



Universidade Estadual de Campinas  
Instituto de Biologia



Victor Marques de Oliveira Gambarini

Análise do metatranscriptoma de rizosfera de  
cana-de-açúcar submetida ao estresse hídrico

Campinas  
2018

Victor Marques de Oliveira Gambarini

Análise do metatranscriptoma de rizosfera de  
cana-de-açúcar submetida ao estresse hídrico

Dissertação apresentada ao Instituto de Biologia da Universidade Estadual de Campinas como parte dos requisitos para a obtenção do título de Mestre em Genética e Biologia Molecular, na área de Bioinformática.

ESTE ARQUIVO DIGITAL CORRESPONDE À  
VERSÃO FINAL DA [TESE / DISSERTAÇÃO]  
DEFENDIDA [PELO ALUNO / PELA ALUNA]  
[NOME DO(A) ALUNO(A)] E ORIENTADA [PELO  
/ PELA] [NOME DO(A) ORIENTADOR(A)].

Orientador: Prof. Dr. Renato Vicentini dos Santos

Campinas  
2018

**Agência(s) de fomento e nº(s) de processo(s):** FAPESP, 2016/08994-3  
**ORCID:** <https://orcid.org/0000-0002-5639-6820>

Ficha catalográfica  
Universidade Estadual de Campinas  
Biblioteca do Instituto de Biologia  
Mara Janaina de Oliveira - CRB 8/6972

G141a                      Gambarini, Victor Marques de Oliveira, 1987-  
                                 Análise do metatranscriptoma de rizosfera de cana-de-açúcar  
                                 submetida ao estresse hídrico / Victor Marques de Oliveira  
                                 Gambarini. – Campinas, SP :  
                                 [s.n.], 2018.

                                 Orientador: Renato Vicentini dos Santos.  
                                 Dissertação (mestrado) – Universidade Estadual de  
                                 Campinas, Instituto de Biologia.

                                 1. Transcriptoma. 2. Rizosfera. 3. Cana-de-açúcar. I. Vicentini,  
                                 Renato, 1979-. II. Universidade Estadual de Campinas.  
                                 Instituto de Biologia. III. Título.

#### Informações para Biblioteca Digital

**Título em outro idioma:** Analysis of the sugarcane rizosphere metatranscriptome subjected to water stress

**Palavras-chave em inglês:**

Transcriptome

Rhizosphere

Sugar-cane

**Área de concentração:** Bioinformática

**Titulação:** Mestre em Genética e Biologia Molecular

**Banca examinadora:**

Renato Vicentini dos Santos [Orientador]

Marcelo Falsarella Carazzolle

Marco Aurelio Takita

**Data de defesa:** 23-07-2018

**Programa de Pós-Graduação:** Genética e Biologia Molecular

Campinas, 23 de julho de 2018.

**COMISSÃO EXAMINADORA**

Prof. Dr. Renato Vicentini dos Santos

Prof. Dr. Marcelo Falsarella Carazzolle

Prof. Dr. Marco Aurelio Takita

*Os membros da Comissão Examinadora acima assinaram a Ata de Defesa, que se encontra no processo de vida acadêmica do aluno.*

Aos meus pais Edson (in memorian) e Ivete,  
à minha sogra Violeta e, principalmente, à  
minha esposa Raquel, pelo apoio,  
força, incentivo e amor.

## **AGRADECIMENTOS**

Agradeço a Deus pela graça imerecida e por toda a capacidade para a realização dessa obra.

À minha esposa Raquel por todo o amor e companheirismo.

À minha mãe Ivete por todo apoio e incentivo.

À minha sogra Violeta pelo carinho e incentivo.

A todos os meus familiares pelo apoio.

Aos colegas de laboratório pela ajuda e amizade.

Ao professor Dr. Renato pela oportunidade, auxílios e paciência.

À Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pelo apoio financeiro, por meio do processo 2016/08994-3. Registro aqui que as opiniões, hipóteses e conclusões ou recomendações expressas neste material são de responsabilidade do(s) autor(es) e não necessariamente refletem a visão da FAPESP.

## Resumo

Em vista das mudanças climáticas, estão previstas mudanças no regime pluviométrico de todo o globo, causando seca em determinadas regiões e aumento de chuvas em outras. Nesse sentido, plantas de grande interesse comercial como a cana-de-açúcar devem ser alvo de pesquisas em busca de maior produtividade e de maiores condições de adaptação frente às futuras alterações do clima. É sabido que os microrganismos que se desenvolvem próximos à raiz das plantas, ou na chamada rizosfera, podem influenciar o crescimento e desenvolvimento destas através de uma variedade de mecanismos, como por exemplo a promoção da mobilização de nutrientes, aumento da absorção de minerais, estímulo ao crescimento vegetal e supressão de patógenos. Dessa forma, esse estudo utilizou métodos de bioinformática para analisar os dados gerados no sequenciamento em larga escala do metatranscriptoma da rizosfera de cana-de-açúcar submetida ao estresse hídrico. Ademais, foi caracterizado o perfil taxonômico e funcional da microbiota ativa, além de ter-se caracterizado a expressão gênica diferencial induzida pelo estresse hídrico e, finalmente, foram geradas redes de co-expressão gênica a partir dos dados obtidos. A montagem *de novo* do metatranscriptoma resultou em 316.958 *contigs* a partir dos quais foram preditas 405.547 *ORFs*. As análises da composição taxonômica do metatranscriptoma revelaram uma predominância dos filos Proteobacteria (52,2%), Actinobacteria (15,1%) e Bacteroidetes (10,6%). Os quatro gêneros mais abundantes que tiveram abundância relativa aumentada após a aplicação do estresse hídrico foram Actinomadura, Dongia, Sphingomonas e Vulgatibacter. A anotação funcional no SEED foi capaz de anotar 208.816 transcritos (51,5%) em 40 categorias funcionais diferentes. As categorias funcionais que apresentaram maior abundância no SEED foram aquelas responsáveis por processos essenciais dos microrganismos, como por exemplo metabolismo de carboidratos, proteínas, aminoácidos,

vitaminas e RNA. Foram descobertas três categorias do SEED com aumento de abundância após a aplicação do estresse hídrico, sendo elas relacionadas a degradação de proteínas para obtenção de energia (degradação de arginina e ornitina) e enzimas oxirredutases atuantes na cadeia respiratória aeróbia (succinato desidrogenase) e em processos fermentativos (cetoisovalerato oxirredutase). A análise da expressão gênica diferencial resultou em 43.716 transcritos diferencialmente expressos e demonstrou que foram *down*-regulados termos GO relacionados a regulação da respiração celular, biogênese de ribossomo, dobramento de proteína, endocitose, resposta ao estresse e tradução. Por outro lado, os principais termos GO enriquecidos nos transcritos *up*-regulados foram catabolismo de benzoato, montagem de flagelo e síntese de tRNA de pseudouridina. Além disso, foi gerada uma rede de co-expressão gênica a partir dos transcritos diferencialmente expressos que foi composta por 582 clusters e, através das análises dessa rede, pode-se observar que com a aplicação do estresse hídrico, houve um aumento da expressão de transcritos relacionados, por exemplo a auto proteólise, degradação de ácidos nucleicos e transporte de sideróforos. Os resultados obtidos nesse trabalho auxiliam na compreensão da diversidade taxonômica e funcional da rizosfera de cana-de-açúcar e os efeitos causados pelo estresse hídrico. Finalmente, este trabalho pode servir de base para o desenvolvimento de aplicações para o conhecimento aqui gerado, visando a adaptação e aumento da produtividade da cana-de-açúcar frente às previstas mudanças climáticas globais.



## Abstract

In view of climate changes, the pluviometric regime of the whole globe is expected to change, causing drought in certain regions and rainfall increase in others. In this sense, plants of great commercial interest such as sugarcane should be the subject of research in pursuit of greater productivity and better adaptation conditions in the face of future climate changes. It is known that microorganisms that develop near the root of plants, or in the so-called rhizosphere, can influence the growth and development of plants through a variety of mechanisms, such as promoting the mobilization of nutrients, increasing mineral absorption, stimulating plant growth and suppressing pathogens. Thus, this study used bioinformatics methods to analyze data generated in large-scale sequencing of the sugarcane rhizosphere metatranscriptome subjected to water stress. In addition, the taxonomic and functional profile of the active microbiota was characterized, the differential gene expression induced by water stress was revealed, and finally, co-expression gene networks were generated from the data obtained. De novo assembly of the metatranscriptome resulted in 316,958 contigs from which 405,547 ORFs were predicted. The analysis of the taxonomic composition of the metatranscriptome revealed a predominance of the phyla Proteobacteria (52,2%), Actinobacteria (15,1%) and Bacteroidetes (10,6%). The four most abundant genera that had increased relative abundance after application of water stress were *Actinomadura*, *Dongia*, *Sphingomonas* and *Vulgatibacter*. The functional annotation in SEED was able to annotate 208,816 transcripts (51.5%) in 40 different functional categories. The functional categories that presented greater abundance in SEED were those responsible for essential microorganism's processes, such as carbohydrate, protein, amino acids, vitamins and RNA metabolism. Three SEED categories with an increase in abundance after the application of water stress were discovered, being related to the degradation of proteins to obtain energy

(degradation of arginine and ornithine) and oxidizing enzymes acting in the aerobic respiratory chain (succinate dehydrogenase) and in fermentation processes (ketoisovalerate oxirreductase). Analysis of differential gene expression resulted in 43,716 differentially expressed transcripts and demonstrated that GO terms related to regulation of cellular respiration, ribosome biogenesis, protein folding, endocytosis, stress response and translation were down-regulated. On the other hand, the major GO terms enriched in up-regulated transcripts were benzoate catabolism, flagellum assembly, and pseudouridine tRNA synthesis. In addition, a gene co-expression network was generated from the differentially expressed transcripts which was composed of 582 clusters and, through the analyzes of this network, it could be observed that with the application of water stress, there was an increase in the expression of transcripts related to, for example, auto proteolysis, degradation of nucleic acids and transport of siderophores. The results obtained in this work help to understand the taxonomic and functional diversity of sugarcane rhizosphere and the effects caused by water stress. Finally, this work can be used as a basis for the development of applications for the knowledge generated here, aiming to adapt and increase the productivity of sugarcane in anticipation of global climate change.

## Lista de Figuras

<b>Figura 1:</b> Projeções médias para as mudanças na temperatura da superfície e precipitação entre os anos de 2081 e 2100. a) Mudança na temperatura média da superfície e b) mudança na média anual de precipitação. (Modificado de IPCC 2014). .....	18
<b>Figura 2:</b> Exemplo das interações planta-microrganismos na zona da rizosfera. As raízes das plantas liberam exsudatos contendo açúcares, ácidos orgânicos e amino ácidos que podem atrair microrganismos. Em troca, esses microrganismos podem liberar moléculas com atividade antimicrobiana, liberar hormônios e aumentar a disponibilidade de nutrientes para as plantas (retirado de Rincon-Florez, Carvalhais, and Schenk 2013). .....	20
<b>Figura 3:</b> O dogma centra da biologia, ilustra o fluxo de informação que ocorre do DNA para o RNA (transcrição) e do RNA para a proteína (tradução). Esse processo ocorre em todas as células vivas, independentemente de sua classificação taxonômica (Alberts et al. 2014). .....	28
<b>Figura 4:</b> Exemplo de rede de co-expressão gênica, mostrando a rede gerada a partir de dados de RNA-Seq. Os nós coloridos representam diferentes funções biológicas associadas a cada um dos transcritos e as arestas cinza representam as relações entre os nós. ....	29
<b>Figura 1:</b> Distribuição de comprimentos dos transcritos montados com o <i>software</i> EMIRGE. Nota-se picos próximos a 1500 e 1800 pb, que são os comprimentos estimados dos genes 16S e 18S, respectivamente. ....	41
<b>Figura 2:</b> Abundância relativa nos níveis de domínio e filo dos transcritos com taxonomia atribuída a partir do RNA ribossomal. ....	44
<b>Figura 3:</b> Abundância relativa nos níveis de classe e ordem dos transcritos com taxonomia atribuída a partir do RNA ribossomal. ....	46
<b>Figura 4:</b> Abundância relativa nos níveis de família e gênero dos transcritos com taxonomia atribuída a partir do RNA ribossomal. ....	47
<b>Figura 5:</b> Análise de componentes principais dos dados a nível de filo para os diferentes tratamentos de estresse hídrico aplicados. ....	48
<b>Figura 6:</b> Proporção média e diferenças preditas para os transcritos de rRNA no nível de gênero para os tratamentos com zero e dois dias de estresse hídrico. ....	49
<b>Figura 7:</b> Proporção média e diferenças preditas para os transcritos de rRNA no nível de gênero para os tratamentos com zero e doze dias de estresse hídrico. ....	49
<b>Figura 8:</b> Comparação entre as atribuições taxonômicas a partir dos dados de mRNA contra os dados de rRNA. ....	51
<b>Figura 9:</b> Desempenho dos pares dos adaptadores testados com o software desenvolvido, chamado VirtualPrimer. ....	52
<b>Figura 10:</b> Número de sequências alinhadas pelo VirtualPrimer aos <i>primers</i> 338F e 1046R. ....	53
<b>Figura 11:</b> Filtragem dos transcritos com comprimento fora da média obtida. A, antes e B, depois da filtragem. ....	54
<b>Figura 12:</b> O mapeamento das <i>reads</i> aos <i>contigs</i> é uma das principais métricas de avaliação de montagens e, pode-se observar que os melhores desempenhos foram do <i>software</i> IDBA-UD e das estratégias de concatenação. ....	58
<b>Figura 13:</b> Resultado obtido através do <i>software</i> BUSCO versão 2 utilizando-se genes normalmente encontrados em genomas de bactérias do filo Proteobacteria. A amostra utilizada para esta análise foi a Bulk 1. ....	59

<b>Figura 14:</b> Genes preditos pelo Prodigal para as diferentes estratégias de montagem aplicadas à amostra Bulk 1. ....	60
<b>Figura 15:</b> Quantidade de transcritos anotados a partir do <i>software</i> SUPERFOCUS para cada uma das principais estratégias de montagem aplicadas à amostra Bulk 1. ....	61
<b>Figura 1:</b> Genes BUSCO encontrados na montagem do IDBA-UD com todas as amostras combinadas. ....	76
<b>Figura 2:</b> Composição taxonômica dos transcritos do metatranscriptoma de rizosfera de cana-de-açúcar em nível de família para os três filos mais abundantes. ....	80
<b>Figura 3:</b> Abundância relativa, em vários níveis taxonômicos, dos transcritos do metatranscriptoma de rizosfera de cana-de-açúcar para cada um dos tratamentos de estresse hídrico aplicados. ....	82
<b>Figura 4:</b> Diferenças na proporção média dos gêneros com diferença significativa ( $p < 0,05$ ) em amostras da rizosfera de cana-de-açúcar com ou sem tratamentos de estresse hídrico. ....	83
<b>Figura 5:</b> Composição funcional dos transcritos do metatranscriptoma de rizosfera de cana-de-açúcar classificados no nível um da hierarquia de subsistemas do SEED. ....	85
<b>Figura 6:</b> Número de transcritos para cada uma das famílias CAZyme encontradas no metatranscriptoma. ....	86
<b>Figura 7:</b> Porcentagem relativa de transcritos putativos de resistência à antibióticos encontrados a partir do banco de dados CARD. ....	87
<b>Figura 8:</b> Abundância diferencial de transcritos induzida pelos tratamentos de estresse hídrico. Os transcritos da microbiota da rizosfera estão sendo avaliados no nível dois da hierarquia SEED. ....	88
<b>Figura 9:</b> Filtragem dos transcritos exclusivos da amostra sem planta. Foram filtrados 24.715 transcritos de total de 405.547 transcritos. ....	90
<b>Figura 10:</b> Heatmap mostrando o perfil de expressão gênica dos 43.716 transcritos diferencialmente expressos entre os diferentes tratamentos de estresse hídrico analisados. ....	92
<b>Figura 11:</b> Análise de enriquecimento de termos GO relacionados a processos biológicos. Os termos GO foram resumidos e visualizados em um <i>treemap</i> pelo servidor Web REVIGO. O tamanho dos quadros indica o $-\log_{(10)}$ do p-valor do enriquecimento dos termos GO. Quadros com cores iguais pertencem a um mesmo termo GO pai. ....	94
<b>Figura 12:</b> Análise de enriquecimento de termos GO relacionados a processos biológicos. Os termos GO foram resumidos e visualizados em um <i>treemap</i> pelo servidor Web REVIGO. O tamanho dos quadros indica o $-\log_{(10)}$ do p-valor do enriquecimento dos termos GO. Quadros com cores iguais pertencem a um mesmo termo GO pai. ....	96
<b>Figura 13:</b> Análise de enriquecimento de termos GO relacionados a processos biológicos. Os termos GO foram resumidos e visualizados em um <i>treemap</i> pelo servidor Web REVIGO. O tamanho dos quadros indica o $-\log_{(10)}$ do p-valor do enriquecimento dos termos GO. Quadros com cores iguais pertencem a um mesmo termo GO pai. ....	97
<b>Figura 14:</b> Ciclo do ácido cítrico no banco de dados KEEG. Em vermelho são os transcritos que tiveram aumento estatisticamente significativo nos valores de expressão, em verde são os transcritos que tiveram diminuição na expressão e em branco são transcritos que não tiveram alteração estatisticamente significativa. ....	98
<b>Figura 15:</b> Self-organizing map (SOM) análise dos 43.715 transcritos diferencialmente expressos na microbiota associada a rizosfera de cana-de-açúcar CTC 9001. Em cada quadro, o eixo x indica amostras (0, 2, 8 e 12 dias de estresse hídrico) e o eixo y indica o valor de centralização da média da expressão gênica de um determinado conjunto de	

transcritos (valor “n” na parte superior). Cores similares na parte superior indicam semelhança no padrão de expressão dos transcritos. Cada quadro contém o nome dos dois GO mais enriquecidos (menor valor de p) e os gêneros mais abundantes relacionados ao perfil de expressão apresentado. .... 100

**Figure 1** Venn diagram of clusters that had transcripts related to one or more features of interest. .... 118

**Figure 2** Co-expression network of 42.941 transcripts clustered in 582 clusters. Nodes in this network represent clusters. Node colors represent different features of clusters, while edges represent significant ( $P < 0.05$ ) associations between the clusters based on co-expression. .... 120

**Figure 3** Biological process GO terms enriched in the clusters up or down-regulated. Go terms were summarized and visualized in a treemap by REVIGO. The area of the rectangles indicate the GO terms enrichment  $-\log(10)$  of the p-value. Equal colors represent GO terms from the same parent term. .... 121

**Figure 4** Network topology of cluster c315. A grey circle represents a node (i.e. transcript). A grey edge connecting two nodes indicates the co-expression relationship between these two nodes. .... 123

**Figure 5** Network topology of cluster c188. A grey circle represents a node (i.e. transcript). A grey edge connecting two nodes indicates the co-expression relationship between these two nodes. .... 124

**Figure 6** Network topology of cluster c421. A grey circle represents a node (i.e. transcript). A grey edge connecting two nodes indicates the co-expression relationship between these two nodes. .... 126

**Figure 7** Network topology of cluster c135. A grey circle represents a node (i.e. transcript). A grey edge connecting two nodes indicates the co-expression relationship between these two nodes. .... 127

**Figure 8** Network topology of cluster c93. A grey circle represents a node (i.e. transcript). A grey edge connecting two nodes indicates the co-expression relationship between these two nodes. .... 128

**Figura S1:** Transcritos relacionados à resistência a antibióticos da classe das sulfonamidas de acordo com busca no banco de dados de resistência à antibióticos CARD. .... 157

## Lista de Tabelas

<b>Tabela 1:</b> Resultado da etapa de controle de qualidade das reads.....	39
<b>Tabela 2:</b> Remoção das <i>reads</i> de RNA ribossomal realizada pelo <i>software</i> riboPicker. *Not used for 16S prediction. **Not used for mRNA assembling. ....	40
<b>Tabela 3:</b> Estatísticas básicas das montagens individuais obtidas através do <i>software</i> EMIRGE para as <i>reads</i> de rRNA de cada uma das amostras. ....	41
<b>Tabela 4:</b> Mapeamento das <i>reads</i> nos transcritos de rRNA através do <i>software</i> Bowtie2. .	42
<b>Tabela 5:</b> Estatísticas das principais estratégias de montagem testadas na amostra Bulk 1. ....	56
<b>Tabela 1:</b> Estatística da montagem do IDBA-UD com todas as amostras.....	74
<b>Tabela 2:</b> <i>Reads</i> mapeadas contra os <i>contigs</i> gerados pelo IDBA-UD através do alinhador Bowtie2.....	75
<b>Tabela 3:</b> Comparação entre Prodigal e FragGeneScan. ....	77
<b>Tabela 4:</b> Porcentagem de alinhamento dos reads nos ORFs preditos por Prodigal e FragGeneScan. Adicionalmente, foram testadas estratégias com diferentes alinhadores: BWA, Bowtie2 e Salmon. ....	78
<b>Table 1</b> Differentially expressed transcripts comparing samples without water stress (0d) against samples with water stress (2d, 8d, and 12d) and the clusters that had 20% or more transcripts up or down-regulated in one of the analyzed comparisons. ....	118
<b>Table 2</b> Clusters that had transcripts related to one or more features of interest.....	119
<b>Table S1</b> Mean values and standard deviation of organic acid exudation rate by CTC 9001 sugarcane roots. The values are expressed in $\mu\text{g.g}$ of root-1.min-1. Adapted from Pereira et al., 2018. ....	158
<b>Table S2</b> Differentially expressed transcripts comparing samples without water stress (0d) against samples with water stress (2d, 8d, and 12d) and the clusters that had 20% or more transcripts up or down-regulated in one of the analyzed comparisons. ....	159

## Sumário

Informações para Biblioteca Digital .....	3
INTRODUÇÃO .....	16
Mudanças climáticas .....	17
Cana-de-açúcar .....	18
Rizosfera .....	19
Microbiota.....	21
Metatranscriptômica .....	22
Montagem dos dados de metatranscriptoma.....	23
Expressão gênica diferencial.....	25
Redes de co-expressão gênica.....	27
OBJETIVOS .....	31
CAPÍTULO I .....	32
Introdução .....	33
Materiais e métodos .....	34
Resultados e discussão.....	38
Conclusão.....	62
Referências bibliográficas.....	64
CAPÍTULO II.....	70
Introdução .....	71
Materiais e métodos .....	72
Resultados e discussão.....	74
Conclusão.....	102
Referências bibliográficas.....	104
CAPÍTULO III.....	112
Introduction.....	113
Materials and methods .....	114
Results and discussion .....	117
Conclusion .....	128
References.....	130
CONCLUSÃO E PERSPECTIVAS .....	134
REFERÊNCIAS BIBLIOGRÁFICAS.....	137
APÊNDICE I .....	157
APÊNDICE II.....	158
ANEXO I.....	160
ANEXO II.....	161

## INTRODUÇÃO

A maioria das plantas terrestres crescem em íntima associação com uma complexa microbiota. Os microrganismos ali presentes podem fornecer nutrientes essenciais ao desenvolvimento vegetal, como por exemplo minerais presentes no solo que encontram-se em formas inacessíveis às plantas, mas que depois de metabolizados por membros do microbioma, podem ser utilizados pelas plantas (Peiffer et al. 2013).

A rizosfera é a zona mais próxima às raízes da planta, esse ambiente é considerado um dos ecossistemas mais complexos da terra e é habitado por bactérias, fungos, oomicetos, nemátodos, protozoários, algas, vírus, arqueias e artrópodes. Esses organismos fazem parte de uma complexa cadeia alimentar que utiliza nutrientes liberados pela planta na forma de exsudatos. Dessa forma, é predito que a planta pode modular o microbioma da rizosfera à seu favor, selecionando microrganismos com características benéficas ao crescimento e à saúde vegetal (Berendsen, Pieterse, and Bakker 2012).

Uma planta de grande interesse econômico para o Brasil é a cana-de-açúcar, que além de ser utilizada na produção de açúcar e etanol, também tem seu alto potencial de produzir biomassa aproveitado para a geração de energia elétrica e térmica (da Costa et al. 2018). Sabendo-se disso, é preciso estudar formas de desenvolver a cultura de cana-de-açúcar buscando maior rendimento, produção e qualidade, além de facilitar a adaptação da planta frente às mudanças climáticas previstas.

Essas mudanças climáticas incluem aumento do CO<sub>2</sub> atmosférico, aumento da temperatura média global e mudança dos regimes pluviométricos, podendo acarretar secas em certas regiões e aumento da precipitação em outras. De acordo com o relatório do Painel Intergovernamental de Mudanças Climáticas (do inglês IPCC) de 2014, um dos principais



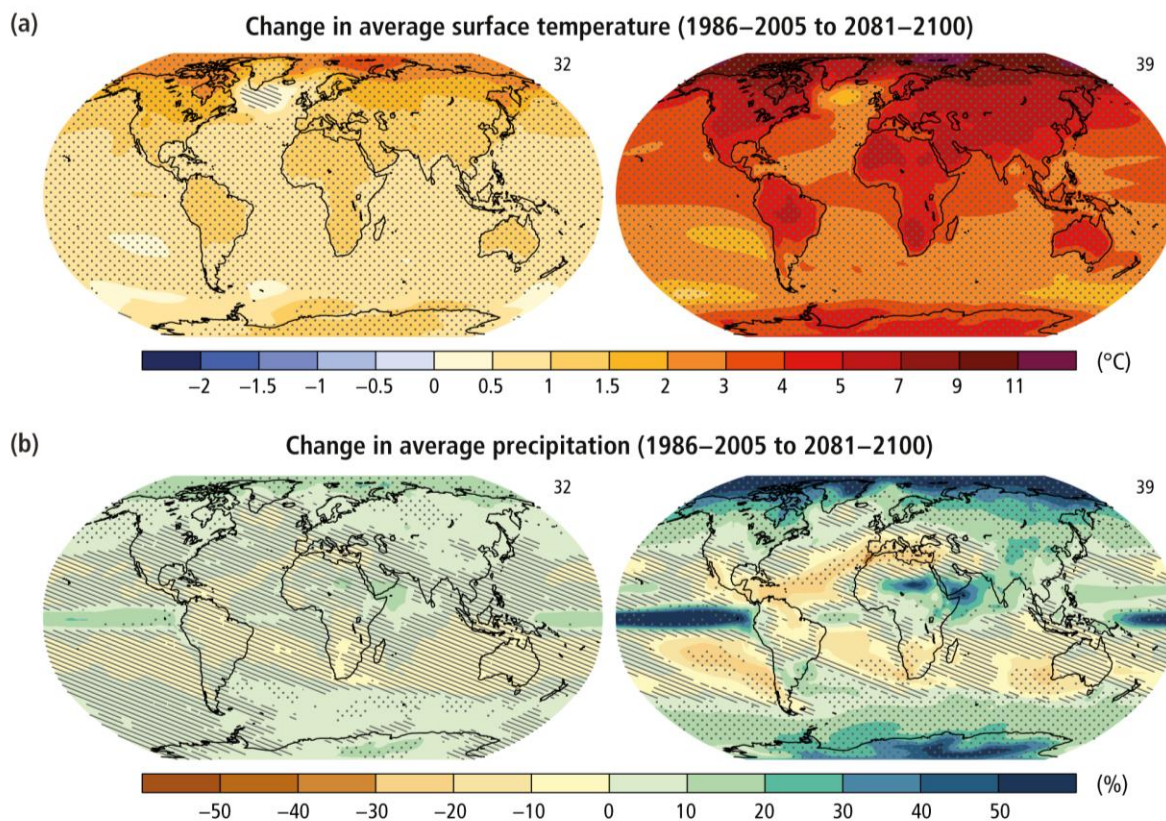
riscos enfrentados por diversas áreas é uma provável diminuição da produção e qualidade dos alimentos, majoritariamente resultante do estresse hídrico e das altas temperaturas (IPCC 2014).

Dessa forma, é importante entender a interação entre a planta e seu microbioma e, também, como modelar esse sistema para otimizar a adaptação da planta às novas condições climáticas previstas. É nesse contexto que se insere o estudo da microbiota da rizosfera da cana-de-açúcar, assim como o foco desse trabalho.

### **Mudanças climáticas**

O aumento das temperaturas globais é influenciado tanto por eventos naturais quanto por atividades humanas, através do aumento da concentração dos gases do efeito estufa, como o dióxido de carbono, o metano e o óxido nitroso. As principais causas do aumento da concentração de gás carbônico na atmosfera são a queima de combustíveis fósseis e as mudanças no uso das terras.

De acordo com o IPCC, o número de dias e noites quentes e a frequência de ondas de calor têm aumentado em grande parte do mundo. Em comparação com o período entre 1850 e 1900, a temperatura da superfície da terra deve subir cerca de 2°C até o fim do século 21. Além do mais, mudanças nos níveis pluviométricos devem ficar mais evidentes, fazendo com que altas latitudes e o Pacífico equatorial provavelmente vejam um aumento em sua média anual de precipitação, enquanto regiões tropicais e de média latitude devem experimentar uma diminuição nos níveis de precipitação (Figura 1). Além disso, as temperaturas médias dos oceanos continuarão subindo durante o século 21, com a maior tendência de aquecimento prevista para regiões tropicais e subtropicais do hemisfério norte (IPCC 2014; Zandalinas et al. 2018).



**Figura 1:** Projeções médias para as mudanças na temperatura da superfície e precipitação entre os anos de 2081 e 2100. a) Mudança na temperatura média da superfície e b) mudança na média anual de precipitação. (Modificado de IPCC 2014).

Esse cenário irá submeter plantações a um grande número de estresses ambientais que podem, inclusive, ocorrer simultaneamente. Uma vez que qualquer alteração no clima terá impactos no desenvolvimento e crescimento de plantas, assim como no relacionamento entre elas e os microrganismos presentes no solo, pode-se prever severas consequências e riscos ao fornecimento de alimento mundial (Gouvêa et al. 2009; Zandalinas et al. 2018).

### Cana-de-açúcar

A cana-de-açúcar (*Saccharum* spp.) é uma planta do tipo C<sub>4</sub>, pertencente à família Poaceae e adaptada a climas tropicais e subtropicais (Daniels and Roach 1987). O gênero *Saccharum* é composto por seis espécies: *S. officinarum*, *S. spontaneum*, *S. robustum*, *S.*

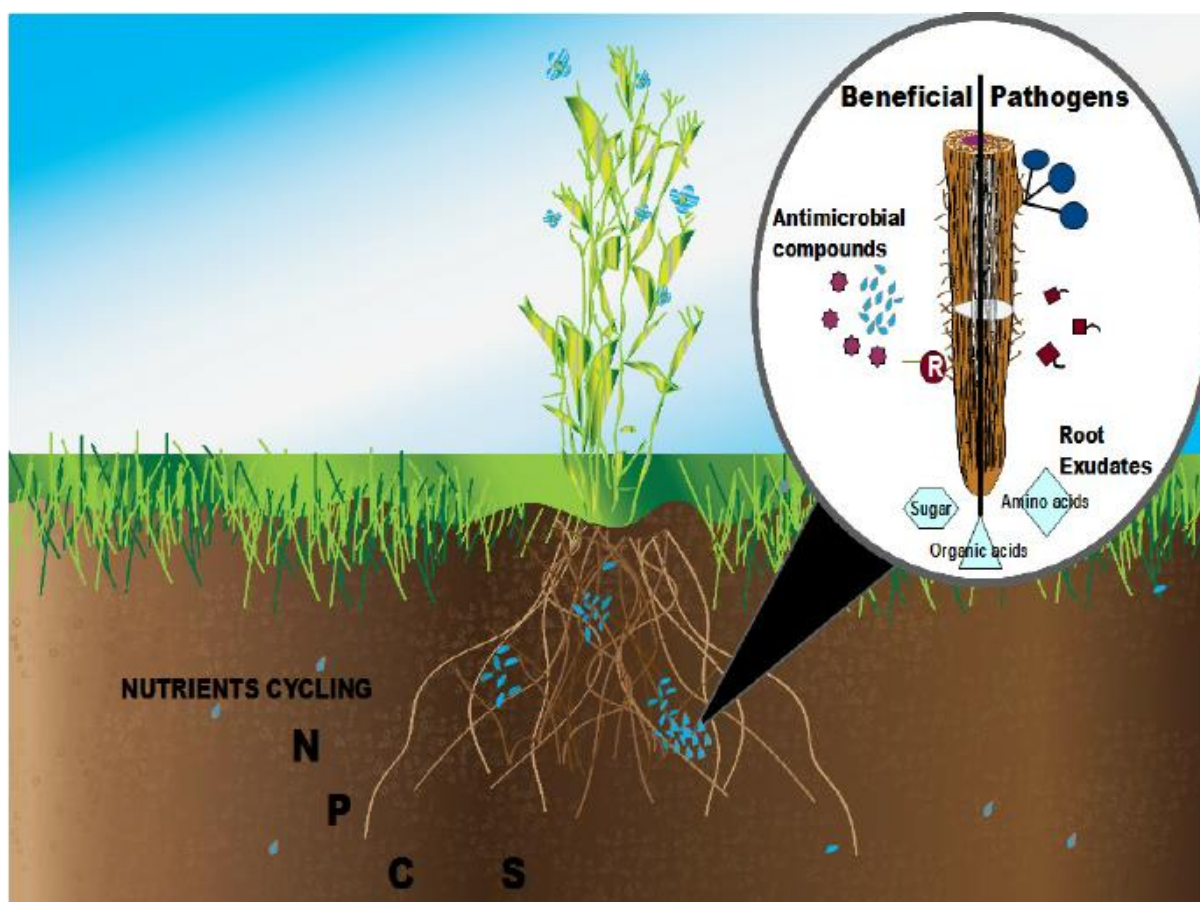
*sinense*, *S. barberi* e *S. edule*. Seu centro de origem, embora existam divergências a respeito, é provavelmente localizado no sudeste da Ásia (Mukherjee 1957). As variedades de cana-de-açúcar encontradas atualmente são híbridos provenientes de cruzamentos interespecíficos realizados no começo do século XX, envolvendo majoritariamente plantas de *S. officinarum* ( $2n = 80$ ), com maior conteúdo de sacarose, e *S. spontaneum* ( $2n = 40-128$ ), que possui maior tolerância à estresses bióticos e abióticos. Esses híbridos são atualmente conhecidos como *Saccharum spp.*, possuem mais de 100 cromossomos ( $2n = 100-130$ ) e são considerados plantas poliplóides e aneuplóides geneticamente complexas (Liu et al. 2016; Landell and Bressiani, 2008).

O colmo da cana-de-açúcar é tido como a principal fonte de matéria-prima para a fabricação de açúcar e etanol, tornando-o a parte da planta de maior relevância econômica. Ele é composto basicamente por fibra (10-12%) e caldo (82-90%), sendo que esta fibra é constituída por celulose, hemicelulose e lignina, enquanto o caldo é composto por água (75-82%) e sólidos solúveis (18-25%). Estes sólidos solúveis, por sua vez, são compostos por sais orgânicos e inorgânicos (1-2%), sacarose (14-24%), glicose (0,2-1%) e frutose (0-0,5%) (Cristina, Ferreira, and Omir 2008).

Por se tratar de uma planta com grande interesse econômico embutido, a cana-de-açúcar é um alvo intenso de pesquisas que buscam aumentar sua produtividade. Além disso, recentemente, um aumento mundial na demanda por energias renováveis fez crescer ainda mais o interesse pela produção da cana-de-açúcar, uma vez que o etanol produzido a partir dela é um combustível de fonte renovável e sua produção é sustentável. Adicionalmente, ele possui um custo de produção inferior ao conseguido com outras culturas, como o etanol do milho nos Estados Unidos da América e o etanol do trigo e da beterraba na Alemanha (Goldemberg 2008).

## **Rizosfera**

Na natureza, plantas e microrganismos do solo parecem ter evoluído juntos para superar estresses ambientais em seus habitats. Muitas interações planta-planta e planta-microbiota propiciam benefícios à ambos. Dessa forma, a rizosfera é normalmente caracterizada como sendo o local onde ocorrem eventos interativos entre plantas e microrganismos, interação essa que ocorre através da percepção de sinais liberados na forma de simples compostos químicos – Figura 2 (Dakora, Matiru, and Kanu 2015).



**Figura 2:** Exemplo das interações planta-microrganismos na zona da rizosfera. As raízes das plantas liberam exsudatos contendo açúcares, ácidos orgânicos e amino ácidos que podem atrair microrganismos. Em troca, esses microrganismos podem liberar moléculas com atividade antimicrobiana, liberar hormônios e aumentar a disponibilidade de nutrientes para as plantas (retirado de Rincon-Florez, Carvalhais, and Schenk 2013).

As plantas podem liberar entre 5 e 21% do carbono fixado pela fotossíntese em sua exsudação radicular, criando, ao seu redor, um ambiente diferente do encontrado no restante do solo (X.-F. Huang et al. 2014). Esse ambiente é ocupado por uma grande quantidade de organismos, entre eles nemátodos, fungos, bactérias, arqueias e herbívoros artrópodes (van Dam and Bouwmeester 2016). O efeito desse microbioma pode ser extremamente benéfico para a planta, como por exemplo a supressão de patógenos, promoção da mobilização de nutrientes, aumento da absorção de minerais e o estímulo ao crescimento vegetal (Chapelle et al. 2016; Dakora, Matiru, and Kanu 2015).

### **Microbiota**

Diversos estudos têm sido realizados na tentativa de identificar os microrganismos e os compostos responsáveis pelo aumento do crescimento da cana-de-açúcar. Dagnaw et al. (Dagnaw et al. 2015) conseguiu isolar 116 microrganismos de rizosfera de cana-de-açúcar e testá-los quanto à capacidade de solubilização de fosfato, tendo encontrado microrganismos capazes de solubilizar até 88,41 mg L<sup>-1</sup> de fosfato. Outro trabalho, este publicado por Beneduzi et al. (Beneduzi et al. 2013) conseguiu isolar 516 cepas dentre a rizosfera, a raiz e o caule de cana-de-açúcar. Dentre as cepas, em relação à promotores de crescimento vegetal, 368 isolados foram capazes de produzir compostos indólicos, 138 foram capazes de solubilizar fosfato e 390 foram produtores de compostos que auxiliam na absorção de ferro.

Embora muito esforço esteja sendo feito na busca pelo entendimento dos processos que ocorrem na rizosfera e interações específicas entre microrganismos e plantas modelos – como a simbiose *Rhizobium*-legumes – sejam bem compreendidos, a maior parte do microbioma das plantas e sua contribuição para o fenótipo estendido delas ainda não está definido. Além disso, o microbioma também é fortemente influenciado pelo genoma da

planta e pode ser considerado como uma extensão do mesmo para formar um segundo genoma ou, coletivamente, para formar um pan-genoma (Turner et al. 2013).

O método tradicional para o estudo desses microrganismos é a microbiologia clássica, a qual envolve o isolamento e cultura destes. Embora a obtenção de uma cultura pura de um organismo seja requerida para estudos mais detalhados de sua genética e fisiologia, técnicas dependentes de cultura não conseguem captar a ampla maioria da biodiversidade microbiana. Para superar este problema, a ecologia de microrganismos faz uso de técnicas moleculares, como o sequenciamento de *amplicon*, onde genes marcadores são amplificados à partir de DNA genômico pela técnica de PCR (*Polymerase Chain Reaction*), diretamente sequenciados e alinhados contra uma base de dados de referência para identificar a composição taxonômica de comunidades microbianas inteiras (Bokulich et al. 2016; Turner et al. 2013).

Entretanto, técnicas baseadas no sequenciamento do gene 16S rRNA são limitadas à bactérias e arqueias, pois esses são os únicos grupos portadores do gene. Para contornar esse problema, desenvolveu-se a técnica de sequenciamento de genomas inteiros ou *shotgun metagenomics*, expandindo, dessa forma, o alcance à grupos não portadores do gene 16S rRNA, como os eucariotos, fungos e até mesmo vírus que possuem DNA. Acoplada com análises computacionais e genomas de referência de alta qualidade, essa técnica possibilita uma alta resolução filogenética onde nenhuma diferença seria detectada usando-se apenas análises do gene 16S rRNA (Forster et al. 2016).

### **Metatranscriptômica**

Outra abordagem, a metatranscriptômica, envolve o mesmo processo de sequenciamento de metagenomas, mas ao invés de usar DNA, o alvo aqui são as moléculas de RNA que são reversamente transcritas à cDNA a partir de amostras de microrganismos combinadas, ou amostras ambientais. O resultado é um perfil da expressão gênica combinado

dos microrganismos ativos presentes na amostra, caracterizando o conjunto de genes sendo expressos por comunidades inteiras e, inclusive, pode-se analisar as mudanças na expressão gênica das comunidades em resposta à diferentes tipos de tratamentos (Bokulich et al. 2016).

Pesquisas recentes têm se utilizado das ferramentas da metatranscriptômica para aperfeiçoar processos dependentes de microrganismos como, por exemplo, o estudo feito por De Filippis et al. (De Filippis et al. 2016) onde a microbiota associada a maturação de queijos italianos tradicionais foi analisada. Estes pesquisadores conseguiram identificar genes diferencialmente expressos em condição de maturação normal e maturação sobre elevada temperatura, que estavam relacionados a proteólise, lipólise e catabolismo de aminoácidos e lipídios. Dessa forma, foi possível sugerir possíveis modificações no processo de fabricação dos queijos a fim de otimizar a eficiência da produção e a qualidade do produto.

Em outro estudo metatranscriptômico, Marzano e Domier (Marzano and Domier 2016) foram capazes de identificar 21 novos micovírus em filosferas de soja, podendo também organizá-los em dez diferentes grupos taxonômicos. Esses micovírus podem ser benéficos para as plantas debilitando fungos patogênicos e reduzindo a severidade de doenças causadas pelos mesmos. Assim, o trabalho abre novas perspectivas, como também a possibilidade de minimização de perdas na produção da soja devido a patógenos.

A metatranscriptômica supera muitas limitações inerentes a outras técnicas, como por exemplo os microarranjos de DNA, propiciando acesso direto ao perfil transcricional de comunidades. Essa abordagem se fez possível através dos avanços na tecnologia do sequenciamento em larga escala que gera dados ordens de magnitude maiores que os métodos de sequenciamento anteriores e com uma fração dos custos. Isso fez com que a metatranscriptômica propiciasse conhecimentos antes inalcançáveis sobre comunidades de microrganismos (Carvalhais et al. 2013).

