ISSN 0102-1110
Março/ 2021

BOLETIM DE PESQUISA E DESENVOLVIMENTO

369

Combinação de abordagens de análises *de novo* e guiadas pelo genoma para explorar dados de RNA-Seq de sementes oleaginosas para anotação de vias de ácidos graxos

Embrapa

ISSN 0102-0110
Março/2021

Empresa Brasileira de Pesquisa Agropecuária
Embrapa Recursos Genéticos e Biotecnologia
Ministério da Agricultura, Pecuária e Abastecimento

BOLETIM DE PESQUISA E DESENVOLVIMENTO 369

Combinação de abordagens de análises *de novo* e guiadas pelo genoma para explorar dados de RNA-Seq de sementes oleaginosas para anotação de vias de ácidos graxos

Vinícius Nattan Silva Lemos
Roberto Coiti Togawa
Marcos Mota do Carmo Costa
Elíbio Leopoldo Rech
Priscila Grynberg

Embrapa Recursos Genéticos e Biotecnologia
Brasília, DF
2021

Exemplares desta publicação podem ser adquiridos na:

Embrapa Recursos Genéticos e Biotecnologia
Parque Estação Biológica
PqEB, Av. W5 Norte (final)
70970-717 , Brasília, DF
Fone: +55 (61) 3448-4700
Fax: +55 (61) 3340-3624
www.embrapa.br
www.embrapa.br/fale-conosco/sac

Supervisão editorial
Ana Flávia do N. Dias Côrtes

Revisão de texto
Priscila Grynberg

Normalização bibliográfica
Ana Flávia do N. Dias Côrtes - (CRB-1999)

Tratamento das ilustrações
Adilson Werneck

Projeto gráfico da coleção
Carlos Eduardo Felice Barbeiro

Editoração eletrônica
Adilson Werneck

Foto da capa
Zineb Benchekchou - BME Embrapa

1ª edição
1ª impressão (ano): tiragem

Dados Internacionais de Catalogação na Publicação (CIP)
Embrapa Recursos Genéticos e Biotecnologia

Combinação de abordagens de análises *de novo* e guiadas pelo genoma para explorar dados de RNA-Seq de sementes oleaginosas para anotação de vias de ácidos graxos / Vinicius Nattan Silva Lemos [et al.]... – Brasília, DF: Embrapa Recursos Genéticos e Biotecnologia, 2021.

35 p. - (Boletim de Pesquisa e Desenvolvimento / Embrapa Recursos Genéticos e Biotecnologia, 369).

ISSN: 0102-0110

Sistema requerido: Adobe Acrobat Reader

Modo de Acesso: World Wide Web

1. *Elaeis guineenses*. 2. *Jatropha curcas*. 3. *Ricinus communis*. 4. Dendê I. Embrapa Recursos Genéticos e Biotecnologia. IX. Série.

Ana Flávia do N. Dias Côrtes(CRB-1999)

Sumário

Combinação de abordagens de análises *de novo* e guiadas pelo genoma para explorar dados de RNA-Seq de sementes oleaginosas para anotação de vias de ácidos graxos

Vinícius Nattan Silva Lemos[1]
Roberto Coiti Togawa[2]
Marcos Mota do Carmo Costa[3]
Elíbio Leopoldo Rech[4]
Priscila Grynberg[5]

Resumo – *Elaeis guineensis* (dendê), *Jatropha curcas* (pinhão-manso) e *Ricinus communis* (mamona) produzem ácidos graxos que podem ser utilizados como fonte renovável na matriz energética, apresentando um grande potencial biotecnológico para as indústrias da área. 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. Os transcritomas foram montados utilizando abordagens *de novo* e guiados pelo genoma e filtrados com o programa *Evidential Gene* para a obtenção de resultados robustos. Uma base de dados contendo 527 sequências de 170 códigos de enzimas únicos de 12 vias de metabolismo de ácidos graxos foi gerada. Um total de 152, 156 e 150 transcritos de *E. guineensis, J. curcas* e *R. communis* referentes às proteínas pertencentes à 12 vias de metabolismo de ácidos graxos foram encontradas respectivamente, sendo 135 em comum. Os resultados ainda sugerem que há proteínas com expressão tecido-específicos pois 23, 6 e 11

[1] Biotecnólogo, mestre em Tecnologia Química e Biológica, bolsista na Embrapa Recursos Genéticos e Biotecnologia, Brasília, DF.

[2] Bioinformata, doutor em Bioinformática Estrutural, analista da Embrapa Recursos Genéticos e Biotecnologia, Brasília, DF.

[3] Matemático, doutor em Ciência da Computação, analista da Embrapa Recursos Genéticos e Biotecnologia, Brasília, DF.

[4] Agrônomo, doutor em Genetic Manipulation, pesquisador da Embrapa Recursos Genéticos e Biotecnologia, Brasília, DF.

[5] Bióloga, doutora em Bioinformática, pesquisadora da Embrapa Recursos Genéticos e Biotecnologia, Brasília, DF.

proteínas de interesse de *E. guineensis, J. curcas* e *R. communis* não foram encontradas nos transcritomas, respectivamente. A estratégia desenvolvida neste trabalho mostrou ser eficiente para a mineração de dados de interesse em dados de RNA-Seq com baixa cobertura pois o uso de diferentes programas de montagem potencializou a identificação das proteínas.

Termos para indexação: *Elaeis guineensis*, *Jatropha curcas*, *Ricinus communis*, dendê, pinhão-manso, mamona, RNA-Seq, vias metabólicas, ácidos graxos, montagem de transcritoma.

Analysis of the de novo and genome-guided approaches in combination to explore RNA-Seq data from oilseeds for fatty acid pathways annotation

Abstract – *Elaeis guineensis* (palm), *Jatropha curcas* (barbados nut) and *Ricinus communis* (castor) are fatty acids producers. They present a major biotechnological potential for biofuel industry because they can be used as a renewable source in the energy matrix. The aim of this work was the identification of fatty acids related transcripts and their presence in metabolic pathways. Total RNA from seeds was extracted and sequenced. The transcriptomes were assembled using de novo and genome guided approaches and filtered with the Evidential Gene program to obtain robust results. A database containing 527 sequences of 170 unique enzyme codes of 12 fatty acid metabolism pathways was generated. A total of 152, 156 and 150 transcripts of *E. guineensis*, J. curcas and *R. communis* of 12 fatty acid metabolism pathways were found respectively, and 135 were in common. These results suggest the presence of tissue-specific proteins because 23, 6 and 11 proteins of interest from *E. guineensis*, J. curcas and *R. communis* were not found in the transcriptomes, respectively. The strategy developed in this work proved to be efficient for low-coverage RNA-Seq data mining because the use of different assembly programs potentiated the identification of proteins.

Index terms: Elaeis guineensis, Jatropha curcas, Ricinus communis, palm oil, jatropha, castor beans, RNA-Seq, metabolic pathways, fatty acids, transcriptome assembly.

Introdução

Plantas oleaginosas são aquelas cuja característica principal é a presença de altas concentrações de ácidos graxos em sua composição. Elas exercem um papel de destaque na agricultura mundial em função do aumento gradativo da demanda por óleos vegetais, tanto para o consumo individual (alimentos, cosméticos, medicamentos) quanto para aplicações industriais, como a produção de biocombustíveis (Samarth et al., 2015).

No ano de 2018, a produção anual de culturas oleaginosas, segundo dados oficiais, estimados ou inferidos da FAO, foi de 1.038 milhões de toneladas (FAO, 2020). O Brasil produziu 127,5 milhões de toneladas, ou seja, 12,3% da produção mundial. Destes, 117 milhões de toneladas correspondem à produção de soja (FAO, 2020). A matéria prima para a produção do óleo vegetal é obtida, principalmente, a partir de monoculturas uma vez que essas têm alta produção e permitem a produção em larga escala.

