Desenvolvimento da Plataforma CaneRegNet para Anotação Funcional e Análises do Transcriptoma da Cana-de-açúcar

Milton Yutaka Nishiyama Junior

Tese Apresentada
ao
Programa InterUnidades de Pós-Graduação em Bioinformática
da
Universidade de São Paulo
para
Obtenção do Título
de
Doutor em Ciências

Área de Concentração: Bioinformática
Orientadora: Profa. Dr. Glaucia Mendes Souza
Co-Orientador: Prof. Dr. João Eduardo Ferreira

Durante o desenvolvimento deste trabalho, o autor recebeu auxílio financeiro da CAPES

São Paulo, Janeiro de 2015

Desenvolvimento da Plataforma CaneRegNet para Anotação Funcional e Análises do Transcriptoma da Cana-de-açúcar

Tese apresentada ao Programa
InterUnidades de Pós-Graduação em
Bioinformática da Universidade de
São Paulo para obtenção do
título de Doutor em Ciências

São Paulo, Janeiro de 2015

À minha mãe Sandra, ao meu pai Milton,
ao meu irmão Márcio
À minha esposa Ana
Dedico

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

Agradecimentos

Primeiramente agradeço a Deus e à Nossa Senhora, por me dar forças, coragem e orientação durante todos estes anos de trabalho e superação no Doutorado.

Agradeço à minha esposa amada Ana Claudia por estar sempre ao meu lado, em todos os momentos alegres e difíceis da Vida e principalmente nesta fase final do Doutorado. Você é uma das principais responsáveis pelo êxito neste trabalho.

Agradeço à Profa. Glaucia, que é um exemplo de profissionalismo e competência, pela excelente conviência durante estes últimos 10 anos, pela orientação e por todo o suporte e oportunidades que proporcionou à minha carreira. Ao Prof. João Eduardo Ferreira pela excelente Co-Orientação e por ser decisivo em momentos importantes deste Doutorado.

Agradeço aos Profs. Sergio Verjovski-Almeida e Aline Maria da Silva, por me darem a oportunidade de conhecer e trabalhar com Bionformática e de aprender a amar a ciência.

Agradeço aos meus pais Milton e Sandra, por me trazerem a este mundo, pela excelente educação e orientação para o caminho do bem, por me mostrarem a alegria de viver e que sempre é possível se superar.

Agradeço ao meu irmão Márcio por me mostrar que a vida é muito importante, e que é possível ser feliz independente dos tombos no caminho, por ser um exemplo de vida e uma referência de perseverança e superação, você sempre estará me surpreendendo e me orientando de alguma maneira.

Agradeço a todos meus familiares, em especial às minhas Tias Mariê, Suely e Toshie e à minha avó Tsuma por me proporcionarem oportunidades para que Eu pudesse chegar onde cheguei, pelos exemplos de vida, de perseverança e bondade, e pelos maravilhosos momentos de alegrias e amor.

Agradeço às centenas de grandes amigos que fiz e passaram pela minha vida, deixando ótimas lembranças, ao longo destes últimos 15 anos no Instituto de Química, Instituto de Computação e Estatística, Instituto Butantan, e Instituto Ludwig.

Agradeço aos amigos Sávio, pelo companheirismo e auxílio nesta fase final do doutorado, ao grande amigo Diorge, pela convivência e inesquecíveis experiências em tantas noites de pizza, butecos e viagens, que serão lembranças e histórias para o resto da vida, ao (R)Júlio pela amizade e discussões, ao amigo Francisco pela grande amizade e longas discussões, ao amigo Apuã, pela orientação e ensinamento nos primeiros anos de Bioinformática e pelas noites de sinuca, feijoadas e discussões.

Agradeço a todo o pessoal do laboratório, principalmente à amiga Luciana, pela amizade e companhia nas longas horas de trabalho e pelas discussões, à Érica, Carolina, Paloma, Sarah e Alessandra pela amizade e pelos momentos divertidos nas festas do cachorro quente e nos almoços do FranBoi, ao Carlos Hotta pela amizade e enriquecedoras discussões científicas, aos amigos Rodrigo, Maxmiller e Humberto pela amizade e discussões do dia a dia.

