as encontradas em regiões semi-áridas, bem como a possibilidade de fomentar a agricultura familiar a partir destes cultivares (NUNES, 2007). Estas plantas mostraram possuir propriedades químicas em seus óleos importantes para a indústria de cosméticos e de biocombustíveis chamando assim a atenção de grupos de pesquisa ao redor do mundo (COSTA et al., 2010; KUMAR; SHARMA, 2008; OGUNNIYI, 2006; OLIVARES-CARRILLO et al., 2016).

A escolha de vias metabólicas de ácidos graxos se dá pelo crescimento e da necessidade do mercado na captação de matéria prima como o óleo vegetal para a produção de lubrificantes, biodiesel e outros derivados mais sustentáveis do que os provenientes do petróleo, bem como possibilitar a diversificação da matriz energética onde há a predominância de soja e algodão, sendo as maiores produções brasileiras nos últimos anos.

No entanto, para o melhoramento futuro não só destas espécies, mas também de outras mais bem estabelecidas, novas informações genômicas e transcritômicas são necessárias, justificando este trabalho para elucidar possíveis gargalos com relação as vias relacionadas a ácidos graxos ou possibilitar o melhoramento genético espécies já domesticadas (soja, canola e girassol) com base nas informações obtidas pelas analises do transcritoma (BADOUIN et al., 2017; COSTA et al., 2010; SCHULZ et al., 2012; WANG et al., 2014).

3 Objetivos

3.1 Objetivo Geral

Explorar dados provenientes de RNA-Seq para identificar e anotar genes relacionados às vias metabólicas de produção, elongação e degradação das cadeias de ácidos graxos em *Jatropha curcas, Ricinus communis* e *Elaeis guineensis*.

3.2 Objetivos Específicos

- Obter os transcritomas das espécies *Jatropha curcas, Ricinus communis* e *Elaeis guineensis* através da comparação de seis programas de montagem (Velveth/Oases, Trinity-GG, HISAT2, SPAdes, SOAP e STAR) utilizando diferentes tamanhos de *k-mers*
- Criar um banco de dados local com proteínas de dez espécies de plantas importantes economicamente e academicamente associadas às vias metabólicas de ácidos graxos de interesse para este estudo
- Identificar e anotar os transcritos associados com as vias metabólicas de ácidos graxos
- Comparar os transcritomas com os genomas disponíveis e com o banco de dados do KEGG com relação às enzimas encontradas
- Verificar a completude dos transcritomas

4 Métodos

4.1 Extração e sequenciamento

A extração e sequenciamento de RNA total de sementes de cada espécie (*J. curcas, E. guineensis* e *R. communis)* foram realizados pela Macrogen, segundo o protocolo da empresa para obtenção do número de número e tamanho de sequencias contratados pelo Laboratório de Biologia Sintética da Embrapa Recursos Genéticos e Biotecnologia. O Laboratório de Biologia Sintética da Embrapa Recursos Genéticos e Biotecnologia foi o responsável pelo envio das sementes maduras para a realização dos procedimentos de extração e sequenciamento. O sequenciamento foi realizado em equipamento Illumina *HiSeq* 2500 protocolo *paired-end* (Figura 8) que gera sequências de cDNA com tamanho máximo de 150 pb (METZKER, 2010).

4.2 Pré-processamento e montagem dos transcritomas

O programa FASTQC avalia a qualidade das sequencias produzidas pelo sequenciamento e com essa finalidade foi executado com parâmetros padrão para produzir relatórios gráficos contendo as seguintes informações: qualidade das sequências por base, escores de qualidade por sequência, conteúdo da sequência por base, qualidade e conteúdo GC por base e por sequência, conteúdo de N por base, distribuição por tamanho de sequência, quantidade de sequências duplicadas e sequências super representadas (ANDREWS, 2014). Para fins de comparação, o programa também foi rodado com as sequências brutas, que são as sequencias que vieram do sequenciador sem nenhum pré-processamento feito.

Os programas *cutadapt* (versão 1.9) e fast-mcf (versão 1.04) (ARONESTY, 2011; MARTIN, 2013) foram usados em conjunto para identificar e retirar o adaptador universal da *Illumina*, utilizando as métricas padrões de cada um dos programas, e as sequências com

baixa qualidade através da clivagem por alinhamento. Estes programas geram resultados com sequências livres de adaptadores e com uma qualidade maior que a encontrada anteriormente.

Como o sequenciamento foi *paired-end*, o número de sequências é o dobro, pois são dois arquivos diferentes, um contendo as sequencias *reverse* e outro contendo a *forward* de um mesmo fragmento de cDNA que foi sequenciado, como pode ser observado na Figura 9.

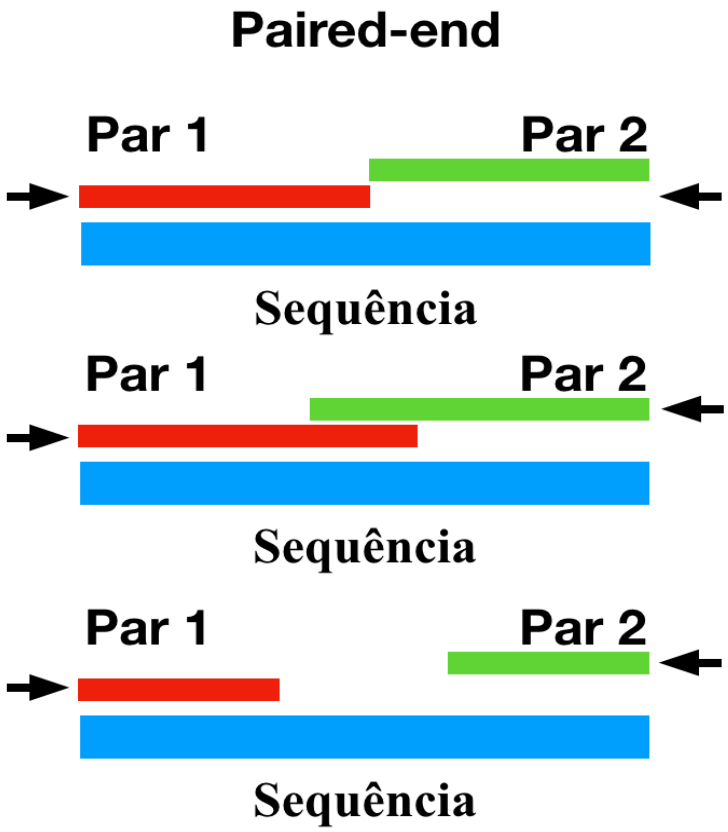


Figura 9 - Representação do protocolo paired-end. É possível observar que segundo o modelo de sequenciamento por síntese, é possível se obter o par senso e antisenso de uma mesma sequencia. A - Os pares se encontram no meio da sequencia. B - Os pares se sobrepõe o que facilita a montagem da sequencia completa. C – Os pares flanqueam a sequencia, que dificultam a terminação de toda sequencia a ser sequenciada.

Fonte: Elaboração do autor (2019).

Para a montagem dos transcritomas seis diferentes programas foram testados (Tabela 2). O HISAT2, o Trinity e o STAR foram usados para montagens com os genomas de referência. Os demais montadores geraram transcritomas *de novo*. Para cada programa (com excessão do HISAT2 e STAR), foram testados diferentes *k-mers* com o objetivo de avaliar a

melhor estratégia. Os *inputs* foram os arquivos de saída (*output*) do *cutadapt* (após retirada dos adaptadores). Para a montagem dos transcritos gerados pelo HISAT2 e STAR, foi necessário rodar o programa *cufflinks* (GHOSH; CHAN, 2016) a partir do arquivo .gtf gerado. Os resultados de cada montador foram arquivos de textos (formato fasta) contendo as sequências dos transcritos gerados. As linhas de comando e parâmetros de todas as etapas descritas acima encontram-se no Apêndice.

Tabela 2 - Programa utilizados para a montagem dos transcritos, os *k-mers* usados e o com o algoritmo do programa e protocolo

Assembler	Tamanho do K-mer	Algoritmo	Protocolo	Referência
Velveth/Oases	17,29,45	Grafo de Bruijn	*de novo*	SCHULZ et al., 2012
SPAdes	17,29,45	Grafo de Bruijn		BANKEVICH et al., 2012
SOAP	31	Grafo de String		XIE et al., 2014
HISAT2	*Default*	Grafo de String	Guiado pelo genoma	KIM; LANGMEAD; SALZBERG, 2015
Trinity	25	Grafo de Bruijn		BRUCE W.; FRIEDMAN, 2013
STAR	*Default*	STAR		DOBIN et al., 2013

Fonte: Elaborada pelo autor, 2019.

O programa *Evidential* Gene (2013 release) (CHEN et al., 2015) foi usado para processar os transcritomas gerados pelas diferentes abordagens dos montadores com o objetivo de eliminar erros e artefatos e aumentar a confiabilidade dos resultados. Ele foi desenvolvido especificamente para ser aplicado em estratégias como a deste trabalho: o uso de diferentes montadores e com *k-mers* variados para ampliar as possibilidades de geração de transcritos (GILBERT, 2019). O *input* foi a junção dos arquivos .fasta gerados pelos montadores. O programa filtra sequências com bases não identificadas (Ns), sem códon iniciador, quando duas sequências possuem o mesmo *locus*, loci duplicados e artefatos de montagem, os transcritos serão considerados de baixa qualidade quando se encaixarem em ao menos um desses parâmetros citados acima. Os códigos dos comandos utilizados encontram-se no Apêndice.

4.3 **Completude dos transcritomas**

O programa BUSCO (versão 3.0) foi usado para entender os transcritomas e avaliar o quão completo eles estão com relação ao banco de dados utilizado. Este programa busca por ortólogos de cópia única de 1.440 genes, ou seja, ele se utiliza apenas deste número de genes, em sua maioria considerados como *core* (existentes em todos os organismos). Quanto maior a porcentagem indicada pelo programa, mais completo o transcritoma está. Ele possui bancos de dados específicos para cada espécie, fornecendo um resultado mais acurado, para este trabalho foi utilizado o referente para plantas (embryophyta.db). O resultado do BUSCO valida o processo de montagem conferindo maior credibilidade (SIMÃO et al., 2015).

Os *inputs* utilizados foram os arquivos de saída (*outputs*) de cada um dos montadores, para cada *k-mer* diferente (17, 29, 31 e 45), e o resultado do *Evidential* gene totalizando 11 arquivos diferentes para cada espécie. O comando encontra-se no Apêndice.

O *output* gerado foi um arquivo de texto contendo a sumarização da informação sobre a completude encontrada, levando em consideração a duplicidade, fragmentação e ausência dos mesmos, trazendo assim informações do quão completo está o conjunto de dados, tendo como referência o banco de dados do próprio programa.

4.4 **Banco de dados local de sequências das vias de ácidos graxos**

Vias metabólicas relacionadas com síntese, elongação e/ou degradação de ácidos graxos em plantas foram selecionadas para serem estudadas neste trabalho segundo já citado na introdução deste estudo devido aos seus potenciais biotecnológicos (ERB; JONES; BAR-EVEN, 2017; ZHOU et al., 2016). As vias foram: biossíntese de ácidos graxos, elongação de ácidos graxos, degradação de ácidos graxos, biossíntese de esteróides, metabolismo de glicolipídeo, metabolismo de glicerofosfolipídeo, metabolismo de éter lipídico, metabolismo de ácido araquidônico, metabolismo de ácido linoleico e ɑ-linoleico, metabolismo de

esfingolipídios e biossíntese de ácidos graxos insaturados (ácido palmítico, oleico, bohemico).

As sequências de nucleotídeos e aminoácidos de proteínas de sete plantas amplamente estudadas (*Glycine max, Oryza sativa, Zea mays, Arabidopsis thaliana, Phoenix dactylifera, Eucalyptus grandis, Arabidopsis lyrata*), que serviriam de padrão para o estudo, e além das três espécies deste trabalho (*J. curcas, R. communis* e *E. guineenses*) pertencentes às 12 vias metabólicas acima citadas foram baixadas do repositório do KEGG (OGATA et al., 1999), totalizando 286 proteínas, sendo 250 únicas e 36 que estão presentes em mais de uma via. Foram desenvolvidos scripts *in-house* para a extração dos dados de interesse (identificador, sequência e EC number) através do uso dos códigos KEGG das vias metabólicas e das espécies (Tabela 3). O banco de dados foi criado com a finalidade de identificar os transcritos relacionados as vias metabólicas de ácidos graxos presentes no transcritoma, utilizando-se o programa *makeblastdb* (AGARWALA et al., 2018).

Essa árvore foi feita com o programa OrthoFinder (versão 2.2.3) (EMMS; KELLY, 2015), onde os proteomas das 10 espécies de plantas foram organizados em uma única pasta e o diretório foi indicado para o Orthofinder, que utiliza a metodologia de BLAST (all-versus-all). Os ortogrupos foram gerados utilizando RBNHs (*Reciprocal Best length - Normalised hit*) que define os limites da similaridade entre sequências, onde apenas serão aceitos para entrar no ortogrupo os que obtiverem um *score* igual ou maior ao RBNHs. Por último, foi feito o agrupamento dos genes em ortogrupos pelo algoritmo do grafo de MCL (*Markov Cluster Algorithm*) (EMMS; KELLY, 2015). O script usado encontra-se no Apêndice. O Orthofinder gerou, dentre diversos resultados, a árvore filogenética (Figura 10).

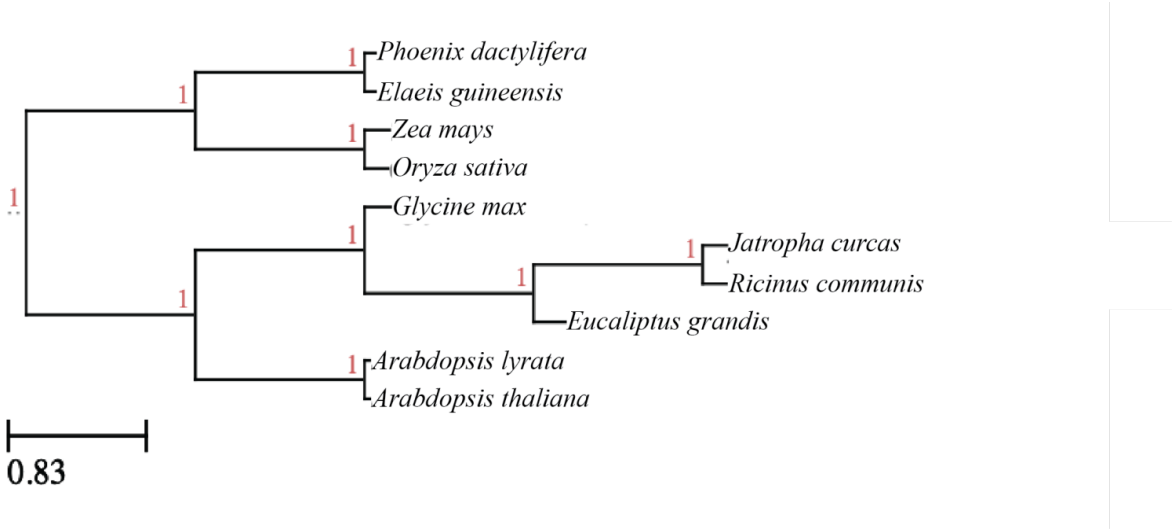


Figura 10 - Árvore filogenética das espécies utilizadas para a confecção do banco de dados, obtida através do Orthofinder.

Fonte: Elaboração do autor (2019).

Tabela 3 - Vias metabólicas e seus respectivos códigos KEGG.

Via metabólica	Código KEGG
Biossíntese de ácidos graxos	00061
Elongação de ácidos graxos	00062
Degradação de ácidos graxos	00071
Biossíntese de esteroides	00100
Metabolismo de glicolipídeo	00561
Metabolismo de glicerofosfolipídeo	00564
Metabolismo de éteres	00565
Metabolismo de ácido araquidônico	00590
Metabolismo de ácido linoleico	00591
Metabolismo de ácido alfa-linolênico	00592
Metabolismo de sphingolipids	00600
Biossíntese de ácidos graxos insaturados	01040

Fonte: Elaboração do autor (2019)

O banco de dados pode ser acessado pelo GitHub: https://github.com/ViniciusNattan/Mestrado-archive.

4.5 **Anotação das sequências e das vias metabólicas**

O Kobas *(KEGG Orthology Based Annotation System*) foi usado para anotar os transcritos de cada espécie com termos de ortologia Kegg (KO) usando, como referência, o proteoma de *A. thaliana*, que foi utilizada por ser uma planta modelo (AI; KONG, 2018; WU et al., 2006; XIE et al., 2011). Para isso, as sequências nucleotídicas dos transcritos foram traduzidas em sequências de aminoácidos usando o *EmbossTrasnseq*, utilizando o protocolo com todos os frames possíveis (CHOJNACKI et al., 2017) e carregadas no website. O organismo modelo e o banco de dados, no caso, o KEGG, foram selecionados. O resultado foi uma tabela com os transcritos, vias metabólicas anotadas e teste estatístico para o enriquecimento de vias metabólicas. As vias metabólicas enriquecidas foram aquelas que apresentaram o valor-p corrigido abaixo de 0,05.

O programa BLAST foi usado com o objetivo de identificar os transcritos associados ao metabolismo de ácidos graxos. Para isso, as sequências nucleotídicas dos transcritos e dos genomas foram alinhadas contra o banco de dados de sequências de proteínas pertencentes às vias de interesse. A linha de comando com os parâmetros encontra-se no Apêndice. Estes dados serviram de *input* para a anotação das vias metabólicas através do KEGG API (*application programming interface*), que permite a personalização das análises. Para isso foram usados os comandos da operação "GET" para obter as doze vias metabólicas de interesse existentes para cada espécie deste trabalho (https://www.kegg.jp/kegg/rest/keggapi.html). De posse dos modelos das vias metabólicas, o próximo passo foi usar o comando "SHOW_PATHWAY" (https://www.kegg.jp/kegg/rest/keggapi.html). O comando exige que sejam informados os códigos de enzimas (EC numbers), que foram obtidos através de scripts *in house* desenvolvidos para recuperar esta informação a partir do resultado do BLAST. O resultado foram imagens destacando, com cores, as enzimas presentes nas vias metabólicas em cada uma das espécies, tanto para o transcritoma quanto para o genoma. A informação foi

completada através do download das doze vias metabólicas de cada espécie, com as enzimas presentes do banco de dados do KEGG destacadas. As imagens foram obtidas no próprio site, nas páginas referentes a cada espécie. O programa Photoshop® foi usado para recriar as imagens das vias metabólicas a partir do modelo do KEGG para compilar os resultados finais dos transcritomas, genomas e KEGG permitindo, assim, que as informações fossem comparadas.

O fluxograma geral do trabalho está disposto na Figura 11.

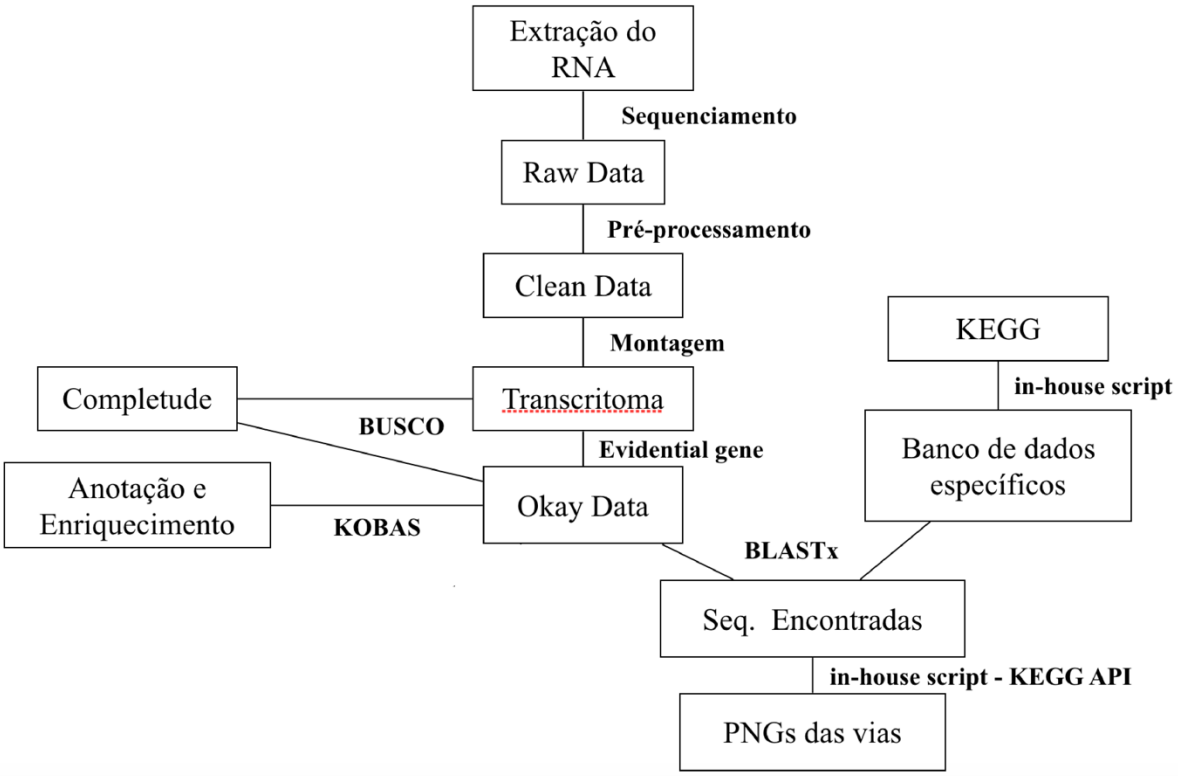


Figura 11 – Fluxograma completo da estratégia desenvolvida. É possível observar todas as analises feitas, e os passos seguidos.

Fonte: Elaboração do autor (2019)

5 Resultados

5.1 Sequenciamento e pré-processamento das sequências

O sequenciamento das sementes produziu 13,5, 13,8 e 11,0 milhões de sequências para *J. curcas*, *R. communis* e *E. guineensis*, respectivamente. O conteúdo GC variou entre 43% e 48% (Tabela 4). As sequências apresentaram boa qualidade, com valores de PHRED acima de 28 (faixa verde), como pode ser visto na coluna "Antes do pré-processamento" (Figura 12). Os protocolos de pré-processamento excluíram as sequências de baixa qualidade (Tabela 4), melhorando ainda mais os valores de PHRED, como pode ser visto na coluna "Depois do pré-processamento (Figura 12). O valor da qualidade média das sequências antes do pré-processamento foi de 36, 37 e 36 e após o pré-processamento foi cerca 37, 38 e 37 para *J. curcas, R. communis* e *E. guineensis,* respectivamente. Isso demonstra que as sequências estão em boa qualidade após o pré-processamento, devido a retirada das porções com baixa qualidade das *reads*, que se encontram principalmente no final do sequenciamento pelo acúmulo de erros ao longo do sequenciamento.

Tabela 4 - Dados gerais do RNA-Seq e resultados do pré-processamento

Espécies	Tamanho do Genoma	Tamanho da Sequência	Quantidade de Sequências	Nº de Sequências Retiradas	Porcentagem de GC
J. curcas	318Mb	100	13,537,243	250,364 (0.018%)	44%
R. communis	351Mb	100	13,801,603	187,170 (0,014%)	43%
E. guineensis	1.5Gb	100	11,040,757	288,159 (0,026%)	48%

Fonte: Elaboração do autor (2019).

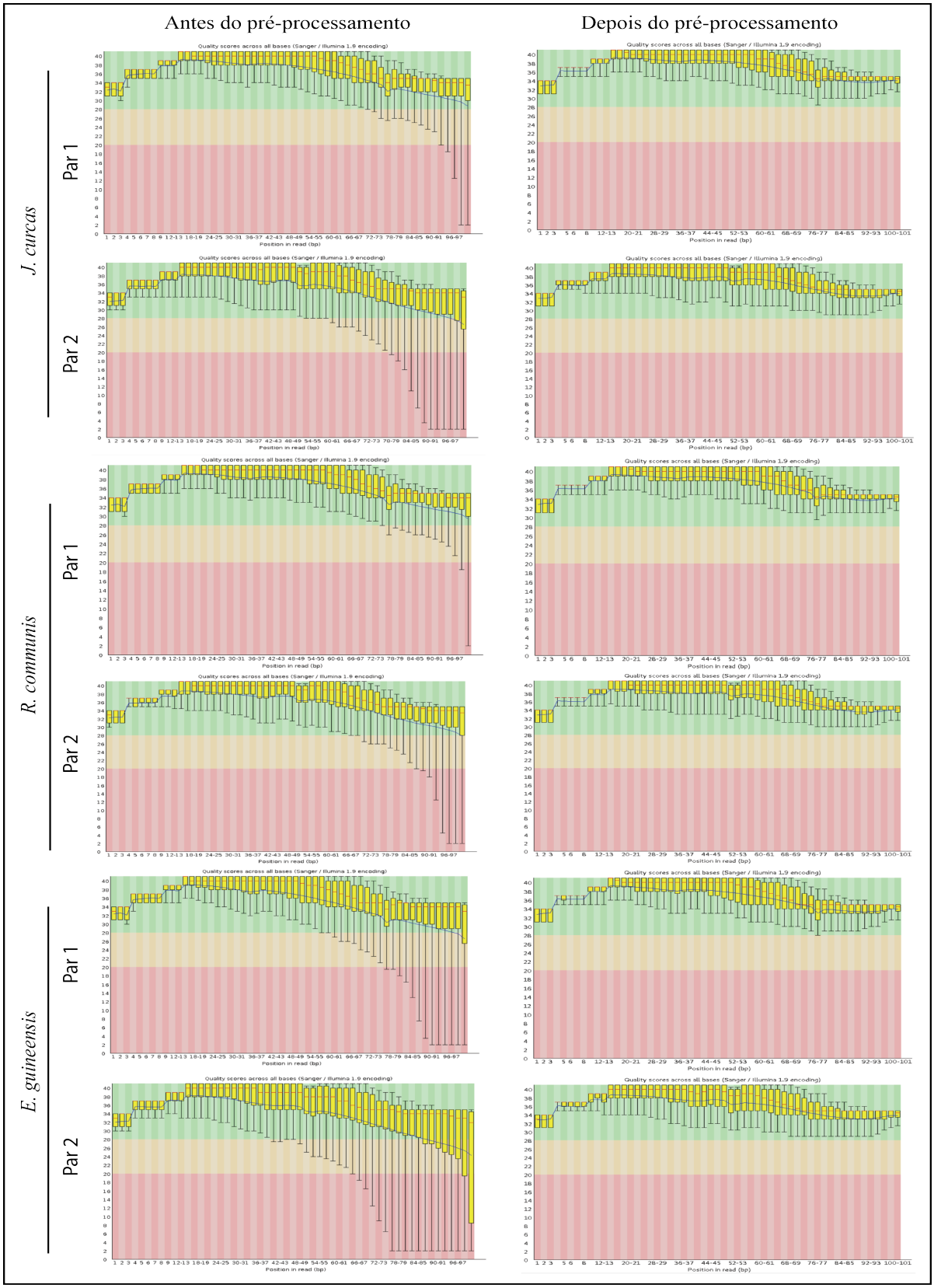


Figura 12 - Arquivo de saída do FASTQC do qual apresenta a qualidade das sequencias do par 1 e par 2 anterior e depois do pré-processamento. O eixo X representa o tamanho das sequencias (101 pares de base), enquanto o eixo Y representa o valor de PHRED. A faixa verde corresponde ao PHRED acima de 28, enquanto a amarela representa PHRED entre 20 e 28, e a faixa vermelha PHRED abaixo de 20.

5.2 **Transcritomas**

Ao todo foram utilizados seis montadores, dos quais geraram entre 32.823 e 135.475 transcritos de *J. curcas* (média de 66.957,4), entre 29.122 e 132.908 transcritos de *R. communis* (média de 55.612,2) e entre 33.521 e 137.334 transcritos de *E. guineensis* (média de 88.055,1). No geral os programas que usam genoma de referência geraram menos transcritos do que os que fizeram montagem *de novo*. Baseado no número total de transcritos gerados pelos seis montadores, o Evidental Gene foi capaz de reduzir entre 20 a 34 vezes (Tabela 8). Assim, a quantidade obtida de transcritos foi de 20.033, 16.309 e 33.189, enquanto a esperada que foi descrita pelos genomas disponíveis era de 23.076, 28.584 e 41.887 para *J. curcas*, R. *communis* e *E. guineensis* respectivamente.

O programa Velveth/Oases, com *k-mer* = 17, montou o maior número de transcritos, com quantidades acima de 130.000 para cada espécie, seguindo pelo SOAP, que montou 137.334 transcritos para *E. guineensis* com *k-mer* = 31. O montador que apresentou o menor número de transcritos foi o STAR com 32.823, 29.122 e 33.521 para *J. curcas, R. communis* e *E. guineensis* respectivamente (Tabela 5). Foram utilizados *k-mers* diferentes para avaliar como a variação iria afetar as montagens, para o SOAP foram utilizados diferentes *k-mers* também, mas não foi verificado nenhuma alteração nos *outputs* gerados.

Tabela 5 - Quantidade de transcritos gerados pelos diferentes montadores e após uso do Evidential Gene (EviGene)

Assembler	***K-mer***	***J. curcas***	***R. communis***	***E. guineensis***
		Número de transcritos		
Trinity	-	62.761	59.738	93.732
Velveth/Oases	17	135.475	132.908	132.369
	29	81.421	53.457	86.535
	45	51.452	42.801	58.578
Spades	17	60.395	60.362	95.414
	29	52.793	46.171	93.782
	45	43.596	38.768	71.44
HISAT2	-	41.247	34.303	38.923
SOAP	17, 31 e 45	107.611	58.49	137.334
STAR	-	32.823	29.122	33.521
Unity-Fasta[1]	*	669.574	556.12	880.551
EviGene	*	20.033	16.309	33.189

1 – Arquivo fasta contendo todos os transcritos das montagens. Verde – Número de transcritos após o uso do Evidential gene. Amarelo – Menor número transcritomas. Azul – Maior número de transcritomas.

Fonte: Elaboração do autor (2019)

5.3 Completude dos transcritomas

A plataforma BUSCO gerou métricas de completude dos transcritomas. Com os resultados foi possível avaliar as diferentes abordagens das montagens com base nos índices de completude a partir do banco de dados específico para plantas, permitindo a comparação entre os softwares, os diferentes *k-mers* e a ação do *Evidential gene* (EviGene).

As porcentagens de completude obtidas por cada programa variaram entre 17,7% e 73,8% para transcritos completos de *J. curcas*. Após o uso do Evidential Gene a porcentagem aumentou para 79,2%. O mesmo resultado se repetiu para as demais espécies, como pode ser visto na Tabela 6. É importante mencionar que as porcentagens de completude obtidas ficaram consideravelmente abaixo das apresentadas pelos genomas, que possuem valores acima dos 90%.

Tabela 6 - Completude obtida por montador e após uso do *Evidential* Gene (EviGene)

Assembler	*k-mer*	Completos			Fragmentados		
		J. curcas	*R. communis*	*E. guineensis*	*J. curcas*	*R. communis*	*E. guineensis*
Trinity	-	60,5%	65,4%	43,2%	14,9%	9,9%	12,3%
Velveth/Oases	17	17,7%	20,1%	12,3%	21,6%	21,2%	17,6%
	29	59,1%	62,4%	39,0%	14,0%	9,5%	11,5%
	45	52,9%	55,4%	32,5%	14,7%	11,8%	12,2%
Spades	17	25,4%	24,6%	12,2%	28,1%	25,9%	21,9%
	29	60,3%	60,5%	41,9%	15,8%	14,0%	14,5%
	45	61,3%	62,8%	41,5%	14,2%	11,9%	13,5%
HISAT2	-	73,8%	71,1%	50,7%	6,8%	6,0%	6,9%
SOAP	31	54,6%	59,6%	33,3%	18,3%	13,7%	17,2%
STAR	-	70,2%	64,7%	47,2%	6,5%	7,9%	6,7%
EviGene	*	79,2%	80,2%	61,4%	6,5%	4,5%	8,1%
Genoma[1]	*	96,2%	93,7%	92,7%	1,5%	2,5%	2,7%

1 – Genoma de referência. Verde – Completudes após o uso do Evidential gene. Amarelo – Menores completudes. Azul – Maiores completudes
Fonte: Elaboração do autor (2019).

A figura 13 apresenta a junção dos dados obtidos do número de transcritos e a completude encontrada para os montadores, genoma e pelo polimento com o Evidential Gene.

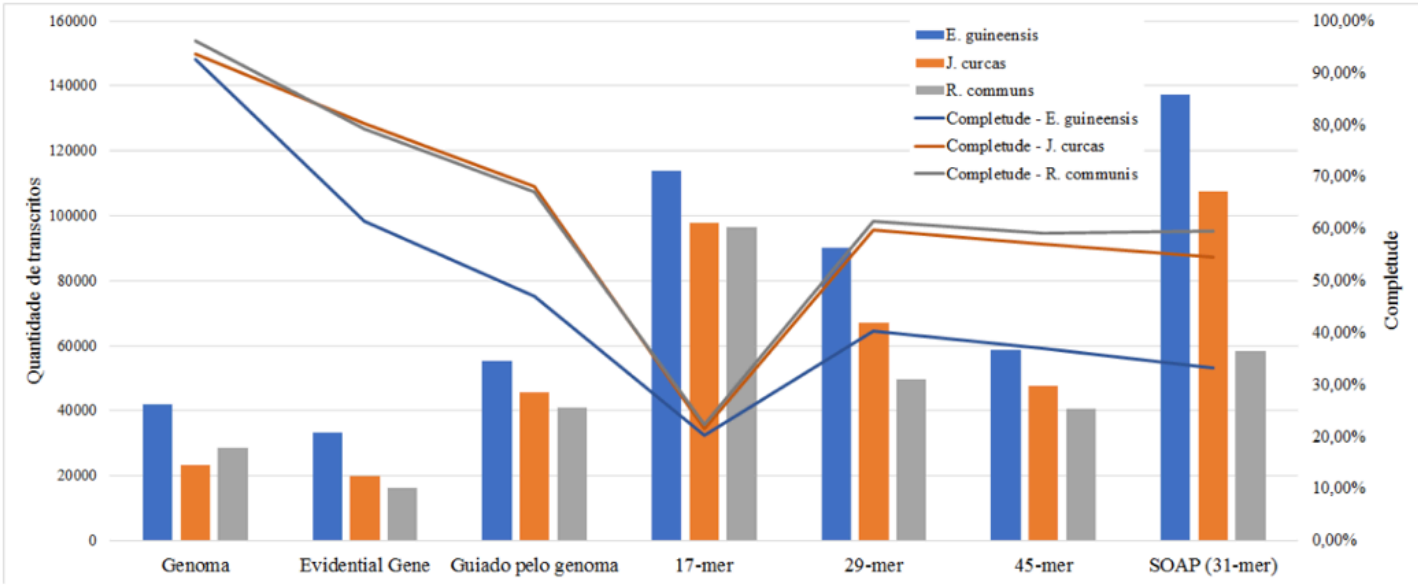


Figura 13 - Comparação entre os montadores e as completudes obtidas. É possível observar que os montadores de novo obtiveram grandes quantidades de transcritos, mas baixas completudes quando comparados com os montadores guiados pelo genoma. E o Evidential Gene obteve a mais completude dentre os montadores individualmente, com uma quantidade de transcritos menor.