A produção de biodiesel no Brasil vem crescendo a cada ano, com quase 6 milhões de m^3 produzidos em 2019 (Agência..., 2019). Os estados das regiões centro-oeste e sul somam mais de 80% da produção nacional de acordo com a Associação Brasileira das Indústrias de Óleos Vegetais – ABIOVE. A variedade de plantas utilizadas para a produção de biocombustíveis pela indústria é relativamente baixa, sendo restrita principalmente à soja nos Estados Unidos, Argentina e Brasil, à colza na Europa e óleo de palma na Indonésia e Malásia (Junginger, Goh e Faaij, 2014). Além destes, podemos ainda citar girassol e algodão. No Brasil, a soja é responsável por aproximadamente 70% do biodiesel fabricado, seguido pelo uso de gordura bovina (Agência..., 2019).

A mamona (*Ricinus communis*), da família Euphorbiaceae é uma planta perene, produzida principalmente no Brasil, China e Índia, países em desenvolvimento e com vulnerabilidade econômica. Juntos, esses países são responsáveis por 95% da produção mundial (FAO, 2020). Cerca de 45% da composição da semente é ácido graxo (Forero, 2005) sendo que aproximadamente 93% do óleo é formado por ácido ricinoleico, que possibilita a produção de derivados de pureza elevada. O dendê (*Elaeis guineensis*), pertencente à família Arecaceae produz o óleo de palma, ou óleo de dendê. Originária do continente africano esta planta completa seu ciclo em um ano, passando pela germinação e florescimento neste período (Sambanthamurthi

et al., 2009). O óleo de palma é composto principalmente por ácido palmítico, esteárico e mirístico (saturados) e ácido oleico e linoleico (insaturados), totalizando aproximadamente 95% (revisado por Mba, Dumont e Ngadi, 2015). O pinhão-manso, outra Euphorbiaceae, é uma pequena árvore semi-perene originária do México. É bastante promissora para ser utilizada na 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., 2008; Fuentes et al., 2018). Alguns fatores são limitantes para o cultivo em larga escala desta planta, como, por exemplo, alto custo para o plantio, falta de conhecimentos avançados sobre a biologia, incidência e manejo de pragas e doenças relacionadas, e ausência de programas governamentais de incentivo (Edrisi et al., 2015).

A disponibilidade de dados genômicos e transcritômicos de inúmeras espécies de plantas tem contribuído fortemente para o entendimento de vias metabólicas de interesse, assim como o descobrimento de novos produtos gênicos de vias metabólicas de interesse. Os esforços têm se concentrado em estudos transcritômicos, pois são mais acessíveis financeiramente e menos desafiantes bioinformaticamente (Owen et al., 2017). Para exemplificar, há ao menos 1.000 entradas de dados entre sequenciamentos de DNA e RNA no SRA (Sequence read archive) do NCBI para cada espécie citada acima, mas quando se trata de genomas montados, a realidade é outra. O dendê é o que possui a maior quantidade de dados genômicos disponíveis tendo tido o primeiro genoma produzido em 2013 e montado em nível de cromossomos com tamanho de 1.5Gb, 41% de conteúdo GC, 40.349 scaffolds e 43.541 proteínas preditas (Singh et al., 2013). Recentemente uma nova montagem em nível de scaffolds foi disponibilizada, com 1.2 Gb, 40,46% de conteúdo GC, e 165 scaffolds (Ong et al., 2020). Os demais genomas desta espécie, publicados em 2016 e 2017 estão em níveis de scaffolds e contigs respectivamente, com tamanhos muito menores que o esperado. Já o pinhão-manso possui quatro montagens, todas em nível de scaffolds, com tamanhos entre 266 e 319Mb, cerca de 35% de conteúdo GC e 27.680 proteínas preditas no último draft, onde foi utilizado a combinação de sequências produzidos por equipamentos Illumina e PacBio (Ha et al., 2019). A mamona possui um genoma não montado com aproximadamente 400 Mb divididos em 25.878 scaffolds e 31.221 transcritos preditos (Chan et al., 2010).

O sequenciamento em larga escala de moléculas de RNA (RNA-Seq) é uma técnica valiosa, pois fornece dados de expressão gênica global, permitindo a compreensão de mecanismos moleculares versáteis e a resposta à diversas questões biológicas. Para organismos não-modelo, na indisponibilidade de um genoma de qualidade, ou seja, com quantidade de dados insuficientes e muito fragmentado, esta técnica permite a montagem e quantificação do transcritoma (Conesa et al., 2016; Wang, Gerstein e Snyder, 2009).

As sequências, que geralmente são curtas caso a origem seja dos equipamentos Illumina, precisam ser ordenados de forma a gerar sequências dos transcritos completas. Na ausência de um genoma de referência, utiliza-se os montadores de novo, ou seja, que utilizam apenas as sequências obtidas para obter os transcritos e gerar um transcritoma. No entanto, quando há genomas disponíveis, as sequências são primeiramente mapeadas para depois gerar os transcritos. Geralmente esses métodos produzem melhores transcritomas. Mesmo com a disponibilidade de genomas, trabalhos relatam os benefícios de combinarem diferentes programas com métodos distintos (Cabau et al., 2017; Holding et al., 2018; Sahraeian et al., 2017) porque os programas possuem eficiências variadas que dependem da origem das amostras, confecção das bibliotecas, equipamentos de sequenciamento e qualidade das sequências (Hölzer e Marz, 2019).

Com o objetivo de identificar as enzimas componentes das vias metabólicas de ácidos graxos no dendê, pinhão-manso e mamona, dados de RNA-Seq de sementes foram analisados associando montadores de novo e baseados em genoma para identificar e anotar genes relacionados à produção, elongação e degradação das cadeias de ácidos graxos, vias que podem ser engenheiradas visando a melhoria de produção de ácidos graxos por métodos de engenharia metabólica e biologia sintética.

Material e Métodos

Os comandos de cada programa usados e os scripts *in-house* desenvolvidos neste trabalho estão organizados em um tutorial na página do Laboratório de Bioinformática na plataforma Github no endereço: https://github.com/lbi-cenargen/oil_plants/.

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. O sequenciamento foi realizado em equipamento Illumina HiSeq 2500 protocolo paired-end (Metzker, 2010).

Pré-processamento e montagem dos transcritomas

A qualidade das sequências foi avaliada pelo programa FASTQC com os parâmetros padrão (Andrews, 2014). Os programas cutadapt (versão 1.9) e fast-mcf (versão 1.04) (Aronesty, 2011; Martin, 2011) 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 foram clivados.

Para a montagem dos transcritomas, seis diferentes programas foram testados (Tabela 1). O StringTie, o Trinity e o STAR foram usados para montagens com os genomas de referência e os demais montadores geraram transcritomas de novo. Para cada programa (com exceção do StringTie e STAR), foram testados diferentes k-mers com o objetivo de avaliar a melhor estratégia. Os inputs foram os arquivos de saída do cutadapt (após retirada dos adaptadores). Para a montagem dos transcritos gerados pelo StringTie e STAR, foi necessário rodar o programa cufflinks (Ghosh e Chan, 2016) a partir do arquivo .gtf gerado. Os resultados de cada montador geraram arquivos contendo as sequências dos transcritos gerados (formato fasta).

Tabela 1: Lista de programas usados para a montagem dos transcritos

Protocolo	Assembler	Tamanho do K-mer	Algoritmo	Referência
de novo	Velveth/Oases	17,29,45	Grafo de Bruijn	(Schulz et al., 2012)
	SPAdes	17,29,45	Grafo de Bruijn	(Bankevich et al., 2012)
	SOAP	31	Grafo de String	(Luo et al., 2013)
Guiado pelo genoma	StringTie	*Default*	Grafo de String	(Pertea et al., 2015)
	Trinity	25	Grafo de Bruijn	(Grabherr et al., 2011)
	STAR	*Default*	STAR	(Dobin et al., 2012)

O programa *Evidential Gene* (Gilbert, 2013) 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. O input foi a junção dos arquivos 'fasta' gerados pelos montadores. O programa filtra sequências com bases não identificadas (N), sem códon iniciador, loci duplicados e artefatos de montagem. Os transcritos são considerados de baixa qualidade quando se encaixam em ao menos um desses parâmetros citados acima.