Agradeço ao Edwin, pelos vários momentos de discussões e pelo grande suporte para que este trabalho pudesse tornar-se muito mais importante.

Agradeço em especial à Patricia Martorelli, secretária do Programa de Bioinformática, pela orientação nas questões burocráticas, paciência e auxílio durante todo o Doutorado.

Agradeço a todos que de alguma forma contribuíram para este trabalho.

Agradeço ao CAPES pelo auxílio financeiro na realização deste trabalho e pelo suporte da FAPESP #2008/52146-0.

Resumo

Nishiyama-Jr, Milton Yutaka. **Desenvolvimento da Plataforma CaneRegNet para Anotação Funcional e Análises do Transcriptoma da Cana-de-açúcar**. Tese (Doutorado em Bioinformática) - Interunidades em Bioinformática, Universidade de São Paulo, São Paulo, 2015.

A identificação de genes alvos, vias de sinalização e vias metabólicas para melhoramento de cana-de-açúcar associados a características de interesse, ainda são pouco conhecidos e estudados. Alguns estudos do transcriptoma através de plataformas de microarranjo têm buscado identificar listas de genes, para experimentos tecido-específico ou submetidos a condições de estresse bióticos e abióticos. Estudos pontuais destes dados tem sido associados a vias metabólicas ou vias de sinalização já descritas na literatura, de forma a identificar alterações relacionadas a padrões de expressão gênica. Porém, estas relações em cana-de-açúcar são pouco conhecidas e estudadas. O estudo e entendimento de cana-de-açúcar por meio da diversidade genética e de sua adaptação ao ambiente é um grande desafio, principalmente pela ausência de um genoma sequenciado e por possuir um genoma complexo. Apresentamos nossos resultados para tentar superar tais limitações e desafios para estudos de expressão gênica. Foram desenvolvidas metodologias para anotação funcional do transcriptoma, centradas na transferência de anotação, identificação de vias metabólicas e enzimas pelo método de similaridade bi-direcional, predição de genes *full-length*, análises de ortologia e desenho de oligonucleotídeos para microarranjos customizados, resultando no ORFeoma de cana-de-açúcar, na identificação e classificação de famílias de fatores de transcrição e identificação de genes ortólogos entre gramíneas. Além disso, desenvolvemos uma plataforma para processamento e análise automatizada de experimentos por microarranjo, para armazenamento, recuperação e integração com a anotação funcional. Adicionalmente desenvolvemos e implementamos métodos para seleção de genes diferencialmente e significativamente expressos, e abordagens para análise de enriquecimento de categorias, e escores de atividade de vias metabólicas. De forma a integrar a anotação funcional do transcriptoma aos estudos por expressão gênica, desenvolvemos a plataforma CaneRegNet e uma interface para integração desta rede de dados biológicos e conhecimentos, composta por aplicativos para consulta e prospecção de dados por análises de agrupamento e correlação entre experimentos de microarranjo, possibilitando a geração de novas hipóteses e previsões dentro da organização da regulação celular.

Palavras-chave: Bioinformática, Banco de Dados, Anotação funcional, integração de dados, prospecção de dados, plataforma de microarranjo, transcriptoma Cana-de-açúcar

Abstract

Nishiyama-Jr, Milton Yutaka. **Development of CaneRegNet Platform for Functional Annotation and Analysis of Sugarcane Transcriptome**. Tese (Doutorado em Bioinformática) - Interunidades em Bioinformática, Universidade de São Paulo, São Paulo, 2015.