Fonte: Elaboração do autor (2019).

5.4 Anotação das sequências

Um total de 15.433, 13.384 e 22.341 transcritos de *J. curcas, R. communis* e *E. guineensis* foram anotados, respectivamente, através do uso do Kobas. As anotações foram baseadas nas informações existentes de *A. thaliana*. As plantas *J. curcas* e *R. communis* apresentaram 13 e 32 vias metabólicas enriquecidas (valor-P corrigido < 0,05), incluindo duas e três vias associadas ao metabolismo de ácidos graxos, respectivamente (Tabela 7). A via metabólica "Metabolismo de ácidos graxos", que é a via geral, apresentou-se enriquecida em ambas as espécies. A via "Metabolismo de glicerolipídeo" foi enriquecida em J. *curcas*,

enquanto as vias "Biossíntese de ácidos graxos" e "Degradação de ácidos graxos" foram enriquecidas em *R. communis*. *E. guineensis* não apresentou nenhuma via metabólica enriquecida.

Tabela 7 - Vias metabólicas de ácidos graxos enriquecidas com base no banco de dados de *A. thaliana*

Via metabólica	**Identificador da via**	**No de sequências anotadas**	**No de sequências em A. thaliana (padrão)**	***J. curcas***	***R. communis***
				Valor P corrigido	
Metabolismo de ácidos graxos	ath01212	64 e 59	67	1,33E-02	5,34E-03
Metabolismo de glicerolipideo	ath00561	49	53	4,95E-02	X
Degradação de ácidos graxos	ath00071	37	41	X	2,84E-02
Biossíntese de ácidos graxos	ath00061	36	41	X	3,95E-02

Fonte: Elaboração do autor (2019).

Enfim, os transcritos e sequências genômicas anotados contra o banco de dados local, juntamente com as informações disponíveis no KEGG sobre as proteínas das três espécies foram compiladas com o objetivo de visualizar os elementos presentes nas vias metabólicas de ácidos graxos.

Um total de 110 transcritos de *J. curcas* e *R. communis* e 109 de *E. guineensis* foram anotados contra o banco de dados local de sequências de proteínas pertencentes às vias metabólicas de ácidos graxos extraídas do KEGG, sendo 107 comuns a todas as espécies (Figura 14). Nenhum transcrito foi exclusivo de alguma espécie. No entanto, 36 produtos ocorrem em mais de uma via, totalizando 146 ocorrências em *J. curcas*, sendo quatro destes transcritos anotados exclusivamente através da metodologia desenvolvida neste trabalho, enquanto anotamos apenas 93 sequências presentes no genoma (sendo que uma não foi encontrada no transcritoma). No entanto, o KEGG possui 158 proteínas destas espécies associadas à essas vias. Em *R. communis*, também encontramos 110 transcritos totalizando 146 ocorrências, sendo dois exclusivos, 118 sequências genômicas (duas não foram anotadas no transcritoma) e 162 proteínas presentes no KEGG para as mesmas vias. E por fim, em *E. guineensis*, foi possível anotar 109 transcritos e 148 ocorrências, sendo um exclusivo, 123

sequências genômicas e 159 proteínas do KEGG (Figura 14). Foram comuns aos transcritomas, aos genomas e ao KEGG 111, 102 e 78 produtos para *J. curcas, R. communis* e *E. guineensis* respectivamente (Figura 15). Não houve nenhum termo anotado de forma exclusiva para os genomas de referência em nenhuma das espécies. Os nomes das cinco enzimas encontradas exclusivamente nos transcritomas analisados encontram-se na Tabela 8.

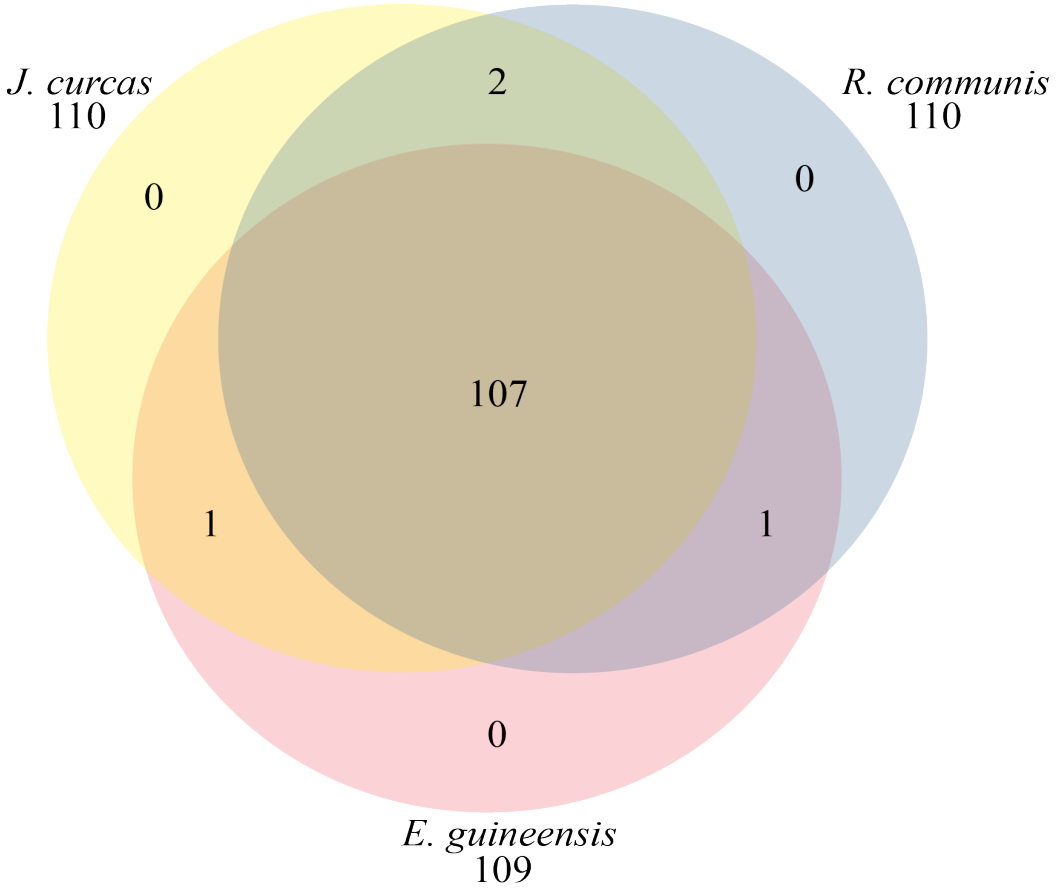


Figura 14 - Diagramas de Venn com o número de transcritos únicos anotados em vias metabólicas de ácidos graxos para cada espécie.

Fonte: Elaboração do autor (2019)

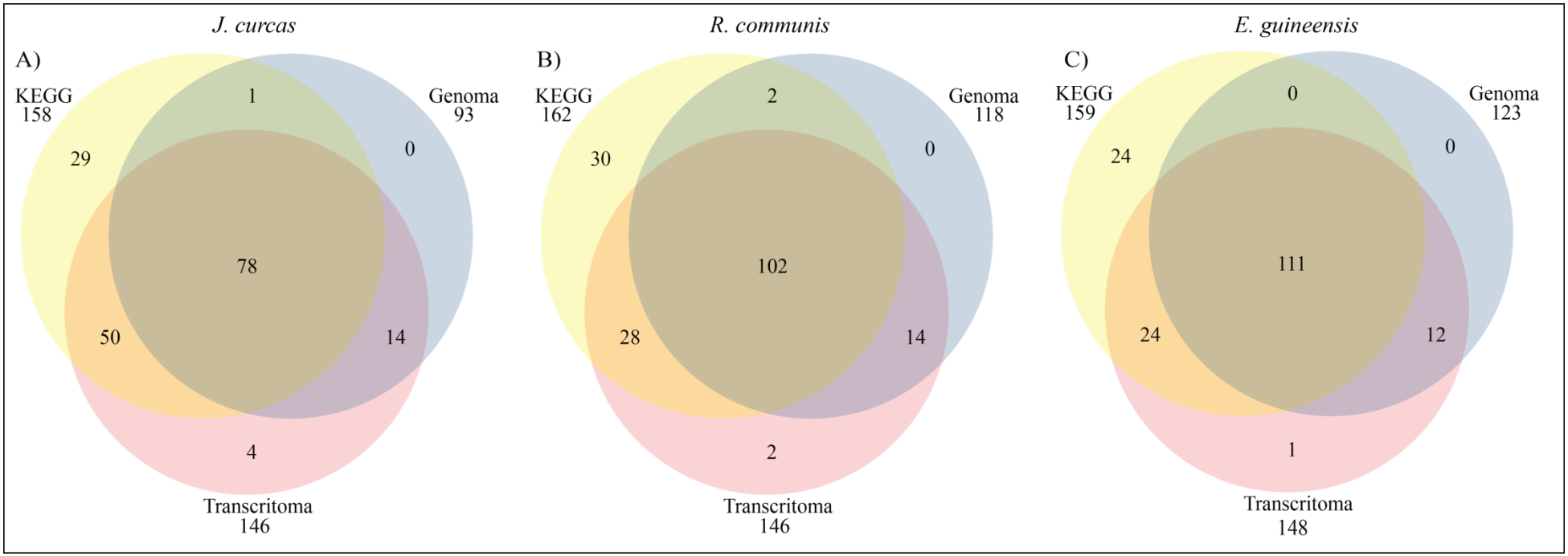


Figura 15 - Diagramas de Venn com o número de enzimas que foram encontradas para o banco de dados KEGG (amarelo), no genoma de referência (azul) e no transcritoma (vermelho) J. curcas (A), R. communis (B) e E. guineensis (C).

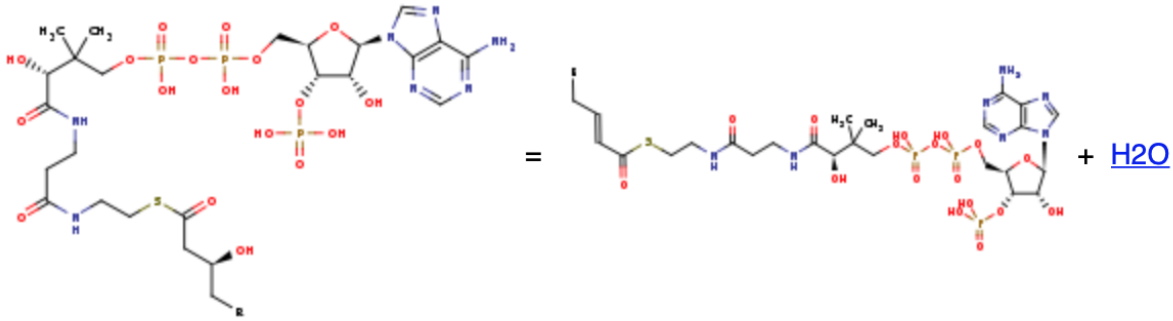
Fonte: Elaboração do autor (2019).

Tabela 8 - Enzimas encontradas que estavam ausentes nos genomas e no KEGG

Nomenclatura	Via metabólica	EC number	Espécie
Enoil-CoA hidratase	Elongação e degradação de ácidos graxos	4.2.1.17	*J. curcas*
Lisofosfolipase	Metabolismo de glicerofosfolipídeo	3.1.1.5	*J. curcas*
Fosfoetanolamine	Metabolismo de glicerofosfolipídeo	3.1.3.75	*J. curcas*
2-acilglicerol O-aciltransferase	Metabolismo de glicerolipídeo	2.3.1.22	*R. communis e J. curcas*
Fosfatidate fosfatase	Metabolismo de glicerolipídeo, glicerofosfolipídeo, esfingolipídeo e éter lipídeo	3.1.3.4	*R. communis e E. guineensis*

Fonte: Elaboração do autor (2019).

A Enoyl-CoA hydratase (Figura 16), encontrada apenas em *J. curcas*, está relacionada com a degradação de ácidos graxos e a inibição ou ausência desta enzima aumenta a concentração de ácidos graxos (ARENT et al., 2010). Em humanos esta diminuição da eficácia enzimática pode acarretar em esteatose hepática (IDE et al., 2017).



(3S)-3-hydroxyacyl-CoA = trans-2(or 3)-enoyl-CoA + H2O

Figura 16 - Reação catalisada pela enoil-CoA hidratase.

Fonte: BRENDA, 2019.

A Lisofosfolipase, enzima encontrada nos dados de RNA-Seq apenas em *J. curcas*, está relacionada com estresse oxidativo (GAO et al., 2010). O produto gerado (glicerofosfocolina) pode ser utilizado na indústria como surfactante e aditivo na indústria alimentícia, além de poder converter o substrato acumulado em ácido graxo (Figura 17).

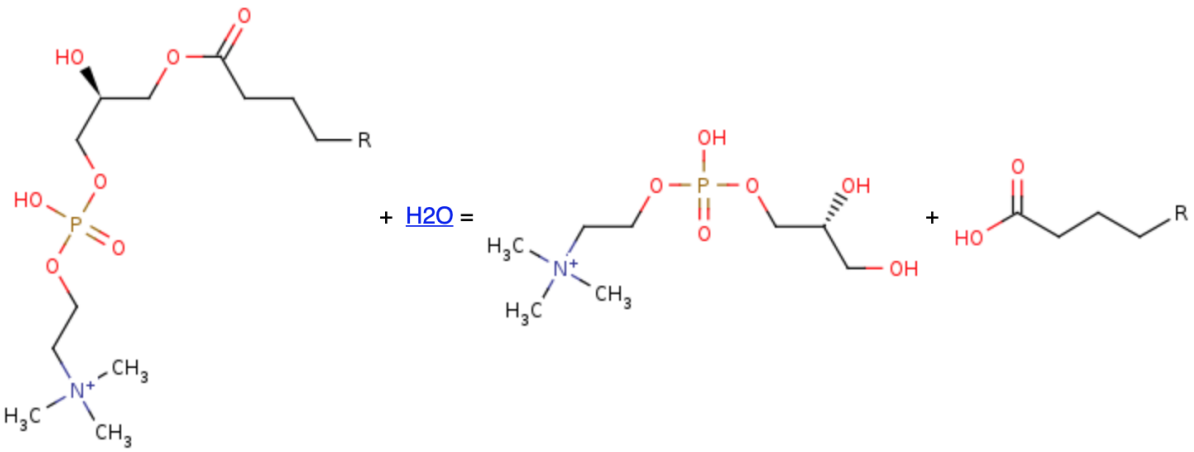


Figura 17 - Reação catalisada pela Lisofosfolipase.

Fonte: BRENDA, 2019.

A fosfoetanolamina, encontrada também em apenas *J. curcas*, é utilizada para reposição de fosfato para a degradação de fosfolipídios (MAY; SPINKA; KÖCK, 2012). Já em mamíferos é possível observar tal enzima na formação óssea, trabalhando também no recrutamento de fosfato (ROBERTS et al., 2004) (Figura 18).

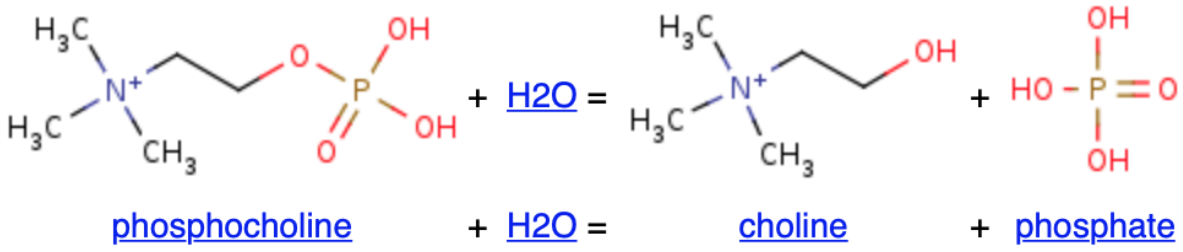


Figura 18 - Reação catalisada pela Fosfoetanolamine.

Fonte: BRENDA, 2019.

A 2-acilglicerol O-aciltransferase, enzima encontrada em *J. curcas* e *R. communis*, foi identificada, purificada e caracterizada em amendoim (*Arachis hypogaea*). Esta enzima está associada à via metabólica de MAG (*monoacylglicerol*), que por sua vez está ligada à regulação na transdução do sinal e à síntese de lipídeos complexos (TUMANEY; SHEKAR;

RAJASEKHARAN, 2001). Suas aplicações estão voltadas para a área médica no combate a obesidade e diabetes (BARLIND et al., 2013) (Figura 19).

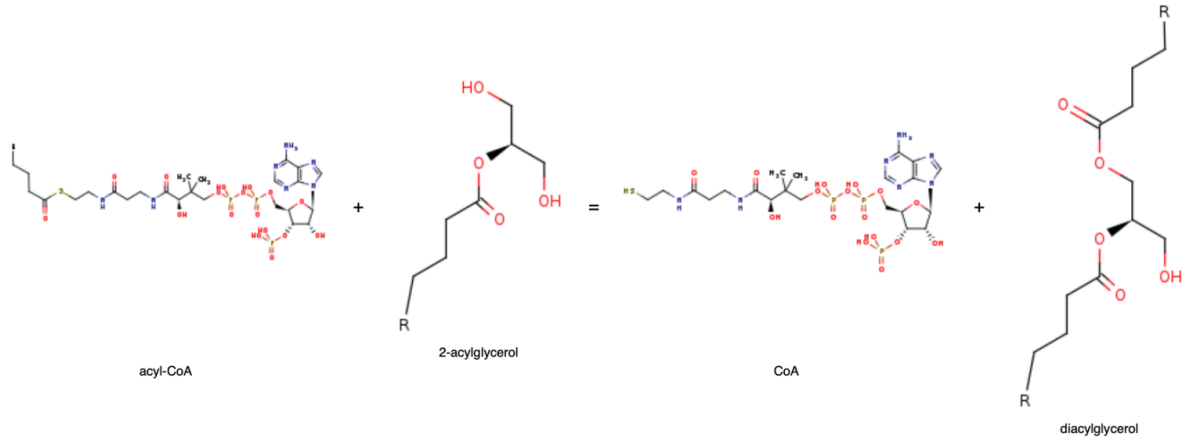


Figura 19 - Reação catalisada pela 2-acilglicerol O-aciltransferase.

Fonte: BRENDA, 2019.

A enzima Fosfatidate fosfatase, encontrada em *R. communis* e *E. guineensis*, possui um papel importante na acumulação lipídica e na transdução do sinal, possibilitando o acúmulo de DAG (diacyl-glycerol), que é um importante passo na via de Kennedy (uma das vias de ácido graxo). Não está claro o mecanismo de ação, nem quais ortólogos possuem maior eficácia no acúmulo de DAG (DENG; CAI; FEI, 2013). A perda de atividade desta enzima pode causar esteatose hepática em mamíferos (HARRIS et al., 2007) (Figura 20).

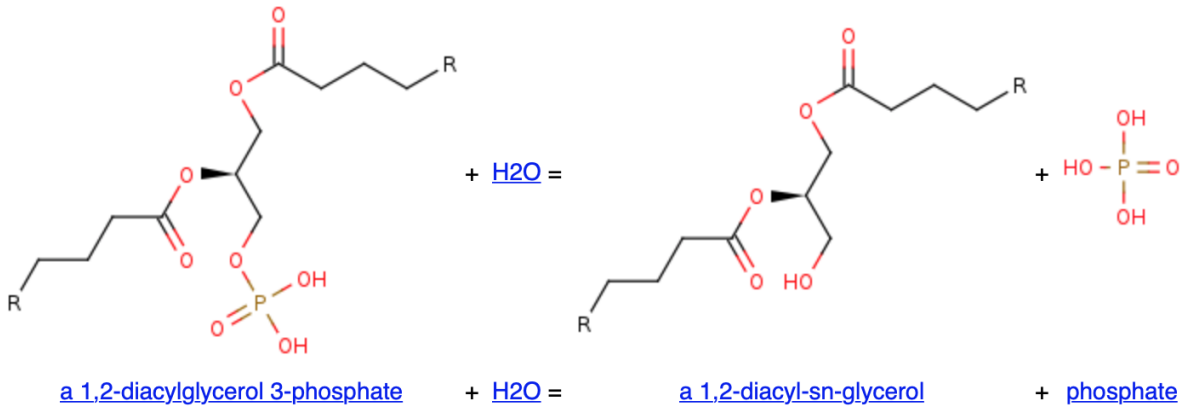


Figura 20 - Reação catalisada pela Fosfatidate fosfatase.

Fonte: BRENDA, 2019.

5.5 Anotação das vias metabólicas de ácido graxo

O extenso trabalho de anotação dos transcritos, juntamente com a inclusão dos dados genômicos e informações do KEGG nos permitiu comparar as vias metabólicas de ácidos graxos entre as espécies. Os mapas metabólicos do KEGG mostram todas as reações já identificadas em ao menos um organismo, independentemente do reino a que pertencem. Algumas vias, em plantas, são bem completas, como a biossíntese, elongação e degradação de ácidos graxos. Estas vias são algumas das principais dentre as 12 analisadas que são associadas à produção de biocombustíveis. Outras vias, como metabolismo de éteres e de ácidos linoleicos apresentam poucas reações identificadas em plantas (Tabela 9).

Tabela 9 – Número de reações de cada via metabólica, presentes no banco de dados e anotados a cada espécie

Nome (Código da via)	Número de reações		Espécies		
	Todos os organismos	**Em plantas (banco de dados local)**	***J. curcas***	***R. communis***	***E. guineensis***
Biossíntese de ácidos graxos (00061)	50	43	42	43	42
Elongação de ácidos graxos (00062)	35	28	26	27	27
Degradação de ácidos graxos (00071)	48	24	25	26	20
Biossíntese de esteróides (00100)	49	31	31	31	30
Metabolismo de glicolipídeo (00561)	37	20	20	19	16
Metabolismo de glicerofosfolipídeo (00564)	77	41	38	41	33
Metabolismo de éteres (00565)	32	8	5	7	7
Metabolismo de ácido araquidônico (00590)	52	16	15	16	15
Metabolismo de ácido linoleico (00591)	14	5	4	5	5
Metabolismo de ácido alfa-linolênico (00592)	26	19	19	18	19
Metabolismo de sphingolipids (00600)	36	23	23	23	23
Biossíntese de ácidos graxos insaturados (01040)	21	11	11	11	11

Células verdes: Totalidade das reações preditas em plantas que foram encontradas. Células vermelhas: Número de reações encontradas foram maiores que as esperadas.
Fonte: Elaboração do autor (2019).

Ainda assim, encontramos diferenças entre as anotações dos transcritos das três espécies deste trabalho. Com exceção da via "Biossíntese de ácidos graxos insaturados", em todas as demais houve diferença na presença de 62 proteínas, que estão presentes em ao menos um transcritoma dentre 250 proteínas únicas (Quadro 1), neste quadro é possível

comparar as três espécies do estudo com relação ao que está presente ou ausente no KEGG, genoma e transcritoma.

Trinta e cinco das 62 proteínas estão ausentes no genoma de *J. curcas*. Para efeito de comparação, 12 e 11 proteínas não foram anotadas nos genomas de *R. communis* e *E. guineensis* respectivamente. Uma possível explicação seria problema na montagem deste genoma.

As proteínas FabZ, citocromo bifuncional P450/NADPH--P450 redutase, lisofosfolipideo aciltransferase, prostaglandina-E sintase e acilglicerol lipase não foram encontradas em *E. guineensis*. Cicloeucalenol cicloisomerase, Diacilglicerol O-aciltransferase 2 e fosfatidilglicerofosfatase foram anotadas no genoma e estão presentes no KEGG desta espécie, mas ausentes do transcritoma.

A proteína Citocromo P450 4A22 não foi encontrada tanto nos genomas quanto nos transcritomas de *J. curcas* e *R. communis*. A proteína fosfolipase A2 está ausente no genoma e transcritoma de *J. curcas*. Por fim, a proteína fosfatase pirofosfate-especifica foi anotada em todos os transcritomas, mas não foi encontrada nos genomas e nem na base KEGG em *J. curcas* e *R. communis*.

Considerando as quatro vias metabólicas de interesse para a indústria de biocombustível (biossíntese, elongação e degradação de ácidos graxos e metabolismo de glicerolipídeo), os resultados mostraram que a metodologia desenvolvida foi capaz de encontrar, no total, 18 enzimas ausentes no genoma anotado de *J. curcas*, sete em *R. communis* e três em *E. guineensis*.

Quadro 1 - Comparação entre *J. curcas*, *R. communis* e *E. guineensis* em relação à presença ou não das enzimas no KEGG, transcritoma e genoma.

Via metabólica	Nome	Codigo / Sigla	*J. curcas*	*R. communis*	*E. guineensis*
Biossíntese de Ácidos graxos	Malonyl CoA-acyl carrier protein transacylase	FabD			
	3-oxoacyl-[acyl-carrier-protein] synthase 2	FabF			
	3-oxoacyl-[acyl-carrier-protein] synthase 3	FabH			
	3-hydroxyacyl-[acyl-carrier-protein] dehydratase	FabZ			
	Very-long-chain 3-oxoacyl-CoA reductase	1.1.1.330			
	Peroxisomal trans-2-enoyl-CoA reductase	1.3.1.38			
	Palmitoyl-protein thioesterase	3.1.2.22			
	Very-long-chain (3R)-3-hydroxyacyl-CoA dehydratase	4.2.1.134			
Degradação de Ácidos graxos	Bifunctional cytochrome P450/NADPH--P450 reductase	1.14.14.1			
	Cytochrome P450 4A22	1.14.14.80			
	Medium-chain specific acyl-CoA dehydrogenase	1.3.8.7			
	Acetyl-CoA acetyltransferase	2.3.1.9			
	Peroxisomal fatty acid beta-oxidation multifunctional protein	5.3.3.8			
Biossíntese de esteróides	Delta(7)-sterol-C5(6)-desaturase 1	1.14.19.20			
	Delta(14)-sterol reductase	1.3.1.70			
	3-beta-hydroxysteroid-Delta(8),Delta(7)-isomerase	5.3.3.5			
	Cycloeucalenol cycloisomerase	5.5.1.9			
	Cytochrome P450 710A1	CYP710A			
	Peptidyl-prolyl cis-trans isomerase	FK			
	cholestenol Delta-isomerase	HYD1			
	Cycloartenol-C-24-methyltransferase	SMT1			
	Delta(7)-sterol-C5(6)-desaturase	STE1			
Metabolismo de glicolipídeo	Aldo-keto reductase	1.1.1.21			
	Diacylglycerol O-acyltransferase 2	2.3.1.22			
	Monogalactosyldiacylglycerol synthase	2.4.1.46			
	Dihydroxyacetone kinase	2.7.1.29			
	glycerol kinase	2.7.1.30			
	Sulfoquinovosyl transferase	SQD2			
Metabolismo de glicerofosfolipídeo	Glycerol-3-phosphate 2-O-acyltransferase 6	2.3.1.198			
	CDP-diacylglycerol--inositol 3-	2.7.8.11			
	CDP-diacylglycerol--glycerol-3-phosphate 3-	2.7.8.5			
	lysophospholipase	3.1.1.5			
	phosphatidylcholine 1-acylhydrolase	3.1.1.32			
	phosphatidylglycerophosphatase	3.1.3.27			
	Pyrophosphate-specific phosphatase	3.1.3.75			
	Phosphatidylserine decarboxylase proenzym	4.1.1.65			
	lysocardiolipin acyltransferase	LCLAT			
	lysophospholipid acyltransferase	LPCAT			
	lysophosphatidylcholine acyltransferase	LPEAT			
	Lysophospholipid acyltransferase 7	LPIAT			
Metabolismo de Ácido araquidônico	carbonyl reductase 1	1.1.1.184			
	carbonyl reductase 1	1.1.1.189			
	glutathione peroxidase 3	1.11.1.9			
	leukotriene-A4 hydrolase	3.3.2.6			
	prostaglandin-E synthase	5.3.99.3			
	long-chain fatty acid omega-monooxygenase	CYP4A			
	long-chain fatty acid omega-monooxygenase	CYP4A11			
Metabolismo de Ácido linoleico	Linoleate 9S-lipoxygenase 1	1.13.11.58			
Metabolismo de Ácido alfa-linolênico	Jasmonate O-methyltransferase	2.1.1.141			
	Allene oxide synthase	4.2.1.92			
	Allene oxide cyclase	5.3.99.6			
	enoyl-CoA hydratase/3-hydroxyacyl-CoA	MPF2			
Metabolismo de sphingolipids	3-dehydrosphinganine	1.1.1.102			
	Sphingolipid delta(4)-desaturase DES1-like	1.14.19.17			
	Acyl-CoA-dependent ceramide synthase	2.3.1.24			
	sphingosine kinase	2.7.1.91			
	shingomyelin synthase	2.7.8.27			
	alkaline ceramidase	3.5.1.23			
Metabolismo de glicolipídeo e Metabolismo de glicerofosfolipídeo	acylglycerol lipase	3.1.1.23			
Metabolismo de glicolipídeo, Metabolismo de esfingolipídeo	phosphatidate phosphatase	3.1.3.4			
Metabolismo de glicerofosfolipídeo, Metabolismo de Éteres, Metabolismo de Ácido araquidônico, Metabolismo de Ácido linoleico e Metabolismo de Ácido alfa-linolênico	secretory phospholipase A2	3.1.1.4			
Elongação de Ácidos graxos e Degradação de Ácidos graxos	enoyl-CoA hydratase	4.2.1.17			

Células amarelas: Ocorrência no KEGG. Células vermelhas: Ocorrência no transcritoma. Células azuis: Ocorrência no genoma.

Fonte: Elaboração do autor (2019).

A Figura 21 apresenta uma das vias, como forma de exemplificação. No Apêndice encontram-se as figuras das demais vias para as três espécies. As figuras das vias podem ser visualizadas com melhor resolução no repositório no GitHub (https://github.com/ViniciusNattan/Mestrado-archive). Nas figuras o código JCU, RCU e EGU são para *J. curcas, R. communis* e *E. guineensis*, respectivamente. As células brancas correspondem a ausência do transcrito no genoma, transcritoma e no banco de dados KEGG, enquanto as células amarelas correspondem a presença do transcrito apenas no KEGG, as células azuis apresentam a ocorrência do transcrito apenas no genoma, células vermelhas representam que o transcrito está presente apenas no transcritoma. É possível observar também a presença do transcrito no genoma, transcritoma e KEGG, que está representada nas células de cores azuis, vermelhas e amarelas. As células vermelhas e azuis demonstram que o transcrito está presente no transcritoma e no genoma apenas, enquanto as células vermelhas e amarelas descrevem que o transcrito ocorre no transcritoma e no KEGG.

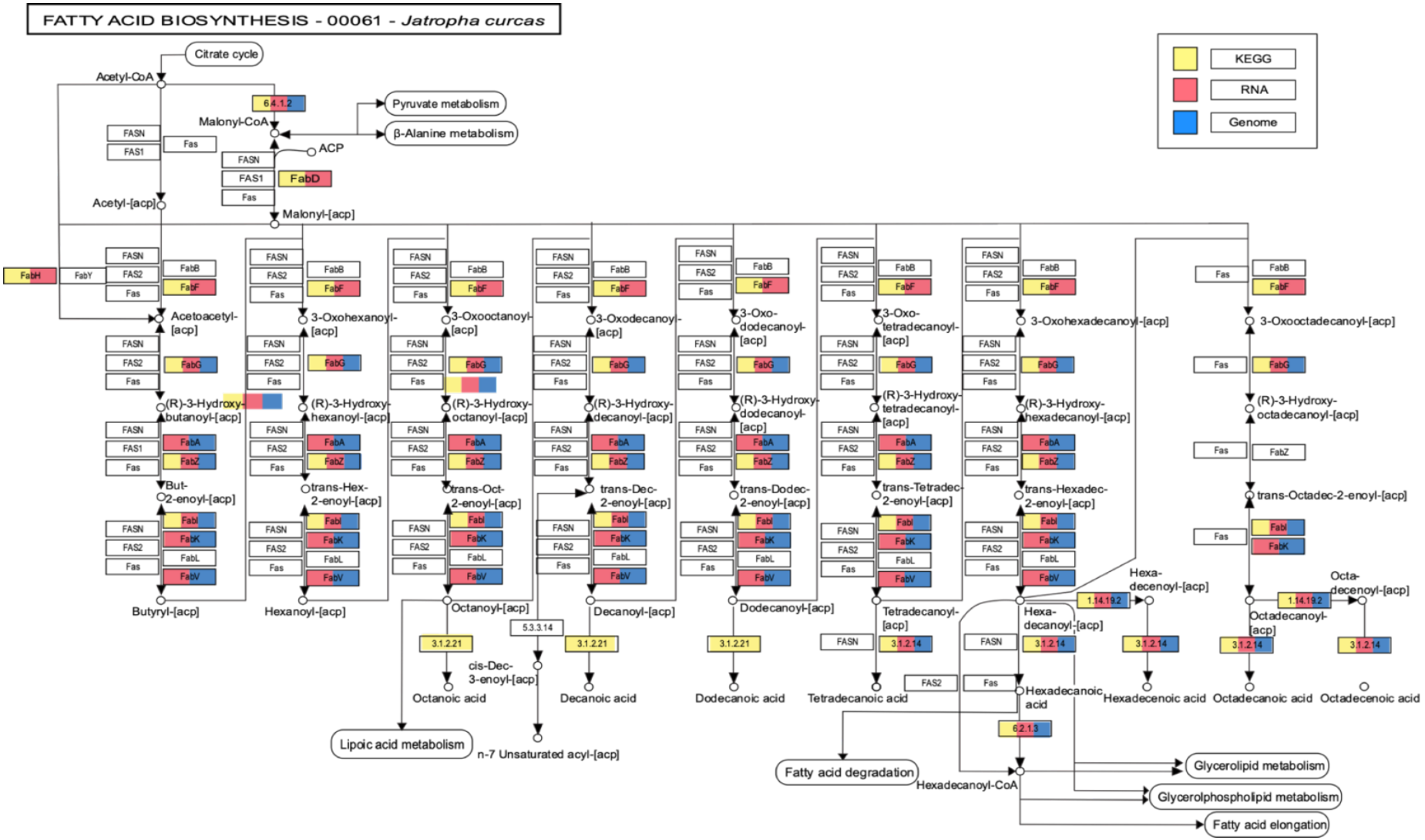


Figura 21 – Via de Biossíntese de ácidos graxos indicando as enzimas de *J. curcas* presentes nos dados do KEGG, transcritoma e/ou genoma

Fonte: Elaboração do autor (2019).

6 Discussão

Neste trabalho foram analisados os transcritomas de sementes maduras de três espécies de plantas de interesse biotecnológico: *J. curcas*, *R. communis* e *E. guineensis*. A estratégia desenvolvida possibilitou a identificação de cinco produtos gênicos conservados até então desconhecidos nestas espécies. Os dados dos transcritomas, juntamente com as informações disponíveis dos genomas e base de dados KEGG foram compilados para a realização da anotação de vias metabólicas de ácidos graxos.