Completude dos transcritomas

O programa BUSCO (versão 3.0) foi usado para avaliar o quão completos os transcritomas gerados estavam (Simão et al., 2015). Este programa busca por ortólogos de cópia única de 1.440 genes considerados essenciais. 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. Utilizamos, neste trabalho, o banco de dados de plantas (embryophyta.db). O resultado do BUSCO valida o processo de montagem conferindo maior credibilidade.

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 (Tabela 2) devido aos seus potenciais biotecnológicos de acordo com a revisão de (Erb, Jones e Bar-Even, 2017) e trabalho de (Zhou et al., 2016). As sequências de aminoácidos de todas as proteínas presentes nas 12 vias metabólicas de cada uma das três espécies deste trabalho foram trazidas do repositório Kegg (Ogata et al., 1999) via Kegg API (*application programming interface*).

Tabela 2: 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 esfingolipídeos	00600
Biossíntese de ácidos graxos insaturados	01040

Estudo das vias de metabolismo de ácidos graxos de interesse

As sequências proteicas de interesse deste estudo foram alinhadas, utilizando-se o programa tblastn, em cada um dos transcritomas gerados neste trabalho: sete com montagens de novo, três com montagens guiadas pelo genoma e uma com o Evidential Gene. A mesma base de dados também foi usada para identificar essas proteínas nos genomas de cada espécie. Para isso, incluímos o parâmetro qcovs, que é a porcentagem de bases da query alinhadas no subject. A ideia central dessa metodologia foi testar a seguinte hipótese: Montagens menos acuradas teriam alinhamentos com porcentagens de bases alinhadas idênticas e porcentagens de bases da query alinhadas no transcrito mais baixas que montagens mais acuradas. Para cada proteína de interesse em cada montagem e nos genomas, os parâmentos pident e qcovs foram multiplicados, gerando um valor único indicativo da qualidade do transcrito. Os valores obtidos foram usados para gerar gráficos dos tipos scatter e boxplot através do programa estatístico e gráfico R (R Core Team, 2019).

Resultados e 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* (Figura 1). 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 estratégia desenvolvida neste trabalho mostrou-se ser eficiente em gerar transcritomas de qualidade, fundamentais para a identificação das proteínas das vias metabólicas de ácidos graxos de interesse.

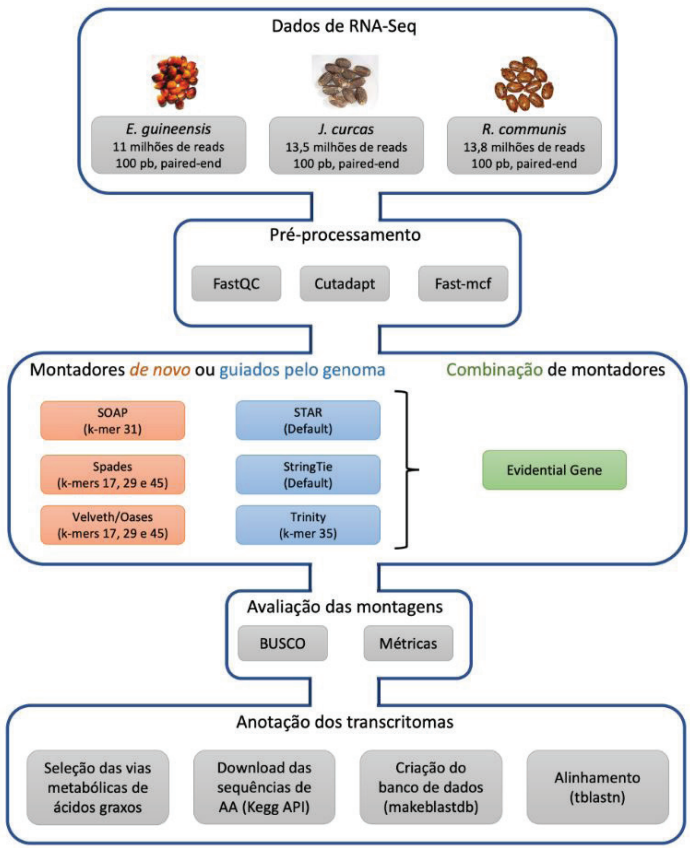

Figura 1: Fluxograma da metodologia desenvolvida e aplicada neste trabalho. Dados de sementes de três espécies foram pré-processados. Seis montadores foram usados para gerar transcritomas e os resultados foram polidos pelo Evidental Gene. A qualidade das montagens foram avaliadas pelo BUSCO e por métricas específicas. Os transcritos foram anotados com foco em vias metabólicas de ácidos graxos.

Sequenciamento e pré-processamento das sequências

Os sequenciamentos em larga escala das sementes produziram 22,6, 27,5 e 27,9 milhões de sequências para *E. guineensis*, *J. curcas* e *R. communis* respectivamente. O conteúdo GC variou entre 43% e 48% (Tabela 3). As sequências brutas e pré-processadas apresentaram valores médios de qualidade (PHRED) de 36, 36 e 37 e 37, 37 e 38 para *E. guineensis*, *J. curcas* e *R. communis*, respectivamente. Tais valores indicam que as bases apresentaram alta qualidade de sequenciamento. Ainda assim, o uso de pré-processadores como o fastq-mcf e cutadapt 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. Os dados também mostram que as coberturas dos sequenciamentos foram adequadas para *J. curcas* (8,06X) e *R. communis* (7,91X). Porém *E. guineensis* teve uma cobertura bem abaixo das demais (1,49X), fato que pode influenciar na identificação das proteínas de interesse.

Tabela 3: Dados gerais do RNA-Seq e resultados do pré-processamento

Espécies	Tamanho do Genoma	Tamanho da Sequência (pb)	Quantidade de Sequências	Nº de Sequências Retiradas	Porcentagem de GC
E. guineensis	1.5Gb	100	22.657.832	288.159 (1,26%)	48%
J. curcas	339Mb	100	27.577.214	250.364 (0,90%)	44%
R. communis	351Mb	100	27.977.546	187.170 (0,66%)	43%

Transcritomas

Seis diferentes programas de montagem foram usados para gerar os transcritomas de sementes das plantas estudadas neste trabalho. Foram gerados entre 33.521 e 137.334 transcritos de *E. guineensis*, entre 32.823 e 135.475 transcritos de *J. curcas* e entre 29.122 e 132.908 transcritos de *R. communis*. Os programas claramente produziram diferentes resultados.

No geral, aqueles que requerem um genoma de referência produziram transcritomas mais enxutos do que os que aplicaram uma abordagem de novo. O programa Velveth/Oases, com k-mer = 17, montou o maior número de transcritos, com quantidades acima de 130.000 para cada espécie,

seguido 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 33.521, 32.823 e 29.122 e *E. guineensis*, *J. curcas* e *R. communis* respectivamente, independente dos valores de k-mers testados. O alinhamento com o genoma de referência diminuiu o número de transcritos gerados erroneamente.

O uso do Evidential Gene, que filtra os possíveis erros da montagem dos transcritos e elimina a redundância, se mostrou eficiente para produzir, a partir dos resultados de todos os montadores, transcritomas 20 a 34 vezes menor, com quantidades de transcritos que se aproximaram do número esperado a partir das informações dos genomas: 33.189, 20.033 e 16.309, enquanto a quantidade esperada é de 41.887, 23.076 e 28.584 para *E. guineensis, J. curcas* e *R. communis* respectivamente. O número menor de transcritos obtidos para as três espécies pode ser explicado pelo fato de serem dados de sementes, ou seja, de tecido-específico.

Completude e qualidade 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.

As porcentagens de completude obtidas por cada programa variaram entre 12,2% e 50,7% para os dados de *E. guineensis*, entre 17,7% e 73,8% para transcritos completos de *J. curcas* e entre 20,1% e 71,1% para os transcritos obtidos de *R. communis*. Após o uso do Evidential Gene as porcentagens atingiram os valores de 61,4% para *E. guineensis,* 79,2% para *J. curcas* e 80,2% para *R. communis*, mostrando o efeito positivo do uso deste programa, ainda que os resultados tenham ficado abaixo dos valores de completude dos respectivos genomas (92,7% para *E. guineensis*, 96,2% para *J. curcas* e 93,7% para *R. communis*), como pode ser visto na Figura 2. 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, genes tecido-específico no transcritoma de cinco tecidos diferentes de *R. communis* para a biossíntese de lipídeos e o enriquecimento das vias de metabolismo,

degradação e biossíntese de ácidos graxos e metabolismo de glicolipídeo nos transcritos anotados de *R. communis* indicaram um viés em função do tecido sequenciado (Brown et al., 2012). Outro trabalho demonstra diferenças de transcritos e vias metabólicas de ácidos graxos de mesocarpo, embrião ou endosperma de *E. guineensis* (Dussert et al., 2013).