The identification of target genes, metabolic and signaling pathways associated with characteristics of interest to the sugarcane improvement are still poorly known and studied. Some transcriptome studies through microarray platforms has tried to identify lists of genes, for tissue-specific experiments or subjected to conditions of biotic and abiotic stress. In the literature specific studies of these data has already been associated with metabolic or signaling pathway, in order to identify changes in these tracks related to patterns of gene expression. However, these relations are still little know and generally defined slightly. The study and understanding of sugarcane by means of genetic diversity and its adaptation to the environment is a major challenge, mainly due to the absence of a sequenced genome and by your complex genome. We present our results to surpass this barrier e challenges for the study of gene expression. Methodologies were developed for the transcriptome functional annotation, focused on the annotation transfer, identification of metabolic pathways and enzymes by the bi-directional method; prediction of full-length genes; ortology analysis and probe design for customized microarrays, resulting in the sugarcane ORFeome, the identification and classification of transcription factor families and identification of ortholog genes between grasses. Besides that, we have developed a plataform for automated processing and analysis for microarray experiments, to store, retrieve and integration with the functional annotation. Additionally, we have developed and implemented methods for identification of differentially and significantly expressed genes, and approaches for over-represented analysis and functional class scoring (FCS). To integrate the functional annotation and the studies by gene expression profile, we have developed the CaneRegNet platform and an interface to integrate this network of biological data and knowledge, composed by searching and data mining tools for clustering and correlations between microarray experiments, enabling the generation of new hypothesis and predictions around the organization of cellular regulation.

Keywords: Bioinformatics, Database, Functional annotation, data integration, data mining, microarray platform, sugarcane transcriptome.

Lista de Figuras

- 1** Representação diagramática do genoma da cana-de-açúcar moderna. As linhas representam os cromossomos homólogos, as barras em branco os cromossomos herdados de *S. officinarum*, as barras em cinza os cromossomos herdados de *S. spontaneum* e as barras com duas cores as recombinações cromossômicas entre as duas espécies ancestrais.....8
- 2** Evolução do número de sequências de mRNA e genômicas depositadas no GenBank.....10
- 3** Árvore construída a partir de organismos pré-selecionados, baseado na classificação taxonômica do banco de dados do NCBI (Fderhen, 2012).....11
- 4** Distribuição do total de proteínas no Reino Eucariota, correspondendo a 33% do Uniprot (Dados de 26-Nov-2014 do UniProtKB/Swiss-Prot).....12
- 5** Representação dos dois grupos de NATs de acordo com sua origem genômica: cis-NATs e trans-NATs. (A) Os cis-NATs são formados no mesmo locus do respectivo par senso, e (B) trans-NATs e seu respectivo par são formados em locus distintos.....15
- 6** Representação do complexo transcricional em eucariotos para genes transcritos pela RNA Pol II (modificado de Maston G.A. et. al., 2006).....18
- 7** Esquema de um experimento de microarranjo. Fragmentos de um gene, ou oligonucleotídeos são sintetizados e imobilizados em lâminas de vidro. Os cDNAs das amostras de duas condições experimentais são marcados com os fluoróforos Cy3 e Cy5 e hibridizados. Após a hibridização das amostras na lâmina, um escaner é utilizado para obter imagens monocromáticas, em seguida são extraídos os níveis de intensidade para cada ponto na lâmina em cada uma das condições e os resultados são utilizados para calcular a razão de expressão para cada fragmento/oligonucleotídeo (i.e. tratamento/controle).....22
- 8** Visão geral dos métodos de análises de vias usando dados de expressão gênica como exemplo, mas pode ser aplicado a dados de outras tecnologias de larga escala. O dado gerado por um ou mais experimentos e anotações funcionais (vias) são a entrada para os métodos de análise.....25
- 9** Utilizando a combinação de tamanhos de insertos curtos e longos com sequenciamento *paired-end* resulta em máxima cobertura para montagens de novo. *Reads* de insertos longos podem cobrir de forma mais eficiente regiões repetitivas. *Reads* de insertos curtos sequenciados com alta cobertura podem preencher *gaps* deixados pelos *reads* longos.....29
- 10** Uma representação parcial de uma via metabólica no KEGG. Os nós são representados por caixas (produtos de genes) e círculos (compostos químicos). Arestas representam interações ou relações moleculares. Em vermelho são os genes de interesse mapeado na Via.....34