A disponibilização de dados genômicos e transcritômicos através do uso de tecnologias de sequenciamento de nova geração tem possibilitado a escolha de genes candidatos que podem ser utilizados em estudos de biologia sintética, engenharia metabólica e melhoramento genético, sendo que estas áreas do conhecimento podem ser responsáveis pelos próximos avanços na produção industrial (LEE et al., 2008; SABLOK et al., 2014). Para que este avanço ocorra é imprescindível que novos conhecimentos acerca do metabolismo de espécies de interesse industrial sejam obtidos e que processos sejam melhorados. A metodologia presente neste trabalho mostrou-se eficiente e encontrou cinco novos genes de vias de metabolismo de ácidos graxos que podem vir a ser candidatos para programas de melhoramento de plantas oleaginosas. A metodologia desenvolvida pode, ainda, ser aplicada a espécies que não possuem genoma de referência.

No geral, os sequenciamentos produziram sequências com boa qualidade. Ainda assim, o uso de pré-processadores como o fastq-mcf e cutadapt, uma etapa obrigatória em qualquer *pipeline* de análise de dados provenientes de metodologias de sequenciamento de nova geração, melhorou a qualidade das sequências e a confiabilidade dos dados, já que é sabido que há perda de qualidade devido ao acúmulo de erros no final da corrida quando se utiliza sequenciadores de nova geração, que podem ser minimizados com uma boa

montagem de biblioteca, e a utilização de pré-processadores melhoraram a qualidade das sequencias obtidas (SCHIRMER et al., 2015; SHIN; PARK, 2016).

Seis programas foram usados para as montagens dos transcritomas. Três usam genomas de referência como guia para os alinhamentos e consequente montagem (HISAT2, Trinity e STAR) e três fazem montagem *de novo* (Velveth/Oasis, Spades e SOAP). Há a possibilidade de se utilizar outros montadores no processo de montagem, os montadores citados foram utilizados por serem amplamente utilizados na literatura para a montagem de diversos organismos diferentes e terem pela sua utilização pelo polimento com *Evidential* Gene (GENIZA; JAISWAL, 2017; GILBERT, 2019; OCKENDON et al., 2016; VISSER et al., 2015). Os programas claramente produziram diferentes resultados, independentemente do parâmetro de tamanho de *k-mer* usado. No geral, aqueles que requerem um genoma de referência produziram transcritomas mais enxutos do que os que aplicaram uma abordagem *de novo*, devido ao fato das montagens guiadas pelo genoma apenas fazerem o alinhamento com o genoma de referência, diminuindo assim o número de transcritos gerados erroneamente e artefatos, e apenas os transcritos que foram anotados estariam presentes no genoma de referencia (MARTIN; WANG, 2011). Nenhum dos seis programas foram eficientes em gerar transcritomas com porcentagem de completude próximas as encontradas pelo genoma. Por isso optamos por usar o *Evidential* Gene, que filtra os possíveis erros da montagem dos transcritos e elimina a redundância, a melhor combinação foi a utilização de todos os montadores, pois devido as especificidades de cada programa não seria perdida nenhuma informação presente. Essa abordagem foi eficiente, pois a completude que chegou mais próxima a do genoma foi o transcritoma filtrado pelo *Evidential* Gene, assim sendo o transcritoma escolhido para dar sequência no trabalho. Corroborando com os nossos resultados, *Nakasugi et al.* (2014) utilizaram quatro montadores *de novo* para dados de *Nicotiana benthamiana*. O *Evidential* Gene foi

responsável pela obtenção de transcritos que apresentaram melhores resultados de similaridade contra os bancos de dados e queda no número de sequências redundantes.

As analises de completude se aplicam ao estudo de transcritoma como é explicitado pelo próprio artigo do BUSCO (SIMÃO et al., 2015). Haak *et al* utilizou essa métrica como estratégia de avaliação da completude do transcritoma de *Croton tiglium* (HAAK et al., 2018). Rana *et al* também utilizou o BUSCO como métrica para avaliar se seu transcritoma montado com Trinity, SPAdes e SOAP estavam em montados e se obtiveram altas porcentagens de completude em relação ao próprio banco de dados do programa (RANA et al., 2016). Bryant *et al* utilizou os as analises de completude para corroborar com os fatores de regeneração encontrados por seu trabalho, conferindo assim uma maior confiabilidade nos dados e analises obtidas.

Ainda assim, os trancritomas tiveram completudes menores (79,2% para *J. curcas*, 80,2% para *R. communis* e 61,4% para *E. guineensis*) quando comparados com os respectivos genomas (96,2% para *J. curcas*, 93,7% para *R. communis* e 92,7% para *E. guineensis*). Uma possível explicação é que, como apenas as sementes maduras foram sequenciadas, os transcritomas não representam a totalidade dos genes presentes em cada espécie. Em acordo com esse resultado, o enriquecimento das vias de metabolismo, degradação e biossíntese de ácidos graxos e metabolismo de glicolipídeo nos transcritos anotados de *J. curcas* e *R. communis* via Kobas é mais um indício deste viés em função do tecido sequenciado (BROWN et al., 2012; HAO et al., 2011). Dussert *et al.* (2013) descreve que há uma diferença de transcritos e vias metabólicas de ácidos graxos ativos para *E. guineensis* dependendo do tecido analisado (mesocarpo, embrião ou endosperma). Brown *et al.* (2012) também comprovaram a presença de genes tecido-específico no transcritoma de cinco tecidos diferentes de *R. communis* para a biossíntese de lipídeos.

É importante ressaltar que a análise de vias metabólicas enriquecidas foi realizada após a anotação dos transcritos utilizando *A. thaliana* como referência. Filogeneticamente

J. curcas e *R. communis* são mais próximas de *A. thaliana*, o que também pode explicar esse resultado.

Analisando especificamente os transcritos associados às vias metabólicas de ácidos graxos, 107 transcritos foram comuns às três espécies, o que evidência que as vias de ácidos graxos são conservadas entre as espécies (BURGAL et al., 2008). Cinco transcritos estão ausentes (FabZ, citocromo bifuncional P450/NADPH--P450 redutase, lisofosfolipideo aciltransferase, prostaglandina-E sintase e acilglicerol lipase) de *E. guineensis* em todos os dados analisados: transcritoma, genoma e KEGG. A ausência pode ser justificada pela distância filogenética entre esta espécie e *J. curcas* e *R. communis*, já que *E. guineensis* se diverge.

É importante comentar o alto número de produtos encontrados apenas no KEGG (29, 30 e 24 proteínas para *J. curcas*, *R. communis* e *E. guineensis*, respectivamente). Isto pode ser explicado pelo fato de que o KEGG anota as sequências automaticamente usando como referência a base de dados RefSeq e GenBank do NCBI. O RefSeq, em teoria, deveria ser um banco de dados de referência com curadoria. No entanto, o grande volume de dados e o baixo número de funcionários ou colaboradores não permite a curadoria manual de todas as entradas (KANEHISA et al., 2016). Além de usar os bancos de dados do NCBI, o KEGG reanota computacionalmente os produtos gênicos em grupos ortólogos baseados em proximidade filogenética (GREEN; KARP, 2005).

As vias metabólicas de ácidos graxos utilizadas neste trabalho foram reconstruídas com base no modelo disponível na página do KEGG e anotadas com as informações disponíveis no KEGG, no genoma e no transcritoma feito, com base na metodologia desenvolvida. Um ponto a ser discutido é o fato de que o genoma de *J. curcas* foi o que apresentou o maior número de ausências de proteínas, chegando a ter mais que o dobro em comparação com as outras espécies do estudo conforme pode ser observado no Quadro 1. Apesar deste genoma ter apresentado a maior completude dentre as três espécies (96,2%), este resultado mostra que

há um problema na montagem deste genoma, uma vez que, com apenas duas exceções, todos os produtos ausentes no genoma estão presentes no transcritoma, podendo assim melhorar a anotação do genoma, trazendo novas informações para está espécie, que não estavam presentes anteriormente. Apesar de apresentar uma cobertura de 189X, o genoma de *J. curcas* foi feito *de novo* com tecnologia Illumina, sem outra tecnologia que produz sequências com tamanhos maiores (superiores a 3kb) para auxiliar na montagem (ALKAN; SAJJADIAN; EICHLER, 2011).

As vias do ciclo básico de biossíntese, elongação, degradação de ácidos graxos e metabolismo de glicerolipídeo, que totalizam 28 enzimas, são importantes para a produção de biocombustíveis por atuarem na regeneração de cofatores necessários para a elongação de cadeias carbônicas que são utilizadas para produção (LU et al., 2011). A metodologia desenvolvida neste trabalho foi capaz de identificar, dentre os transcritos, 26, 27 e 24 em *J. curcas*, *R. communis* e *E. guineensis* respectivamente das 28 enzimas esperadas, enquanto até então a informação existente, através das anotações dos genomas, contabilizava 10, 21 e 22 enzimas para em *J. curcas*, *R. communis* e *E. guineensis* respectivamente. A combinação de informações ômicas e técnicas de edição genômica podem ser utilizadas para o desenho sintético de vias metabólicas em plantas oleaginosas a partir de novas informações, como as encontradas neste trabalho (HASLAM et al., 2016).

7 Conclusões

As principais conclusões deste trabalho foram:

- A utilização de seis montadores com protocolos *de novo* e guiados pelo genoma, associados com diferentes *k-mers* e finalizados com o Evidential gene gerou transcritomas com completudes de 79,2% para *J. curcas*, 80,2% para *R. communis* e 61,4% para *E. guineensis*, que são melhores que quando comparados com o resultado de cada montador individualmente.
- Foi possível identificar 45 vias metabólicas enriquecidas nas três espécies, as vias de metabolismo de ácidos graxos, metabolismo de glicerolipideo, degradação de ácidos graxos e biossíntese de ácidos graxos foram as vias enriquecidas que são associadas ao metabolismo de ácidos graxos.
- A estratégia de anotação, com a criação de um banco de dados local de proteínas de 12 vias metabólicas associadas à ácidos graxos de 10 espécies de plantas, identificou 111 transcritos únicos de *J. curcas*, *R. communis* e *E. guineensis*, sendo 107 comuns às três espécies.
- Tais informações são importantes para aumentar a compreensão das vias metabólicas de ácidos graxos de *J. curcas*, *R. communis* e *E. guineensis* permitindo novos progressos nos programas de melhoramento genético e metabólico, com foco na produção de biocombustíveis, químicos, fármacos e cosméticos de interesse industrial.
- A estratégia possibilitou a identificação de cinco possíveis transcritos não anotados anteriormente nos genomas de *J. curcas*, *R. communis* e *E. guineensis*, estes transcritos se confirmados podem ser utilizados como alvos de melhoramento genético, desta forma possibilitando novas utilizações destas

plantas na escala produtiva e das informações obtidas no melhoramento de plantas já utilizadas.

8 Perspectivas

Como perspectivas, vamos analisar dados de ortologia gênica a partir dos dados gerados pelo programa Orthofinder, o programa que construiu a árvore filogenética. Os dados podem trazer novas informações quanto à especificidade de ortogrupos com presença de proteínas que participam das vias metabólicas de ácidos graxos.

Pretendemos, também, gerar um protocolo automatizado através da plataforma BPipe que é uma ferramenta para execução e gerenciamento de fluxogramas de analises em bioinformática (SADEDIN; POPE; OSHLACK, 2012).

Um boletim de pesquisa e desenvolvimento será preparado para publicação ainda este ano.

9 Referências

ACHTEN, W. M. J. et al. Jatropha biodiesel fueling sustainability? **Biofuels, Bioprod. Bioref.**, v. 1, n. 4, p. 283–291, dez. 2007.

ACHTEN, W. M. J. et al. Jatropha bio-diesel production and use. **Biomass and Bioenergy**, v. 32, n. 12, p. 1063–1084, dez. 2008.

AGARWALA, R. et al. Database resources of the National Center for Biotechnology Information. **Nucleic Acids Research**, v. 46, n. D1, p. D8–D13, 2018.

AHSAN, H.; AHAD, A.; SIDDIQUI, W. A. A review of characterization of tocotrienols from plant oils and foods. **J Chem Biol**, v. 8, n. 2, p. 45–59, abr. 2015.

AI, C.; KONG, L. CGPS: A machine learning-based approach integrating multiple gene set analysis tools for better prioritization of biologically relevant pathways. **Journal of Genetics and Genomics**, v. 45, n. 9, p. 489–504, 2018.

ALKAN, C.; SAJJADIAN, S.; EICHLER, E. E. Limitations of next-generation genome sequence assembly. **Nature Methods**, v. 8, n. 1, p. 61–65, 2011.

ALVIM, A. M. Biocombustíveis: uma análise da evolução do biodiesel no Brasil. **Economia & Tecnologia**, v. 25, 2011.

ANDREWS, S. FastQC A Quality Control tool for High Throughput Sequence Data. **http://www.bioinformatics.babraham.ac.uk/projects/fastqc/**, 2014.

ANNEKEN, D. J. et al. Fatty Acids. In: **Ullmann's Encyclopedia of Industrial Chemistry**. [s.l.] American Cancer Society, 2006.

ARENT, S. et al. The Multifunctional Protein in Peroxisomal ␣ -Oxidation. v. 285, n. 31, p. 24066–24077, 2010.

ARONESTY, E. **ea-utils: Command-line tools for processing biological sequencing data**Durham, NC, , 2011.

BACENETTI, J. et al. Biodiesel production from unconventional oilseed crops (Linum usitatissimum L. and Camelina sativa L.) in Mediterranean conditions: Environmental sustainability assessment. **Renew Energy**, v. 112, p. 444–456, nov. 2017.

BADOUIN, H. et al. The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. **Nature**, v. 546, p. 148, 22 maio 2017.

BARAJAS FORERO, C. L. Biodiesel from Castor Oil: A Promising Fuel for Cold Weather. **Renewable Energy and Power Quality Journal**, v. 1, n. 03, p. 59–62, 2005.

BARLIND, J. G. et al. Identification and design of a novel series of MGAT2 inhibitors. **Bioorganic & Medicinal Chemistry Letters**, v. 23, n. 9, p. 2721–2726, 2013.

BENNER, S. A.; SISMOUR, A. M. Synthetic biology. **Nature Reviews Genetics**, v. 6, n. 7, p. 533–543, 2005.

BILAL, M. et al. Metabolic engineering and enzyme-mediated processing: A biotechnological venture towards biofuel production – A review. **Renewable and Sustainable Energy Reviews**, v. 82, p. 436–447, fev. 2018.

BRITTAINE, R.; LUTALADIO, N. Jatropha: a smallholder bioenergy crop: the potential for pro-poor development. **Integrated Crop Management**, v. 8, p. xv + 96 pp., 2010.

BROWN, A. P. et al. Tissue-specific whole transcriptome sequencing in castor, directed at understanding triacylglycerol lipid biosynthetic pathways. **PLoS ONE**, v. 7, n. 2, 2012.

BURGAL, J. et al. Metabolic engineering of hydroxy fatty acid production in plants: RcDGAT2 drives dramatic increases in ricinoleate levels in seed oil. **Plant Biotechnology Journal**, v. 6, n. 8, p. 819–831, 2008.

CANVIN, D. T. THE EFFECT OF TEMPERATURE ON THE OIL CONTENT AND FATTY ACID COMPOSITION OF THE OILS FROM SEVERAL OIL SEED CROPS. **Canadian Journal of Botany**, v. 43, n. 1, p. 63–69, jan. 1965.

CASTRO GONZÁLES, N. F. International experiences with the cultivation of Jatropha curcas for biodiesel production. **Energy**, v. 112, n. 2016, p. 1245–1258, 2016.

CHEN, S. et al. Optimizing Transcriptome Assemblies for Leaf and Seedling by Combining Multiple Assemblies from Three De Novo Assemblers. **Plant Genome**, v. 8, n. 1, p. 0, 2015.

CHOJNACKI, S. et al. Programmatic access to bioinformatics tools from EMBL-EBI update : 2017. **Nucleic Acids Research**, v. 45, n. April, p. 550–553, 2017.

COMPEAU, P. E. C.; PEVZNER, P. A.; TESLER, G. How to apply de Bruijn graphs to genome assembly. **Nat Biotechnol**, v. 29, n. 11, p. 987–991, nov. 2011.

COSTA, G. G. L. et al. Transcriptome analysis of the oil-rich seed of the bioenergy

crop Jatropha curcas L. 2010.

COVINGTON, M. F. et al. Global transcriptome analysis reveals circadian regulation of key pathways in plant growth and development. **Genome Biology**, v. 9, n. 8, 2008.

DEMIRBAS, M. F.; BALAT, M. Recent advances on the production and utilization trends of bio-fuels: A global perspective. **Energy Conversion and Management**, v. 47, n. 15–16, p. 2371–2381, set. 2006.

DENG, X.; CAI, J.; FEI, X. Involvement of phosphatidate phosphatase in the biosynthesis of triacylglycerols in Chlamydomonas reinhardtii. v. 14, n. 12, p. 1121–1131, 2013.

DERISI, J. L.; IYER, V. R.; BROWN, P. O. Exploring the metabolic and genetic control of gene expression on a genomic scale. **Science**, v. 278, n. 5338, p. 680–686, out. 1997.

DEVAPPA, R. K.; MAKKAR, H. P. S.; BECKER, K. Jatropha toxicity-A review. **Journal of Toxicology and Environmental Health - Part B: Critical Reviews**, v. 13, n. 6, p. 476–507, 2010.

DURAND-GASSELIN, T. et al. Breeding for sustainable palm oil. **International Seminar on Breeding for Sustainability in Oil Palm. 18 November 2011. Kuala Lumpur, Malaysia**, n. November, p. 178–193, 2011.

EDEM, D. O. Palm oil: biochemical, physiological, nutritional, hematological, and toxicological aspects: a review. **Plant Foods Hum Nutr**, v. 57, n. 3–4, p. 319–341, 2002.

EDRISI, S. A. et al. Jatropha curcas L.: A crucified plant waiting for resurgence. **Renewable and Sustainable Energy Reviews**, v. 41, p. 855–862, jan. 2015.

EMMS, D. M.; KELLY, S. {OrthoFinder}: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. **Genome Biol**, v. 16, p. 157, ago. 2015.

ERB, T. J.; JONES, P. R.; BAR-EVEN, A. Synthetic metabolism: metabolic engineering meets enzyme design. **Curr Opin Chem Biol**, v. 37, p. 56–62, abr. 2017.

GAO, W. et al. Acyl-CoA-binding protein 2 binds lysophospholipase 2 and lysoPC to promote tolerance to cadmium-induced oxidative stress in transgenic Arabidopsis. p. 989–1003, 2010.

GENIZA, M.; JAISWAL, P. Tools for building de novo transcriptome assembly.

Current Plant Biology, v. 11–12, n. December, p. 41–45, 2017.

GHOSH, S.; CHAN, C. K. Analysis of RNA-Seq Data Using TopHat and Cuffl inks. **Methods in molecular biology**, v. 1374, p. 339–361, 2016.

GILBERT, D. G. Genes of the pig, Sus scrofa , reconstructed with EvidentialGene. **PeerJ**, v. 7, p. e6374, 1 fev. 2019.

GOODWIN, S.; MCPHERSON, J. D.; MCCOMBIE, W. R. Coming of age: ten years of next-generation sequencing technologies. **Nat Rev Genet**, v. 17, n. 6, p. 333–351, maio 2016.

GREEN, M. L.; KARP, P. D. Genome annotation errors in pathway databases due to semantic ambiguity in partial EC numbers. **Nucleic Acids Research**, v. 33, n. 13, p. 4035–4039, 2005.

HAAK, M. et al. High Quality de Novo Transcriptome Assembly of Croton tiglium. **Frontiers in Molecular Biosciences**, v. 5, n. July, p. 1–5, 2018.

HANSEN, A. S. L. et al. Systems biology solutions for biochemical production challenges. **Curr Opin Biotechnol**, v. 45, p. 85–91, mar. 2017.

HAO, D. C. et al. The first insight into the tissue specific taxus transcriptome via illumina second generation sequencing. **PLoS ONE**, v. 6, n. 6, 2011.

HARRIS, T. E. et al. Insulin Controls Subcellular Localization and Multisite Phosphorylation of the Phosphatidic Acid Phosphatase , Lipin 1 *. v. 282, n. 1, p. 277–286, 2007.

HARVEY, A. L.; EDRADA-EBEL, R.; QUINN, R. J. The re-emergence of natural products for drug discovery in the genomics era. **Nat Rev Drug Discov**, v. 14, n. 2, p. 111–129, fev. 2015.

HASLAM, R. P. et al. Synthetic redesign of plant lipid metabolism. **The Plant journal : for cell and molecular biology**, v. 87, n. 1, p. 76–86, 2016.

HILDEBRAND, D. Lipid Biosynthesis. **Plant Metabolism and Biotechnology**, v. 7, n. July, p. 27–65, 2011.

HONG, D. Y.; BLACKMORE, S. **Plants of China: A Companion to the Flora of China**. [s.l.] Cambridge University Press, 2015.

HORNER, D. S. et al. Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. **Brief Bioinformatics**, v. 11, n. 2, p. 181–197,

mar. 2010.

IDE, T. et al. Physiological effects of γ-linolenic acid and sesamin on hepatic fatty acid synthesis and oxidation. **The Journal of Nutritional Biochemistry**, v. 41, p. 42–55, 2017.

IQBAL, Z. et al. De novo assembly and genotyping of variants using colored de Bruijn graphs. v. 44, n. 2, p. 226–232, 2012.

JOHNSON, D. K. et al. Biomass Recalcitrance: Engineering Plants and Enzymes for Biofuels Production. **Science**, v. 315, n. 5813, p. 804–807, 2007.

KANEHISA, M. et al. {KEGG} for linking genomes to life and the environment. **Nucleic Acids Res**, v. 36, n. Database issue, p. D480-4, jan. 2008.

KANEHISA, M. et al. {KEGG} as a reference resource for gene and protein annotation. **Nucleic Acids Res**, v. 44, n. D1, p. D457-62, jan. 2016.

KANTAR, M. B. et al. Perennial grain and oilseed crops. **Annu Rev Plant Biol**, v. 67, p. 703–729, abr. 2016.

KHALIL, A. S.; COLLINS, J. J. Synthetic biology: applications come of age. **Nature Reviews Genetics**, v. 11, p. 367, 1 maio 2010.

KUMAR, A.; SHARMA, S. An evaluation of multipurpose oil seed crop for industrial uses (Jatropha curcas L.): A review. **Industrial Crops and Products**, v. 28, n. 1, p. 1–10, 2008.

LEE, S. K. et al. Metabolic engineering of microorganisms for biofuels production: from bugs to synthetic biology to fuels. **Current Opinion in Biotechnology**, v. 19, n. 6, p. 556–563, 2008.

LU, C. et al. New frontiers in oilseed biotechnology: meeting the global demand for vegetable oils for food, feed, biofuel, and industrial applications. **Curr Opin Biotechnol**, v. 22, n. 2, p. 252–259, abr. 2011.

LUO, R. et al. SOAPdenovo2 : an empirically improved memory-efficient short-read de novo assembler. p. 1–6, 2012.

MAMANOVA, L. et al. Target-enrichment strategies for next-generation sequencing. **Nat Methods**, v. 7, n. 2, p. 111–118, fev. 2010.

MANZONI, C. et al. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. **Brief Bioinformatics**, v. 19, n. 2, p. 286–302, mar.

2018.

MARGULIES, M. et al. Genome Sequencing in Open Microfabricated High Density Picoliter Reactors Marcel. v. 437, n. 7057, p. 376–380, 2006.

MARTIN, J. A.; WANG, Z. Next-generation transcriptome assembly. **Nature Reviews Genetics**, v. 12, n. 10, p. 671–682, 2011.

MARTIN, M. Cutadapt removes adapter sequences from high-throughput sequencing reads kenkyuhi hojokin gan rinsho kenkyu jigyo. **EMBnet.journal**, v. 17, n. 1, p. 10–12, 2013.

MAY, A.; SPINKA, M.; KÖCK, M. Arabidopsis thaliana PECP1 — Enzymatic characterization and structural organization of the fi rst plant phosphoethanolamine / phosphocholine phosphatase. **BBA - Proteins and Proteomics**, v. 1824, n. 2, p. 319–325, 2012.

MBA, O. I.; DUMONT, M.-J.; NGADI, M. Palm oil: Processing, characterization and utilization in the food industry – A review. **Food Biosci**, v. 10, p. 26–41, jun. 2015.

METZKER, M. L. Sequencing technologies - the next generation. **Nat Rev Genet**, v. 11, n. 1, p. 31–46, jan. 2010.

MIAZEK, K. et al. Sphingolipids: Promising lipid-class molecules with potential applications for industry. A review | Sphingolipides: Des molécules lipidiques à haut potentiel de valorisation présentant de nombreuses applications industrielles (synthèse bibliographique). **Biotechnology, Agronomy and Society and Environment**, v. 20, n. S1, p. 321–336, 2016.

MOGHE, G. D. et al. Multi-omic analysis of a hyper-diverse plant metabolic pathway reveals evolutionary routes to biological innovation Impact statement. 2017.

MONTES, J. M.; MELCHINGER, A. E. Domestication and Breeding of Jatropha curcas L. **Trends Plant Sci**, v. 21, n. 12, p. 1045–1057, set. 2016.

MORAIS, S. et al. Synthesis and Stabilization of Gold Nanoparticles in Castor Oil. **Revista Virtual de Química**, v. 5, 1 jan. 2013.

MUTZ, K.-O. et al. Transcriptome analysis using next-generation sequencing. **Curr Opin Biotechnol**, v. 24, n. 1, p. 22–30, fev. 2013.

NAPIER, J. A.; GRAHAM, I. A. Tailoring plant lipid composition: designer oilseeds come of age. **Curr Opin Plant Biol**, v. 13, n. 3, p. 330–337, jun. 2010.

NATARAJAN, P.; PARANI, M. De novo assembly and transcriptome analysis of five major tissues of Jatropha curcas L. using GS FLX titanium platform of 454 pyrosequencing. **BMC Genomics**, v. 12, 2011.

NAVARRO-PINEDA, F. S. et al. Advances on the processing of Jatropha curcas towards a whole-crop biorefinery. **Renewable and Sustainable Energy Reviews**, v. 54, p. 247–269, 2016.

NIELSEN, J.; KEASLING, J. D. Engineering Cellular Metabolism. **Cell**, v. 164, n. 6, p. 1185–1197, mar. 2016.

NÜTZMANN, H. W.; HUANG, A.; OSBOURN, A. Plant metabolic clusters – from genetics to genomics. **New Phytologist**, v. 211, n. 3, p. 771–789, 2016.

OBAHIAGBON, F. I. A Review: Aspects of the African Oil Palm (Elaeis guineesis jacq.) and the Implications of its Bioactives in Human Health. **American Journal of Biochemistry and Molecular Biology**, v. 2, n. 3, p. 106–119, 1 mar. 2012.

OCKENDON, N. F. et al. Optimization of next-generation sequencing transcriptome annotation for species lacking sequenced genomes. **Molecular Ecology Resources**, v. 16, n. 2, p. 446–458, 2016.

OGATA, H. et al. KEGG: kyoto encyclopedia of genes and genomes. **Nucleic Acids Res**, v. 27, n. 1, p. 29–34, jan. 1999.

OGUNNIYI, D. S. Castor oil: a vital industrial raw material. **Bioresour Technol**, v. 97, n. 9, p. 1086–1091, jun. 2006.

OLIVARES-CARRILLO, P. et al. High-yield production of biodiesel by non-catalytic supercritical methanol transesteri fi cation of crude castor oil (Ricinus communis) n Rom a Celi a. v. 107, p. 165–171, 2016.

OSAKI, M. & BATALHA, M. O. Produção de biodiesel e óleo vegetal no Brasil: Realidade e desafio. **Organizações Rurais & Agroindustriais**, v. 12, n. 2, p. 227–242, 2011.

PATEL, V. R. et al. Castor Oil : Properties , Uses , and Optimization of Processing Parameters in Commercial Production. p. 1–12, 2016.

PRAMANIK. Properties and use of jatropha curcas oil and diesel fuel blends in compression ignition engine. **Renewable Energy and Power Quality Journal**, v. 28, p. 239–248, 2003.

RANA, S. B. et al. Comparison of de Novo transcriptome assemblers and k-mer

strategies using the killifish, fundulus heteroclitus. **PLoS ONE**, v. 11, n. 4, p. 1–16, 2016.

ROBERTS, S. J. et al. Human PHOSPHO1 exhibits high specific phosphoethanolamine and phosphocholine phosphatase activities. v. 65, p. 59–65, 2004.

SABLOK, G. et al. Fuelling genetic and metabolic exploration of C 3 bioenergy crops through the first reference transcriptome of Arundo donax L . p. 554–567, 2014.

SADEDIN, S. P.; POPE, B.; OSHLACK, A. Bpipe: A tool for running and managing bioinformatics pipelines. **Bioinformatics**, v. 28, n. 11, p. 1525–1526, 2012.

SAMBANTHAMURTHI, R. et al. Opportunities for the oil palm via breeding and biotechnology. In: JAIN, S. M.; PRIYADARSHAN, P. M. (Eds.). . **Breeding plantation tree crops: tropical species**. New York, {NY}: Springer New York, 2009. p. 377–421.

SAVADI, S. et al. Genetic engineering approaches to enhance oil content in oilseed crops. **Plant Growth Regul**, p. 1–16, nov. 2016.

SAVADI, S. et al. Genetic engineering approaches to enhance oil content in oilseed crops. **Plant Growth Regulation**, v. 83, n. 2, p. 207–222, 2017.

SCHIRMER, M. et al. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. **Nucleic Acids Research**, v. 43, n. 6, 2015.

SCHULZ, M. H. et al. Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. **Bioinformatics**, v. 28, n. 8, p. 1086–1092, 2012.

SEVERINO, L. S. et al. A review on the challenges for increased production of castor. **Agron J**, v. 104, n. 4, p. 853, 2012.

SHIN, S.; PARK, J. Characterization of sequence-specific errors in various next-generation sequencing systems. **Molecular BioSystems**, v. 12, n. 3, p. 914–922, 2016.

SHU, Y. Z. Recent natural products based drug development: a pharmaceutical industry perspective. **J Nat Prod**, v. 61, n. 8, p. 1053–1071, ago. 1998.

SIMÃO, F. A. et al. {BUSCO}: assessing genome assembly and annotation completeness with single-copy orthologs. **Bioinformatics**, v. 31, n. 19, p. 3210–3212, out. 2015.

SINGH, R. et al. Oil palm genome sequence reveals divergence of interfertile species in Old and New worlds. **Nature**, v. 500, n. 7462, p. 335–339, ago. 2013.

STEPHANOPOULOS, G. Metabolic Fluxes and Metabolic Engineering. **Metabolic Engineering**, v. 11, p. Pages 1-11, 1999.

STICKLEN, M. B. Plant genetic engineering for biofuel production: Towards affordable cellulosic ethanol. **Nature Reviews Genetics**, v. 9, n. 6, p. 433–443, 2008.

TAKATSUTO, S. et al. Erect leaves caused by brassinosteroid deficiency increase biomass production and grain yield in rice. **Nature Biotechnology**, v. 24, n. 1, p. 105–109, 2005.

TAN, Q. G. et al. Three terpenoids and a tocopherol-related compound from Ricinus communis. **Helvetica Chimica Acta**, v. 92, n. 12, p. 2762–2768, 2009.

TEH, H. F. et al. Review: Omics and Strategic Yield Improvement in Oil Crops. **JAOCS, Journal of the American Oil Chemists' Society**, v. 94, n. 10, p. 1225–1244, 2017.

TUMANEY, A. W.; SHEKAR, S.; RAJASEKHARAN, R. Identification , Purification , and Characterization of Monoacylglycerol Acyltransferase from Developing Peanut Cotyledons *. v. 276, n. 14, p. 10847–10852, 2001.

UTHAIPAISANWONG, P. et al. Characterization of the chloroplast genome sequence of oil palm (Elaeis guineensis Jacq.). **Gene**, v. 500, n. 2, p. 172–180, jun. 2012.

VARSHNEY, R. K. et al. Next-generation sequencing technologies and their implications for crop genetics and breeding. **Trends Biotechnol**, v. 27, n. 9, p. 522–530, set. 2009.

VERHEYE, W. Growth and Production of Oil Palm. **Soils, Plant Growth and Crop Production - Vol. II**, p. 10, 2010.

VIGEOLAS, H. et al. Increasing seed oil content in oil-seed rape (Brassica napus L.) by over-expression of a yeast glycerol-3-phosphate dehydrogenase under the control of a seed-specific promoter. **Plant Biotechnology Journal**, v. 5, n. 3, p. 431–441, 2007.

VISSER, E. A. et al. Combined de novo and genome guided assembly and annotation of the Pinus patula juvenile shoot transcriptome. **{BMC} Genomics**, v. 16, p. 1057, dez. 2015.

WANG, L. et al. Genome sequencing of the high oil crop sesame provides insight into oil biosynthesis. 2014.

WAY, J. C. et al. Integrating biological redesign: where synthetic biology came from and where it needs to go. **Cell**, v. 157, n. 1, p. 151–161, mar. 2014.

WU, J. et al. KOBAS server: A web-based platform for automated annotation and pathway identification. **Nucleic Acids Research**, v. 34, n. WEB. SERV. ISS., p. 720–724,

2006.

XIA, E. H. et al. Transcriptome analysis of the oil-rich tea plant, Camellia oleifera, reveals candidate genes related to lipid metabolism. **PLoS ONE**, v. 9, n. 8, 2014.

XIE, C. et al. KOBAS 2.0: A web server for annotation and identification of enriched pathways and diseases. **Nucleic Acids Research**, v. 39, n. SUPPL. 2, p. 316–322, 2011.

XU, C.; SHANKLIN, J. Triacylglycerol metabolism, function, and accumulation in plant vegetative tissues. **Annu Rev Plant Biol**, v. 67, p. 179–206, abr. 2016.

YANG, C. Y. et al. Review and prospects of Jatropha biodiesel industry in China. **Renewable and Sustainable Energy Reviews**, v. 16, n. 4, p. 2178–2190, 2012.

YAO, D. et al. Transcriptome analysis reveals salt-stress-regulated biological processes and key pathways in roots of cotton (Gossypium hirsutum L.). **Genomics**, v. 98, n. 1, p. 47–55, 2011.

ZERBINO, D. R. Using Velvet de novo assembler for short-reading sequence technologies. **Current Protocols Bioinformatics**, p. 1–13, 2010.