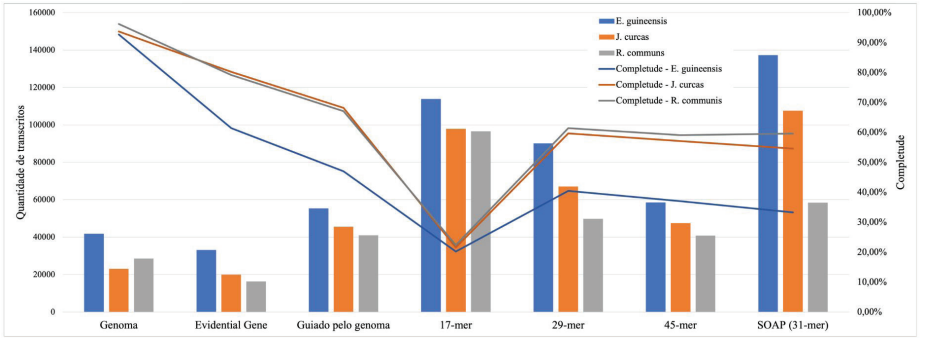

Figura 2: Comparação entre os montadores e as completudes obtidas. É possível observar que os montadores de novo obtiveram grandes quantidades de transcritos, mas baixas completudes quando comparados com os montadores guiados pelo genoma. O Evidential Gene obteve os maiores índices de completude e uma quantidade menor de transcritos.

Os montadores de novo geraram transcritos com tamanho médio de 636 nt enquanto os que usaram genomas de referência produziram transcritos com tamanho médio de 1.492 nt. Quando usou-se o k-mer 17, o tamanho médio dos transcritos foram de apenas 384 nt. Os valores médios dos transcritos produzidos pelo Evidential Gene foram 1.135, 1.305 e 1.463 nt em *E. guineensis*, *J. curcas* e *R. communis* respectivamente. Todos esses valores foram abaixo das médias obtidas pelos montadores guiados pelo genoma para cada espécie (teste-t não pareado - $p<0,003$), indicando que, para este parâmetro, os montadores de novo contribuem negativamente para o desempenho do evidential gene (Figua 3-A). Os montadores não influenciam o tamanho do maior transcrito montado (teste-t não pareado – $p>0,05$) (Figura 3-B). Os valores de N50 também foram influenciados pelo tipo de montador, sendo significativamente menor para os montadores de novo ($p<0,05$).

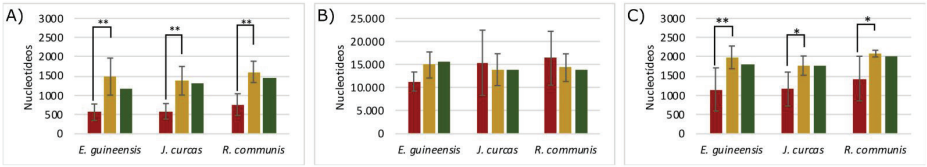

Figura 3: Tamanho (± desvio-padrão) dos transcritos gerados pelas diferentes classes de montadores (de novo = vermelho, guiado pelo genoma = amarelo) e pelo Evidental Gene (verde). A) Média, B) Maior sequência obtida, C) N50. Vermelho * = p<0,05; ** = p<0,01

Montadores baseado em genomas geram transcritos com maior extensão e fidedignidade

Em seguida usamos as proteínas de interesse, obtidas em cada via metabólica selecionada, para alinhar contra os transcritos gerados por cada montador. A ideia por trás dessa análise foi verificar a capacidade dos diversos montadores em gerar transcritos completos e com poucos erros, fatos que refletem nos parâmetros pident (% de identidade) e qcovs (% de bases da proteína de interesse que se alinharam no transcrito). Transcritos bem montados apresentariam altos valores para os dois parâmetros. Os resultados mostraram que os montadores de novo com k-mers de 17 produziram transcritos bastante incompletos, com porcentagens de identidade e cobertura de alinhamento mais baixos (Figura 4). Os montadores de novo com k-mers de 29 e 45 se comportaram de forma bem similar, com resultados bem melhores quando comparados com os anteriores e bem próximo do resultado obtido pelo montador Trinity, baseado em genoma. No entanto, os montadores baseados em genoma Star e StringTie foram os que produziam transcritos mais completos e com menos erros de montagem.

Os dados de *E. guineensis*, cujo sequenciamento obteve uma cobertura aquém da desejável, se beneficiou bastante com o uso do Evidential gene. Para esta espécie, o programa gerou um transcritoma mais enxuto com transcritos mais completos e com menos erros. Já para das demais espécies, que apresentaram uma cobertura maior, o Evidential gene obteve um resultado muito semelhante ao StringTie e/ou Star.

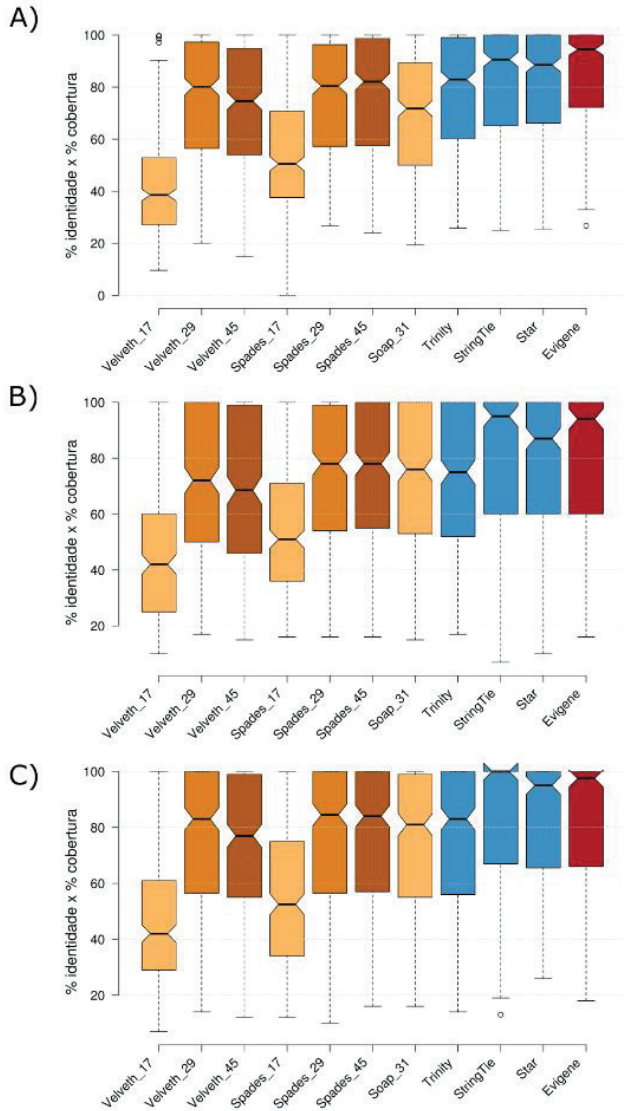

Figura 4: Boxplots dos valores obtidos para a porcentagem de identidade e porcentagem de sequência alinhada entre as proteínas de interesse e os transcritomas obtidos. A) *Elaeis guineensis*, B) *Ricinus communis* e C) *Jatropha curcas*. Os primeiros sete boxplots, em tons de laranjas, são referentes a montadores de novo. Os boxplots azuis são os que usam genoma de referência. Os boxplots em vermelho são referentes aos dados do EviGene.

