UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus de São José do Rio Preto

Relatório Final de Pós-doutoramento

Análise do transcriptoma de um consórcio degradador de biomassa lignocelulósica

Pós-doutoranda: Camila Cesário Fernandes Sartini

Supervisora: Profª. Drª. Eleni Gomes

São José do Rio Preto - SP

Novembro - 2024

Resumo: A produção de etanol de segunda geração, a partir da biomassa lignocelulósica tem sido indicada como uma das alternativas mais promissoras e ambientalmente sustentáveis para a substituição dos combustíveis fósseis. Uma das formas de desconstruir a biomassa e utilizar os açúcares fermentescíveis é através da utilização de consórcios bacterianos que produzem enzimas altamente específicas que possuem a capacidade de quebrar a estrutura da lignocelulose. Para abordar esse desafio tecnológico foram isolados dois consórcios bacterianos com características importantes relacionadas a degradação de lignocelulose. Esses consórcios utilizam como fonte de carbono, apenas resíduos lignocelulósicos (bagaço de cana-de-açúcar, palha de milho ou casca de amendoim), e liberam no meio de cultivo glicose. Além disso, esses estudos demonstraram que o consórcio isolado do solo e cultivado com bagaço de cana-de-açúcar mostra uma diminuição significativa do teor de celulose e hemicelulose, além de mostrar teores estáveis de lignina. Adicionalmente, estudos realizados para caracterizar a população bacteriana que constituem um dos consórcios conseguiu a recuperação de 51 genomas de organismos diferentes e com potencialidade de degradação da biomassa vegetal. Observou-se também a presença de mais de 11000 módulos de genes relacionados às enzimas úteis na desconstrução da lignocelulose. Por outro lado, observa-se a baixa liberação de glicose no meio livre, apesar dos organismos sobreviverem *in vitro* com apenas bagaço de cana-de-açúcar como fonte de carbono. Desta forma, o presente projeto objetiva desvendar as atividades enzimáticas da população microbiana desse consórcio e identificar potenciais organismos, genes, enzimas lignocelulolíticas e vias metabólicas envolvidas nesse processo através de análises metatranscriptomicas. Tais conhecimentos e informações poderão contribuir efetivamente para gerar avanços tecnológicos necessários que permitam a síntese de um conjunto de enzimas que possam aumentar a eficiência dos processos de uso da biomassa vegetal como fonte de energia renovável, especificamente a produção de etanol.
Palavras-chave: consórcio microbiano, metatranscriptoma, lignocelulose

1. INTRODUÇÃO

A biomassa vegetal ou o resíduo lignocelulósico, como o bagaço de cana-de-açúcar, palha de milho, casca de arroz, e outros resíduos da agroindústria, é um importante recurso de carbono renovável para a implantação de biorrefinaria e, portanto, é considerada uma alternativa sustentável e ambientalmente correta à substituição do uso de petróleo (KAMM; KAMM, 2004).

Nos últimos anos, tem-se observado um aumento gradual no preço do petróleo, determinado pela redução na oferta prevista para médio e longo prazo. Do mesmo modo o domínio de áreas produtoras tem levado ao aparecimento conflitos políticos e sociais que causam grande instabilidade nas relações internacionais e sofrimento de grande parte da população. Além disso, o aquecimento global do planeta, determinado pelo acúmulo de gases do efeito estufa na atmosfera, é uma realidade que impulsiona o desenvolvimento de processos industriais à base de matérias-primas renováveis (GUPTA; GAUR, 2019).

Seguindo a tendência mundial, a indústria sucroalcooleira brasileira tem manifestado interesse em tecnologias sustentáveis que possam ser agregadas à sua cadeia produtiva. Esforços estão atualmente sendo direcionados para incluir seus resíduos industriais, a palha e o bagaço da cana-de-açúcar, no ciclo de produção. Tal incorporação tecnológica permitirá que esses materiais sejam então utilizados para a produção de etanol (de segunda geração) e de outros insumos renováveis. Benefícios econômicos, sociais e ambientais são esperados como consequência.

Os açúcares do bagaço, assim como aqueles de qualquer outro material lignocelulósico, encontram-se na forma de polímeros (celulose e hemicelulose) associados entre si e recobertos por uma macromolécula aromática complexa (lignina), formando a microfibrila celulósica. Esta, por sua vez, constitui a parede celular (fibra) vegetal, uma estrutura recalcitrante difícil de ser desestruturada e convertida em monossacarídeos fermentescíveis

(CANILHA et al., 2010). A utilização desses açúcares presentes na biomassa para a produção de biocombustíveis requer várias etapas o que aumenta significativamente os custos da produção (HEMANSI et al., 2019). Antes do processo de sacarificação da celulose e hemicelulose é necessário fazer um pré-tratamento do material, para a retirada da lignina e desestruturação das fibras para tornar a celulose e hemicelulose mais suscetíveis à hidrólise.

Existem pré-tratamentos físicos, químicos e microbiológicos utilizados para o início do processo de utilização da celulose e hemicelulose na produção de bioprodutos, incluindo o etanol de segunda geração (HEMANSI et al., 2019). Entretanto, nenhum dos processos utilizados atualmente é considerado ideal para o preparo da lignocelulose para posterior degradação enzimática. Um pré-tratamento adequado deve ter como características: ter um custo baixo, principalmente sem grandes custos energéticos; maximizar a área de contato das partículas, minimizar a perda de açúcares úteis no processo fermentativo e aumentar a capacidade de hidrólise enzimática e também diminuir a produção de inibidores do processo fermentativo (ADITIYA et al., 2016). Dessa forma, o pré-tratamento biológico/enzimático traz vantagens aos processos químicos ou físicos, como a não utilização de reagentes químicos e o baixo consumo de energia, além de ser um processo específico que tem como objetivo dissociar o complexo lignina-celulose, diminuir o grau de cristalinidade da celulose, aumentar área superficial da biomassa e minimizar formação de compostos que inibem a hidrolise da biomassa e a fermentação dos açúcares liberados (RABELO, 2015).

A extensa variedade de carboidratos encontrados na parede celular das plantas requer uma multiplicidade de enzimas para sua degradação (ARISTIDOU & PENTILLÂ, 2000). Desta forma, atualmente, uma das etapas mais caras durante a produção de álcool a partir da biomassa é a hidrólise enzimática da lignocelulose e liberação de açúcares capazes de serem usados no processo fermentativo (WONGWILAIWALIN et al., 2013). Isso porque a hidrólise desse substrato requer a ação conjunta de vários tipos de enzimas. A celulose contém glicose e a hemicelulose e diversas unidades de açúcar tais como manose, xilose, glicose, galactose e arabinose. Celulase compreende um grupo de pelo menos três enzimas, as endo e exoglucanase e as β-glucosidases. Endoglucanase (endo1,4-D hidrolase glucana, E.C. 3.2.1.4) ataca as regiões de baixa cristalinidade da fibra da celulose, exoglucanase (celobiohidrolase glucano 1,4-β-D, E.C. 3.2.1.91) remove as unidades de celobiose a partir das extremidades livres da cadeia e, finalmente, unidades de celobiose são hidrolisados em glicose pela β-glucosidase (E.C.3.2.1.21). As enzimas líticas da hemicelulose formam um grupo mais complexo e são uma mistura de, pelo menos, oito enzimas tais como endo-1,4-β-D-xilanases, xylocuronidases exo-1,4-β-D, α-L-arabinofuranosidase, mananases endo-1,4-D-β, p-manosidases, acetilo xylanesterases, α-glucoronidases and α- galactosidases (GUPTA et al., 2015).

A lignina é o segundo componente mais abundante da lignocelulose depois da celulose. É um polímero aromático amorfo com alto peso molecular que é formado com fenilpropano como base unidade estrutural (CHEN et al., 2018). A estrutura complexa da lignina a torna extremamente difícil a degradação, sendo que a degradação eficiente da lignina é necessária para a utilização eficiente dos recursos de biomassa. A lignina pode ser degradada por muitas enzimas, incluindo a lacase (Lac), lignina peroxidase (LiP), peroxidase de manganês (MnP), celobiose desidrogenase, álcool arílico oxidase e álcool arílico desidrogenase (BILAL & IQBAL, 2019). Lac, LiP e MnP em conjunto degradam lignina particularmente eficaz como demonstrado recentemente por Zhang e colaboradores

(2020). Nesse trabalho os autores tinham como objetivo geral melhorar a metodologia enzimática de degradação da lignina para remediação de resíduos lignocelulósicos da agroindústria e aproveitamento tecnológico dos açúcares da celulose e hemicelulose.

Considerando a complexidade estrutural da lignocelulose é conhecido que as enzimas de conversão de biomassa têm que trabalhar em conjunto, para se beneficiar de sinergismo entre a sua especificidade (para diferentes componentes e regiões de lignocelulose), bem como a atenuação da sua inibição (por diferentes componentes lignocelulósicos ou produtos de degradação). Muitos organismos, como archaeas, bactérias, fungos, protistas, plantas e animais possuem enzimas lignocelulolíticas que permitem a quebra da biomassa e utilização de seus açúcares. Muitas destas enzimas são secretadas, quer isoladamente ou formando complexos supramoleculares, celulossomas, como pode ser observado em revisão do assunto publicado por Jouzani e Taherzadeh (2015). A degradação de resíduos agroindustriais lignocelulósico por meio de comunidades microbianas complexas é uma abordagem promissora que proporciona uma decomposição eficiente da biomassa para posterior conversão em produtos de valor agregado (JIMENEZ et al 2017).

Conhecendo os aspectos relacionados às dificuldades da degradação da lignocelulose e à baixa eficiência nos processos atuais na degradação enzimática da biomassa vegetal nós iniciamos em 2010 estudos utilizando para a execução dessas atividades consórcios bacterianos. Os resultados obtidos permitiram verificar a grande potencialidade biotecnológica que um consórcio bacteriano tem, principalmente quanto aos genes e enzimas encontrados nessa pequena comunidade. Desta forma, em nosso laboratório existem dois consórcios bacterianos em manutenção que conseguem sobreviver *in vitro* usando como fonte de carbono apenas material lignocelulótico (bagaço de cana-de-açúcar, palha de milho ou casca de amendoim). Esses consórcios foram isolados de solo sob um monte de bagacilhos de usina de açúcar e álcool ou de solo de cultivo da cana-de-açúcar, com colheita mecanizada e apresentando palhada em decomposição. Os primeiros resultados mostraram a capacidade desses organismos em usarem a biomassa vegetal como fonte de carbono e liberarem glicose no meio de cultivo. Posteriormente, passamos a estudar essa população bacteriana, detectamos por sequenciamento dos genes 16SrRNA e do DNA total dos cultivos que a populações não era maior que 100 organismos e constituída de organismos com capacidade de degradar a celulose e que variavam com o tempo de cultivo. (Constâncio et al, 2020)

Ensaios mais recentes mostraram, por microscopias eletrônicas de varreduras, que a estrutura do bagaço de cana-de-açúcar é alterada pelo cultivo com o consórcio, causando assim a desestruturação parcial da fibra pela ação bacteriana. A análise da estrutura do bagaço, submetido ao cultivo por 5 dias ou 10 dias na presença dos consórcios mostrou a desconstrução da estrutura planar e compacta do bagaço, presença de fissuras e descamações, assim como a aderência de vários tipos bacterianos na superfície da fibra.