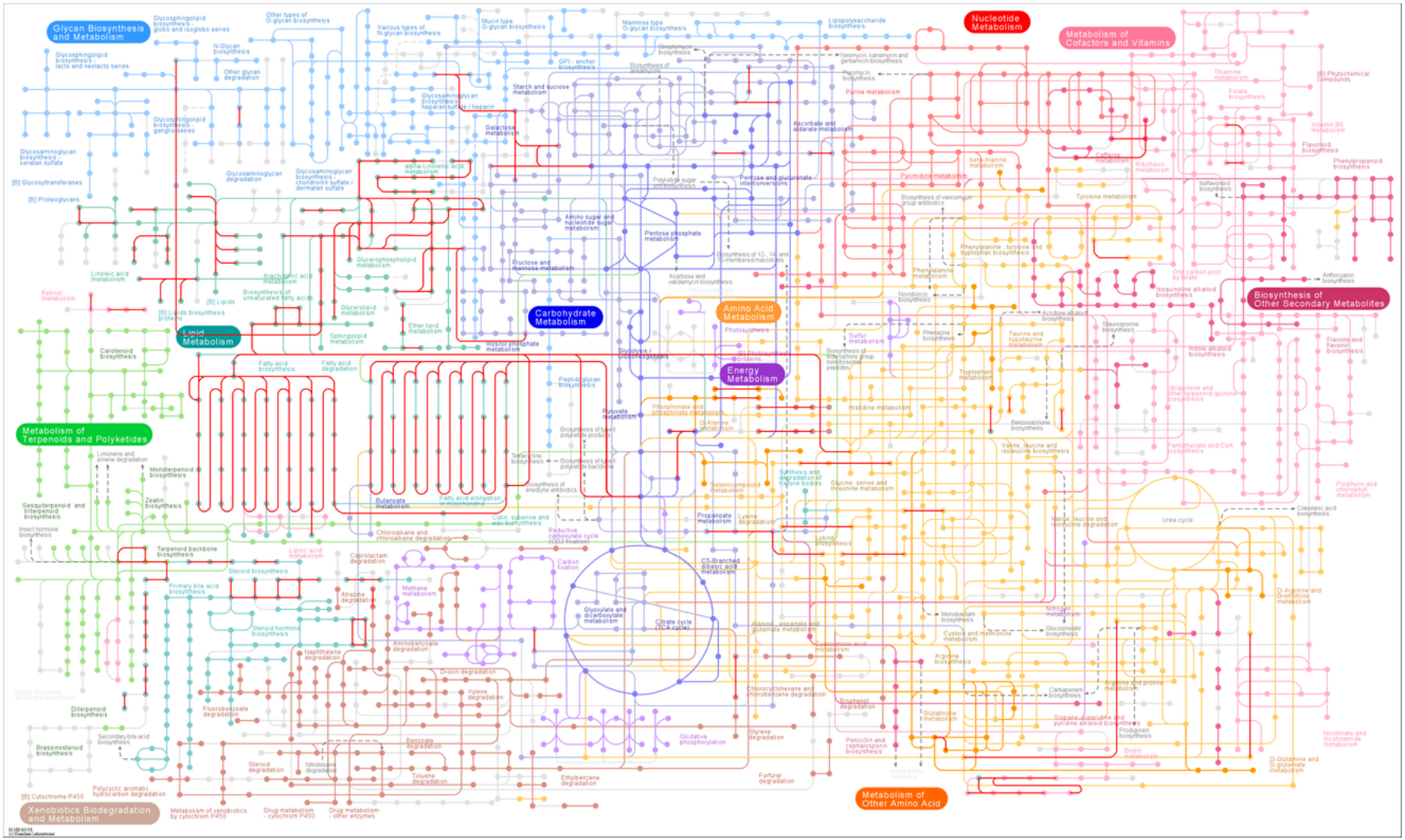
ZHOU, Y. J. et al. Production of fatty acid-derived oleochemicals and biofuels by synthetic yeast cell factories. **Nat Commun**, v. 7, p. 11709, maio 2016.

ZHU, L.-H. et al. Dedicated industrial oilseed crops as metabolic engineering platforms for sustainable industrial feedstock production. **Sci Rep**, v. 6, p. 22181, fev. 2016.

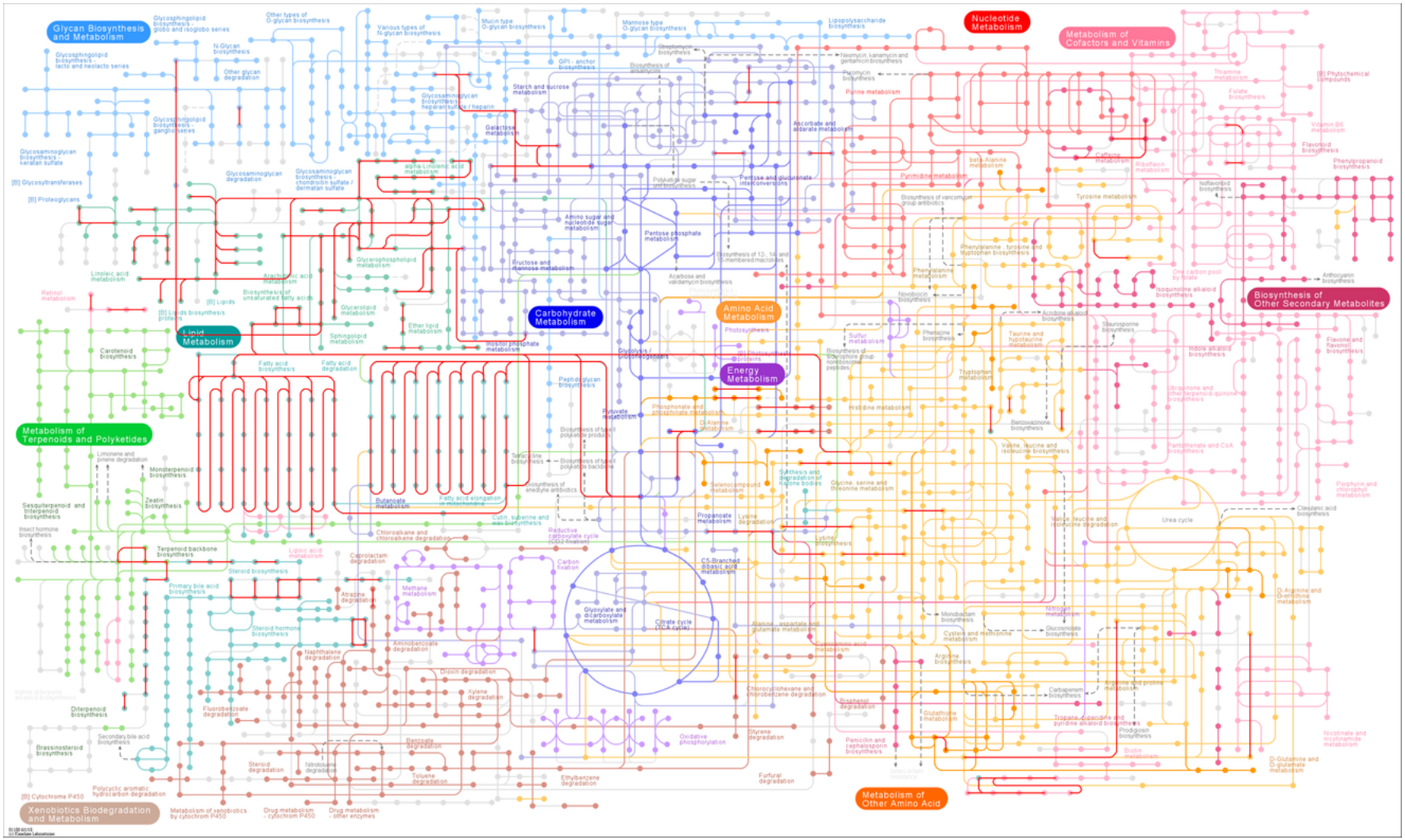
Apêndice

Anotação às vias metabólicas de ácidos graxos

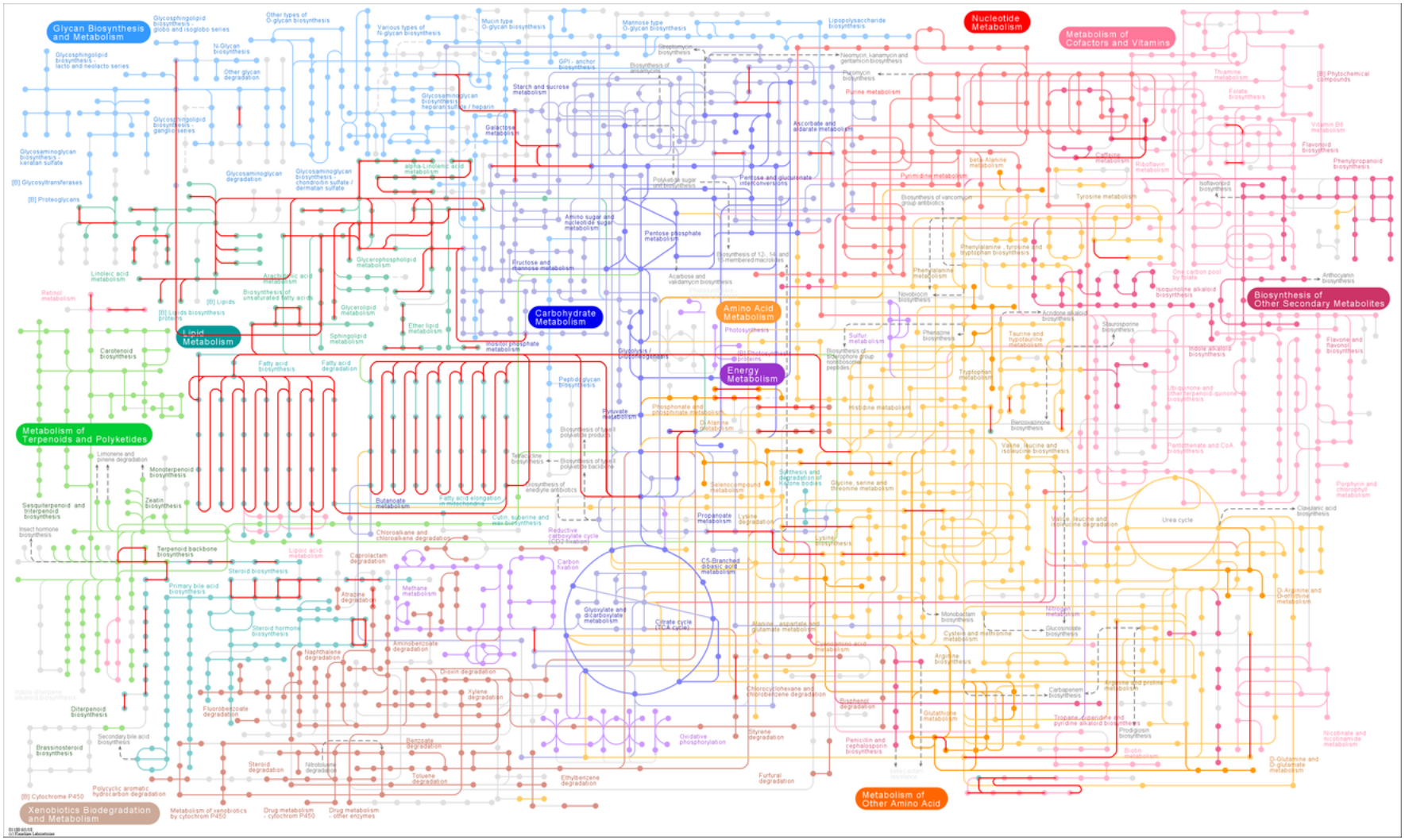
As figuras A1-A3 correspondem a apresentação do mestabolismo completo, com ela é possível observar no contexto geral do metabolismo onde estão localizados os transcritos (em vermelho) encontrados para cada uma das espécies. Figuras referentes às vias metabólicas de ácidos graxos anotadas com os transcritomas estão dispostas abaixo (A4-A39), apresentadas na ordem: *Jatropha curcas, Ricinus communis* e *Elaeis guineensis*.



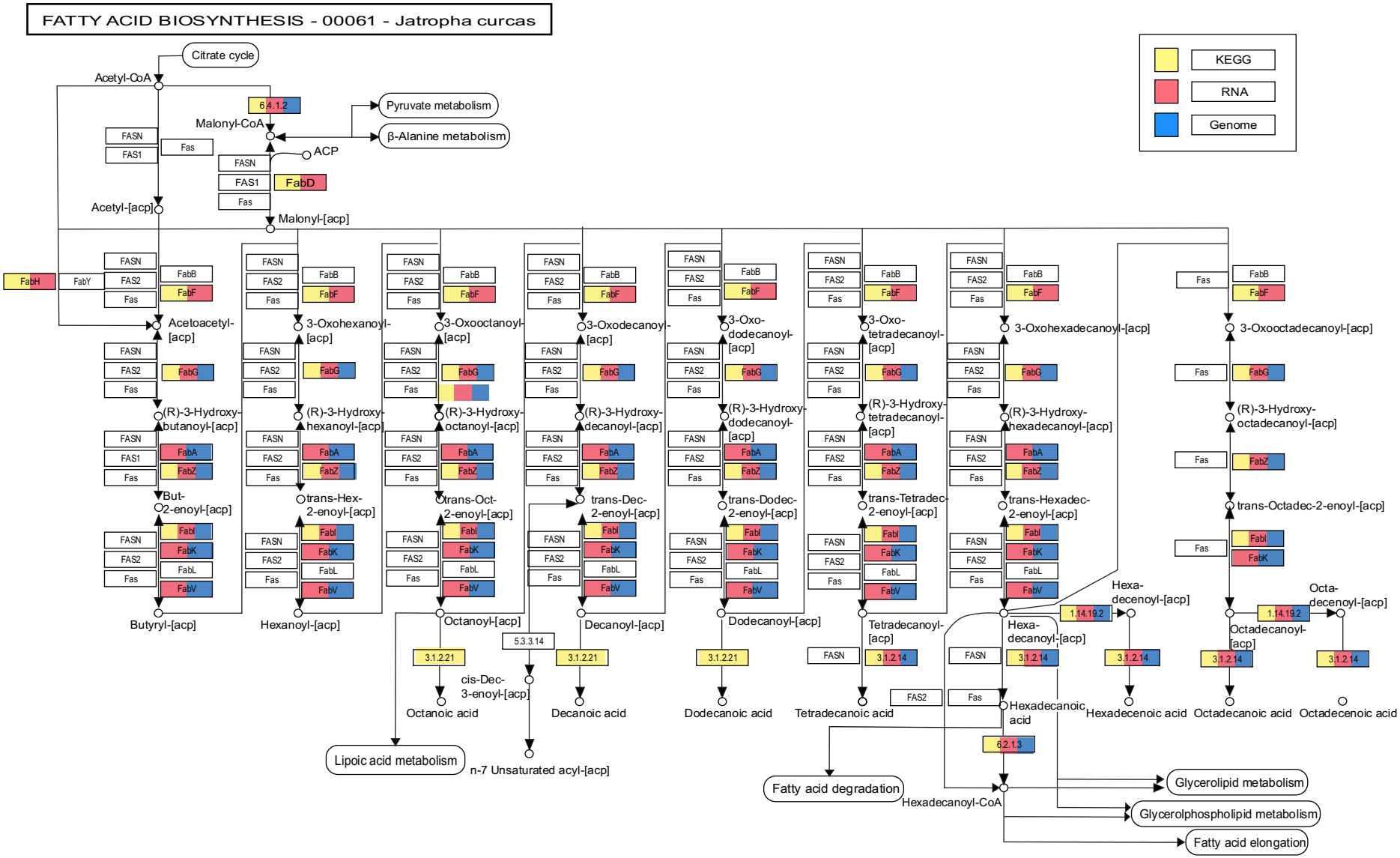
A 1 - Distribuição do transcritoma de *Jatropha curcas* às vias de metabolismo de ácido graxo. Linhas vermelhas são as representam as vias que foram anotadas.



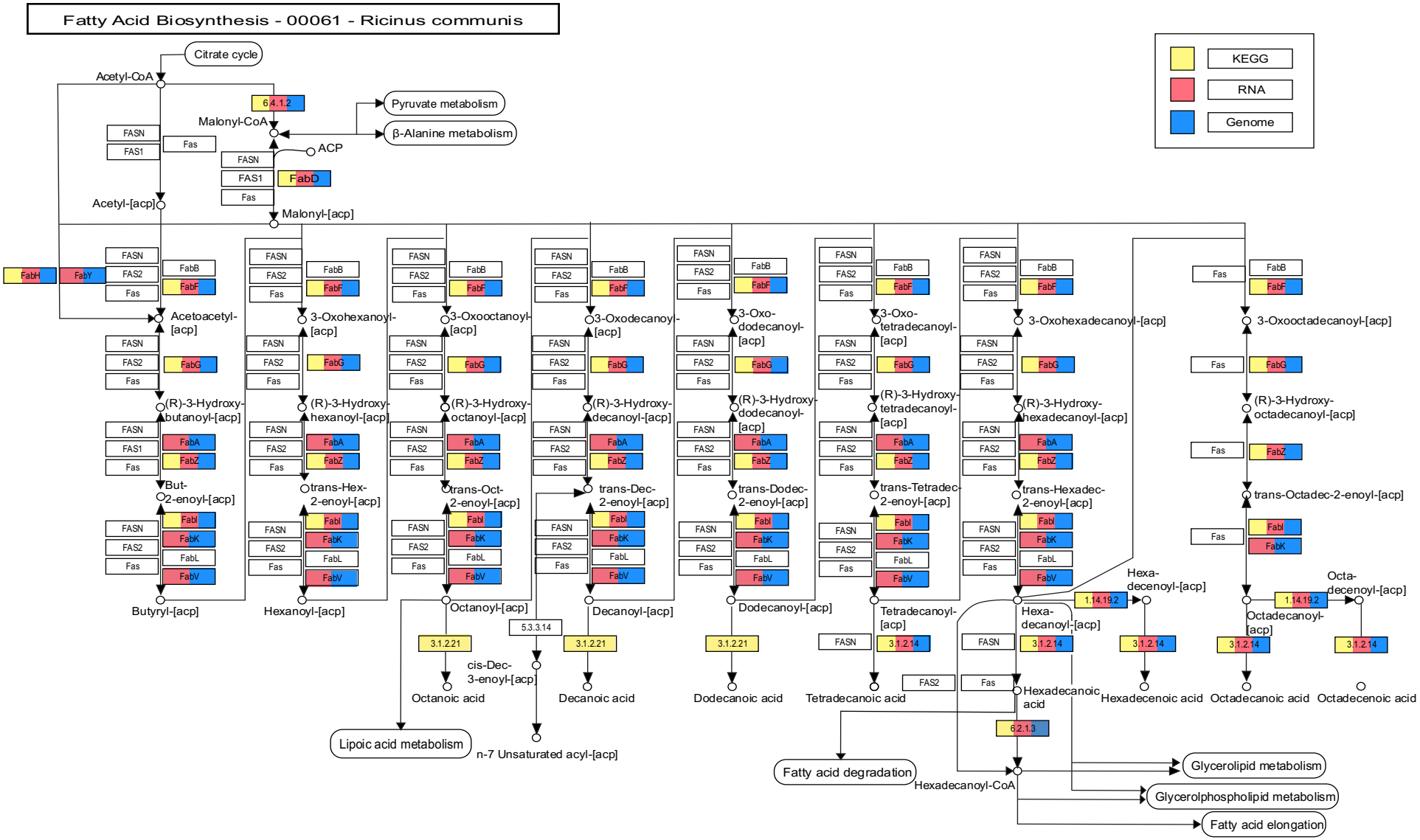
A 2 - Distribuição do transcritoma de *Ricinus communis* às vias de metabolismo de ácido graxo. Linhas vermelhas são as representam as vias que foram anotadas.



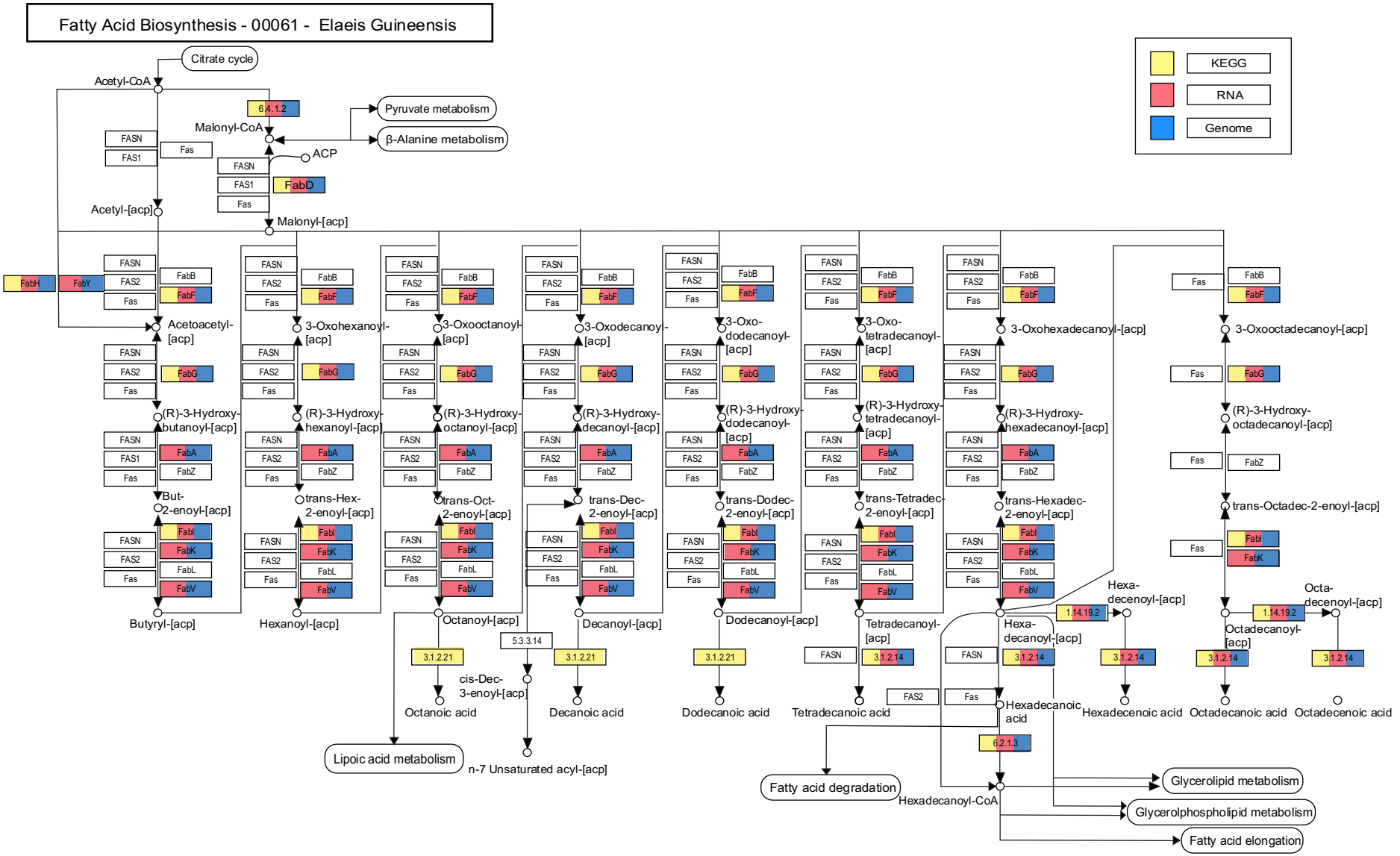
A 3 - Distribuição do transcritoma de *Elaeis guineensis* às vias de metabolismo de ácido graxo. Linhas vermelhas são as representam as vias que foram anotadas.



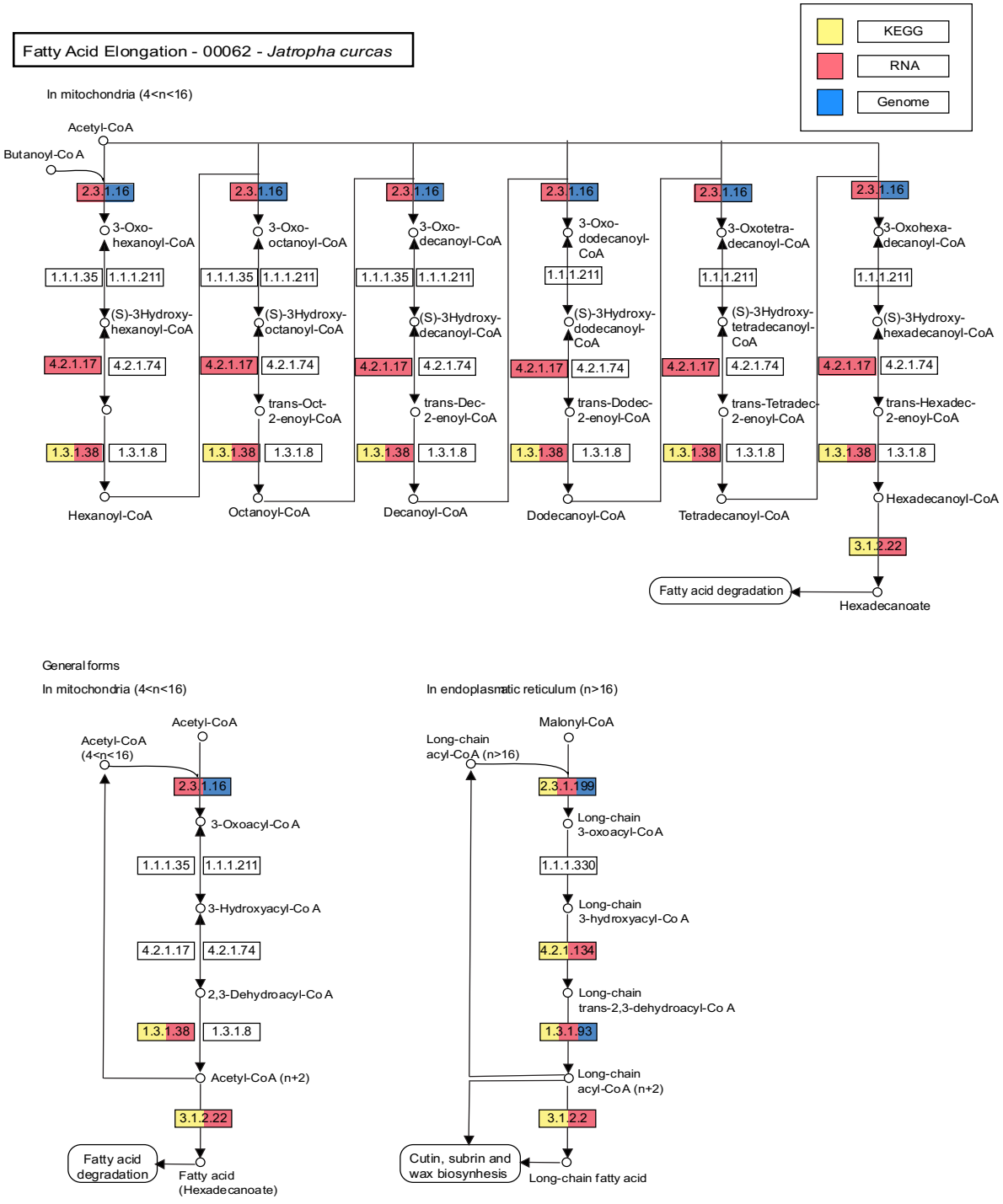
A 4 - Via de Biossíntese de ácidos graxos indicando as enzimas de *J. curcas* presentes nos dados do KEGG, transcritoma e/ou genoma



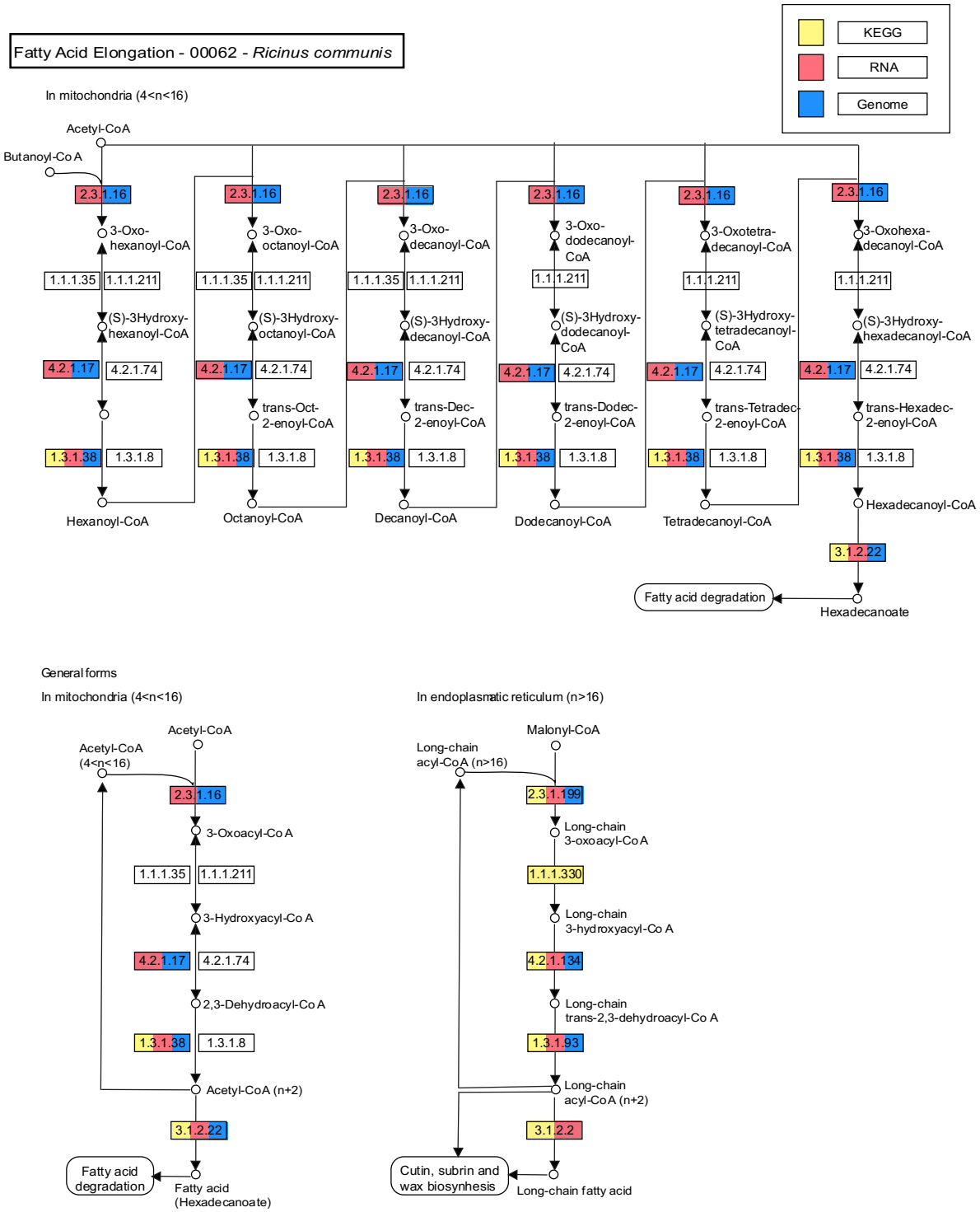
A 5 - Via de Biossíntese de ácidos graxos indicando as enzimas de *R. communis* presentes nos dados do KEGG, transcritoma e/ou genoma



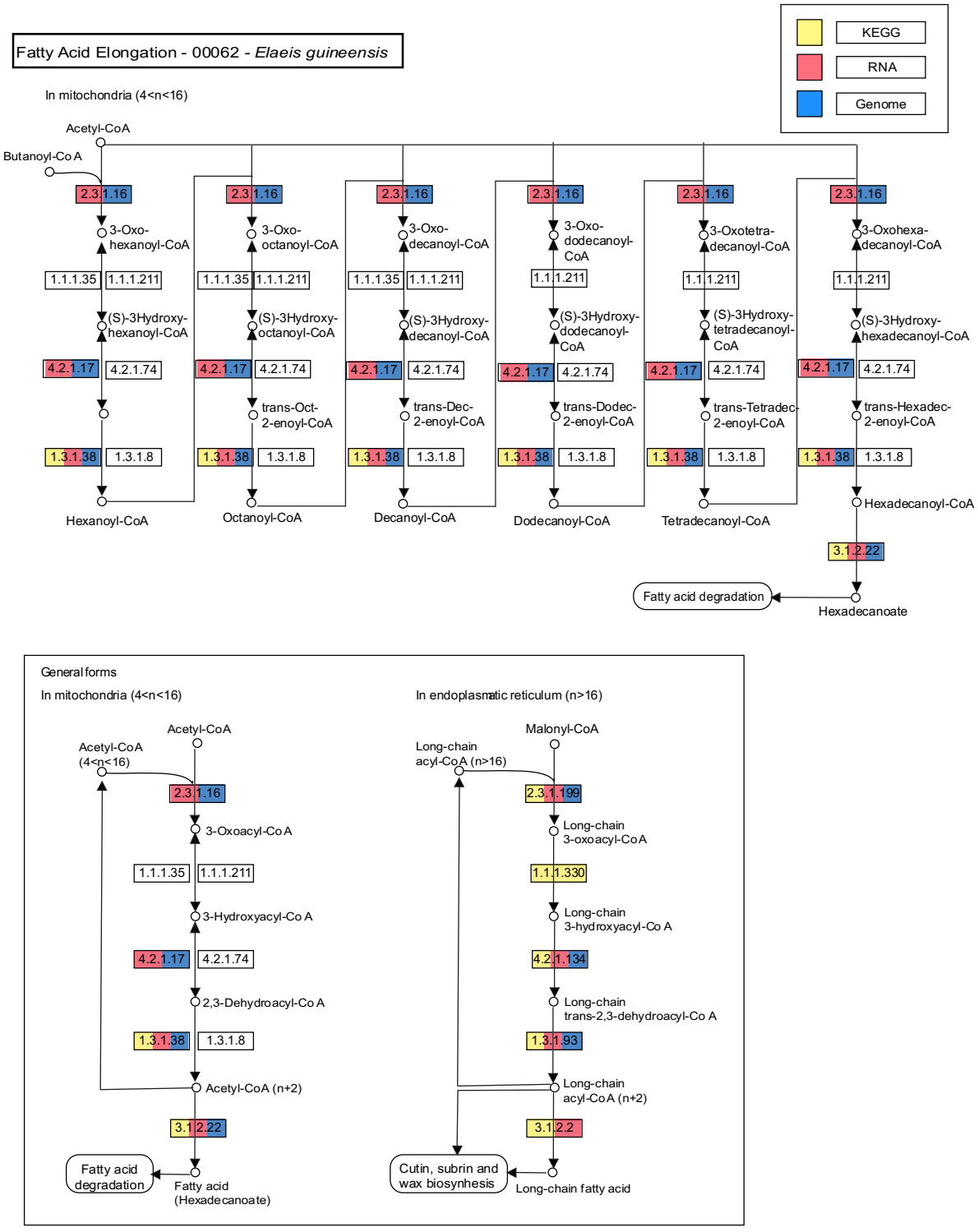
A 6 – Via de Biossíntese de ácidos indicando as enzimas de *E. guineensis* presentes nos dados do KEGG, transcritoma e/ou genoma.