11	Estrutura do mRNA, e suas características como Cap 5', sequência codificadora, regiões 5' e 3' UTR não traduzidas e cauda Poli-A.....	42
12	. <i>Pipeline</i> para anotação do transcriptoma de cana-de-açúcar e identificação de genes ortólogos.....	46
13	Esquema da metodologia desenvolvida para identificação de transcritos completos (<i>full-length</i>) composto de três abordagens distintas, utilizando os genes completos de gramíneas (Sorgo, Milho, Arroz, <i>Brachypodium</i> , <i>Panicum</i> e Setária). Na Análise I, o CDS deve ser totalmente coberto pelo transcrito com melhor <i>hit</i> ; na Análise II, o gene é totalmente coberto por dois ou mais transcritos; na Análise III o transcrito é classificado como completo se possuir tamanho médio da CDS > 1000pb e apresentar códon de início e fim.....	57
14	Diagrama de fluxo descrevendo os passos envolvidos na classificação dos Fatores de transcrição.....	61
15	Estratégia do <i>CaneOligoDesign</i> , para desenho de oligonucleotídeos e validação, composto por cinco 5 etapas: 1) Seleção dos SAS de interesse pela anotação funcional, e Identificação dos SAS codificadores, não codificadores e da orientação; 2) Desenho de todos os oligonucleotídeos senso e antisenso para cada SAS selecionado, descartando oligonucleotídeos que não atendam aos critérios de hibridação; 3) Seleção dos oligonucleotídeos únicos de acordo com os critérios de especificidade; 4) Classificação dos SAS pelos tipos de oligonucleotídeo com orientação (senso, e o par senso e antisenso).....	64
16	Esquema do desenho dos oligonucleotídeos pelo programa <i>CaneOligoDesign</i>	66
17	Representação gráfica da frequência da análise do <i>Score bit</i> do segundo <i>hit</i> dos oligonucleotídeos mapeados contra os ESTs decana-de-açúcar por BLASTN. A linha vertical vermelha indica a posição do <i>Score bit</i> 41.....	68
18	Representação gráfica da frequência da análise da cobertura do segundo <i>hit</i> dos oligonucleotídeos mapeados contra os ESTs de cana-de-açúcar, através do BLASTN.....	68
19	Correlação entre os sinais corrigidos dos fluoróforos Cy3 e Cy5 para os oligonucleotídeos senso (azul), antisenso (vermelho) e controles da lâmina (amarelo).....	75
20	Gráfico do logaritmo da razão de expressão dos oligonucleotídeos senso (verde) e antisenso (laranja) vs logaritmo da média dos sinais. A linha vermelha delimita o intervalo de credibilidade com p=90%, identificando os genes induzidos/reprimidos pelo método Htself adaptado.....	76

21	Diagrama da Plataforma <i>CaneRegNet</i> e Abstração dos Dados, baseado em seis tópicos principais (Categorização de genes, Anotação Funcional, Estudos de perfil de expressão, Sequenciamento, Fisiologia e Plantas Transgênicas).....	89
22	Diagrama de Entidade Relacionamento do banco de dados CaneRegNetDB.....	90
23	Diagrama de Entidade Relacionamento representando o controle de acesso da plataforma CaneRegNetDB.....	93
24	Representação geral das vias metabólicas, integradas a genes e experimentos de microarranjo no banco de grafos.....	96
25	Arquitetura da plataforma CaneRegNet.....	97
26	Representação detalhada da arquitetura da plataforma <i>CaneRegNet</i> , exemplificada por uma aplicação de busca de <i>reads</i> , mas segue o mesmo padrão para outras buscas mudando apenas as páginas JSP.....	100
27	Fluxograma de busca de um gene, mostrando as etapas de consulta e processamento da informação e montagem da estrutura de exibição da nova seção.....	104
28	Fluxograma representando a lógica das aplicações para aquisição do conhecimento através da integração de dados de diferentes fontes. Aplicações para consulta de dados do Transcriptoma, Anotação Funcional e Análises por dados de microarranjo.....	106
29	Representação da porcentagem de genes pelo alinhamento contra os transcritos, utilizando o algoritmo para calcular a cobertura total de todos os hits em um mesmo gene, distribuídos em faixas de porcentagem de cobertura.....	111
30	Representação da porcentagem de genes (gramíneas) e transcritos (RNA-Seq) identificados nos alinhamentos entre as gramíneas e transcritos do <i>RNA-Seq</i> , com critério de e-value 1e-20.....	112
31	Distribuição das Ontologias de Processos Biológicos mais frequentes no SUCEST e <i>RNA-Seq</i>	115
32	Distribuição das Ontologias de Funções Moleculares mais frequentes no SUCEST e <i>RNA-Seq</i>	115
33	Distribuição das Ontologias de Componentes Celulares mais frequentes no SUCEST e <i>RNA-Seq</i>	116
34	Diagrama de Venn da distribuição dos transcritos anotados do SUCEST (A) e <i>RNA-Seq</i> (B) com as anotações por GO entre as três diferentes ontologias.....	116
35	Vias metabólicas do KEGG mais representadas com os transcritos anotados do SUCEST e <i>RNA-Seq</i> pelo KAAS, for removidas desta figura as Vias <i>Metabolic Pathways</i> e <i>Biosynthesis of secundar metabolism</i>	119
36	O gráfico representa as proporções dos 7 domínios mais abundantes no PFAM para o SUCEST.....	122
37	O gráfico representa as proporções dos 7 domínios mais abundantes no PFAM para o <i>RNA-Seq</i>	122