Para compreender melhor como a relação entre a porcentagem de identidade e a porcentagem de sequência alinhada entre as proteínas de interesse e os transcritomas obtidos pode ajudar a definir estratégias de análise, calculamos os valores de regressão linear entre os valores de identidade e cobertura para cada método aplicado (Figura 5). É fácil notar que as performances dos métodos de montagem de novo foram limitadas quando comparados com os métodos baseado em genomas para todas as espécies. Esses algoritmos produziram, com grande frequência, transcritos parciais que se alinham pouco às sequências de referência. Esse fenômeno foi especialmente claro para os métodos Velveth_17 e Spades_17, a relação entre os dois parâmetros foi negativa para as três espécies (transcritos parciais com altas porcentagens de identidades e transcritos completos com baixa qualidade de montagem). Já os resultados dos montadores baseado em genomas foi melhor, pois a maioria dos transcritos formados apresentaram altos valores de porcentagem de identidade e cobertura. Dentre os três, os gráficos indicam que o StringTie foi o que obteve os melhores resultados, como já havíamos mostrado na Figura 4. Em *E. guineensis* (Figura 5-A), o Evidential Gene teve a melhor performance, compilando um genoma com altas coberturas do transcrito ainda que as porcentagens de identidade não sejam muito altas.

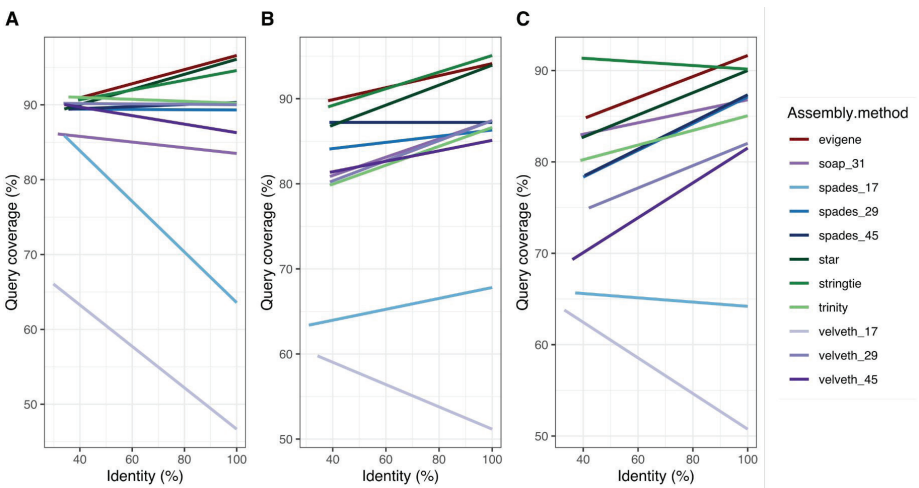

Figura 5: Modelos de regressão linear para cada montador testado, além do EviGene. A = *E. guineensis*, B = *J. curcas*, C = *R. communis*

As proteínas pertencentes às vias de metabolismo de ácidos graxos são amplamente expressas em sementes de oleaginosas

O próximo passo foi confirmar a presença das enzimas das vias metabólicas de ácidos graxos selecionadas neste trabalho. Para isso, criamos um banco de dados local com 437, 315 e 318 proteínas únicas de *E. guineensis*, *J. curcas* e *R. communis* obtidas a partir da base de dados KEGG para as doze vias selecionadas. O número total de enzimas únicas pesquisadas nas doze vias de interesse foi 170 organizadas em 142 ortologias kegg (K). A maioria das enzimas são codificadas por pelo menos dois genes diferentes.

De forma geral os transcritos referentes às enzimas de interesse foram encontrados nas montagens realizadas com altos valores de identidade e cobertura (acima de 70%): 152, 156 e 150 em *E. guineensis*, *J. curcas* e *R. communis* respectivamente, sendo que 135 são em comum. No entanto, algumas enzimas não foram encontradas em uma ou mais espécies, outras apresentaram valores baixos de identidade e/ou cobertura (Tabela 4). Quatro proteínas não foram encontradas em nenhum dos transcritomas: a enzima 2.4.1.337 (1,2-diacilglicerol 3-alfa-glicosiltransferase), do metabolismo de glicerolipídeos, converte a 1,2-Diacil-sn-glicerol em 1,2-Diacil-3-alfa-D-glicosil-sn-glicerol, foi encontrada no genoma de *R. communis*; a enzima 3.1.1.5 (Lisofosfolipase III), de metabolismo de glicerolipídeo, converte a 1-Acil-sn-glicero-3-fosfocolina glicerofosfocolina e está presente no transcritoma de folha de *E. guineensis* (Vieira, 2019); a enzima 2.7.8.8 (CDP-diacilglicerol--serina O-fosfatidiltransferase), também do metabolismo de glicerolipídeo, converte a CDP-diacilglicerol em Fosfatidil-L-serina, e encontra-se anotada apenas no genoma de *E. guineensis*; e a classe de enzima 1.14.99.-, que são oxidoredutases, atuando em doadores pareados, com incorporação ou redução de molécula de oxigênio, dentro do metabolismo de metabolismo de alfa-linoleico, está presente no transcritoma de folha de *E. guineensis* (Vieira, 2019). As enzimas 2.7.7.9 (UTP--glicose-1-fosfato uridililtransferase), do metabolismo de glicerolipídeo, que atua na reação da conversão da D-Glicose 1-fosfato que veio da via de glícolise/gliconeogênse em UDP-glicose, e a enzima 1.1.1.189 (Prostaglandina-E(2) 9-redutase), do metabolismo de ácido aracdônico, que atua na síntese da prostaglandin F2-alpha, foram encontradas com baixos níveis de conservação nos transcritomas das três espécies o que indica que estas proteínas possivelmente não são expressas em sementes de plantas. Sete enzimas foram encontradas apenas em *E. guineensis*, duas em *J. curcas* e uma em *R. communis* (Tabela 4).

Tabela 4: Enzimas pouco conservadas ou ausentes em pelo menos uma espécie.

Código da enzima	Enzima	Código da via	Via	E. guineensis (%)	J. curcas (%)	R. communis (%)
1.3.1.9/ 1.3.1.10	Enoil-(Proteína de Transporte de Acila) Redutase (NADH e NADPH)	00061	Biossíntese de ácidos graxos	27*	100	100
1.3.1.-/ 1.3.1.38	Trans-2-enoil-CoA redutase (NAPPH)	00061/ 00062	Biossíntese de ácidos graxos/ Elongação de ácidos graxos	NE*	98	100
3.1.2.22	Proteína Palmitoil hidrolase	00062	Elongação de ácidos graxos	96	100	53
3.1.2.2	Palmitoil-CoA hidrolase	00062/ 01040	Elongação de ácidos graxos/ Biossíntese de ácidos graxos insaturados	100	100	47
1.14.14.80	Ácido graxo de cadeia longa omega-monooxigenase	00071	Degradação de ácidos graxos	100	63	44
5.4.99.8	Cicloartenol sintase	00100	Biossíntese de esteróides	100	42	58
5.5.1.9	Cicloeucalenol cicloisomerase	00100	Biossíntese de esteróides	NE*	48	100
1.14.19.41	Esterol 22-desaturase	00100	Biossíntese de esteróides	100	NE+	NE
1.14.18.11	Planta 4-alfa-monometilsterol monooxigenase	00100	Biossíntese de esteróides	86	54	49
2.7.1.31	Glicerato 3-kinase	00561	Metabolismo de glicerolipídeo	NE*	100	100
3.1.3.21	Glicerol-1-fosfatase	00561	Metabolismo de glicerolipídeo	94	100	73
2.7.7.9	UTP--glicose-1-fosfato uridililtransferase	00561	Metabolismo de glicerolipídeo	57*	18	66
2.4.1.184	Galactolipídeo galactosiltransferase	00561	Metabolismo de glicerolipídeo	91	14	46
2.4.1.337	1,2-diacilglicerol 3-alfa-glicosiltransferase	00561	Metabolismo de glicerolipídeo	NE+	NE+	NE
3.6.1.16	CDP-glicerol difosfatase	00564	Metabolismo de glicerofosfolipídeo	100	51	NE
3.1.1.5	Lisofosfolipase III	00564	Metabolismo de glicerofosfolipídeo	NE*	NE	NE
2.3.1.15	Glicerol-3-fosfato 1-O-aciltransferase	00564	Metabolismo de glicerofosfolipídeo	100	100	34
2.3.1.198	Glicerol-3-fosfato 2-O-aciltransferase	00564	Metabolismo de glicerofosfolipídeo	100	100	34
2.1.1.103	Fosfoetanolamina N-metiltransferase	00564	Metabolismo de glicerofosfolipídeo	100	91	72
3.1.3.27	Fosfatidilglicerofosfatase	00564	Metabolismo de glicerofosfolipídeo	NE*	100	95
2.7.8.8	CDP-diacilglicerol--serina O-fosfatidiltransferase	00564	Metabolismo de glicerofosfolipídeo	NE	NE+	NE+
3.4.19.14	Leucotrieno-C(4) hidrolase	00590	Metabolismo de ácido aracdônico	100	51	61
1.1.1.184	Carbonil redutase (NADPH)	00590	Metabolismo de ácido araquidônico	70	100	100
1.1.1.189	Prostaglandina-E(2) 9-redutase	00590	Metabolismo de ácido araquidônico	35*	38	50
3.3.2.10	Epóxido solúvel hidrolase	00590	Metabolismo de ácido araquidônico	NE+	98	NE+
1.13.11.58	Linoleato 9S-lipoxigenase	00591	Metabolismo de ácido araquidônico	100	94	43
1.14.99.-	Oxidoredutase	00592	Metabolismo de ácido linoleico	NE*	NE	NE
2.1.1.141	Jasmonato O-metilltransferase	00592	Metabolismo de ácido linoleico	NE	91	NE
1.14.19.17	Sphingolipídeo 4-desaturase	00600	Metabolismo de esfingolipídeo	50*	94	95
2.7.1.138	Ceramida kinase	00600	Metabolismo de esfingolipídeo	100	19	90