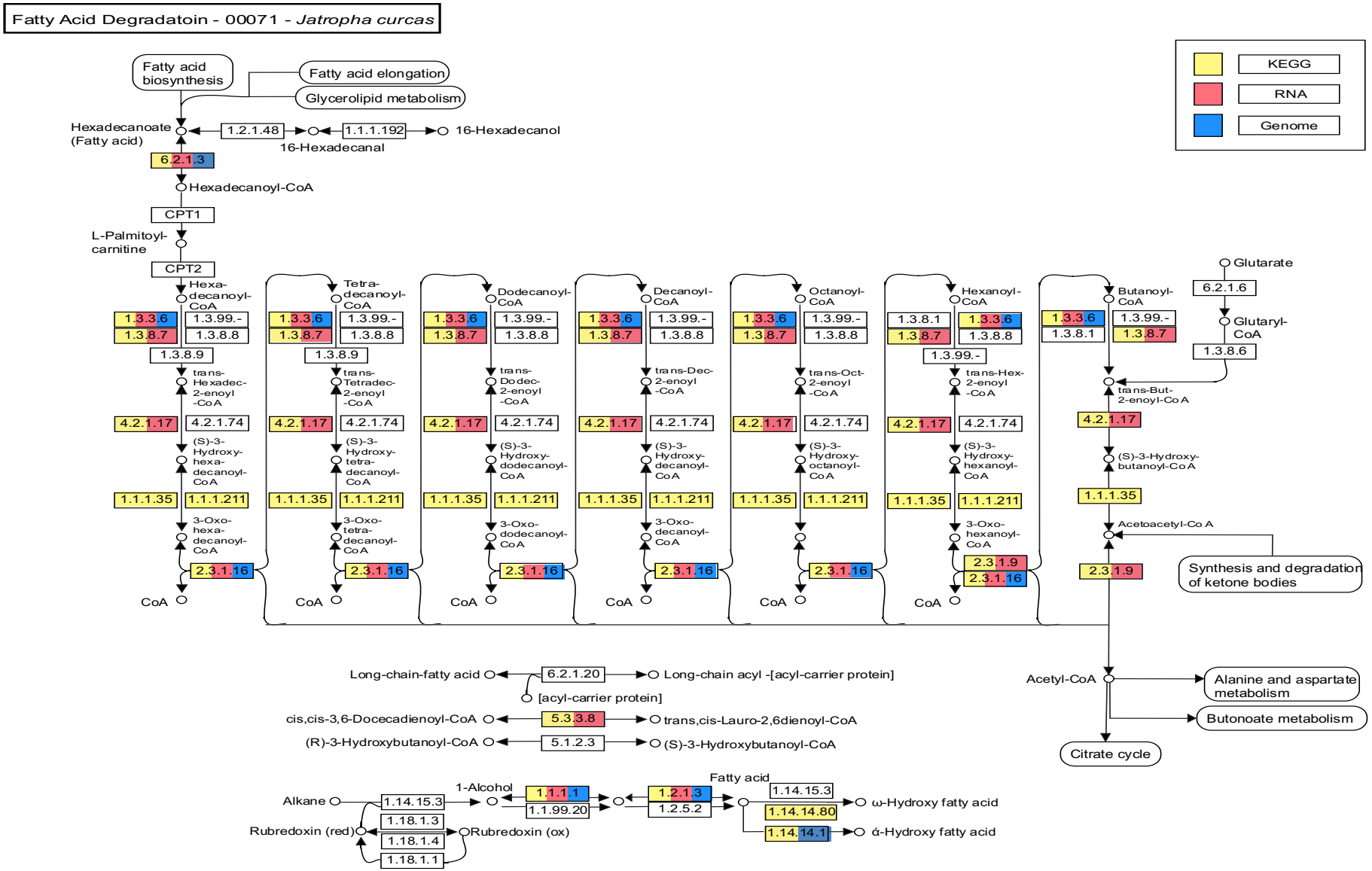
A 7 - Via de Elongação de ácidos graxos indicando as enzimas de *J. curcas* presentes nos dados do KEGG, transcritoma e/ou genoma



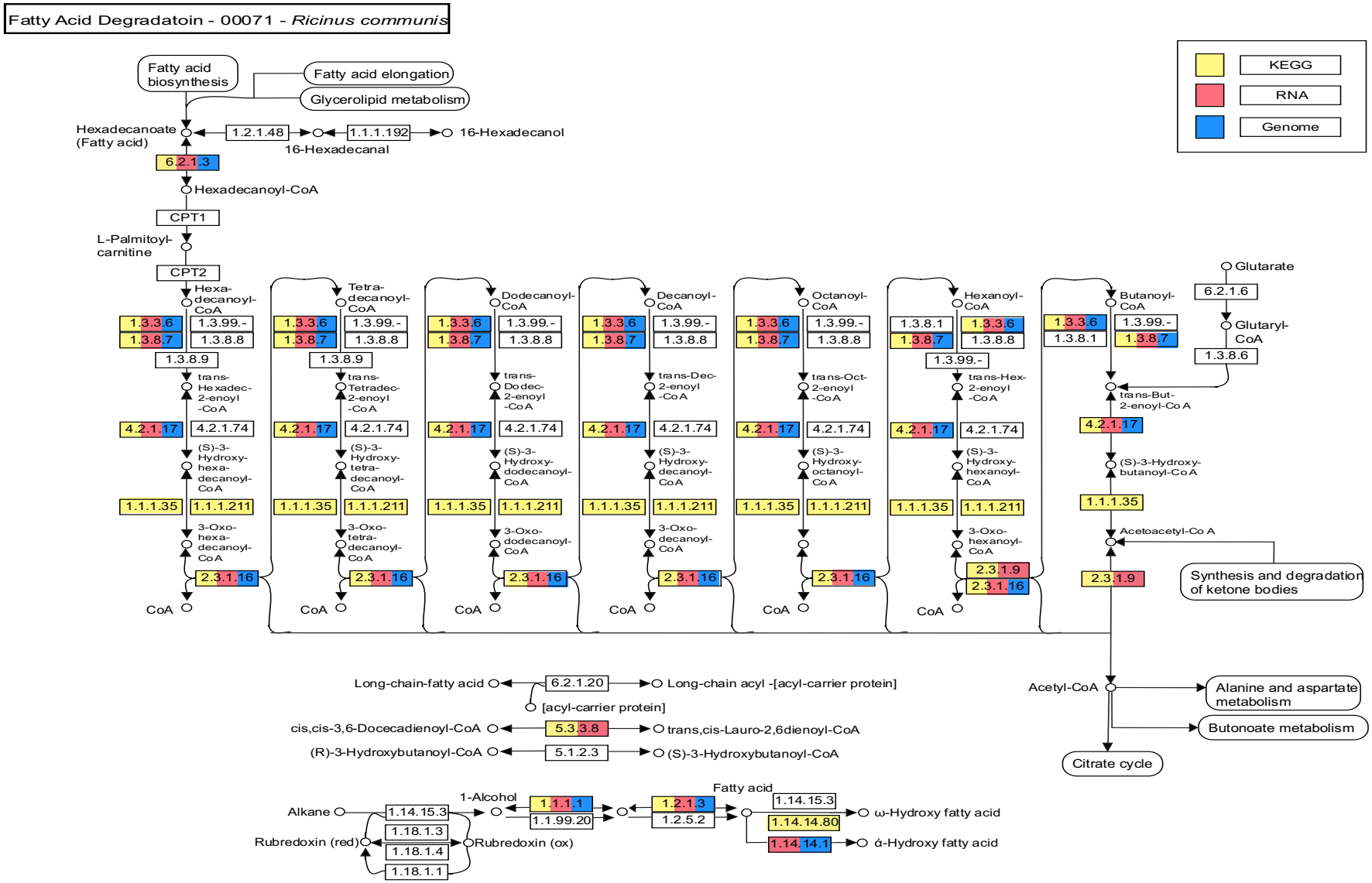
A 8 - Via de Elongação de ácidos graxos indicando as enzimas de *R. communis* presentes nos dados do KEGG, transcritoma e/ou genoma



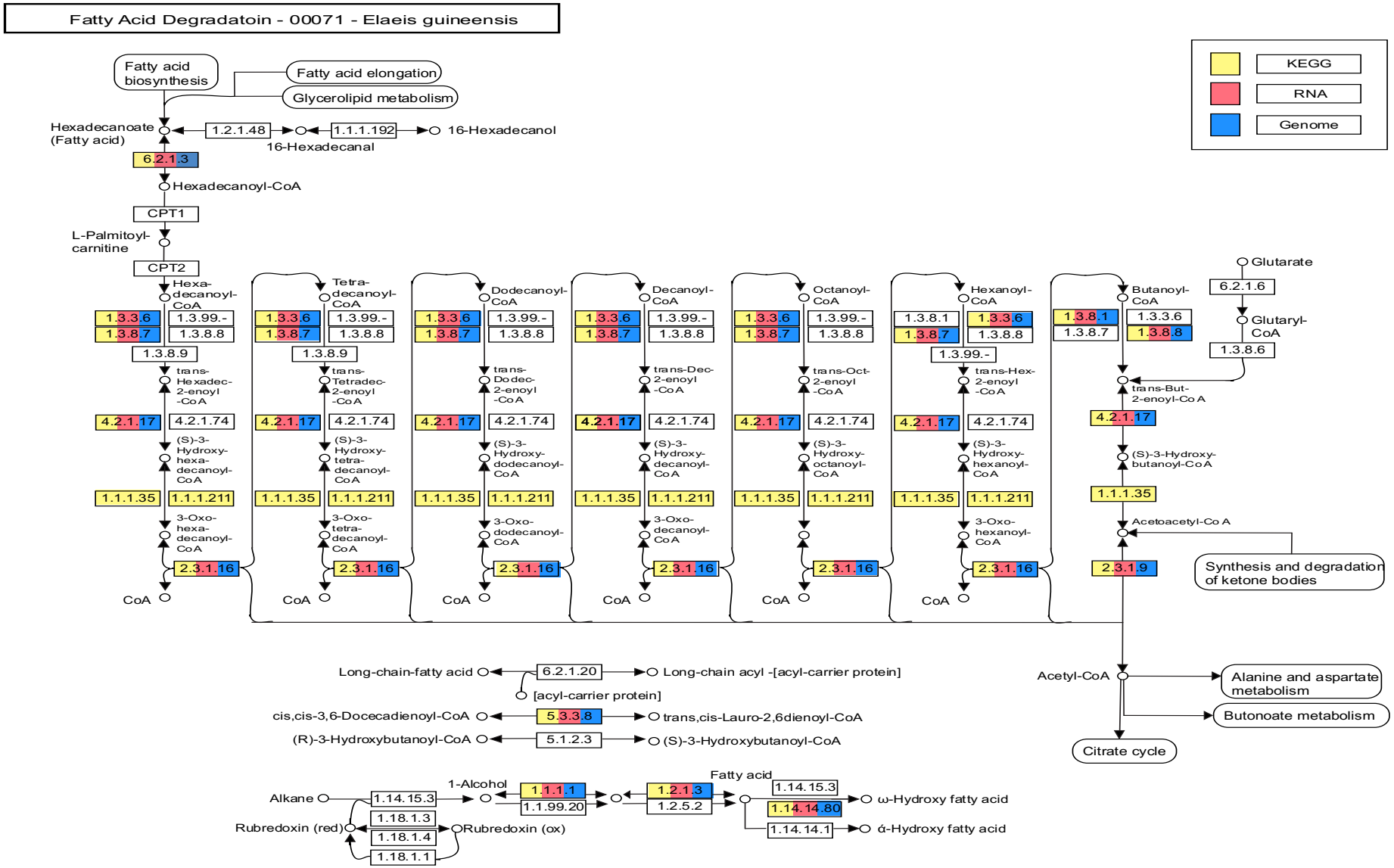
A 9 - Via de Elongação de ácidos graxos indicando as enzimas de *E. guineensis* presentes nos dados do KEGG, transcritoma e/ou genoma.



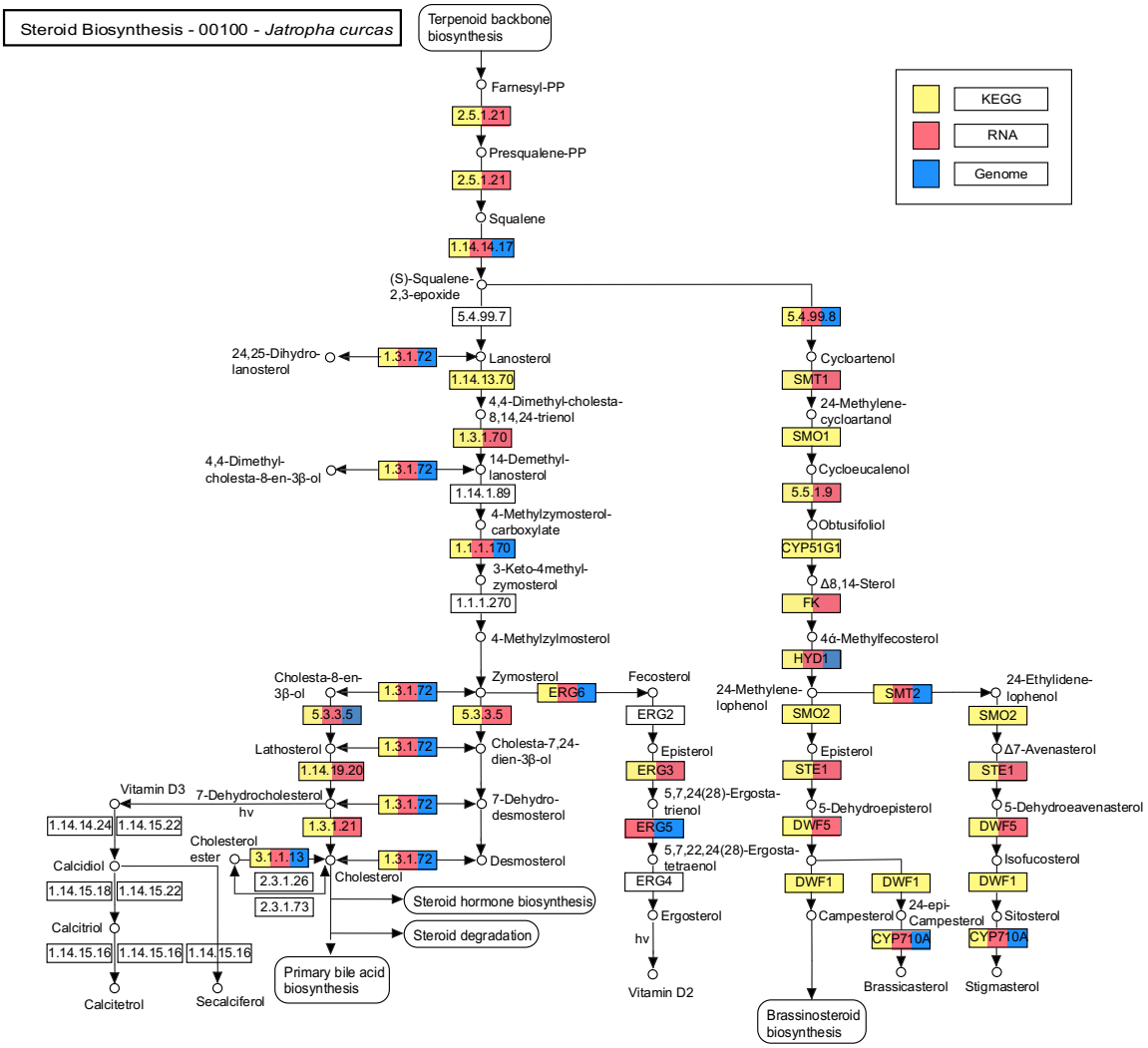
A 10 - Via de Degradação de ácidos graxos indicando as enzimas de *J. curcas* presentes nos dados do KEGG, transcritoma e/ou genoma.



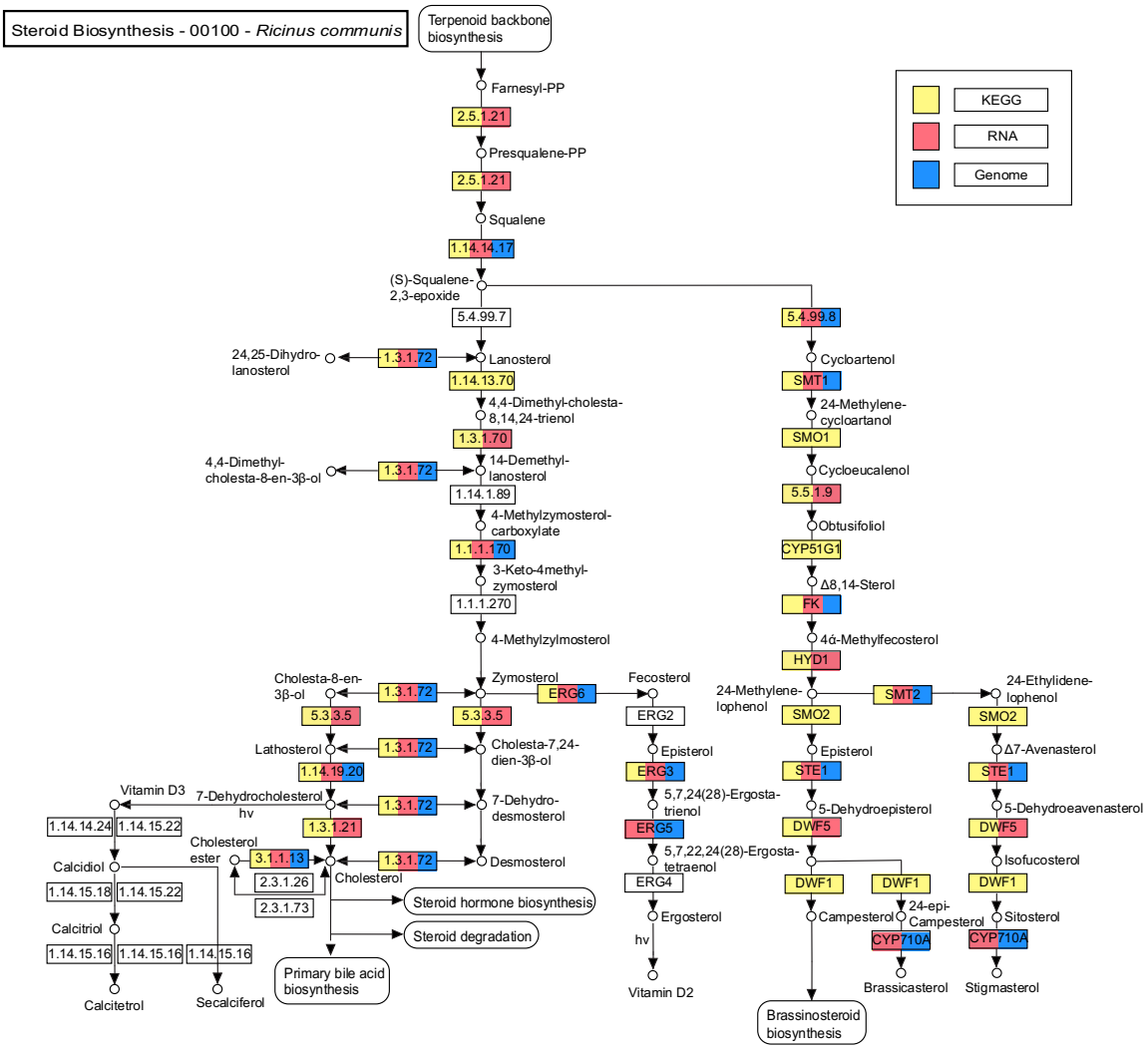
A 11 - Via de Degradação de ácidos graxos indicando as enzimas de *R. communis* presentes nos dados do KEGG, transcritoma e/ou genoma



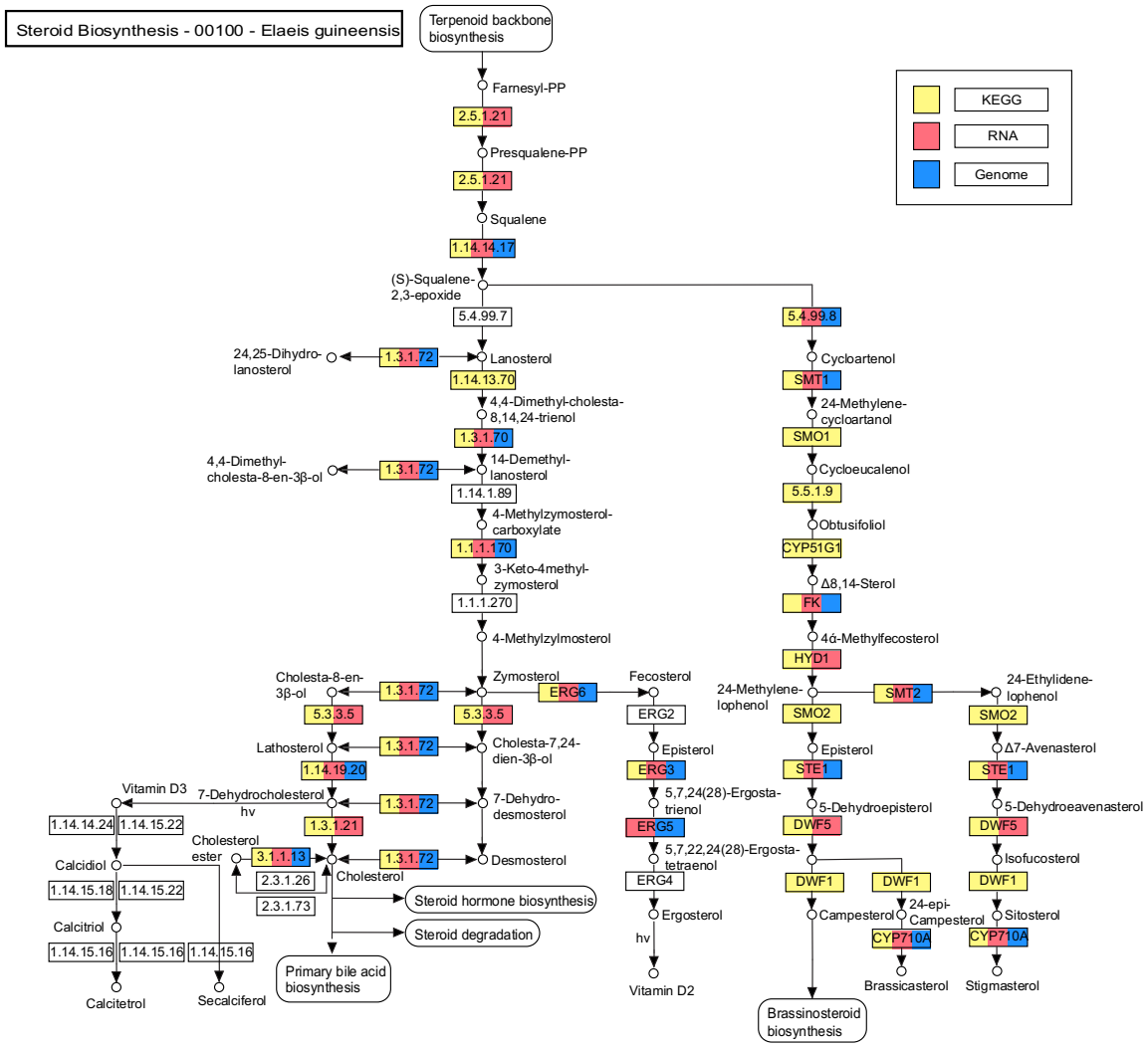
A 12 - Via de Degradação de ácidos graxos indicando as enzimas de *E. guineensis* presentes nos dados do KEGG, transcritoma e/ou genoma.



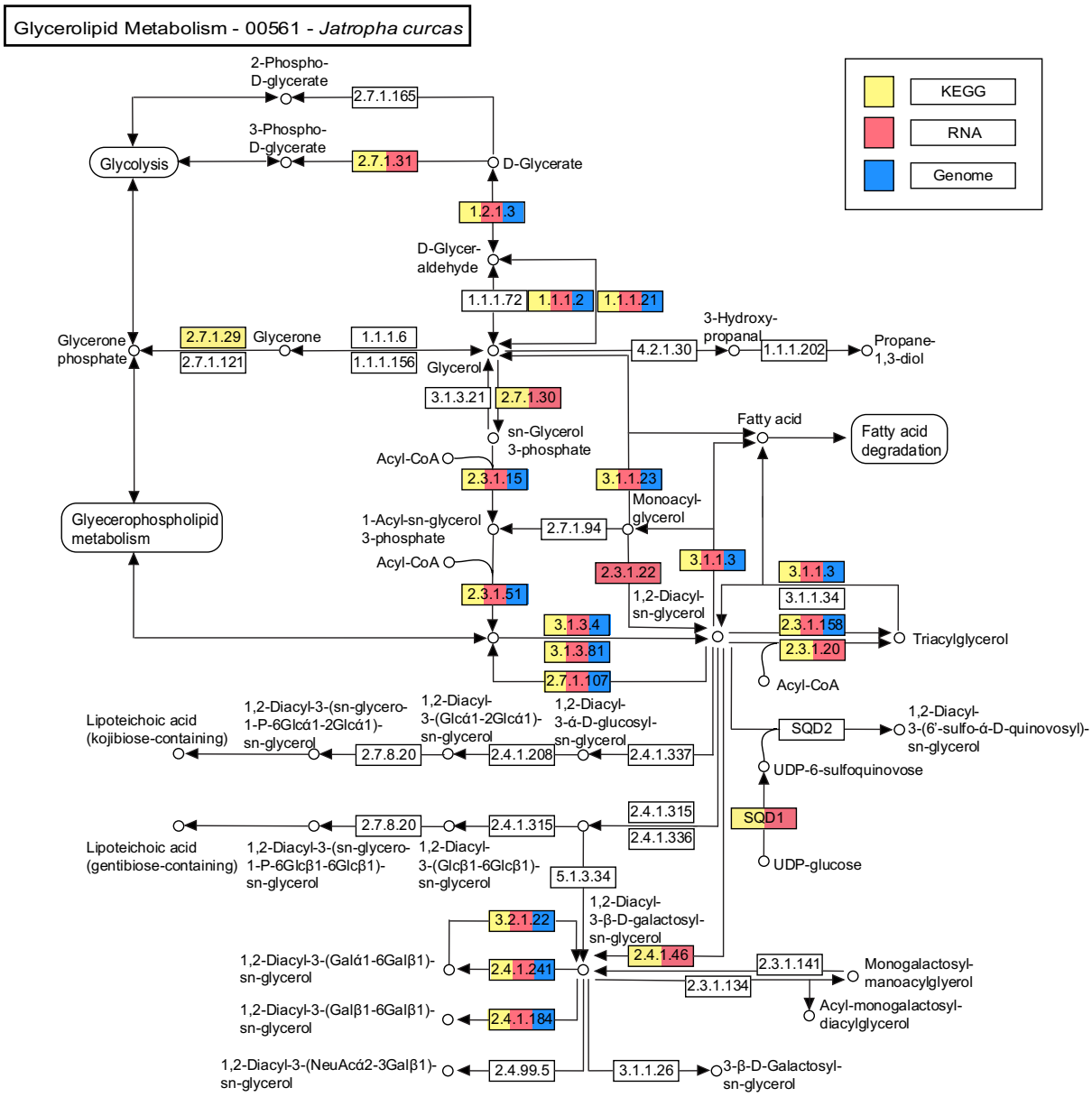
A 13 - Via de biossíntese de esteroides indicando as enzimas de *J. curcas* presentes nos dados do KEGG, transcritoma e/ou genoma



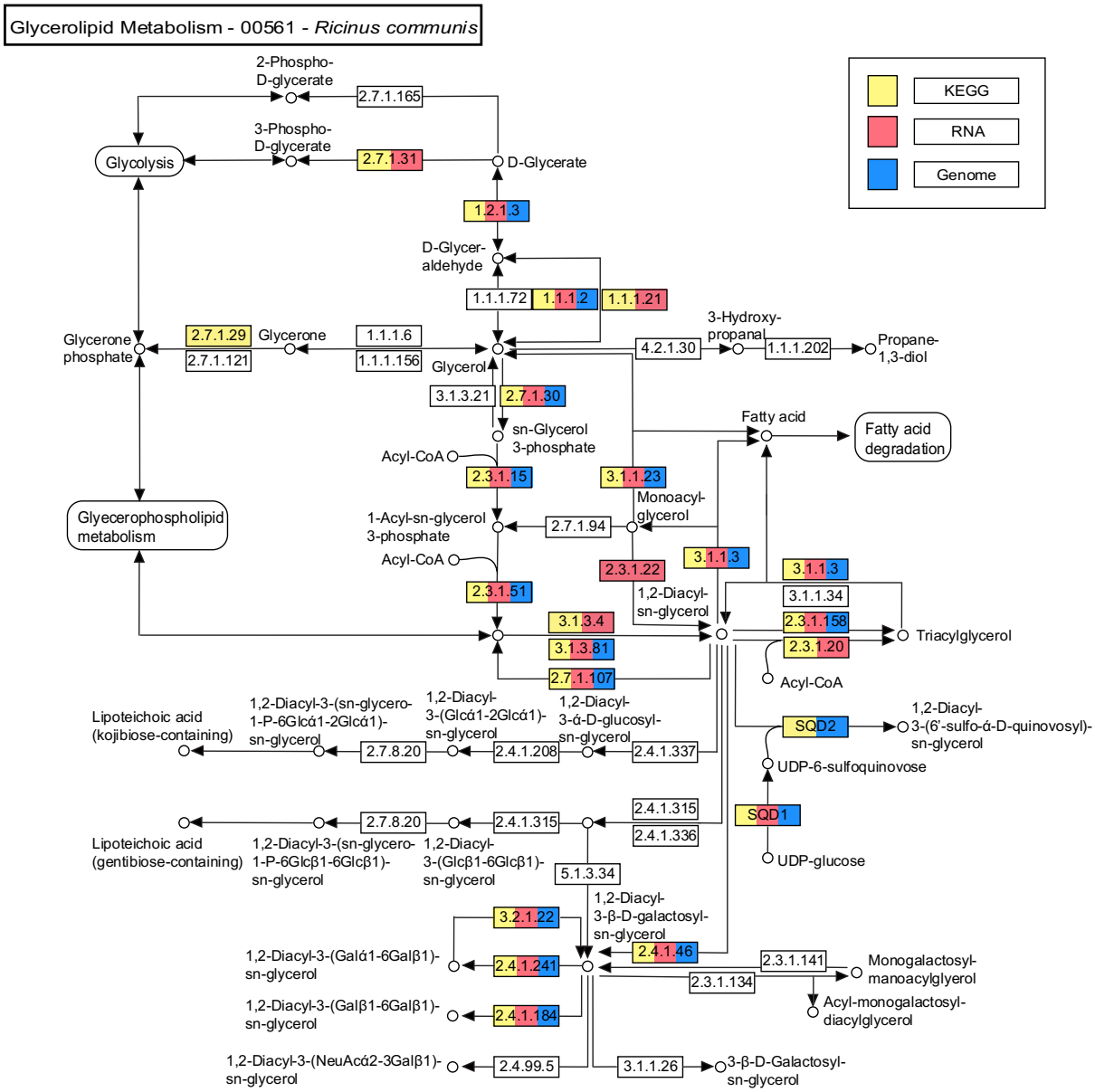
A 14 - Via de biossíntese de esteroides indicando as enzimas de *R. communis* presentes nos dados do KEGG, transcritoma e/ou genoma



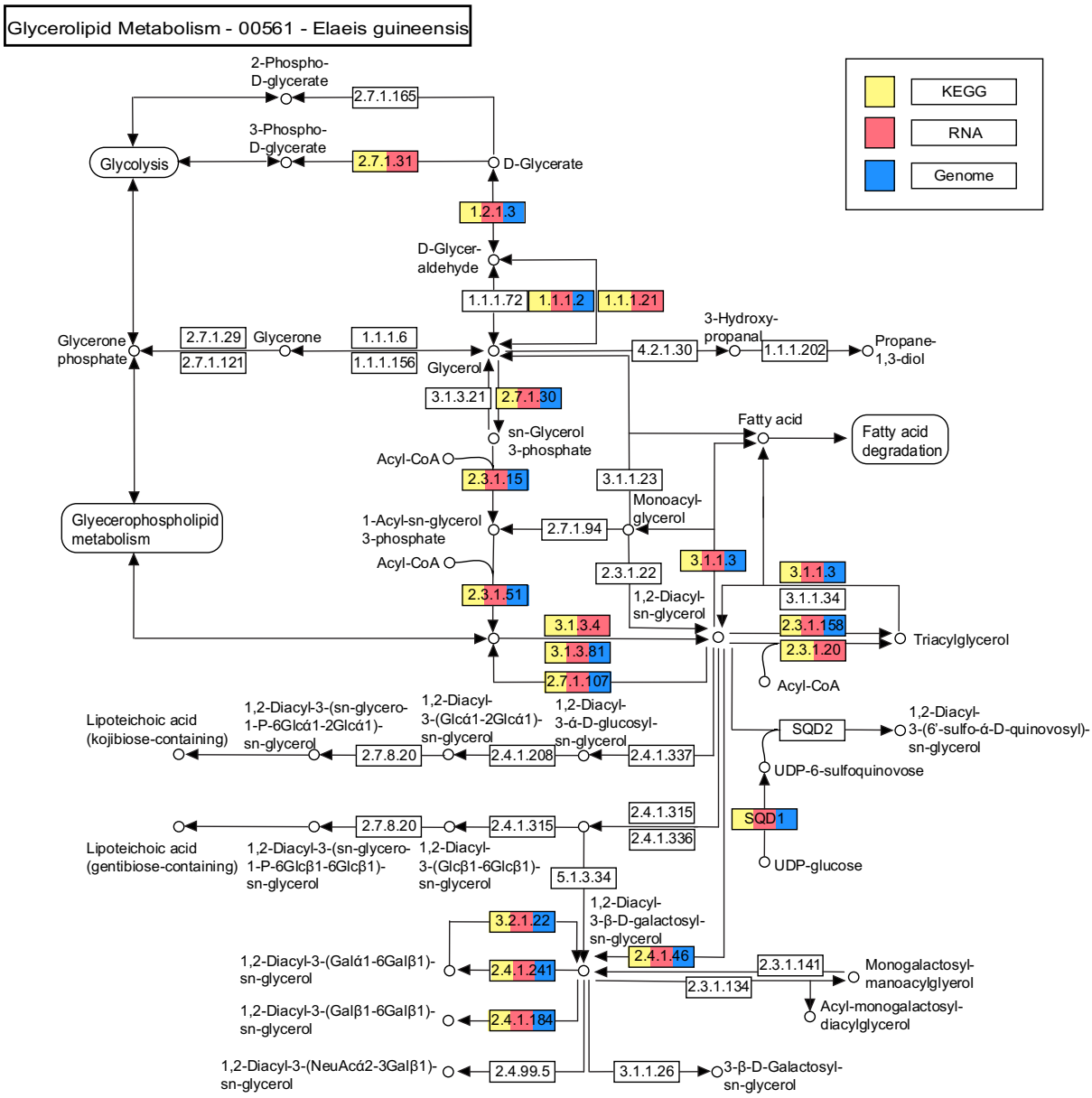
A 15 - Via de biossíntese de esteroides indicando as enzimas de *E. guineensis* presentes nos dados do KEGG, transcritoma e/ou genoma.