38	Distribuição das proteínas de cana-de-açúcar em grupos ortólogos compartilhados com outras espécies. Os valores se referem ao total de grupos de ortólogos e de proteínas identificadas como ortólogos com cana-de-açúcar, e parálogos dentro da própria espécie.....	123
39	Alinhamento entre genes ortólogos e parálogos com Clustalw.....	125
40	Árvore filogenética para os genes ortólogos e parálogos construída com Clustalw.....	126
41	Anotação Funcional dos transcritos <i>full-length</i> (40.407). O gráfico mostra as categorias mais frequentes da anotação do Phytozome. Foram identificadas 5.038 categorias.....	131
42	Distribuição do comprimento das ORFs e cana-de-açúcar, transcritos <i>full-length</i> (FLs), e UTRs. Os gráficos representam a comparação da distribuição (barra cinza) do tamanho das ORFs (A) e transcritos <i>full-length</i> (B) com outras gramíneas (linhas coloridas). A distribuição do comprimento das UTRs 5' e 3' (C, barras preto e branco) dos transcritos <i>full-length</i> em cana-de-açúcar são apresentados também.....	132
43	Estimativa do número de TFs em genomas de gramíneas. Correlação entre o número de TFs e o número total de genes em genomas foi baseado nos genomas completamente anotados de <i>Arabidopsis</i> , arroz, poplar e <i>Chlamydomonas</i> (pontos pretos). A curva de regressão linear ($r = 0.87$) foi usada para estimar o número de TFs preditos de milho, sorgo, e cana-de-açúcar (Tabela 17).....	134
44	Tabela com a distribuição do número de TFs por família identificados no SUCEST (linhas verdes), no <i>RNA-Seq</i> (linhas azuis), e os TFs <i>full-length</i> (linhas vermelhas).....	136
45	Distribuição das categorias mais representativas na lâmina de microarranjo separada em transcritos senso e antisenso.....	138
46	Gráfico da densidade da distribuição da intensidade dos sinais dos oligonucleotídeos senso abaixo e acima do <i>background</i>	139
47	Gráfico da densidade da distribuição da intensidade dos sinais dos oligonucleotídeos antisenso abaixo e acima do <i>background</i>	139
48	Padrão de expressão dos pares de oligonucleotídeos senso e antisenso com sinal acima do sinal de fundo. (A) O padrão de expressão dos pares separados pelo tempo sem irrigação. O eixo Y indica o número de oligonucleotídeos, o eixo X indica o padrão de expressão do senso e antisenso. Os tons de cinza indicam o padrão de expressão entre os pares.....	140
49	Padrão de expressão dos pares senso e antisenso identificados pelo método adaptado HTSelf. O eixo X indica o padrão de expressão do oligonucleotídeo antisenso e o eixo Y o número de oligonucleotídeos. As cores na legenda indicam o padrão de expressão dos oligonucleotídeos senso do par.....	141