### **Montagem dos dados de metatranscriptoma**

Para estudar-se o metatranscriptoma de uma determinada comunidade, é necessário que seja feito o sequenciamento do RNA presente nas amostras de interesse. A tecnologia mais utilizada atualmente é o RNA-seq que, mais frequentemente, leva à geração de *reads* com tamanhos aproximados entre 100 e 150 pares de base. Essas *reads* são utilizadas por programas de computador chamados montadores, que, a partir delas, reconstroem os transcritos originais. O processo de montagem é normalmente comparado à uma quebra cabeças onde as *reads* são as peças e os transcritos são as imagens obtidas pela junção das peças individuais. A montagem do metatranscriptoma é um processo computacionalmente custoso e é a parte que mais exige recursos computacionais em um *pipeline* de análise de dados de RNA-seq (Celaj et al. 2014; Pop 2009).

Os programas que realizam a montagem de metatranscriptomas se baseiam em uma estrutura de dados denominada grafos de Bruijn (dBG), que foi originalmente introduzida na área de montagem de genomas no ano de 2001 por Pevzner et al. em um montador chamado EULER (Pevzner, Tang, and Waterman 2001). Entretanto, algoritmos baseados em dBG só passaram a ser amplamente empregados com o surgimento das tecnologias de sequenciamento de nova geração, quando o número de *reads* produzidos atingiu a casa dos milhões de *reads* mas, por outro lado, o tamanho das *reads* caiu cerca de uma ordem de magnitude em comparação com a tecnologia anterior, que era o sequenciamento pela tecnologia de Sanger.

Juntamente com o desenvolvimento promovido pelas tecnologias de sequenciamento de nova geração e seu impacto no barateamento dos custos do sequenciamento, surgiram também novos desafios em relação ao processo de montagem das sequências, uma vez que a montagem de transcriptomas difere da montagem de genomas em alguns pontos chaves: (i) alguns transcritos estão presentes em baixa abundância, enquanto outros podem apresentar milhares de cópias; (ii) a cobertura de *reads* pode ser desigual ao longo do comprimento do



transcrito; (iii) *reads* com erros de sequenciamento provenientes de transcritos altamente expressos podem ser mais abundantes que *reads* corretas, porém de transcritos de baixa abundância; (iv) transcritos provenientes de loci adjacentes podem conter sobreposições e serem erroneamente fundidos, formando transcritos quiméricos; (v) um gene pode expressar vários transcritos diferentes por meio de *splicing* alternativo; (vi) ambiguidade produzida por sequências repetidas em diferentes transcritos (Grabherr et al. 2011). Dessa forma, foi necessário o desenvolvimento de algoritmos específicos para a montagem de transcriptomas.

Diversos *softwares* foram desenvolvidos para a montagem de transcriptomas, dentre eles Trans Abyss (Birol et al. 2009), Trinity (Grabherr et al. 2011), Oases (Schulz et al. 2012), Soapdenovo-Trans (R. Luo et al. 2012), IDBA-tran (Peng et al. 2013) e Bridger (Chang et al. 2015) e todos eles utilizam estruturas de dados baseados em grafos de Bruijn. Embora o campo de montagem de transcriptomas tenha se desenvolvido na última década, a montagem de metatranscriptomas ainda é pouco explorada em termos de algoritmos específicos para abordagem do problema, contando com apenas uma ferramenta especificamente desenvolvida para este fim, o montador IDBA-MT (Leung et al. 2013). Embora este seja o único montador específico para dados metatranscriptômico, sua performance ainda não o levou a ser comumente empregado na literatura.

### **Expressão gênica diferencial**

A análise da expressão gênica diferencial se trata do estudo de genes que apresentam diferenças no nível de expressão entre diferentes condições ou tratamentos. No caso mais simples, o objetivo dessa análise é comparar o nível de expressão de genes entre duas condições, por exemplo, tratado versus não tratado ou *wild type* versus mutante. Também é possível analisar-se desenhos experimentais mais complexos, com múltiplos níveis, por

exemplo, múltiplos mutantes, diferentes doses de uma droga ou uma série temporal (S Anders et al. 2013).

A estratégia usada para a detecção de genes diferencialmente expressos baseia-se na contagem de reads que se alinham a cada um dos genes, gerando uma tabela de contagem na qual aplica-se análises estatísticas visando o descobrimento de mudanças quantitativas nos níveis de expressão gênica entre grupos experimentais (S Anders et al. 2013). Apesar do poder analítico gerado pela possibilidade de se analisar o nível de expressão de todo o conjunto de genes de um organismo de forma simultânea, a análise desse tipo de dados apresenta uma série de dificuldades.

Algumas dessas dificuldades são inerentes à própria tecnologia de sequenciamento, por exemplo, a tecnologia Illumina apresenta erros de substituição a uma taxa entre 0,5 a 2,5% das bases sequenciadas, sendo que essa taxa aumenta ainda mais em direção ao final 3' das *reads* (Kelley, Schatz, and Salzberg 2010). Outros tipos de erro são vistos “entre amostras” em um experimento de RNA-Seq. Primeiro, a profundidade de sequenciamento ou o tamanho das bibliotecas são normalmente diferentes entre diferentes amostras, o que implica no fato de que a contagem observada não pode ser diretamente comparada entre as amostras. Dessa forma, mesmo na ausência de qualquer expressão diferencial, se uma amostra é sequenciada duas vezes mais que a outra, é esperado que todos os seus genes apresentem uma contagem duas vezes maior (Soneson and Delorenzi 2013).

A forma mais fácil de abordar a questão de bibliotecas de tamanhos diferentes é simplesmente alterar a escala ou fazer uma reamostragem das contagens para obter bibliotecas de tamanhos iguais para todas as amostras. Entretanto, mesmo que as bibliotecas tenham tamanhos iguais, as contagens provenientes de RNA-seq representam abundâncias relativas entre os seus genes, ou seja, alguns poucos genes altamente expressos podem contribuir com uma grande parte das *reads* sequenciadas, deixando apenas algumas *reads*

para os genes restantes. Para lidar com essa problemática, esquemas de normalização que levem em consideração fatores de correção amostra específicos se fazem necessários. Um exemplo de método de normalização que leva isso em consideração é a método de normalização TMM (do inglês *Trimmed mean of M-values*) (Robinson and Oshlack 2010) que estima fatores de escala entre amostras que podem ser incorporados nos modelos estatísticos já utilizados para análise de expressão diferencial (Soneson and Delorenzi 2013).

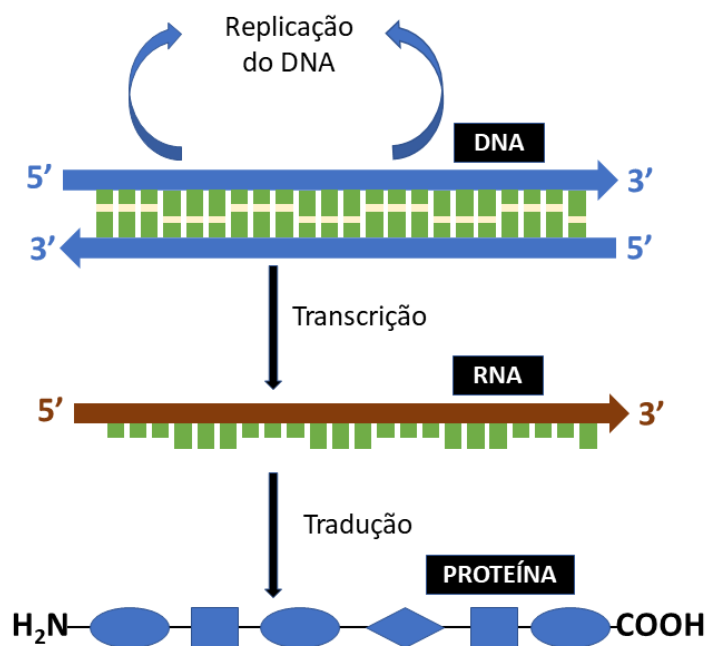
O método de normalização por TMM, juntamente com mais oito métodos, foram analisados em um amplo estudo (Dillies et al. 2013) onde apenas o TMM e um método de normalização empregado no software DESeq (Simon Anders and Huber 2010), demonstrou resultados satisfatórios em todas as métricas analisadas durante o estudo (Soneson and Delorenzi 2013), demonstrando sua acurácia e robustez em experimentos de RNA-seq.

Dentre as diversas implementações existentes para a análise de expressão diferencial, o pacote do R denominado NOISeq (Tarazona et al. 2011) destaca-se por modelar empiricamente o ruído nos dados de contagem, é robusto em relação à profundidade de sequenciamento e pode funcionar na ausência de réplicas. Ao contrário da maioria dos softwares existente para análise de expressão diferencial, o NOISeq não se baseia em premissas iniciais feitas à respeito da distribuição probabilística dos dados, fato este que nem sempre é verdadeiro e pode levar à resultados errôneos. Esse método foi utilizado em uma série de estudos com dados de metatranscriptoma (Yost et al. 2015; F. Luo et al. 2014; Stewart et al. 2016; Duran-Pinedo et al. 2014) e, também, foi comparado com os softwares de expressão diferencial mais comumente empregados na literatura e obteve bons resultados (Soneson and Delorenzi 2013; Nookaew et al. 2012).

### **Redes de co-expressão gênica**

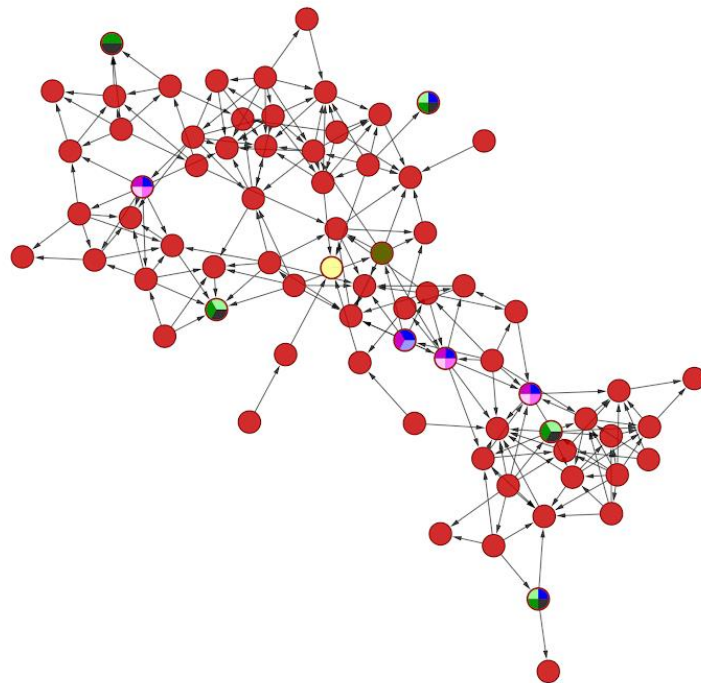
De acordo com o dogma central da biologia molecular (Figura 3), o fluxo de informação genética ocorre a partir do DNA, que serve de molde para a transcrição do RNA

que, por sua vez, é traduzido em proteína (Hirao 2006). Esse fluxo de informação pode ser regulado em cada uma das etapas existentes, dessa maneira, as proteínas que serão sintetizadas podem ser reguladas (1) com um controle de quando e com que frequência um gene é transcrito, (2) com um controle sobre o processamento e splicing dos transcritos de RNA, (3) com a seleção de quais RNAs mensageiros serão exportados do núcleo para o citosol, (4) com a seleção de quais mRNAs serão traduzidos pelo ribossomo, (5) com a degradação de determinadas moléculas de mRNA no citoplasma ou (6) ativando, desativando, degradando ou colocando as proteínas já traduzidas em locais específicos da célula (Alberts et al. 2014).



**Figura 3:** O dogma central da biologia, ilustra o fluxo de informação que ocorre do DNA para o RNA (transcrição) e do RNA para a proteína (tradução). Esse processo ocorre em todas as células vivas, independentemente de sua classificação taxonômica (Alberts et al. 2014).

Essa relação de interdependência entre os genes em um complexo sistema de interação e regulação pode ser prevista por meio de métodos estatísticos e computacionais, gerando uma representação da influência recíproca de genes. Uma rede de co-expressão (GCN do inglês *Gene Co-expression Network*) conecta pares de genes (nós) que são significativamente co-expressos através de diferentes tratamentos em um determinado estudo (Figura 4). Dessa forma, o primeiro passo para a criação de uma GCN é contabilizar a correlação entre cada um dos pares de genes analisados em uma tabela matriz de correlação. O segundo passo é escolher um valor de corte acima do qual todos os pares de genes são conectados. GCNs são de grande interesse biológico, uma vez que genes co-expressos provavelmente são controlados pelos mesmos mecanismos de regulação da transcrição, e assim provavelmente têm funções relacionadas ou são membros de uma mesma via metabólica ou complexo protéico (Weirauch 2011).



**Figura 4:** Exemplo de rede de co-expressão gênica, mostrando a rede gerada a partir de dados de RNA-Seq. Os nós coloridos representam diferentes funções biológicas associadas a cada um dos transcritos e as arestas cinza representam as relações entre os nós.

Por fim, a dissertação aqui apresentada se organiza de forma que no primeiro capítulo serão apresentados os testes de softwares e pipelines utilizados para se definir a melhor estratégia de análise dos dados. Além disso, serão apresentados breves resultados obtidos com a análise do RNA ribossomal. Já o capítulo dois trata da análise o RNA mensageiro através das estratégias estabelecidas no primeiro capítulo. Finalmente, o terceiro capítulo disserta sobre a rede de co-expressão gênica construída com os transcritos diferencialmente expressos em resposta aos diferentes tratamentos de estresse hídrico aplicados.

## OBJETIVOS

O objetivo geral foi avaliar e determinar o perfil taxonômico e funcional da microbiota ativa associada à rizosfera de cana-de-açúcar submetida ao estresse hídrico. Adicionalmente, foram gerados dados que podem dar subsídio a estudos futuros que busquem descrever a modulação da expressão dos genes da microbiota de cana-de-açúcar em diferentes cenários.

Finalmente, os objetivos específicos deste estudo foram:

1. Caracterizar a comunidade microbiana ativa da rizosfera de cana-de-açúcar submetida a tratamentos de estresse hídrico;
2. Caracterizar o perfil funcional da comunidade microbiana da rizosfera de cana-de-açúcar nas condições de estresse hídrico;
3. Analisar a expressão gênica diferencial gerada pelos tratamentos de estresse hídrico;
4. Analisar as redes de co-expressão gênica construídas através dos dados de expressão gênica diferencial gerados nos diferentes tratamentos.

# CAPÍTULO I

---

## Análises Preliminares



## Introdução

Microrganismos presentes no solo tem um papel muito importante na aquisição de nutrientes pelas plantas. Com as previsões futuras de mudança no padrão de chuvas em todo o globo, aumento da temperatura global e aumento da concentração de CO<sub>2</sub> na atmosfera derivadas das mudanças climáticas, existe um crescente interesse nas comunidades microbianas do solo com o objetivo de aumentar a produção e diminuir perdas causadas por estresses bióticos e abióticos derivados das mudanças climáticas (León-Sánchez et al. 2018).

Rizobactérias promotoras do crescimento vegetal (PGPR – do inglês *Plant Growth Promoting Rhizobacteria*) são um grupo heterogêneo de bactérias que colonizam as raízes das plantas e beneficiam as mesmas através da produção e secreção de um arsenal de moléculas regulatórias no ambiente da rizosfera. As PGPR podem estimular o crescimento vegetal de forma direta ou indireta. O estímulo indireto incluiu supressão de microrganismos deletérios ou patógenos através da produção de substâncias antagônicas como cianeto, proteases, quitinases, antibióticos, competindo por nutrientes e espaço na região da rizosfera e induzindo resistência sistêmica nas plantas contra um amplo espectro de patógenos. Por outro lado, o estímulo direto envolve a disponibilização, para as plantas, de compostos como nitrogênio fixado, fitohormônios como ácido indol-3-acético, ácido indol butírico, citocininas, giberelinas ou solubilização de nutrientes como ferro, fósforo e zinco do solo (Kalam et al. 2017).

Ao invés de focar-se em uma única resposta funcional à um determinado tratamento, análises de metatranscriptomas medem a expressão de milhares de transcritos de uma única vez através da detecção do RNA mensageiro. A metatranscriptômica tem sido usada para investigar a diversidade funcional de comunidades microbianas em vários ambientes, assim como examinar a resposta da comunidade microbiana à mudanças em determinados fatores

ambientais (Newman et al. 2016). Por exemplo, Urich et al. (Urich et al. 2008) utilizou estratégias metatranscriptômica para estudar simultaneamente a diversidade taxonômica e funcional de uma comunidade de solo arenoso de uma área de conservação. Já Shrestha et al. (Shrestha et al. 2009) fez uso de técnicas de metatranscriptômica para comparar a atividade microbiana entre solos de arroz com e sem oxigênio. Esses estudos exemplificam a viabilidade e utilidade de estratégias metatranscriptômicas no estudo da atividade de comunidade microbianas de solo. Dessa forma, o objetivo das análises empregadas nesse capítulo, denominado “Análises Preliminares”, foi definir as melhores estratégias e softwares para analisar os dados provenientes do sequenciamento. Além disso, fez-se uma breve análise dos dados de RNA ribossomal que não entraram nas análises do capítulo II.

## **Materiais e métodos**

### **Experimento de estresse hídrico e coleta das amostras**

A variedade selecionada para o experimento foi o cultivar comercial CTC9001, fornecido pelo Centro de Tecnologia Canavieira (CTC, Piracicaba). Essa variedade foi escolhida, pois apresenta boa adaptação às condições de seca, e pode ter uma associação preferencial a determinadas bactérias. Dessa forma, plantas de cana-de-açúcar de quatro meses foram mantidas em casa de vegetação por três meses sob irrigação normal. As plantas foram, então, submetidas ao estresse hídrico, permanecendo sem irrigação durante um período de 2, 8 ou 12 dias. Foi realizada a coleta das amostras de rizosfera de plantas na condição inicial, não submetida ao estresse hídrico (WS0d), e das plantas não irrigadas por 2 (WS2d), 8 (WS8d) e 12 dias (WS12d) a partir do início do experimento, assim como amostras de solo de vasos sem plantas (Bulk).

Para a coleta das amostras, as plantas foram retiradas do solo e levemente agitadas para a retirada de aglomerações de solos não associados às raízes. O solo ainda aderido às raízes,

então, foi retirado por agitação vigorosa. As amostras foram imediatamente estocadas em solução de conservação de RNA de solo LifeGuard™ (MoBio Laboratories, Inc, Carlsbad, CA).

### **Preparação das amostras**

Em seguida, foi realizado o isolamento do RNA total do solo da rizosfera. Para o isolamento do RNA foi utilizado o RNA PowerSoil® Total RNA Isolation Kit (MoBio Laboratories, Carlsbad, CA, USA). A quantificação do RNA total foi feita no Qubit (Invitrogen) e a qualidade do RNA verificada em gel desnaturante de agarose. O RNA foi armazenado a -80°C.

### **Sequenciamento do metatranscriptoma**

A partir do RNA extraído das amostras foi feito o sequenciamento do metatranscriptoma da microbiota associada à rizosfera de cana-de-açúcar submetida a estresse hídrico. Aproximadamente 2 µg de RNA total de cada amostra foram usados para o ‘enriquecimento’ do RNA mensageiro bacteriano (mRNA) utilizando o MICROBExpress™ Bacterial mRNA Enrichment Kit (Ambion®). O mRNA foi posteriormente fragmentado utilizando-se cátions divalentes a 94°C e a seguir, a primeira fita de cDNA foi sintetizada utilizando-se transcriptase reversa e *primers* aleatórios. A síntese da segunda fita de cDNA foi feita com a DNA polimerase I. Adaptadores específicos foram inseridos nos cDNAs. As bibliotecas foram então enriquecidas por 12 ciclos de amplificação e, a seguir, os clusters foram gerados. As bibliotecas foram quantificadas utilizando-se o kit Agilent DNA high-sensitivity (Agilent) em um Bioanalyzer. Quantidades equimolares das bibliotecas foram reunidas em uma *flowcell* e o sequenciamento (*paired-end*) feito na plataforma HiSeq 2500® (Illumina, Inc.).

### **Controle de qualidade do sequenciamento**

Os reads gerados no sequenciamento foram filtrados através do programa NGS QC Toolkit (Patel and Jain 2012) com valores de *cutoff* para qualidade mínima de 20 e tamanho mínimo de 70 pares de base e analisados com o programa FASTQC (Andrews 2010), em seguida foi realizada a remoção de *reads* contaminadas por adaptadores através do programa FASTX-Toolkit (Gordon and Hannon 2010).

### **Identificação do RNA ribossomal**

A identificação *in silico* do rRNA foi feita através do *software* riboPicker versão 0.4.3 (Schmieder, Lim, and Edwards 2012), utilizando o banco de dados de sequências ribossomais não redundantes disponibilizado pelo software. As *reads* provenientes de RNA ribossomal foram montadas com o programa EMIRGE versão 0.61.1 (Miller et al. 2011) e estatísticas básicas de montagem foram geradas para cada uma das montagens individuais através do software Transrate (Smith-Unna et al. 2016). Em seguida, o *software* Bowtie2 (Langmead and Salzberg 2012) foi utilizado para mapear as reads nos transcritos montados, gerando arquivos no formato BAM que foram utilizados como *input* para o *software* featureCounts (Liao, et al., 2013), esse segundo gerou uma tabela de contagem que representa a abundância de cada um dos transcritos.

Os transcritos montados foram alinhados contra o banco de dados de sequências de RNA ribossomal Silva (Pruesse et al. 2007) através do software BLASTN (Altschul et al. 1990) com os parâmetros *-outfmt 6, -num\_alignments 30*. Os valores de contagem para cada um dos transcritos e os alinhamentos do BLASTN contra o banco de dados Silva foram passados para o programa MEGAN6 Community Edition (Huson et al. 2016) através de um *script* Python desenvolvido *in house*, com os parâmetros *Min Score=120, Top Percent=10* e

*Min Support=5*, que possibilitam uma classificação taxonômica conservativa como sugerido por Mitra et al. (Mitra, Stärk, and Huson 2011).

### **Análise do RNA mensageiro**

As *reads* da amostra Bulk não classificadas como provenientes de RNA ribossomal foram montadas com diversos montadores e estratégias de montagem: IDBA-UD (Peng et al. 2012) versão 1.1.3 com os parâmetros *--mink=20*, *--maxk=100* e *--step=20*; SOAPdenovo-Trans (Y. Xie et al. 2014) versão 1.03 com o parâmetro *-k* variando de 21 a 81; Trinity (Grabherr et al. 2011) versão 2.6.6 com os parâmetros *default*; Bridger (Chang et al. 2015) versão r2014-12-01 com os parâmetros *default*; Oases (Schulz et al. 2012) versão 0.2.09 com o parâmetro *-k* variando de 21 a 81. Algumas montagens foram analisadas como software Transrate versão 1.0.3 com os parâmetros *default* e as melhores montagens de cada montador foram mescladas usando-se os *softwares* CAP3 (X. Huang and Madan 1999) versão 02-10-15 com os parâmetros *default* e o *software* CD-HIT-EST (W. Li and Godzik 2006) versão 4.6.6 com os parâmetros *-c=1* e *-M=30000*, conforme descrito em (Nakasugi et al. 2014).

Comparações entre as montagens foram feitas para a amostra Bulk com os programas Transrate versão 1.0.3 com os parâmetros *default*; Bowtie2 versão 2.3.1 com os parâmetros *default*; Busco (Simão et al. 2015) versão 2.0.1 com o parâmetro *-m=tran*; SUPER-FOCUS (G. G. Z. Silva et al. 2015) versão 0.27 com os parâmetros *-p=1*, *-n=0*, *-mi=60*, *-ml=15*, *-e=0.00001*, *-db=DB\_100* e *-a=rapsearch*; Prodigal (Hyatt et al. 2010) versão 2.6.3 com o parâmetro *-p=meta*. A montagem com todas as amostras foi feita com o montador IDBA-UD com os parâmetros *--mink=20*, *--maxk=100* e *--step=20*.

Para a predição de ORFs em cima da montagem com todas as amostras, testou-se dois *softwares*, FragGeneScan (Rho, Tang, and Ye 2010) versão 1.30 com os parâmetros *default* e o *software* Prodigal versão 2.6.3 com o parâmetro *-p=meta*.

A determinação da melhor forma de fazer o mapeamento das *reads* nos transcritos montados foi feita através da comparação de três *softwares*: Bowtie2 versão 2.3.1 com o parâmetro `--local`; BWA (H. Li and Durbin 2009) versão 0.7.17 com os parâmetros *default*; Salmon versão 0.8.2 com o parâmetro `--meta`. Sendo que tanto para Bowtie2 quanto para BWA, as tabelas de contagem foram extraídas dos arquivos BAM gerados com a utilização do *software* featureCounts versão 1.5.3 com os parâmetros `-p` e `-f`.

Por fim, para determinar o conjunto de transcritos diferencialmente expressos, foi utilizado o *software* NOISeq (Tarazona et al. 2011) versão 2.22.1.

## **Resultados e discussão**

### **Controle de qualidade do sequenciamento**

O controle de qualidade das *reads* foi feito de forma a retirar *reads* com qualidade menor que 20 (Q-score  $\geq 20$ ) e comprimento menor que 70 pares. Além disso, retirou-se *reads* com alto conteúdo de sequências de adaptadores Illumina. Das dez bibliotecas sequenciadas, sete apresentaram boa qualidade, sendo possível um aproveitamento de mais de 87% das *reads* dessas bibliotecas. Entretanto, três amostras, a segunda réplica da amostra Bulk (Bulk 2), e as duas amostras sem irrigação por 8 dias (WS8d 1 e WS8d 2) obtiveram uma grande proporção de *reads* de baixa qualidade ou contaminados por adaptadores, de forma que apenas puderam ser utilizadas 3,29%, 11,04% e 15,87% das *reads* dessas amostras, respectivamente (Tabela 1). Embora essas três últimas amostras tenham obtido *reads* com menor qualidade, o número de *reads* que passaram pela filtragem ainda foi alto o suficiente para que tanto WS8d 1 quanto WS8d 2 pudessem ser utilizadas na etapa seguinte das análises. Entretanto, a amostra Bulk 2 não pode ser utilizada nas etapas posteriores pois restaram poucas *reads* de alta qualidade.

Sample	Total number of paired reads	Number of paired reads Q-score<20 & size<70bp	Q-score (% , Q < 20)	Number of high adapter content paired reads	Adapter content (%)	Total number of useful paired reads	Total number of useful paired reads (%)
Bulk 1	30.825.567	1.157.310	3,75	2.564.159	8,32	27.104.098	87,93
Bulk 2	21.260.830	13.076.354	61,50	7.484.500	35,20	699.976	3,29
WS0d 1	21.157.956	929.634	4,39	1.302.070	6,15	18.926.252	89,45
WS0d 2	31.611.155	1.172.459	3,71	2.766.753	8,75	27.671.943	87,54
WS2d 1	20.994.459	768.629	3,66	1.728.085	8,23	18.497.745	88,11
WS2d 2	30.346.136	1.945.011	6,41	1.611.396	5,31	26.789.729	88,28
WS8d 1	74.928.514	43.058.628	57,47	23.596.270	31,49	8.273.616	11,04
WS8d 2	61.877.643	18.604.977	30,07	33.449.643	54,06	9.823.023	15,87
WS12d 1	32.842.412	2.389.503	7,28	1.113.922	3,39	29.338.987	89,33
WS12d 2	21.764.254	989.893	4,55	1.359.810	6,25	19.414.551	89,20

**Tabela 1:** Resultado da etapa de controle de qualidade das reads.

### Identificação do RNA ribossomal

A identificação de reads provavelmente provenientes de RNA ribossomal foi realizada pelo *software* riboPikcer e o resultado pode ser observado na Tabela 2. Por terem resultado em uma pequena quantidade de reads se comparadas com as outras amostras, as amostras Bulk 1, WS8d 1 e WS8d 2 não foram usadas na etapa de montagem das *reads* de RNA ribossomal. A diferença na quantidade de RNA ribossomal nas amostras pode ter sido influenciada pela qualidade do RNA utilizado, uma vez que o kit de remoção de RNA ribossomal utilizado é fortemente influenciado pelo estado de degradação do RNA utilizado (He et al. 2010).

Finalmente, a amostra Bulk 2 teve 61,5% das reads removidas nas etapas de filtragem por baixa qualidade e trimagem, além disso, outros 35,2% foram removidas por apresentarem alta quantidade de adaptadores Illumina. Isso fez que apenas 3,29% das *reads* sequenciadas fossem aproveitáveis e, dessa forma, essa amostra foi removida das análises subsequentes.

Sample	Total number of useful paired reads	rRNA paired reads	rRNA (%)	mRNA paired reads	mRNA (%)
Bulk 1	27.104.098	22.585.544	83,33	4.518.554	16,67
Bulk 2	699.976	9.477 *	1,35	690.499 **	98,65
WS0d 1	18.926.252	14.809.099	78,25	4.117.153	21,75
WS0d 2	27.671.943	23.587.928	85,24	4.084.015	14,76
WS2d 1	18.497.745	15.623.393	84,46	2.874.352	15,54
WS2d 2	26.789.729	15.841.622	59,13	10.948.107	40,87
WS8d 1	8.273.616	98.632 *	1,19	8.174.984	98,81
WS8d 2	9.823.023	1.527.743 *	15,55	8.295.280	84,45
WS12d 1	29.338.987	11.140.091	37,97	18.198.896	62,03
WS12d 2	19.414.551	13.813.556	71,15	5.600.995	28,85

**Tabela 2:** Remoção das *reads* de RNA ribossomal realizada pelo *software* riboPicker.

\*Not used for 16S prediction. \*\*Not used for mRNA assembling.

## Análise do RNA ribossomal

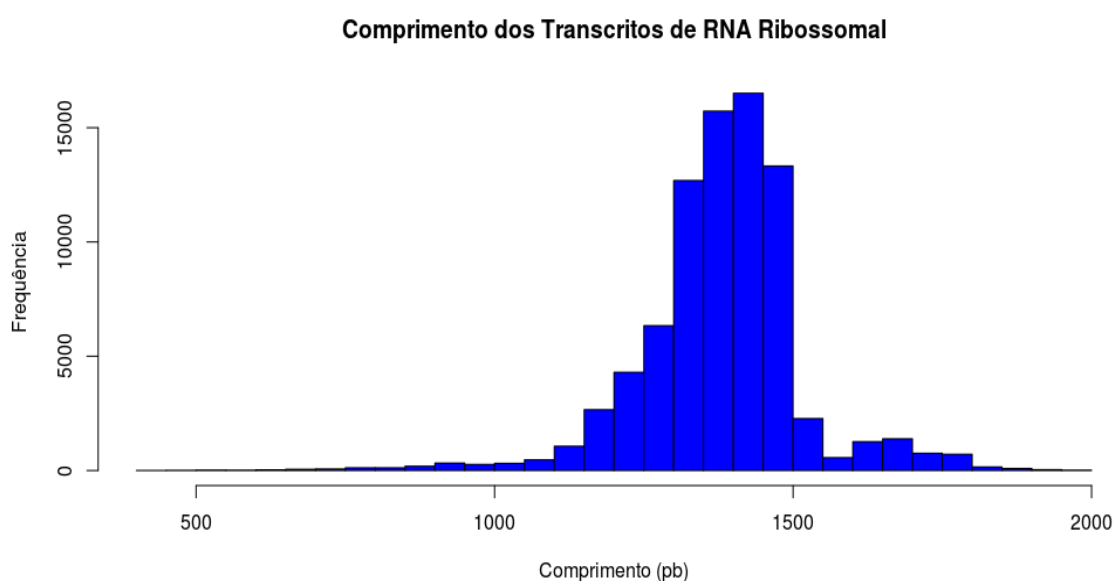
### Montagem

As *reads* identificadas pelo riboPicker como provenientes de RNA ribossomal foram montados pelo software EMIRGE, um montador especializado em montagem de genes de RNA ribossomal a partir de *reads* curtas. Ao total, foram montados por volta de 11 a 14 mil transcritos de rRNA diferentes para cada uma das amostras. Esses transcritos apresentaram comprimento relativamente semelhante, tendo o N50 variando de 1395 a 1401, como mostra a Tabela 3. No total, foram montados 81.881 transcritos com distribuição de tamanho conforme a Figura 1 e comprimento médio de 1.381 pares de base. É possível notar que a maioria dos transcritos montados pelo EMIRGE teve um comprimento próximo a 1500 pb, que é o comprimento estimado de um gene 16S, tendo também transcritos com comprimentos próximos a 1800 pb que é o comprimento estimados do gene 18S.



Sample	N. Sequences	N50	Average Size	Largest	Smallest	GC Content
Bulk 1	9822	1399	1385	1976	473	0.56
Ws 0d 1	11179	1391	1372	1972	509	0.55
Ws 0d 2	13620	1401	1387	1976	470	0.55
Ws 2d 1	11205	1395	1387	1988	501	0.55
Ws 2d 2	12796	1396	1378	1967	449	0.55
Ws 12d 1	11381	1400	1383	1963	487	0.56
Ws 12d 2	11878	1396	1376	1957	490	0.56

**Tabela 3:** Estatísticas básicas das montagens individuais obtidas através do *software* EMIRGE para as *reads* de rRNA de cada uma das amostras.



**Figura 1:** Distribuição de comprimentos dos transcritos montados com o *software* EMIRGE. Nota-se picos próximos a 1500 e 1800 pb, que são os comprimentos estimados dos genes 16S e 18S, respectivamente.

#### *Mapeamento das Reads nos Contigs de RNA Ribossomal*

O *software* Bowtie2 foi utilizado para mapear as *reads* de RNA ribossomal contra os transcritos montados pelo EMIRGE e, dessa forma, obter-se uma medida de abundância de cada um dos transcritos montados. Foi possível mapear por volta de 28 a 41% das *reads* nos transcritos montados (Tabela 4). Os dados do mapeamento das *reads* nos transcritos gerados pelo Bowtie2, na forma de arquivos SAM, foram contabilizados na forma de uma tabela de

contagem pelo *software* featureCounts e utilizados na etapa seguinte de classificação taxonômica dos transcritos de RNA ribossomal.

Sample	N. Sequences	N. Reads	Mapping %
Bulk 1	9.822	22.585.544	28.55%
Ws 0d 1	11.179	14.809.099	28.57%
Ws 0d 2	13.620	23.587.928	31.88%
Ws 2d 1	11.205	15.623.393	41.04%
Ws 2d 2	12.796	15.841.622	31.95%
Ws 12d 1	11.381	11.140.091	32.18%
Ws 12d 2	11.878	13.813.556	30.28%

**Tabela 4:** Mapeamento das *reads* nos transcritos de rRNA através do *software* Bowtie2.

A variação entre 28 e 41% da porcentagem de reads que mapearam nos transcritos indica que o *software* EMIRGE provavelmente não utilizou todas as *reads* classificadas como *reads* de RNA ribossomal para a montagem dos transcritos. Isso pode ser explicado pelo fato de que o EMIRGE apenas monta as sequências da subunidade menor do ribossomo - 16S para procariotos e 18S para eucariotos - e não utiliza as reads provenientes da subunidade maior - 5S e 23S para procariotos e 5.8S e 28S para eucariotos (Miller et al. 2011).

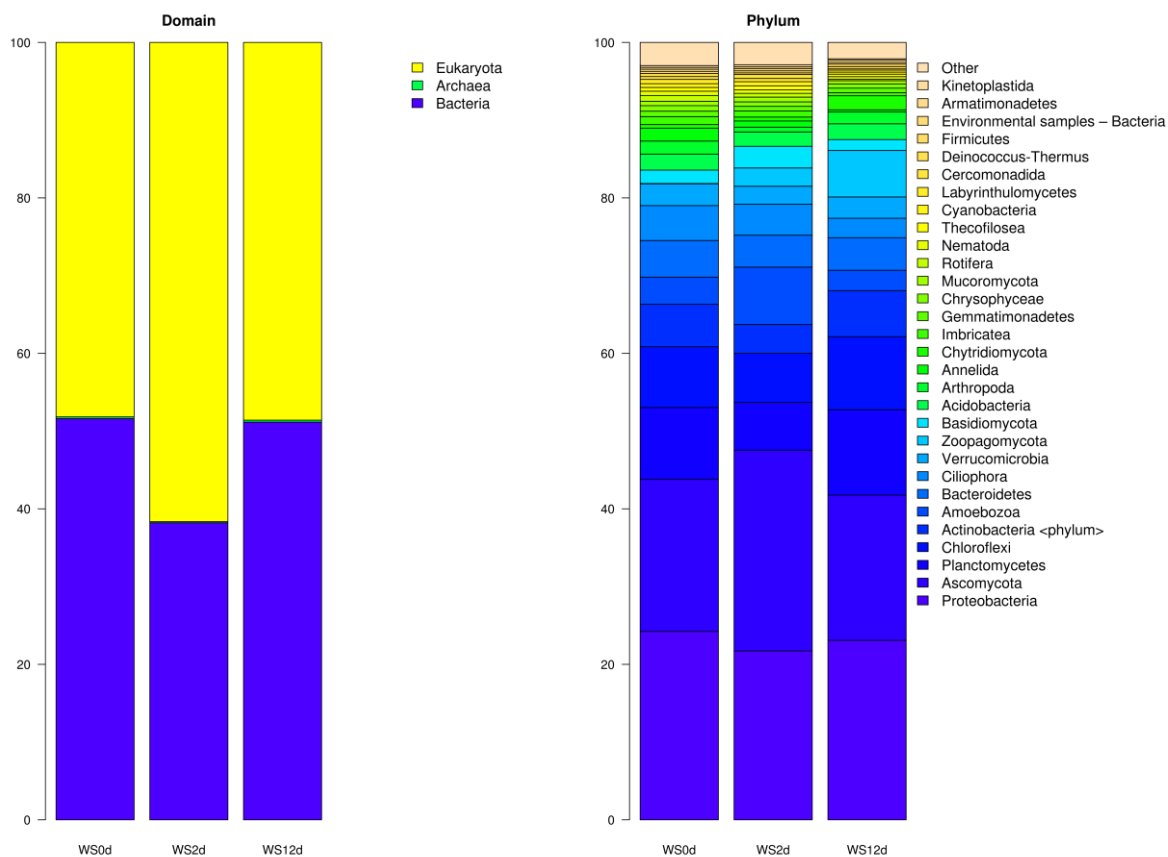
#### *Análise taxonômica do RNA ribossomal*

A maioria dos transcritos provenientes de rRNA anotados são de eucariotos e bactérias, com apenas uma pequena quantidade de transcritos sendo anotados como arqueia (Figura 2). Essa maior proporção de eucariotos em relação aos dados do RNA mensageiro (resultado apresentado no capítulo II) provavelmente se deve ao fato de que o kit de depleção de RNA ribossomal usado, apenas remove sequências provenientes de bactérias, inclusive removendo sequências de forma desproporcional uma vez que o kit funciona por meio de hibridização subtrativa e a remoção ou não das sequências depende da existência de

sequências complementares no kit. Dessa forma, as proporções do RNA ribossomal restante são totalmente alteradas.

Essa característica de alteração das abundâncias do rRNA foi descrita na literatura (He et al. 2010), onde os autores compararam a eficiência de dois métodos diferentes para a remoção de RNA ribossomal. No estudo, foram criadas duas comunidades sintética com cinco espécies diferentes em cada a partir da junção de quantidades iguais de RNA extraído independentemente de isolados microbianos. À essa junção de RNA, foi aplicado o kit de depleção seguido de sequenciamento. Ao final do estudo, pôde-se observar que as abundâncias relativas entre as cinco espécies da comunidade sequenciada foram totalmente alteradas.

Embora a abundância relativa de transcritos de bactérias e arqueias seja alterada, espera-se que não ocorram grandes alterações nas abundâncias de transcritos provenientes de eucariotos, uma vez que o kit não possuiu sequências específicas para esse grupo de organismos.



**Figura 2:** Abundância relativa nos níveis de domínio e filo dos transcritos com taxonomia atribuída a partir do RNA ribossomal.

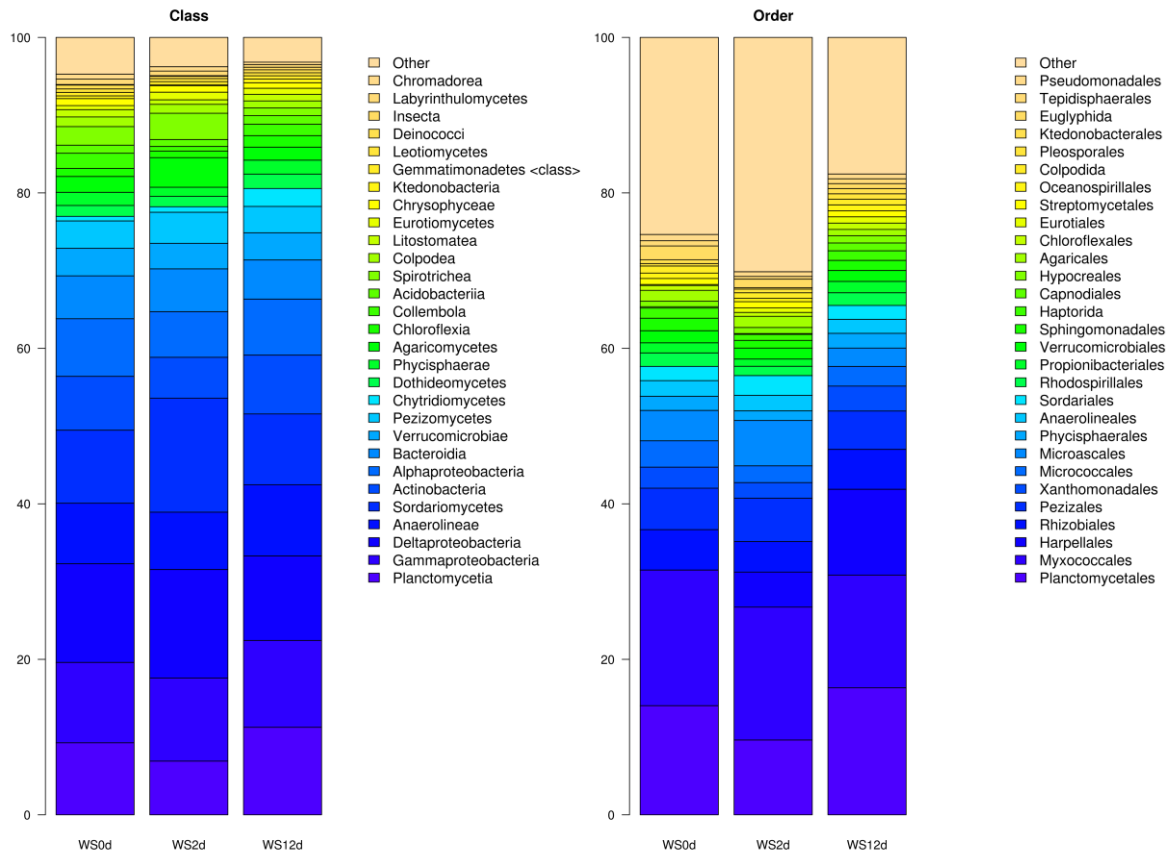
Dentre os 11 filios mais abundantes nas amostras de RNA ribossomal (Figura 2), que juntos contabilizam mais de 80% dos transcritos com taxonomia atribuída, cinco são eucariotos, sendo eles Ascomycota, Amoebozoa, Ciliophora, Zoopagomycota e Basidiomycota e seis são bactérias, sendo elas Proteobacteria, Planctomycetes, Chloroflexi, Actinobacteria, Bacteroidetes e Verrucomicrobia.