Os teores de celulose, hemicelulose e lignina residuais obtidos do cultivo do consórcio com bagaço de cana-de-açúcar diminuíram significativamente após 5 e 10 dias em relação ao tempo 0. Concomitantemente, nesse período, pode-se observar a liberação de glicose no meio decorrente da hidrólise das fibras, o que corresponde a concentrações de aproximadamente 5,6 mg/mL, 1,9 mg/mL para os dias 5 e 10 de cultivo. Quando foi analisado o rendimento do processo de hidrólise através do cálculo do balanço de massa, como proposto por Zhu e

colaboradores (2011), observou-se que o tratamento que apresentou melhor rendimento (36% de eficiência) foi o promovido pelo consórcio com 5 dias de cultivo.

Os resultados obtidos mostram que de um lado observamos a desconstrução de fibras de bagaço de cana-de-açúcar e a diminuição nos teores de celulose, hemicelulose e lignina residuais de cultivos pelas bactérias do consórcio. Do outro lado observamos liberação de glicose no meio de cultivo, portanto tais dados confirmam que os microrganismos do consórcio bacteriano estudado apresentam capacidade de desconstruir a biomassa vegetal e retirar desse material carbono para sua sobrevivência. Desta forma, outra abordagem foi executada nos projetos acima descritos: 1) Qual a população bacteriana que constitui o consórcio bacteriano e qual os organismos envolvidos no processo de desconstrução da lignocelulose? 2) Existe diferença na população bacteriana que está livre no meio de cultivo e a que está aderida às fibras de lignocelulose?

Para se responder tais questões inicialmente a população bacteriana (livre e aderida) foi estudada pelo sequenciamento do gene 16S rRNA em diferentes tempos (5 e 10 dias). Essa avaliação não mostrou significativa diferença entre os isolados detectados aderidos às fibras ou livres no meio de cultivo. A análise dos dados e comparação das sequências no Banco de Dados Silva revelou a presença total de aproximadamente 70 gêneros diferentes com abundância dos gêneros *Caulobacter, Pseudoxanthomonas, Chryseobacterium, Sphingomonas, Rhizobium, Dokdonella, Chitinophaga, Pseudomonas, Flavobacterium*, dentre outras. A análise da capacidade de utilizar a biomassa vegetal desses organismos foi avaliada por dados obtidos na literatura e pode-se verificar que nessa população existe maior proporção de gêneros com organismos degradadores da biomassa, aproximadamente 3:1. Desta forma essa população bacteriana é uma fonte importante de genes/enzimas envolvidas no processo de desconstrução da lignocelulose.

Um total de 52 genomas foram recuperados do metagenoma do consórcio, no qual ~ 46% das espécies não mostraram nenhuma modificação relevante em sua abundância durante as 20 semanas de cultivo, sugerindo um consórcio principalmente estável. Seu repertório de CAZYmes indicou que muitas das espécies mais abundantes são conhecidas por desconstruir a lignina e carregar sequências relacionadas à desconstrução da lignocelulose (Weiss, et al, 2021). Juntos, nossos resultados desvendaram a diversidade bacteriana, o potencial enzimático e a eficácia dessa decomposição da lignocelulose, sugerindo fortemente que esse consórcio pode ser explorado ainda mais em esforços de engenharia e biologia sintética.

Infelizmente resultados iniciais do uso do consórcio para a degradação da biomassa e utilização dos processos fermentativos, tanto simultaneamente como em separado, não mostraram resultados positivos. Esses resultados podem ser decorrentes de diversos fatores, dentre os quais a baixa taxa de hidrólise ou o consumo da glicose liberada pela própria população bacteriana em competição com as leveduras fermentativas ou ainda a liberação de oligômeros de glicose que não são usados pelas leveduras.

Entretanto, os consórcios estudados já mostraram serem capazes de promover a degradação da biomassa (bagaço de cana-de-açúcar) pela ação conjunta de seus diversos microrganismos, os quais possuem genes codificadores de diversas enzimas envolvidas na degradação da biomassa vegetal. Desta forma, esse novo projeto tem como objetivo buscar mais informações metabólicas dessa comunidade identificando os genes que estão sendo expressos, através da análise de metatranscriptoma e determinar se os microrganismos que constituem o consorcio

possuem características outras para serem usados em processos biotecnológicos, em biorremediação ou como promotores de crescimento em plantas. Desta forma o conhecimento a ser adquirido contribuirá para três vertentes de pesquisa: 1) Considerando que os estudos anteriores mostram que os componentes do consórcio bacteriano são espécies que podem ser cultivadas poderemos isolar os microrganismos e recriar um consórcio bacteriano mais eficiente na hidrólise e com menos organismos consumidores de glicose; 2) os isolados bacterianos obtidos separadamente do consorcio poderão ser utilizados individualmente, de acordo com suas características em processos biotecnológicos em diversas áreas da produção; 3) Os resultados de transcriptoma serão importantes para que no futuro se possa realizar a clonagem e expressão de genes das enzimas envolvidas na desconstrução da lignocelulose, que atuam em conjunto e portanto que possam constituir um coquetel enzimático mais eficiente para o aproveitamento da biomassa vegetal dentro do conceito de biorrefinarias. Tanto com relação à degradação da lignina como com relação à liberação dos açúcares que possam ser usados no processo fermentativo.

2. OBJETIVOS

Sequenciar e analisar o metatranscriptoma do consórcio bacteriano obtido a partir de solo sob bagaço de cana-de-açúcar, estudar as relações metabólicas na população bacteriana que está constituindo o consórcio, determinar a expressão das enzimas que estão sendo liberadas para o consórcio conseguir utilizar a lignocelulose como fonte de carbono e conhecer os organismos que estão promovendo a hidrólise da biomassa vegetal e os que estão agindo apenas como consumidores do açúcar liberado, além de quantificar a expressão dos genes associados à degradação da lignocelulose, e avaliar as diferenças significativas de abundância desses genes, nos tempos de amostragem.

3. MATERIAL E MÉTODOS

3.1.Transcriptoma do consórcio

O consórcio bacteriano utilizado para o desenvolvimento deste trabalho foi isolado do solo, de uma área de cultivo comercial de cana-de-açúcar com colheita mecanizada. A área possuía palhada em decomposição e está localizado em propriedade da Usina São Martinho (LAT: 21º19´23.51220” S / LOG: 48º09´12.24864” O / ALT: 534,100 m). A coleta das amostras foi realizada em áreas aleatórias do solo com profundidade de 0 a 20 cm, sempre em zigue-zague, após a coleta, as amostras foram reunidas e homogeneizadas, tornando assim uma única amostra, a qual foi utilizada para o isolamento do consórcio (CONSTANCIO et Al, 2020).

O estoque dos consórcios da primeira e decima nona semana de obtenção foram usados como inoculo (1000µl) para o cultivo em meio BHB (BUSHNELL, 1941) constituído (g.L-1) por: sulfato de magnésio, 0,2; cloreto de cálcio, 0,02; fosfato monopotássico, 1,0; fosfato de amônio, 0,02, nitrato de potássio, 1,0 e cloreto férrico, 0,05, contendo como fonte de carbono 2,0 mg/mL de bagaço de cana-de-açúcar lavado, seco, triturado e peneirado (32 mesh) a 0,5 mm. As culturas foram mantidas a temperatura de 30ºC e agitação de 120 rpm por 5 e 10 dias, para ambos os tempos foram realizados cultivos em quadruplicata. Essas culturas foram usadas para a extração do RNA e como controle negativo usou-se meio de cultivo, sem inoculo bacterianos, mas com a mesma agitação e tempo de cultivo dos meios inoculados

A extração do RNA das culturas acima descritas foi realizada usando o RNeasy PowerSoil Total RNA Kit (QIAGEN) a partir do sedimento de 10mL do material cultivado. O restante do material cultivado foi centrifugado e o sobrenadante e sedimentos congelados. O sobrenadante será usado para análise de glicose livre e o sedimento para avaliação do teor de celulose, hemicelulose e lignina. Os resultados das extrações de RNA estão observados na **Tabela 1**, a qualidade (RIN) e quantidade do material extraído foi realizada pela análise em Bioanalyzer e Qubit respectivamente. O controle sem bactéria serviu apenas para demonstrar a baixa contaminação com material vegetal do meio de cultivo. Para o sequenciamento foram usadas as 4 amostras de cada tempo de cultivo. O preparo das bibliotecas foi realizado utilizando o Kit Illumina Stranded Total RNA Pre ligation with Ribo-Zero Plus (1-100ng, Illumina). A clusterização e sequenciamento (NextSeq 2x100pb) foi realizado na NGS Soluções Genomicas.

Tabela 1. Qualificação e quantificação do RNA extraído das amostras de 10mL do cultivo dos consórcios da segunda e vigésima semanas por 5 e 10 dias de cultivo

Amostras (semana 2)	**Concentração RNA (ng/µl)**	**RIN**	**Amostras (semana 20)**	**Concentração RNA (ng/µl)**	**RIN**
5 dias			**10 dias**		
controle	0,05	0,0	controle	6,3	8,1
replica1	510	0,0	replica1	265	0,0
replica 2	320	0,0	replica 2	166	7,5
replica 3	174	7,0	replica 3	285	7,5
replica 4	137	7,0	replica 4	356	6,0
10 dias			**10 dias**		
controle	7	0,0	controle	0,0	0,0
replica1	177	7,2	replica1	100	0,0
replica 2	157	7,2	replica 2	114	8,5
replica 3	156	7,2	replica 3	153	8,8
replica 4	152	7,2	replica 4	45	8,3

3.2. Análise *in silico*

3.2.1. Processamento dos dados

Para a aferição inicial dos dados sequenciados, utilizou-se o programa "FastQC" (v.0.11.9; Andrews, 2010), o qual forneceu relatórios da distribuição de tamanhos e qualidades das sequências. Os relatórios foram sumarizados com o programa MultiQC (v.1.17 ; Ewels et al., 2016). Em seguida, o programa "Atropos" (v.1.1.24; Didion et al., 2017) foi utilizado para remoção de adaptadores residuais de sequenciamento. Devido a utilização da estratégia "paired-end", implementou-se dois processamentos com o "Atropos", sendo o primeiro baseado no modo "insert", o qual checa por adaptadores nas sobressalências excedentes da sobreposição dos pares, e o segundo no modo "adapter", busca padrão por alinhamento entre a sequência do adaptador e a sequência alvo, de modo semi-global. Para a remoção de sequências e bases de baixa qualidade, utilizou-se o programa"PRINSEQ++ " (v.1.2; Cantu et al., 2019), onde removeu-se bases cuja janela composta por 3 bases consecutivas ("-trim_qual_window 3"), tivessem valor de qualidade Phred (Q) inferior a 30 ("-trim_qual_right 30"), bem como sequências completas cuja média de qualidade fosse inferior a Q30 ("-min_qual_mean 30") ou menores que 35 bases ("-min_len 35"). Sequências cujos seus respectivos pares foram perdidos durante as etapas de controle de qualidade ("singletons"), foram reservadas e utilizadas para a etapa de montagem.

3.2.2. Montagem do metatranscriptoma e determinação de expressão

As sequências pré-processadas foram agrupadas e montadas com o montador Trinity (v.2.15.1;Grabherr et al., 2011) usando os valores padrão. Para remover transcritos pequenos, apenas contigs com mais de 200pb foram selecionados para inclusão no metatranscriptoma de referência ("--min_contig_length 200"). As leituras de todas as amostras foram mapeadas para o metatranscriptoma de referência, e a expressão foi avaliada usando scripts de utilidade do Trinity ("align_and_estimate_abundance.pl" e "abundance_estimates_to_matrix.pl") conforme recomendado pelos autores do programa (https://github.com/ trinityrnaseq/trinityrnaseq/wiki/).