Células laranjas escuras: NE = Não encontrado; células laranjas: proteínas pouco conservadas (< 70% de identidade); células laranja claras: proteínas conservadas (aproximadamente 70% de identidade); células verdes: proteínas muito conservadas (> 90% de identidade). *Transcrito encontrado em outros tecidos (Vieira, 2019). +Enzima não encontrada no genoma e na base de dados KEGG

Tabela 5: Número de proteínas ausentes por montador

	Conservação baixa	Conservação média	Conservação alta	Ausentes	Total
E. guineensis					
Star	3	2	3	23	31
StringTie	4	3	1	23	31
Trinity	1	0	2	23	25
Evidential Gene	0	0	3	23	26
J. curcas					
Star	4	4	5	6	19
StringTie	5	3	4	6	18
Trinity	2	1	2	6	11
Evidential Gene	2	0	1	6	9
R. communis					
Star	10	5	5	11	31
StringTie	9	5	1	11	26
Trinity	5	2	1	11	19
Evidential Gene	0	1	1	11	13

O uso de múltiplos montadores auxilia a identificação e anotação de proteínas de interesse

O ponto chave deste trabalho, como já mencionado anteriormente, foi o uso de seis diferentes montadores, sendo três baseado em alinhamento com o genoma e outros três que geram transcritomas de novo. Todos os resultados resultantes dos seis montadores foram usados como input para o programa Evidential Gene gerar um transcritoma robusto, sem redundância e completo. Após analisar via a via, com a identificação de todas as proteínas de interesse nos três transcritomas, nós questionamos se o transcritoma gerado pelo Evidential Gene foi, de fato, aquele que trouxe os melhores resultados. Para isso, contabilizamos o número de proteínas ausentes nos transcritomas gerados pelos montadores Star, StringTie, Trinity e Evidential Gene. Em seguida, identificamos os falso-negativos, ou seja, transcritos que não foram montados por um determinado montador, mas presente nos demais (um ou mais) com alta identidade e cobertura.

Um total de 23, 6 e 11 transcritos não foram montados por nenhum programa em *E. guineensis*, *J. curcas* e *R. communis* respectivamente (Tabela

5, Ausentes). Se tivéssemos considerado apenas o transcritoma gerado pelo Evidential Gene, teríamos afirmado, erroneamente, que três enzimas de *E. guineensis*, uma de *J. curcas* e uma de *R. communis* não estavam presentes. No entanto, elas foram montadas com altos índices de conservação por outro (s) montador (es), como pode ser visualizado na Tabela 5, Conservação alta.

Por outro lado, essa análise também evidencia que muitos transcritos com baixa conservação foram incluídos nos transcritomas de diversos montadores. Para exemplificar, os programas Star e StringTie, em *R. communis*, não incluíram em suas montagens 10 e 9 transcritos com baixa conservação ou que foram identificados apenas fragmentos dos mesmos.

Os transcritos obtidos variaram bastante em função dos seus protocolos. Mas não é só isso que influencia a performance desses programas. Cada programa pode se comportar de forma diferente dependendo da espécie estudada e de como o experimento foi realizado. Há muitos trabalhos publicados que comparam montadores de transcritomas com conclusões acerca do uso de apenas um programa baseado na acurácia, precisão ou outras métricas. Mas um trabalho comparou 10 diferentes programas, entre protocolos de novo e guiados pelo genoma em nove bibliotecas de RNA-Seq pertencentes a sete espécies diferentes, incluindo mamíferos, planta, fungo e bactéria e mostrou que nenhum programa foi o melhor montador para todas as amostras (Hölzer e Marz, 2019).

Muitos outros trabalhos vão além, onde mostram que o uso de apenas um programa ou um método de reconstrução pode não ser o melhor para todos os loci gênicos, e recomendam a combinação de diferentes programas e métodos de reconstrução para obter resultados mais robustos e precisos (Cabau et al., 2017; Holding et al., 2018; Sahraeian et al., 2017). Em situações onde não há um genoma de referência de qualidade, o uso de alguns programas de novo pode gerar transcritomas de qualidade, sendo o resultado claramente dependente da escolha do valor de k-mer (quando for o caso). No entanto, mesmo que tenha disponível um bom genoma para ser usado como referência, a junção dos resultados obtidos pelos montadores de novo por programas como o Evidential Gene, podem trazer melhorias substanciais para o transcritoma de interesse (Huang, Chen e Armbruster, 2016), corroborando com os nossos resultados. De fato, ao compilar os resultados obtidos neste trabalho em cada uma das métricas e parâmetros analisados, os programas STAR, StringTie, Trinity e Evidential Gene se alternam entre os que tiveram

o melhor desempenho (Tabela 6). Os montadores baseados em genomas tiveram ótimos desempenhos em relação aos parâmetros relacionados aos transcritos analisados de forma individual, como o tamanho médio e maior transcrito obtido, o N50 e a quantidade de transcritos completos, métrica avaliada de acordo com o alinhamento com as proteínas do banco de dados local. No entanto, quando avaliamos a melhor estratégia para a obtenção da quantidade real de transcritos (sem redundância), a completude e a quantidade de transcritos não montados (ausentes), não há dúvida do desempenho excelente do Evidential Gene.

Tabela 6: Indicação do melhor programa usado para a montagem dos transcritomas por parâmetro analisado

	Número de transcritos (proximidade com o número real)	Completude (BUSCO)	Tamanho médio do transcrito	Maior transcrito	N50	Transcritos completos (%pident x %qcovs)	Menor quantidade de transcritos ausentes
E. guineensis	Evidential Gene	Evidential Gene	STAR	STAR	StringTie	EvidentialGene	Trinity
J. curcas	Evidential Gene	Evidential Gene	StringTie	STAR	StringTie	StringTie	Evidental Gene
R. communis	Evidential Gene	Evidential Gene	STAR	StringTie	StringTie	StringTie	Evidental Gene

Estratégias como esta foram aplicadas em trabalhos envolvendo dados de plantas cujos autores concluem, de forma similar, que a combinação de dados melhorou de forma substancial os transcritomas obtidos (Petek et al., 2020; Sadat-Hosseini et al., 2020; Visser et al., 2015). Especificamente em um trabalho com Nicotiana benthamiana, os autores utilizaram quatro montadores de novo. 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 (Nakasugi et al., 2014).