Dentre os filios de fungo encontrados com maior abundância, Zoopagomycota ainda não havia sido descrito em estudos de rizosfera de cana-de-açúcar. Por outro lado, os filios Ascomycota e Basidiomycota já foram encontrados em grande abundância em estudos de rizosfera de cana-de-açúcar. Por exemplo o estudo feito por Paungfoo-Lonhienne et al. (Paungfoo-Lonhienne et al. 2015) investigou o efeito de diferentes dosagens de fertilizante de

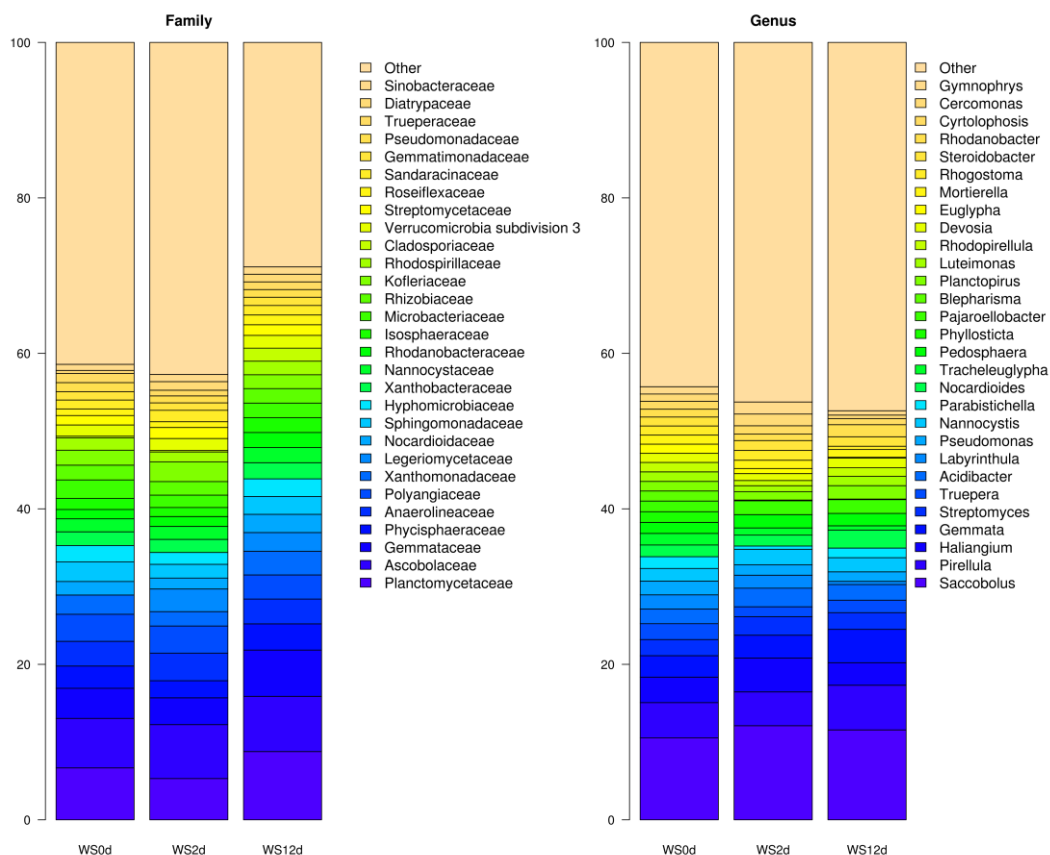
nitrogênio na comunidade de fungos da rizosfera de cana-de-açúcar e encontrou uma diversidade taxonômica envolvendo basicamente dois filos, Ascomycota e Basidiomycota, com o primeiro aumentando em abundância com maiores quantidades de fertilizante e o segundo diminuindo.

Já no caso dos filos de bactéria encontrados, o filo Proteobacteria é normalmente o filo mais abundante em amostras de rizosfera de cana-de-açúcar (e também foi o mais abundante nos dados do RNA mensageiro apresentados no capítulo II), por exemplo, Pisa et al. (Pisa et al. 2011) analisou sequências de 16S de rizosfera de cana-de-açúcar de 12 e 13 meses de idade, com ou sem o uso de fertilizante de nitrogênio. Foi observado que Proteobacteria era o filo mais abundante em solo fertilizado, perdendo a primeira posição para o filo Actinobacteria em solo não fertilizado.

As Figuras 3 e 4 apresentam a abundância relativa dos níveis taxonômicos mais inferiores. Dentre os cinco gêneros mais abundantes, *Saccobolus* e *Pezizales* são fungos e *Pirellula*, *Haliangium* e *Rhizobiales* são bactérias. Dentre os dois fungos, somente *Pezizales* já foi citado na literatura sobre cana-de-açúcar, tendo sido encontrado em um estudo sobre a comunidade de fungos atuando na degradação de folhas de cana-de-açúcar e a resposta da comunidade à diferentes tipos de manejo do solo (Miura et al. 2015). Nesse estudo o fungo identificado como *Pezizales* respondeu negativamente ao tratamento de cobertura do solo com bagaço e foi proposto como um tratamento que favoreceu organismos de crescimento lento e que preferem ambientes com poucos nutrientes.



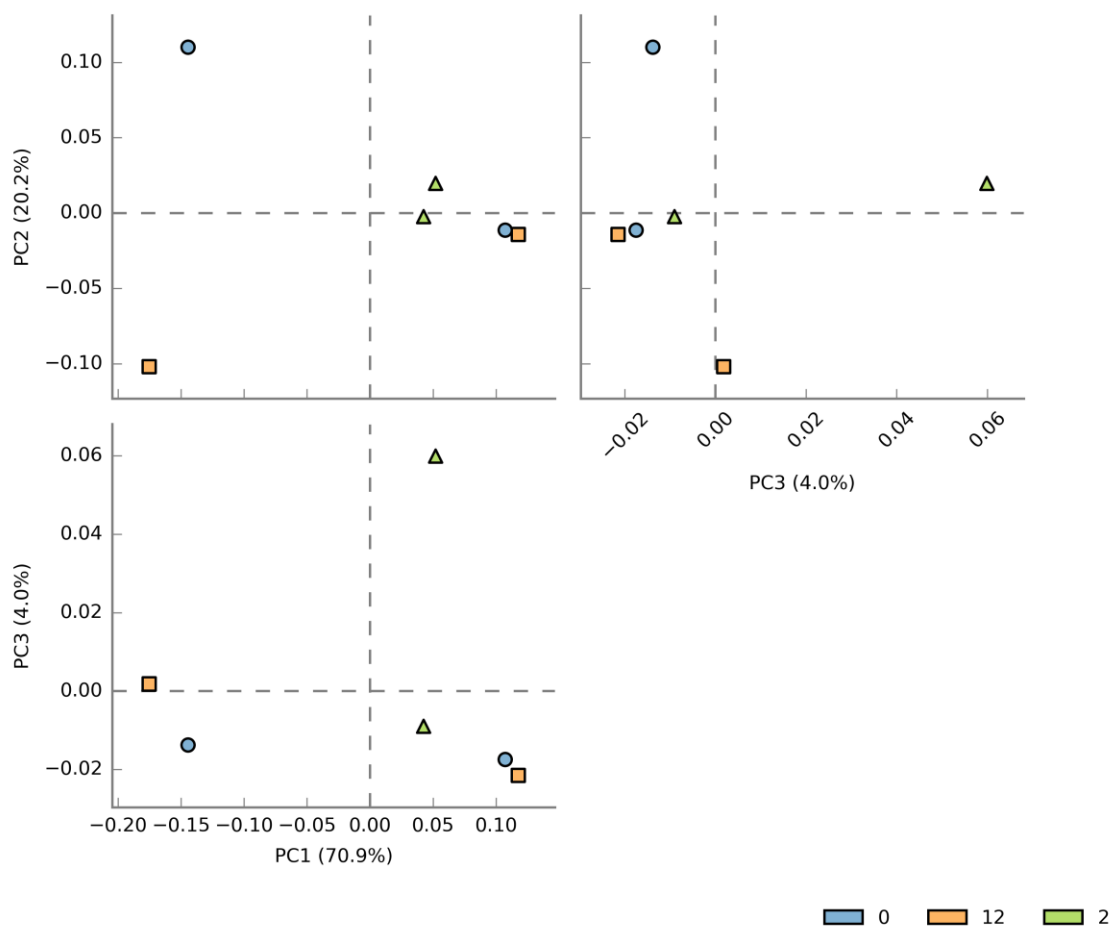
**Figura 3:** Abundância relativa nos níveis de classe e ordem dos transcritos com taxonomia atribuída a partir do RNA ribossomal.



**Figura 4:** Abundância relativa nos níveis de família e gênero dos transcritos com taxonomia atribuída a partir do RNA ribossomal.

Para analisar a abundância diferencial dos táxons identificados a partir do rRNA, os dados foram filtrados no programa MEGAN de forma a manter apenas os dados de eucariotos, uma vez que depois da utilização do kit para a remoção do RNA ribossomal bacteriano, as abundâncias relativas dos transcritos de bactérias e arqueias provavelmente foram totalmente alteradas. Assim, os dados de abundância dos transcritos de rRNA de eucariotos foram exportados do MEGAN e inseridos no *software* STAMP (Parks et al. 2014) para determinação de grupos que responderam aos diferentes tratamentos de estresse hídrico aplicados. Uma análise de componentes principais foi feita para os dados a nível de filo (Figura 5). Nessa análise, a junção dos dois primeiros componentes PC1 e PC2 explica cerca de 91% dos dados e, embora não seja possível observar uma clara separação entre as

amostras sem estresse hídrico e com doze dias de estresse, é possível observar uma separação das amostras com dois dias de estresse hídrico. Adicionalmente, as análises de componente principal dos níveis taxonômicos inferiores seguiram o mesmo padrão.

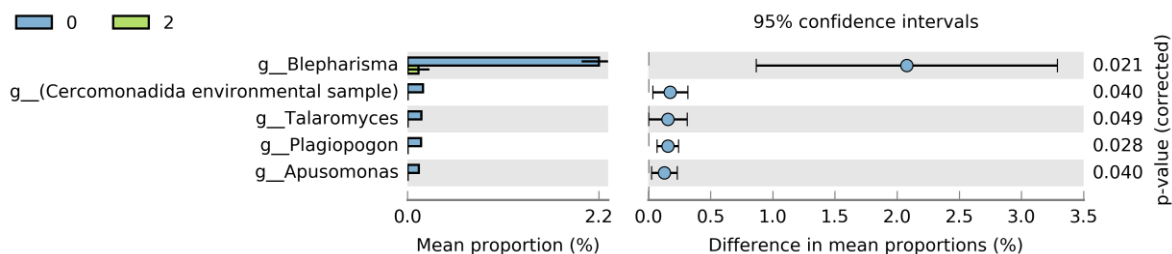


**Figura 5:** Análise de componentes principais dos dados a nível de filo para os diferentes tratamentos de estresse hídrico aplicados.

As análises estatísticas do *software* STAMP encontraram cinco gêneros diferencialmente abundantes entre os tratamentos com zero e dois dias de estresse hídrico e sete gêneros entre os tratamentos com zero e doze dias de estresse hídrico (Figuras 6 e 7). Entre zero e dois dias apenas foram encontrados gêneros com diminuição da proporção média no tratamento com estresse hídrico, sendo o gênero *Blepharisma* o gênero com maior queda

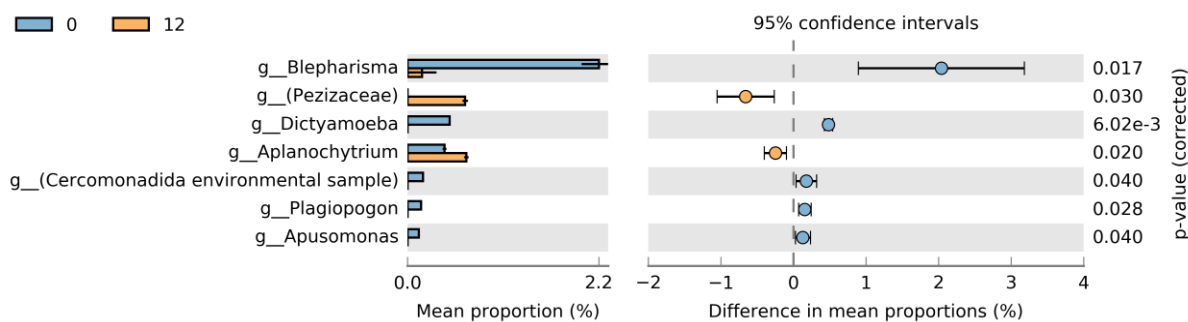


nos valores de proporção média. *Blepharisma* é um gênero de ciliados amplamente distribuído e tem sido reportado no mais variados ambientes, como ambientes de água doce, ambientes de água salgada, solos e até ambientes extremos (Fernandes et al. 2013).



**Figura 6:** Proporção média e diferenças previstas para os transcritos de rRNA no nível de gênero para os tratamentos com zero e dois dias de estresse hídrico.

Já no tratamento com doze dias de estresse hídrico, com exceção do gênero *Talarocyces*, todos os outros gêneros que tiveram queda de abundância entre zero e dois dias de estresse hídrico também tiveram queda de proporção média entre zero e doze dias. Além disso, o grupo *Pezizaceae*, que na verdade é uma família e o gênero *Aplanochytrium* apresentaram aumento nos valores de proporção média.

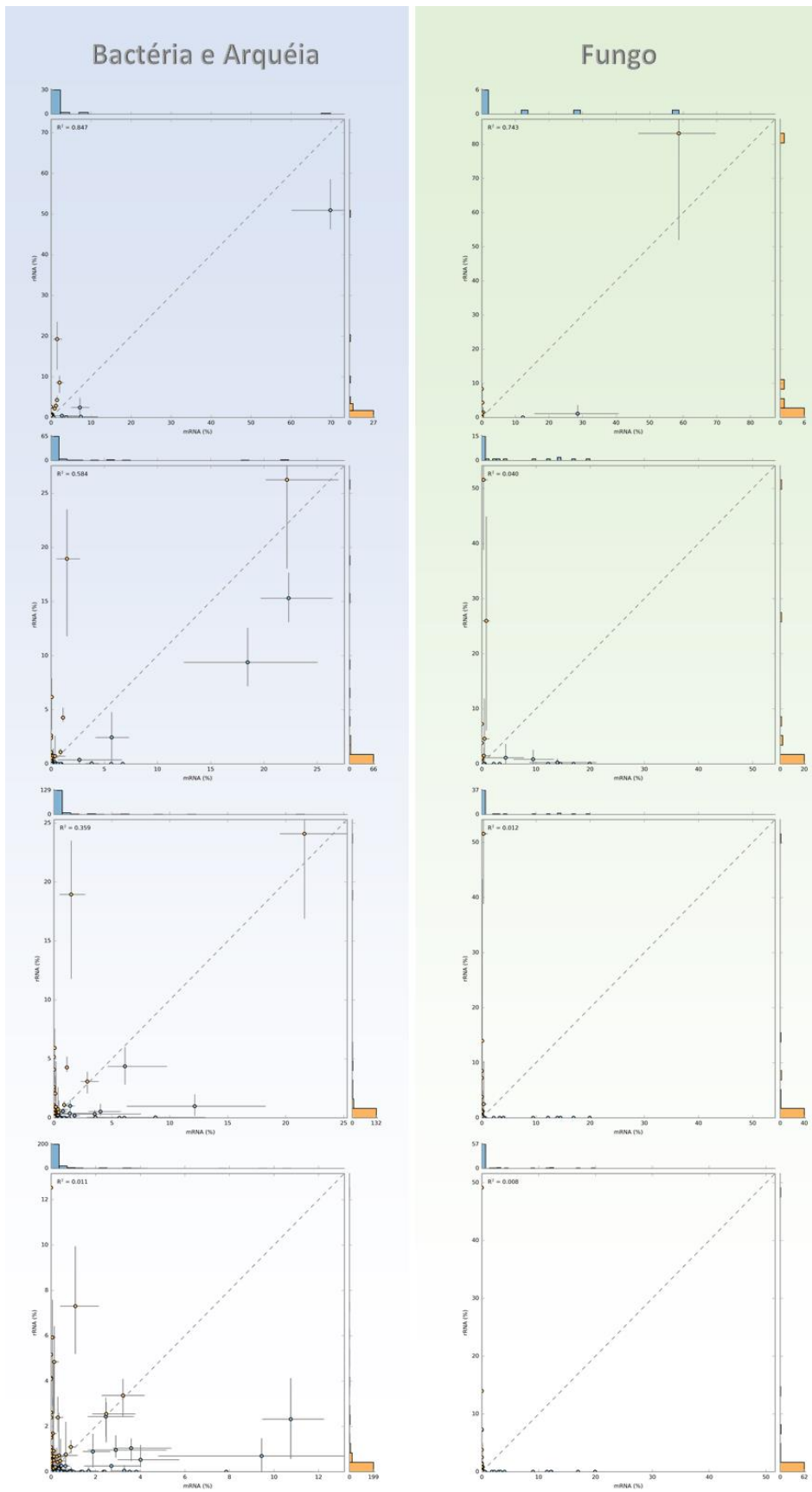


**Figura 7:** Proporção média e diferenças previstas para os transcritos de rRNA no nível de gênero para os tratamentos com zero e doze dias de estresse hídrico.

Por fim, foram comparadas as atribuições taxonômicas obtidas para os dados de rRNA contra os dados de mRNA. Todos os dados foram analisados no *software* MEGAN e

exportados para o software STAMP que realizou as análises estatísticas. Os resultados podem ser visualizados na Figura 8.

A nível de filo obteve-se uma boa correlação entre a atribuição taxonômica do rRNA e do mRNA, os dados de rRNA tiveram uma correlação com os dados de mRNA de 0.847 e 0.743, para bactérias e arqueias e para fungos, respectivamente. Entretanto, essa correlação caiu para 0.584 e 0.040 no nível de classe para bactérias e arqueias e para fungos, respectivamente. Dessa forma, é possível concluir-se que existiu uma boa correlação entre os dados de rRNA e mRNA para o nível mais geral de filo e que os dados de bactéria e arqueias tiveram uma correlação maior que os dados de fungo, mas que as atribuições taxonômicas para ambos os grupos diferiram bastante para níveis taxonômicos mais inferiores.

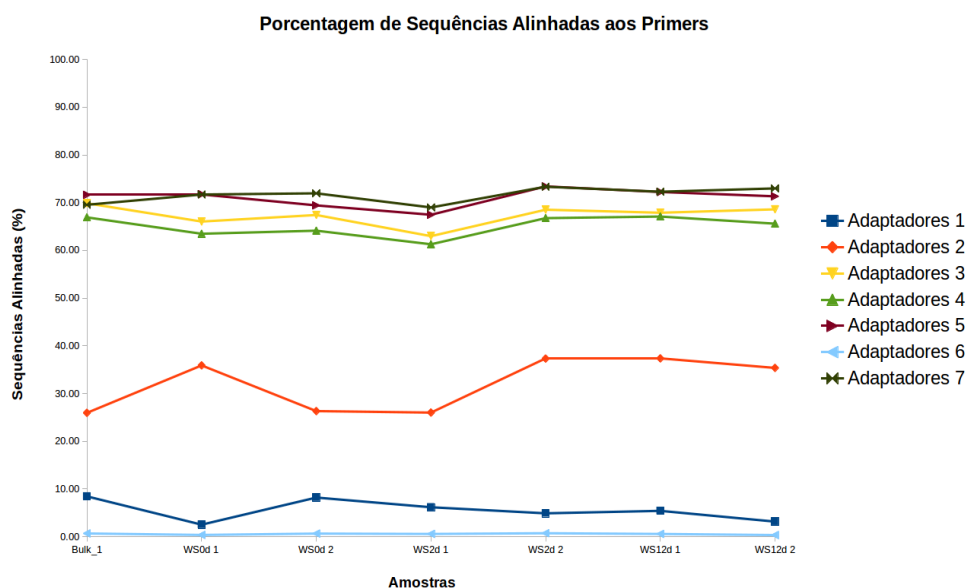


**Figura 8:** Comparação entre as atribuições taxonômicas a partir dos dados de mRNA contra os dados de rRNA.

### *VirtualPrimer*

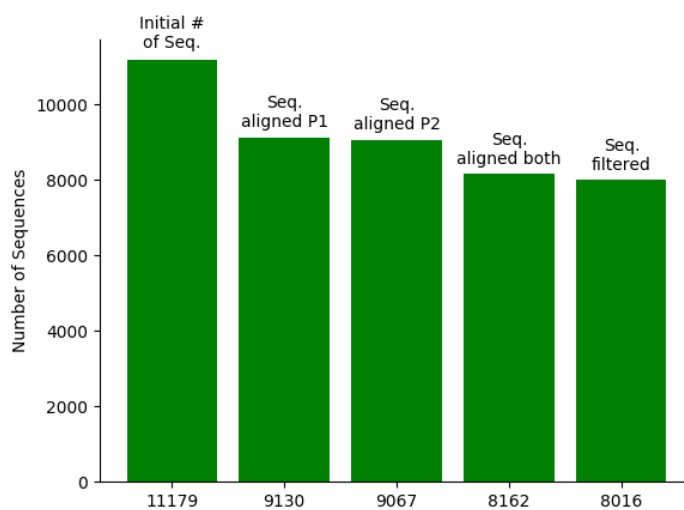
Paralelamente à análise das sequências de RNA ribossomal, foi desenvolvido um *software* chamado *VirtualPrimer* - disponível para *download* em <https://github.com/VictorGambarini/VirtualPrimer> - com o intuito de melhorar a comparabilidade dos dados. Em resumo, o *VirtualPrimer* busca por sequências que contenham regiões que seriam amplificadas por um par de *primers* disponibilizado e as filtra, seguido por um filtro de comprimento e, por último, as sequências são cortadas, gerando um arquivo apenas com as sequências filtradas.

Utilizando o *software* desenvolvido, foi possível testar uma variedade de *primers* do gene ribossomal 16S comumente empregados. Observou-se que os *primers* que geraram um maior número de sequências foram os *primers* 338F (5'-ACTCCTACGGGAGGCAGCAG-3') e 1046R (5'-CGACAGCCATGCANCACT-3'), denominados *Adaptadores 7* na Figura 9.

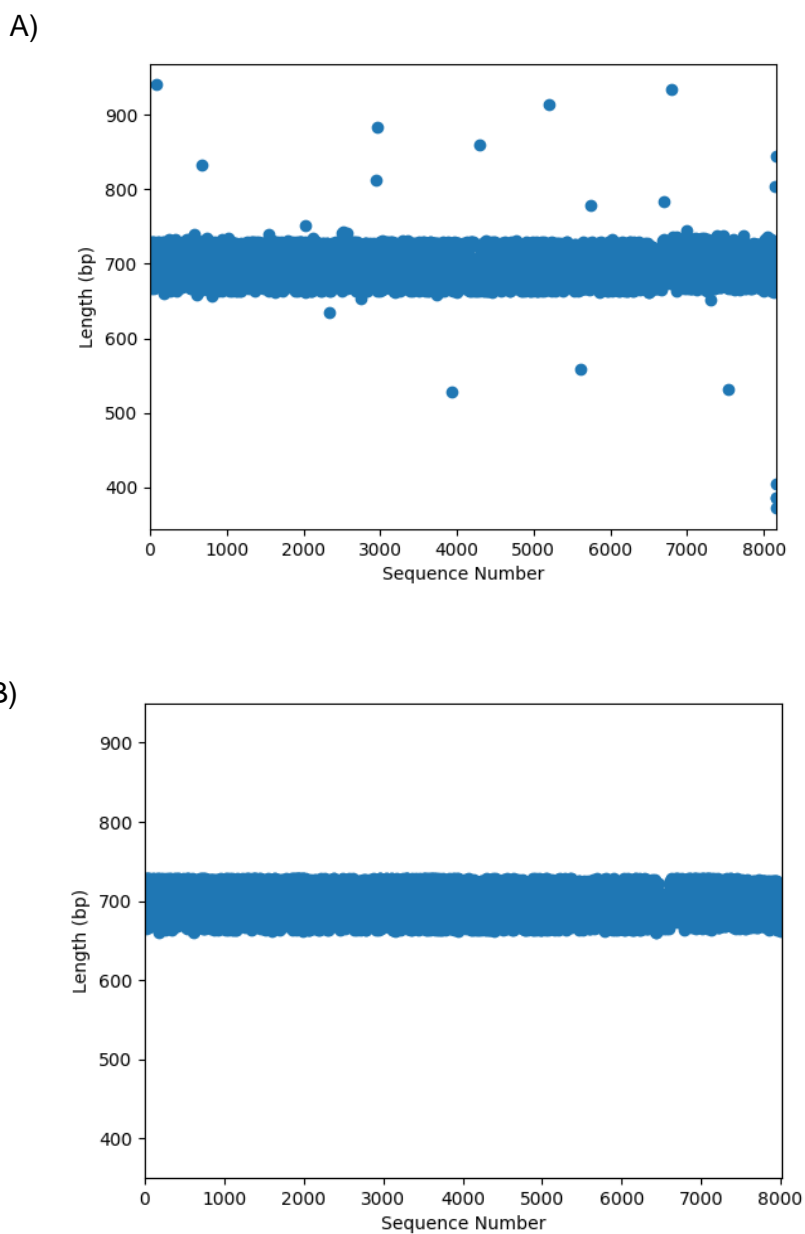


**Figura 9:** Desempenho dos pares dos adaptadores testados com o software desenvolvido, chamado *VirtualPrimer*.

Como exemplo de resultado obtido com o VirtualPrimer temos a análise da amostra WS0d 1, onde, dos 11.179 transcritos iniciais, 8.162 apresentam alinhamento contra os dois *primers*, estando 8.016 dentro do comprimento médio obtido (Figuras 10 e 11).



**Figura 10:** Número de sequências alinhadas pelo VirtualPrimer aos *primers* 338F e 1046R.



**Figura 11:** Filtragem dos transcritos com comprimento fora da média obtida. A, antes e B, depois da filtragem.

Apesar do método utilizado para a análise taxonômica das sequências de rRNA não ter utilizado o VirtualPrimer, os resultados obtidos através dele foram utilizados para selecionar o melhor par de primers, para fazer a caracterização taxonômica das bactérias isoladas pela aluna de doutorado Letícia Pereira. Dessa forma, o par de Adaptadores 7,

apresentado na Figura 9, foi utilizado com sucesso e conseguiu-se caracterizar taxonomicamente a maioria dos isolados bacterianos.

### **Análise do RNA mensageiro**

#### *Teste de Diferentes Estratégias de Montagem*

Segundo Celaj et al. (2014) que testou vários montadores diferentes em dados de metatranscriptoma, o melhor desempenho foi obtido pelo montador Trinity (Grabherr et al. 2011). Por outro lado, Rana et al. (Rana et al. 2016) também testou vários montadores no transcriptoma de *Fundulus heteroclitus* e o montador Bridger (Chang et al. 2015) foi o que obteve o melhor desempenho. Em outra tentativa para buscar a melhor montagem possível, Nakasugi et al. (Nakasugi et al. 2014) combinou a montagem individual de vários montadores em uma única montagem e utilizou softwares como CAP3 (X. Huang and Madan 1999) e CD-HIT-EST (W. Li and Godzik 2006) para eliminar redundância dos dados e, dessa forma, considerou essa montagem combinada como a melhor dentre as testadas.

Sabendo-se disso, a montagem do RNA não filtrado pelo riboPicker, ou RNA mensageiro, foi feita com vários montadores normalmente empregados na literatura, sendo eles IDBA-UD (Peng et al. 2012), Bridger, Trinity, Velvet (Zerbino and Birney 2008), Oases (Schulz et al. 2012) e SOAPdenovo-Trans (Y. Xie et al. 2014). Para essa avaliação dos montadores, utilizou-se um subconjunto dos dados que consistiu da amostra Bulk 1, diminuindo-se, dessa forma, a complexidade e tamanho dos dados, o que facilitou a análise e otimização de vários parâmetros e fatores envolvidos no processo. Além disso, testou-se também a estratégia de montagens individuais combinadas, como sugerido por Nakasugi et al. (Nakasugi et al. 2014). Finalmente, para obter métricas por onde avaliar as montagens obtidas, foram utilizadas diversas ferramentas e estratégias que levaram vários pontos importantes em consideração.

### *Transrate*

O programa Transrate (Smith-Unna et al. 2016) foi utilizado para gerar estatísticas básicas das montagens analisadas (Tabela 5). É possível observar que o número de sequências obtidas com cada montador variou bastante, sendo o montador IDBA-UD o que gerou a menor quantidade de *contigs* e a montagem com múltiplos k-mers combinada em uma única montagem do software SOAPdenovo-Trans a que gerou o maior número de *contigs*.

	IDBA-UD	Bridger	Trinity	Oases 51	SOAPdenovo-Trans MK	Merged-CDHIT-Cap3	Merged-Cap3
Sequências	32306	44357	72508	43230	85990	32177	55686
Maior	9211	13074	9215	14052	8333	14685	15280
Tamanho médio	524	436	314	395	376	653	529
Maiores de 1000 pb	2098	2463	787	1601	1853	4810	5214
Com ORF	5674	6237	5261	2289	6364	6160	8540
N50	522	488	298	397	384	773	604
N10	1438	2027	711	1449	930	2237	972
Conteúdo GC	0,5809	0,6006	0,60095	0,56131	0,56723	0,57605	0,58255

**Tabela 5:** Estatísticas das principais estratégias de montagem testadas na amostra Bulk 1.

Dentre as montagens individuais (IDBA-UD, Bridger, Trinity, Oases 51 e SOAPdenovo-Trans MK), a montagem do IDBA-UD apresentou o melhor resultado nas métricas de tamanho médio e N50. Além disso, a montagem do IDBA-UD obteve o segundo melhor resultado na métrica de contigs maiores de 1000 pb e resultados próximos aos outros montadores nas outras métricas.

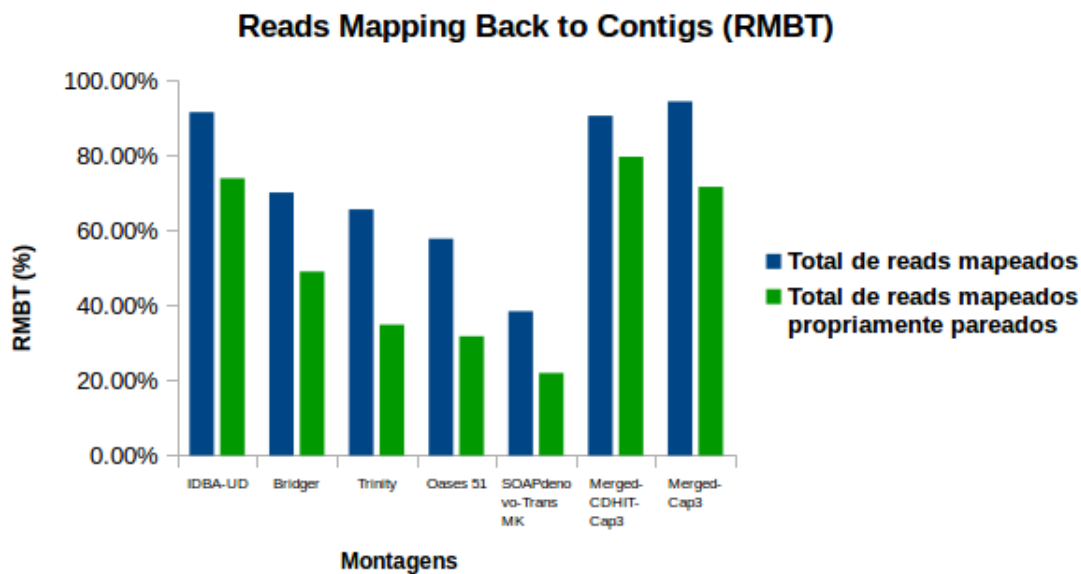
As duas montagens combinadas (Merged-CDHIT-Cap3 e Merged-Cap3) foram feitas combinando-se o resultado das outras cinco montagens individuais, com a etapa de redução



de redundância feita pelo software CD-HIT-EST feita em apenas uma delas. De acordo com o resultado obtido pelas duas montagens, pode-se notar que a montagem que utilizou o *software* CD-HIT-EST obteve aproximadamente metade do número de contigs da outra montagem, entretanto essa diferença não se repetiu nas outras métricas como por exemplo o número de contigs com *ORFs* que caiu apenas 28%. Além disso, outras métricas ficaram melhores nessa primeira montagem mais compacta, como por exemplo o N50 quase 170 pb mais longo na primeira montagem, que de um modo geral obteve o melhor desempenho entre todas as montagens, tanto individuais quanto combinadas.

#### *RMBT (Reads Mapping Back to Contigs)*

O RMBT foi calculado a partir do alinhamento das *reads* contra os *contigs* realizado pelo *software* Bowtie2 (Langmead and Salzberg 2012) para a amostra Bulk 1. O desempenho das principais estratégias de montagem pode ser acompanhado na Figura 12. Dentre as montagens individuais, o montador IDBA-UD obteve o melhor desempenho com cerca de 75% dos pares de reads mapeando corretamente aos contigs montados contra cerca de 50% e 35% dos montadores seguintes com melhor desempenho, respectivamente os montadores Bridger e Trinity.



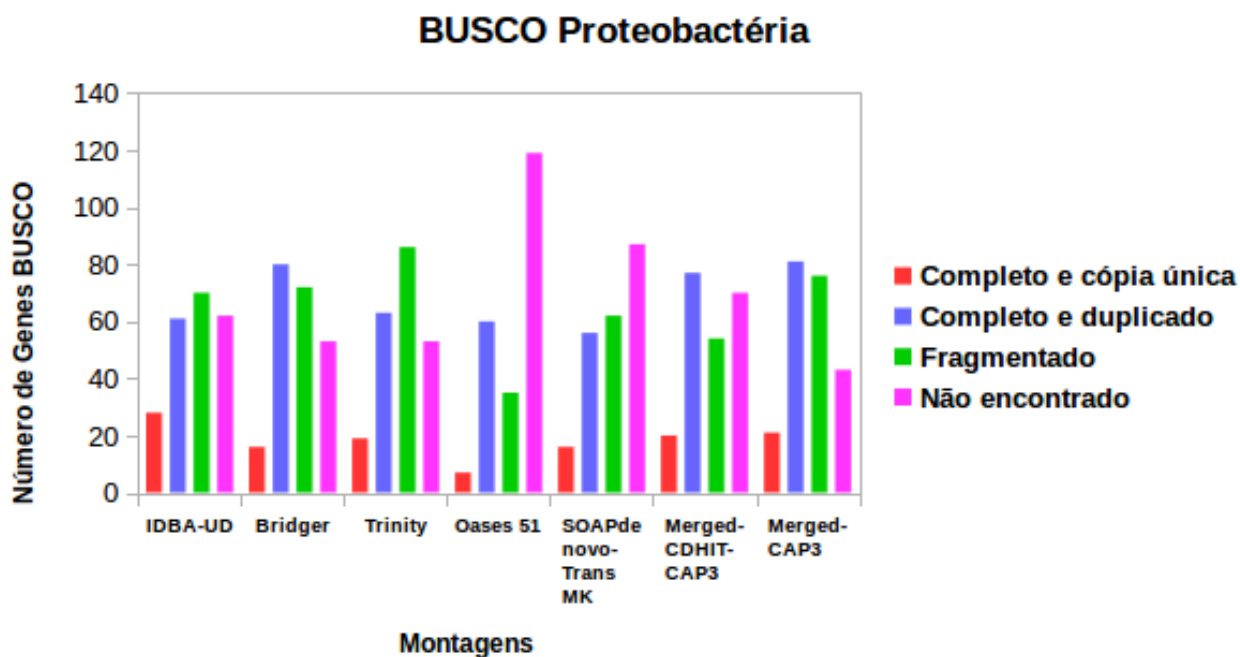
**Figura 12:** O mapeamento das *reads* aos *contigs* é uma das principais métricas de avaliação de montagens e, pode-se observar que os melhores desempenhos foram do *software* IDBA-UD e das estratégias de concatenação.

Dentre as montagens combinadas, novamente a montagem que passou pela etapa de retirada de redundância com o *software* CD-HIT-EST obteve o melhor desempenho. Além disso, nota-se que apesar da primeira montagem ter cerca de metade do número de *contigs* da segunda montagem, ainda assim ela contou com uma maior porcentagem de pares de *reads* alinhando nos *contigs*. Fato esse que reforça ainda mais a qualidade da primeira montagem em relação à segunda.

### *Genes BUSCO*

Outra métrica importante na avaliação de montagens de *reads* é a presença de genes que são normalmente encontrados nos grupos taxonômicos estudados. Por meio do *software* BUSCO foi possível obter uma medida da completude do metatranscriptoma montado (Figura 13). Foram analisados genes do banco de dados do BUSCO para Proteobacteria, pois

análises paralelas já demonstravam que esse era o grupo taxonômico mais abundantes nos dados gerados.



**Figura 13:** Resultado obtido através do *software* BUSCO versão 2 utilizando-se genes normalmente encontrados em genomas de bactérias do filo Proteobacteria. A amostra utilizada para esta análise foi a Bulk 1.

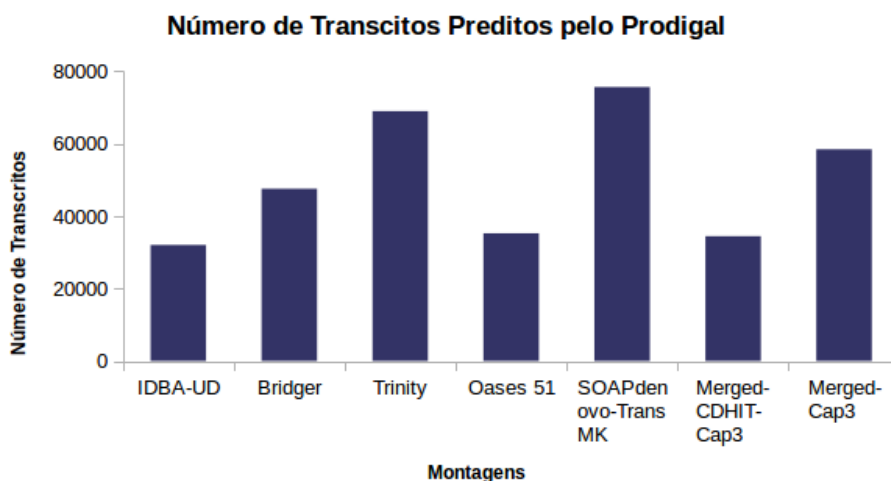
A melhor montagem individual nessa análise foi a do montador Bridger que obteve apenas 53 genes não encontrados e 72 fragmentados, seguido pelo montador Trinity que também obteve apenas 53 genes não encontrados, mas que montou 86 genes fragmentados. Por sua vez, o montador IDBA-UD apresentou 62 genes não encontrados e 70 fragmentados, ficando em uma posição intermediária entre as cinco montagens individuais. Embora a montagem do IDBA-UD não tenha sido a melhor em relação aos genes BUSCO com as sequências da amostra Bulk, no capítulo II será apresentado o resultado da montagem do IDBA-UD com todas as amostras juntas onde ele foi capaz de encontrar todos os genes BUSCO para bactérias e deixou de encontrar apenas um gene para o grupo Proteobacteria,

tendo praticamente todos os genes completos, demonstrando assim a qualidade da montagem realizada por este montador.

Já entre as montagens combinadas, pela primeira vez o resultado das duas se alterou, de forma que a segunda montagem, ou seja, a montagem sem remoção de redundância pelo *software* CD-HIT-EST obteve o melhor resultado. Dos 221 genes BUSCO para Proteobacteria buscados, a segunda montagem obteve apenas 43 genes não encontrados contra 70 genes não encontrados da primeira montagem.