3.2.3. Identificação de transcritos diferencialmente expressos

Os transcritos diferencialmente expressos (DE) foram identificados com o programa "DESeq2" (v. 1.34.0; Love et al., 2014) implementado via script de utilidade do Trinity: "run_DE_analysis.pl". Como limiar para ser considerado DE, foi estabelecido um q-valor (p-valor ajustado) de 0.01. Adicionalmente, filtrou-se DEs de baixa diferença de expressão, estabelecendo um valor mínimo de Log_2 Fold-Change de ± 2.

3.2.4. Anotação funcional

A anotação funcional do metatranscriptoma foi conduzida usando a ferramenta eggNOG-mapper (v.2.1.12; Huerta-Cepas et al., 2017) em relação ao banco de dados de proteínas eggNOG (v.5.0). Os dados de entrada foram projetados como metagenoma, utilizando o parâmetro "--itype metagenome", e a predição de genes foi realizada com o Prodigal ("--genepred prodigal"). Os genes preditos foram pesquisados no banco de dados

usando o alinhador de sequências DIAMOND ("-m diamond") com a configuração de maior sensibilidade ("--sensmode ultra-sensitive"). Somente alinhamentos significativos, conforme indicado por um limiar de "-- evalue" de 0,001, foram considerados uma anotação bem-sucedida. Essa estratégia de anotação abrangeu vários repositórios funcionais, incluindo COG (Clusters de Grupos Ortólogos), GO (Gene Ontology), KO (KEGG Orthology), CAZy (Carbohydrate-Active enZYmes) e Pfam.

3.2.5. Análise de Componentes Principais e representações gráficas

Para reduzir a dimensionalidade e definir os componentes principais no que se refere à variação dos perfis de expressão das amostras, utilizou-se uma Análise de Componentes Principais (PCA), implementada pela função " PCA" do pacote R " FactoMineR" (v.2.7; Lê et al., 2008). O mapa de calor foi gerado com o pacoteR "pheatmap" (v.1.0.12; Kolde, 2019) e os demais, com o pacote R "ggplot2" (v.3.4.4; Wickham, 2016).

4. RESULTADOS E DISCUSSÃO

4.1.Processamento dos dados

Para as análises *in silico* das sequencias obtidas e algumas comparações os resultados foram divididos em tabelas e visualizações relacionadas as etapas de: processamento dos dados, montagem e anotação dos transcritos, e análise de expressão diferencial. Para cada resultado foi feita uma breve explicação da sua importância e os principais destaques observados.

A Tabela 2 fornece uma visão detalhada da contagem de leituras de sequenciamento de RNA para cada amostra em estágios críticos de pré-processamento. Cada coluna representa uma etapa de processamento distinta, começando com as leituras brutas obtidas a partir do sequenciamento e concluindo com a classificaçãodas leituras em categorias de RNA não-ribossômico e RNA ribossômico. As leituras não-ribossômicas foram as utilizadas para a montagem e todas as etapas subsequentes.

Na etapa inicial de pré-processamento de dados, um total de 1.273.001.119 leituras foram geradas. Após a aplicação de medidas de controle de qualidade, 94,02% das leituras foram mantidas, resultando em1.196.915.695 leituras de alta qualidade. Apesar da depleção de rRNA, apenas 20,23% (257.553.582 leituras) foram identificadas como não-rRNA e consideradas adequadas para análises subsequentes. Esse pré-processamento garantiu a remoção de contaminação de rRNA indesejado e o uso de dados de alta qualidade para a montagem do transcriptoma e análises adicionais.

A Figura 1 ilustra a qualidade das amostras de RNA, avaliada pelos escores Phred em cada posição da sequência, tanto antes (A) quanto depois (B) das etapas de pré-processamento. O gráfico fornece uma comparação visual da qualidade da sequência, mostrando a qualidade inicial (A) e a final (B), destacando melhorias pós-processamento. A qualidade dos dados de sequenciamento de RNA estava inicialmente boa logo após o sequenciamento, mas as etapas de pré-processamento melhoraram ainda mais ligeiramente a qualidade dos dados. Isso é especialmente notado na porção inicial das sequências. Montagem, mapeamento e quantificação dos

transcritos.

Na Tabela 2, que fornece um resumo da análise metatranscriptômica, componentes-chave como medidas da montagem e anotação funcional do metatranscriptoma, são destacados. A montagem do metatranscriptoma envolve a reconstrução de transcritos de RNA a partir de dados de sequenciamento para compreender a diversidade da expressão gênica. A anotação funcional associa informações biológicas aos transcritos identificados, revelando suas funções e vias potenciais. Cada seção na Tabela 3 contém estatísticas específicas relacionadas a esses componentes, oferecendo insights sobre a qualidade e importância dos resultados da análise metatranscriptômica.

Foram montados um total de 268.898 transcritos, representando isoformas relacionadas a 241.931 genes. Notavelmente, 16,55% desses transcritos apresentaram níveis de expressão relevantes, com mais de 1 transcrito por milhão (TPM) em 2 ou mais amostras, enquanto a maioria, correspondendo a 83,45%, exibiu baixa abundância e prevalência nas amostras. Quanto à anotação, uma parte substancial dos transcritos, totalizando 86,67%, foi efetivamente anotada. Essa anotação incluiu associações com Clusters de Grupos Ortólogos (COG) para 82% dos transcritos, enquanto 24,81% foram vinculados a termos do Gene Ontology (GO). Além disso, 51,53% dos transcritos receberam anotações de KEGG Orthology (KO), e 2,02% foram classificados com base em Carbohydrate-Active enZYmes (CAZy).

Tabela 2. Resumo da contagem de leituras em cada etapa de pré-processamento após o sequenciamento de RNA. "Leituras brutas": Contagem inicial de leituras após o sequenciamento, "c/ Adaptador": Leituras que contêm adaptadores, mantidas após o processamento com Atropos, "Filt. Qualidade": Leituras que atendemaos padrões de qualidade após o PRINSEQ, "non-rRNA": Leituras de RNA não-ribossômico após identificação pelo SortMeRNA, "rRNA": Leituras de RNA ribossômico identificadas pelo SortMeRNA. 02W e 20W (consorcio da semana 2 e 20) e 05D e 10D (tempo de cultivo 5 e 10 dias).

Amostra	Leituras Brutas	c/ Adaptadores	Filt. Qualidade	non-rRNA	rRNA
02W05D.1	83.708.798	83.660.023	79.200.392	10.280.184	68.920.208
02W05D.2	74.513.123	74.477.582	70.590.633	20.774.549	49.816.084
02W05D.3	72.720.175	72.594.824	68.625.267	9.551.710	59.073.557
02W05D.4	76.080.252	74.781.804	70.346.974	9.705.178	60.641.796
02W10D.1	80.465.597	80.336.938	75.942.634	11.194.532	64.748.102
02W10D.2	75.177.395	75.013.651	70.433.787	11.605.572	58.828.215
02W10D.3	108.338.589	107.986.455	101.719.544	21.072.055	80.647.489
02W10D.4	80.088.516	79.634.889	74.084.726	13.812.489	60.272.237
20W05D.1	74.448.281	74.142.352	70.090.721	19.279.791	50.810.930
20W05D.2	76.785.010	76.086.809	71.748.425	25.095.501	46.652.924
20W05D.3	82.427.551	81.982.509	77.314.087	20.144.408	57.169.679
20W05D.4	72.224.103	72.083.044	68.615.891	15.862.956	52.752.935
20W10D.1	81.397.512	81.211.955	76.962.791	20.830.813	56.131.978
20W10D.2	82.086.517	81.849.031	77.782.618	16.175.377	61.607.241
20W10D.3	74.050.863	73.958.041	70.402.683	14.904.550	55.498.133
20W10D.4	78.488.837	77.149.516	73.054.522	17.263.917	55.790.605

Figura 1. Comparação dos escores Phred por Posição em amostras de RNA pré (A) e pós (B) processamentos.

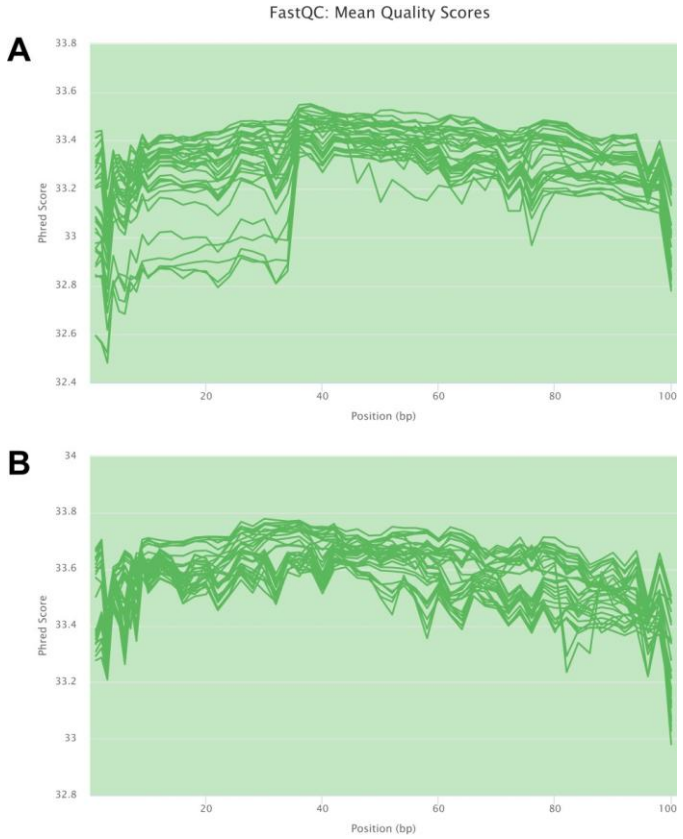


Tabela 3. Resumo das estatísticas de montagem e anotação funcional do metatranscriptoma.

Aspecto	Quantificação
Geral	
Número de genes	241.931
Número de transcritos	268.898
Número de transcritos com >1 TPM* em 2 ou mais amostras	44.512
Montagem: todas as Isoformas	
Contig: N50**	1.178
Tamanho médio de Contig	738
Total de bases montadas	198.334.670
Montagem: apenas Isoformas mais longas por Gene	
Contig: N50	998
Tamanho médio de Contig	675
Total de bases montadas	163.354.963
Anotação	
Transcritos anotados	233.059
Anotados com Clusters of Orthologous Groups (COGs) of proteins	220.494
Anotados com Gene Ontology (GO)	66.702
Anotados com KEGG Orthology (KO)	138.572
Anotados com Carbohydrate-Active enZYmes (CAZy)	5.426

*TPM: Transcritos por Milhão. ** N50: A sequência de comprimento em que 50% do comprimento total das sequênciasno transcriptoma está contido em sequências desse tamanho ou maiores

O mapeamento de reads alinha as leituras de sequenciamento a uma referência (aqui é o metatranscriptoma), permitindo a quantificação da expressão gênica. A Tabela 4 apresenta um detalhamento das estatísticas de mapeamento para leituras de RNA, destacando diferentes categorias de saídas do processo de mapeamento. A coluna "Entrada" representa o número total de leituras inseridas no processo de mapeamento, enquanto a coluna "Descartadas" considera as leituras que foram descartadas devido a razões como múltiplos mapeamentos, mapeamentos discordantes ou não mapeamento. "Mapeadas (Qual. Baixa)" indica mapeamentos que foram descartados devido à baixa qualidade de mapeamento, e "Mapeadas (Qual. Alta)" inclui mapeamentos de alta qualidade retidos para subsequente quantificação. A coluna "Mapeamentos (%)" fornece a porcentagem de leituras que foram mapeadas e quantificadas com sucesso, oferecendo uma visão sobre a eficiência do processo de mapeamento.