Conclusão

Ao menos 150 proteínas, de 170 previamente identificadas nas vias metabólicas de ácidos graxos de plantas, foram encontradas nas sequências obtidas a partir de sementes de *E. guineensis*, *J. curcas* e *R. communis*. A combinação dos resultados obtidos pelos três montadores de novo e pelos três montadores guiados pelo genoma através do programa Evidential Gene mostrou ser eficiente em produzir transcritomas enxutos, sem redundância e com os mais altos valores de completude quando comparados com os montadores individualmente. Os montadores guiados pelo genoma, juntamente com o Evidential Gene, geraram transcritos mais completos e com menos erros. A estratégia aplicada neste trabalho mostrou ser eficiente e com alta aplicabilidade em dados de organismos com ou sem genomas de referência.

Tais informações são importantes para aumentar a compreensão das vias metabólicas de ácidos graxos de *E. guineensis*, *J. curcas* e *R. communis* permitindo novos progressos nos programas de melhoramento genético e metabólico, com foco na produção de biocombustíveis, químicos, fármacos e cosméticos de interesse industrial.

Agradecimentos

Ao Programa de Pós-graduação em Tecnologias Química e Biológica da Universidade de Brasília (UNB) pela formação acadêmica do discente Vinícius Nattan Lemos Silva e pela oportunidade de desenvolver este trabalho.

Referência Bibliográfica

ACHTEN, W. M. J.; VERCHOT, L.; FRANKEN, Y. J.; MATHIJS, E.; SINGH, V. P.; AERTS R.; MUYS, B. Jatrophabio-diesel production and use. **Biomass and Bioenergy**, v. 32, n. 12, p. 1063-1084, 2008.

ANDREWS, S. **FastQC:** a quality control tool for high throughput sequence data.2014. Disponível em: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>. Acesso em: 15 de abril de 2018.

AGÊNCIA NACIONAL DO PETRÓLEO (Brasil). Disponível em: < https://www.gov.br/anp/pt-br>. Acesso em: 17 de dezembro de 2020.

ARONESTY, E. **ea-utils**: Command-line tools for processing biological sequencing data. 2011. Disponivel em: < https://expressionanalysis.github.io/ea-utils/>. Acesso em: 15 de abril de 2018.

BANKEVICH, A.; NURK, S.; ANTIPOV, D.; GUREVICH, A. A.; DVORKIN, M.; KULIKOV, A. S.; LESIN, V. M.; NIKOLENKO, S. I.; PHAM, S.; PRJIBELSKI, A. D.; PYSHKIN, A. V.; SIROTKIN, A. V.; VYAHHI, N.; TESLER, G.; ALEKSEYEV, M. A.; PEVZNER, P. A. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. **Journal of computational biology**: a journal of computational molecular cell biology, v.19, n. 5, 455-477, 2021.

BROWN, A. P.; JOHAN, T. M; KROON, J. T. M.; SWARBRECK, D.; FEBRER, M.; LARSON, T. R.; GRAHAM, I. A.; CACCAMO, M.; SLABAS, A. R. Tissue-specific whole transcriptome sequencing in castor, directed at understanding triacylglycerol lipid biosynthetic pathways. **PLoS ONE**, v. 3, 2012.

CABAU, C. FRÉDÉRIC ESCUDIÉ, F.; DJARI, A.; GUIGUEN, Y.; BOBE, J. KLOPP, C. Compacting and correcting Trinity and Oases RNA-Seq de novo assemblies. **PeerJ**, 2017.

CHAN, A. P.; CRABTREE, J.; ZHAO, Q.; LORENZI, H.; ORVIS, J.; PUIU, D.; MELAKE-BERHAN, A.; JONES, K. M.; REDMAN, J.; CHEN, G.; CAHOON, E. B.; GEDIL, M.; STANKE, M.; HAAS, B. J.; WORTMAN, J. R.; FRASER-LIGGETT, C. M.; RAVEL, J.; RABINOWICZ, P. D. Draft genome sequence of the oilseed species *Ricinus communis*. **Nature Biotechnology**, v. 28, n. 9, p. 951-956, set. 2010.

CONESA, A. Conesa, A., Madrigal, P., Tarazona, S. GOMEZ-CABRERO, D.; CERVERA, A.; MCPHERSON, A.; SZCZEŚNIAK, M. W.; GAFFNEY, D. J.;

ELO, L. L.; ZHANG, X.; MORTAZAVI, A. A survey of best practices for RNA-seq data analysisGenome Biology, **Genome Biology**, v. 17, n. 13, 2016.

DOBIN, A.; DAVIS, C. A.; SCHLESINGER, F.; DRENKOW, J.; ZALESKI, C.; JHA, S.; BATUT, P.; CHAISSON, M.; GINGERAS, T. R. STAR: ultrafast universal RNA-seq aligner. **Bioinformatics**, v. 29, n. 1, p. 15-21, 2013.

DUSSERT, S.; GUERIN, C.; ANDERSSON, M.; JOËT, T.; TRANBARGER, T. J.; PIZOT, M.; SARAH, G.; OMORE, A.; DURAND-GASSELIN, T.; MORCILLO, F. Comparative transcriptome analysis of three oil palm fruit and seed tissues that differ in oil content and fatty acid composition. **Plant Physiology**, v. 162, n. 3, p. 1337–1358, 1 jul. 2013.

EDRISI, S. A. DUBEY, R. K.; TRIPATHI, V.; BAKSHI, M.; SRIVASTAVA, P.; JAMIL, S.; SINGH, H. B.; SINGH, N.; ABHILASH, P. C. *Jatropha curcas* L.: A crucified plant waiting for resurgence. **Renewable and Sustainable Energy Reviews**, v. 41, p. 855-862, 2015.

ERB, T. J.; JONES, P. R.; BAR-EVEN, A. Synthetic metabolism: metabolic engineering meets enzyme design. **Current Opinion in Chemical Biology**, v. 37, p. 56-62, 2017.

FAO. **FAOSTAT**. Disponível em: < http://www.fao.org/faostat/en/#home>. Acesso em: abril de 2019.

FORERO, C. L. B. Biodiesel from castor oil: a promising fuel for cold weather. **Renewable Energy and Power Quality Journal**, v. 1, n. 3, p. 59–62, 1 mar. 2005.

FUENTES, A.; GARCIA, C.; HENNECKE, A.; MASERA, O. Life cycle assessment of *Jatropha curcas* biodiesel production: a case study in Mexico. **Clean Technologies and Environmental Policy**, v. 20, n. 7, p. 1721–1733, 2018.

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. **Gene-omes built from mRNA-seq not genome DNA:** simple, quick, accurate, less cost, more complete gene sets. Indiana University, Bloomington. 2013. Disponível em: <https://doi.org/10.7490/f1000research.1112594.1>. Acesso em: 22 de maio de 2018 .

GRABHERR, M. G.; HAAS, B.; YASSOUR, M. LEVIN, J. Z.; THOMPSON, D. A.; AMIT, I.; ADICONIS, X.; FAN, L.; RAYCHOWDHURY, R.; ZENG, Q.; CHEN, Z.; MAUCELI, E.; HACOHEN, N.; GNIRKE, A.; RHIND, N.; DI PALMA,

F.; BIRREN, B. W.; NUSBAUM, C.; LINDBLAD-TOH, K.; FRIEDMAN, N.; REGEV, N. Full-length transcriptome assembly from RNA-Seq data without a reference genome. **Nature biotechnology**, v. 29, n. 7, p. 644–652, 15 maio 2011.

HA, J.; JUNGMIN HÁ, J.; SHIM, S.; LEE, T.; KANG, Y. J.; HWANG, W. J.; LAOSATIT, K.; LEE, J.; KIM, S. K.; SATYAWAN, D.; LESTARI, P.; YOON, M. Y.; KIM, M. Y.; CHITIKINENI, A.; TANYA, P.; SOMTA, P.; SRINIVES, P.; VARSHNEY, R. K.; SUK-HA LEE. Genome sequence of *Jatropha curcas* L., a non-edible biodiesel plant, provides a resource to improve seed-related traits. **Plant Biotechnology Journal**, v. 17, n. 2, p. 517–530, 1 fev. 2019.