### *Prodigal*

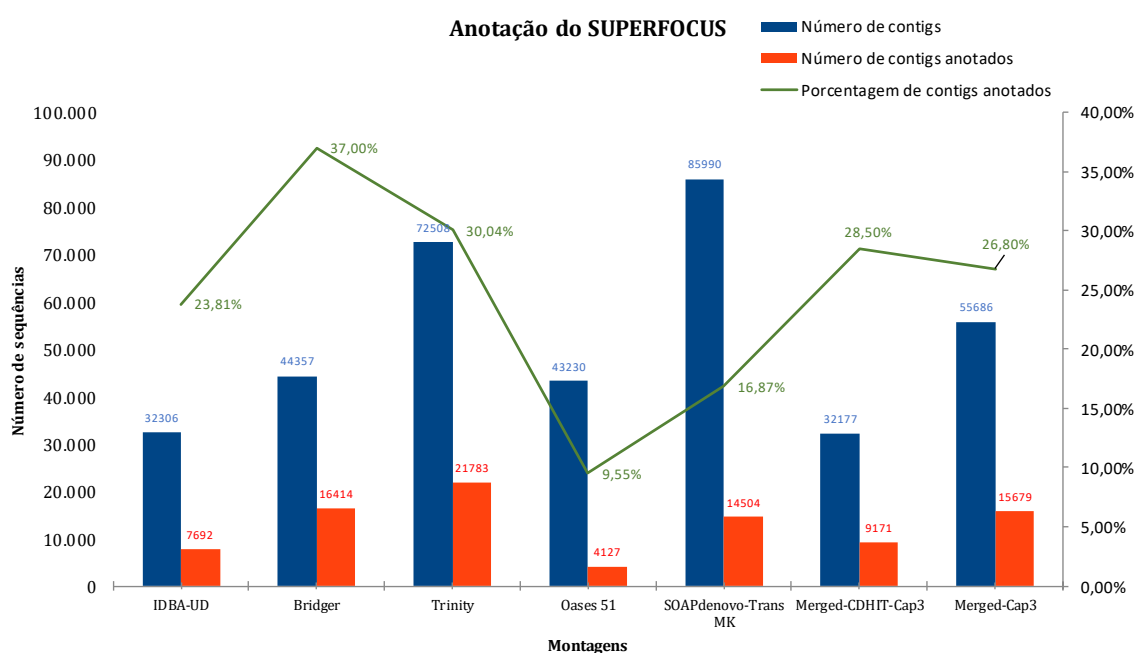
Também foi utilizado como métrica o resultado obtido pelo *software* Prodigal (Hyatt et al. 2010), que prediz *Open Read Frames* dentro dos *contigs* montados (Figura 14). De forma geral, o número de *ORFs* encontradas pelo Prodigal seguiu o padrão do número de *contigs* de cada amostra. Dessa forma, não se pôde extrair nenhum resultado conclusivo em favor de alguma das montagens analisadas por essa métrica.



**Figura 14:** Genes preditos pelo Prodigal para as diferentes estratégias de montagem aplicadas à amostra Bulk 1.

## SUPERFOCUS

Por fim foi feita a anotação funcional dos contigs obtidos através das várias estratégias de montagem testadas. Para isso, foi utilizado o *software* SUPERFOCUS (G. G. Z. Silva et al. 2015) que anota os transcritos de maneira rápida, porém comparável a outros métodos computacionalmente mais custosos, fornecendo assim, uma medida da quantidade de transcritos que puderam receber alguma anotação funcional (Figura 15).



**Figura 15:** Quantidade de transcritos anotados a partir do *software* SUPERFOCUS para cada uma das principais estratégias de montagem aplicadas à amostra Bulk 1.

Dentre as montagens individuais e de acordo com os resultados do SUPERFOCUS, observa-se que o melhor desempenho foi obtido pela montagem do *software* Bridger, que anotou 37% do contigs montados, seguido por Trinity com 30,04% e IDBA-UD com 23,81% dos *contigs* anotados. Os softwares SOAPdenovo-Trans e Oases obtiveram resultados bastante inferiores aos demais montadores com 16,87% e 9,55%, respectivamente.

Já dentre as montagens combinadas, as duas estratégias obtiveram resultados semelhantes, com uma pequena vantagem para a montagem não redundante com 28,50% contra 26,80% da outra estratégia de montagem.

## **Conclusão**

Através das análises preliminares realizadas foi possível filtrar as *reads* de baixa qualidade ou contaminadas por adaptadores e prosseguir para as etapas seguintes apenas com as *reads* de alta qualidade. Nesse ponto, acabou-se descartando a amostra Bulk 2 tanto para a montagem do mRNA quanto para a montagem do rRNA, enquanto as amostras WS8d 1 e WS8d 2 foram descartadas apenas para a montagem do rRNA.

Também foi possível identificar, montar e analisar taxonomicamente as *reads* provenientes de RNA ribossomal, verificando-se a eficiência do kit de depleção de RNA ribossomal bacteriano pelo aumento da proporção relativa de eucariotos encontrados nos *reads* de rRNA quando comparados com os *reads* de mRNA. Paralelamente a essa etapa, desenvolveu-se um *software* para filtragem de sequências provenientes de genes 16S ou 18S chamado VirtualPrimer que foi, inclusive, útil para outras análises.

Adicionalmente, foram testados diversos *softwares* e estratégias de montagem e foram desenvolvidas diversas formas de avaliação das montagens obtidas. O *software* IDBA-UD obteve o melhor desempenho na maioria das métricas analisadas e, principalmente, na métrica de mapeamento dos *reads* nos *contigs* montados. Essa métrica é uma das principais formas de avaliação da montagem, uma vez que as *reads* são uma representação do dado biológico analisado e, dessa forma, quanto mais concordantes os *contigs* estiverem em relação às *reads*, melhor. Sobre as estratégias de junção de várias montagens individuais em uma única montagem, essas estratégias não foram escolhidas pois embora elas tenham obtido bons resultados na maioria das métricas analisadas, verificou-se que boa parte das métricas

estavam sendo infladas pela presença de quimeras formadas pelo *software* Cap3. Dessa forma, se optou pela utilização da montagem do IDBA-UD para montar todos os dados de RNA mensageiro de todas as amostras combinadas. Esses resultados serão apresentados no Capítulo II.

## Referências bibliográficas

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), 403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2).
- Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. <https://doi.org/citeulike-article-id:11583827>.
- Chang, Z., Li, G., Liu, J., Zhang, Y., Ashby, C., Liu, D., ... Huang, X. (2015). Bridger: A new framework for de novo transcriptome assembly using RNA-seq data. *Genome Biology*, 16(1). <https://doi.org/10.1186/s13059-015-0596-2>.
- Fernandes, Noemi Mendes, Roberto Júnio Pedroso Dias, Marcus Vinicius Xavier Senra, Carlos Augusto Gomes Soares, and Inácio Domingos da Silva Neto. 2013. “Morphology and 18S RDNA Gene Sequence of *Blepharisma Sinuosum* Sawaya, 1940 (Ciliophora: Heterotrichea) from Brazil.” *European Journal of Protistology* 49 (4): 623–33. <https://doi.org/10.1016/j.ejop.2013.04.003>.
- Gordon, A., & Hannon, G. J. (2010). Fastx-toolkit. FASTQ/A short-reads pre-processing tools. Unpublished [Http://Hannonlab.Cshl.Edu/Fastx\\_Toolkit](http://Hannonlab.Cshl.Edu/Fastx_Toolkit).
- Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., ... 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>.
- He, S., Wurtzel, O., Singh, K., Froula, J. L., Yilmaz, S., Tringe, S. G., ... Hugenholtz, P. (2010). Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. *Nature Methods*, 7(10), 807–812. <https://doi.org/10.1038/nmeth.1507>.



- Huang, X., & Madan, A. (1999). CAP3: A DNA sequence assembly program. *Genome Research*, 9(9), 868–877. <https://doi.org/10.1101/gr.9.9.868>.
- Huson, D. H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., ... Tappu, R. (2016). MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLoS Computational Biology*, 12(6). <https://doi.org/10.1371/journal.pcbi.1004957>.
- Hyatt, D., Chen, G. L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11. <https://doi.org/10.1186/1471-2105-11-119>.
- Kalam, Sadaf, Subha Narayan Das, Anirban Basu, and Appa Rao Podile. 2017. "Population Densities of Indigenous Acidobacteria Change in the Presence of Plant Growth Promoting Rhizobacteria (PGPR) in Rhizosphere." *Journal of Basic Microbiology* 57 (5): 376–85. <https://doi.org/10.1002/jobm.201600588>.
- Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4), 357–359. <https://doi.org/10.1038/nmeth.1923>.
- León-Sánchez, Lupe, Emilio Nicolás, Marta Goberna, Iván Prieto, Fernando T. Maestre, and José Ignacio Querejeta. 2018. "Poor Plant Performance under Simulated Climate Change Is Linked to Mycorrhizal Responses in a Semi-Arid Shrubland." *Journal of Ecology* 106 (3): 960–76. <https://doi.org/10.1111/1365-2745.12888>.
- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14), 1754–1760. <https://doi.org/10.1093/bioinformatics/btp324>.
- Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13), 1658–1659. <https://doi.org/10.1093/bioinformatics/btl158>.

- Liao, Y., Smyth, G. K., & Shi, W. (2014). FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7), 923–930. <https://doi.org/10.1093/bioinformatics/btt656>.
- Miller, C. S., Baker, B. J., Thomas, B. C., Singer, S. W., & Banfield, J. F. (2011). EMIRGE: Reconstruction of full-length ribosomal genes from microbial community short read sequencing data. *Genome Biology*, 12(5). <https://doi.org/10.1186/gb-2011-12-5-r44>.
- Mitra, S., Stärk, M., & Huson, D. H. (2011). Analysis of 16S rRNA environmental sequences using MEGAN. In 10th Int. Conference on Bioinformatics - 1st ISCB Asia Joint Conference 2011, InCoB 2011/ISCB-Asia 2011: Computational Biology - Proceedings from Asia Pacific Bioinformatics Network (APBioNet) (Vol. 12). <https://doi.org/10.1186/1471-2164-12-S3-S17>.
- Miura, Toshiko, Ainin Niswati, I. G. Swibawa, Sri Haryani, Heru Gunito, Satoshi Shimano, Koichi Fujie, and Nobuhiro Kaneko. 2015. “Diversity of Fungi on Decomposing Leaf Litter in a Sugarcane Plantation and Their Response to Tillage Practice and Bagasse Mulching: Implications for Management Effects on Litter Decomposition.” *Microbial Ecology* 70 (3): 646–58. <https://doi.org/10.1007/s00248-015-0620-9>.
- Nakasugi, K., Crowhurst, R., Bally, J., & Waterhouse, P. (2014). Combining transcriptome assemblies from multiple de novo assemblers in the allo-tetraploid plant *Nicotiana benthamiana*. *PLoS ONE*, 9(3). <https://doi.org/10.1371/journal.pone.0091776>.
- Newman, Molli M., Nicola Lorenz, Nigel Hoilett, Nathan R. Lee, Richard P. Dick, Mark R. Liles, Cliff Ramsier, and Joseph W. Kloepper. 2016. “Changes in Rhizosphere Bacterial Gene Expression Following Glyphosate Treatment.” *Science of the Total Environment* 553: 32–41. <https://doi.org/10.1016/j.scitotenv.2016.02.078>.

- Parks, Donovan H., Gene W. Tyson, Philip Hugenholtz, and Robert G. Beiko. 2014. "STAMP: Statistical Analysis of Taxonomic and Functional Profiles." *Bioinformatics* 30 (21): 3123–24. <https://doi.org/10.1093/bioinformatics/btu494>.
- Patel, R. K., & Jain, M. (2012). NGS QC toolkit: A toolkit for quality control of next generation sequencing data. *PLoS ONE*, 7(2). <https://doi.org/10.1371/journal.pone.0030619>.
- Paungfoo-Lonhienne, Chanyarat, Yun Kit Yeoh, Naga Rup Pinaki Kasinadhuni, Thierry G.A. Lonhienne, Nicole Robinson, Philip Hugenholtz, Mark A. Ragan, and Susanne Schmidt. 2015. "Nitrogen Fertilizer Dose Alters Fungal Communities in Sugarcane Soil and Rhizosphere." *Scientific Reports* 5. <https://doi.org/10.1038/srep08678>.
- Peng, Y., Leung, H. C. M., Yiu, S. M., & Chin, F. Y. L. (2012). IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, 28(11), 1420–1428. <https://doi.org/10.1093/bioinformatics/bts174>.
- Pisa, G., Magnani, G. S., Weber, H., Souza, E. M., Faoro, H., Monteiro, R. A., ... Cruz, L. M. (2011). Diversity of 16S rRNA genes from bacteria of sugarcane rhizosphere soil. *Brazilian Journal of Medical and Biological Research*, 44(12), 1215–1221. <https://doi.org/10.1590/S0100-879X2011007500148>.
- Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., & Glöckner, F. O. (2007). SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research*, 35(21), 7188–7196. <https://doi.org/10.1093/nar/gkm864>.
- Rana, S. B., Zadlock, F. J., Zhang, Z., Murphy, W. R., & Bentivegna, C. S. (2016). Comparison of de Novo transcriptome assemblers and k-mer strategies using the killifish, *fundulus heteroclitus*. *PLoS ONE*, 11(4). <https://doi.org/10.1371/journal.pone.0153104>.

- Rho, M., Tang, H., & Ye, Y. (2010). FragGeneScan: Predicting genes in short and error-prone reads. *Nucleic Acids Research*. <https://doi.org/10.1093/nar/gkq747>.
- Schmieder, R., Lim, Y. W., & Edwards, R. (2012). Identification and removal of ribosomal RNA sequences from metatranscriptomes. *Bioinformatics*, 28(3), 433–435. <https://doi.org/10.1093/bioinformatics/btr669>.
- Schulz, M. H., Zerbino, D. R., Vingron, M., & Birney, E. (2012). Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, 28(8), 1086–1092. <https://doi.org/10.1093/bioinformatics/bts094>.
- Shrestha, Pravin Malla, Michael Kube, Richard Reinhardt, and Werner Liesack. 2009. “Transcriptional Activity of Paddy Soil Bacterial Communities.” *Environmental Microbiology* 11 (4): 960–70. <https://doi.org/10.1111/j.1462-2920.2008.01821.x>.
- Silva, G. G. Z., Green, K. T., Dutilh, B. E., & Edwards, R. A. (2015). SUPER-FOCUS: A tool for agile functional analysis of shotgun metagenomic data. *Bioinformatics*, 32(3), 354–361. <https://doi.org/10.1093/bioinformatics/btv584>.
- Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19), 3210–3212. <https://doi.org/10.1093/bioinformatics/btv351>.
- Smith-Unna, R., Boursnell, C., Patro, R., Hibberd, J. M., & Kelly, S. (2016). TransRate: Reference-free quality assessment of de novo transcriptome assemblies. *Genome Research*, 26(8), 1134–1144. <https://doi.org/10.1101/gr.196469.115>.
- Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A., & Conesa, A. (2011). Differential expression in RNA-seq: A matter of depth. *Genome Research*, 21(12), 2213–2223. <https://doi.org/10.1101/gr.124321.111>.

- Urich, Tim, Anders Lanzén, Ji Qi, Daniel H. Huson, Christa Schleper, and Stephan C. Schuster. 2008. "Simultaneous Assessment of Soil Microbial Community Structure and Function through Analysis of the Meta-Transcriptome." *PLoS ONE* 3 (6). <https://doi.org/10.1371/journal.pone.0002527>.
- Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., ... Wang, J. (2014). SOAPdenovo-Trans: De novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics*, 30(12), 1660–1666. <https://doi.org/10.1093/bioinformatics/btu077>.
- Zerbino, D. R., & Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18(5), 821–829. <https://doi.org/10.1101/gr.074492.107>.

## CAPÍTULO II

---

### Análise do RNA Mensageiro do Metatranscriptoma de Rizosfera de Cana-de- Açúcar Submetida ao Estresse Hídrico

## **Introdução**

Apesar da cana-de-açúcar ter sido tradicionalmente cultivada para a produção de açúcar, nas últimas décadas ela passou a ter grande importância para a indústria dos biocombustíveis. A alta demanda por açúcar e biocombustível faz com que a indústria global da cana-de-açúcar busque cada vez mais aumentar a sua produção e mitigar os danos causados por pragas, doenças e fatores climáticos (Yeoh et al. 2016). Os três principais fatores ambientais que limitam a produtividade da cana-de-açúcar são secas, alta salinidade e baixas temperaturas. Esses estresses ambientais induzem uma série de respostas bioquímicas, fisiológicas e de desenvolvimento nas plantas (Wu et al. 2008).

Por causa da natureza irregular das chuvas, plantações de cana-de-açúcar necessitam de grandes quantidades de irrigação para alcançar as metas de produção. Entretanto, água para irrigação é limitada, aumenta os custos da produção e diminui a sustentabilidade do processo (M. D. A. Silva et al. 2007). Além disso, estão previstas alterações nos ciclos e padrões de chuvas em todo o globo em virtude das mudanças climáticas acarretadas pelo aumento da concentração de gases do efeito estufa na atmosfera e consequente aquecimento global (S. P. Xie et al. 2015).

Essa necessidade de aumentar a produtividade, porém de forma sustentável, gerou o desenvolvimento de alternativas para o incremento da produção menos danosas ao meio ambiente como por exemplo a manipulação do microbioma rizosférico, que além de propiciar tolerância à estresses abióticos, também proporciona aumento do crescimento, defesa contra patógenos e aquisição de nutrientes para as plantas hospedeiras (De Souza et al. 2016).

Dessa forma, o objetivo desse estudo foi analisar a microbiota ativa da rizosfera de cana-de-açúcar submetida ao estresse hídrico através da análise do metatranscriptoma dos microrganismos rizosféricos.

## **Materiais e métodos**

### **Controle de qualidade do sequenciamento e filtragem do RNA ribossomal**

Os reads gerados no sequenciamento foram filtrados através do programa NGS QC Toolkit (Patel and Jain 2012) com valores de *cutoff* para qualidade mínima de 20 e tamanho mínimo de 70 pares de base e analisados com o programa FASTQC (Andrews 2010), em seguida foi realizada a remoção de *reads* contaminadas por adaptadores através do programa FASTX-Toolkit (Gordon and Hannon 2010).

A identificação *in silico* do rRNA foi feita através do *software* riboPicker versão 0.4.3 (Schmieder, Lim, and Edwards 2012), utilizando o banco de dados de sequências ribossomais não redundantes disponibilizado pelo software.

### **Análise do RNA mensageiro**

As *reads* de todas as amostras foram combinadas em uma montagem combinada utilizando o *software* IDBA-UD (Peng et al. 2012) versão 1.1.3 com os parâmetros --mink=20, --maxk=100, --step=20 e --min\_contig 200. O resultado da montagem foi analisado com o *software* Transrate versão 1.0.3 com os parâmetros *default*, com o *software* Bowtie2 versão 2.3.1 com os parâmetros *default* e com o *software* Busco (Simão et al. 2015) versão 2.0.1 com o parâmetro -m=tran.

Para a predição de ORFs nos contigs montados, testou-se dois *softwares*, FragGeneScan (Rho, Tang, and Ye 2010) versão 1.30 com os parâmetros -w 0 para sequências fragmentadas e -t illumina\_5 para sequenciamento Illumina com taxa de erro de cerca de 0,5%.

A determinação da melhor forma de fazer o mapeamento das *reads* nos transcritos montados foi feita através da comparação de três *softwares*: Bowtie2 versão 2.3.1 com o



parâmetro `--local`; BWA (H. Li and Durbin 2009) versão 0.7.17 com os parâmetros *default*; Salmon versão 0.8.2 com o parâmetro `--meta`. Sendo que tanto para Bowtie2 quanto para BWA, as tabelas de contagem foram extraídas dos arquivos BAM gerados com a utilização do *software* featureCounts versão 1.5.3 com os parâmetros `-p` e `-f`.

Para a anotação taxonômica e funcional, as *ORFs* foram alinhadas contra toda a coleção de sequências de proteínas de bactérias, fungos e arqueias do banco de dados RefSeq do NCBI versão 84 (Pruitt et al., 2006) usando o *software* DIAMOND v0.9.10.111 com parâmetros *default* (Buchfink et al., 2015), seguido por atribuição taxonômica usando o algoritmo do menor ancestral comum implementado no *software* MEGAN Community Edition versão v6.8.20 (Huson et al., 2011). Os termos GO (Gene Ontology) foram atribuídos mapeando-se os resultados do RefSeq contra registros do banco de dados do Gene Ontology (Ashburner et al., 2000).

Além disso, todos os transcritos foram alinhados contra sequências do banco de dados de enzimas com atividades sobre carboidratos (CAZy, do inglês *Carbohydrate-Active enZymes*) (Lombard et al. 2014) usando o *pipeline* de anotação automatizado do servidor Web dbCAN (Yin et al. 2012). Finalmente, a anotação de transcritos de resistência a antibióticos foi feita através do RGI (do inglês, *Resistance Gene Identifier*) no servidor Web do banco CARD (do inglês, *Comprehensive Antibiotic Resistance Database*) com os parâmetros *Protein sequence, Perfect, Strict and Loose hits e Low quality;coverage* (Jia et al. 2017).

O pacote do R denominado NOIseq Tarazona et al., 2012) foi usado para normalizar a tabela de contagem gerada pelo *software* featureCounts usando a técnica de normalização denominada TMM (do inglês, *Trimmed Mean of M-values*) (Robinson & Oshlack, 2010) e, também, para identificar os transcritos diferencialmente expressos entre todas as comparações par a par com um cutoff de  $q = 0.95$ .

## Resultados e discussão

### Montagem alinhamento das *reads* de RNA mensageiro

A partir das análises preliminares realizadas no capítulo I, a melhor montagem com os dados de RNA mensageiro foi gerada pelo software IDBA-UD (Peng et al. 2012). Dessa forma, ele foi utilizado para montar todas as amostras em uma única montagem combinada. Essa montagem gerou um total de 316,958 contigs com um N50 de 626 (Tabela 1). Em comparação com a literatura, Hayden et al (Hayden et al. 2018) conseguiu cerca de dois milhões de *contigs* em um metatranscriptoma de rizosfera de trigo montado com o *software* Trinity (Grabherr et al. 2011), entretanto o tamanho médio dos contigs foi de apenas 296 pb e o N50 foi de 278 pb, praticamente metade do valor alcançado pela montagem do IDBA-UD no metatranscriptoma aqui apresentado.

Measure	Values
N. Sequences	316.958
Smallest	200
Largest	18.993
N. base pairs	1,9E+08
Average size	612,592
Larger 1,000 bp	33.563
Larger 10,000 bp	48
N. ORFs (Prodigal)	380.635
ORFs average size	413
N90	345
N70	454
N50	626
N30	997
N10	2221
GC Content	0,64

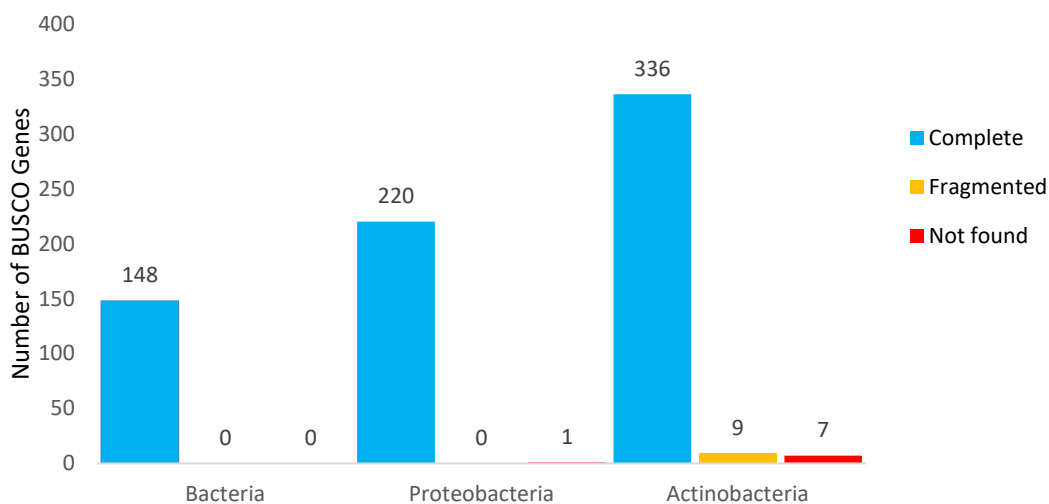
**Tabela 1:** Estatística da montagem do IDBA-UD com todas as amostras.

Com o intuito de analisar a qualidade da montagem obtida, as *reads* de cada uma das amostras foram mapeadas com o *software* Bowtie2 (Langmead and Salzberg 2012) contra os *contigs* gerados pelo IDBA-UD (Tabela 2), além disso, o nível de completude da montagem foi avaliado através da comparação com genes universalmente encontrados no domínio Bacteria e nos filos Actinobacteria e Proteobacteria através do *software* BUSCO (Figura 1) (Simão et al. 2015).

Amostra	Reads mapeados (%)
Bulk 1	57,48%
WS0d 1	34,77%
WS0d 2	55,61%
WS2d 1	52,80%
WS2d 2	20,32%
WS8d 1	11,64%
WS8d 2	27,58%
WS12d 1	17,39%
WS12d 2	30,77%

**Tabela 2:** *Reads* mapeadas contra os *contigs* gerados pelo IDBA-UD através do alinhador Bowtie2.

## BUSCO Assessment



**Figura 1:** Genes BUSCO encontrados na montagem do IDBA-UD com todas as amostras combinadas.

Para a predição de ORFs (Open Read Frames) a partir dos contigs montados pelo IDBA-UD, testou-se dois diferentes softwares comumente utilizados: Prodigal (Hyatt et al. 2010) e FragGeneScan (Rho, Tang, and Ye 2010). Tanto o Prodigal quanto o FragGeneScan obtiveram resultados bastante semelhantes, mas as ORFs do FragGeneScan ficaram um pouco maiores que as ORFs do Prodigal (Tabela 3).

	Prodigal	FragGeneScan
Número de <i>ORFs</i>	416,179	405,547
Média de <i>reads</i> mapeadas (Bowtie2 → featureCounts)	28.25%	38.29%
Busco bactéria	143C/4F/1M	143C/4F/1M
Busco actinobactéria	331C/13F/8M	329C/14F/9M
Busco proteobactéria	210C/7F/4M	210C/7F/4M
Busco firmicutes	192C/8F/32M	191C/9F/32M
Tamanho dos <i>ORFs</i> 0-200 pb	83280 (20,01%)	76211 (18,79%)
Tamanho dos <i>ORFs</i> 201-400 pb	171684 (41,25%)	162403 (40,05%)
Tamanho dos <i>ORFs</i> 401-600 pb	95826 (23,03%)	96671 (23,84%)
Tamanho dos <i>ORFs</i> 601-800 pb	33507 (8,05%)	33872 (8,35%)
Tamanho dos <i>ORFs</i> 801-1000 pb	15318 (3,68%)	15899 (3,92%)
Tamanho dos <i>ORFs</i> 1001-... pb	16564 (3,98%)	20491 (5,05%)

**Tabela 3:** Comparação entre Prodigal e FragGeneScan.

Paralelamente os *ORFs* obtidos através do Prodigal e do FragGeneScan foram utilizados para testar os *softwares* de alinhamento Bowtie2, BWA (H. Li and Durbin 2009) e Salmon (Patro, Duggal, and Kingsford 2015) e definir-se a melhor estratégia tanto para a etapa de predição de *ORFs* quanto para a etapa de mapeamento das *reads* nas *ORFs*. Essa comparação entre as diferentes estratégias pode ser observada na Tabela 4.

	Prodigal			FragGeneScan		
	BWA	Bowtie2	Salmon	BWA	Bowtie2	Salmon
<b>Bulk 1</b>	38,9 %	36,4 %	35,8 %	54,5 %	52,4 %	43,5 %
<b>WS0d 1</b>	34,9 %	26,0 %	23,7 %	47,0 %	37,7 %	27,7 %
<b>WS0d 2</b>	37,9 %	35,1 %	33,2 %	57,5 %	55,2 %	40,2 %
<b>WS2d 1</b>	35,0 %	32,3 %	30,5 %	52,8 %	50,2 %	37,8 %
<b>WS2d 2</b>	34,5 %	22,8 %	17,9 %	39,5 %	27,5 %	19,6 %
<b>WS8d 1</b>	32,2 %	21,4 %	16,3 %	36,5 %	25,3 %	16,3 %
<b>WS8d 2</b>	31,0 %	24,9 %	22,4 %	38,2 %	31,6 %	26,5 %
<b>WS12d 1</b>	40,0 %	25,4 %	18,3 %	42,1 %	27,4 %	19,3 %
<b>WS12d 2</b>	42,3 %	30,0 %	25,0 %	49,7 %	37,3 %	27,2 %
<b>Média</b>	36,3 %	28,3 %	24,8 %	46,4 %	38,2 %	28,7 %

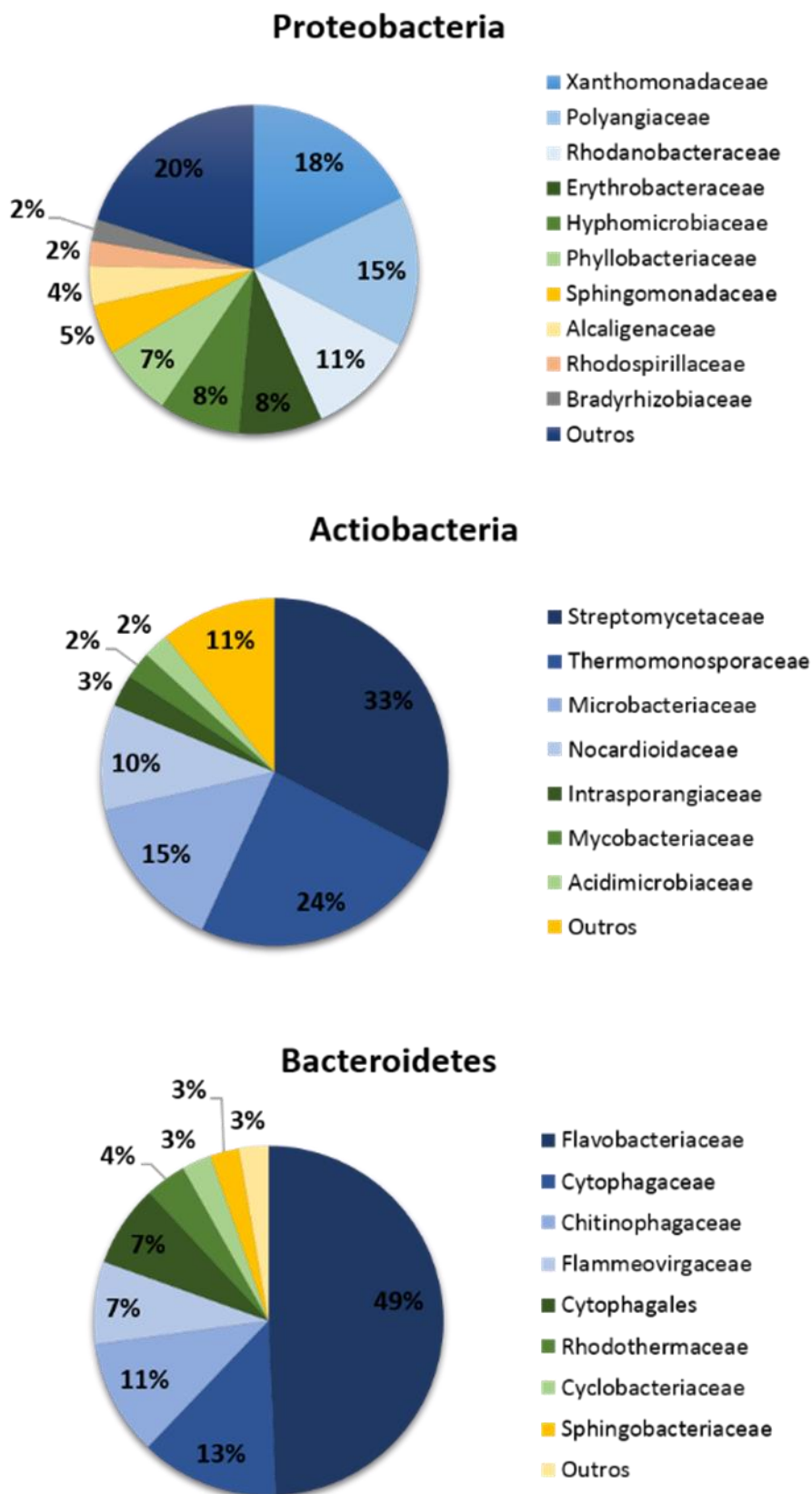
**Tabela 4:** Porcentagem de alinhamento dos reads nos ORFs preditos por Prodigal e FragGeneScan. Adicionalmente, foram testadas estratégias com diferentes alinhadores: BWA, Bowtie2 e Salmon.

Através da comparação dessas diferentes estratégias, pode-se verificar que a combinação entre FragGeneScan e BWA conseguiu utilizar uma maior porcentagem das *reads*. Dessa forma, essa estratégia foi escolhida e o resultado gerado, na forma de uma *count table* contendo 405.547 transcritos, foi utilizado na etapa seguinte.

### Composição taxonômica do metatranscriptoma

Os transcritos preditos pelo *software* FragGeneScan foram alinhados contra um subconjunto do banco de dados de sequências referência do NCBI (RefSeq) (Pruitt, Tatusova, and Maglott 2007), contendo sequências de bactérias, arqueias e fungos. O alinhamento foi feito através do *software* Diamond (Buchfink, Xie, and Huson 2015) e os resultados foram importados no *software* MEGAN (Huson et al. 2016) para atribuição taxonômica e funcional dos transcritos. Dos 405.547 transcritos, 298.726 (73,7%) foram anotados em algum nível taxonômico.

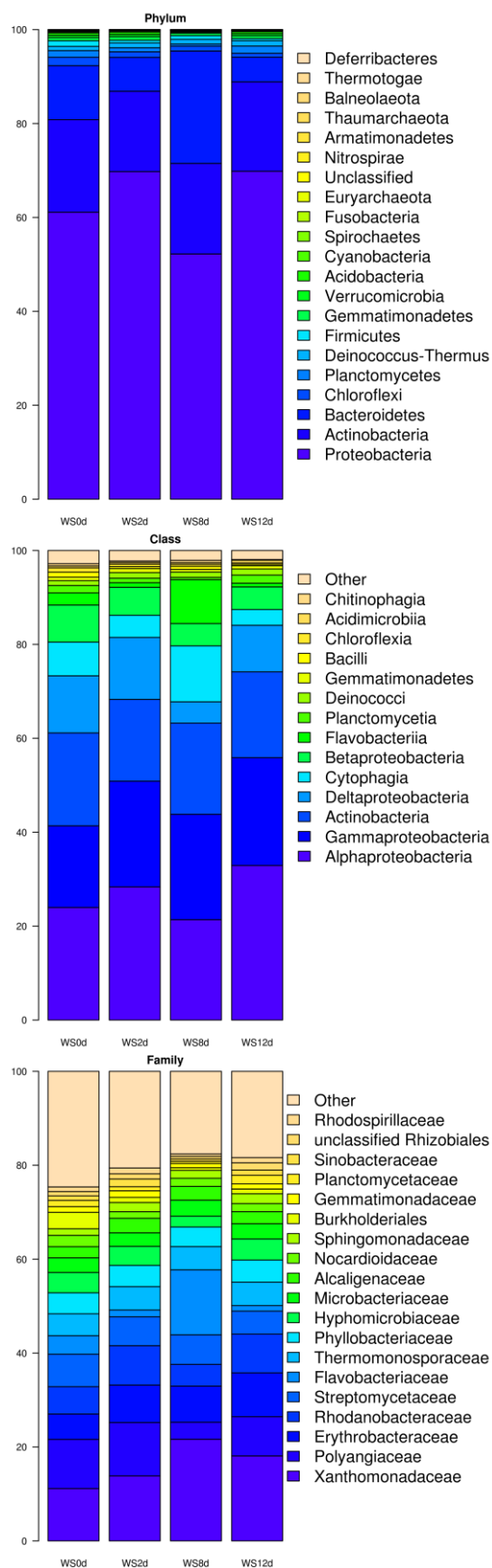
As análises da composição taxonômica do metatranscriptoma montado revelaram uma predominância dos filos Proteobacteria (52,2%), Actinobacteria (15,1%) e Bacteroidetes (10,6%), as famílias mais abundantes de cada um desses filos podem ser observadas na Figura 2. Em outros estudos de microbiomas de rizosfera, esses filos também foram encontrados em grande abundância, como por exemplo na rizosfera de sorgo (Xu et al. 2018), na rizosfera de milho (Yang et al. 2017) e na própria rizosfera de cana-de-açúcar (da Costa et al. 2018; Pisa et al. 2011).



**Figura 2:** Composição taxonômica dos transcritos do metatranscriptoma de rizosfera de cana-de-açúcar em nível de família para os três filos mais abundantes.

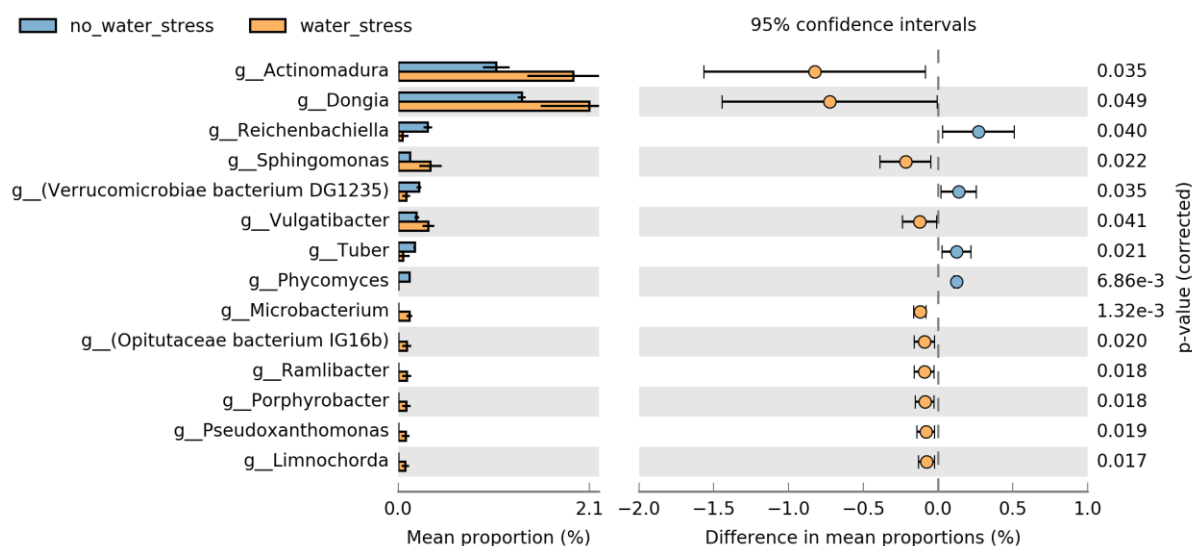


As abundâncias de cada transcrito foram estimadas alinhando-se as *reads* nos transcritos com o *software* BWA seguido da contabilização dos alinhamentos feita pelo *software* featureCounts (Liao, Smyth, and Shi 2014). Os valores de abundância de cada transcrito foram passados ao *software* MEGAN que normalizou e estimou a abundância de cada transcrito em cada um dos tratamentos (WS0d, WS2d, WS8d e WS12d), além de fazer a contabilização das abundâncias dos grupos taxonômicos identificados nas análises (Figura 3).



**Figura 3:** Abundância relativa, em vários níveis taxonômicos, dos transcritos do metatranscriptoma de rizosfera de cana-de-açúcar para cada um dos tratamentos de estresse hídrico aplicados.

Os valores de abundância de cada grupo taxonômico foram exportados do MEGAN para o *software* de análises estatísticas STAMP (Parks et al. 2014). Dessa forma, foi possível observar as alterações taxonômicas causadas pelo estresse hídrico na microbiota associada à rizosfera de cana-de-açúcar. Para avaliar detalhadamente as bactérias que tiveram sua expressão gênica alterada em situações irrigadas e de estresse hídrico, o programa STAMP foi utilizado para determinar os gêneros que apresentaram diferença entre essas duas condições (Figura 4).



**Figura 4:** Diferenças na proporção média dos gêneros com diferença significativa ( $p < 0,05$ ) em amostras da rizosfera de cana-de-açúcar com ou sem tratamentos de estresse hídrico.

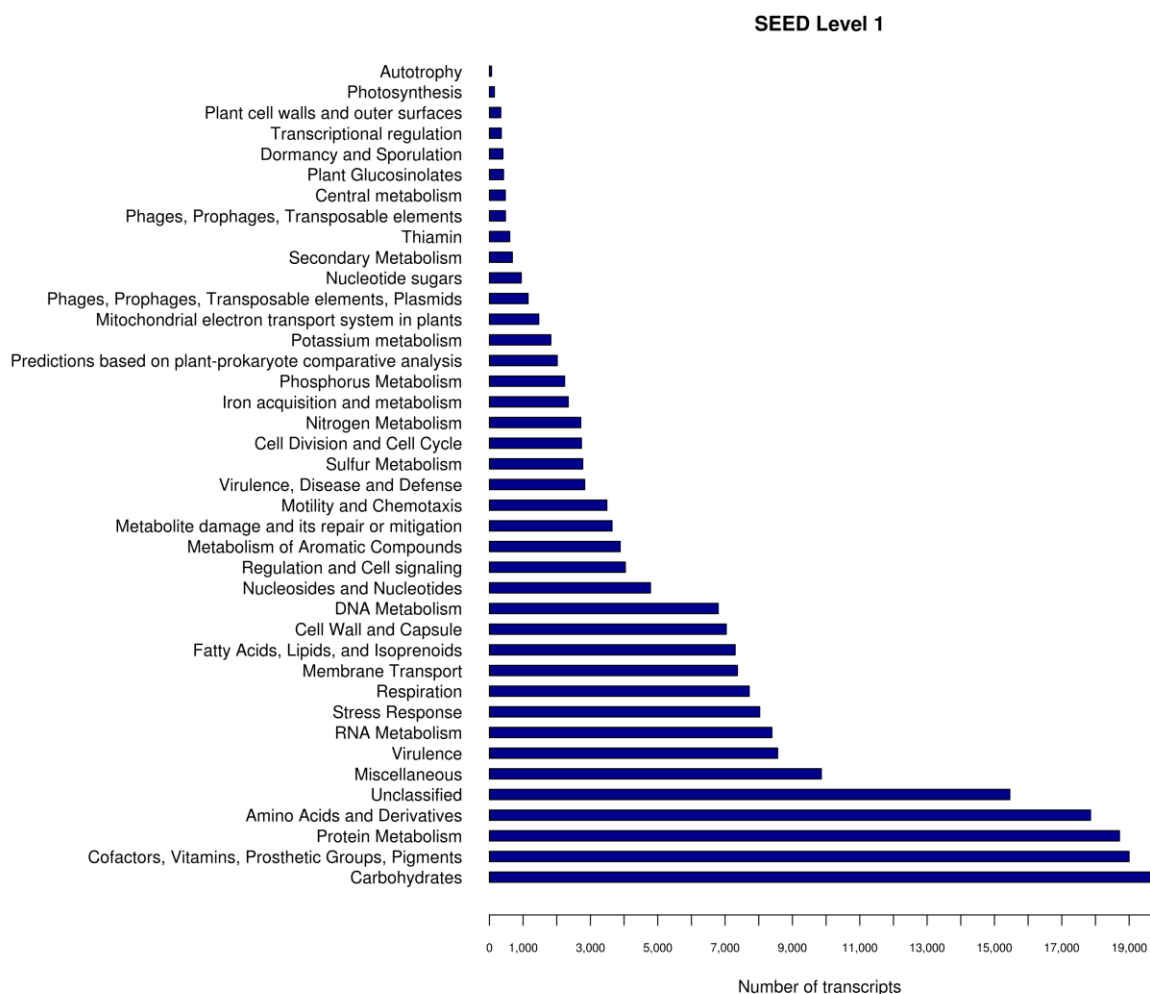
Dentre os 14 gêneros que apresentaram diferença significativa (teste de Welch,  $p < 0,05$ ) entre os tratamentos com e sem estresse hídrico, quatro foram mais abundantes nas amostras sem estresse hídrico e dez foram mais abundantes nas amostras que foram submetidas ao estresse hídrico. Dos quatro gêneros com diminuição de abundância ao longo do estresse hídrico, dois são fungos (Tuber e Phycomyces) e dois são bactérias (Reichenbachiella e Verrucomicrobiae bacterium DG1235). Já entre os gêneros que tiveram um aumento de abundância após a aplicação do estresse hídrico, os quatro que apresentaram

os maiores níveis de abundância foram *Actinomadura*, *Dongia*, *Sphingomonas* e *Vulgatibacter*.