50	Número de <i>reads</i> identificados como NATs, baseado na orientação do alinhamento com as gramíneas.....	142
51	Porcentagem dos <i>reads</i> senso e NATs alinhados contra cada espécie em relação ao total de <i>reads</i> sequenciados no PGM (29 milhões de <i>reads</i>).....	144
52	Visão geral da plataforma de integração para estudo do transcriptoma de cana-de-açúcar.....	146
53	Interface inicial da plataforma CaneRegNet.....	151
54	Busca e Visualização para transcritos através da interface do CaneRegNet. A página inicial permite a busca por expressão regular, do nome do transcritos, parte dele, por anotações , ESTs ou oligonucleotídeos, enviando as consultas ao bancos de dados à camada de controle. A página exibida é estruturada como descrito na metodologia para apresentar os resultados de cada módulo vinculado ao transcrito e ao clicar em uma aba diferente as informações são exibidas dinamicamente.....	156
55	Visualização da busca para uma lista de transcritos e página de resultado com as informações presents no CaneRegNetDB e a apresentação da figura da Via Metabólica com os transcritos representados graficamente, através das enzimas.....	157
56	Página de Visualização para todas as informações para o transcriptoma do transcrito consultado. A aba possui sub-itens que podem ser abertos para exibir toda informação, como anotações por várias abordagens, cartegorizações, clones, e expressão por Northern Digital e níveis de expressão em todos os experimentos grafica e textualmente, para os três níveis de análise de expressão: <i>probe_intensity</i> , contendo informações dos níveis de expressão normalizado por oligonucleotídeo, <i>probe_statistics</i> , com níveis de razão de expressão entre as amostras tratado e controle, e <i>probe_expression</i> , contendo apenas transcritos DE presentes em todas as réplicas, e exibidos de acordo com o nível de acesso do usuário.....	158
57	Fluxograma de Subissão de um experimento de microarranjo. E criação de experimentos híbridos.....	161
58	Fluxograma do processo de cruzamento entre múltiplos experimentos a partir da razão de expressão obtida pelo método HTSelf adaptado.....	163
59	Apresentação da interface de entrada (A) e saída (B) do aplicativo para seleção de genes SE.....	165
60	Apresentação da interface da entrada (A) e saída (B) do aplicativo para seleção de genes DE.	166
61	Interface para análise de enriquecimento de termos, recebe como entrada um arquivo com uma lista de genes para a análise, seleção do banco de dados para o estudo de enriquecimento, número de representantes por categoria a ser exibida e filtros adicionais.	167
62	Interface de saída do aplicativo para análise de enriquecimento de termos. Apresentando o gráfico do escore de enriquecimento de cada categoria (em azul), o gráfico da	

representação populacional e do conjunto de estudo e a tabela com os escores e genes de cada termo.	168
63 Comparação do escore de enriquecimento das famílias de TFs entre os três pontos experimentais para estresse por seca (24h, 72h e 120h) sem irrigação.....	169
64 Interface para análise comparativa de escore de atividade das vias entre diferentes condições experimentais. Inicialmente é escolhida a forma de seleção dos experimentos (A), por arquivo ou por projeto, em seguida são selecionados dois ou mais experimentos disponíveis ao usuário no CaneRegNetDB (B) e o banco de vias desejado, com o método para agrupamento hierárquico das vias e o método para correlação entre as condições experimentais (C).	170
65 Representação do log2 da razão do escore de atividade entre as duas condições experimentais, com o p-value do teste de Wilcoxon.....	171
66 Comparação do nível de atividade das vias em diferentes condições experimentais.....	174
67 Grupos de vias extraídas do agrupamento hierárquico usando <i>R-statistics</i> . Linhas cinzas indicam vias individuais, linha verde nível de atividade média por grupo e no topo o total de vias identificadas no cluster.....	175
68 <i>Heatmap</i> mostrando a matriz de correlação de Spearman hierarquicamente agrupada resultante da comparação entre os níveis de atividade das vias para cada par de amostras.....	175
69 Representação do escore de atividade das vias nas três condições experimentais de estresse por seca.	178
70 Representação parcial do caminho mais longo na via de metabolismo de amido e sacarose, considerando os caminhos em qualquer direção passando pela enzima 5.4.2.2.....	179
71 Representação parcial do caminho mais longo na via de metabolismo de amido e sacarose, considerando os caminhos em direção à enzima 5.4.2.2.....	180
72 Representação parcial do caminho mais curto na via de metabolismo de amido e sacarose, considerando o caminho em qualquer direção entre as enzimas 2.4.1.7 e 3.2.1.74.	181
73 Representação parcial do caminho mais curto na via de metabolismo de amido e sacarose, considerando o caminho mais curto na direção <i>incoming</i> entre as enzimas 2.4.1.7 e 3.2.1.74.....	181
74 Representação parcial do caminho mais curto na via de metabolismo de amido e sacarose, considerando o caminho <i>outgoing</i> entre as enzimas 2.4.1.7 e 3.2.1.74.....	182