HOLDING, M. L.; HOLDING, M. L.; MARGRES, M. J.; MASON, A. J.; PARKINSON, C. L.; ROKYTA, D. R. *et al.* Evaluating the performance of de novo assembly methods for venom-gland transcriptomics. **Toxins**, 2018.

HÖLZER, M.; MARZ, M. De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers. **GigaScience**, 2019.

HUANG, X.; CHEN, X. G.; ARMBRUSTER, P. A. Comparative performance of transcriptome assembly methods for non-model organisms. **BMC Genomics**, v. 17, n. 1, p. 1–14, 2016.

JUNGINGER, M.; GOH, C. S.; FAAIJ, A. **International bioenergy trade:** history status & outlook on securing sustainable bioenergy supply, demand and markets. [s.l]: Springer, 2014. 233 p.

LEE, S. K.; KUKLEE, S.; CHOU, H.; SHAM, T.; SOONLEE, T.; DKEASLING, J. 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.

LUO, R.; WONG, T.; ZHU, J.; LIU, C.-M.;ZHU, X.; WU, E.; LEE, L-K.; LIN, H.; ZHU, W.; CHEUNG, D. W.; TING, H. F.; YIU, S.-M.; PENG, S.; YU, C.; LI, Y.; LI, R.; LAM, T. W. SOAP3-dp: Fast, Accurate and Sensitive GPU-Based Short Read Aligner. **PLoS ONE**, v. 8, n. 5, 31 maio 2013.

MARTIN, M. Cutadapt removes adapter sequence from high-throughput sequencing reads. **EMBnet.journal**, v. 17, n. 1, p. 10–12, 2011.

MBA, O. I.; DUMONT, M. J.; NGADI, M. Palm oil: processing, characterization and utilization in the food industry - a review. **Food Bioscience,** v. 10, p. 26-41, 2015.

METZKER, M. L. Sequencing technologies: the next generation. **Nature reviews**. Genetics, v. 11, n. 1, p. 31-46, jan. 2010.

NAKASUGI, K.; CROWHURST, R.; BALLY, J.; WATERHOUSE, P. Combining transcriptome assemblies from multiple *De Novo* assemblers in the allo-tetraploid plant nicotiana benthamiana. **PLoS ONE**, 2014.

OGATA, H. GOTO, S.; SATO, K.; FUJIBUCHI, W.; BONO, H.; KANEHISA, M. KEGG: Kyoto Encyclopedia of Genes and Genomes. **Nucleic Acids Research**, v. 27, n. 1, p. 29–34, 1 jan. 1999.

ONG, A. L.; TEH, C.-K.; MAYES, S.; MASSAWE, F.; APPLETON, D. R.; KULAVEERASINGAM, H. An improved oil palm genome assembly as a valuable resource for crop improvement and comparative genomics in the Arecoideae subfamily. **Plants**, v. 9, n. 11, 2020.

OWEN, C.; PATRON, N. J.; HUANG, A.; OSBOURN, A. Harnessing plant metabolic diversity. **Current Opinion in Chemical Biology**, v. 40, p. 24-30, 2017.

PERTEA, M. PERTEA, G. M.; ANTONESCU, C. M.; CHANG, T.C.; MENDELL, J. T.; SALZBERG, S. L. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. **Nature Biotechnology**, v. 33, p. 290–295, 2015.

PETEK, M.; ZAGORŠČAK, M.; RAMŠAK, Z.; SANDERS, S.; TOMAŽ, S.; TSENG, E.; ZOUINE, M.; COLL, A.; GRUDEN. K. Cultivar-specific transcriptome and pan-transcriptome reconstruction of tetraploid potato **Scientific Data**, v. 7, n. 249, 2020.

R Core Team. **R**: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. Disponível em: <https://www.R-project.org>. Acesso em: :16 de agosto de 2020.

SABLOK, G. FU, Y.; BOBBIO, V.; LAURA, M.; ROTINO, G. L.; BAGNARESI, P.; ALLAVENA, A.; VELIKOVA, V.; VIOLA, R.; LORETO, F.; LI, M.; VAROTTO, C. Fuelling genetic and metabolic exploration of C3 bioenergy crops through the first reference transcriptome of *Arundo donax* L. **Plant Biotechnology Journal**, v.12, n. 5, 2014.

SADAT-HOSSEINI, M.; BAKHTIARIZADEH, M. R.; BOROOMAND, N.; TOHIDFAR, M.; VAHDATI, K. Combining independent de novo assemblies to optimize leaf transcriptome of Persian walnut. **PLoS ONE**, 2020.

SAHRAEIAN, S. M. E. MOHIYUDDIN, M.; SEBRA, R.; TILGNER, H.; AFSHAR, P. T.; AU, K. F.; ASADI, N. B.; GERSTEIN, M. B.; WONG, W. H.; SNYDER, M. P.; SCHADT, E.; LAM, H. Y. K. Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. **Nature Communications**, v. 8, n. 59, 2017.

SAMARTH, N. B.; MAHANWAR, O. A. Modified vegetable oil based additives as a future polymeric material - review. **Open Journal of Organic Polymer Materials**, v. 5, n. 1, p. 1-22, 9 jan., 2015.

SAMBANTHAMURTHI, R.; SINGH, R.; KADIR, A. P. G.; ABDULLAH, M. O.; KUSHAIRI, A. Opportunities for the oil palm via breeding and biotechnology. In: JAIN, S. M.; PRIYADARSHAN, P. M. (Eds). **Breeding plantation tree crops**: tropical species. Springer, New York, NY, 2009. p. 377-421.

SCHULZ, M. H.; SCHULZ, M. H; ZERBINO, D. R.; VINGRON, M.; BIRNEY, E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. **Bioinformatics**, v. 28, n. 8, p. 1086-1092, 24 fev. 2012.

SIMÃO, F. A.; SIMÃO, F. A.; WATERHOUSE, R. M.; IOANNIDIS, P.; KRIVENTSEVA, E. V.; ZDOBNOV, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. **Bioinformatics**, v. 31, n. 19, p. 3210-3212, out. 2015.

SINGH, R.; ONG-ABDULLAH, M.; LOW, E. T. L.; MANAF, M. A. A.; ROSLI, R.; NOOKIAH, R.; OOI, L. C. L.; OOI, S. E.; CHAN, K. L.; HALIM, M. A.; AZIZI, N.; NAGAPPAN, J.; BACHER, B.; LAKEY, N.; SMITH, S. W.; HE, D.; HOGAN, M.; BUDIMAN, M. A.; LEE, E. K.; DeSalle, R.; KUDRNA, D.; GOICOECHEA, J. L.; WING, R. A.; WILSON, R. K.; FULTON, R. S.; ORDWAY, J. M.; MARTIENSSEN, R. A.; SAMBANTHAMURTHI, R. Oil palm genome sequence reveals divergence of interfertile species in Old and New worlds. **Nature**, v. 500, n. 7462, p. 335-339, 24 jul. 2013.

VIEIRA, L. R. **Respostas morfofisiológicas, metabolômicas e transcritômicas da palma de óleo (Elaeis guinensis) aos estresses abióticos de seca e salinidade**. (Mestrado em Biotecnologia Vegetal). Lavras: Universidade de Lavras, 2019.

VISSER, E. A.; WEGRZYN, J. L.; STEENKMAP, E. T.; MYBURG, A. A.; NAIDOO, S. Combined de novo and genome guided assembly and annotation of the Pinus patula juvenile shoot transcriptome. **BMC Genomics**, v. 16, n. 1, p. 1-13, 2015.

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

ZHOU, Y. J.; BUIJS, N. A.; ZHU, Z.; QIN, J.; SIEWERS, V.; NIELSEN, J. Production of fatty acid-derived oleochemicals and biofuels by synthetic yeast cell factories. **Nature communications**, v. 7, n. 1, p. 11709, 25 maio 2016.

Embrapa

Recursos Genéticos e Biotecnologia

MINISTÉRIO DA AGRICULTURA, PECUÁRIA E ABASTECIMENTO

PÁTRIA AMADA BRASIL GOVERNO FEDERAL