O gênero *Actinomadura* foi o gênero que apresentou o maior aumento de abundância após a aplicação do estresse hídrico. Bactérias desse gênero já foram isoladas de solos de deserto (Norovsuren, Zenova, and Mosina 2007), solos rizosféricos (Su Yien Ting, Hoon Tan, and Kay Wai 2009), solos salinos (Lubsanova et al. 2014) e mangue (Ara et al. 2008), o que indica uma provável tolerância desse grupo de organismos à estresses hídricos e salinos. Além disso, bactérias desse gênero já foram isoladas de rizosfera de cana-de-açúcar e demonstraram serem capazes de solubilizar fosfato e produzir vários compostos promotores do crescimento vegetal como ácido indolacético (maior produção dentre todos os isolados testados), amônia e sideróforos (Bhosale & Kadam, 2015). Finalmente, esse grupo de bactérias pode ser um futuro alvo em estudos de para otimização do crescimento e resistência à seca através da microbiota da rizosfera em cana-de-açúcar.

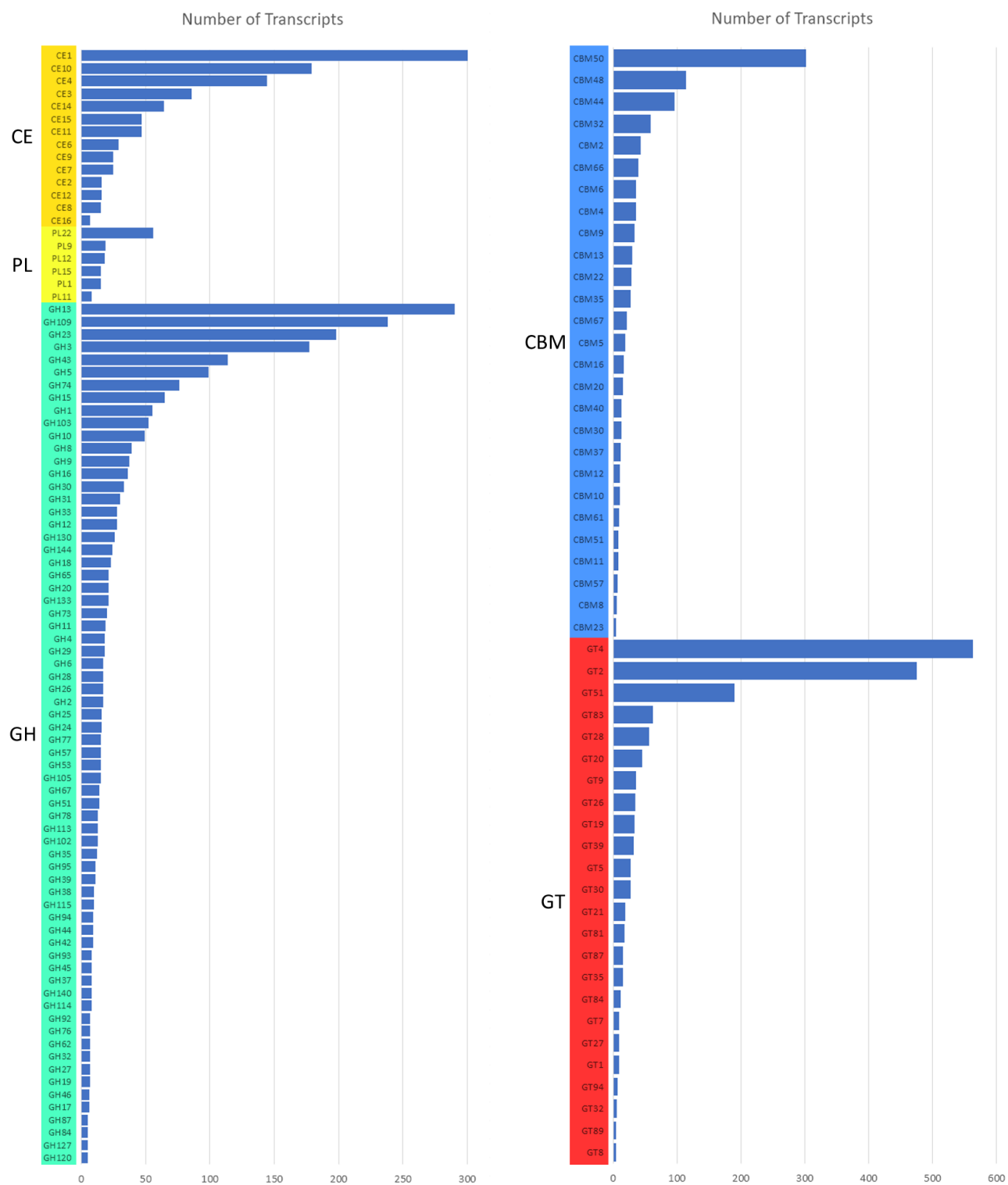
### **Composição funcional do metatranscriptoma**

A anotação funcional dos transcritos foi feita de várias formas, uma delas foi feita através do *software* MEGAN que mapeou os resultados do alinhamento dos transcritos contra o banco de dados RefSeq na hierarquia de subsistemas SEED (Overbeek et al. 2005). Dessa forma, 208.816 transcritos (51.5%) foram anotados em 40 categorias funcionais diferentes dos subsistemas do SEED (Figura 5). As categorias funcionais que apresentaram maior abundância no SEED foram aquelas responsáveis por processos essenciais dos micro-organismos, como por exemplo metabolismo de carboidratos, proteínas, aminoácidos, vitaminas e RNA.



**Figura 5:** Composição funcional dos transcritos do metatranscriptoma de rizosfera de cana-de-açúcar classificados no nível um da hierarquia de subsistemas do SEED.

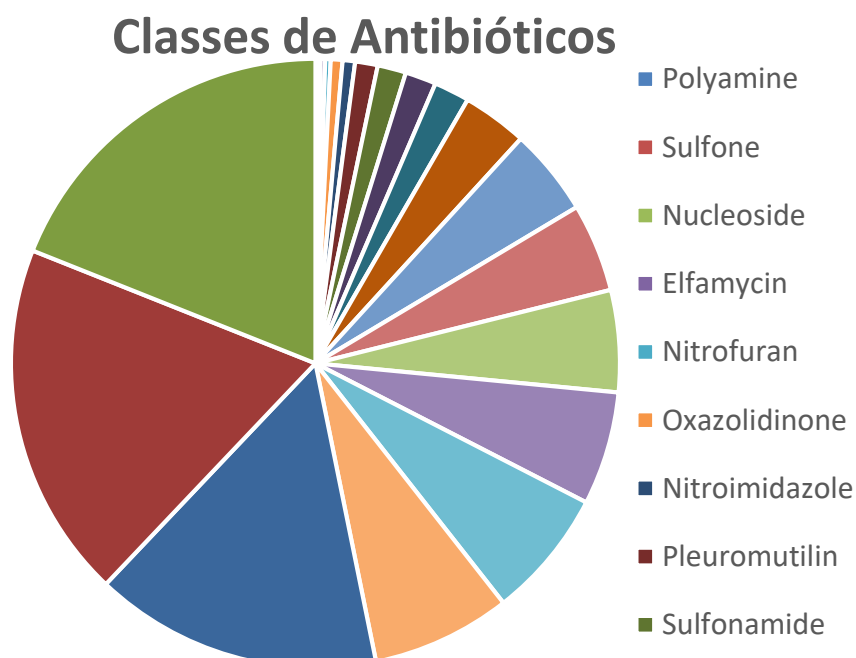
Uma vez que a categoria funcional mais abundante no SEED foi o metabolismo de carboidratos, resolve-se aprofundar um pouco mais a visão dessa função no metatranscriptoma. Dessa forma, todos os transcritos foram alinhados contra sequências do banco de dados de enzimas com atividades sobre carboidratos (CAZy, do inglês *Carbohydrate-Active enZymes*) (Lombard et al. 2014) usando o pipeline de anotação automatizado do servidor Web dbCAN (Yin et al. 2012). O resultado pode ser observado na Figura 6.



**Figura 6:** Número de transcritos para cada uma das famílias CAZyme encontradas no metatranscriptoma.

No total foram encontradas 233 famílias CAZyme no metatranscriptoma, sendo 101 GHs, 50 CBMs, 49 GTs, 17 PLs e 16 CEs. As enzimas da família GH hidrolisam ligações glicosídicas entres dois ou mais carboidratos, ou entre um carboidrato e uma molécula não carboidrato, como uma proteína ou um lipídeo (Cantarel et al. 2009). As três famílias GH mais abundantes no metatranscriptoma foram GH13, GH109 e GH23, sendo que a primeira age em substratos que contém ligações  $\alpha$ -glicosídicas, a segunda tem atividade de  $\alpha$ -acetilgalactosaminidase e a terceira tem atividade sobre peptidoglicanos (CAZypedia Consortium 2018).

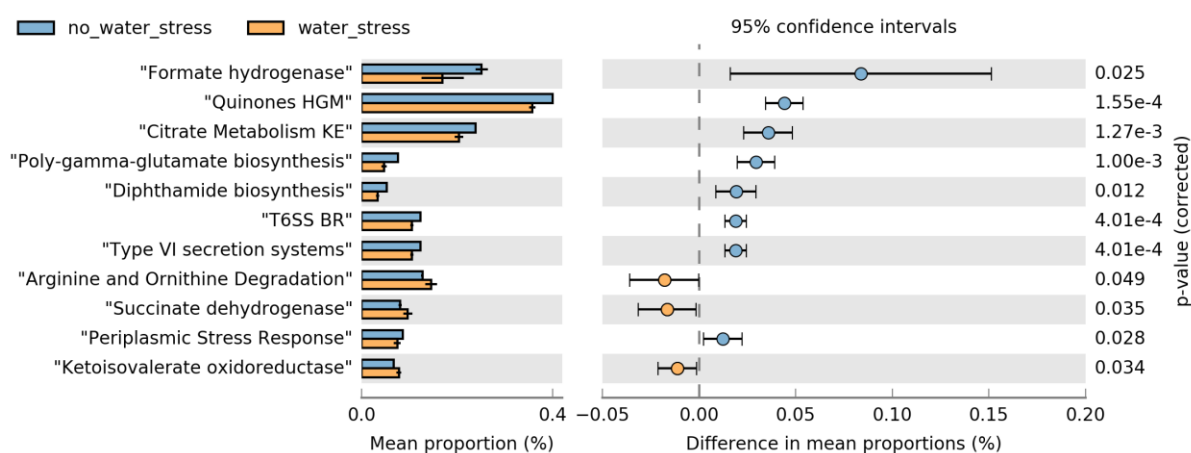
Além da anotação dos transcritos relacionados ao metabolismo de carboidratos, também foi feita a anotação de transcritos relacionados a resistência a antibióticos (Figura 7). Essa anotação foi feita através da *pipeline* RGI (do inglês, *Resistance Gene Identifier*) no servidor Web do banco CARD (do inglês, *Comprehensive Antibiotic Resistance Database*) (Jia et al. 2017).



**Figura 7:** Porcentagem relativa de transcritos putativos de resistência à antibióticos encontrados a partir do banco de dados CARD.

Foram encontrados transcritos putativos de resistência à antibióticos de 21 classes diferentes. Os mais abundantes foram resistência à tetraciclina (5.384 transcritos), macrolida (5.381 transcritos), fluoroquinolona (4.338 transcritos), antibiótico peptídico (2.101) e fenicol (1.948).

Para poder analisar a mudança funcional causada pelo estresse hídrico, foi feito o mapeamento das anotações funcionais na hierarquia dos subsistemas SEED foi exportada do *software* MEGAN para o *software* de análises estatísticas STAMP. As análises estatísticas revelaram que o estresse hídrico estimulou uma abundância diferencial (teste de Welch,  $p < 0,05$ ) de várias funções na microbiota associada à rizosfera de cana-de-açúcar (Figura 8).



**Figura 8:** Abundância diferencial de transcritos induzida pelos tratamentos de estresse hídrico. Os transcritos da microbiota da rizosfera estão sendo avaliados no nível dois da hierarquia SEED.

A microbiota que não foi submetida ao estresse hídrico (*no\_wter\_stress*) apresentou uma maior abundância de algumas categorias funcionais essenciais, as quais estão envolvidas com o metabolismo energético (formato hidrogenase, metabolismo de citrato), transporte de elétrons (quinonas) e fidelidade tradução (biossíntese de diftamida). Interessantemente, a falta de diftamida pode causar elevada mudança do quadro de leitura (mudança de -1) na síntese de proteínas de células eucarióticas. Além disso, durante certas infecções bacterianas,

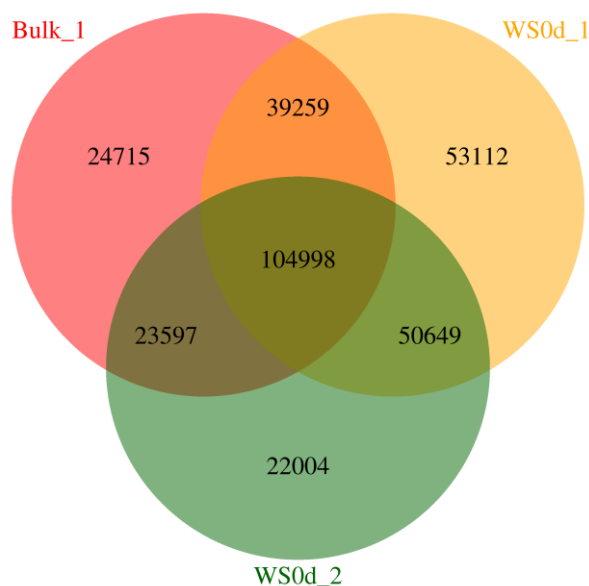


diftamida é especificamente reconhecida por certas toxinas produzidas por bactérias, como por exemplo toxina diftérica, Pseudomonas exotoxina A e toxina diftamida cholix (Su, Lin, and Lin 2013). Dessa forma, uma queda na abundância de transcritos relacionados à biossíntese de diftamida pode indicar uma maior susceptibilidade à infecção bacteriana após a aplicação do estresse hídrico.

Foram descobertas três categorias do SEED com aumento de abundância após a aplicação do estresse hídrico, sendo elas relacionadas a degradação de proteínas para obtenção de energia (degradação de arginina e ornitina) e enzimas oxirredutases atuantes na cadeia respiratória aeróbia (succinato desidrogenase) e processos fermentativos (cetoisovalerato oxirredutase). A degradação dos aminoácidos arginina e ornitina, por exemplo, leva à produção de putrescina e posteriormente succinato, este último sendo usado no ciclo do ácido cítrico para produção de energia (Caspi et al. 2014). Assim sendo, pode-se inferir uma provável utilização de fontes alternativas de energia (aminoácidos) por parte da microbiota da rizosfera para poder sobreviver às condições de estresse hídrico aplicadas nesse experimento.

### **Expressão diferencial dos transcritos em resposta ao estresse hídrico**

Antes da etapa de detecção dos transcritos diferencialmente expressos, foi feita uma filtragem dos transcritos exclusivos da amostra de solo sem planta (Bulk 1) em comparação com as amostras com planta (WS0d 1 e WS0d2), dessa forma retirou-se das análises os transcritos que não estavam relacionados com a planta. Finalmente, foram filtrados 24.715 transcritos do total de 405.547 transcritos, restando um universo de 380.832 transcritos para as análises seguintes (Figura 9).



**Figura 9:** Filtragem dos transcritos exclusivos da amostra sem planta. Foram filtrados 24.715 transcritos de total de 405.547 transcritos.

Para as análises de expressão gênica diferencial foi utilizado o pacote do R denominado NOIseq (Tarazona et al. 2012). Juntamente com as funções necessárias para o cálculo da expressão gênica diferencial, o pacote NOIseq também disponibiliza ferramentas para análise e filtragem dos dados. Dessa forma, verificou-se uma falta de profundidade da amostra WS8d 2 e a utilização da mesma poderia influenciar negativamente o desempenho das análises, assim essa amostra foi retirada das análises seguintes.

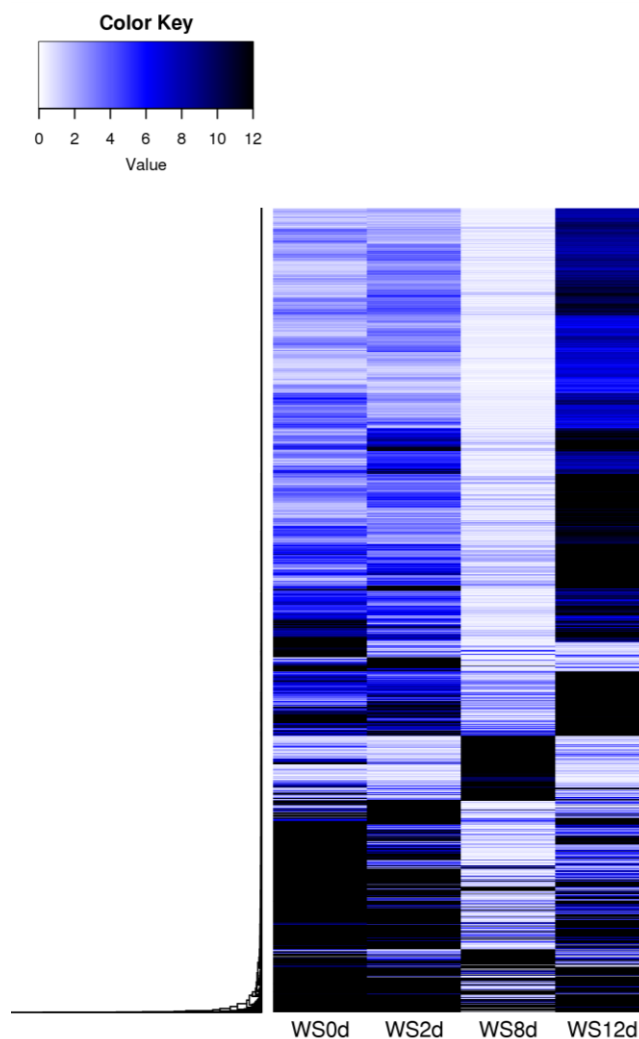
Nas análises da expressão gênica diferencial, foram feitas todas as comparações par a par entre os diferentes tratamentos de estresse hídrico aplicados nesse estudo. de todos os possíveis pares das amostras da rizosfera revelou um total de 43.716 transcritos diferencialmente expressos em pelo menos um dos tratamentos de estresse hídrico avaliado (Tabela 5). Pode-se observar uma tendência de queda nos valores de expressão gênica ao longo do estresse hídrico, uma vez que dos 43.716 transcritos diferencialmente expressos 77,5% deles foram *down* regulados nas amostras com maior tempo de duração do estresse hídrico, ao passo que apenas 22,5% foram *up* regulados.

Comparações	Transcritos DE	<i>Up regulated*</i>	<i>Down regulated*</i>
WS0d x WS2d	924	140 (15,2%)	784 (84,8%)
WS0d x WS8d	8.216	1.371 (16,7%)	6.845 (83,3%)
WS0d x WS12d	6.556	53 (0,8%)	6.503 (99,2%)
WS2d x WS8d	7.334	2.611 (35,6%)	4.723 (64,4%)
WS2d x WS12d	2.766	41 (1,5%)	2.725 (98,5%)
WS8d x WS12d	35.178	30.845 (87,7%)	4.333 (12,3%)
Total (únicos)	43.716	33.896 (77,5%)	15.549 (22,5%)

**Tabela 5:** Número de transcritos diferencialmente expressos nas comparações par a par dos diferentes tratamentos de estresse hídrico aplicados.

\**Up* ou *down* regulados no segundo termo da comparação.

Os 43.716 transcritos diferencialmente expressos foram plotados em um heatmap para que se possa observar os padrões mais gerais da resposta da microbiota ao estresse hídrico aplicado. Os valores de expressão gênica podem ser observados na Figura 10, os dados foram normalizados pela estratégia TMM do pacote NOIseq. É possível observar uma maior similaridade entre as duas primeiras amostras, WS0d e WS2d em relação às demais amostras. Também se nota que a maioria dos transcritos diferencialmente expressos apresentam valores baixos de expressão na amostra WS8d. Finalmente, nota-se um grupo de transcritos que apenas apresentaram aumento de expressão na amostra WS12d. Esse padrão não linear de expressão gênica demonstra uma resposta complexa da microbiota da rizosfera de cana-de-açúcar frente ao estresse hídrico imposto pelo experimento.



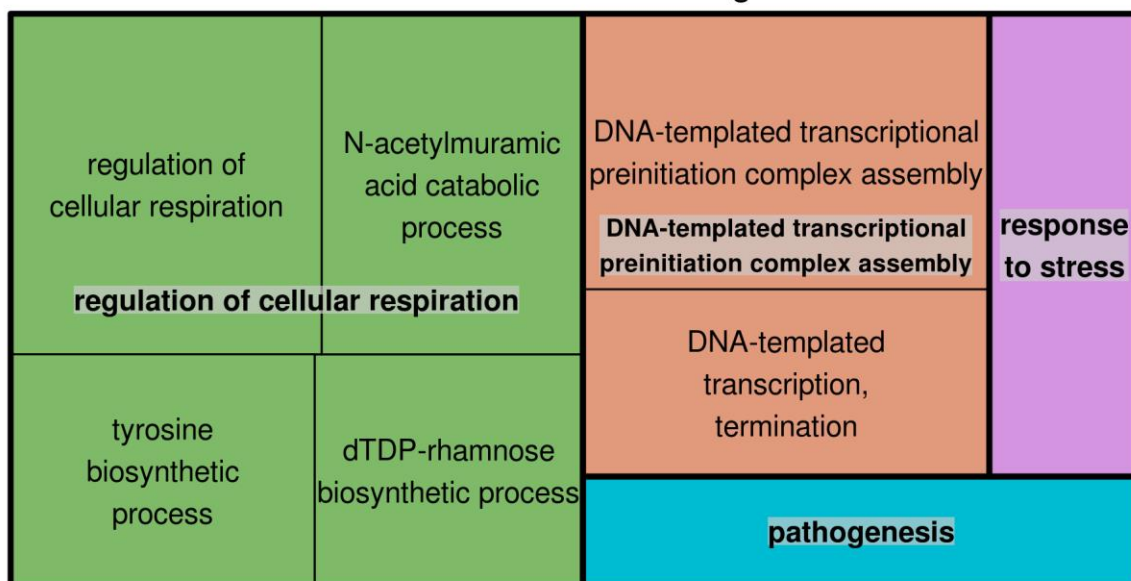
**Figura 10:** Heatmap mostrando o perfil de expressão gênica dos 43.716 transcritos diferencialmente expressos entre os diferentes tratamentos de estresse hídrico analisados.

Para se analisar as respostas funcionais da microbiota da rizosfera aos tratamentos de estresse hídrico, os transcritos diferencialmente expressos foram analisados sobre a perspectiva de processos biológicos possibilitada pelo enriquecimento de termos GO (Gene Ontology). Essa análise foi feita pelo pacote do R denominado topGO (Alexa and Rahnenfuhrer 2010), que comparou determinados subconjuntos de transcritos do metatranscriptoma em relação a todo o metatranscriptoma, verificando-se processos biológicos estatisticamente enriquecidos nos subconjuntos de transcritos testados. Os resultados do topGO foram submetidos ao servidor Web REVIGO (Supek et al. 2011) que

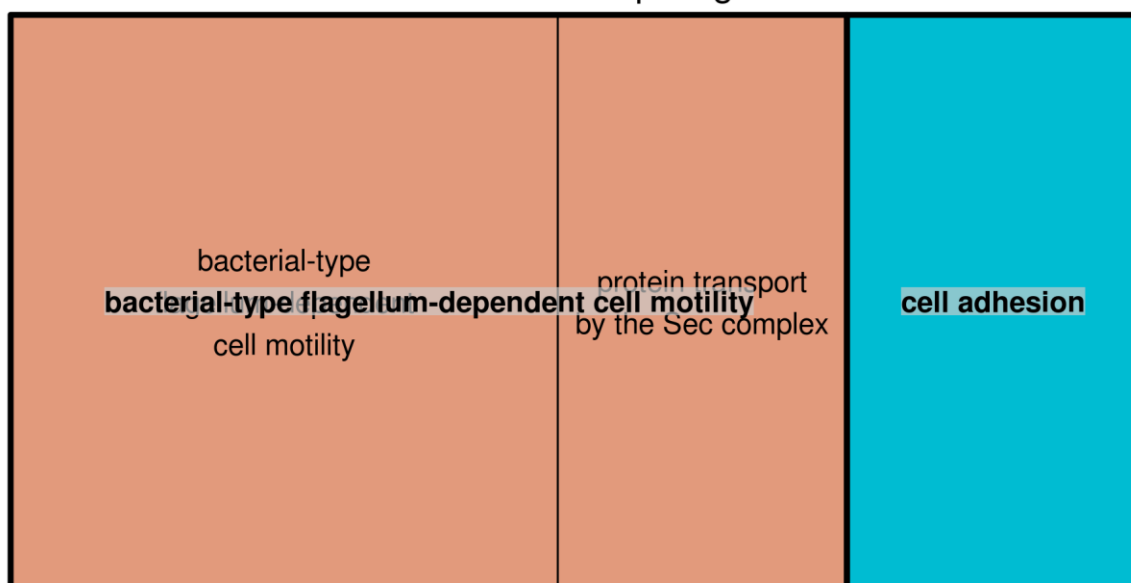
sumarizou os termos GO e retirou termos redundantes, posteriormente plotando os resultados em *treemaps* para a visualização dos dados.

Os principais termos GO enriquecidos nos transcritos *down* regulados na comparação entre WS0d e WS2d foram regulação da respiração celular, montagem do complexo de pré-iniciação transcricional DNA-moldada, resposta ao estresse e patogenicidade. Por outro lado, os principais termos GO enriquecidos nos transcritos *up* regulados entre WS0d e WS2d foram motilidade celular dependente de flagelo e adesão celular (Figura 11).

## WS0d vs WS2d - Down Regulated



## WS0d vs WS2d - Up Regulated

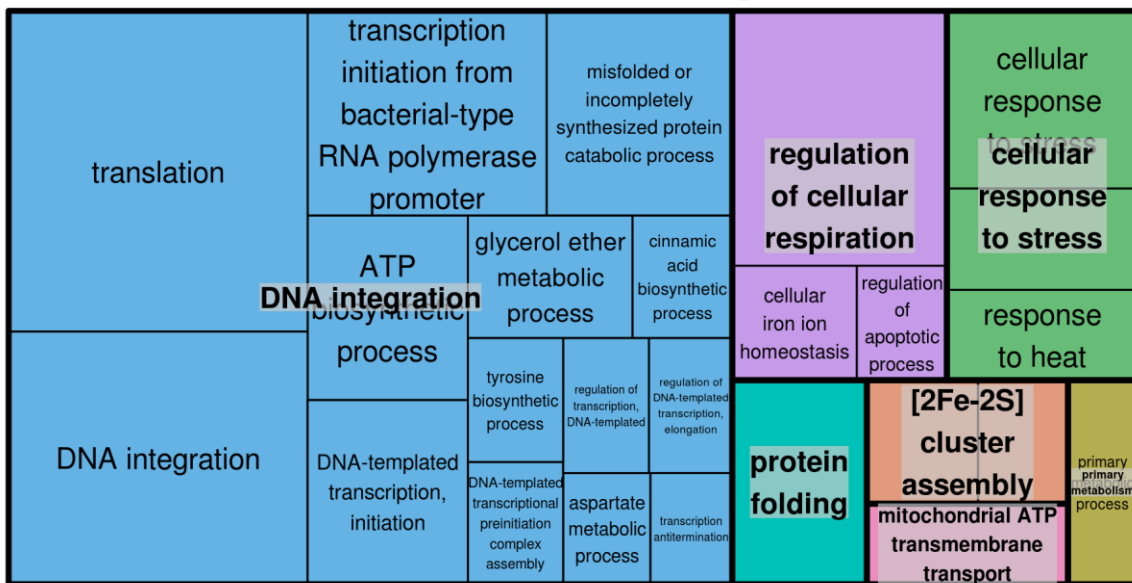


**Figura 11:** Análise de enriquecimento de termos GO relacionados a processos biológicos. Os termos GO foram resumidos e visualizados em um *treemap* pelo servidor Web REVIGO. O tamanho dos quadros indica o  $-\log_{(10)}$  do p-valor do enriquecimento dos termos GO. Quadros com cores iguais pertencem a um mesmo termo GO pai.

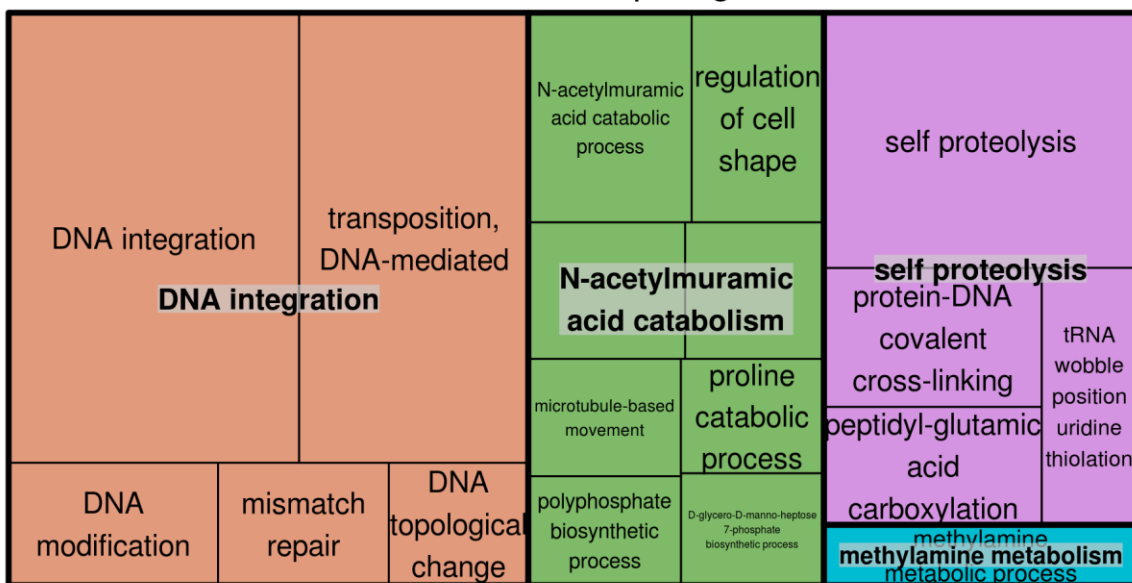
Para a comparação entre WS0d e WS8d, os principais termos GO enriquecidos nos transcritos *down* regulados foram integração ao DNA, regulação da respiração celular,

dobramento de proteínas, resposta celular ao estresse, montagem do cluster 2Fe-2S, transporte transmembrana de ATP mitocondrial e metabolismo primário. Por outro lado, os principais termos GO enriquecidos nos transcritos *up* regulados entre WS0d e WS8d foram integração ao DNA, catabolismo de ácido N-acetilmuramico, auto proteólise e metabolismo de metilamina (Figura 12).

WS0d vs WS8d - Down Regulated



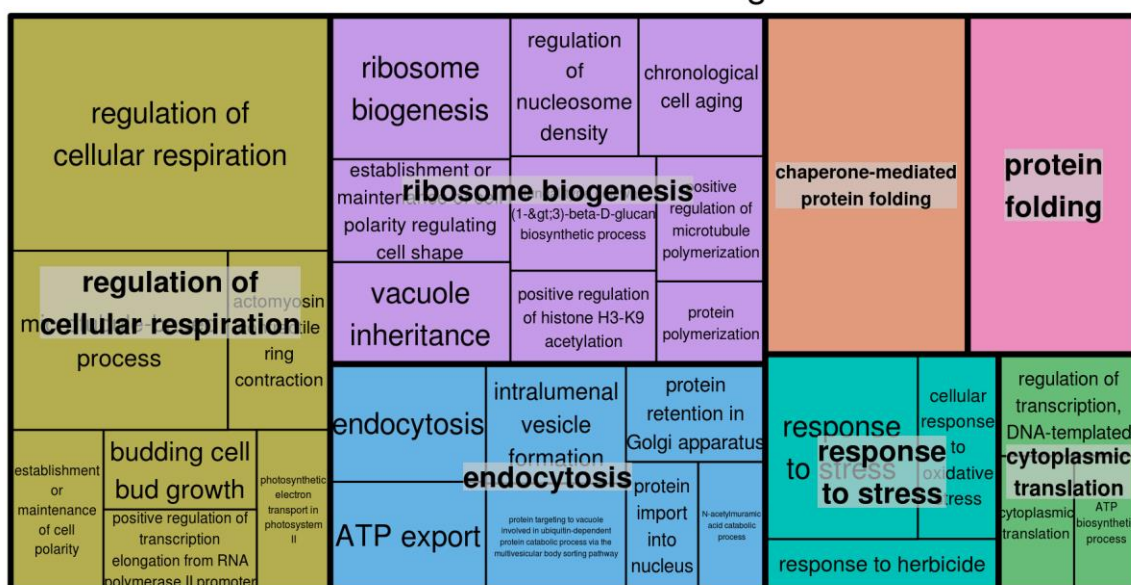
WS0d vs WS8d - Up Regulated



**Figura 12:** Análise de enriquecimento de termos GO relacionados a processos biológicos. Os termos GO foram resumidos e visualizados em um *treemap* pelo servidor Web REVIGO. O tamanho dos quadros indica o  $-\log_{(10)}$  do p-valor do enriquecimento dos termos GO. Quadros com cores iguais pertencem a um mesmo termo GO pai.

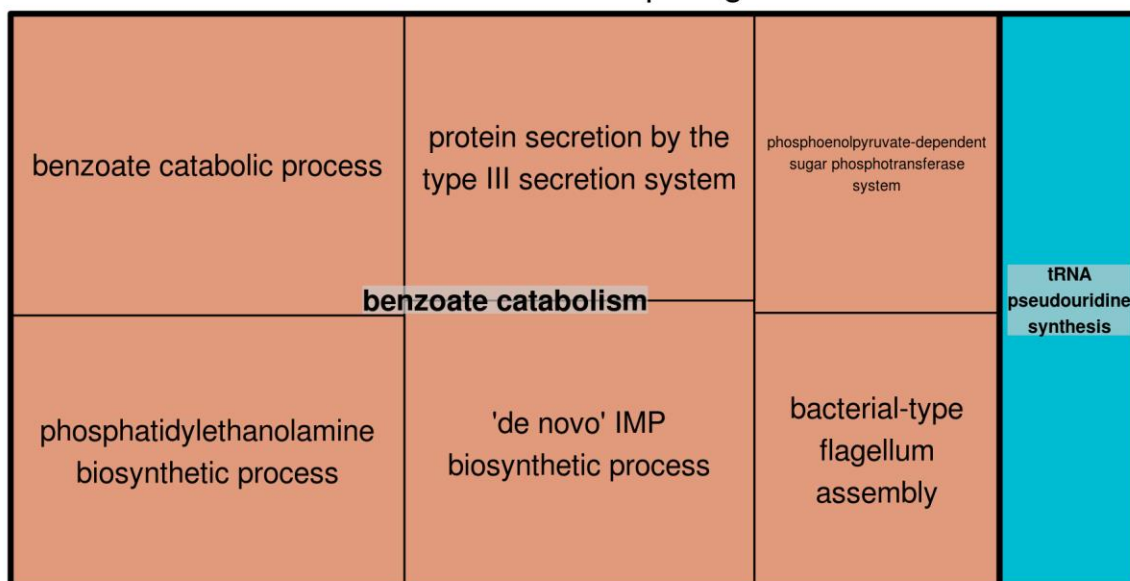
Finalmente, para a comparação entre WS0d e WS12d, os principais termos GO enriquecidos nos transcritos *down* regulados foram regulação da respiração celular, biogênese de ribossomo, dobramento de proteína mediado por chaperona, dobramento de proteína, endocitose, resposta ao estresse e tradução citoplasmática. Por outro lado, os principais termos GO enriquecidos nos transcritos *up* regulados entre WS0d e WS12d catabolismo de benzoato, montagem de flagelo do tipo bactéria e síntese de tRNA de pseudouridina (Figura 13).

### WS0d vs WS12d - Down Regulated



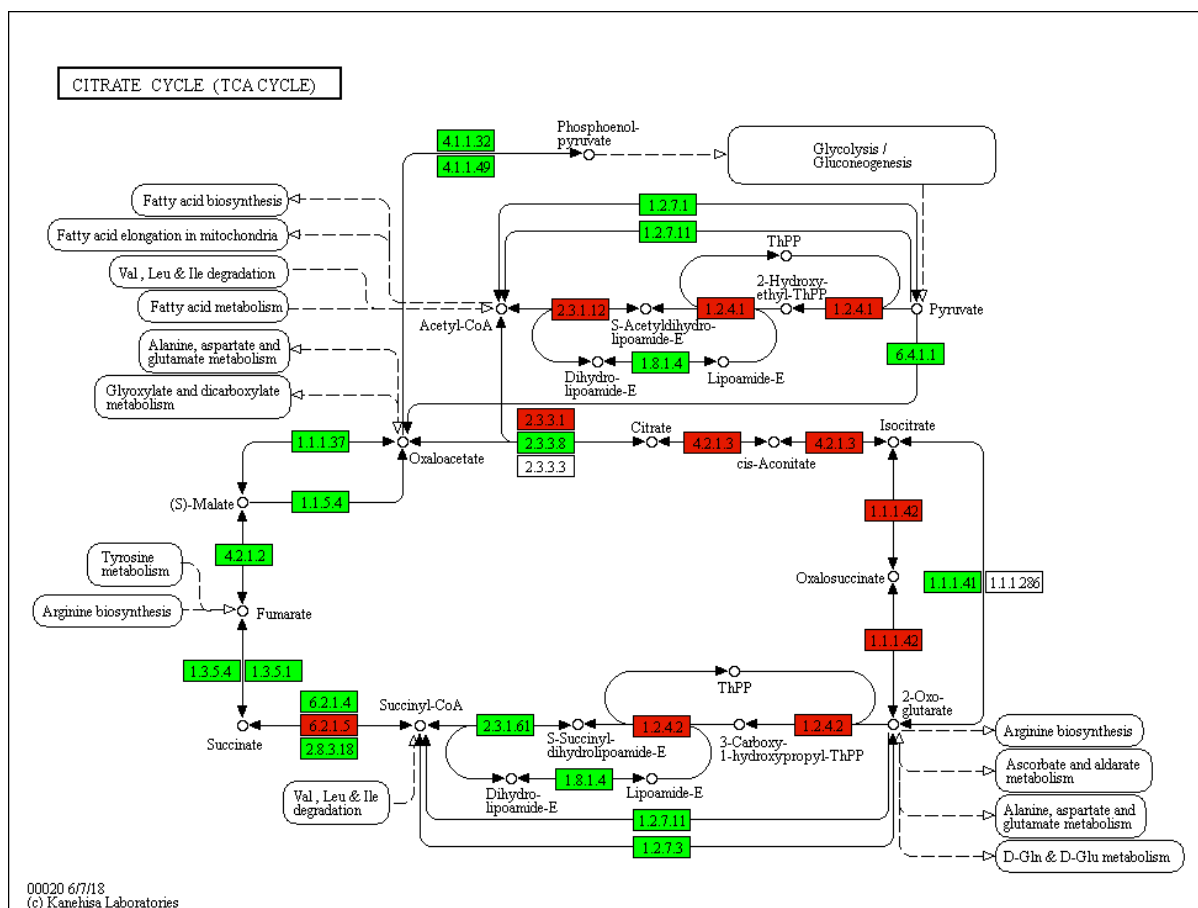


## WS0d vs WS12d - Up Regulated



**Figura 13:** Análise de enriquecimento de termos GO relacionados a processos biológicos. Os termos GO foram resumidos e visualizados em um treemap pelo servidor Web REVIGO. O tamanho dos quadros indica o  $-\log_{(10)}$  do p-valor do enriquecimento dos termos GO. Quadros com cores iguais pertencem a um mesmo termo GO pai.

Baseando-se nas informações obtidas até agora, pode-se observar que, dentre as várias alterações funcionais causadas pelo estresse hídrico, várias dessas alterações dizem respeito ao metabolismo energético. Dessa forma, transcritos relacionados ao ciclo do ácido cítrico foram visualizados através do banco de dados de vias metabólicas KEGG, para obter-se uma visão geral das mudanças que ocorreram na citada via metabólica – Figura 14.