O mapeamento das leituras não-rRNA revelou que 66,98% das 257.553.582 leituras foram mapeadas com sucesso e alta qualidade. Curiosamente, houve variações nas taxas de mapeamento entre as amostras, com amostras da "semana 2" apresentando uma taxa média de mapeamento de 59,01%, enquanto as amostrasda "semana 20" exibiram uma taxa média de mapeamento de 72,73%. Essas diferenças sugerem algum tipo de variação ou viés temporal no transcriptoma. Os mapeamentos de alta qualidade foram subsequentemente usados para a quantificação da expressão com a ferramenta Salmon.

Tabela 4. Estatísticas de mapeamento para leituras de sequenciamento de RNA, divididas por amostra.

Amostra	Entrada	Descartadas	Mapeadas (Baixa Qual.)	Mapeadas (Alta Qual.)	Mapeamentos (%)
02W05D.1	10.280.184	2.090.812	1.358.096	6.831.276	66,45
02W05D.2	20.774.549	4.295.994	2.735.062	13.743.493	66,16
02W05D.3	9.551.710	3.414.127	817.963	5.319.620	55,69
02W05D.4	9.705.178	3.021.626	1.321.075	5.362.477	55,25
02W10D.1	11.194.532	3.889.618	525.974	6.778.940	60,56
02W10D.2	11.605.572	3.869.629	780.049	6.955.894	59,94
02W10D.3	21.072.055	8.283.445	925.544	11.863.066	56,30
02W10D.4	13.812.489	6.041.088	900.321	6.871.080	49,75
20W05D.1	19.279.791	2.512.821	2.938.936	13.828.034	71,72
20W05D.2	25.095.501	3.530.412	3.861.902	17.703.187	70,54
20W05D.3	20.144.408	2.482.341	3.332.953	14.329.114	71,13
20W05D.4	15.862.956	2.007.145	2.846.830	11.008.981	69,40
20W10D.1	20.830.813	2.980.705	2.623.515	15.226.593	73,10
20W10D.2	16.175.377	2.338.741	1.747.474	12.089.162	74,74
20W10D.3	14.904.550	2.015.384	1.755.623	11.133.543	74,70
20W10D.4	17.263.917	1.460.924	2.349.745	13.453.248	77,93

4.2.Anotação funcional

O mapa de calor (Figura 2) fornece uma representação intuitiva dos níveis de expressão de transcritos e suas relações, auxiliando na identificação de padrões e agrupamentos transcricionais nos dados. Foi observada uma separação notável entre as amostras da "semana 2" e da "semana 20", indicando padrões de expressão distintos. As amostras da "semana 2" apresentaram um perfil de expressão mais amplo, com uma média de 74.913,87 transcritos expressos, enquanto as amostras da "semana 20" exibiram uma média de 18.585,87 transcritos expressos (dados presentes na aba "***Transcritos***" do arquivo "***Tabelas.xlsx***". No entanto, é importante observar que as amostras da "semana 2" do "dia 5" pareciam um tanto inconsistentes e não agrupadas, sugerindo uma influência potencial de viés de amostragem ou sequenciamento.

Em contraste, as amostras da "semana 20" revelaram padrões de expressão distintos com uma clara separação entre as amostras do "dia 5" e do "dia 10". Essa divisão resultou em dois clusters (blocos) distintos:um contendo transcritos expressos em ambos os dias (localizados no canto superior direito do gráfico) e outro predominantemente expressando transcritos no dia 10 (situado logo abaixo do cluster comum).

O gráfico de PCA (Figura 3) revela a estrutura dos dados, auxiliando na identificação de agrupamentos de amostras e fatores potenciais que contribuem para a variabilidade observada na expressão de transcritos. A Análise de Componentes Principais (PCA) dos perfis de transcritos explicou efetivamente a variabilidade do conjunto de dados, com o PCA1 e o PCA2 representando 55,07% e 13,2% da variância total, respectivamente (total: 68,27%). Os resultados do PCA coincidiram com a análise do mapa de calor, confirmando padrões de expressãodistintos entre as amostras de diferentes semanas. As amostras da "semana 20" apresentaram uma forte coesão, enquanto as amostras da "semana 2" mostraram dois agrupamentos distintos no "dia 5" e um perfil mais consistente no "dia 10".

O banco de dados COG categoriza proteínas de diversos organismos em grupos funcionais com base em suas relações ortólogas, oferecendo insights sobre a diversidade funcional de genes e proteínas. Neste histograma, as cores representam diferentes categorias do COG, com a frequência de transcritos atribuídos a cada categoria. Assim, o banco de dados COG auxilia na compreensão dos papéis funcionais prevalentes em diferentes amostras.

O perfil geral de anotação do metatranscriptoma com o COG (Clusters de Grupos Ortólogos) (Figura 4A) revelaram que 16,54% dos transcritos foram categorizados como "Função desconhecida" (S), destacando a necessidade de experimentos adicionais de validação e identificação funcional. As principais categorias de interesse incluíam "Transporte e metabolismo de aminoácidos" (E) (8,39% dos transcritos anotados), "Produção e conversão de energia" (C) (7,13%) e "Transporte e metabolismo de carboidratos" (G) (6,69%).

A análise dos perfis de anotação do COG por amostra (Figura 4B) revelou diferenças significativas entre as diferentes semanas: Primeiramente, a categoria "Função desconhecida" (S) exibiu uma frequência relativa média de 22,25% nas amostras da "semana 2", contrastando com reduzidos 4,64% observados nas amostras da "semana 20". A categoria "Mecanismos de transdução de sinal" (T) apresentou uma discrepância substancial, com as amostras da "semana 2" representando relativamente baixos 5,04%, enquanto as amostras da "semana20" exibiam

uma média consideravelmente maior, de 61,55%. Por fim, a categoria "Transporte e metabolismo de aminoácidos" (E) tinha uma média de 7,54% nas amostras da "semana 2", mas diminuiu para 1,98% nas amostras da "semana 20".

Figura 2. Visualização em mapa de calor de transcritos, destacando os padrões de expressão. Os dados foram padronizados com escores Z para destacar variações entre as amostras e o agrupamento hierárquicofoi realizado usando o algoritmo "Ward" para agrupar transcritos (linhas) e amostras (colunas) semelhantes.Para reduzir o ruído, transcritos com baixa abundância e prevalência (com base em um mínimo de >1 TPM em 2 ou mais amostras) foram desconsiderados.

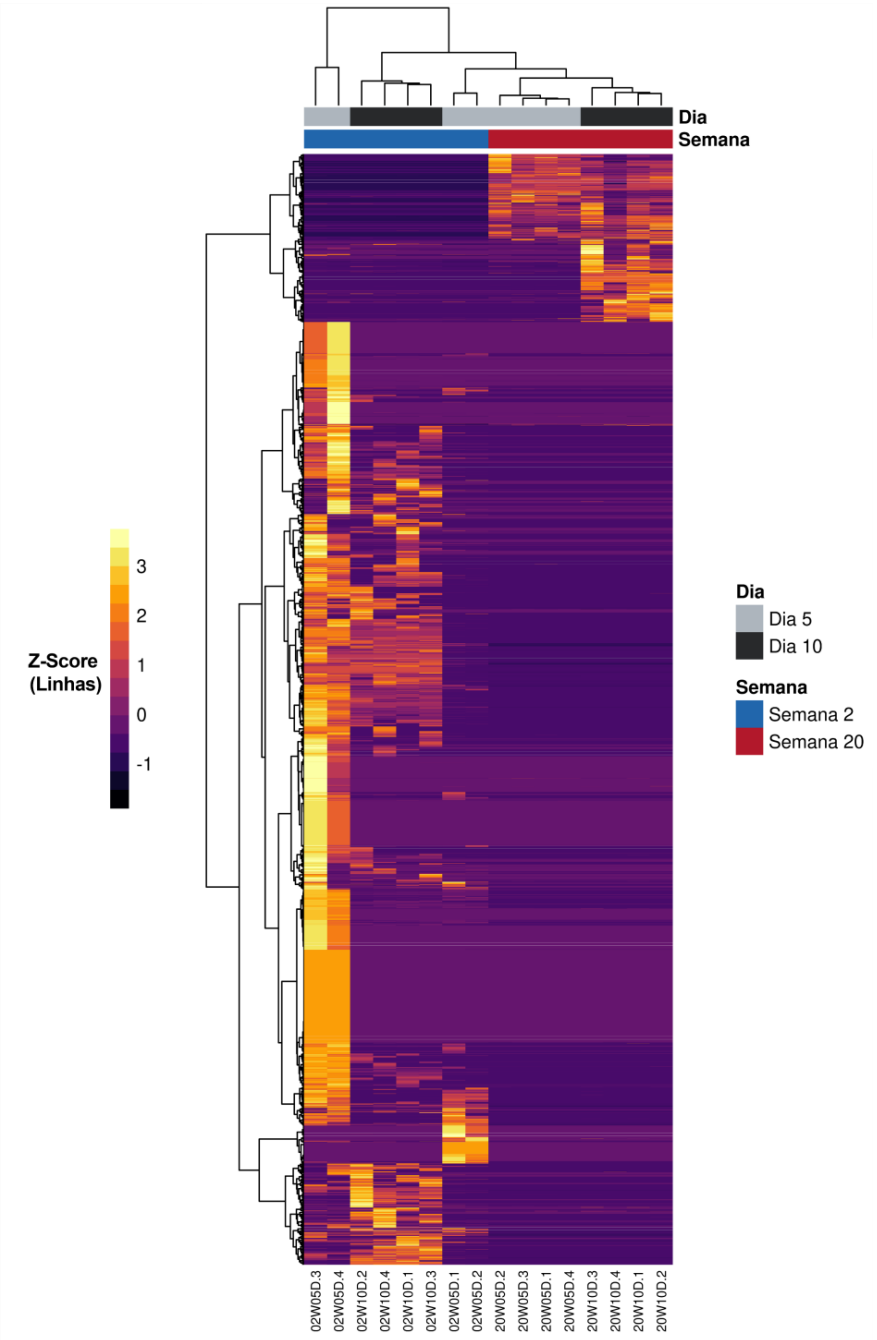


Figura 3. Representação da Análise de Componentes Principais (PCA) de transcritos para demonstrar o agrupamento de amostras, com visualizações apresentadas a partir das perspectivas dos fatores: "Semana" (A) e "Dia" (B). Transcritos com baixa abundância e prevalência (com base em um mínimo de >1 TPM em 2 ou mais amostras) foram desconsiderados.