Lista de Tabelas

1	Classificação das famílias de fatores de transcrição dentro das superclasses de DBDs de acordo com suas características.....	20
2	Códigos de Evidência para a proveniência das anotações do GO definidas pelo Consórcio de Ontologia de Genes (<i>The Gene Ontology Consortium</i>).....	49
3	Número de genes, completos e tamanho médio da CDS em cada espécie do Phytozome v9.0.....	59
4	Lista dos módulos criados no CaneRegNet e referência para o gerenciamento de toda a informação.....	94
5	Dados do sequenciamento por <i>RNA-Seq</i> e da montagem com o <i>software</i> Trinity.....	106
6	Representação do mapeamento dos <i>reads</i> de cada amostra nos transcritos, com o total de transcritos presentes em cada grupo. As linhas cinza representam transcritos específicos em SP80-3280; Amarelo, genótipos ancestrais; Laranja, <i>S. spontaneum</i> ; Verde, <i>S. officinarum</i>	107
7	Número total de genes por espécie e em cada banco de dados utilizado.....	109
8	Números absolutos e porcentagens do total de hits entre os transcritos de cana-de-açúcar do SUCEST e RNA-Seq contra os bancos das 6 gramíneas, <i>Gene Ontology</i> e <i>KEGG</i> , através do método de alinhamento Uni-direcional e Bi-direcional, com e-value 1e-05, score bit > 30, SBH > 0.95 ou BHR > 0.95.....	109
9	Número de transcritos anotados pelo banco do <i>Gene Ontology</i> , pela abordagem bi-direcional e agrupados nas categorias de código de evidência.....	113
10	Número total de Vias, enzimas, <i>KEGG</i> ortólogos e transcritos do SUCEST e RNA-Seq identificados com o <i>software</i> KAAS.....	117
11	As sete Enzimas mais representadas por transcritos do SUCEST e <i>RNA-Seq</i>	118
12	Número total de proteínas preditas com a matriz de <i>codon-usage</i> de cana-de-açúcar para os transcriptoma do SUCEST e RNA-Seq.....	119
13	Número de genes ortólogos putativos entre cana, gramíneas e <i>Arabidopsis</i> , baseado no InParanoid (comparação pareada) e MultiParanoid (grupos de ortólogos entre múltiplas espécies usando a saída do InParanoid).....	122
14	Número de genes ortólogos putativos entre cana, gramíneas e <i>Arabidopsis</i> , dentro de cada nó da árvore utilizando o Hieranoid.....	126
15	Número de genes, genes <i>full-length*</i> , genes com 5' e 3' UTR e tamanho médio da CDS em cada espécie.	127
16	Número de transcritos <i>full-length</i> identificados pela em cada análise.....	129
17	Número de TFs em gramíneas. O número esperado de TFs foi estimado da Figura 37.....	133

18	Comparação do GrassTFDB e TFs obtidos pelo RNA-Seq com outros bancos disponíveis de TFs.....	135
19	Lista de algumas aplicações desenvolvidos para o gerenciamento de dados, processamento e análises, separados por módulos.....	147
20	Número de oligonucleotídeos senso e antisenso identificados como SE nos experimentos por estresse hídrico, com total em relação à lâmina (A) e por condição experimental (B).	164