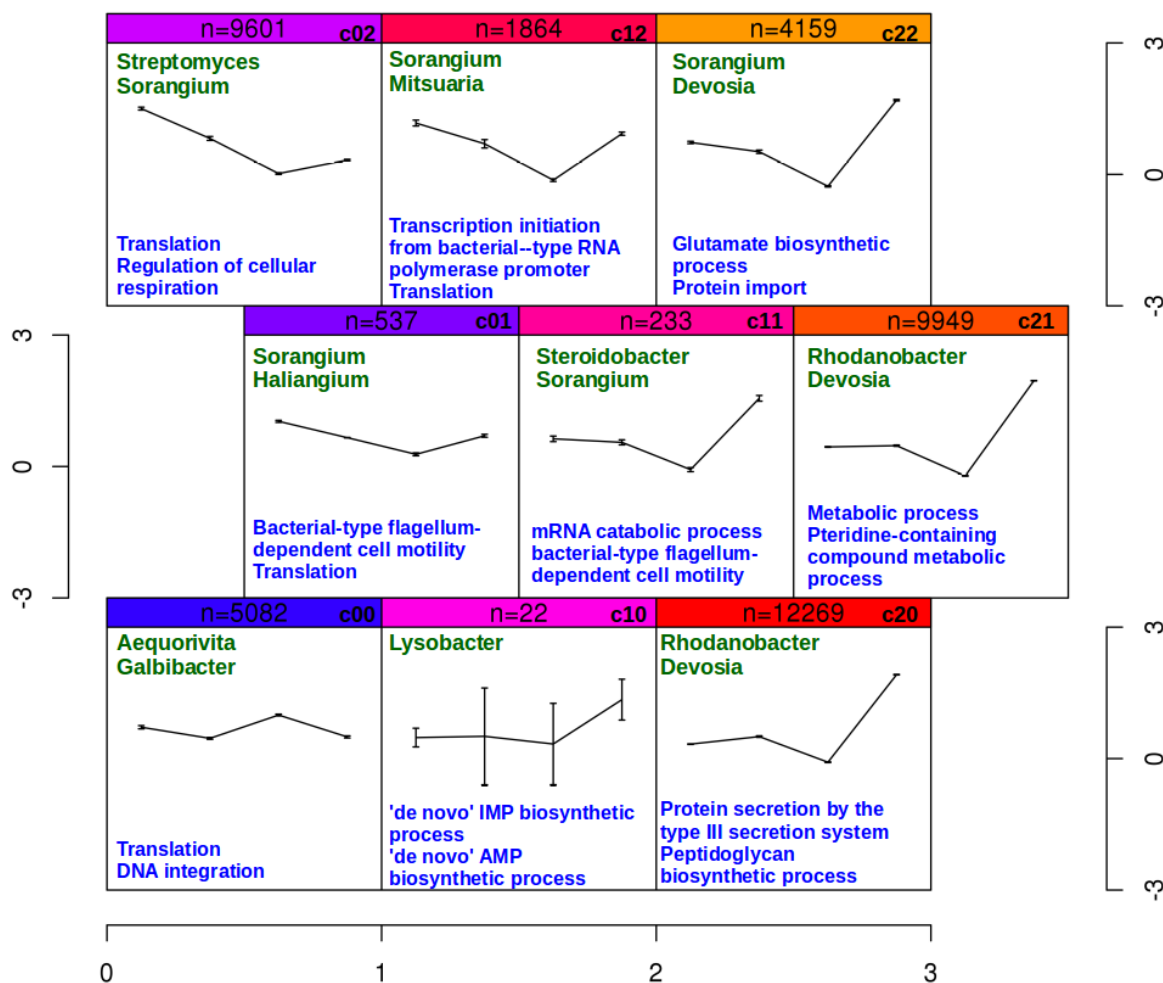


**Figura 14:** Ciclo do ácido cítrico no banco de dados KEEG. Em vermelho são os transcritos que tiveram aumento estatisticamente significativo nos valores de expressão, em verde são os transcritos que tiveram diminuição na expressão e em branco são transcritos que não tiveram alteração estatisticamente significativa.

A análise do ciclo do ácido cítrico do KEEG revelou um aumento de expressão gênica dos transcritos relacionados, principalmente, ao ácido cítrico e isocítrico. Essa informação pode ter relação com o aumento da liberação desses ácidos orgânicos pela planta na forma de exsudatos. Esse aumento foi analisado por Pereira et al., 2018, onde constatou-se um aumento significativo da exsudação desses dois ácidos na rizosfera durante os tratamentos de estresse hídrico. Dessa forma, isso pode significar que os microrganismos capazes de sobreviver às condições de estresse impostas foram aqueles com capacidade de moldar seu metabolismo energético de forma a utilizar as fontes de carbono oferecidas pela planta. Por

outro lado, isso também corrobora com a hipótese de que a planta pode selecionar os microrganismos benéficos à ela através dos compostos químicos liberados na rizosfera.

Por fim, uma vez que este experimento revelou mais de 40 mil transcritos diferencialmente expressos, se fez necessária a adoção de estratégias de análises de dados em larga escala. Dessa maneira, foi utilizada uma categoria de algoritmos baseados em redes neurais artificiais denominado mapas auto organizáveis (SOM, do inglês Self-Organizing Maps). Para isso, foi utilizado um pacote do R chamado SOM (J. Yan, 2010) para clusterizar os mais de 40 mil transcritos diferencialmente expressos em nove clusters diferentes de acordo com a similaridade no perfil de expressão dos transcritos ao longo dos diferentes tratamentos de estresse hídrico (Figura 15). Para cada cluster identificado pelo SOM, foram realizadas análises de enriquecimento de termos GO com o pacote do R denominado topGO, além disso os gêneros mais abundantes de cada cluster foram identificados.



**Figura 15:** Self-organizing map (SOM) análise dos 43.715 transcritos diferencialmente expressos na microbiota associada a rizosfera de cana-de-açúcar CTC 9001. Em cada quadro, o eixo x indica amostras (0, 2, 8 e 12 dias de estresse hídrico) e o eixo y indica o valor de centralização da média da expressão gênica de um determinado conjunto de transcritos (valor “n” na parte superior). Cores similares na parte superior indicam semelhança no padrão de expressão dos transcritos. Cada quadro contém o nome dos dois GO mais enriquecidos (menor valor de p) e os gêneros mais abundantes relacionados ao perfil de expressão apresentado.

A clusterização do SOM gerou nove clusters, desses nove, os clusters c20, c21 e c02 reuniram um grande número de transcritos, com 12.269, 9.949 e 9.601 transcritos em cada um, respectivamente. Outros três clusters reuniram um número intermediário de transcritos, sendo eles c00, c12 e c22, com 5.082, 1.864 e 4.159 transcritos, respectivamente. Finalmente,

três clusters reuniram apenas um pequeno número de transcritos, sendo eles c10, c01 e c11, com 22, 537 e 233 transcritos, respectivamente.

Dentre os três maiores clusters, c20 e c21 apresentaram tendência à aumento de expressão ao longo do estresse hídrico e c02 demonstrou um perfil de queda nos valores de expressão gênica ao longo do estresse hídrico. Os dois grandes clusters com aumento de expressão, c20 e c21, tiveram os mesmos dois gêneros como mais abundantes, *Rhodanobacter* e *Devosia*. As duas bactérias são Gram negativas, sendo *Rhodanobacter* uma gammaproteobacteria e *Devosia* uma alphaproteobacteria. O gênero *Rhodanobacter* já foi identificado em raízes de cana-de-açúcar (Beneduzi et al. 2013) e várias vezes foi ligado à processos de desnitrificação (Green et al. 2012; Lee et al. 2007; Prakash et al. 2012). A outra bactéria abundante nos clusters c20 e 21 pertence ao gênero *Devosia*, descrito em alguns estudos sobre a rizosfera de cana-de-açúcar (Dong et al. 2018; Yeoh et al. 2016) e identificada como uma bactéria formadora de nódulos de fixação de nitrogênio no legume aquático *Neptunia natans* (Rivas et al. 2002).

Os quatro processos biológicos mais enriquecidos nesses dois clusters, c20 e c21, foram processo metabólico, processo metabólico de composto contendo pteridina, secreção de proteínas pelo sistema de secreção do tipo III e processo biossintético de peptidoglicano. A pteridina é precursora na via metabólica de produção de ácido fólico, que por sua vez é utilizado pelas células para produzir DNA e RNA (DellaPenna 2007). A síntese de ácido fólico depende da ação da enzima diidropteroato sintetase que é responsável pela incorporação da pteridina ao ácido diidrofólico. Essa enzima é o alvo dos antibióticos da classe das sulfonamidas, que pela inibição da mesma impedem a síntese de ácidos nucleicos pelas bactérias (Haasum et al. 2001). Adicionalmente, em alinhamento feito contra o banco de dados de resistência à antibióticos CARD (McArthur et al. 2013), foram encontrados 430 transcritos no metatranscriptoma relacionados à resistência a antibióticos da classe das

sulfonamidas e a expressão desses transcritos aumentou em direção ao final do estresse hídrico (Figura S1). Esse fato poderia indicar um possível aumento na produção de pteridina para reduzir os efeitos de sulfonamidas secretadas por membros da comunidade rizosférica.

Outro processo biológico altamente enriquecido nos clusters c20 e c21 foi o processo biossintético de peptidoglicano. Zhai et al. (Zhai et al. 2014) encontrou que super expressão de uma lipoproteína associada a peptidoglicano aumentou a tolerância à estresse salino de *Escherichia coli*, uma bactéria Gram-negativa. Esse resultado sugere uma possível alteração da microbiota da rizosfera de cana-de-açúcar na composição da parede celular em resposta ao estresse hídrico aplicado.

O cluster c02, foi o cluster com maior número de transcritos dentre os clusters com perfil de diminuição da expressão ao longo do estresse. Os dois gêneros mais abundantes desse cluster foram *Streptomyces* e *Sorangium* e os dois processos biológicos mais enriquecidos foram tradução e regulação da respiração celular. A diminuição na expressão desses dois processos pode indicar uma possível diminuição da atividade microbiana relacionada ao estresse hídrico aplicado.

## **Conclusão**

Através das análises preliminares desenvolvidas no capítulo I e dos métodos estabelecidos no capítulo II, foi possível fazer uma caracterização taxonômica e funcional a microbiota da rizosfera de cana-de-açúcar e descrever as mudanças causadas nesse microbioma pela aplicação do estresse hídrico. Taxonomicamente, foram encontradas bactérias com aumento de abundância relativa após o tratamento de estresse hídrico, como por exemplo *Actinomadura*, *Dongia*, *Sphingomonas* e *Vulgatibacter*. Funcionalmente, foram descobertas três categorias do segundo nível da hierarquia do banco de dados SEED com aumento de abundância após a aplicação do estresse hídrico, sendo elas relacionadas a

degradação de proteínas para obtenção de energia (degradação de arginina e ornitina) e enzimas oxirredutases atuantes na cadeia respiratória aeróbia (succinato desidrogenase) e em processos fermentativos (cetoisovalerato oxirredutase).

Também foi realizada a análise da expressão gênica diferencial estimulada pelo tratamento de estresse hídrico ao qual a microbiota foi submetida, o que resultou em 43.716 transcritos diferencialmente expressos e demonstrou que foram *down*-regulados termos GO relacionados a regulação da respiração celular, biogênese de ribossomo, dobramento de proteína, endocitose, resposta ao estresse e tradução. Por outro lado, os principais termos GO enriquecidos nos transcritos *up*-regulados foram catabolismo de benzoato, montagem de flagelo e síntese de tRNA de pseudouridina.

Finalmente, o capítulo II demonstrou o pipeline desenvolvido nesse trabalho para as análises de dados de metatranscriptoma e também a aplicação desse pipeline nos dados de metatranscriptoma de rizosfera de cana-de-açúcar com sucesso.

## Referências bibliográficas

- Alexa, A., & Rahnenfuhrer, J. (2010). topGO: enrichment analysis for gene ontology. R Package Version. Retrieved from <http://bioconductor.uib.no/2.7/bioc/html/topGO.html>
- Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. <https://doi.org/citeulike-article-id:11583827>
- Ara, I., Matsumoto, A., Abdul Bakir, M., Kudo, T., Omura, S., & Takahashi, Y. (2008). *Actinomadura maheshkhaliensis* sp. nov., a novel actinomycete isolated from mangrove rhizosphere soil of Maheshkhali, Bangladesh. *The Journal of General and Applied Microbiology*, 54(6), 335–342. <https://doi.org/10.2323/jgam.54.335>
- Beneduzi, A., Moreira, F., Costa, P. B., Vargas, L. K., Lisboa, B. B., Favreto, R., ... Passaglia, L. M. P. (2013). Diversity and plant growth promoting evaluation abilities of bacteria isolated from sugarcane cultivated in the South of Brazil. *Applied Soil Ecology*, 63, 94–104. <https://doi.org/10.1016/j.apsoil.2012.08.010>
- Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12(1), 59–60. <https://doi.org/10.1038/nmeth.3176>
- Cantarel, B. I., Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V., & Henrissat, B. (2009). The Carbohydrate-Active EnZymes database (CAZy): An expert resource for glycogenomics. *Nucleic Acids Research*, 37(SUPPL. 1). <https://doi.org/10.1093/nar/gkn663>
- Caspi, R., Altman, T., Billington, R., & Dreher, K. (2014). The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases -- Caspi et al. 36 (Supplement 1): D623 -- *Nucleic Acids Research*. *Nucleic Acids Research*, 36(Database), D623–D631. <https://doi.org/10.1093/nar/gkm900>



- CAZypedia Consortium. (2018). Ten years of CAZypedia: a living encyclopedia of carbohydrate-active enzymes. *Glycobiology*, 28(1), 3–8. <https://doi.org/10.1093/glycob/cwx089>
- da Costa, D. P., Dias, A. C. F., Cotta, S. R., Vilela, D., de Andrade, P. A. M., Pellizari, V. H., & Andreote, F. D. (2018). Changes of bacterial communities in the rhizosphere of sugarcane under elevated concentration of atmospheric CO<sub>2</sub>. *GCB Bioenergy*, 10(2), 137–145. <https://doi.org/10.1111/gcbb.12476>
- De Souza, R. S. C., Okura, V. K., Armanhi, J. S. L., Jorrín, B., Lozano, N., Da Silva, M. J., ... Arruda, P. (2016). Unlocking the bacterial and fungal communities assemblages of sugarcane microbiome. *Scientific Reports*, 6. <https://doi.org/10.1038/srep28774>
- DellaPenna, D. (2007). Biofortification of plant-based food: enhancing folate levels by metabolic engineering. *Proceedings of the National Academy of Sciences of the United States of America*, 104(10), 3675–3676. <https://doi.org/10.1073/pnas.0700640104>
- Dong, M., Yang, Z., Cheng, G., Peng, L., Xu, Q., & Xu, J. (2018). Diversity of the bacterial microbiome in the roots of four saccharum species: *S. spontaneum*, *S. robustum*, *S. barberi*, and *S. officinarum*. *Frontiers in Microbiology*, 9(FEB). <https://doi.org/10.3389/fmicb.2018.00267>
- Gordon, A., & Hannon, G. J. (2010). Fastx-toolkit. FASTQ/A short-reads pre-processing tools. Unpublished [Http://Hannonlab.Cshl.Edu/Fastx\\_Toolkit](http://Hannonlab.Cshl.Edu/Fastx_Toolkit).
- Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., ... 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>

- Green, S. J., Prakash, O., Jasrotia, P., Overholt, W. A., Cardenas, E., Hubbard, D., ... Kostka, J. E. (2012). Denitrifying bacteria from the genus *Rhodanobacter* dominate bacterial communities in the highly contaminated subsurface of a nuclear legacy waste site. *Applied and Environmental Microbiology*, 78(4), 1039–1047. <https://doi.org/10.1128/AEM.06435-11>
- Haasum, Y., Ström, K., Wehelie, R., Luna, V., Roberts, M. C., Maskell, J. P., ... Swedberg, G. (2001). Amino acid repetitions in the dihydropteroate synthase of *Streptococcus pneumoniae* lead to sulfonamide resistance with limited effects on substrate Km. *Antimicrobial Agents and Chemotherapy*, 45(3), 805–809. <https://doi.org/10.1128/AAC.45.3.805-809.2001>
- Hayden, H. L., Savin, K. W., Wadeson, J., Gupta, V. V. S. R., & Mele, P. M. (2018). Comparative metatranscriptomics of wheat Rhizosphere microbiomes in disease suppressive and non-suppressive soils for *Rhizoctonia solani* AG8. *Frontiers in Microbiology*, 9(MAY). <https://doi.org/10.3389/fmicb.2018.00859>
- Huson, D. H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., ... Tappu, R. (2016). MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLoS Computational Biology*, 12(6). <https://doi.org/10.1371/journal.pcbi.1004957>
- Hyatt, D., Chen, G. L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11. <https://doi.org/10.1186/1471-2105-11-119>
- Jia, B., Raphenya, A. R., Alcock, B., Waglechner, N., Guo, P., Tsang, K. K., ... McArthur, A. G. (2017). CARD 2017: Expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Research*, 45(D1), D566–D573. <https://doi.org/10.1093/nar/gkw1004>

- Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4), 357–359. <https://doi.org/10.1038/nmeth.1923>
- Lee, C. S., Kim, K. K., Aslam, Z., & Lee, S. T. (2007). *Rhodanobacter thiooxydans* sp. nov., isolated from a biofilm on sulfur particles used in an autotrophic denitrification process. *International Journal of Systematic and Evolutionary Microbiology*, 57(8), 1775–1779. <https://doi.org/10.1099/ijs.0.65086-0>
- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14), 1754–1760. <https://doi.org/10.1093/bioinformatics/btp324>
- Liao, Y., Smyth, G. K., & Shi, W. (2014). FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7), 923–930. <https://doi.org/10.1093/bioinformatics/btt656>
- Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M., & Henrissat, B. (2014). The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Research*, 42(D1). <https://doi.org/10.1093/nar/gkt1178>
- Lubsanova, D. A., Zenova, G. M., Kozhevin, P. A., Manucharova, N. A., & Shvarov, A. P. (2014). Filamentous actinobacteria of the saline soils of arid territories. *Moscow University Soil Science Bulletin*, 69(2), 88–92. <https://doi.org/10.3103/S0147687414020057>
- McArthur, A. G., Waglechner, N., Nizam, F., Yan, A., Azad, M. A., Baylay, A. J., ... Wright, G. D. (2013). The comprehensive antibiotic resistance database. *Antimicrobial Agents and Chemotherapy*, 57(7), 3348–3357. <https://doi.org/10.1128/AAC.00419-13>
- Norovsuren, Z., Zenova, G. M., & Mosina, L. V. (2007). Actinomycetes in the rhizosphere of semidesert soils of Mongolia. *Eurasian Soil Science*, 40(4), 415–418. <https://doi.org/10.1134/S1064229307040072>

- Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H. Y., Cohoon, M., ... Vonstein, V. (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Research*, 33(17), 5691–5702. <https://doi.org/10.1093/nar/gki866>
- Parks, D. H., Tyson, G. W., Hugenholtz, P., & Beiko, R. G. (2014). STAMP: Statistical analysis of taxonomic and functional profiles. *Bioinformatics*, 30(21), 3123–3124. <https://doi.org/10.1093/bioinformatics/btu494>
- Patel, R. K., & Jain, M. (2012). NGS QC toolkit: A toolkit for quality control of next generation sequencing data. *PLoS ONE*, 7(2). <https://doi.org/10.1371/journal.pone.0030619>
- Patro, R., Duggal, G., & Kingsford, C. (2015). Accurate, fast, and model-aware transcript expression quantification with Salmon. *BioRxiv*, 021592. <https://doi.org/10.1101/021592>
- Peng, Y., Leung, H. C. M., Yiu, S. M., & Chin, F. Y. L. (2012). IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, 28(11), 1420–1428. <https://doi.org/10.1093/bioinformatics/bts174>
- Pisa, G., Magnani, G. S., Weber, H., Souza, E. M., Faoro, H., Monteiro, R. A., ... Cruz, L. M. (2011). Diversity of 16S rRNA genes from bacteria of sugarcane rhizosphere soil. *Brazilian Journal of Medical and Biological Research*, 44(12), 1215–1221. <https://doi.org/10.1590/S0100-879X2011007500148>
- Prakash, O., Green, S. J., Jasrotia, P., Overholt, W. A., Canion, A., Watson, D. B., ... Kostka, J. E. (2012). *Rhodanobacter denitrificans* sp. nov., isolated from nitrate-rich zones of a contaminated aquifer. *International Journal of Systematic and Evolutionary Microbiology*, 62(10), 2457–2462. <https://doi.org/10.1099/ijs.0.035840-0>

- Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2007). NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, 35(SUPPL. 1). <https://doi.org/10.1093/nar/gkl842>
- Rho, M., Tang, H., & Ye, Y. (2010). FragGeneScan: Predicting genes in short and error-prone reads. *Nucleic Acids Research*. <https://doi.org/10.1093/nar/gkq747>
- Rivas, R., Velázquez, E., Willems, A., Vizcaíno, N., Subba-Rao, N. S., Mateos, P. F., ... Martínez-Molina, E. (2002). A new species of *Devosia* that forms a unique nitrogen-fixing root-nodule symbiosis with the aquatic legume *Neptunia natans* (L.f.) Druce. *Applied and Environmental Microbiology*, 68(11), 5217–5222. <https://doi.org/10.1128/AEM.68.11.5217-5222.2002>
- Schmieder, R., Lim, Y. W., & Edwards, R. (2012). Identification and removal of ribosomal RNA sequences from metatranscriptomes. *Bioinformatics*, 28(3), 433–435. <https://doi.org/10.1093/bioinformatics/btr669>
- Silva, M. D. A., Jifon, J. L., Da Silva, J. A. G., & Sharma, V. (2007). Use of physiological parameters as fast tools to screen for drought tolerance in sugarcane. *Brazilian Journal of Plant Physiology*, 19(3), 193–201. <https://doi.org/10.1590/S1677-04202007000300003>
- Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19), 3210–3212. <https://doi.org/10.1093/bioinformatics/btv351>
- Su, X., Lin, Z., & Lin, H. (2013). The biosynthesis and biological function of dipthamide. *Critical Reviews in Biochemistry and Molecular Biology*. <https://doi.org/10.3109/10409238.2013.831023>

- Su Yien Ting, A., Hoon Tan, S., & Kay Wai, M. (2009). Isolation and Characterization of Actinobacteria with Antibacterial Activity From Soil and Rhizosphere Soil. *Australian Journal of Basic and Applied Sciences*, 3(4), 4053–4059.
- Supek, F., Bošnjak, M., Škunca, N., & Šmuc, T. (2011). Revigo summarizes and visualizes long lists of gene ontology terms. *PLoS ONE*, 6(7). <https://doi.org/10.1371/journal.pone.0021800>
- Tarazona, S., García, F., Ferrer, A., Dopazo, J., & Conesa, A. (2012). NOIseq: a RNA-seq differential expression method robust for sequencing depth biases. *EMBnet.Journal*, 17(B), 18. <https://doi.org/10.14806/ej.17.B.265>
- Wu, Y., Zhou, H., Que, Y. X., Chen, R. K., & Zhang, M. Q. (2008). Cloning and identification of promoter Prd29A and its application in sugarcane drought resistance. *Sugar Tech*, 10(1), 36–41. <https://doi.org/10.1007/s12355-008-0006-0>
- Xie, S. P., Deser, C., Vecchi, G. A., Collins, M., Delworth, T. L., Hall, A., ... Watanabe, M. (2015). Towards predictive understanding of regional climate change. *Nature Climate Change*. <https://doi.org/10.1038/nclimate2689>
- Xu, L., Naylor, D., Dong, Z., Simmons, T., Pierroz, G., Hixson, K. K., ... Coleman-Derr, D. (2018). Drought delays development of the sorghum root microbiome and enriches for monoderm bacteria. *Proceedings of the National Academy of Sciences*, 201717308. <https://doi.org/10.1073/pnas.1717308115>
- Yang, Y., Wang, N., Guo, X., Zhang, Y., & Ye, B. (2017). Comparative analysis of bacterial community structure in the rhizosphere of maize by highthroughput pyrosequencing. *PLoS ONE*, 12(5). <https://doi.org/10.1371/journal.pone.0178425>
- Yeoh, Y. K., Paungfoo-Lonhienne, C., Dennis, P. G., Robinson, N., Ragan, M. A., Schmidt, S., & Hugenholtz, P. (2016). The core root microbiome of sugarcane cultivated

under varying nitrogen fertilizer application. *Environmental Microbiology*, 18(5), 1338–1351. <https://doi.org/10.1111/1462-2920.12925>

Yin, Y., Mao, X., Yang, J., Chen, X., Mao, F., & Xu, Y. (2012). DbCAN: A web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Research*, 40(W1). <https://doi.org/10.1093/nar/gks479>

Zhai, L., Xue, Y., Song, Y., Xian, M., Yin, L., Zhong, N., ... Ma, Y. (2014). Overexpression of AaPal, a peptidoglycan-associated lipoprotein from *Alkalomonas amylolytica*, improves salt and alkaline tolerance of *Escherichia coli* and *Arabidopsis thaliana*. *Biotechnology Letters*, 36(3), 601–607. <https://doi.org/10.1007/s10529-013-1398-9>

## CAPÍTULO III

---

### Co-expression Network of Sugarcane Rhizosphere Subjected to Water Stress



## Introduction

As a crop of global importance for sugar and bioenergy production, Sugarcane (*Saccharum* spp.) tolerance and adaptation to foreseen climate changes should be a topic of intensive research. Among the expected changes, precipitation patterns are predicted to become more irregular, making rainfall levels likely to increase in high latitudes and the equatorial Pacific and decrease in many mid-latitude and subtropical dry regions (Zandalinas et al. 2018). Drought is one of the most important environmental stress factors for sugarcane yield once it affects tillering and culm height, culminating in reduced sucrose production and generating losses worldwide (M. D. A. Silva et al. 2007; Masoabi et al. 2018).

Plants and soil microorganisms are thought to have co-evolved to overcome environmental stress in their habitats in a plant–bacterial interaction that is beneficial to both partners. Hence, the rhizosphere is recognized as the hotspot of interactive events between soil microbes and plants and that interaction occur through release of chemical molecules from both sides (Dakora, Matiru, and Kanu 2015). These plant-microbe interactions significantly impact plant performance and microbial community structure (Saleem et al. 2018). The rhizosphere has a complex community of microbes that are distinct from the surrounding soil and is estimated to have hundreds of thousands of different species (Haney et al., 2018), however only a small proportion (< 0.5%) of bacteria in soil or plant rhizosphere have been isolated and cultured with traditional laboratory phenotypic tests (Monreal et al. 2018).

Some efforts to understand the dynamics of sugarcane rhizosphere were done by 16S rRNA sequencing (Pisa et al. 2011) traditional isolation techniques (Sellamuthu and Govindaswamy 2003), 16S rRNA and traditional isolation together (De Souza et al. 2016),

and 16S rRNA accompanied with metagenomics (Yeoh et al. 2016). Although these studies gave us a general picture about the sugarcane rhizosphere community composition and genetic potential, none of them were able to demonstrate which microorganisms were really active and what functional role they were playing in the rhizosphere community. In addition, the changes in the sugarcane rhizosphere community under drought was not evaluated yet. To fulfill this necessity, we conducted a metatranscriptome study of sugarcane rhizosphere subjected to water stress and constructed a co-expression network to better understand the high amount of data generated.

## **Materials and methods**

### **Water stress experiment and sample collection**

Four-month-old sugarcane cultivar CTC9001 plants were provided by the Sugarcane Technology Center (CTC, Piracicaba, Brazil), and were kept in greenhouse for three months under normal irrigation before being subjected to water stress. Irrigation was interrupted 12 (WS12d), 8 (WS8d) and 2 days (WS2d) prior to rhizosphere sample collection. We also collected rhizosphere samples from vases in which irrigation was not interrupted (WS0d) and soil samples from vases without plants (Bulk).

To collect the rhizosphere samples, plants were removed from soil and slightly shaken to remove agglomerations of soil not associated with roots. Soil that kept adhered to the roots was then removed by vigorous stirring. Samples were immediately stocked in LifeGuard™ soil RNA conservation solution (MoBio Laboratories, Inc, Carlsbad, CA, USA).

### **RNA extraction and metatranscriptome sequencing**

Soil RNA isolation was performed using RNA PowerSoil® Total RNA Isolation Kit (MoBio Laboratories), purified RNA was quantified using a Qubit fluorometer (Invitrogen, Carlsbad, CA, USA), and RNA quality was evaluated on agarose gel electrophoresis. Approximately 2 µg of total RNA was enriched for bacterial mRNA using MICROBExpress™ Bacterial mRNA Enrichment Kit (Ambion, Thermo Fisher Scientific, NY, USA).

Construction of cDNA libraries, constructions of clusters, libraries quantification using Agilent DNA high-sensitivity kit on Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA), and paired-end 2 × 100 bp sequencing on the Illumina HiSeq 2500 (Illumina, San Diego, CA, USA) was performed by the Central Laboratory of High Performance Technologies (LaCTAD, UNICAMP, Campinas, SP, Brazil).

### **Bioinformatics analysis**

Quality of reads was analyzed with FASTQC v0.11.6 (Andrews 2010) and low-quality reads with  $Q < 20$  and size  $< 70$  were filtered out by NGS QC Toolkit v2.3.3 (Patel and Jain 2012). Ribosomal RNA-like reads were filtered out by riboPicker v0.4.3 (Schmieder, Lim, and Edwards 2012) with the comprehensive riboPicker database “rrnadb” provided by the software authors. Putative mRNA reads were de novo assembled using IDBA-UD v1.1.3 (Peng et al. 2012) with the following parameters: —mink 20, —maxk 100, —step 20, —min\_contig 200. Prediction of open reading frames (ORFs) was performed using FragGeneScan v1.30 (Rho, Tang, and Ye 2010) with the parameters -w 0 for fragmented sequences and -t illumina\_5 for Illumina sequencing reads with about 0.5% error rate. Reads were mapped to ORFs by BWA mem v0.7.12 (Li & Durbin, 2009), the output was converted to sorted BAM by Samtools (H. Li and Durbin 2009), and the BAM files were

used as input to the software featureCounts v1.5.3 (Liao, Smyth, and Shi 2014) to summarize the features of all RNA-seq libraries.

For taxonomic and functional annotation, ORFs were aligned against entire collection of bacterial, fungal and archaeal protein sequences in the NCBI RefSeq database release 84 (Pruitt, Tatusova, and Maglott 2007) using DIAMOND v0.9.10.111 with default settings (Buchfink, Xie, and Huson 2015), following taxonomy assignment using the lowest-common ancestor algorithm as implemented in MEGAN Community Edition v6.8.20 (Huson et al. 2016). GO (Gene Ontology) terms were acquired by mapping RefSeq matches to the Gene Ontology database (Ashburner et al. 2000).

NOISeq package (Tarazona et al. 2012) was used to normalize the count table generated by featureCounts using the TMM (trimmed mean of M-values) normalization technique (Robinson and Oshlack 2010) and to identify differentially expressed transcripts for all pairwise comparisons with a cutoff of  $q = 0.95$ .

Co-expression network was initially constructed from the gene dataset through calculations of co-expression values (all versus all) with Pearson correlation cutoff of 0.8 (van Noort, Snel, and Huynen 2004) in order to capture linear relations of expression among any given pair of genes. Then, the Highest Reciprocal Ranking (HRR) and the Heuristic Cluster Chiseling Algorithm (HCCA) (Mutwil et al. 2010) were implemented to i. set the maximum number of interactions (edges) per transcript (node) to 30 and ii. generate putative clusters within the initial network with the ability of reinsert these clusters into the original network, overcoming data discontinuity due to fragmentation. The final HCCA network was rendered using the software Cytoscape v.3.6.1 (Shannon et al. 2003), also used for GO enrichment analysis, performed on all network clusters, with the BiNGO plugin v3.0.3 (Maere, Heymans, and Kuiper 2005), using up-to-date databases and applying a

hypergeometric test as well as a Benjamini and Hochberg false discovery rate (FDR) correction ( $<0.1$ ).

## Results and discussion

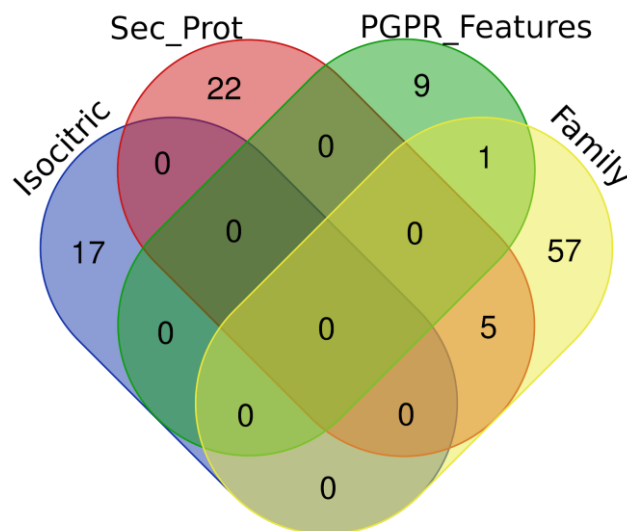
We inferred a microbiome-wide co-expression network by analyzing seven metatranscriptome samples from sugarcane rhizosphere subjected to 2 (2 samples), 8 (1 sample) and 12 (2 samples) days of water stress, as well as two samples without water stress. From the 43,716 differentially expressed transcripts among pairwise comparisons of the treatments, 42,941 could be used to construct the network, generating 582 clusters with sizes between 40 and 154 transcripts and average size of 73 transcripts. Our study aimed to determine co-expression relationships among microorganisms and functions present in the sugarcane rhizosphere metatranscriptome that could respond to the water stress treatments. In addition, we also measured the release of organic acids by the plant into the rhizosphere such as citric, isocitric, and lactic acids (Pereira et al., 2018). These exudates were tested for correlation to the water stress applied – Table S1.

In order to find clusters related to drought, we first searched through the network clusters that had 20% or more of its transcripts up or down-regulated in the pairwise comparisons of the samples with no water stress against all the other samples - Table 1. It revealed 30 unique clusters up-regulated and 160 unique clusters down-regulated.

Comparison	Transcripts up-regulated	Transcripts down-regulated	Clusters up-regulated	Clusters down-regulated
0d vs 2d	140	784	3	14
0d vs 8d	1,371	6,845	27	127
0d vs 12d	53	6,503	0	124
Total unique	1,552	10,506	30	160

**Table 1** Differentially expressed transcripts comparing samples without water stress (0d) against samples with water stress (2d, 8d, and 12d) and the clusters that had 20% or more transcripts up or down-regulated in one of the analyzed comparisons.

We also searched the network for four specific characteristics, (1) PGPR-features, known features related to drought stress such as transcripts related to plant growth hormones, nitrogen fixation, oxidative stress response, siderophores, and protein secretion; (2) isocitric acid, clusters that correlated to isocitric acid which was found to positively correlate with the water stress in our experiment; (3) secreted proteins, clusters that had 10% or more of its transcripts with a signal peptide according to SignalP software and did not have a transmembrane domain according to TMHMM software; (4) family, clusters that had 20% or more of its transcripts taxonomically assigned to the same family - Figure 1, Table 2, and Table S2.

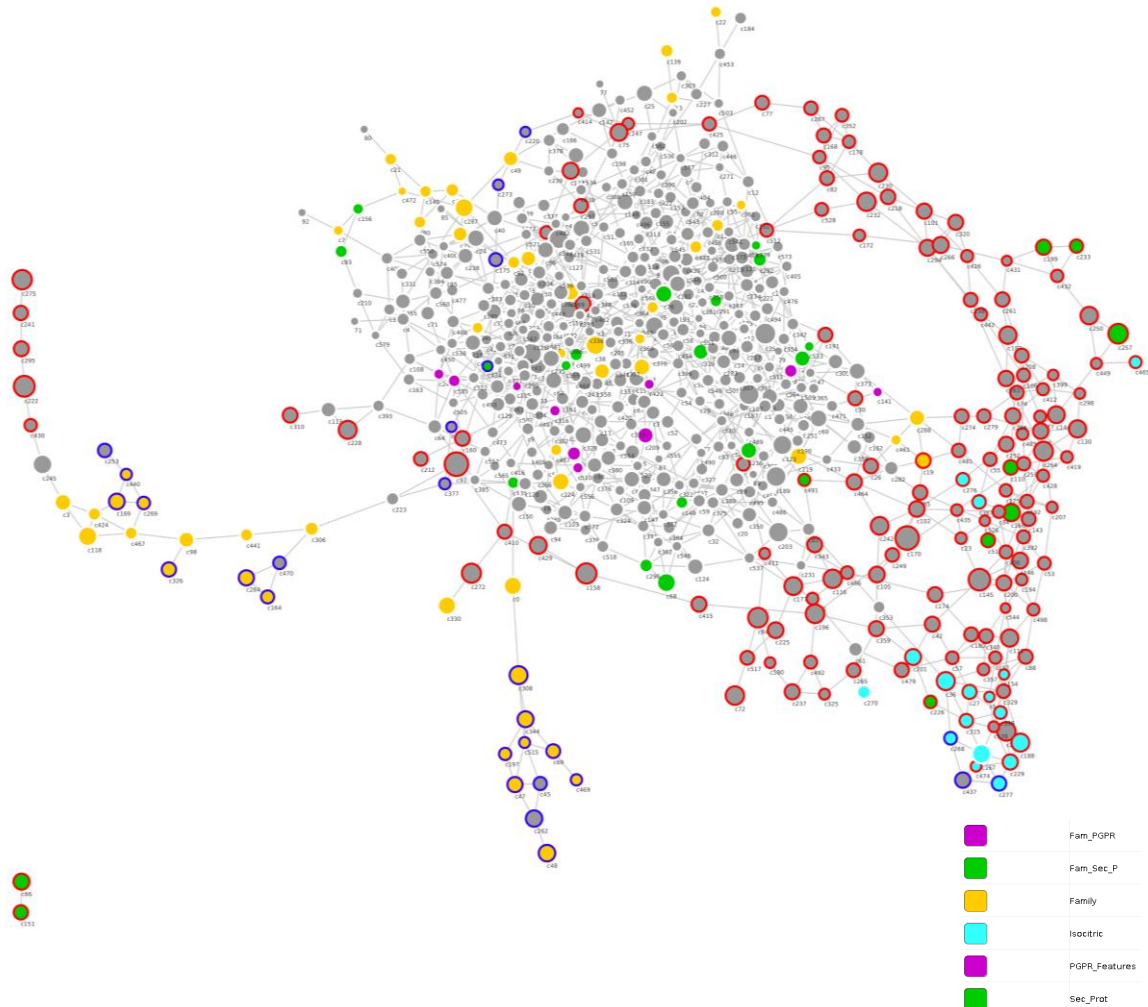


**Figure 1** Venn diagram of clusters that had transcripts related to one or more features of interest.

Features	Number of clusters	Clusters
Family & Secreted proteins	5	c156 c135 c93 c226 c499
Family & PGPR-features	1	c115
Isocitric acid	17	c277 c465 c276 c229 c167 c188 c201 c268 c154 c36 c398 c474 c27 c315 c270 c5 c387
Secreted proteins	22	c319 c491 c216 c110 c434 c569 c233 c257 c66 c339 c76 c296 c533 c149 c151 c395 c292 c291 c68 c166 c51 c199
PGPR-features	9	c535 c141 c248 c100 c423 c209 c421 c510 c316
Family	57	c472 c49 c219 c485 c461 c98 c214 c47 c136 c140 c515 c80 c369 c344 c361 c21 c424 c444 c360 c19 c469 c78 c326 c7 c0 c269 c22 c345 c457 c330 c306 c197 c288 c92 c441 c459 c440 c139 c164 c438 c48 c3 c169 c118 c308 c69 c122 c542 c287 c224 c38 c467 c13 c297 c284 c285 c286

**Table 2** Clusters that had transcripts related to one or more features of interest.

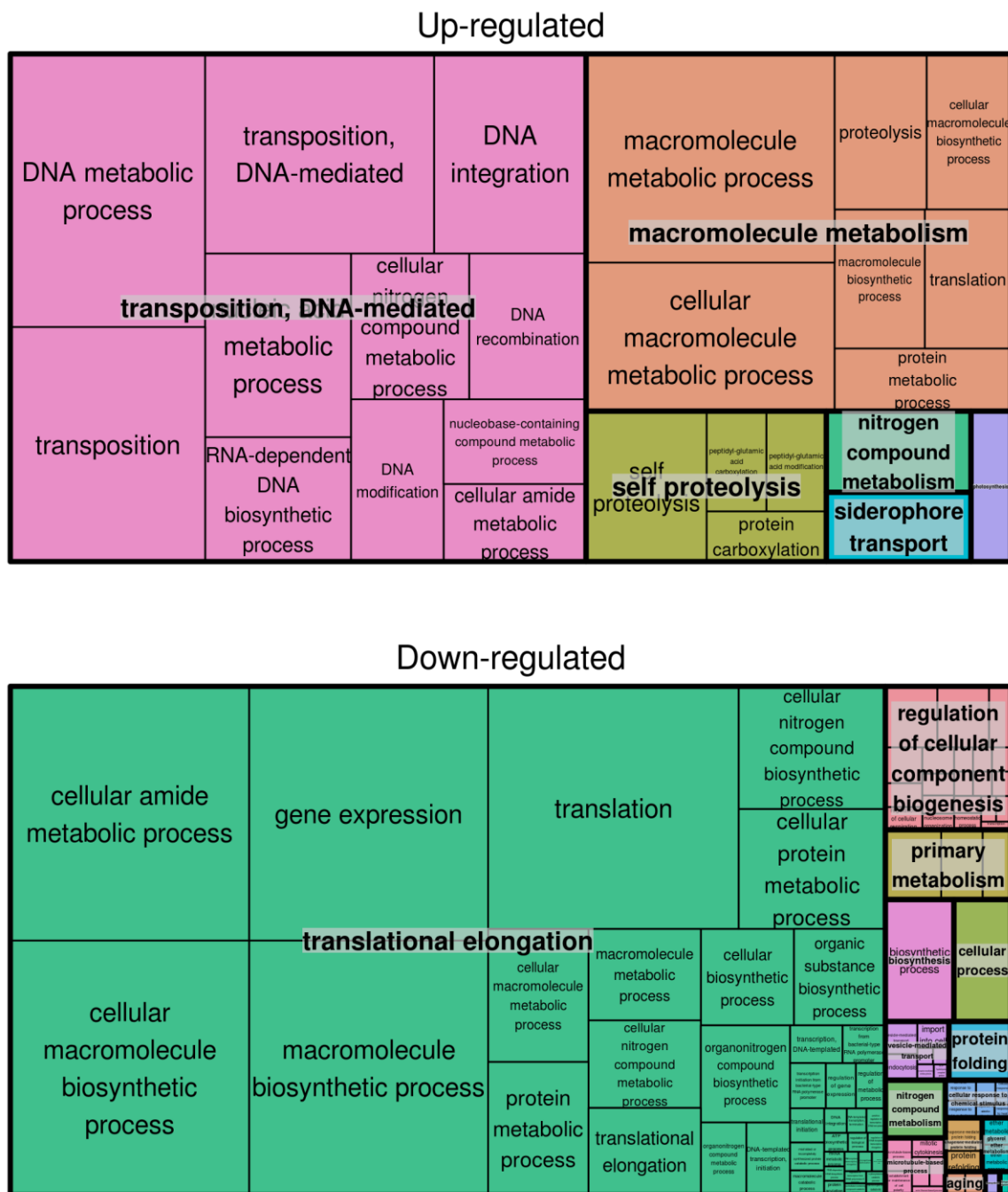
There were 63 clusters with the family attribute, 10 PGPR-features clusters, 17 isocitric acid clusters, and 27 secreted proteins clusters. All those features and also the up or down-regulated clusters were plotted in the Figure 2.