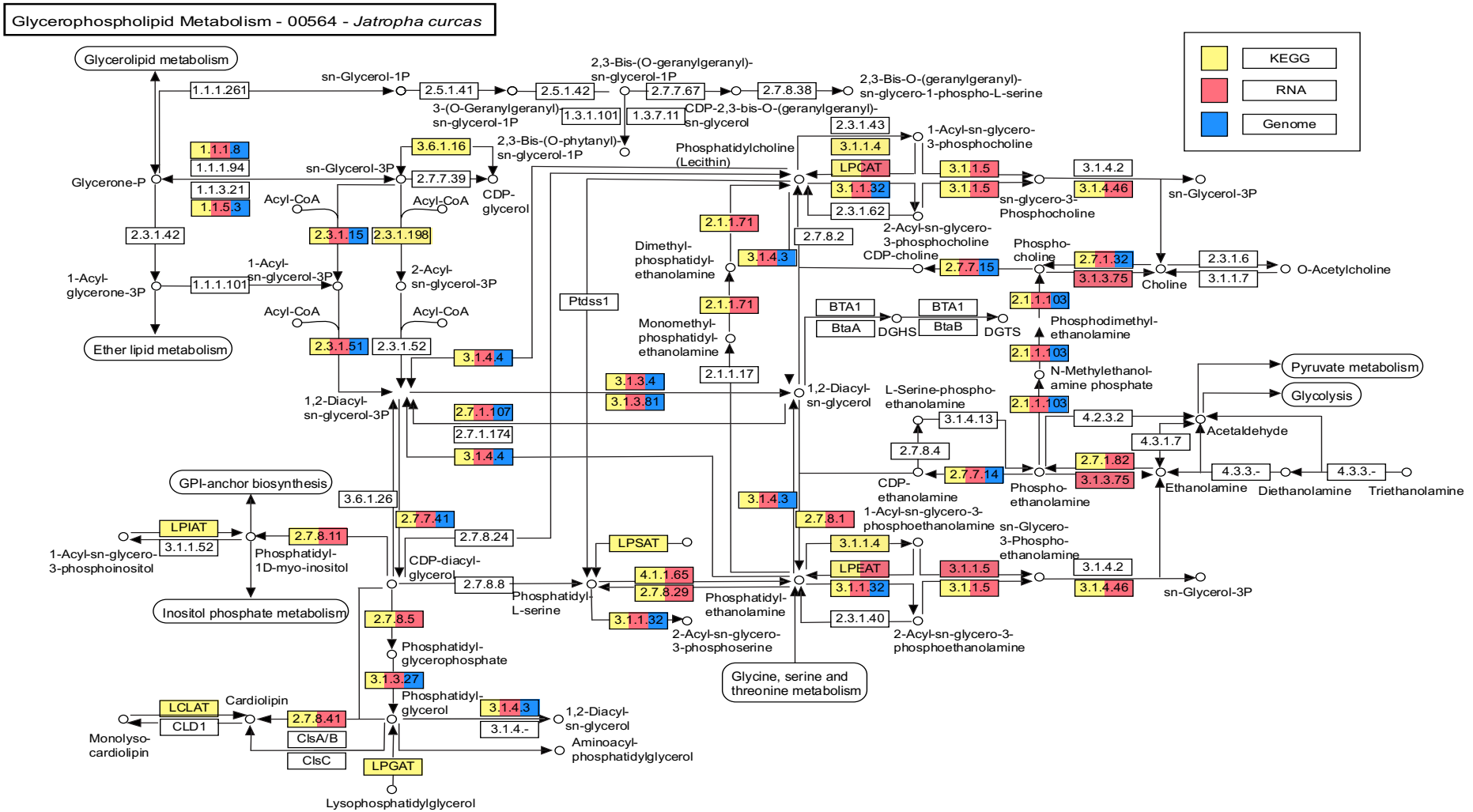
A 16 - Via de metabolismo de glicerolipídeo indicando as enzimas de *J. curcas* presentes nos dados do KEGG, transcritoma e/ou genoma



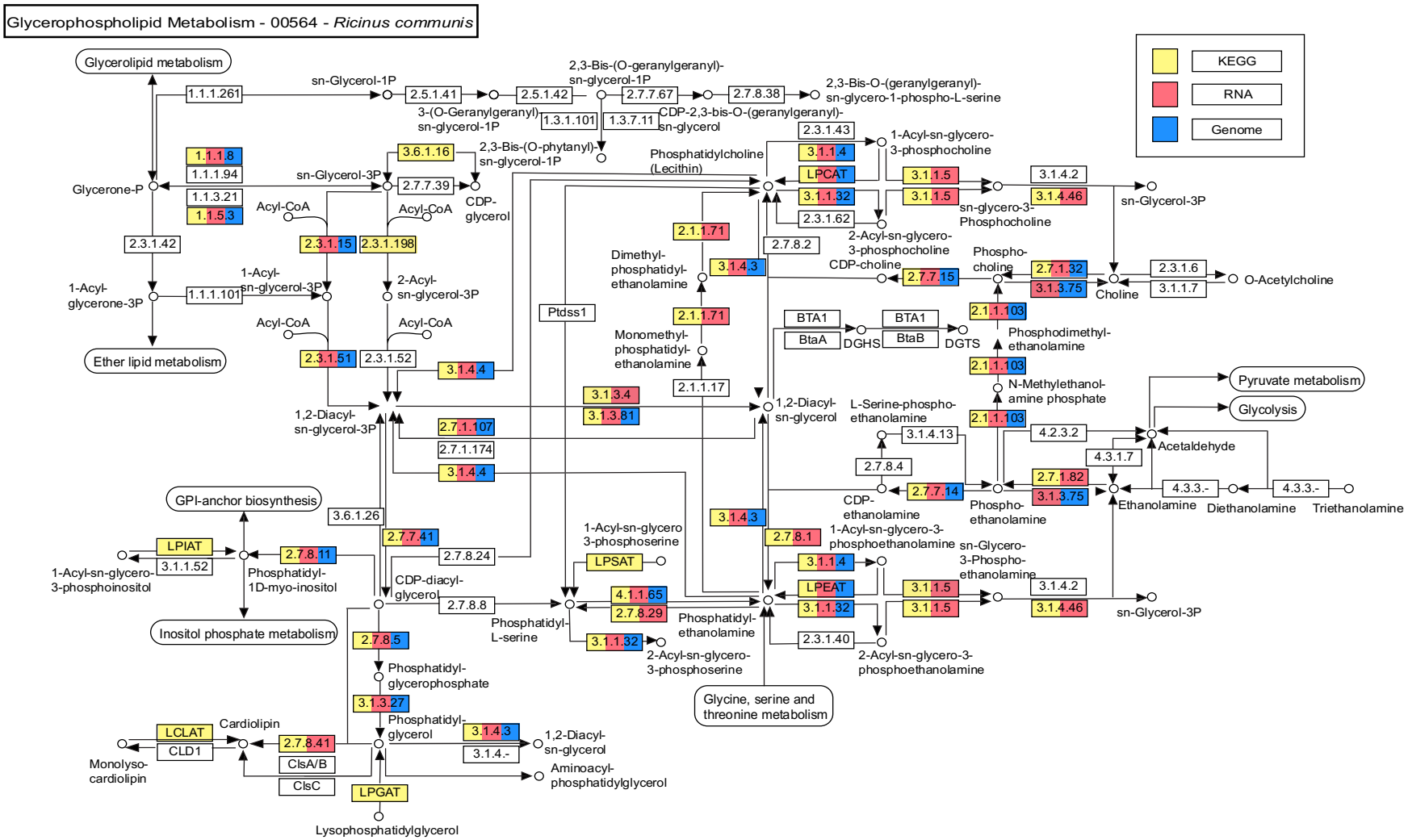
A 17 - Via de metabolismo de glicerolipídeo indicando as enzimas de *R. communis* presentes nos dados do KEGG, transcritoma e/ou genoma



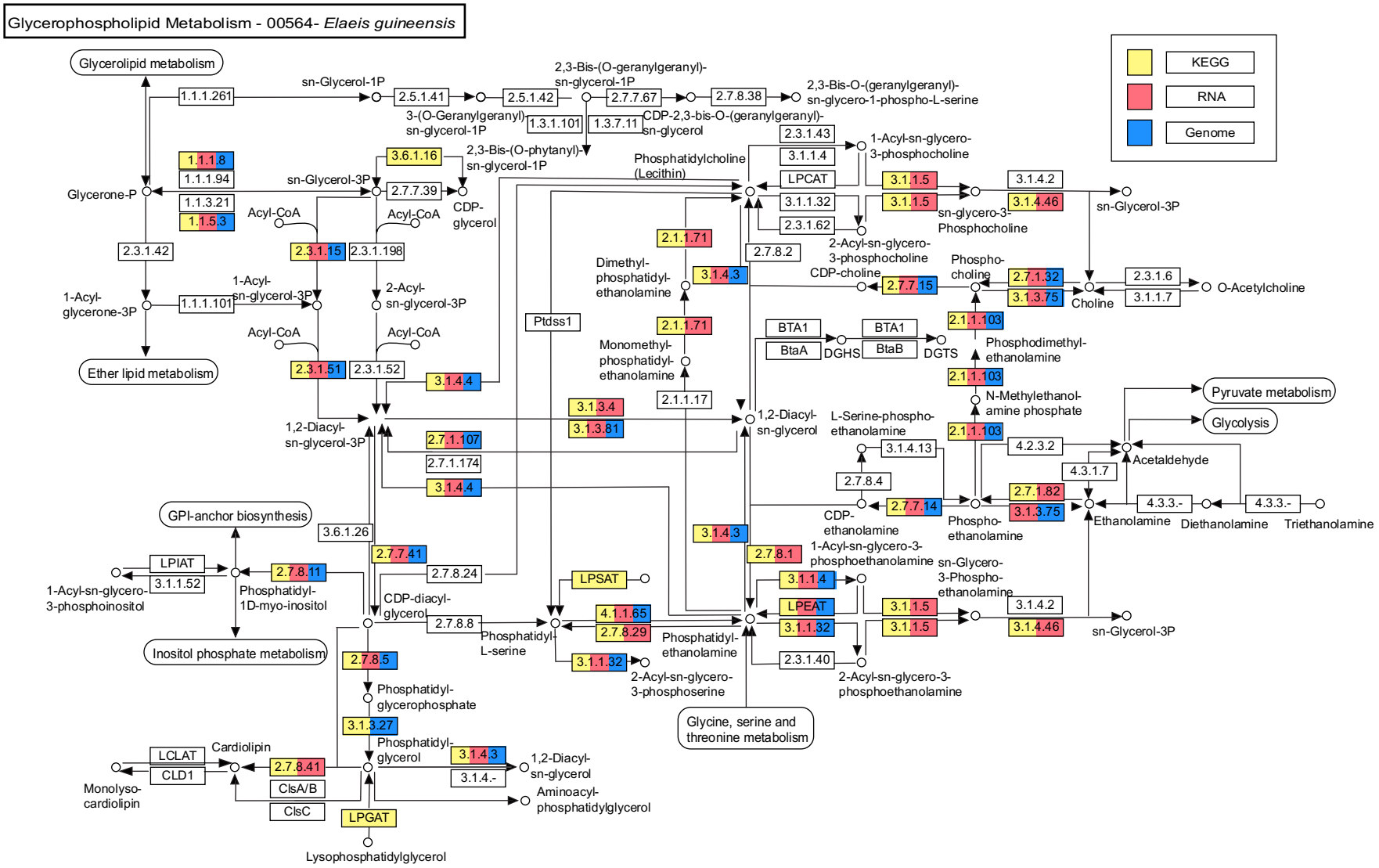
A 18 - Via de metabolismo de glicerolipídeo indicando as enzimas de *E. guineensis* presentes nos dados do KEGG, transcritoma e/ou genoma.



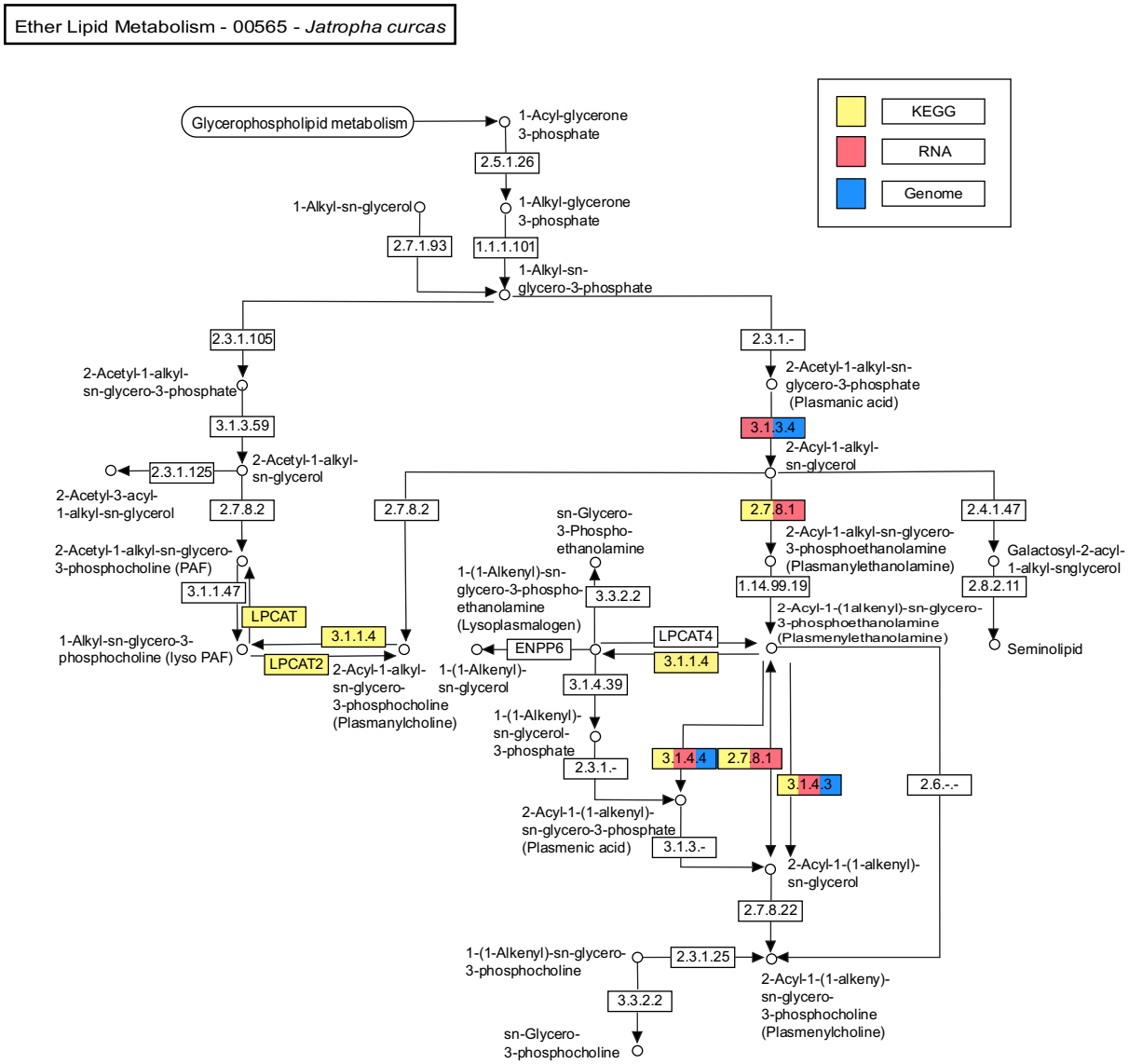
A 19 - Via de metabolismo de glicerofosfolipídeo indicando as enzimas de *J. curcas* presentes nos dados do KEGG, transcritoma e/ou genoma


Glycerophospholipid Metabolism - 00564 - Ricinus communis
KEGG
RNA
Genome
Glycerolipid metabolism
Glycerone-P
sn-Glycerol-1P
sn-Glycerol-3P
1-Acyl-glycerone-3P
1-Acyl-sn-glycerol-3P
2-Acyl-sn-glycerol-3P
1,2-Diacyl-sn-glycerol-3P
1,2-Diacyl-sn-glycerol
Acyl-CoA
CDP-glycerol
Ether lipid metabolism
1.1.1.261
2.5.1.41
2.5.1.42
2.7.7.67
2.7.8.38
2,3-Bis-(O-geranylgeranyl)-sn-glycerol-1P
2,3-Bis-O-(geranylgeranyl)-sn-glycero-1-phospho-L-serine
3-(O-Geranylgeranyl)-sn-glycerol-1P
1.3.1.101
1.3.7.11
CDP-2,3-bis-O-(geranylgeranyl)-sn-glycerol
2,3-Bis-(O-phytanyl)-sn-glycerol-1P
3.6.1.16
1.1.1.8
1.1.1.94
1.1.3.21
1.1.5.3
2.7.7.39
2.3.1.42
2.3.1.15
2.3.1.198
1.1.1.101
2.3.1.51
2.3.1.52
3.1.4.4
2.7.1.107
2.7.1.174
Phosphatidylcholine (Lecithin)
2.3.1.43
3.1.1.4
LPCAT
3.1.1.32
2.3.1.62
1-Acyl-sn-glycero-3-phosphocholine
2-Acyl-sn-glycero-3-phosphocholine
3.1.1.5
3.1.4.2
3.1.4.46
sn-glycero-3-Phosphocholine
2.1.1.71
3.1.4.3
Dimethyl-phosphatidyl-ethanolamine
Monomethyl-phosphatidyl-ethanolamine
Ptdss1
2.7.8.2
CDP-choline
2.7.7.15
Phospho-choline
2.7.1.32
3.1.3.75
Choline
2.3.1.6
3.1.1.7
O-Acetylcholine
BTA1
BtaA
DGHS
BtaB
DGTS
2.1.1.103
Phosphodimethyl-ethanolamine
N-Methylethanol-amine phosphate
3.1.3.4
3.1.3.81
2.1.1.17
1,2-Diacyl-sn-glycerol
L-Serine-phospho-ethanolamine
3.1.4.13
2.7.8.4
4.2.3.2
4.3.1.7
Acetaldehyde
Pyruvate metabolism
Glycolysis
2.7.7.14
CDP-ethanolamine
Phospho-ethanolamine
2.7.1.82
3.1.3.75
4.3.3.-
Ethanolamine
Diethanolamine
Triethanolamine
GPI-anchor biosynthesis
3.6.1.26
2.7.7.41
2.7.8.24
LPIAT
3.1.1.52
2.7.8.11
1-Acyl-sn-glycero-3-phosphoinositol
Phosphatidyl-1D-myo-inositol
Inositol phosphate metabolism
CDP-diacyl-glycerol
2.7.8.8
Phosphatidyl-L-serine
1-Acyl-sn-glycero-3-phosphoserine
LPSAT
4.1.1.65
2.7.8.29
3.1.1.32
2-Acyl-sn-glycero-3-phosphoserine
Phosphatidyl-ethanolamine
1-Acyl-sn-glycero-3-phosphoethanolamine
sn-Glycero-3-Phospho-ethanolamine
LPEAT
2.3.1.40
2-Acyl-sn-glycero-3-phosphoethanolamine
2.7.8.1
3.1.4.3
Glycine, serine and threonine metabolism
2.7.8.5
Phosphatidyl-glycerophosphate
3.1.3.27
Phosphatidyl-glycerol
3.1.4.3
3.1.4.-
1,2-Diacyl-sn-glycerol
Aminoacyl-phosphatidylglycerol
Cardiolipin
LCLAT
CLD1
2.7.8.41
ClsA/B
ClsC
LPGAT
Monolyso-cardiolipin
Lysophosphatidylglycerol

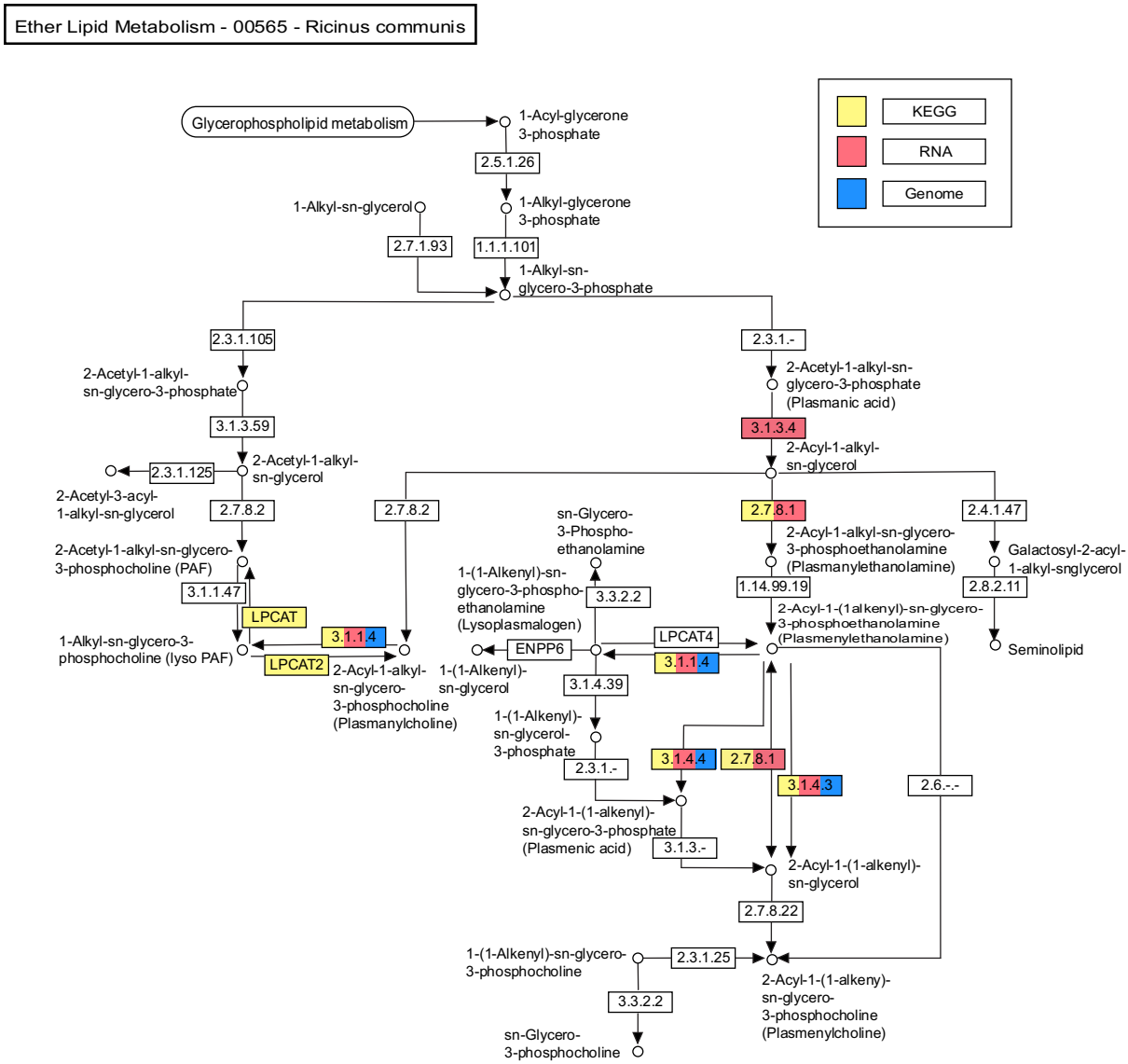
A 20 - Via de metabolismo de glicerofosfolipídeo indicando as enzimas de *R. communis* presentes nos dados do KEGG, transcritoma e/ou genoma.


Glycerophospholipid Metabolism - 00564- Elaeis guineensis
KEGG
RNA
Genome
Glycerolipid metabolism
Glycerone-P
sn-Glycerol-1P
sn-Glycerol-3P
1-Acyl-glycerone-3P
1-Acyl-sn-glycerol-3P
2-Acyl-sn-glycerol-3P
Acyl-CoA
CDP-glycerol
Ether lipid metabolism
1,2-Diacyl-sn-glycerol-3P
1,2-Diacyl-sn-glycerol
3-(O-Geranylgeranyl)-sn-glycerol-1P
2,3-Bis-(O-geranylgeranyl)-sn-glycerol-1P
2,3-Bis-(O-phytanyl)-sn-glycerol-1P
2,3-Bis-O-(geranylgeranyl)-sn-glycero-1-phospho-L-serine
CDP-2,3-bis-O-(geranylgeranyl)-sn-glycerol
Phosphatidylcholine (Lecithin)
1-Acyl-sn-glycero-3-phosphocholine
2-Acyl-sn-glycero-3-phosphocholine
sn-glycero-3-Phosphocholine
LPCAT
Dimethyl-phosphatidyl-ethanolamine
Monomethyl-phosphatidyl-ethanolamine
Ptdss1
CDP-choline
Phospho-choline
Choline
O-Acetylcholine
BTA1
BtaA
BtaB
DGHS
DGTS
Phosphodimethyl-ethanolamine
N-Methylethanol-amine phosphate
L-Serine-phospho-ethanolamine
CDP-ethanolamine
Phospho-ethanolamine
Ethanolamine
Diethanolamine
Triethanolamine
Acetaldehyde
Pyruvate metabolism
Glycolysis
GPI-anchor biosynthesis
LPIAT
1-Acyl-sn-glycero-3-phosphoinositol
Phosphatidyl-1D-myo-inositol
Inositol phosphate metabolism
CDP-diacyl-glycerol
Phosphatidyl-L-serine
Phosphatidyl-ethanolamine
1-Acyl-sn-glycero-3-phosphoethanolamine
sn-Glycero-3-Phospho-ethanolamine
LPSAT
LPEAT
2-Acyl-sn-glycero-3-phosphoserine
2-Acyl-sn-glycero-3-phosphoethanolamine
Glycine, serine and threonine metabolism
Phosphatidyl-glycerophosphate
Phosphatidyl-glycerol
Cardiolipin
LCLAT
CLD1
Monolyso-cardiolipin
ClsA/B
ClsC
LPGAT
Lysophosphatidylglycerol
Aminoacyl-phosphatidylglycerol

A 21 - Via de metabolismo de glicerofosfolipídeo indicando as enzimas de *E. guineensis* presentes nos dados do KEGG, transcritoma e/ou genoma.

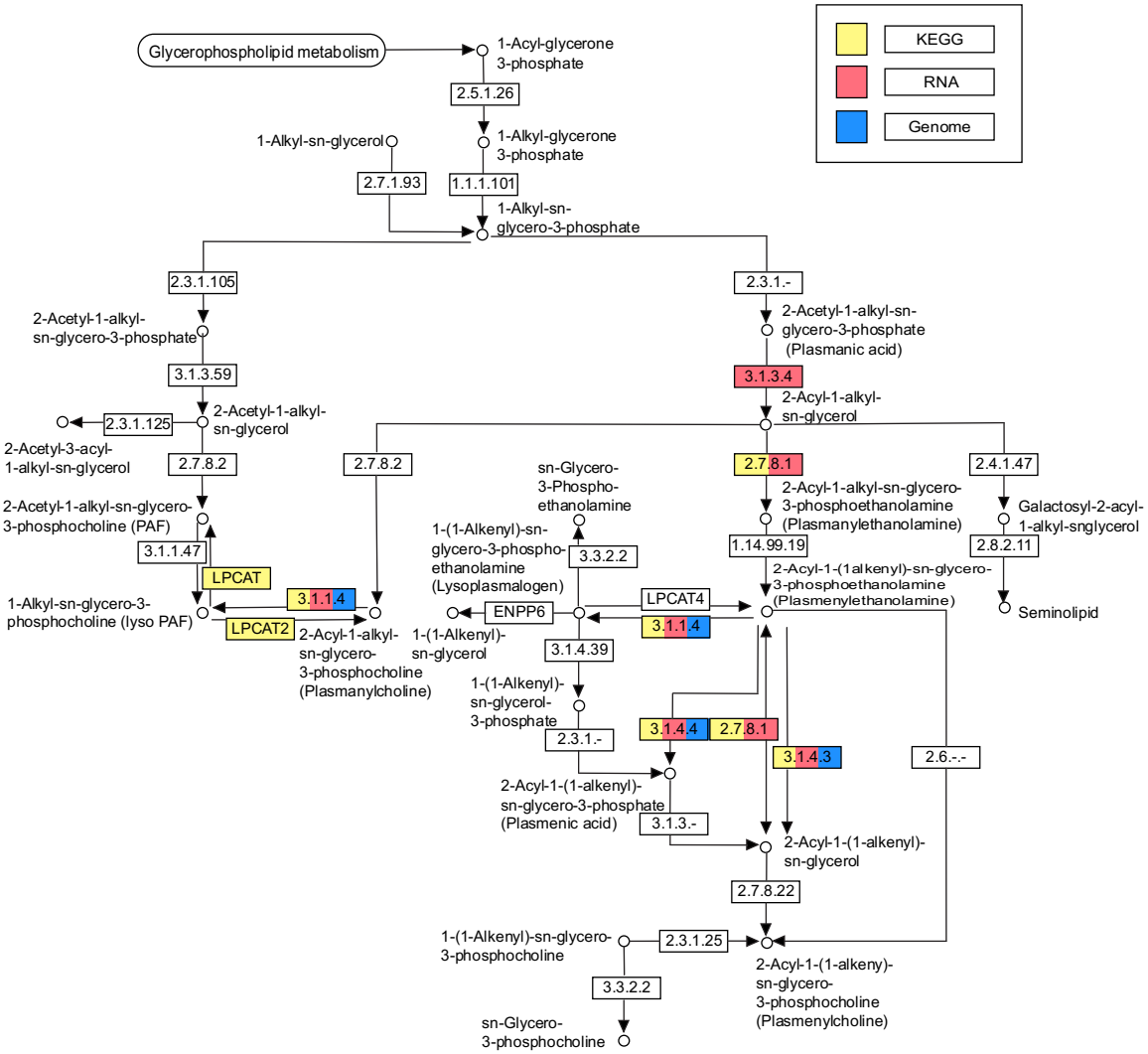


A 22 - Via de metabolismo de éter lipidico indicando as enzimas de *J. curcas* presentes nos dados do KEGG, transcritoma e/ou genoma

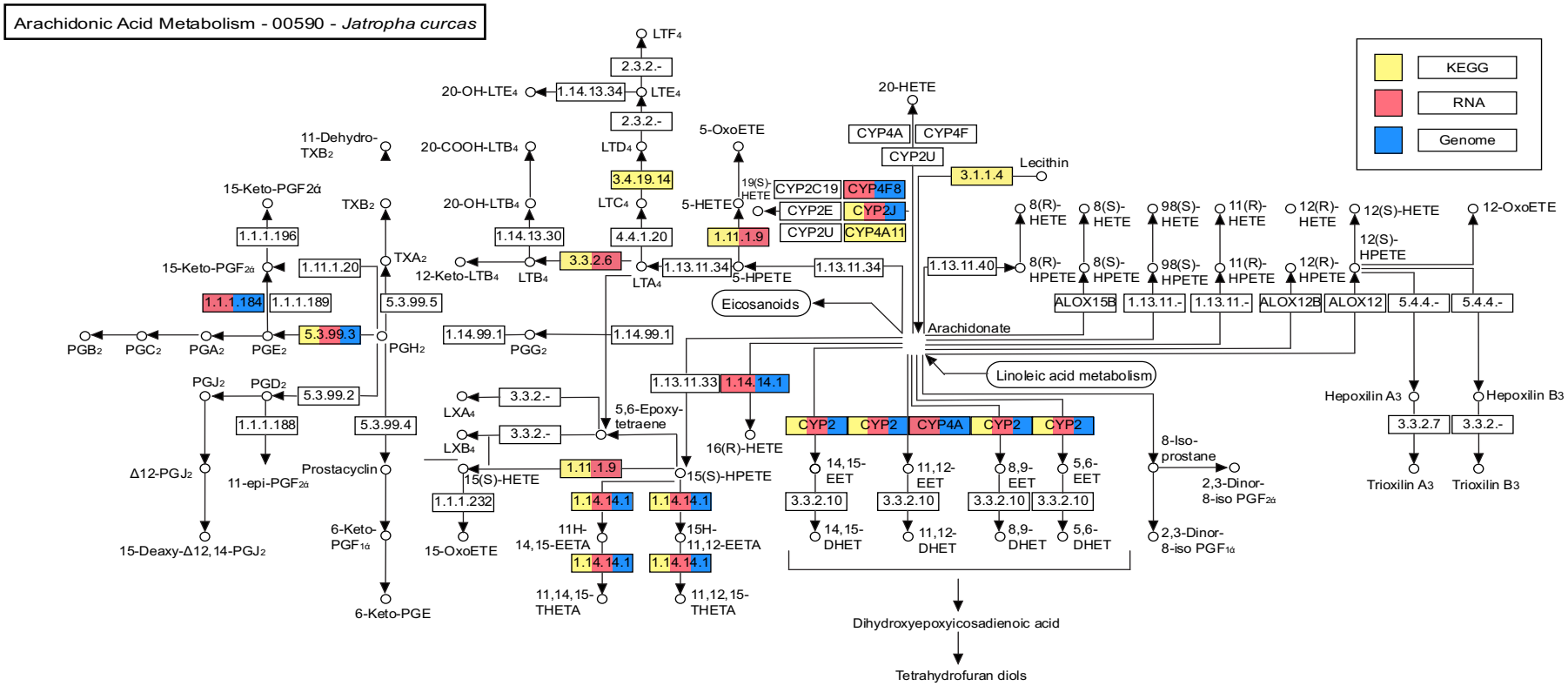


A 23 - Via de metabolismo de éter lipidico indicando as enzimas de *R. communis* presentes nos dados do KEGG, transcritoma e/ou genoma