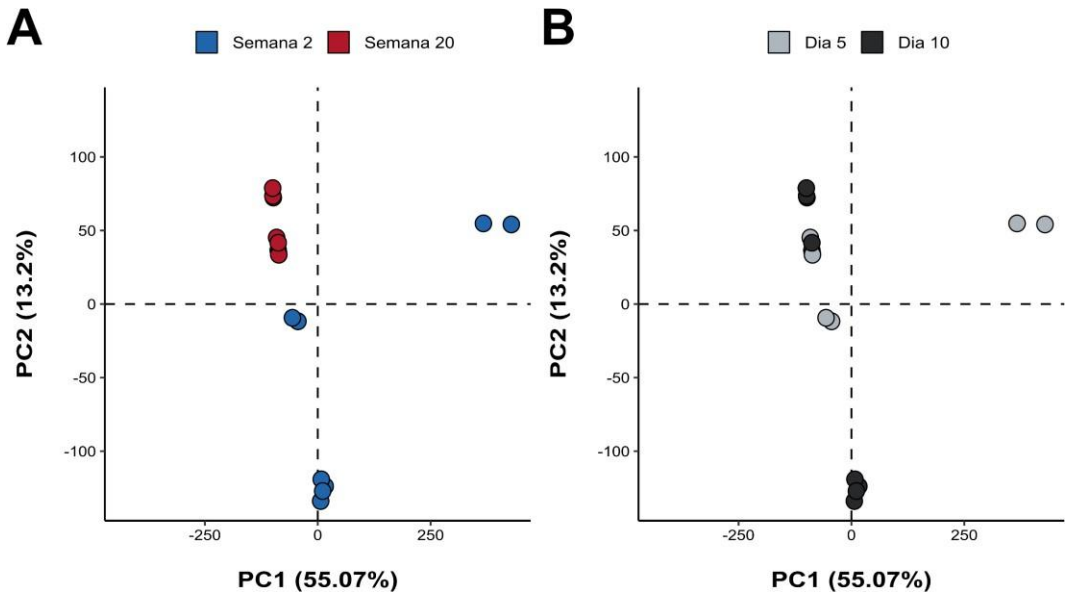
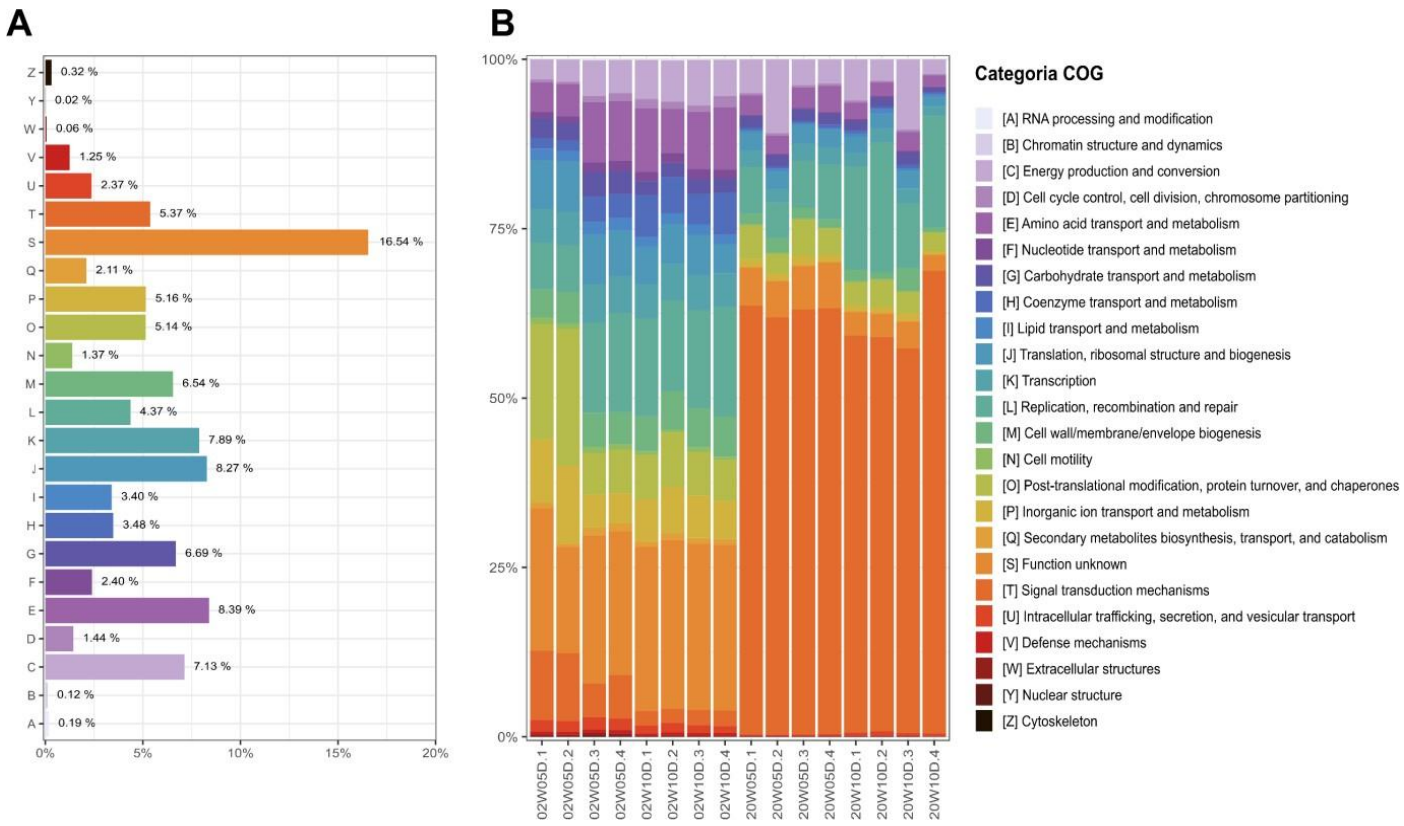


Figura 4. Ilustra a distribuição das anotações do COG ("Clusters of Orthologous Groups") ao longo do metatranscriptoma (A) e das amostras (B).



O banco de dados KEGG Orthology categoriza genes e proteínas em categorias funcionais, ajudando a elucidar seus papéis em várias vias e processos biológicos. Essa análise contribui para nossa compreensão da diversidade funcional e das vias metabólicas potenciais presentes no metatranscriptoma e podem ser vistas na Figura 5.

O perfil de anotação KEGG para o metatranscriptoma forneceu uma visão geral do perfil funcional, com uma presença notável de transcritos associados ao metabolismo de carboidratos (15,26%), metabolismo de aminoácidos (10,99%) e metabolismo de energia (6,90%).

A categoria de vias metabólicas do KEGG abrange vias e processos biológicos, concentrando-se principalmente no metabolismo celular. O KEGG Metabolismo engloba uma ampla gama de reações bioquímicas envolvidas na síntese, degradação e transformação de diversas moléculas, incluindo carboidratos, lipídios, aminoácidos e ácidos nucleicos.

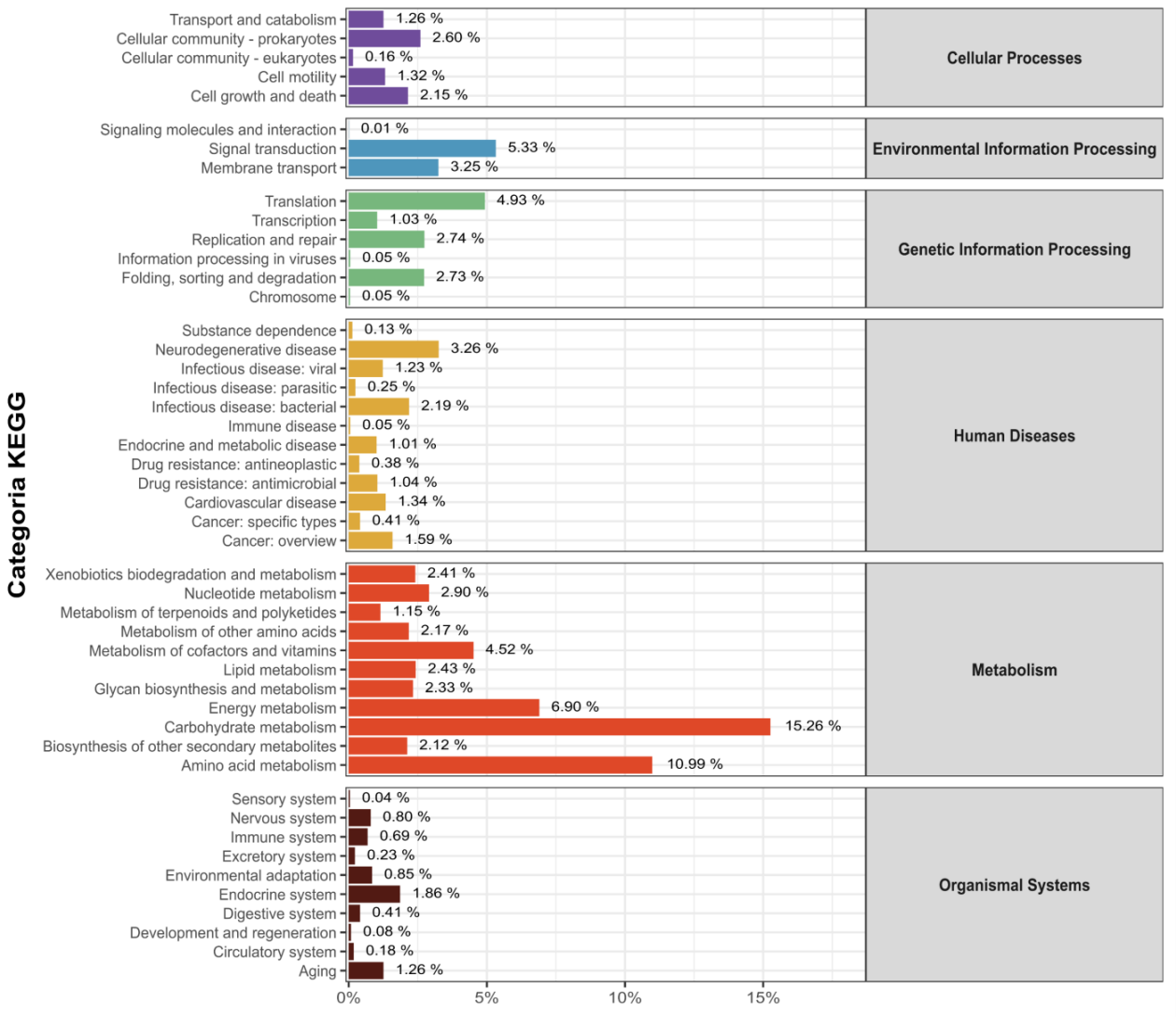
Explorando à fundo a categoria de Metabolismo do KEGG, foram observadas variações significativas nos perfis das amostras entre as semanas (Figura 6). O metabolismo de aminoácidos representou, em média, 29,87% nas amostras da "semana 2" e 21,23% nas amostras da "semana 20". O metabolismo de carboidratos apresentou uma média menor de 25,46% na "semana 2" em comparação com 39,72% na "semana 20". Além disso, a biossíntese e o metabolismo de glicanos corresponderam a 2,40% nas amostras da "semana 2" e 7,89% nas

amostras da "semana 20". Dentro das amostras da "semana 2", nota-se também diferenças entre amostras do "dia 5" e do "dia 10". Especificamente, o metabolismo de aminoácidos, que representou 23,26% dos transcritos nas amostras do "dia 5", em contraste com 36,47% nas amostras do "dia 10".

O banco de dados CAZy é um recurso especializado dedicado à classificação de enzimas envolvidas no metabolismo e na modificação de carboidratos, glicoproteínas e glicolipídios. Essas enzimas desempenham papéis fundamentais em diversos processos biológicos, incluindo síntese de parede celular, metabolismo de energia e degradação de carboidratos complexos, como celulose, hemicelulose e quitina.