**Figure 2** Co-expression network of 42,941 transcripts clustered in 582 clusters. Nodes in this network represent clusters. Node colors represent different features of clusters, while edges represent significant ( $P < 0.05$ ) associations between the clusters based on co-expression.

The transcripts that belonged to the clusters up or down-regulated were tested for enrichment of GO (Gene Ontology) terms using the Cytoscape plugin BINGO and the result was summarized by the web server REVIGO (Supek et al. 2011) (Figure 3).





**Figure 3** Biological process GO terms enriched in the clusters up or down-regulated. Go terms were summarized and visualized in a treemap by REVIGO. The area of the rectangles indicate the GO terms enrichment  $-\log(10)$  of the p-value. Equal colors represent GO terms from the same parent term.

Among the GO terms enriched in the down-regulated clusters, there were several GO terms related to basic cellular functions such as translational elongation, gene expression,

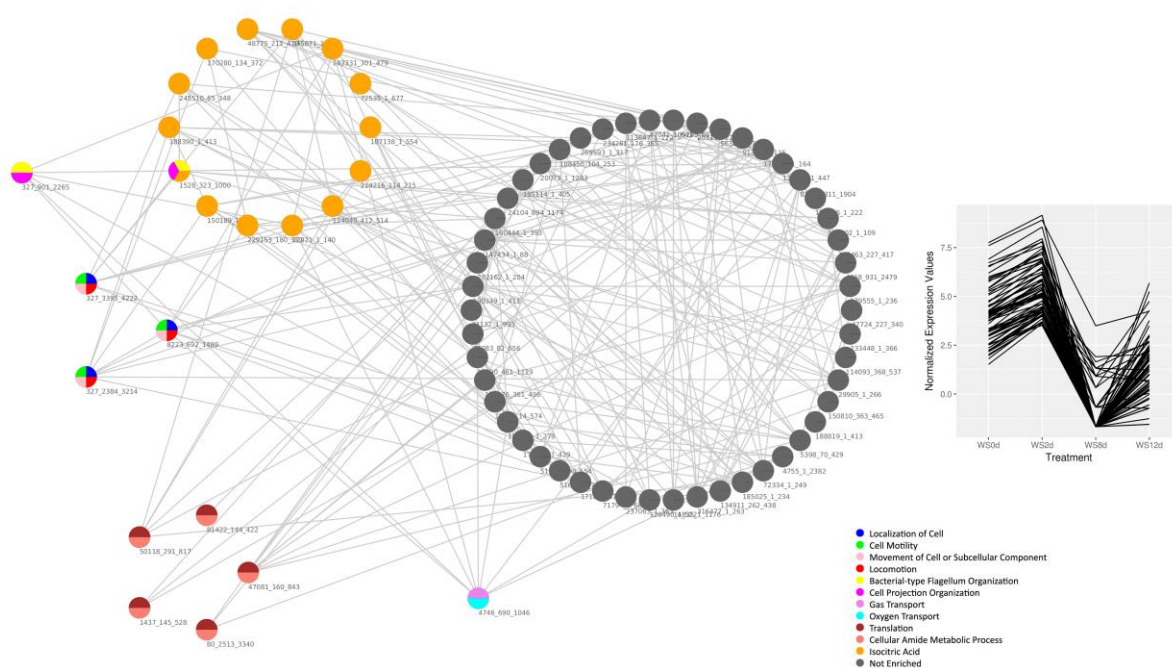
macromolecule biosynthetic process, regulation of cellular component biogenesis, primary metabolism, biosynthesis, protein folding and so forth. Down-regulation of such primary functions may indicate that the drought treatment that was applied was severely inhibiting the metabolism of the rhizosphere microbiota.

On the other hand, the GO terms enriched in the up-regulated clusters were related to DNA metabolic process, transposition, DNA modification, DNA recombination, macromolecule metabolism, self-proteolysis, nitrogen metabolism, siderophore transport and so forth. Up-regulation of terms such as self-proteolysis may corroborate with the hypothesis of inhibition of the microorganism's metabolism by the water stress applied. Additionally, the GO term siderophore transport was enriched in the up-regulated clusters, and most transcripts belonged to the Bacteroidetes family Flavobacteriales. There were 5,876 transcripts assigned to this family in the metatranscriptome and a GO term enrichment analysis of these transcripts showed enrichment for some primary functions but also for proteolysis (corrected p-value =  $3.5e-6$ ) and nucleic acid phosphodiester bond hydrolysis (corrected p-value =  $3.4e-2$ ). It may indicate that under water stress these bacteria may be secreting siderophores which is a desirable PGPR characteristic.

To deeply explore the rhizosphere network of sugarcane subjected to water stress, some clusters - c315, c188, c421, c135, and c93 - that were found enriched for features of interest, such as the features in Figure 2, were individually analyzed.

Cluster c315, the cluster with highest percentage of isocitric acid correlated transcripts, has 74 nodes and 218 edges. The significantly over-represented biological process GO terms detected in cluster c315 and the transcripts correlated to isocitric acid are depicted in Figure 4. They were summarized using REVIGO into 10 major GO terms and among them cell motility is the most over-represented biological process in cluster c315.

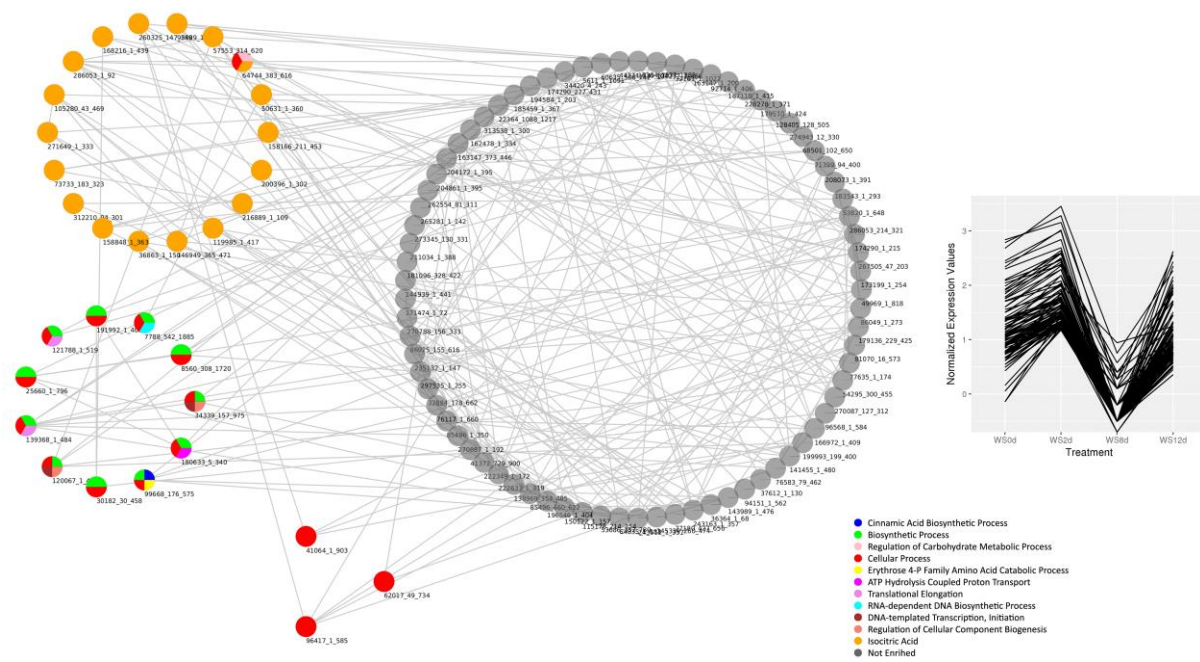
A total of 14 transcripts (19% of the cluster) were found to correlate with isocitric acid and one of them was also part of two of the biological process enriched in this cluster, bacterial-type flagellum organization and cell projection organization. Among the major enriched GO terms, there were five transcripts involved in translation and cellular amide metabolic process, three transcripts involved in cell motility and locomotion, two transcripts involved in bacterial-type flagellum organization, and one transcript involved in oxygen transport.



**Figure 4** Network topology of cluster c315. A grey circle represents a node (i.e. transcript). A grey edge connecting two nodes indicates the co-expression relationship between these two nodes.

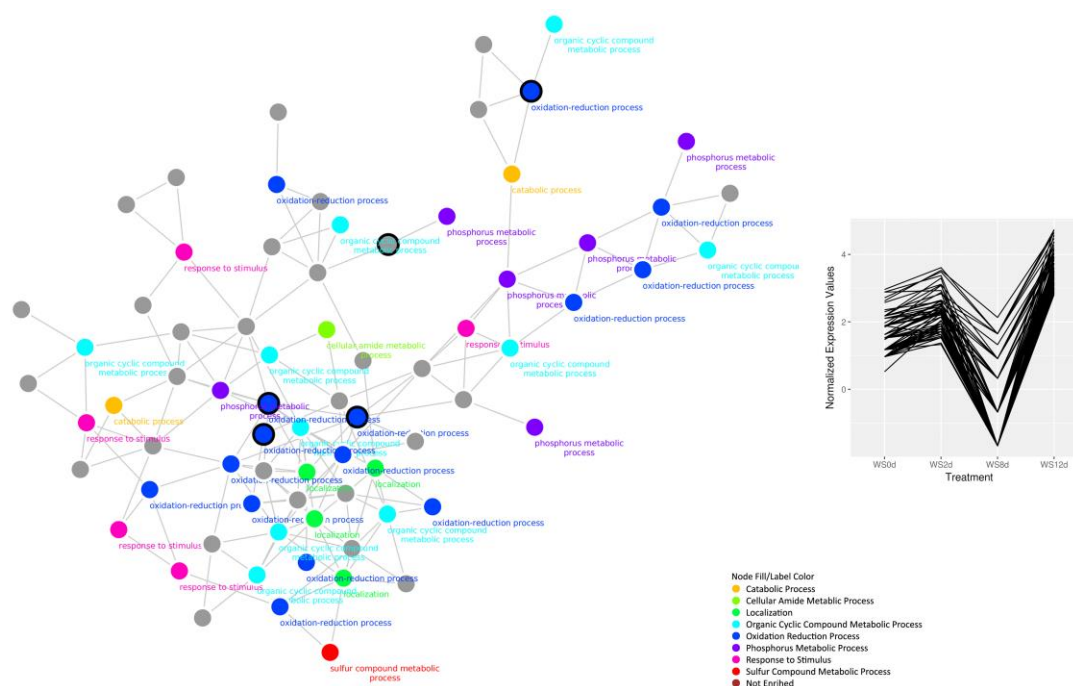
With 17% of its transcripts correlated to isocitric acid content in the rhizosphere, cluster c188 has 109 nodes and 246 edges. Transcripts correlated to isocitric acid and transcripts annotated as biological process GO terms that were significantly over-represented

in cluster c188 are shown in Figure 5. They were summarized using REVIGO into 10 major GO terms and among them cinnamic acid biosynthetic process is the most over-represented biological process in cluster c188. A total of 18 transcripts (17% of the cluster) were found to correlate with isocitric acid and one of them was also part of two of the biological process enriched in this cluster, regulation of carbohydrate metabolic process and cellular process. There were another 11 transcripts involved in biosynthetic process, two transcripts involved in translational elongation, two transcripts involved in DNA-templated transcription, initiation, one transcript involved in erythrose 4-phosphate/phosphoenolpyruvate family amino acid catabolic process, one transcript involved in ATP hydrolysis coupled proton transport, and one transcript involved in RNA-dependent DNA biosynthetic process.



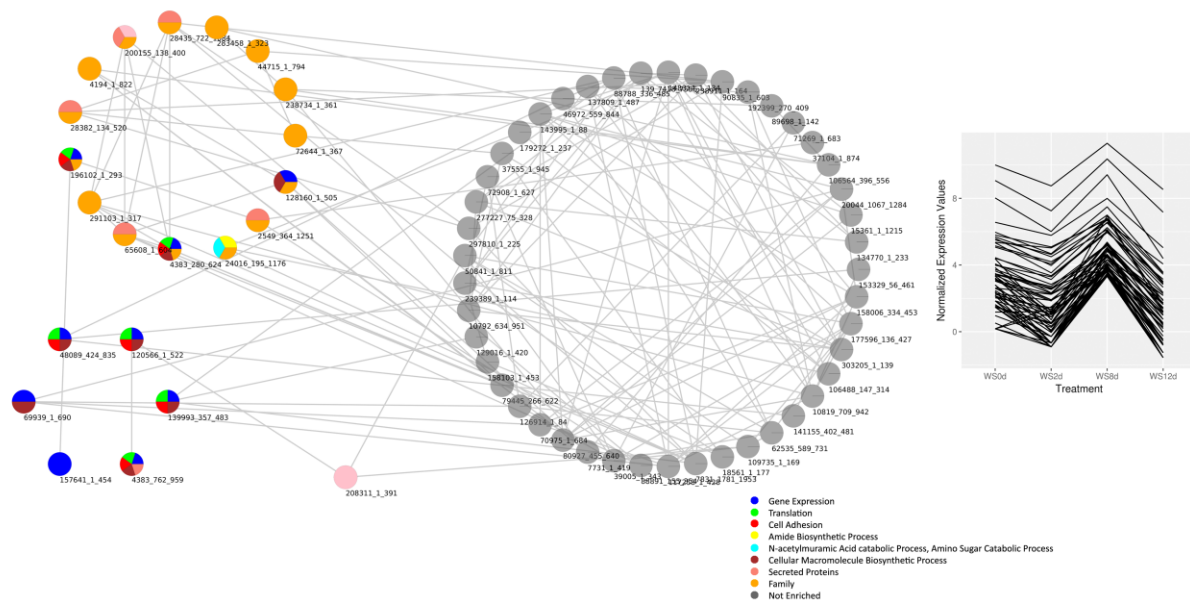
**Figure 5** Network topology of cluster c188. A grey circle represents a node (i.e. transcript). A grey edge connecting two nodes indicates the co-expression relationship between these two nodes.

Cluster c421 has 74 nodes, 161 edges, and five transcripts identified in our search for PGPR-like transcripts (7% of the cluster), three of them are peroxidases and two are superoxide dismutases, indicating a possible role in reactive oxygen species response. The GO terms annotated in cluster c421 and the putative PGPR-like transcripts are depicted in Figure 6. GO terms were summarized using REVIGO and a single representative biological process was assigned to each node. Although most transcripts in this cluster had GO terms assigned, no GO terms were enriched in this cluster ( $FDR < 0.1$ ). Four of the PGPR-like transcripts were assigned to the biological function oxidation-reduction process and one was assigned to the biological process organic cyclic compound metabolic process. In the whole cluster, there were 15 transcripts involved in oxidation-reduction process, ten transcripts involved in organic cyclic compound metabolic process, nine transcripts involved in response to stimulus, eight transcripts involved in phosphorus metabolic process, six transcripts involved in localization, five transcripts involved in catabolic process, and two transcripts involved in sulfur compound metabolic process.



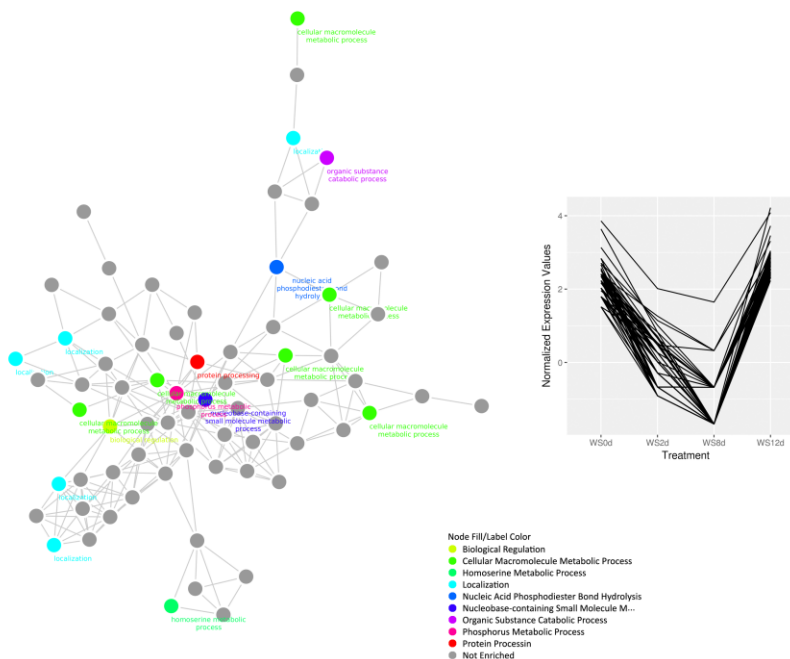
**Figure 6** Network topology of cluster c421. A grey circle represents a node (i.e. transcript). A grey edge connecting two nodes indicates the co-expression relationship between these two nodes.

Cluster c135 has 67 nodes, 146 edges, 15 transcripts assigned to Flavobacteriaceae family, and six transcripts coding for putative secreted proteins. The significantly over-represented biological process GO terms detected in cluster c135, the Flavobacteriaceae family transcripts, and transcripts coding for putative secreted proteins are shown in Figure 7. They were summarized using REVIGO into 7 major GO terms and among them gene expression is the most over-represented biological process in cluster c135. There were nine transcripts involved in gene expression, eight transcripts involved in cellular macromolecule biosynthetic process, six transcripts involved in translation, six transcripts involved in amide biosynthetic process, two transcripts involved in cell adhesion, one transcript involved in N-acetylmuramic acid catabolic process, and one transcript involved in amino sugar catabolic process.



**Figure 7** Network topology of cluster c135. A grey circle represents a node (i.e. transcript). A grey edge connecting two nodes indicates the co-expression relationship between these two nodes.

Cluster c93 has 70 nodes and 205 edges, 5 transcripts coding for putative secreted proteins, and 28 transcripts assigned to Planctomycetaceae family. Only one biological process GO term was found significantly over-represented, which is homoserine metabolic process. The transcripts with GO terms annotated in cluster c93, the transcripts coding for putative secreted proteins, and the Planctomycetaceae family assigned transcripts are depicted in Figure 8. There were six transcripts involved in cellular macromolecule metabolic process, five transcripts involved in localization, one transcript involved in organic substance catabolic process, one transcript involved in nucleic acid phosphodiester bond hydrolysis, one transcript involved in protein processing, one transcript involved in phosphorus metabolic process, and one transcript involved in nucleobase-containing small molecule metabolic process.



**Figure 8** Network topology of cluster c93. A grey circle represents a node (i.e. transcript). A grey edge connecting two nodes indicates the co-expression relationship between these two nodes.

## Conclusion

With the aim of determining the co-expression relationships of transcripts among microorganisms and functions present in the sugarcane rhizosphere metatranscriptome that could respond to the water stress treatments, in this chapter III, a co-expression network was generated from the differentially expressed transcripts. The network was composed by 582 clusters and the average size of clusters was 73 transcripts.

With the analysis of down-regulated clusters, it was possible to find several GO terms enriched such as translational elongation, gene expression, macromolecule biosynthetic process, regulation of cellular component biogenesis, primary metabolism, biosynthesis, and protein folding. Also, GO terms enriched in the up-regulated clusters were related to DNA



metabolic process, transposition, DNA modification, DNA recombination, macromolecule metabolism, self-proteolysis, nitrogen metabolism, and siderophore transport.

## References

- Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. <https://doi.org/citeulike-article-id:11583827>
- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., ... Sherlock, G. (2000). Gene ontology: Tool for the unification of biology. *Nature Genetics*. <https://doi.org/10.1038/75556>
- Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12(1), 59–60. <https://doi.org/10.1038/nmeth.3176>
- Dakora, F. D., Matiru, V. N., & Kanu, A. S. (2015). Rhizosphere ecology of lumichrome and riboflavin, two bacterial signal molecules eliciting developmental changes in plants. *Frontiers in Plant Science*, 6. <https://doi.org/10.3389/fpls.2015.00700>
- De Souza, R. S. C., Okura, V. K., Armanhi, J. S. L., Jorrín, B., Lozano, N., Da Silva, M. J., ... Arruda, P. (2016). Unlocking the bacterial and fungal communities assemblages of sugarcane microbiome. *Scientific Reports*, 6. <https://doi.org/10.1038/srep28774>
- Huson, D. H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., ... Tappu, R. (2016). MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLoS Computational Biology*, 12(6). <https://doi.org/10.1371/journal.pcbi.1004957>
- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14), 1754–1760. <https://doi.org/10.1093/bioinformatics/btp324>
- Liao, Y., Smyth, G. K., & Shi, W. (2014). FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7), 923–930. <https://doi.org/10.1093/bioinformatics/btt656>

- Maere, S., Heymans, K., & Kuiper, M. (2005). BiNGO: A Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. *Bioinformatics*, 21(16), 3448–3449. <https://doi.org/10.1093/bioinformatics/bti551>
- Masoabi, M., Lloyd, J., Kossmann, J., & van der Vyver, C. (2018). Ethyl Methanesulfonate Mutagenesis and In Vitro Polyethylene Glycol Selection for Drought Tolerance in Sugarcane (*Saccharum* spp.). *Sugar Tech*, 20(1), 50–59. <https://doi.org/10.1007/s12355-017-0524-8>
- Monreal, C. M., Zhang, J., Koziel, S., Vidmar, J., González, M., Matus, F., ... Etcheverria, P. (2018). Bacterial community structure associated with the addition of nitrogen and the dynamics of soluble carbon in the rhizosphere of canola (*Brassica napus*) grown in a Podzol. *Rhizosphere*, 5, 16–25. <https://doi.org/10.1016/j.rhisph.2017.11.004>
- Mutwil, M., Usadel, B., Schutte, M., Loraine, A., Ebenhoh, O., & Persson, S. (2010). Assembly of an Interactive Correlation Network for the Arabidopsis Genome Using a Novel Heuristic Clustering Algorithm. *PLANT PHYSIOLOGY*, 152(1), 29–43. <https://doi.org/10.1104/pp.109.145318>
- Patel, R. K., & Jain, M. (2012). NGS QC toolkit: A toolkit for quality control of next generation sequencing data. *PLoS ONE*, 7(2). <https://doi.org/10.1371/journal.pone.0030619>
- Peng, Y., Leung, H. C. M., Yiu, S. M., & Chin, F. Y. L. (2012). IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, 28(11), 1420–1428. <https://doi.org/10.1093/bioinformatics/bts174>
- Pisa, G., Magnani, G. S., Weber, H., Souza, E. M., Faoro, H., Monteiro, R. A., ... Cruz, L. M. (2011). Diversity of 16S rRNA genes from bacteria of sugarcane rhizosphere soil. *Brazilian Journal of Medical and Biological Research*, 44(12), 1215–1221. <https://doi.org/10.1590/S0100-879X2011007500148>

- Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2007). NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, 35(SUPPL. 1). <https://doi.org/10.1093/nar/gkl842>
- Rho, M., Tang, H., & Ye, Y. (2010). FragGeneScan: Predicting genes in short and error-prone reads. *Nucleic Acids Research*. <https://doi.org/10.1093/nar/gkq747>
- Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3). <https://doi.org/10.1186/gb-2010-11-3-r25>
- Saleem, M., Law, A. D., Sahib, M. R., Pervaiz, Z. H., & Zhang, Q. (2018). Impact of root system architecture on rhizosphere and root microbiome. *Rhizosphere*. <https://doi.org/10.1016/j.rhisph.2018.02.003>
- Schmieder, R., Lim, Y. W., & Edwards, R. (2012). Identification and removal of ribosomal RNA sequences from metatranscriptomes. *Bioinformatics*, 28(3), 433–435. <https://doi.org/10.1093/bioinformatics/btr669>
- Sellamuthu, K. M., & Govindaswamy, M. (2003). Effect of fertiliser and humic acid on rhizosphere microorganisms and soil enzymes at an early stage of sugarcane growth. *Sugar Tech*, 5(4), 273–277. <https://doi.org/10.1007/BF02942484>
- Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., ... Ideker, T. (2003). Cytoscape : A Software Environment for Integrated Models of Biomolecular Interaction Networks Cytoscape : A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, (Karp 2001), 2498–2504. <https://doi.org/10.1101/gr.1239303>
- Silva, M. D. A., Jifon, J. L., Da Silva, J. A. G., & Sharma, V. (2007). Use of physiological parameters as fast tools to screen for drought tolerance in sugarcane. *Brazilian Journal*

of *Plant Physiology*, 19(3), 193–201. <https://doi.org/10.1590/S1677-04202007000300003>

Supek, F., Bošnjak, M., Škunca, N., & Šmuc, T. (2011). Revigo summarizes and visualizes long lists of gene ontology terms. *PLoS ONE*, 6(7). <https://doi.org/10.1371/journal.pone.0021800>

Tarazona, S., García, F., Ferrer, A., Dopazo, J., & Conesa, A. (2012). NOIseq: a RNA-seq differential expression method robust for sequencing depth biases. *EMBnet.Journal*, 17(B), 18. <https://doi.org/10.14806/ej.17.B.265>

van Noort, V., Snel, B., & Huynen, M. A. (2004). The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *EMBO Reports*, 5(3), 280–284. <https://doi.org/10.1038/sj.embor.7400090>

Yeoh, Y. K., Paungfoo-Lonhienne, C., Dennis, P. G., Robinson, N., Ragan, M. A., Schmidt, S., & Hugenholtz, P. (2016). The core root microbiome of sugarcane cultivated under varying nitrogen fertilizer application. *Environmental Microbiology*, 18(5), 1338–1351. <https://doi.org/10.1111/1462-2920.12925>

Zandalinas, S. I., Mittler, R., Balfagón, D., Arbona, V., & Gómez-Cadenas, A. (2018). Plant adaptations to the combination of drought and high temperatures. *Physiologia Plantarum*. <https://doi.org/10.1111/ppl.12540>.

## CONCLUSÃO E PERSPECTIVAS

Através desse trabalho, foi possível estabelecer as melhores estratégias para cada uma das etapas de análise de um metatranscriptoma. Por exemplo, foi possível estabelecer que o melhor montador para os dados gerados foi o IDBA-UD, que o melhor software para predição ORFs foi o FragGeneScan ou que o melhor alinhamento das *reads* nos *contigs* foi conseguido através do BWA e do software featureCounts. Além disso, foi possível estabelecer um pipeline completo de análise de metatranscriptomas que pode ser aplicado em futuros estudos.

Com o pipeline desenvolvido, foi possível caracterizar taxonômica e funcionalmente a microbiota da rizosfera de cana-de-açúcar e descrever as mudanças causadas na rizosfera pelo estresse hídrico. Foram encontradas bactérias com aumento de abundância relativa após o tratamento de estresse hídrico, como por exemplo *Actinomadura*, *Dongia*, *Sphingomonas* e *Vulgatibacter*. Também, foram descobertas três categorias do segundo nível da hierarquia do banco de dados SEED com aumento de abundância após a aplicação do estresse hídrico, sendo elas relacionadas a degradação de proteínas para obtenção de energia (degradação de arginina e ornitina) e enzimas oxirredutases atuantes na cadeia respiratória aeróbia (succinato desidrogenase) e em processos fermentativos (cetoisovalerato oxirredutase).

Também foi realizada a análise da expressão gênica diferencial desencadeada pelo estresse hídrico ao qual a microbiota foi submetida, o que resultou em 43.716 transcritos diferencialmente expressos e demonstrou que foram *down*-regulados termos GO relacionados a regulação da respiração celular, biogênese de ribossomo, dobramento de proteína, endocitose, resposta ao estresse e tradução. Por outro lado, os principais termos GO enriquecidos nos transcritos *up*-regulados foram catabolismo de benzoato, montagem de flagelo e síntese de tRNA de pseudouridina.

Adicionalmente, foi gerada uma rede de co-expressão gênica a partir dos transcritos diferencialmente expressos aplicando-se uma correlação de Pearson e consequente utilização do algoritmo HCCA. Essa rede foi composta por 582 clusters e, através das análises da mesma, pode-se observar que com a aplicação do estresse hídrico, houve um aumento da expressão de transcritos relacionados, por exemplo a auto proteólise, degradação de ácidos nucleicos e transporte de sideróforos.

Através de todas as análises realizadas e apresentadas neste trabalho, foram encontrados vários microrganismos que responderam ao estresse hídrico, como os gêneros *Dongia*, *Actinomadura*, *Sphingomonas*, *Vulgatibacter*, *Sorangium*, *Devosia*, *Steroidobacter*, *Rhodanobacter* e *Aequorivita*. Com exceção dos gêneros *Actinomadura* que é uma Actinobactéria e *Aequorivita* que pertence ao filo Bacteroidetes, todos os outros gêneros aqui destacados são Proteobactérias, das classes Alfa, Gama e Delta. Dessa forma, pode-se observar que o filo Proteobactéria, além de ser o mais abundante na rizosfera de cana-de-açúcar, é também o filo com mais membros que responderam ao estresse hídrico.

Por meio das análises taxonômicas e funcionais, foi possível ter uma ideia de como provavelmente ocorre uma parte da interação entre a planta e os microrganismos da rizosfera. Foi verificado que a planta aumentou a liberação de certos exsudatos no ambiente rizosférico durante o estresse hídrico, como por exemplo os ácidos cítrico, isocítrico e málico, além disso, observou-se uma mudança na via metabólica do ciclo do ácido cítrico, indicando um aumento na expressão de transcritos relacionados ao metabolismo desses mesmos ácidos orgânicos e uma diminuição da expressão dos transcritos relacionados às outras partes da via. Assim sendo, isso pode indicar uma provável mudança no metabolismo energético desses microrganismos em resposta a exsudação da planta, de forma a utilizar essa energia que estava sendo disponibilizada pela mesma.

Finalmente, os resultados obtidos nesse trabalho auxiliam na compreensão da diversidade taxonômica e funcional da rizosfera de cana-de-açúcar e os efeitos causados pelo estresse hídrico ao qual a microbiota da rizosfera foi submetida. Desta forma, este trabalho pode servir de base para o desenvolvimento de aplicações e tecnologias a partir do conhecimento da dinâmica desse sistema biológico que foi aqui gerado, com vistas para a adaptação e aumento da produtividade da cana-de-açúcar frente às previstas mudanças climáticas globais.



## REFERÊNCIAS BIBLIOGRÁFICAS

- Alberts, Bruce, Alexander Johnson, Julian Lewis, David Morgan, Martin Raff, Keith Roberts, and Peter Walter. 2014. *Molecular Biology of the Cell 6e*. Garland Science. Vol. 6. [https://doi.org/10.1002/1521-3773\(20010316\)40:6<9823::AID-ANIE9823>3.3.CO;2-C](https://doi.org/10.1002/1521-3773(20010316)40:6<9823::AID-ANIE9823>3.3.CO;2-C).
- Alexa, A, and J Rahnenfuhrer. 2010. “TopGO: Enrichment Analysis for Gene Ontology.” *R Package Version*. <http://bioconductor.uib.no/2.7/bioc/html/topGO.html>.
- Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. 1990. “Basic Local Alignment Search Tool.” *Journal of Molecular Biology* 215 (3): 403–10. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2).
- Anders, S, D J McCarthy, Y S Chen, M Okoniewski, G K Smyth, W Huber, and M D Robinson. 2013. “Count-Based Differential Expression Analysis of RNA Sequencing Data Using R and Bioconductor.” *Nature Protocols* 8 (9): 1765–86. <https://doi.org/10.1038/nprot.2013.099>.
- Anders, Simon, and Wolfgang Huber. 2010. “Differential Expression Analysis for Sequence Count Data.” *Genome Biology* 11 (10). <https://doi.org/10.1186/gb-2010-11-10-r106>.
- Andrews, Simon. 2010. “FastQC: A Quality Control Tool for High Throughput Sequence Data.” <Http://Www.Bioinformatics.Babraham.Ac.Uk/Projects/Fastqc/>. 2010. <https://doi.org/citeulike-article-id:11583827>.
- Ara, Ismet, Atsuko Matsumoto, Mohammad Abdul Bakir, Takuji Kudo, Satoshi Omura, and Yoko Takahashi. 2008. “Actinomadura Maheshkhaliensis Sp. Nov., a Novel Actinomycete Isolated from Mangrove Rhizosphere Soil of Maheshkhali, Bangladesh.” *The Journal of General and Applied Microbiology* 54 (6): 335–42. <https://doi.org/10.2323/jgam.54.335>.
- Ashburner, Michael, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J.

- Michael Cherry, Allan P. Davis, et al. 2000. "Gene Ontology: Tool for the Unification of Biology." *Nature Genetics*. <https://doi.org/10.1038/75556>.
- Beneduzi, Anelise, Fernanda Moreira, Pedro B. Costa, Luciano K. Vargas, Bruno B. Lisboa, Rodrigo Favreto, José Ivo Baldani, and Luciane Maria P Passaglia. 2013. "Diversity and Plant Growth Promoting Evaluation Abilities of Bacteria Isolated from Sugarcane Cultivated in the South of Brazil." *Applied Soil Ecology* 63: 94–104. <https://doi.org/10.1016/j.apsoil.2012.08.010>.
- Berendsen, Roeland L., Corné M.J. Pieterse, and Peter A.H.M. Bakker. 2012. "The Rhizosphere Microbiome and Plant Health." *Trends in Plant Science*. <https://doi.org/10.1016/j.tplants.2012.04.001>.
- Bhosale, H. J., and Kadam, T. A. 2015. Generic diversity and a comparative account on plant growth promoting characteristics of actinomycetes in roots and rhizosphere of *Saccharum officinarum*. *Int J Curr Microbiol Appl Sci*, 4, 230-244.
- Birol, Inanç, Shaun D. Jackman, Cydney B. Nielsen, Jenny Q. Qian, Richard Varhol, Greg Stazyk, Ryan D. Morin, et al. 2009. "De Novo Transcriptome Assembly with ABySS." *Bioinformatics (Oxford, England)* 25 (21): 2872–77. <https://doi.org/10.1093/bioinformatics/btp367>.
- Bokulich, Nicholas A., Zachery T. Lewis, Kyria Boundy-Mills, and David A. Mills. 2016. "A New Perspective on Microbial Landscapes within Food Production." *Current Opinion in Biotechnology*. <https://doi.org/10.1016/j.copbio.2015.12.008>.
- Buchfink, Benjamin, Chao Xie, and Daniel H Huson. 2015. "Fast and Sensitive Protein Alignment Using DIAMOND." *Nature Methods* 12 (1): 59–60. <https://doi.org/10.1038/nmeth.3176>.
- Cantarel, Brandi I., Pedro M. Coutinho, Corinne Rancurel, Thomas Bernard, Vincent Lombard, and Bernard Henrissat. 2009. "The Carbohydrate-Active EnZymes

- Database (CAZy): An Expert Resource for Glycogenomics.” *Nucleic Acids Research* 37 (SUPPL. 1). <https://doi.org/10.1093/nar/gkn663>.
- Carvalhais, Lilia C., Paul G. Dennis, Gene W. Tyson, and Peer M. Schenk. 2013. “Rhizosphere Metatranscriptomics: Challenges and Opportunities.” In *Molecular Microbial Ecology of the Rhizosphere*, 2:1137–44. <https://doi.org/10.1002/9781118297674.ch109>.
- Caspi, R., T. Altman, R. Billington, and K. Dreher. 2014. “The MetaCyc Database of Metabolic Pathways and Enzymes and the BioCyc Collection of Pathway/Genome Databases -- Caspi et Al. 36 (Supplement 1): D623 -- *Nucleic Acids Research*.” *Nucleic Acids Research* 36 (Database): D623–31. <https://doi.org/10.1093/nar/gkm900>.
- CAZypedia Consortium. 2018. “Ten Years of CAZypedia: A Living Encyclopedia of Carbohydrate-Active Enzymes.” *Glycobiology* 28 (1): 3–8. <https://doi.org/10.1093/glycob/cwx089>.
- Celaj, Albi, Janet Markle, Jayne Danska, and John Parkinson. 2014. “Comparison of Assembly Algorithms for Improving Rate of Metatranscriptomic Functional Annotation.” *Microbiome* 2 (1). <https://doi.org/10.1186/2049-2618-2-39>.
- Chang, Zheng, Guojun Li, Juntao Liu, Yu Zhang, Cody Ashby, Deli Liu, Carole L. Cramer, and Xiuzhen Huang. 2015. “Bridger: A New Framework for de Novo Transcriptome Assembly Using RNA-Seq Data.” *Genome Biology* 16 (1). <https://doi.org/10.1186/s13059-015-0596-2>.
- Chapelle, Emilie, Rodrigo Mendes, Peter A.Hm Bakker, and Jos M. Raaijmakers. 2016. “Fungal Invasion of the Rhizosphere Microbiome.” *ISME Journal* 10 (1): 265–68. <https://doi.org/10.1038/ismej.2015.82>.
- Costa, Diogo P. da, Armando C.F. Dias, Simone R. Cotta, Daniella Vilela, Pedro A.M. de Andrade, Vivian H. Pellizari, and Fernando Dini Andreote. 2018. “Changes of

- Bacterial Communities in the Rhizosphere of Sugarcane under Elevated Concentration of Atmospheric CO<sub>2</sub>.” *GCB Bioenergy* 10 (2): 137–45. <https://doi.org/10.1111/gcbb.12476>.
- Cristina, Izabel, Bruno Ferreira, and Marcos Omir. 2008. “A Qualidade Da Cana-de-Açúcar Como Matéria-Prima Para Produção de Álcool.” *Nucleus*, 82–92. <https://doi.org/10.3738/1982.2278.93>.
- Dagnaw, Feredegn, Fassil Assefa, Heluf Gebrekidan, and Anteneh Argaw. 2015. “Characterization of Plant Growth Promoting Bacteria from Sugarcane (*Saccharum Officinarum* L) Rhizosphere of Wonji-Shoa Sugar Estate and Farmers Landraces of Ethiopia.” *Biotechnology* 14 (2): 58–64. <https://doi.org/10.3923/biotech.2015.58.64>.
- Dakora, Felix D., Vivienne N. Matiru, and Alfred S. Kanu. 2015. “Rhizosphere Ecology of Lumichrome and Riboflavin, Two Bacterial Signal Molecules Eliciting Developmental Changes in Plants.” *Frontiers in Plant Science* 6. <https://doi.org/10.3389/fpls.2015.00700>.
- Dam, Nicole M. van, and Harro J. Bouwmeester. 2016. “Metabolomics in the Rhizosphere: Tapping into Belowground Chemical Communication.” *Trends in Plant Science*. <https://doi.org/10.1016/j.tplants.2016.01.008>.
- Daniels, J, and B Roach. 1987. “Taxonomy and Evolution.” In *Sugarcane Improvement through Breeding*, 7–81. <https://doi.org/10.1016/b978-0-444-42769-4.50007-2> LB - Daniels1987.
- DellaPenna, Dean. 2007. “Biofortification of Plant-Based Food: Enhancing Folate Levels by Metabolic Engineering.” *Proceedings of the National Academy of Sciences of the United States of America* 104 (10): 3675–76. <https://doi.org/10.1073/pnas.0700640104>.
- Dillies, Marie Agnès, Andrea Rau, Julie Aubert, Christelle Hennequet-Antier, Marine

- Jeanmougin, Nicolas Servant, Céline Keime, et al. 2013. “A Comprehensive Evaluation of Normalization Methods for Illumina High-Throughput RNA Sequencing Data Analysis.” *Briefings in Bioinformatics* 14 (6): 671–83. <https://doi.org/10.1093/bib/bbs046>.
- Dong, Meng, Zongtao Yang, Guangyuan Cheng, Lei Peng, Qian Xu, and Jingsheng Xu. 2018. “Diversity of the Bacterial Microbiome in the Roots of Four *Saccharum* Species: *S. Spontaneum*, *S. Robustum*, *S. Barberi*, and *S. Officinarum*.” *Frontiers in Microbiology* 9 (FEB). <https://doi.org/10.3389/fmicb.2018.00267>.
- Duran-Pinedo, Ana E., Tsute Chen, Ricardo Teles, Jacqueline R. Starr, Xiaoshan Wang, Keerthana Krishnan, and Jorge Frias-Lopez. 2014. “Community-Wide Transcriptome of the Oral Microbiome in Subjects with and without Periodontitis.” *ISME Journal* 8 (8): 1659–72. <https://doi.org/10.1038/ismej.2014.23>.
- Fernandes, Noemi Mendes, Roberto Júnio Pedroso Dias, Marcus Vinicius Xavier Senra, Carlos Augusto Gomes Soares, and Inácio Domingos da Silva Neto. 2013. “Morphology and 18S rDNA Gene Sequence of *Blepharisma Sinuosum* Sawaya, 1940 (Ciliophora: Heterotrichea) from Brazil.” *European Journal of Protistology* 49 (4): 623–33. <https://doi.org/10.1016/j.ejop.2013.04.003>.
- Ficklin, S. P., and F. A. Feltus. 2011. “Gene Coexpression Network Alignment and Conservation of Gene Modules between Two Grass Species: Maize and Rice.” *PLANT PHYSIOLOGY* 156 (3): 1244–56. <https://doi.org/10.1104/pp.111.173047>.
- Filippis, Francesca De, Alessandro Genovese, Pasquale Ferranti, Jack A. Gilbert, and Danilo Ercolini. 2016. “Metatranscriptomics Reveals Temperature-Driven Functional Changes in Microbiome Impacting Cheese Maturation Rate.” *Scientific Reports* 6. <https://doi.org/10.1038/srep21871>.
- Forster, Samuel C., Hilary P. Browne, Nitin Kumar, Martin Hunt, Hubert Denise, Alex

- Mitchell, Robert D. Finn, and Trevor D. Lawley. 2016. "HPMCD: The Database of Human Microbial Communities from Metagenomic Datasets and Microbial Reference Genomes." *Nucleic Acids Research* 44 (D1): D604–9. <https://doi.org/10.1093/nar/gkv1216>.
- Goldemberg, José. 2008. "The Brazilian Biofuels Industry." *Biotechnology for Biofuels*. <https://doi.org/10.1186/1754-6834-1-6>.
- Gordon, A, and G J Hannon. 2010. "Fastx-Toolkit. FASTQ/A Short-Reads Pre-Processing Tools." *Unpublished Http://Hannonlab. Cshl. Edu/Fastx\_Toolkit*.
- Gouvêa, Júlia Ribeiro Ferreira, Paulo Cesar Sentelhas, Samuel Thomazella Gazzola, and Marcelo Cabral Santos. 2009. "Climate Changes and Technological Advances: Impacts on Sugarcane Productivity in Tropical Southern Brazil." *Scientia Agricola* 66 (October): 593–605. <https://doi.org/10.1590/S0103-90162009000500003>.
- Grabherr, Manfred G., Brian J. Haas, Moran Yassour, Joshua Z. Levin, Dawn A. Thompson, Ido Amit, Xian Adiconis, et al. 2011. "Full-Length Transcriptome Assembly from RNA-Seq Data without a Reference Genome." *Nature Biotechnology* 29 (7): 644–52. <https://doi.org/10.1038/nbt.1883>.
- Green, Stefan J., Om Prakash, Puja Jasrotia, Will A. Overholt, Erick Cardenas, Daniela Hubbard, James M. Tiedje, et al. 2012. "Denitrifying Bacteria from the Genus *Rhodanobacter* Dominate Bacterial Communities in the Highly Contaminated Subsurface of a Nuclear Legacy Waste Site." *Applied and Environmental Microbiology* 78 (4): 1039–47. <https://doi.org/10.1128/AEM.06435-11>.
- Haasum, Y., K. Ström, R. Wehelie, V. Luna, M. C. Roberts, J. P. Maskell, L. M.C. Hall, and G. Swedberg. 2001. "Amino Acid Repetitions in the Dihydropteroate Synthase of *Streptococcus Pneumoniae* Lead to Sulfonamide Resistance with Limited Effects on Substrate Km." *Antimicrobial Agents and Chemotherapy* 45 (3): 805–9.