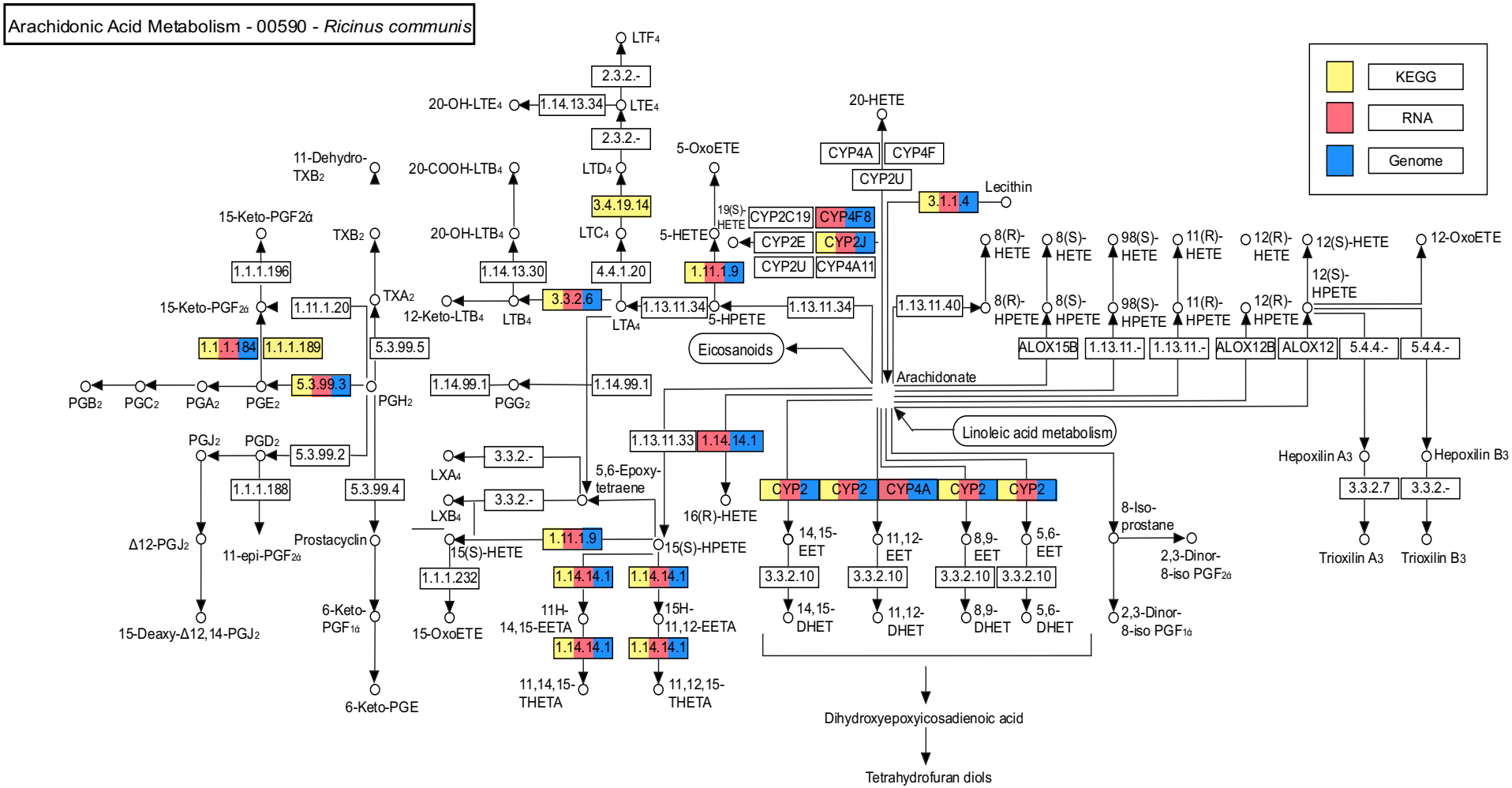
Ether Lipid Metabolism - 00565 - *Elaeis guineensis*



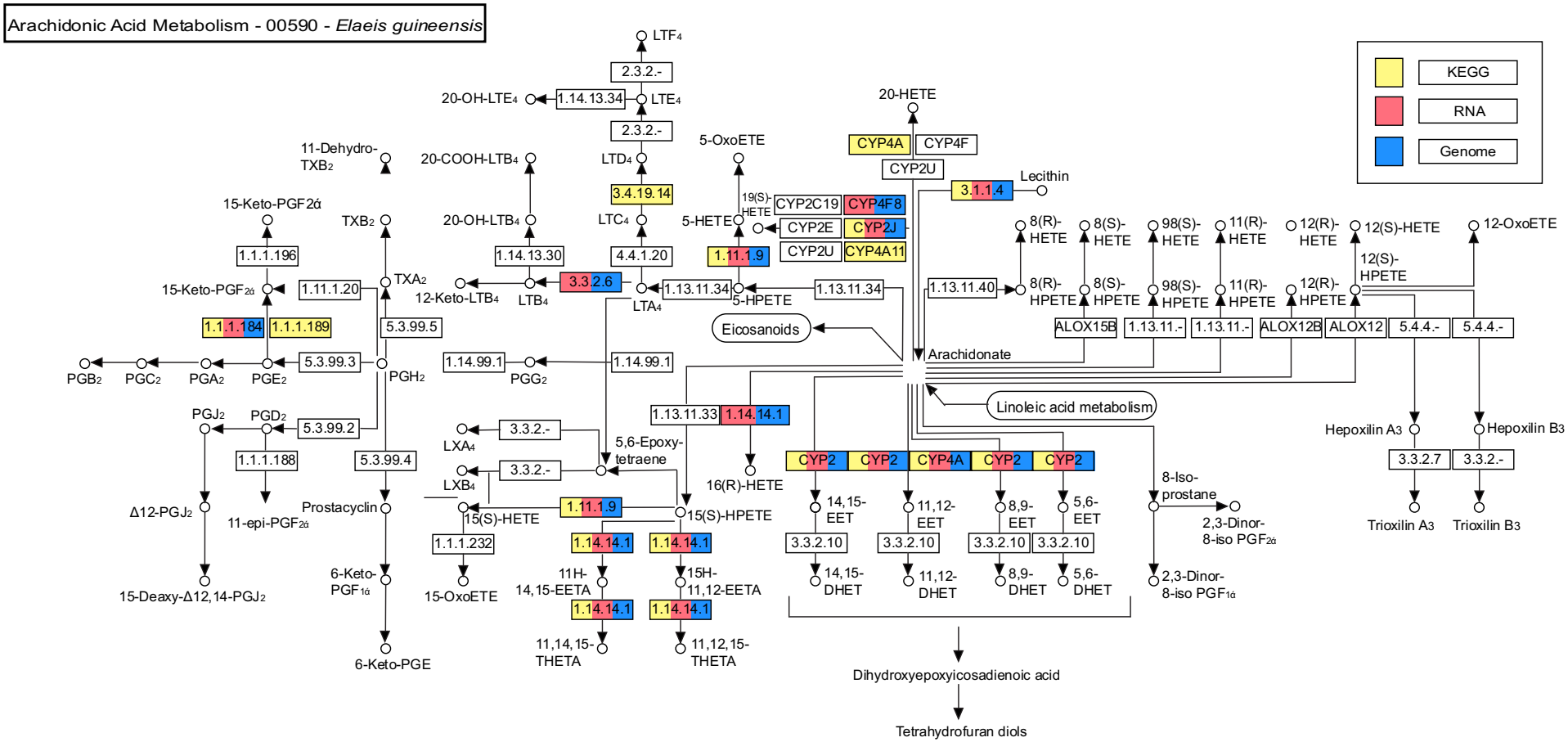
A 24 - Via de metabolismo de éter lipidico indicando as enzimas de *E. guineensis* presentes nos dados do KEGG, transcritoma e/ou genoma.



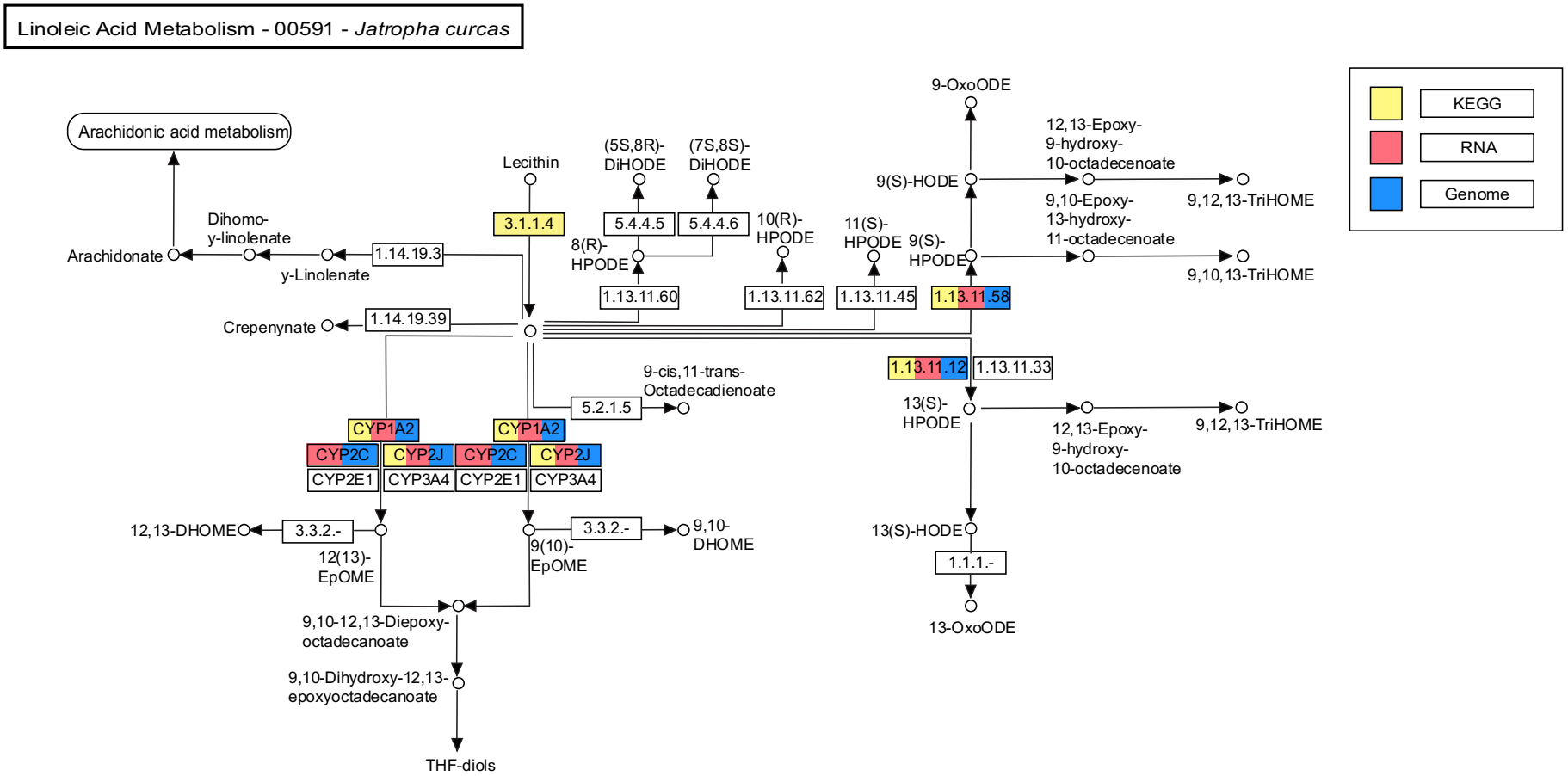
A 25 - Via de metabolismo de araquidônico indicando as enzimas de *J. curcas* presentes nos dados do KEGG, transcritoma e/ou genoma



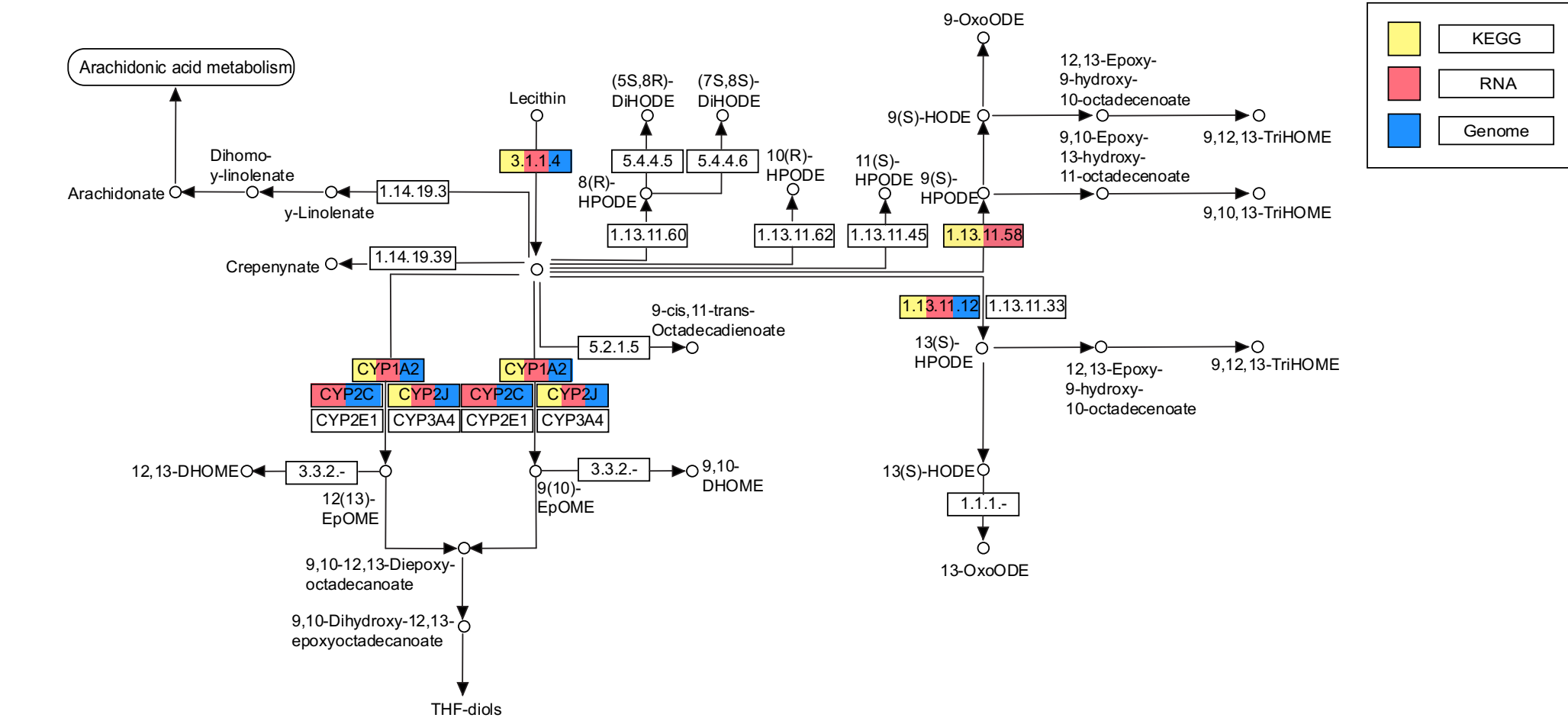
A 26 - Via de metabolismo de araquidônico indicando as enzimas de *R. communis* presentes nos dados do KEGG, transcritoma e/ou genoma.



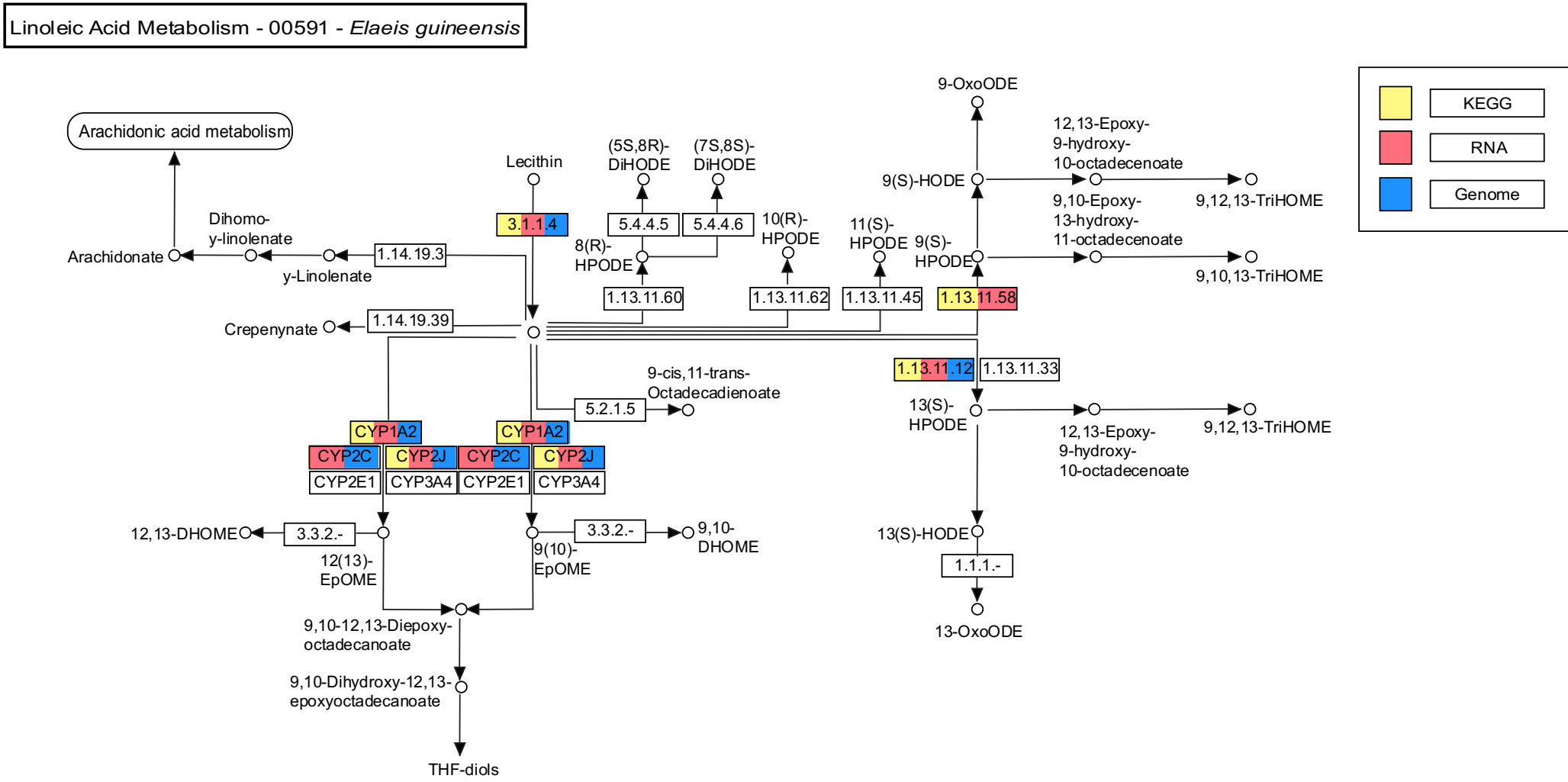
A 27 - Via de metabolismo de araquidônico indicando as enzimas de *E. guineensis* presentes nos dados do KEGG, transcritoma e/ou genoma.



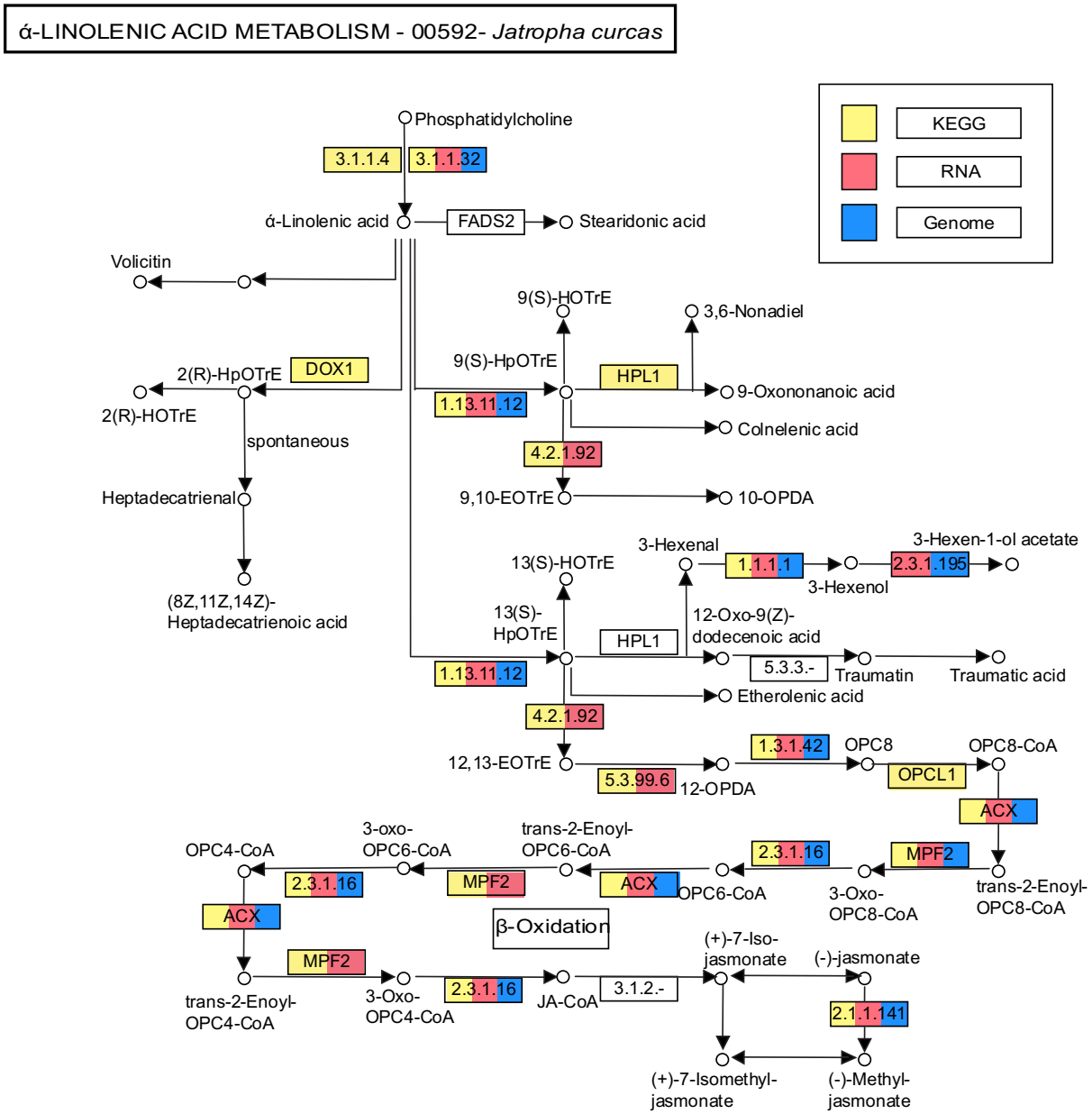
A 28 - Via de metabolismo de linoleico indicando as enzimas de *J. curcas* presentes nos dados do KEGG, transcritoma e/ou genoma.



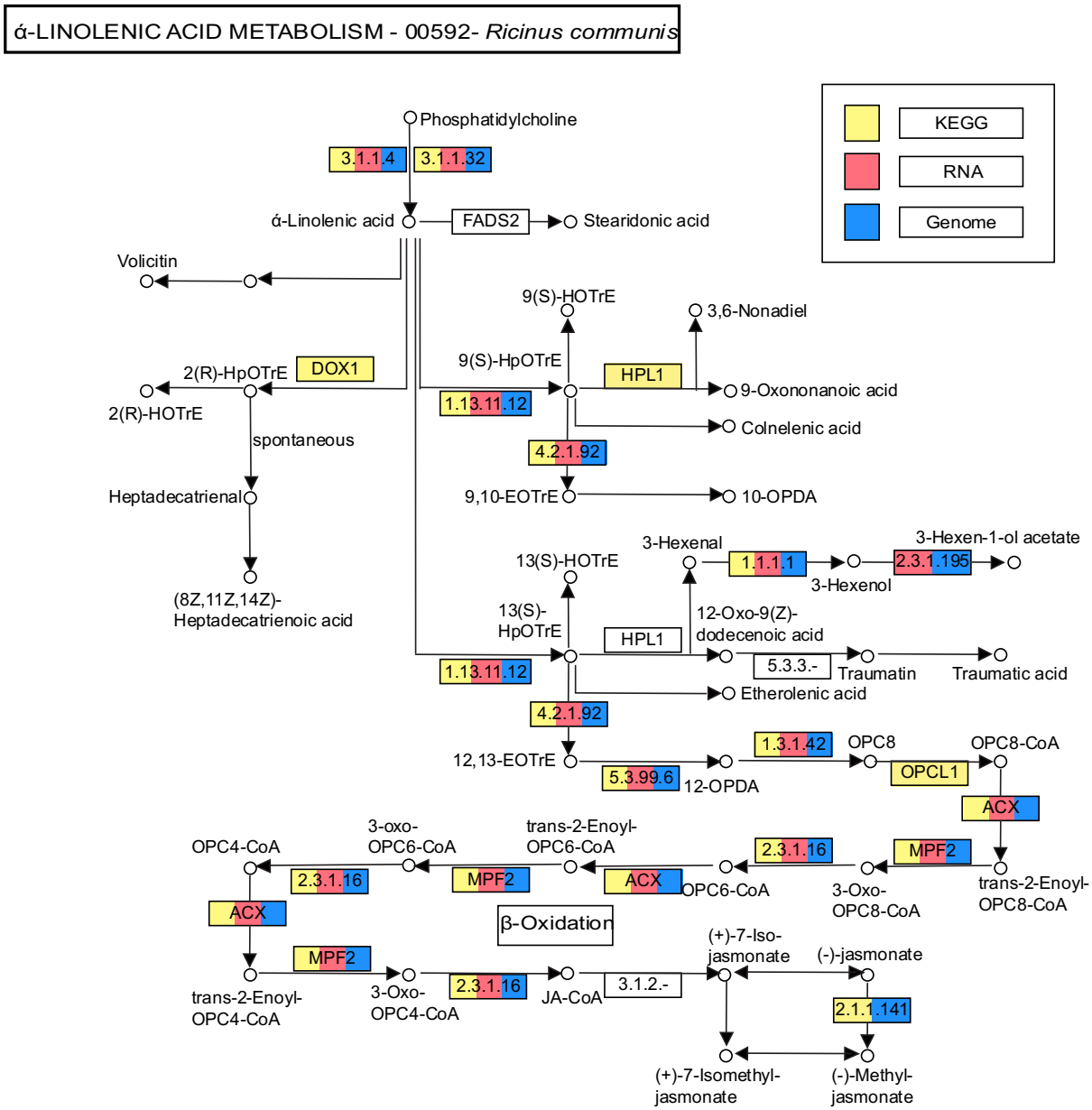
A 29 - Via de metabolismo de linoleico indicando as enzimas de *R. communis* presentes nos dados do KEGG, transcritoma e/ou genoma.



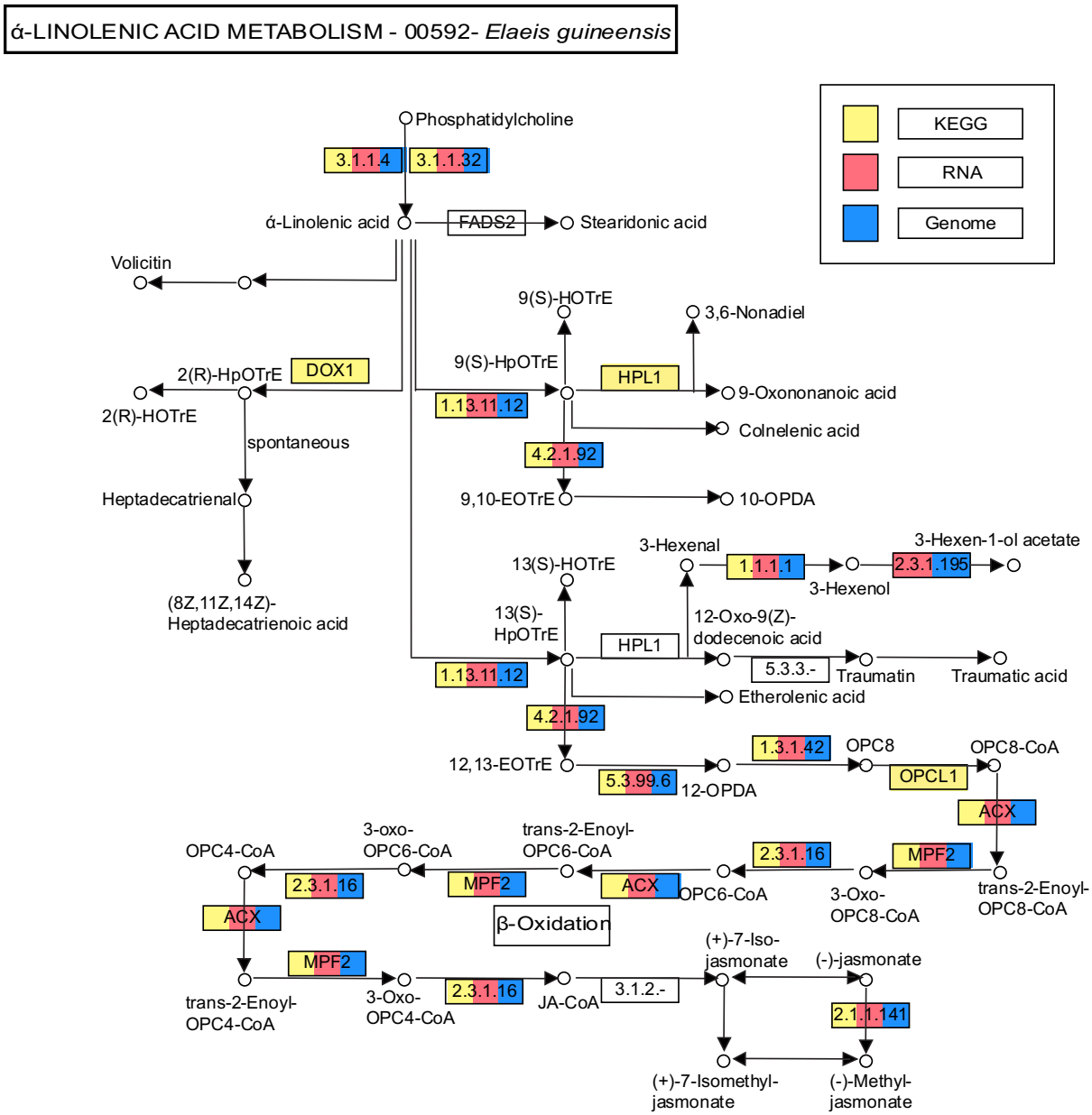
A 30 - Via de metabolismo de linoleico indicando as enzimas de *E. guineensis* presentes nos dados do KEGG, transcritoma e/ou genoma



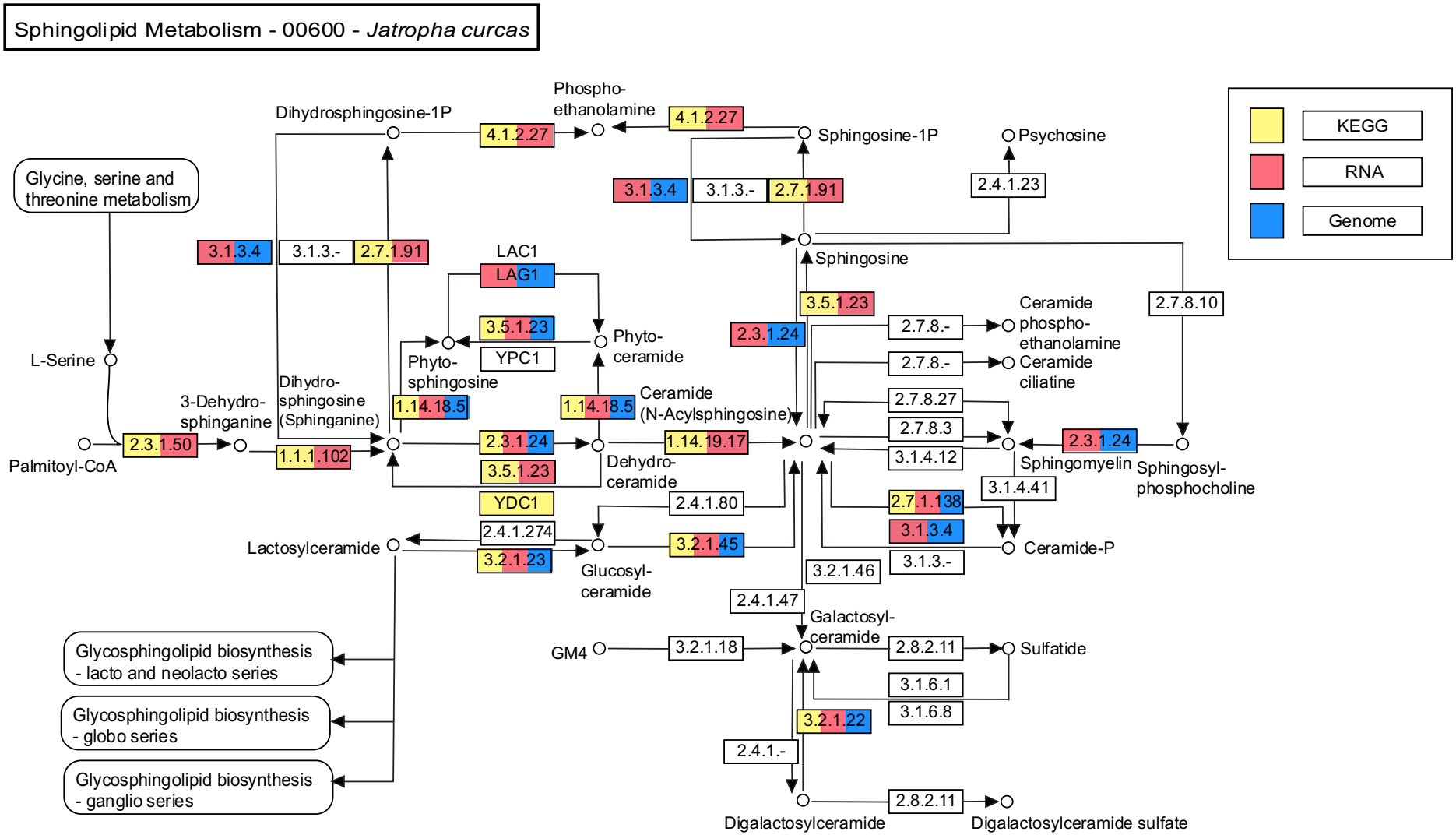
A 31 - Via de metabolismo de ácido ɑ-linolenico indicando as enzimas de *J. curcas* presentes nos dados do KEGG, transcritoma e/ou genoma



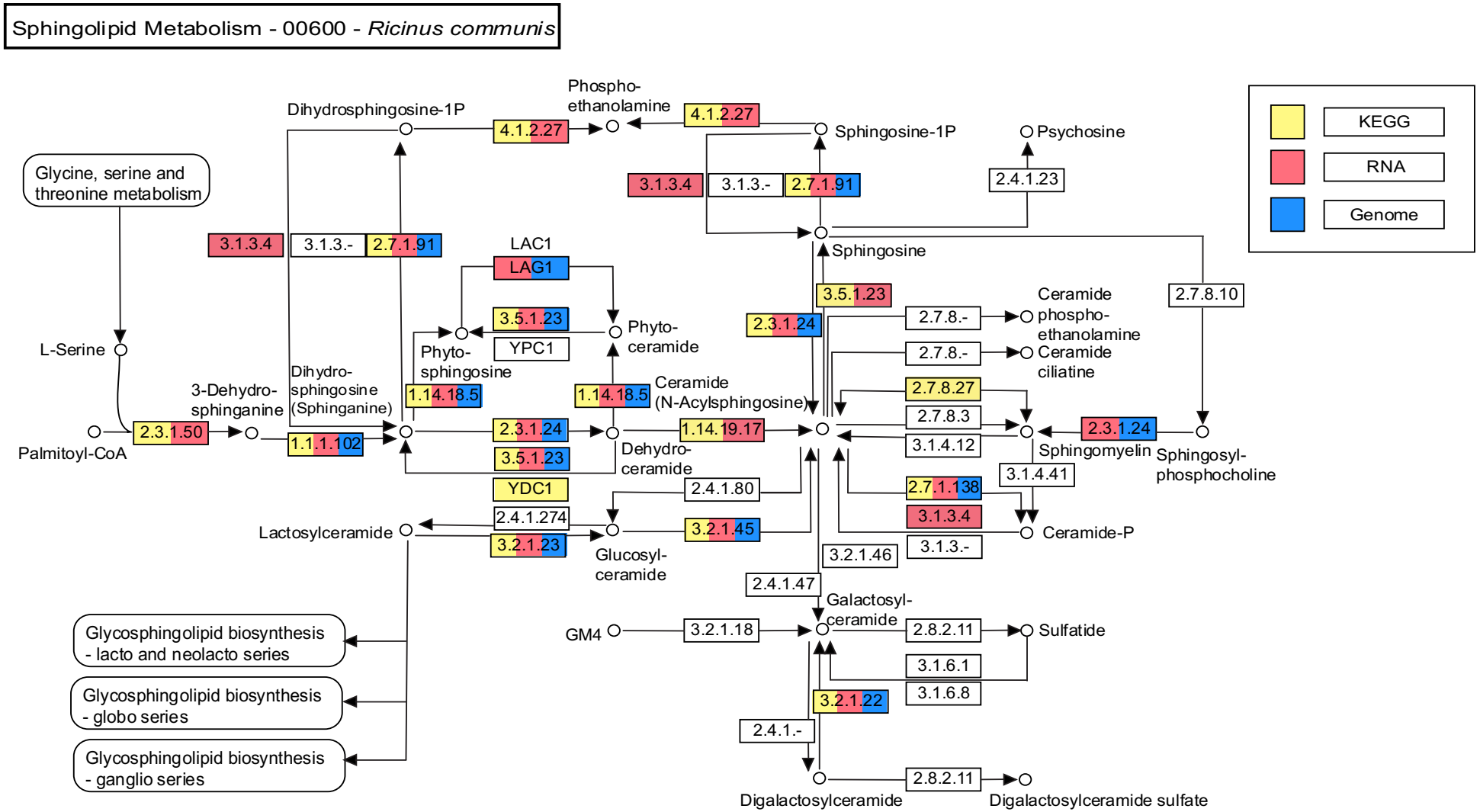
A 32 - Via de metabolismo de ácido ɑ-linolenico indicando as enzimas de *R. communis* presentes nos dados do KEGG, transcritoma e/ou genoma