Na Figura 7A, a classe "Hidrolase de Glicosídeos" (GH) foi a mais representada, compreendendo 87,58% dos transcritos anotados com o CAZy. Em seguida, encontra-se a classe "GlicosilTransferase" (GT) com 7,83%, "Esterase de Carboidratos" (CE) com 2,22%, "Liase de Polissacarídeos" (PL) com 1,11%, "Módulosde Ligação a Carboidratos" (CBM) com 0,68%, e "Atividades Auxiliares" (AA) com 0,59%. A Figura 7B revelou diferenças na prevalência de GH nas amostras, entre as semanas, com 74,55% na "semana 2" e um aumento notável para 95,39% na "semana 20". Dentro da "semana 2", as amostras do "dia 5" tiveram uma média de 84,01% de representação de GH, enquanto as amostras do "dia 10" tiveram uma média menor, de 65,09%. Além disso, foram observadas grandes variações nas proporções da categoria GT, com 21,30% na "semana 2"e 2,81% na "semana 20". Ainda nessa classe, dentro da "semana 2", as amostras do "dia 5" tiveram uma médiade 11,81%, enquanto as amostras do "dia 10" tiveram uma média mais alta, de 30,79%.

Figura 5. Ilustra a distribuição das classificações do KEGG Orthology (KO) ao longo do metatranscriptoma.



Essa visualização fornece informações interessantes sobre processos de coexpressão e regulação das famílias CAZy dentro do conjunto de dados metatranscriptômico, auxiliando na exploração das capacidades de utilização de carboidratos e no potencial enzimático. O mapa de calor (Figura 8) apresentou padrões um tanto similares ao perfil geral (Figura 2), com uma clara separação entre as semanas. As amostras da "semana 20" são mais coesas, enquanto as amostras da "semana 2" mostram uma separação distinta, especialmente entre as amostras do "dia 5". Na "semana 20", são observadas diferenças mínimas entre as amostras do "dia 5" e do "dia 10". Nenhum cluster específico de famílias se destacou proeminentemente nessa comparação. Em contraste, as amostras da "semana 2" exibem coesão entre as amostras do "dia 10", mas as amostras do "dia 5" apresentam perfis distintos e espalhados. É importante notar que os clusters observados no mapa de calor podem ser influenciados por transcritos altamente expressos anotados em múltiplas famílias CAZy. Os genes podem ter uma relação de um-para-muitos de acordo com a anotação do CAZy, contribuindo para a complexidade dos padrões

de agrupamento em nível de família.

Figura 6. Ilustra a distribuição das classificações do KEGG Orthology (KO) na categoria de vias metabólicas, ao longo das amostras.

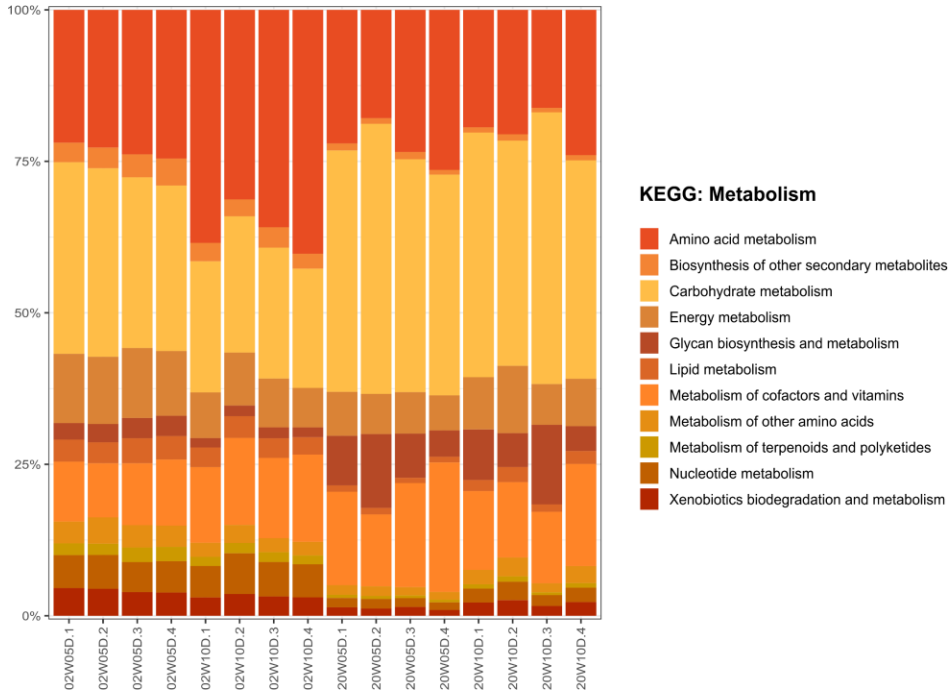


Figura 7. Ilustra a distribuição das anotações do CAZy (Carbohydrate-Active enZYmes) ao longo do metatranscriptoma (A) e das amostras (B).

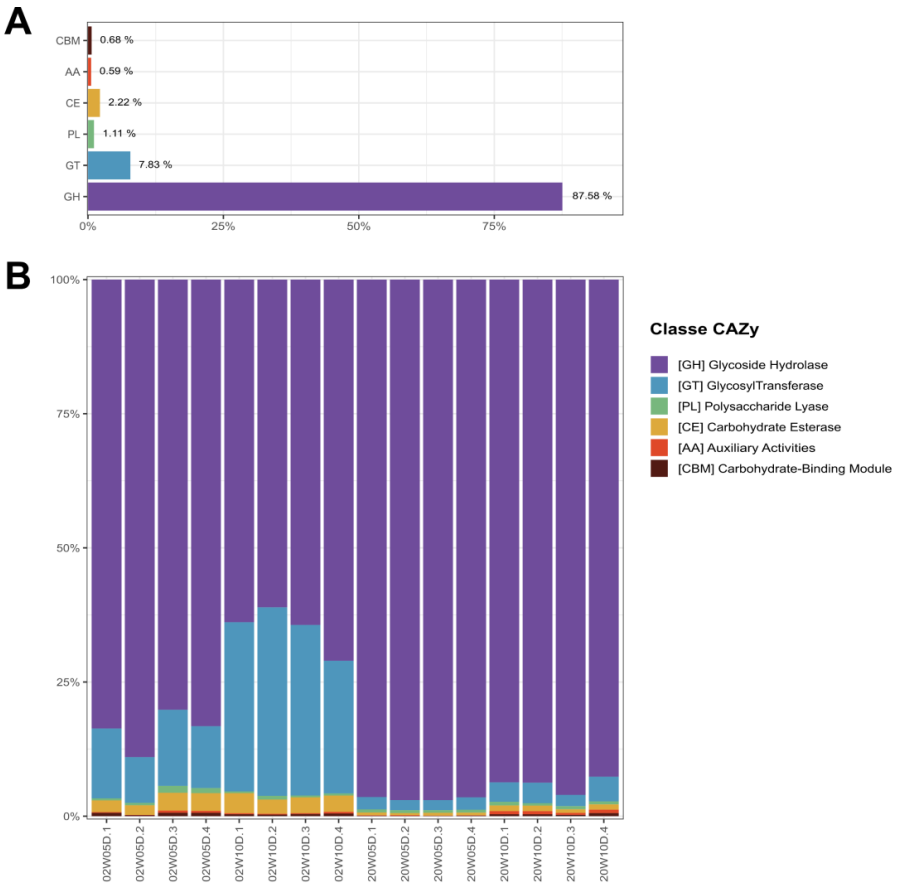
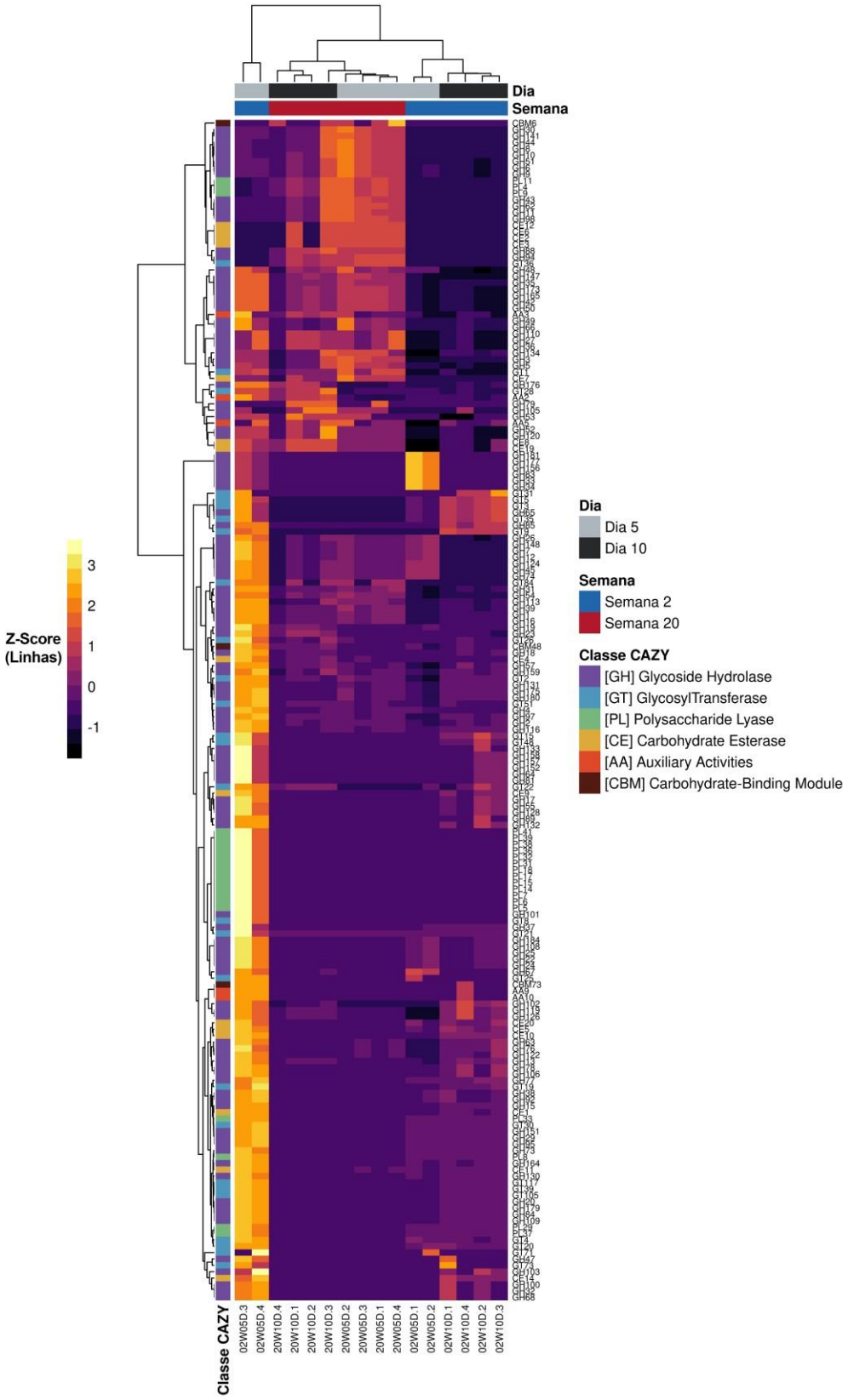


Figura 8. Visualização em mapa de calor de transcritos anotados com o CAZy, destacando padrões de abundância de famílias CAZy. Os dados foram escalonados com escores Z para destacar variações entre as amostras, e o agrupamento hierárquico foi realizado usando o algoritmo "Ward" para agrupar famílias CAZy com expressão similar (linhas) e amostras (colunas) semelhantes.



4.3.Expressão diferencial

Os transcritos DE são aqueles que exibem diferenças estatisticamente significativas na expressão entre as condições definidas. A Tabela 5 fornece uma visão geral do total de transcritos DE, destacando os genes que são diferencialmente expressos no contexto das "semanas" e dos "dias". Os gráficos de "Volcano" (Figura 9) fornecem uma visão abrangente dos transcritos DE ao representar simultaneamente sua intensidade da diferença (Fold-Change) e significância estatística (p-valores). Os pontos no gráfico representam transcritos individuais, facilitando a identificação da significância e magnitude das mudanças na expressão em diferentes condições.