<https://doi.org/10.1128/AAC.45.3.805-809.2001>.

- Hayden, Helen L., Keith W. Savin, Jenny Wadeson, Vadakattu V.S.R. Gupta, and Pauline M. Mele. 2018. "Comparative Metatranscriptomics of Wheat Rhizosphere Microbiomes in Disease Suppressive and Non-Suppressive Soils for *Rhizoctonia Solani* AG8." *Frontiers in Microbiology* 9 (MAY). <https://doi.org/10.3389/fmicb.2018.00859>.
- He, Shaomei, Omri Wurtzel, Kanwar Singh, Jeff L. Froula, Suzan Yilmaz, Susannah G. Tringe, Zhong Wang, et al. 2010. "Validation of Two Ribosomal RNA Removal Methods for Microbial Metatranscriptomics." *Nature Methods* 7 (10): 807–12. <https://doi.org/10.1038/nmeth.1507>.
- Hirao, Ichiro. 2006. "Unnatural Base Pair Systems for DNA/RNA-Based Biotechnology." *Current Opinion in Chemical Biology*. <https://doi.org/10.1016/j.cbpa.2006.09.021>.
- Huang, Xiaoqiu, and Anup Madan. 1999. "CAP3: A DNA Sequence Assembly Program." *Genome Research* 9 (9): 868–77. <https://doi.org/10.1101/gr.9.9.868>.
- Huang, Xing-Feng, Jacqueline M. Chaparro, Kenneth F. Reardon, Ruifu Zhang, Qirong Shen, and Jorge M. Vivanco. 2014. "Rhizosphere Interactions: Root Exudates, Microbes, and Microbial Communities<sup>1</sup>." *Botany* 92 (4): 267–75. <https://doi.org/10.1139/cjb-2013-0225>.
- Huson, Daniel H., Sina Beier, Isabell Flade, Anna Górska, Mohamed El-Hadidi, Suparna Mitra, Hans Joachim Ruscheweyh, and Rewati Tappu. 2016. "MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data." *PLoS Computational Biology* 12 (6). <https://doi.org/10.1371/journal.pcbi.1004957>.
- Hyatt, Doug, Gwo Liang Chen, Philip F. LoCascio, Miriam L. Land, Frank W. Larimer, and Loren J. Hauser. 2010. "Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification." *BMC Bioinformatics* 11. <https://doi.org/10.1186/1471->

2105-11-119.

IPCC. 2014. “IPCC Fifth Assessment Synthesis Report-Climate Change 2014 Synthesis Report.” *IPCC Fifth Assessment Synthesis Report-Climate Change 2014 Synthesis Report*, pages: 167.

Jia, Baofeng, Amogelang R. Raphenya, Brian Alcock, Nicholas Waglechner, Peiyao Guo, Kara K. Tsang, Briony A. Lago, et al. 2017. “CARD 2017: Expansion and Model-Centric Curation of the Comprehensive Antibiotic Resistance Database.” *Nucleic Acids Research* 45 (D1): D566–73. <https://doi.org/10.1093/nar/gkw1004>.

Kalam, Sadaf, Subha Narayan Das, Anirban Basu, and Appa Rao Podile. 2017. “Population Densities of Indigenous Acidobacteria Change in the Presence of Plant Growth Promoting Rhizobacteria (PGPR) in Rhizosphere.” *Journal of Basic Microbiology* 57 (5): 376–85. <https://doi.org/10.1002/jobm.201600588>.

Kelley, David R., Michael C. Schatz, and Steven L. Salzberg. 2010. “Quake: Quality-Aware Detection and Correction of Sequencing Errors.” *Genome Biology* 11 (11). <https://doi.org/10.1186/gb-2010-11-11-r116>.

Landell, M. D. A., and Bressiani, J. A. 2008. Melhoramento genético, caracterização e manejo varietal. In: Dinardo-Miranda, L. L., Vasconcelos, A. C. M., Landell, M. G. A. Cana-de-açúcar. Campinas, IAC, 101-155.

Langmead, Ben, and Steven L. Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2.” *Nature Methods* 9 (4): 357–59. <https://doi.org/10.1038/nmeth.1923>.

Lee, Chang Soo, Kwang Kyu Kim, Zubair Aslam, and Sung Taik Lee. 2007. “Rhodanobacter Thiooxydans Sp. Nov., Isolated from a Biofilm on Sulfur Particles Used in an Autotrophic Denitrification Process.” *International Journal of Systematic and Evolutionary Microbiology* 57 (8): 1775–79. <https://doi.org/10.1099/ijs.0.65086-0>.

León-Sánchez, Lupe, Emilio Nicolás, Marta Goberna, Iván Prieto, Fernando T. Maestre, and



- José Ignacio Querejeta. 2018. “Poor Plant Performance under Simulated Climate Change Is Linked to Mycorrhizal Responses in a Semi-Arid Shrubland.” *Journal of Ecology* 106 (3): 960–76. <https://doi.org/10.1111/1365-2745.12888>.
- Leung, Henry C.M., Siu-Ming Yiu, John Parkinson, and Francis Y.L. Chin. 2013. “IDBA-MT: De Novo Assembler for Metatranscriptomic Data Generated from Next-Generation Sequencing Technology.” *Journal of Computational Biology* 20 (7): 540–50. <https://doi.org/10.1089/cmb.2013.0042>.
- Li, Heng, and Richard Durbin. 2009. “Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform.” *Bioinformatics* 25 (14): 1754–60. <https://doi.org/10.1093/bioinformatics/btp324>.
- Li, Weizhong, and Adam Godzik. 2006. “Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences.” *Bioinformatics* 22 (13): 1658–59. <https://doi.org/10.1093/bioinformatics/btl158>.
- Liao, Yang, Gordon K. Smyth, and Wei Shi. 2014. “FeatureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features.” *Bioinformatics* 30 (7): 923–30. <https://doi.org/10.1093/bioinformatics/btt656>.
- Liu, Pingwu, Amaresh Chandra, Youxiong Que, Ping Hua Chen, Michael P. Grisham, William H. White, Caleb D. Dalley, Thomas L. Tew, and Yong Bao Pan. 2016. “Identification of Quantitative Trait Loci Controlling Sucrose Content Based on an Enriched Genetic Linkage Map of Sugarcane (*Saccharum* Spp. Hybrids) Cultivar ‘LCP 85-384.’” *Euphytica* 207 (3): 527–49. <https://doi.org/10.1007/s10681-015-1538-5>.
- Lombard, Vincent, Hemalatha Golaconda Ramulu, Elodie Drula, Pedro M. Coutinho, and Bernard Henrissat. 2014. “The Carbohydrate-Active Enzymes Database (CAZy) in 2013.” *Nucleic Acids Research* 42 (D1). <https://doi.org/10.1093/nar/gkt1178>.

- Lubsanova, D. A., G. M. Zenova, P. A. Kozhevin, N. A. Manucharova, and A. P. Shvarov. 2014. "Filamentous Actinobacteria of the Saline Soils of Arid Territories." *Moscow University Soil Science Bulletin* 69 (2): 88–92. <https://doi.org/10.3103/S0147687414020057>.
- Luo, Fei, Roya Gitiafroz, Cheryl E. Devine, Yunchen Gong, Laura A. Hug, Lutgarde Raskin, and Elizabeth A. Edwards. 2014. "Metatranscriptome of an Anaerobic Benzene-Degrading, Nitrate-Reducing Enrichment Culture Reveals Involvement of Carboxylation in Benzene Ring Activation." *Applied and Environmental Microbiology* 80 (14): 4095–4107. <https://doi.org/10.1128/AEM.00717-14>.
- Luo, Ruibang, Binghang Liu, Yinlong Xie, Zhenyu Li, Weihua Huang, Jianying Yuan, Guangzhu He, et al. 2012. "SOAPdenovo2: An Empirically Improved Memory-Efficient Short-Read *de Novo* Assembler." *GigaScience* 1: 18. <https://doi.org/10.1186/2047-217X-1-18>.
- Maere, Steven, Karel Heymans, and Martin Kuiper. 2005. "BiNGO: A Cytoscape Plugin to Assess Overrepresentation of Gene Ontology Categories in Biological Networks." *Bioinformatics* 21 (16): 3448–49. <https://doi.org/10.1093/bioinformatics/bti551>.
- Marzano, Shin Yi Lee, and Leslie L. Domier. 2016. "Novel Mycoviruses Discovered from Metatranscriptomics Survey of Soybean Phyllosphere Phytobiomes." *Virus Research* 213: 332–42. <https://doi.org/10.1016/j.virusres.2015.11.002>.
- Masoabi, M., J. Lloyd, J. Kossmann, and C. van der Vyver. 2018. "Ethyl Methanesulfonate Mutagenesis and In Vitro Polyethylene Glycol Selection for Drought Tolerance in Sugarcane (*Saccharum* Spp.)." *Sugar Tech* 20 (1): 50–59. <https://doi.org/10.1007/s12355-017-0524-8>.
- McArthur, Andrew G., Nicholas Wagglechner, Fazmin Nizam, Austin Yan, Marisa A. Azad, Alison J. Baylay, Kirandeep Bhullar, et al. 2013. "The Comprehensive Antibiotic

- Resistance Database.” *Antimicrobial Agents and Chemotherapy* 57 (7): 3348–57.  
<https://doi.org/10.1128/AAC.00419-13>.
- Miller, Christopher S., Brett J. Baker, Brian C. Thomas, Steven W. Singer, and Jillian F. Banfield. 2011. “EMIRGE: Reconstruction of Full-Length Ribosomal Genes from Microbial Community Short Read Sequencing Data.” *Genome Biology* 12 (5).  
<https://doi.org/10.1186/gb-2011-12-5-r44>.
- Mitra, Suparna, Mario Stärk, and Daniel H. Huson. 2011. “Analysis of 16S RRNA Environmental Sequences Using MEGAN.” In *10th Int. Conference on Bioinformatics - 1st ISCB Asia Joint Conference 2011, InCoB 2011/ISCB-Asia 2011: Computational Biology - Proceedings from Asia Pacific Bioinformatics Network (APBioNet)*. Vol. 12. <https://doi.org/10.1186/1471-2164-12-S3-S17>.
- Miura, Toshiko, Ainin Niswati, I. G. Swibawa, Sri Haryani, Heru Gunito, Satoshi Shimano, Koichi Fujie, and Nobuhiro Kaneko. 2015. “Diversity of Fungi on Decomposing Leaf Litter in a Sugarcane Plantation and Their Response to Tillage Practice and Bagasse Mulching: Implications for Management Effects on Litter Decomposition.” *Microbial Ecology* 70 (3): 646–58. <https://doi.org/10.1007/s00248-015-0620-9>.
- Monreal, C. M., J. Zhang, S. Koziel, J. Vidmar, M. González, F. Matus, S. Baxi, S. Wu, M. DeRosa, and P. Etcheverria. 2018. “Bacterial Community Structure Associated with the Addition of Nitrogen and the Dynamics of Soluble Carbon in the Rhizosphere of Canola (*Brassica Napus*) Grown in a Podzol.” *Rhizosphere* 5: 16–25.  
<https://doi.org/10.1016/j.rhisph.2017.11.004>.
- Mukherjee, Sunil Kumar. 1957. “Origin and Distribution of *Saccharum*.” *Botanical Gazette* 119 (1): 55–61. <https://doi.org/10.1086/335962>.
- Mutwil, M., B. Usadel, M. Schutte, A. Loraine, O. Ebenhoh, and S. Persson. 2010. “Assembly of an Interactive Correlation Network for the *Arabidopsis* Genome Using

- a Novel Heuristic Clustering Algorithm.” *PLANT PHYSIOLOGY* 152 (1): 29–43.  
<https://doi.org/10.1104/pp.109.145318>.
- Nakasugi, Kenlee, Ross Crowhurst, Julia Bally, and Peter Waterhouse. 2014. “Combining Transcriptome Assemblies from Multiple de Novoassemblers in the Allo-Tetraploid Plant *Nicotiana Benthamiana*.” *PLoS ONE* 9 (3).  
<https://doi.org/10.1371/journal.pone.0091776>.
- Newman, Molli M., Nicola Lorenz, Nigel Hoilett, Nathan R. Lee, Richard P. Dick, Mark R. Liles, Cliff Ramsier, and Joseph W. Klopper. 2016. “Changes in Rhizosphere Bacterial Gene Expression Following Glyphosate Treatment.” *Science of the Total Environment* 553: 32–41. <https://doi.org/10.1016/j.scitotenv.2016.02.078>.
- Nookaew, Intawat, Marta Papini, Natapol Pornputtpong, Gionata Scalcinati, Linn Fagerberg, Matthias Uhlén, and Jens Nielsen. 2012. “A Comprehensive Comparison of RNA-Seq-Based Transcriptome Analysis from Reads to Differential Gene Expression and Cross-Comparison with Microarrays: A Case Study in *Saccharomyces Cerevisiae*.” *Nucleic Acids Research* 40 (20): 10084–97.  
<https://doi.org/10.1093/nar/gks804>.
- Noort, Vera van, Berend Snel, and Martijn A. Huynen. 2004. “The Yeast Coexpression Network Has a Small-World, Scale-Free Architecture and Can Be Explained by a Simple Model.” *EMBO Reports* 5 (3): 280–84.  
<https://doi.org/10.1038/sj.embor.7400090>.
- Norovsuren, Zh., G. M. Zenova, and L. V. Mosina. 2007. “Actinomycetes in the Rhizosphere of Semidesert Soils of Mongolia.” *Eurasian Soil Science* 40 (4): 415–18.  
<https://doi.org/10.1134/S1064229307040072>.
- Overbeek, Ross, Tadhg Begley, Ralph M. Butler, Jomuna V. Choudhuri, Han Yu Chuang, Matthew Cohoon, Valérie de Crécy-Lagard, et al. 2005. “The Subsystems Approach

- to Genome Annotation and Its Use in the Project to Annotate 1000 Genomes.” *Nucleic Acids Research* 33 (17): 5691–5702. <https://doi.org/10.1093/nar/gki866>.
- Parks, Donovan H., Gene W. Tyson, Philip Hugenholtz, and Robert G. Beiko. 2014. “STAMP: Statistical Analysis of Taxonomic and Functional Profiles.” *Bioinformatics* 30 (21): 3123–24. <https://doi.org/10.1093/bioinformatics/btu494>.
- Patel, Ravi K., and Mukesh Jain. 2012. “NGS QC Toolkit: A Toolkit for Quality Control of next Generation Sequencing Data.” *PLoS ONE* 7 (2). <https://doi.org/10.1371/journal.pone.0030619>.
- Patro, Rob, Geet Duggal, and Carl Kingsford. 2015. “Accurate, Fast, and Model-Aware Transcript Expression Quantification with Salmon.” *BioRxiv*, 021592. <https://doi.org/10.1101/021592>.
- Paungfoo-Lonhienne, Chanyarat, Yun Kit Yeoh, Naga Rup Pinaki Kasinadhuni, Thierry G.A. Lonhienne, Nicole Robinson, Philip Hugenholtz, Mark A. Ragan, and Susanne Schmidt. 2015. “Nitrogen Fertilizer Dose Alters Fungal Communities in Sugarcane Soil and Rhizosphere.” *Scientific Reports* 5. <https://doi.org/10.1038/srep08678>.
- Peiffer, J. A., A. Spor, O. Koren, Z. Jin, S. G. Tringe, J. L. Dangl, E. S. Buckler, and R. E. Ley. 2013. “Diversity and Heritability of the Maize Rhizosphere Microbiome under Field Conditions.” *Proceedings of the National Academy of Sciences* 110 (16): 6548–53. <https://doi.org/10.1073/pnas.1302837110>.
- Peng, Yu, Henry C M Leung, S. M. Yiu, and Francis Y L Chin. 2012. “IDBA-UD: A de Novo Assembler for Single-Cell and Metagenomic Sequencing Data with Highly Uneven Depth.” *Bioinformatics* 28 (11): 1420–28. <https://doi.org/10.1093/bioinformatics/bts174>.
- Peng, Yu, Henry C M Leung, Siu Ming Yiu, Ming Ju Lv, Xin Guang Zhu, and Francis Y L Chin. 2013. “IDBA-Tran: A More Robust de Novo de Bruijn Graph Assembler for

- Transcriptomes with Uneven Expression Levels.” In *Bioinformatics*. Vol. 29.  
<https://doi.org/10.1093/bioinformatics/btt219>.
- Pereira, Letícia Bianca. 2018. “Analysis of the metatranscriptome of sugarcane rhizosphere subjected to water stress.” Manuscript in preparation.
- Pevzner, P. A., H. Tang, and M. S. Waterman. 2001. “An Eulerian Path Approach to DNA Fragment Assembly.” *Proceedings of the National Academy of Sciences* 98 (17): 9748–53. <https://doi.org/10.1073/pnas.171285098>.
- Pisa, G., G. S. Magnani, H. Weber, E. M. Souza, H. Faoro, R. A. Monteiro, E. Daros, et al. 2011. “Diversity of 16S rRNA Genes from Bacteria of Sugarcane Rhizosphere Soil.” *Brazilian Journal of Medical and Biological Research* 44 (12): 1215–21. <https://doi.org/10.1590/S0100-879X2011007500148>.
- Pop, Mihai. 2009. “Genome Assembly Reborn: Recent Computational Challenges.” *Briefings in Bioinformatics* 10 (4): 354–66. <https://doi.org/10.1093/bib/bbp026>.
- Prakash, Om, Stefan J. Green, Puja Jasrotia, Will A. Overholt, Andy Canion, David B. Watson, Scott C. Brooks, and Joel E. Kostka. 2012. “*Rhodanobacter Denitrificans* Sp. Nov., Isolated from Nitrate-Rich Zones of a Contaminated Aquifer.” *International Journal of Systematic and Evolutionary Microbiology* 62 (10): 2457–62. <https://doi.org/10.1099/ijs.0.035840-0>.
- Pruesse, Elmar, Christian Quast, Katrin Knittel, Bernhard M. Fuchs, Wolfgang Ludwig, Jörg Peplies, and Frank Oliver Glöckner. 2007. “SILVA: A Comprehensive Online Resource for Quality Checked and Aligned Ribosomal RNA Sequence Data Compatible with ARB.” *Nucleic Acids Research* 35 (21): 7188–96. <https://doi.org/10.1093/nar/gkm864>.
- Pruitt, Kim D., Tatiana Tatusova, and Donna R. Maglott. 2007. “NCBI Reference Sequences (RefSeq): A Curated Non-Redundant Sequence Database of Genomes, Transcripts

- and Proteins.” *Nucleic Acids Research* 35 (SUPPL. 1).  
<https://doi.org/10.1093/nar/gkl842>.
- Rana, Satshil B., Frank J. Zadlock, Ziping Zhang, Wyatt R. Murphy, and Carolyn S. Bentivegna. 2016. “Comparison of de Novo Transcriptome Assemblers and K-Mer Strategies Using the Killifish, *Fundulus Heteroclitus*.” *PLoS ONE* 11 (4).  
<https://doi.org/10.1371/journal.pone.0153104>.
- Rho, Mina, Haixu Tang, and Yuzhen Ye. 2010. “FragGeneScan: Predicting genes in short and error-prone reads.” *Nucleic Acids Research*.  
<https://doi.org/10.1093/nar/gkq747>.
- Rincon-Florez, Vivian A., Lilia C. Carvalhais, and Peer M. Schenk. 2013. “Culture-Independent Molecular Tools for Soil and Rhizosphere Microbiology.” *Diversity*.  
<https://doi.org/10.3390/d5030581>.
- Rivas, Raul, Encarna Velázquez, Anne Willems, Nieves Vizcaíno, Nanjappa S. Subba-Rao, Pedro F. Mateos, Monique Gillis, Frank B. Dazzo, and Eustoquio Martínez-Molina. 2002. “A New Species of *Devosia* That Forms a Unique Nitrogen-Fixing Root-Nodule Symbiosis with the Aquatic Legume *Neptunia Natans* (L.f.) Druce.” *Applied and Environmental Microbiology* 68 (11): 5217–22.  
<https://doi.org/10.1128/AEM.68.11.5217-5222.2002>.
- Robinson, Mark D., and Alicia Oshlack. 2010. “A Scaling Normalization Method for Differential Expression Analysis of RNA-Seq Data.” *Genome Biology* 11 (3).  
<https://doi.org/10.1186/gb-2010-11-3-r25>.
- Saleem, Muhammad, Audrey D. Law, Mohammad Radhi Sahib, Zahida H. Pervaiz, and Qingming Zhang. 2018. “Impact of Root System Architecture on Rhizosphere and Root Microbiome.” *Rhizosphere*. <https://doi.org/10.1016/j.rhisph.2018.02.003>.
- Schmieder, Robert, Yan Wei Lim, and Robert Edwards. 2012. “Identification and Removal

- of Ribosomal RNA Sequences from Metatranscriptomes.” *Bioinformatics* 28 (3): 433–35. <https://doi.org/10.1093/bioinformatics/btr669>.
- Schulz, Marcel H., Daniel R. Zerbino, Martin Vingron, and Ewan Birney. 2012. “Oases: Robust de Novo RNA-Seq Assembly across the Dynamic Range of Expression Levels.” *Bioinformatics* 28 (8): 1086–92. <https://doi.org/10.1093/bioinformatics/bts094>.
- Sellamuthu, K. M., and M. Govindaswamy. 2003. “Effect of Fertiliser and Humic Acid on Rhizosphere Microorganisms and Soil Enzymes at an Early Stage of Sugarcane Growth.” *Sugar Tech* 5 (4): 273–77. <https://doi.org/10.1007/BF02942484>.
- Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. 2003. “Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks.” *Genome Research*, no. Karp 2001: 2498–2504. <https://doi.org/10.1101/gr.1239303>.
- Shrestha, Pravin Malla, Michael Kube, Richard Reinhardt, and Werner Liesack. 2009. “Transcriptional Activity of Paddy Soil Bacterial Communities.” *Environmental Microbiology* 11 (4): 960–70. <https://doi.org/10.1111/j.1462-2920.2008.01821.x>.
- Silva, Genivaldo Gueiros Z., Kevin T. Green, Bas E. Dutilh, and Robert A. Edwards. 2015. “SUPER-FOCUS: A Tool for Agile Functional Analysis of Shotgun Metagenomic Data.” *Bioinformatics* 32 (3): 354–61. <https://doi.org/10.1093/bioinformatics/btv584>.
- Silva, Marcelo De A., John L. Jifon, Jorge A.G. Da Silva, and Vivek Sharma. 2007. “Use of Physiological Parameters as Fast Tools to Screen for Drought Tolerance in Sugarcane.” *Brazilian Journal of Plant Physiology* 19 (3): 193–201. <https://doi.org/10.1590/S1677-04202007000300003>.



- Simão, Felipe A., Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, and Evgeny M. Zdobnov. 2015. "BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs." *Bioinformatics* 31 (19): 3210–12. <https://doi.org/10.1093/bioinformatics/btv351>.
- Smith-Unna, Richard, Chris Bournell, Rob Patro, Julian M. Hibberd, and Steven Kelly. 2016. "TransRate: Reference-Free Quality Assessment of de Novo Transcriptome Assemblies." *Genome Research* 26 (8): 1134–44. <https://doi.org/10.1101/gr.196469.115>.
- Soneson, Charlotte, and Mauro Delorenzi. 2013. "A Comparison of Methods for Differential Expression Analysis of RNA-Seq Data." *BMC Bioinformatics* 14. <https://doi.org/10.1186/1471-2105-14-91>.
- Souza, Rafael Soares Correa De, Vagner Katsumi Okura, Jaderson Silveira Leite Armanhi, Beatriz Jorrín, Núria Lozano, Márcio José Da Silva, Manuel González-Guerrero, et al. 2016. "Unlocking the Bacterial and Fungal Communities Assemblages of Sugarcane Microbiome." *Scientific Reports* 6. <https://doi.org/10.1038/srep28774>.
- Stewart, Frank J., Sherry L. Seston, Neha Sarode, Sangita Ganesh, and Darren J. Parris. 2016. "Generation and Analysis of Microbial Metatranscriptomes." *Manual of Environmental Microbiology, 4th Edition*, 2.4.5-1-2.4.5-19. <https://doi.org/10.1128/9781555818821.ch2.4.5>.
- Su, Xiaoyang, Zhewang Lin, and Hening Lin. 2013. "The Biosynthesis and Biological Function of Diphthamide." *Critical Reviews in Biochemistry and Molecular Biology*. <https://doi.org/10.3109/10409238.2013.831023>.
- Su Yien Ting, Adeline, Siew Hoon Tan, and Mei Kay Wai. 2009. "Isolation and Characterization of Actinobacteria with Antibacterial Activity From Soil and Rhizosphere Soil." *Australian Journal of Basic and Applied Sciences* 3 (4): 4053–59.

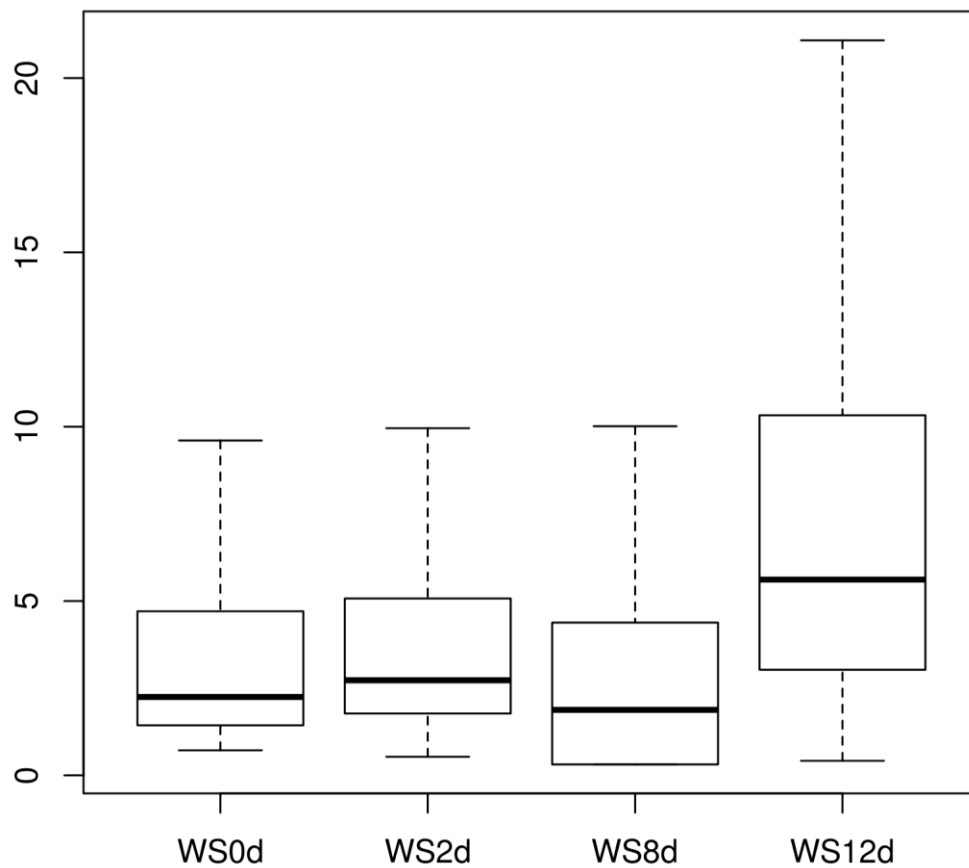
- Supek, Fran, Matko Bošnjak, Nives Škunca, and Tomislav Šmuc. 2011. “Revigo Summarizes and Visualizes Long Lists of Gene Ontology Terms.” *PLoS ONE* 6 (7). <https://doi.org/10.1371/journal.pone.0021800>.
- Tarazona, Sonia, Fernando García-Alcalde, Joaquín Dopazo, Alberto Ferrer, and Ana Conesa. 2011. “Differential Expression in RNA-Seq: A Matter of Depth.” *Genome Research* 21 (12): 2213–23. <https://doi.org/10.1101/gr.124321.111>.
- Tarazona, Sonia, Fernando García, Alberto Ferrer, Joaquín Dopazo, and Ana Conesa. 2012. “NOIseq: A RNA-Seq Differential Expression Method Robust for Sequencing Depth Biases.” *EMBnet.Journal* 17 (B): 18. <https://doi.org/10.14806/ej.17.B.265>.
- Turner, Thomas R, Euan K James, Philip S Poole, JA Gilbert, F Meyer, J Jansson, J Gordon, et al. 2013. “The Plant Microbiome.” *Genome Biology* 14 (6): 209. <https://doi.org/10.1186/gb-2013-14-6-209>.
- Urich, Tim, Anders Lanzén, Ji Qi, Daniel H. Huson, Christa Schleper, and Stephan C. Schuster. 2008. “Simultaneous Assessment of Soil Microbial Community Structure and Function through Analysis of the Meta-Transcriptome.” *PLoS ONE* 3 (6). <https://doi.org/10.1371/journal.pone.0002527>.
- Weirauch, MT. 2011. “Gene Coexpression Networks for the Analysis of DNA Microarray Data.” *Applied Statistics for Network Biology: Methods*. <http://onlinelibrary.wiley.com/doi/10.1002/9783527638079.ch11/summary>.
- Wu, Y., H. Zhou, Y. X. Que, R. K. Chen, and M. Q. Zhang. 2008. “Cloning and Identification of Promoter Prd29A and Its Application in Sugarcane Drought Resistance.” *Sugar Tech* 10 (1): 36–41. <https://doi.org/10.1007/s12355-008-0006-0>.
- Xie, Shang Ping, Clara Deser, Gabriel A. Vecchi, Matthew Collins, Thomas L. Delworth, Alex Hall, Ed Hawkins, et al. 2015. “Towards Predictive Understanding of Regional Climate Change.” *Nature Climate Change*. <https://doi.org/10.1038/nclimate2689>.

- Xie, Yinlong, Gengxiong Wu, Jingbo Tang, Ruibang Luo, Jordan Patterson, Shanlin Liu, Weihua Huang, et al. 2014. "SOAPdenovo-Trans: De Novo Transcriptome Assembly with Short RNA-Seq Reads." *Bioinformatics* 30 (12): 1660–66. <https://doi.org/10.1093/bioinformatics/btu077>.
- Xu, Ling, Dan Naylor, Zhaobin Dong, Tuesday Simmons, Grady Pierroz, Kim K. Hixson, Young-Mo Kim, et al. 2018. "Drought Delays Development of the Sorghum Root Microbiome and Enriches for Monoderm Bacteria." *Proceedings of the National Academy of Sciences*, 201717308. <https://doi.org/10.1073/pnas.1717308115>.
- Yang, Yi, Na Wang, Xinyan Guo, Yi Zhang, and Boping Ye. 2017. "Comparative Analysis of Bacterial Community Structure in the Rhizosphere of Maize by Highthroughput Pyrosequencing." *PLoS ONE* 12 (5). <https://doi.org/10.1371/journal.pone.0178425>.
- Yeoh, Yun Kit, Chanyarat Paungfoo-Lonhienne, Paul G. Dennis, Nicole Robinson, Mark A. Ragan, Susanne Schmidt, and Philip Hugenholtz. 2016. "The Core Root Microbiome of Sugarcanes Cultivated under Varying Nitrogen Fertilizer Application." *Environmental Microbiology* 18 (5): 1338–51. <https://doi.org/10.1111/1462-2920.12925>.
- Yin, Yanbin, Xizeng Mao, Jincai Yang, Xin Chen, Fenglou Mao, and Ying Xu. 2012. "DbCAN: A Web Resource for Automated Carbohydrate-Active Enzyme Annotation." *Nucleic Acids Research* 40 (W1). <https://doi.org/10.1093/nar/gks479>.
- Yost, Susan, Ana E. Duran-Pinedo, Ricardo Teles, Keerthana Krishnan, and Jorge Frias-Lopez. 2015. "Functional Signatures of Oral Dysbiosis during Periodontitis Progression Revealed by Microbial Metatranscriptome Analysis." *Genome Medicine* 7 (1). <https://doi.org/10.1186/s13073-015-0153-3>.
- Yun, J. 2010. Package 'som'. R-project CRAN.
- Zandalinas, Sara I., Ron Mittler, Damián Balfagón, Vicent Arbona, and Aurelio Gómez-

- Cadenas. 2018. "Plant Adaptations to the Combination of Drought and High Temperatures." *Physiologia Plantarum*. <https://doi.org/10.1111/ppl.12540>.
- Zerbino, Daniel R., and Ewan Birney. 2008. "Velvet: Algorithms for de Novo Short Read Assembly Using de Bruijn Graphs." *Genome Research* 18 (5): 821–29. <https://doi.org/10.1101/gr.074492.107>.
- Zhai, Lei, Yanfen Xue, Yuehao Song, Mingjie Xian, Liang Yin, Naiqin Zhong, Guixian Xia, and Yanhe Ma. 2014. "Overexpression of AaPal, a Peptidoglycan-Associated Lipoprotein from *Alkalomonas Amylolytica*, Improves Salt and Alkaline Tolerance of *Escherichia Coli* and *Arabidopsis Thaliana*." *Biotechnology Letters* 36 (3): 601–7. <https://doi.org/10.1007/s10529-013-1398-9>.

## APÊNDICE I

Material suplementar referente ao capítulo II.



**Figura S1:** Transcritos relacionados à resistência a antibióticos da classe das sulfonamidas de acordo com busca no banco de dados de resistência à antibióticos CARD.

## APÊNDICE II

Material suplementar referente ao capítulo III.

Organic acid	WS0d	WS2d	WS8d	WS12d	Chi <sup>2</sup>	P-value
Citric	6,56 (± 1,95)	2,88 (± 1,50)	11,24 (± 4,40)	21,8 (± 11,3)	12,86	<b>0,005*</b>
Isocitric	0,37 (± 0,06)	0,13 (± 0,03)	0,79 (± 0,14)	4,16 (± 2,22)	10,08	<b>0,018*</b>
Lactic <sup>1</sup>	0,87 (± 0,07)	2,32 (± 1,31)	11,52 (± 1,77)	0,19 (± 0)	ND	ND
Maleic-Fumaric	0,49 (± 0,10)	0,25 (± 0,14)	0,59 (± 0,24)	1,18 (± 0,90)	5,35	0,148
Malic	9,39 (± 1,39)	3,04 (± 0,84)	17,26 (± 2,46)	51,35 (± 19,26)	12,53	<b>0,023*</b>
Succinic	1,22 (± 0,08)	0,78 (± 0,53)	1,39 (± 0,46)	1,06 (± 0,86)	2,18	0,534
Malonic <sup>1</sup>	ND	ND	ND	ND	ND	ND

<sup>1</sup> Statistical analysis not performed (n < 3)

ND: not determined

\* Statistical significance at 5%

**Table S1** Mean values and standard deviation of organic acid exudation rate by CTC 9001 sugarcane roots. The values are expressed in µg.g of root-1.min-1. Adapted from Pereira et al., 2018.

Comparison	Expression	Clusters
0d vs 2d	up-regulated	c268 c277 c437
0d vs 2d	down-regulated	c517 c66 c260 c254 c199 c228 c387 c225 c151 c233 c257 c51 c72 c166
0d vs 8d	up-regulated	c434 c220 c377 c269 c469 c244 c164 c273 c175 c214 c197 c136 c69 c470 c326 c169 c515 c440 c47 c284 c308 c344 c78 c45 c253 c262 c48
0d vs 8d	down-regulated	c114 c474 c527 c491 c578 c425 c160 c398 c466 c226 c26 c431 c178 c19 c154 c352 c357 c101 c234 c36 c229 c57 c310 c180 c212 c315 c274 c30 c5 c544 c188 c236 c88 c138 c218 c325 c329 c146 c294 c464 c174 c191 c415 c426 c157 c435 c483 c232 c298 c449 c276 c411 c230 c580 c207 c249 c278 c399 c53 c23 c526 c428 c66 c465 c125 c34 c492 c51 c194 c432 c442 c265 c266 c320 c481 c261 c27 c498 c340 c170 c406 c116 c192 c517 c419 c55 c225 c279 c208 c252 c143 c412 c42 c199 c387 c237 c259 c392 c105 c200 c359 c65 c233 c343 c111 c74 c102 c151 c201 c263 c256 c228 c110 c50 c242 c250 c177 c72 c144 c182 c196 c84 c166 c130 c257
0d vs 12d	up-regulated	NA
0d vs 12d	down-regulated	c273 c220 c544 c443 c474 c578 c329 c157 c414 c57 c229 c410 c580 c154 c281 c110 c293 c399 c180 c315 c5 c430 c233 c65 c442 c398 c435 c27 c53 c91 c105 c23 c236 c133 c166 c465 c479 c50 c199 c272 c249 c265 c41 c412 c428 c483 c170 c34 c182 c247 c257 c325 c431 c492 c143 c415 c145 c177 c192 c207 c406 c432 c144 c295 c298 c449 c146 c111 c201 c200 c276 c252 c425 c172 c174 c359 c426 c102 c218 c241 c260 c527 c512 c528 c95 c256 c352 c320 c42 c208 c254 c517 c178 c278 c138 c263 c294 c106 c237 c429 c55 c222 c250 c196 c266 c267 c232 c387 c261 c51 c82 c130 c77 c168 c275 c101 c75 c225 c230 c264 c158 c72 c206 c84

**Table S2** Differentially expressed transcripts comparing samples without water stress (0d) against samples with water stress (2d, 8d, and 12d) and the clusters that had 20% or more transcripts up or down-regulated in one of the analyzed comparisons.

## ANEXO I



COORDENADORIA DE PÓS-GRADUAÇÃO  
INSTITUTO DE BIOLOGIA  
Universidade Estadual de Campinas  
Caixa Postal 6109. 13083-970, Campinas, SP, Brasil  
Fone (19) 3521-6378. email: cpqib@unicamp.br



## DECLARAÇÃO

Em observância ao §5º do Artigo 1º da Informação CCPG-UNICAMP/001/15, referente a Bioética e Biossegurança, declaro que o conteúdo de minha Dissertação de Mestrado, intitulada "*Análise do Metatranscriptoma de Rizosfera de Cana-de-açúcar Submetida ao Estresse Hídrico*", desenvolvida no Programa de Pós-Graduação em Genética e Biologia Molecular do Instituto de Biologia da Unicamp, não versa sobre pesquisa envolvendo seres humanos, animais ou temas afetos a Biossegurança.

Assinatura: \_\_\_\_\_

Handwritten signature of Victor Marques de Oliveira Gambarini in blue ink.

Nome do(a) aluno(a): Victor Marques de Oliveira Gambarini

Assinatura: \_\_\_\_\_

Handwritten signature of Renato Vicentini dos Santos in blue ink.

Nome do(a) orientador(a): Renato Vicentini dos Santos

Data: 24/08/2018



**ANEXO II****Declaração**

As cópias de artigos de minha autoria ou de minha co-autoria, já publicados ou submetidos para publicação em revistas científicas ou anais de congressos sujeitos a arbitragem, que constam da minha Dissertação/Tese de Mestrado/Doutorado, intitulada **Análise do Metatranscriptoma de Rizosfera de Cana-de-açúcar Submetida ao Estresse Hídrico**, não infringem os dispositivos da Lei n.º 9.610/98, nem o direito autoral de qualquer editora.

Campinas, 28/08/2018

Assinatura : Victor Gambarini  
Nome do(a) autor(a): **Victor Marques de Oliveira Gambarini**  
RG n.º 32.479.495-2

Assinatura : Renato  
Nome do(a) orientador(a): **Renato Vicentini dos Santos**  
RG n.º 29.995.011-6