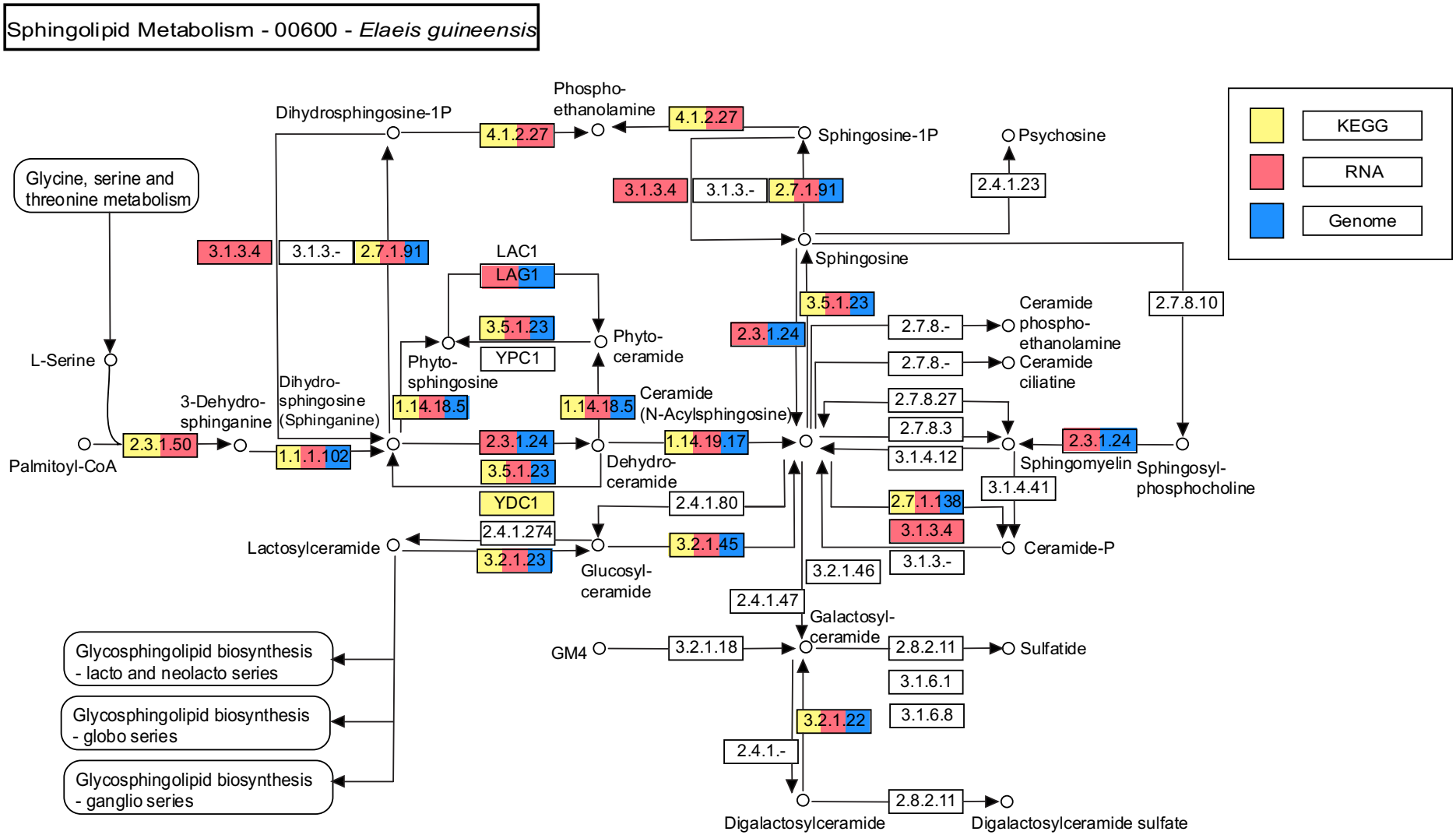
A 33 - Via de metabolismo de ácido ά-linolenico indicando as enzimas de *E. guineensis* presentes nos dados do KEGG, transcritoma e/ou genoma.



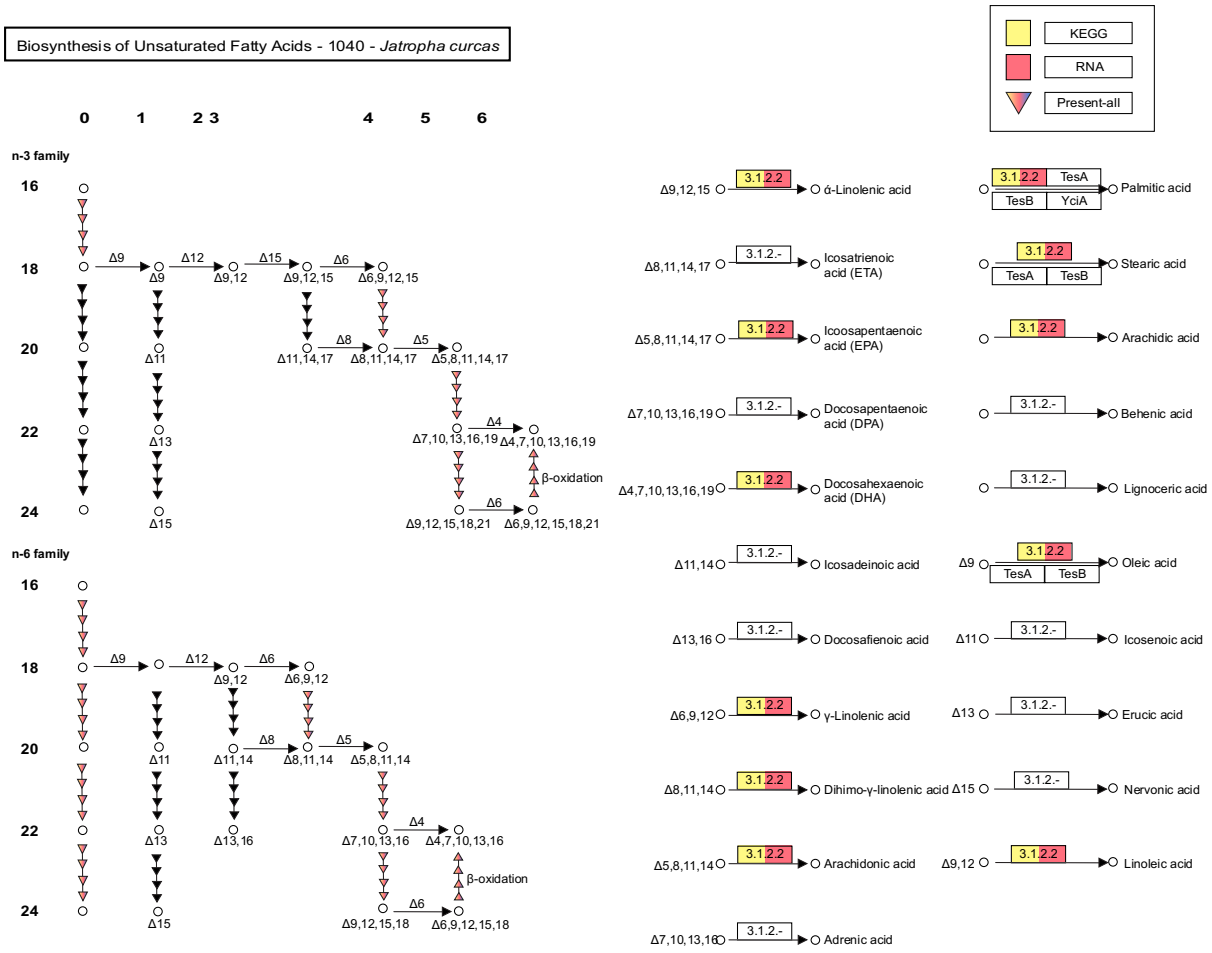
A 34 - Via de metabolismo de esfingolipídio indicando as enzimas de *J. curcas* presentes nos dados do KEGG, transcritoma e/ou genoma.



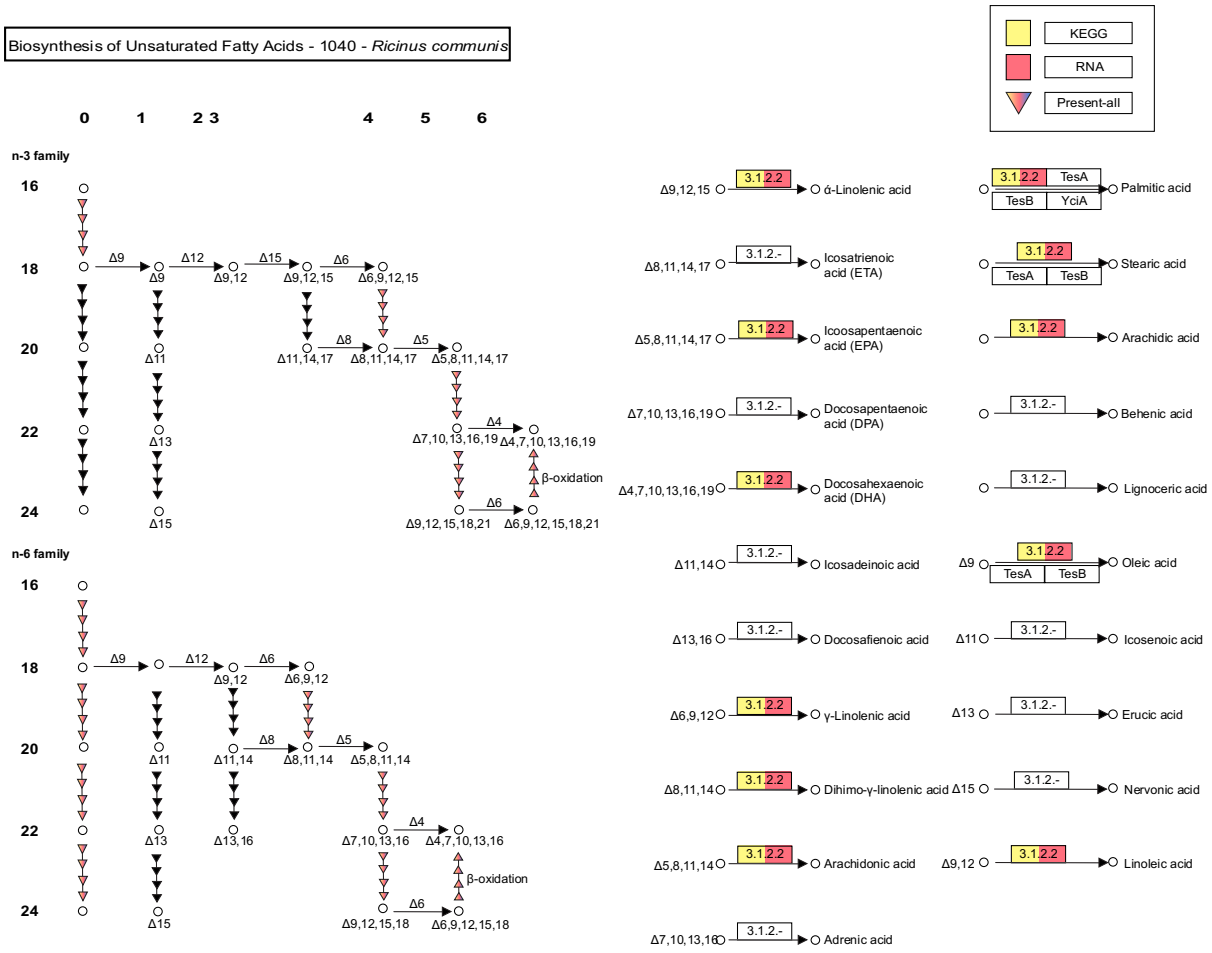
A 35 - Via de metabolismo de esfingolipídio indicando as enzimas de *R. communis* presentes nos dados do KEGG, transcritoma e/ou genoma.



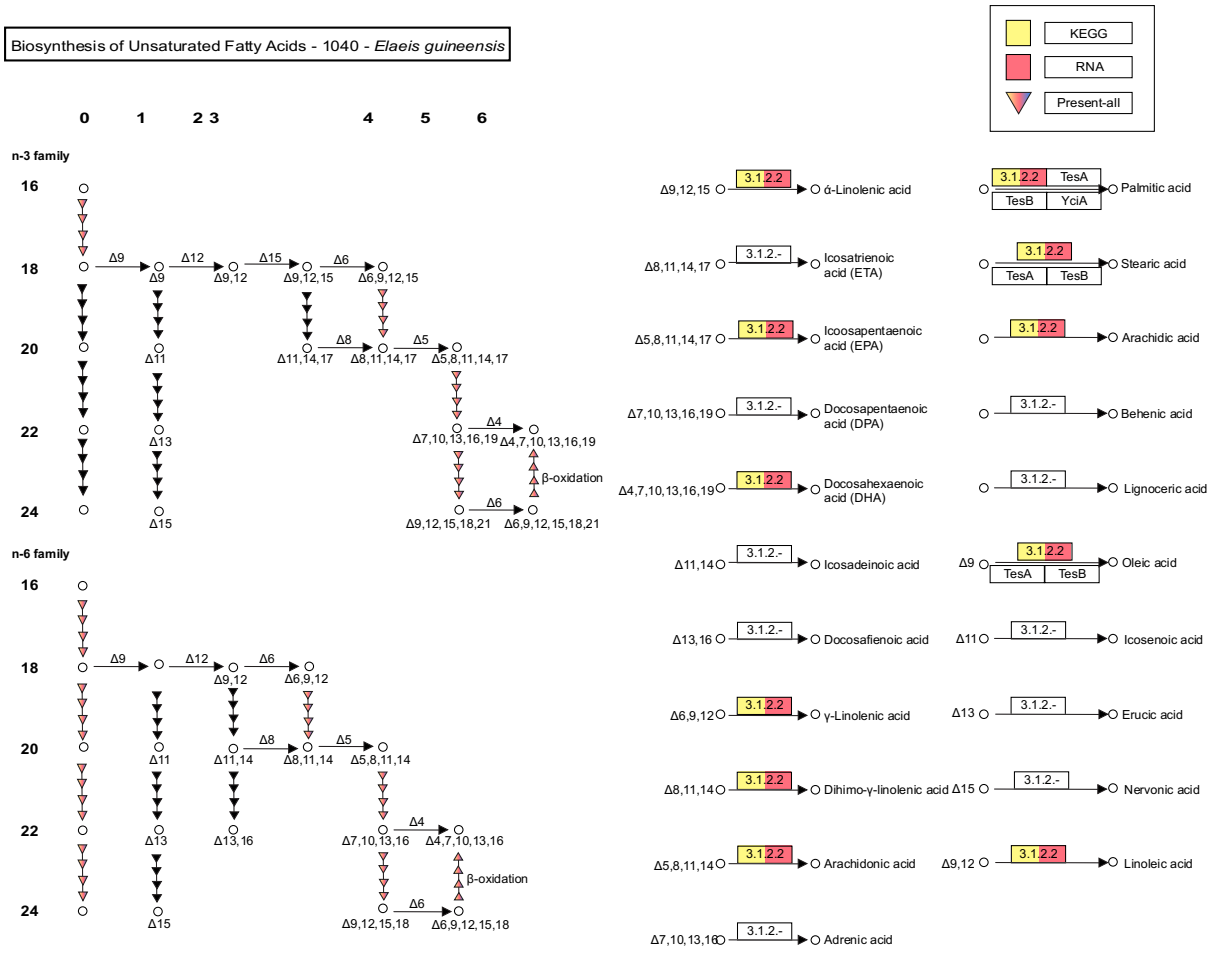
A 36 - Via de metabolismo de esfingolipídio indicando as enzimas de *E. guineensis* presentes nos dados do KEGG, transcritoma e/ou genoma



A 37 - Via de biossíntese de ácidos graxos insaturados indicando as enzimas de *J. curcas* presentes nos dados do KEGG, transcritoma e/ou genoma



A 38 - Via de biossíntese de ácidos graxos insaturados indicando as enzimas de *R. communis* presentes nos dados do KEGG, transcritoma e/ou genoma.



A 39 - Via de biossíntese de ácidos graxos insaturados indicando as enzimas de *E. guineensis* presentes nos dados do KEGG, transcritoma e/ou genoma

Comandos e parâmetros utilizados

Pré-processamento

Os INPUTs foram os arquivos:

- *E. guineensis*: elaeis_guinnensis_1.fastqc e elaeis_guinnensis_2.fastqc
- *J. curcas*: jatropha_sp_1.fastqc e jatropha_sp_2.fastqc
- *R. communis*: ricinus_communis_1.fastqc e ricinus_communis_2.fastqc

FastQC

O FASTQC foi feito com parâmetros *default* utilizando como *input* os arquivos gerados pela etapa de retirada dos adaptadores e também foram incluídos os arquivos antes da retirada dos adaptadores para efeito de comparação, os outputs gerados são arquivos HTML mostrando a qualidade das sequencias presente em cada arquivo.

Cutadapt

Na utilização do Cutadapt foi utilizado a linha de comando abaixo:

/cutadapt-1.9/cutadapt -a AGATCGGAAGAG -A AGATCGGAAGAG -o INPUT_1_clipped_cutadapt.fastq -p INPUT_2_clipped_cutadapt.fastq -f fastq --minimum-length=16 INPUT_1_clipped.fastq INPUT_2_clipped.fastq

-a/-A adaptador a ser encontrado e retirado

-o output

-p arquivo da sequência 5’

-f formato do arquivo

--minimum-length tamanho mínimo da sequência após remoção do adaptador

Fast-mcf

Para a retirada dos adaptadores foi utilizado primeiro o *fastq-mcf* com os parâmetros a seguir: /fastq-mcf -P 33 -w 4 -q 30 adap.fa INPUT_1.fastq INPUT_2.fastq -o INPUT_1_clipped.fastq -o INPUT_2_clipped.fastq

-P Phred-scale

-w o tamanho da janela de qualidade

-q número de corte para a remoção de bases

adap.fa arquivo contendo os adaptadores

INPUT_1 arquivo das sequências 3’

INPUT_2 arquivo das sequências 5’

-o output

Montagem do transcritoma

Os INPUTs foram os arquivos:

- *E. guineensis*: elaeis_guinnensis_1_clipped.fastq e elaeis_guinnensis_2_clipped.fastq
- *J. curcas*: jatropha_sp_1_clipped.fastq e jatropha_sp_2_clipped.fastq
- *R. communis*: ricinus_communis_1_clipped.fastq e ricinus_communis_2_clipped.fastq

Velveth/Oases

velveth [PATH para a pasta do output] [nº k-mer] -fastq -shortPaired [PATH INPUT_1] [PATH INPUT_2]

HISAT2

hisat2 -x [PATH hisat index] -1 [PATH INPUT_1] -2 [PATH INPUT_2]

-x index é feito com base no genoma de referência

-1 INPUT fita *forward*

-2 INPUT fita *reverse*

SPAdes

python SPADES/SPAdes-3.11.1-Linux/bin/spades.py -o [PATH output] --rna -1 [PATH INPUT_1] -2 [PATH INPUT_2] -k [nº k-mer]

Trinity

/trinityrnaseq-Trinity-v2.5.1/Trinity --seqType fq --max_memory 80G --full_cleanup --left [PATH INPUT_1] --right [PATH INPUT_2] --CPU 10 --output [PATH output]

SOAP

SOAP (versão 1.03): ./SOAPdenovo-Trans-31mer all -s [PATH arquivo de configuração] -o [PATH output] -p 10 -M 1

-s o arquivo de configuração é necessário pois é nele que irá se encontrar os arquivos de INPUT e informações adicionais para a montagem ser conduzida

STAR

STAR (versão 2.4.0.1):

Geração de index para o STAR

STAR --runThreadN 10 --runMode genomeGenerate --genomeDir [PATH diretório do output] --genomeFastaFiles [PATH fasta do genoma referência]

Mapeamento

STAR --runThreadN 10 --genomeDir [PATH diretório do output] --sjdbGTFfile [PATH do arquivo GTF] --sjdbOverhang 100 --readFilesIn [PATH INPUT_1] [PATH INPUT_2] --outSAMtype BAM SortedByCoordinate Unsorted --outReadsUnmapped Fastx --outFileNamePrefix [prefixo do output] --quantMode TranscriptomeSAM

Montagem

cufflinks -p 10 --library-type fr-secondstrand -o [output]_cuff -g [PATH do arquivo gtf de referência] [PATH do arquivo BAM gerado no mapeamento]

gffread -w transcripts.fa -g /path/to/genome.fa transcripts.gtf

Polimento dos transcritomas

Evidential gene

Os INPUTs foram os arquivos obtidos a partir da união dos arquivos fastas gerados por cada montador para cada espécie.

evigene/scripts/rnaseq/trformat.pl [PATH dos fastas unidos] [PATH dos fastas unidos]_tratados

Este script é utilizado para regularizar os identificadores das sequências, assegurando que não há duplicatas.

evigene/scripts/prot/tr2aacds2.pl [PATH dos fastas unidos]_tratados [PATH dos fastas unidos]_tratados_okay

Este script por sua vez filtra os fastas com base em sua qualidade.

Analises de completude

BUSCO

Os INPUTs foram os arquivos obtidos dos montadores, após o polimento do Evidential gene e do genoma de referência de cada espécie.

BUSCO_V3/scripts/run_BUSCO.py -m tran -o [PATH para o output] -i [PATH do input] -l /BUSCO_V3/datasets/embryophyta_odb9 -c 20

-m modo da corrida, podendo ser alterado para gen (DNA), tran (RNA) ou prot (AA)

-o output

-i input

-l o banco de dados para a espécie, existem 5 na atual versão

Confecção do banco de dados

Download dos arquivos contendo as sequencias de aminoácidos

```
echo

echo 'extraindo lista para dbget'

echo

echo

for i in `ls -1 *.txt|cut -d'.' -f1`; do awk '{print $4}' $i.txt |grep dbget|awk -F"?" '{print $2}'|sed s/\"//g|sed s/+/\\n/g|sort -u |grep ':' >list_$i;done


echo

echo

echo 'Trazendo arquivos com sequencias'

echo
```

for i **in** `ls -1 list*`; **do** echo $i;**for** j **in** `cat $i`; **do** echo $j;curl -s http://www.genome.jp/dbget-bin/www_bget?-f+-n+a+$j >> $i.aa.raw;**done**;**done**

echo

echo

echo 'convertendo para fasta'

echo

for s **in** `ls -1 *aa.raw` ; **do** echo $s; ../scripts/extrai_fasta_keggDB.pl $s > $s.fasta;**done**

Extraindo as sequencias de aminoácidos

#!/usr/bin/perl

#

 # Le arquivo de sequencia recuperado do keggDB e salva em formato fasta

 # Roberto Togawa - fev/2018

#

Existe argumentos?

if($#ARGV < 0){

print "Digitar: extrai_fasta_keggDB.pl <arquivo de sequencia>\n";**exit**

}

Testa primeiro argumento - Arquivo para separacao

$file = $ARGV[0];

if (! -e $file) {

print "\n\nArquivo '$file' inexistente\n\n";

```
  exit;
}

@arq = `cat $file`;


###### Loop de leitura #######
for ($i=0; $i<=$#arq; $i++) {
  $s = $arq[$i];
  chop $s;
  if ($s =~ "<pre>") {
          $seq_name=$arq[$i+1];
          ($inicio,$desc) = split(/\>\>/,$seq_name);
          print ">$desc";
          for ($k=$i+2; $k<=$#arq;$k++) {
                  $p = $arq[$k];
                  chop $p;
                  if ($p =~ "</pre>") {
                          last;
                  }
                  print "$p\n";
        }
  }

}
```

Árvore filogenética

Orthofinder

/OrthoFinder/OrthoFinder-2.2.3/orthofinder -f [PATH da pasta] -t 10 -S diamond

-f local onde se encontram os arquivos

-S utilização do Diamond ao invés do BLAST para a comparação das sequências

Anotação dos transcritos

BLAST

blastx -query [PATH input] -out [PATH output] -num_threads 10 -evalue 1e-20 -db [PATH banco de dados] -outfmt "6 qseqid sseqid pident qlen slen length mismatch gapopen evalue bitscore"

-query arquivo de saída do evidential gene

-out output do BLAST

-num_threads número de processadores a serem utilizados

-evalue valor utilizado como *cut off* para filtrar alinhamentos mais precisos

-db banco de dados criado contendo as sequências das vias de metabolismo de ácidos graxos

-outfmtformato do output gerado, 6 = tabular

Com o resultado do BLAST foi executado o comando abaixo para cada espécie para obtermos apenas os *EC number* encontrados no alinhamento.

```
#!usr/bin/bash


##USAGE: $1 input file .tab/.txt; $2 output files .txt
##Inicialmente separar a coluna do arquivo que vc quer procurar

cut –f2 $1  > cut_$2
```

```
    ##Grep com o arquivo contendo os ECs

    grep                        -f                        cut_$2
/st70/acgt/projetos/metabolomica_elibio/annotation/Blastx_data
_vs_viaskegg/ECs/EC_IDS.txt > grep_$2

    ##Sort para obtermos ECs unicos apenas
    sort -u -k2 grep_$2 > sorted_$2

    ##Cut para obter apenas os ECs para adc ao link e plotar
as vias no KEGG
    cut -f2 sorted_$2 | sed 's/ //g' > cut_sorted_$2
    sed ':a;N;$!ba;s/\n/+/g' cut_sorted_$2 > plot_sorted_$2
```

Anotação dos transcritos às vias metabolismo de ácidos graxos

Com os *EC numbers* encontrados para cada espécie foi possível utilizá-los para anotar as vias no KEGG.

Jatropha curcas:

```
    #!usr/bin/bash

    for     i     in     `cat     list.txt`;     do     curl
https://www.genome.jp/kegg-
bin/show_pathway?ec$i+1.1.1.1+1.1.1.100+1.1.1.102+1.1.1.170+1.
1.1.184+1.11.1.9+1.1.1.2+1.1.1.21+1.1.1.284+1.1.1.62+1.1.1.8+1
.13.11.12+1.13.11.58+1.14.13.70+1.14.14.1+1.14.14.17+1.14.18.5
+1.14.19.17+1.14.19.2+1.14.19.20+1.14.19.41+1.1.5.3+1.2.1.3+1.
2.1.31+1.3.1.21+1.3.1.38+1.3.1.42+1.3.1.70+1.3.1.72+1.3.1.9+1.
3.1.93+1.3.3.6+1.3.8.7+2.1.1.103+2.1.1.141+2.1.1.143+2.1.1.41+
2.1.1.71+2.3.1.15+2.3.1.158+2.3.1.16+2.3.1.179+2.3.1.180+2.3.1
.195+2.3.1.199+2.3.1.20+2.3.1.22+2.3.1.23+2.3.1.24+2.3.1.39+2.
3.1.50+2.3.1.51+2.3.1.9+2.3.2.2+2.4.1.184+2.4.1.241+2.4.1.46+2
.5.1.21+2.7.1.107+2.7.1.138+2.7.1.28+2.7.1.30+2.7.1.31+2.7.1.3
2+2.7.1.82+2.7.1.91+2.7.7.14+2.7.7.15+2.7.7.41+2.7.8.1+2.7.8.1
1+2.7.8.29+2.7.8.41+2.7.8.5+3.1.1.13+3.1.1.23+3.1.1.3+3.1.1.32
+3.1.1.5+3.1.2.14+3.1.2.2+3.1.2.22+3.13.1.1+3.1.3.27+3.1.3.4+3
.1.3.75+3.1.3.81+3.1.4.3+3.1.4.4+3.1.4.46+3.2.1.22+3.2.1.23+3.
2.1.45+3.3.2.6+3.5.1.23+3.6.1.13+4.1.1.65+4.1.2.27+4.2.1.134+4
.2.1.17+4.2.1.59+4.2.1.92+5.3.3.5+5.3.3.8+5.3.99.3+5.3.99.6+5.
4.99.8+5.5.1.9+6.2.1.3+6.4.1.2 | grep .png | grep input >>
jatropha.url; done

    sed -E 's/^.{51}//' jatropha.url | sed -E 's/.$//' >
jatropha_url_png.txt
```

```
for u in `cat jatropha_url_png.txt`; do wget https://www.genome.jp/tmp$u; done
```

Ricinus communis:

```
#!usr/bin/bash

for i in `cat list.txt`; do curl https://www.genome.jp/kegg-bin/show_pathway?ec$i+1.1.1.1+1.1.1.100+1.1.1.102+1.1.1.170+1.1.1.184+1.11.1.9+1.1.1.2+1.1.1.21+1.1.1.284+1.1.1.62+1.1.1.8+1.13.11.12+1.13.11.58+1.14.13.70+1.14.14.1+1.14.14.17+1.14.18.5+1.14.19.17+1.14.19.2+1.14.19.20+1.14.19.41+1.1.5.3+1.2.1.3+1.2.1.31+1.3.1.21+1.3.1.38+1.3.1.42+1.3.1.70+1.3.1.72+1.3.1.9+1.3.1.93+1.3.3.6+1.3.8.7+2.1.1.103+2.1.1.143+2.1.1.41+2.1.1.71+2.3.1.15+2.3.1.158+2.3.1.16+2.3.1.179+2.3.1.180+2.3.1.195+2.3.1.199+2.3.1.20+2.3.1.22+2.3.1.23+2.3.1.24+2.3.1.39+2.3.1.50+2.3.1.51+2.3.1.9+2.3.2.2+2.4.1.184+2.4.1.241+2.4.1.46+2.5.1.21+2.7.1.107+2.7.1.138+2.7.1.28+2.7.1.30+2.7.1.31+2.7.1.32+2.7.1.82+2.7.1.91+2.7.7.14+2.7.7.15+2.7.7.41+2.7.8.1+2.7.8.11+2.7.8.29+2.7.8.41+2.7.8.5+3.1.1.13+3.1.1.23+3.1.1.3+3.1.1.32+3.1.1.4+3.1.1.5+3.1.2.14+3.1.2.2+3.1.2.22+3.13.1.1+3.1.3.27+3.1.3.4+3.1.3.75+3.1.3.81+3.1.4.3+3.1.4.4+3.1.4.46+3.2.1.22+3.2.1.23+3.2.1.45+3.3.2.6+3.5.1.23+3.6.1.13+4.1.1.65+4.1.2.27+4.2.1.134+4.2.1.17+4.2.1.59+4.2.1.92+5.3.3.5+5.3.3.8+5.3.99.3+5.3.99.6+5.4.99.8+5.5.1.9+6.2.1.3+6.4.1.2 | grep .png | grep input >> ricinus.url; done

sed –E 's/^.{51}//' ricinus.url | sed –E 's/.$//' > ricinus_url_png.txt

for u in `cat ricinus_url_png.txt`; do wget https://www.genome.jp/tmp$u; done
```

***Elaeis guineensis*:**

```
#!usr/bin/bash

for i in `cat list.txt`; do curl https://www.genome.jp/kegg-bin/show_pathway?ec$i+1.1.1.1+1.1.1.100+1.1.1.102+1.1.1.170+1.1.1.184+1.11.1.9+1.1.1.2+1.1.1.21+1.1.1.284+1.1.1.62+1.1.1.8+1.13.11.12+1.13.11.58+1.14.13.70+1.14.14.1+1.14.14.17+1.14.18.5+1.14.19.17+1.14.19.2+1.14.19.20+1.14.19.41+1.1.5.3+1.2.1.3+1.2.1.31+1.3.1.21+1.3.1.38+1.3.1.42+1.3.1.70+1.3.1.72+1.3.1.9+1.3.1.93+1.3.3.6+1.3.8.7+2.1.1.103+2.1.1.141+2.1.1.143+2.1.1.41+2.1.1.71+2.3.1.15+2.3.1.158+2.3.1.16+2.3.1.179+2.3.1.180+2.3.1.195+2.3.1.199+2.3.1.20+2.3.1.22+2.3.1.23+2.3.1.24+2.3.1.39+2.3.1.50+2.3.1.51+2.3.1.9+2.3.2.2+2.4.1.184+2.4.1.241+2.4.1.46+2.5.1.21+2.7.1.107+2.7.1.138+2.7.1.28+2.7.1.30+2.7.1.31+2.7.1.3
```

```
2+2.7.1.82+2.7.1.91+2.7.7.14+2.7.7.15+2.7.7.41+2.7.8.1+2.7.8.1
1+2.7.8.29+2.7.8.41+2.7.8.5+3.1.1.13+3.1.1.23+3.1.1.3+3.1.1.32
+3.1.1.4+3.1.1.5+3.1.2.14+3.1.2.2+3.1.2.22+3.13.1.1+3.1.3.4+3.
1.3.75+3.1.3.81+3.1.4.3+3.1.4.4+3.1.4.46+3.2.1.22+3.2.1.23+3.2
.1.45+3.3.2.6+3.5.1.23+3.6.1.13+4.1.1.65+4.1.2.27+4.2.1.134+4.
2.1.17+4.2.1.59+4.2.1.92+5.3.3.5+5.3.3.8+5.3.99.3+5.3.99.6+5.4
.99.8+6.2.1.3+6.4.1.2 | grep .png | grep input >> elaeis.url;
done

    sed -E 's/^.{51}//' elaeis.url | sed -E 's/.$//' >
elaeis_url_png.txt

    for u in `cat elaeis_url_png.txt`; do wget
https://www.genome.jp/tmp$u; done
```