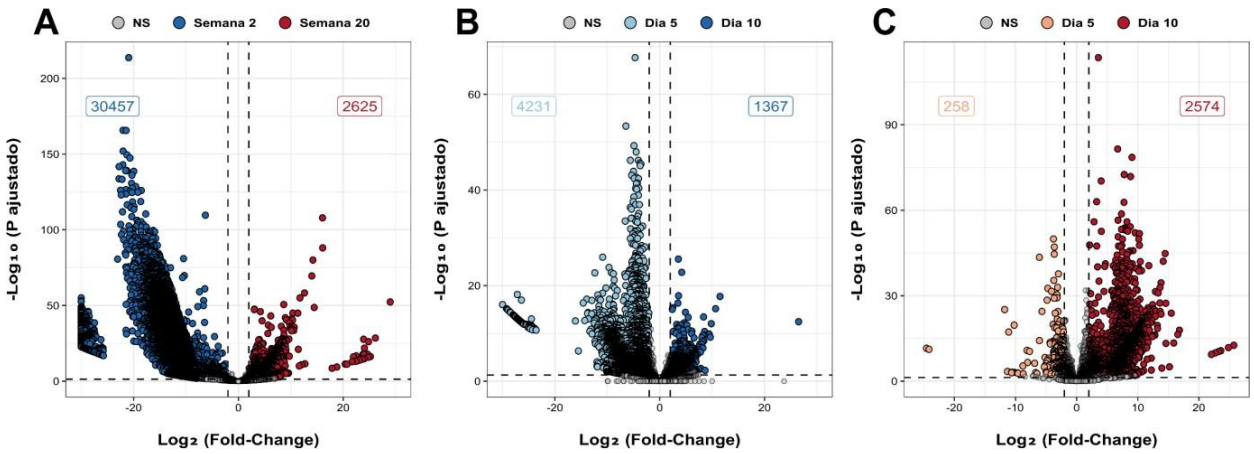
Foi observado um número substancial de transcritos diferencialmente expressos (DE) na comparação entre as semanas (Figura 9A), totalizando 33.082 transcritos DE. Dentre esses, 30.457 transcritos foram significativamente mais expressos e/ou exclusivos da "semana 2". Essa observação está alinhada com os padrões de expressão de transcritos observados no perfil do mapa de calor (Figura 2), confirmando os perfis de expressão distintos entre as duas semanas. Nas comparações dentro das semanas, foram observados comportamentos contrastantes. Na "semana 2", o "dia 5" apresentou o maior número de transcritos DE (Figura9B). Por outro lado, na "semana 20", o maior número de transcritos DE foi associado ao "dia 10" (Figura 9C), sugerindo dinâmicas de expressão distintas durante a linha do tempo experimental dentro de cada semana.

A Tabela 5 e a Figura 9 são mutualmente exclusivas, já que apresentam essencialmente a mesma informação, ainda que os gráficos forneçam uma visão mais profunda. Uma exploração mais profunda dos transcritos DE deverá ser realizada em um segundo momento. Contudo, a tabela 5 contém os top (até 20) transcritos DE, de cada condição e contraste,que foram anotados com o CAZy. O critério de ordenação foi a expressão total. Logo, esses seriam os transcritos que: (1.) Diferiram entre as condições; (2.) foram anotados com o CAZy; e (3.) Possuem nível de expressão relevante nas amostras. Tais características fazem desses os melhores candidatos para explorações futuras e validações via RT-qPCR, só restando comparar qual condição possuiu melhor desempenho nos testes de liberação de glicose e degradação do material lignocelulósico *in vitro*.

Tabela 5. Resumo dos transcritos Diferencialmente Expressos (DE) de acordo com a análise DESeq2. As comparações foram realizadas nos níveis de expressão dos transcritos entre as "Semanas" e os "Dias" (com subdivisão por semana).

Fator de contraste	Condição 1		Condição 2		Total
	Grupo	# de DEs	Grupo	# de DEs	
Semanas	Semana: 2	30.457	Semana: 20	2.625	33.082
Dias					
Semana: 2	Dia: 5	4.231	Dia: 10	1.367	5.598
Semana: 20	Dia: 5	258	Dia: 10	2.574	2.832

Figura 9. Gráficos do tipo "Volcano" que ilustram os transcritos Diferencialmente Expressos (DE) identificados por meio da análise DESeq2. Os gráficos representam análises comparativas, incluindo: (A) comparações entre "Semanas", (B) comparações de "Dias" dentro da "Semana 2" e (C) dentro da "Semana 20". NS = Não-significativo.



5. Referências Bibliográficas

Andrews, S. (2010). *FastQC: a quality control tool for high throughput sequence data*. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.

Cantu, V. A., Sadural, J., & Edwards, R. (2019). PRINSEQ++, a multi-threaded tool for fast and efficient quality control and preprocessing of sequencing datasets. *PeerJ Preprints*, *7*, e27553v1.

Didion, J. P., Martin, M., & Collins, F. S. (2017). Atropos: Specific, sensitive, and speedy trimming of sequencing reads. *PeerJ*, *5*, e3720. https://doi.org/10.7717/peerj.3720

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047–3048.

Grabherr, M. G., Haas, B. J., 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., … Regev, A. (2011). Full-length transcriptome assembly from RNA-

Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–652. https://doi.org/10.1038/nbt.1883

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., von Mering, C., & Bork, P. (2017). Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Molecular Biology and Evolution*, *34*(8), 2115–2122. https://doi.org/10.1093/molbev/msx148

Kolde, R. (2019). *pheatmap: Pretty Heatmaps*. https://CRAN.R-project.org/package=pheatmap

Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: A Package for Multivariate Analysis. *Journal of Statistical Software*, *25*(1), 1–18. https://doi.org/10.18637/jss.v025.i01

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

ANEXO 1. Listagem dos 20 mais expressivos transcritos (DE) que foram anotados com o CAZy, comparação entre as semanas e os dias dentro de cada semana

Contraste: Semanas							
	ID do transcrito	**Média (Semana: 2)**	**Média (Semana: 20)**	**log2FC**	**Nome (gene)**	**Descrição**	**4**
Semana: 2	TRINITY_DN7349_c0_g1_i1	117,2	< 0.01	20,32	*glgP*	Carbohydrate phosphorylase	GH (65), GT (3, 5, 35)
	TRINITY_DN1135_c0_g1_i2	24,61	0	17,04	-	Lipase (class 3)	CE (5)
	TRINITY_DN210_c0_g1_i1	17,05	0	19,32	-	Glycosyl transferase family 2	CE (4)
	TRINITY_DN1742_c0_g1_i1	16,66	0	19,21	-	Alpha-L-fucosidase	GH (29, 95, 141, 151)
	TRINITY_DN208790_c0_g1_i1	16,11	0	14,43	-	COG3387 Glucoamylase and related glycosyl	GH (15, 97)
	TRINITY_DN2969_c0_g1_i1	13,72	0	30	*celA3*	Endoglucanase	GH (5, 6, 7, 8, 9, 10, 12, 26, 44, 45, 48, 51, 74, 124, 148)
	TRINITY_DN488_c0_g1_i1	13,25	0	17,5	-	Phage lysozyme	GH (19, 22, 23, 24, 25, 108, 184)
	TRINITY_DN195_c0_g3_i2	12,54	< 0.01	13,51	*guxA*	Belongs to the glycosyl hydrolase family 6	GH (5, 6, 9, 48, 51)
	TRINITY_DN27842_c0_g1_i1	10,38	1,45	6,48	*lpxC*	Catalyzes the hydrolysis of UDP-3-O-myristoyl-N- acetylglucosamine (…)	CE (11)

	TRINITY_DN897_c0_g1_i3	11,3	0	16,22	*aguA*	Belongs to the glycosyl hydrolase 67 family	GH (4, 67)
	TRINITY_DN1094_c0_g1_i1	10,01	0	15,78	*glgA*	PFAM Starch synthase catalytic domain	GT (4, 5)
	TRINITY_DN996_c0_g1_i3	7,47	< 0.01	17,16	-	Belongs to the glycosyl hydrolase family 6	GH (5, 6, 7, 8, 9, 10, 12, 26, 44, 45, 48, 51, 74, 124, 148)
	TRINITY_DN4369_c0_g1_i1	7,45	0	15,51	*lnt*	Carbon-nitrogen hydrolase	GT (2)
	TRINITY_DN6063_c0_g1_i1	6,4	0	14,87	-	metallopeptidase activity	GH (3, 5, 6, 8, 9, 10, 11, 30, 43, 44, 51, 62, 98, 141)
	TRINITY_DN151068_c0_g1_i1	5,98	0	13,02	-	beta-1,4-mannooligosaccharide phosphorylase	GH (130)
	TRINITY_DN635_c0_g1_i2	5,66	0	18,36	*amyS*	Alpha amylase, catalytic domain	GH (13, 57, 119, 126)
	TRINITY_DN1667_c0_g2_i1	5,56	0	13,78	-	PFAM lipase class 3	CE (5)
	TRINITY_DN240921_c0_g1_i1	5,39	0	14,25	*otsB*	Removes the phosphate from trehalose 6-phosphate to produce free trehalose	GT (20)
	TRINITY_DN635_c0_g1_i1	5,17	0	18,32	*amyS*	Alpha amylase, catalytic domain	GH (13, 57, 119, 126)
	TRINITY_DN188733_c0_g1_i1	4,82	0	10,96	-	Belongs to the glycosyl hydrolase 31 family	GH (4, 13, 31, 63, 76, 97, 122)
Semana: 20	TRINITY_DN5360_c1_g1_i1	0,02	30,43	3,49	*guxA*	Belongs to the glycosyl hydrolase family 6	GH (5, 6, 9, 48, 51)
	TRINITY_DN95_c0_g2_i3	0	25,67	7,96	-	Glycosyl transferase	GH (94), GT (36)

TRINITY_DN999_c0_g1_i44	0,03	12,25	3,99	*bgaA*	beta-galactosidase	GH (1, 2, 3, 5, 16, 35, 39, 42, 50, 54, 147, 165, 173)
TRINITY_DN1171_c0_g2_i1	0	11,13	6,69	-	Belongs to the glycosyl hydrolase 31 family	GH (31)
TRINITY_DN669_c1_g3_i1	0,06	11,02	3,37	*mdoH*	Involved in the biosynthesis of osmoregulated periplasmic glucans (OPGs)	GT (2)
TRINITY_DN5098_c0_g1_i2	0	10,92	6,45	*chiA1*	chitinase	GH (18, 19, 23, 48)
TRINITY_DN5253_c0_g1_i1	0	10,38	6,56	*melA7*	alpha-galactosidase	GH (4, 27, 31, 36, 57, 97, 110)
TRINITY_DN396_c1_g5_i1	0	8,43	8,24	*lnt*	Transfers the fatty acyl group on membrane lipoproteins	GT (2)
TRINITY_DN22447_c0_g1_i1	< 0.01	8,36	5,69	*bcsA*	Cellulose synthase	GT (2)
TRINITY_DN60_c0_g5_i1	0,06	8,08	3,47	*malS*	alpha-amylase	GH (13, 57, 119, 126)
TRINITY_DN1831_c0_g2_i2	0,02	6,09	3,3	*ascB*	Belongs to the glycosyl hydrolase 1 family	GH (1, 2, 3, 4, 5, 16, 30, 39, 116, 131, 175, 180), GT (1)
TRINITY_DN152_c0_g5_i1	0	5,59	7,4	-	Penicillin-binding protein, 1A family	GT (51)
TRINITY_DN1585_c0_g5_i1	0	4,64	5,76	-	cell wall organization	PL (4, 9, 11)
TRINITY_DN1778_c0_g1_i1	0	4,4	5,59	-	deoxyhypusine monooxygenase activity	GT (28)
TRINITY_DN95_c0_g2_i2	< 0.01	4,35	5	*chvB3*	Cellobiose phosphorylase	GH (94), GT (36)

	TRINITY_DN1935_c1_g2_i1	0,03	3,97	3,01	*lnt*	Transfers the fatty acyl group on membrane lipoproteins	GT (2)
	TRINITY_DN150879_c0_g1_i1	0,02	3,73	3,67	*ybhC*	Pectinesterase	CE (8, 19)
	TRINITY_DN62_c0_g1_i13	0	3,74	6,4	*glgB*	Catalyzes the formation of the alpha-1,6-glucosidic linkages (…)	GH (13, 57), CBM (48)
	TRINITY_DN10464_c1_g1_i1	< 0.01	3,68	4,07	-	cell wall organization	PL (4, 9, 11)
	TRINITY_DN1866_c0_g2_i4	< 0.01	3,61	4,83	-	Cellulase N-terminal ig-like domain	GH (5, 6, 7, 8, 9, 10, 12, 26, 44, 45, 48, 51, 74, 124, 148)

Contraste: Dias (Semana: 2)							
	ID do transcrito	**Média (Dia: 5)**	**Média (Dia: 10)**	**log2FC**	**Nome (gene)**	**Descrição**	**Família(s) CAZy**
Dia 5	TRINITY_DN631_c0_g2_i3	0,56	0,05	5,49	-	Belongs to the glycosyl hydrolase 11 (cellulase G) family	GH (3, 5, 6, 8, 9, 10, 11, 30, 43, 44, 51, 62, 98, 141)
	TRINITY_DN1742_c0_g1_i1	30,63	2,69	3,03	-	Alpha-L-fucosidase	GH (29, 95, 141, 151)
	TRINITY_DN2969_c0_g1_i1	27,44	< 0.01	12,97	*celA3*	Endoglucanase	GH (5, 6, 7, 8, 9, 10, 12, 26, 44, 45, 48, 51, 74, 124, 148)
	TRINITY_DN488_c0_g1_i1	25,03	1,48	4,12	-	Phage lysozyme	GH (19, 22, 23, 24, 25, 108, 184)

TRINITY_DN195_c0_g3_i2	25,06	< 0.01	12,49	*guxA*	Belongs to the glycosyl hydrolase family 6	GH (5, 6, 9, 48, 51)
TRINITY_DN897_c0_g1_i3	22,59	0,01	11,82	*aguA*	Belongs to the glycosyl hydrolase 67 family	GH (4, 67)
TRINITY_DN1094_c0_g1_i1	15,56	4,46	3,01	*glgA*	PFAM Starch synthase catalytic domain	GT (4, 5)
TRINITY_DN996_c0_g1_i3	14,93	0,01	11,45	-	Belongs to the glycosyl hydrolase family 6	GH (5, 6, 7, 8, 9, 10, 12, 26, 44, 45, 48, 51, 74, 124, 148)
TRINITY_DN1667_c0_g2_i1	10,85	0,28	4,44	-	PFAM lipase class 3	CE (5)
TRINITY_DN240921_c0_g1_i1	9,92	0,86	4,67	*otsB*	Removes the phosphate from trehalose 6-phosphate to produce free trehalose	GT (20)
TRINITY_DN152909_c0_g1_i1	8,29	0,35	4,04	*waaA*	3-deoxy-D-manno-octulosonic acid transferase	GT (30)
TRINITY_DN205137_c0_g1_i1	7,13	0,92	2,36	*amyS*	Glycosyl hydrolase family 70	GH (13, 57, 119, 126)
TRINITY_DN12418_c1_g1_i1	6,85	0,11	4,98	*bax*	FlgJ-related protein	GH (73)
TRINITY_DN237234_c0_g1_i1	5,14	0,16	4,29	-	cytochrome C peroxidase	AA (2)
TRINITY_DN4053_c0_g1_i3	5,11	0,04	7,96	*pbpF*	penicillin-binding protein	GT (51)
TRINITY_DN16484_c0_g1_i6	5,09	0	10,72	*cda1*	xylanase chitin deacetylase	CE (4)
TRINITY_DN1316_c2_g2_i2	4,92	< 0.01	10,61	*aguA*	Belongs to the glycosyl hydrolase 67 family	GH (4, 67)

	TRINITY_DN127547_c0_g1_i1	4,1	0,26	3,48	-	Belongs to the glycosyl hydrolase 13 family	GH (13, 57)
	TRINITY_DN36996_c0_g1_i1	4,28	0,03	8,83	-	Beta-xylanase	GH (3, 5, 6, 8, 9, 10, 11, 30, 43, 44, 51, 62, 98, 141)
	TRINITY_DN1316_c2_g1_i1	2,99	0	7,78	*aguA*	Belongs to the glycosyl hydrolase 67 family	GH (4, 67)
Dia: 10	**TRINITY_DN635_c0_g1_i2**	0,41	10,91	4,6	*amyS*	Alpha amylase, catalytic domain	GH (13, 57, 119, 126)
	TRINITY_DN635_c0_g1_i1	2,09	8,25	2,16	*amyS*	Alpha amylase, catalytic domain	GH (13, 57, 119, 126)
	TRINITY_DN151310_c0_g1_i1	1,41	8,08	3,69	-	Chitin synthase	GT (2)
	TRINITY_DN36501_c0_g3_i1	0,06	1,41	5,22	*lpxC*	Catalyzes the hydrolysis of UDP-3-O-myristoyl-N- acetylglucosamine (…)	CE (11)
	TRINITY_DN14801_c0_g4_i1	0,04	0,61	5,05	*MA20_09515*	GtrA-like protein	GT (2)
	TRINITY_DN103293_c0_g1_i1	0,05	0,78	4,33	-	Glycosyl transferase	GT (2, 4)
	TRINITY_DN8124_c0_g2_i1	0,05	0,73	4,33	*lpxC*	Catalyzes the hydrolysis of UDP-3-O-myristoyl-N- acetylglucosamine (…)	CE (11)
	TRINITY_DN59805_c0_g1_i1	0,06	0,7	3,53	*bcsA*	Cellulose synthase catalytic subunit (UDP-forming)	GT (2)
	TRINITY_DN12744_c0_g1_i1	0,04	0,5	4,87	*glgE*	Maltosyltransferase that uses maltose 1-phosphate (M1P) (…)	GH (13)
	TRINITY_DN16925_c0_g2_i1	0,01	0,32	5,04	*mltA*	MltA specific insert domain	GH (102)

	TRINITY_DN177871_c0_g1_i1	0,03	0,27	3,48	*gnl*	protein import	GH (5, 6, 7, 8, 9, 10, 12, 26, 44, 45, 48, 51, 74, 124, 148)
	TRINITY_DN4311_c3_g1_i1	< 0.01	0,14	3,73	-	Glycosyltransferases involved in cell wall biogenesis	GT (2, 4)

Contraste: Dias (Semana: 20)							
	ID do transcrito	**Média (Dia: 5)**	**Média (Dia: 10)**	**log2FC**	**Nome (gene)**	**Descrição**	**Família(s) CAZy**
Dia 5	TRINITY_DN2829_c0_g3_i1	2,17	0,44	2,28	*glgB*	Catalyzes the formation of the alpha-1,6-glucosidic linkages (…)	GH (13, 57), CBM (48)
	TRINITY_DN3860_c0_g1_i2	0,95	0,13	3,11	-	Belongs to the glycosyl hydrolase 26 family	GH (5, 26, 113, 134)
	TRINITY_DN124996_c0_g1_i1	0,68	0,12	2,52	*pbpG*	Carboxypeptidase	GT (51)
	TRINITY_DN5799_c0_g1_i1	0,67	0,11	2,59	*acm1*	lysozyme activity	GH (19, 22, 23, 24, 25, 108, 184)
	TRINITY_DN3684_c0_g8_i1	0,26	0,05	2,39	-	Endoglucanase	GH (5, 6, 7, 8, 9, 10, 12, 26, 44, 45, 48, 51, 74, 124, 148)
Dia: 10	**TRINITY_DN1778_c0_g1_i1**	0,04	8,76	7,75	-	deoxyhypusine monooxygenase activity	GT (28)
	TRINITY_DN15317_c0_g1_i1	0,05	7,97	7,33	-	cytochrome c peroxidase	AA (2)

TRINITY_DN62_c0_g1_i13	0,63	6,85	3,44	*glgB*	Catalyzes the formation of the alpha-1,6-glucosidic linkages (…)	GH (13, 57), CBM (48)
TRINITY_DN104206_c0_g1_i1	0	7,39	9,22	*treY*	Alpha amylase, catalytic domain	GH (13)
TRINITY_DN212171_c0_g1_i1	0	4,42	11,62	-	Belongs to the glycosyl hydrolase 13 family	GH (13, 176), CBM (48)
TRINITY_DN26510_c0_g1_i1	0,01	4,12	8,88	*mltC*	Transglycosylase SLT domain	GH (23)
TRINITY_DN151634_c0_g1_i1	< 0.01	3,74	9,52	*bamD*	cell envelope organization	GH (23)
TRINITY_DN236518_c0_g1_i1	0,03	2,26	6,27	-	cytochrome c peroxidase	AA (2)
TRINITY_DN5400_c0_g2_i2	0	1,99	10,17	*XK27_07910*	Polysaccharide deacetylase	CE (4)
TRINITY_DN1519_c1_g1_i1	0	1,82	9,76	*glgP*	Phosphorylase is an important allosteric enzyme in carbohydrate metabolism. (…)	GT (35)
TRINITY_DN6238_c1_g1_i2	0	1,73	8,99	-	Lysozyme C-like	GH (19, 22, 23, 24, 25, 108, 184)
TRINITY_DN236528_c0_g1_i1	0	1,73	7,9	-	Belongs to the glycosyl hydrolase 11 (cellulase G) family	GH (3, 5, 6, 8, 9, 10, 11, 30, 43, 44, 51, 62, 98, 141)
TRINITY_DN182751_c0_g1_i2	0,01	1,54	7,11	-	cytochrome c peroxidase	AA (2)
TRINITY_DN206557_c0_g1_i1	0	1,49	7,9	*bamD*	cell envelope organization	GH (23)
TRINITY_DN897_c0_g1_i1	0	1,48	24,83	*aguA*	Belongs to the glycosyl hydrolase 67 family	GH (4, 67)
TRINITY_DN84772_c0_g2_i1	< 0.01	1,47	7,41	-	Alg9-like mannosyltransferase family	GT (22)

TRINITY_DN68133_c0_g3_i2	< 0.01	1,41	7,71	*mrcA*	penicillin-binding protein	GT (51)
TRINITY_DN206626_c0_g1_i1	0	1,38	8,71	*bamD*	cell envelope organization	GH (23)
TRINITY_DN158116_c0_g1_i1	0,09	1,21	4,17	-	Belongs to the glycosyl hydrolase 11 (cellulase G) family	GH (3, 5, 6, 8, 9, 10, 11, 30, 43, 44, 51, 62, 98, 141)
TRINITY_DN239208_c0_g1_i1	0,11	0,62	2,6	*mrcB*	Peptidoglycan polymerase that catalyzes glycan chain elongation (…)	GT (51)