Sumário

Resumo	vii
Abstract	ix
Lista de Figuras	x
Lista de Tabelas	xvi
1. Introdução	1
2. Motivação	4
3. Estado da Arte - Fundamentos Biológicos e Computacionais	7
3.1. A Cana-de-açúcar	7
3.2. Anotação Funcional do Transcriptoma de Cana-de-açúcar.....	9
3.3. Detecção de Ortólogos putativos funcionais.....	14
3.4. Expressão de Transcritos senso-antisenso.....	15
3.5. Fatores de Transcrição em Gramíneas e Cana-de-açúcar.....	17
3.6. Análise de Expressão gênica em larga escala.....	21
3.7. Abordagens de mineração sobre dados de larga escala.....	24
3.8. Sequenciamento de Nova-Geração	28
3.9. Vias Metabólicas.....	32
3.10. Plataforma para Integração de dados e estudo do Transcriptoma de Cana-de-açúcar.....	34
3.11. Bancos de Dados Relacional e de Grafos.....	36
4. Objetivos	41
4.1.1. Objetivos Gerais.....	41
4.1.2. Objetivos Específicos.....	41
5. Materiais e Métodos	42
5.1. Anotação Funcional do Transcriptoma de Cana-de-açúcar.....	42
5.2. Predição de proteínas em Cana-de-açúcar e Identificação de domínios.....	50
5.3. Identificação de ortólogos entre Cana-de-açúcar e Gramíneas.....	52
5.4. Sequenciamento e Anotação do transcriptoma de Cana-de-açúcar por RNA-Seq.....	54

5.5.	Predição de transcritos completos (<i>full-length</i>) em Cana-de-açúcar.....	56
5.6.	Identificação e Classificação de Fatores de Transcrição em Gramíneas e Cana-de-açúcar.....	60
5.7.	Estudo do Transcriptoma de Cana-de-açúcar por microarranjo e sequenciamento em larga escala.....	62
5.7.1.	<i>Pipeline</i> para desenho e validação de oligonucleotídeos senso e antisenso.....	62
5.7.2.	Processamento e Quantificação de sinais de fluorescência de microarranjos.....	68
5.7.3.	Ajuste de Sinal e correção do sinal de fundo.....	69
5.7.4.	Normalização de microarranjos.....	71
5.7.5.	Identificação de genes significativamente expressos.....	73
5.7.6.	Identificação de genes diferencialmente expressos.....	76
5.7.7.	Análise de Co-expressão entre condições experimentais.....	78
5.7.8.	Análise de Enriquecimento de Termos.....	79
5.7.9.	Análise de escore de classes funcionais.....	81
5.7.10.	Análise de vias metabólicas em grafos.....	83
5.8.	Plataforma CaneRegNet para Integração do Transcriptoma de Cana-de-açúcar.....	87
5.8.1.	Estrutura do Banco de dados Relacional.....	89
5.8.2.	Controle de Acesso da Plataforma <i>CaneRegNet</i>	92
5.8.3.	Estrutura do Banco de dados de Grafos.....	94
5.8.4.	Implementação e Arquitetura da Plataforma CaneRegNet.....	96
5.8.5.	Interface e Estrutura de Integração	101

6. Resultados e Discussão 107

6.1.	Anotação Funcional do transcriptoma de Cana-de-açúcar.....	107
6.1.1.	Anotação Funcional por Gramíneas.....	109
6.1.2.	Anotação por <i>Gene Ontology</i>	113
6.1.3.	Anotação Funcional por Vias Metabólicas e Enzimas (KEGG).....	117
6.1.4.	Predição de proteínas e Identificação de domínios PFAM.....	120
6.2.	Identificação de proteínas Ortólogas entre Cana-de-açúcar e Gramíneas.....	122
6.3.	Predição de genes <i>Full-Length</i>	128
6.4.	Identificação e Classificação de Famílias de Fatores de Transcrição.....	133
6.5.	Análise de Expressão senso e antisenso em Cana-de-açúcar.....	137

6.5.1.	Expressão de pares senso-antisense por microarranjo.....	137
6.5.2.	Estudo de transcritos antisense por <i>RNA-Seq</i>	141
6.6.	Plataforma CaneRegNet para Estudo e Análise do Transcriptoma de Cana-de-açúcar.....	144
6.6.1.	Principais Interfaces de Consulta e Visualização.....	155
6.6.2.	Interface para Submissão e Processamento de lâminas de Microarranjo.....	159
6.6.3.	Interface para Cruzamento entre réplicas de experimento por microarranjo.....	161
6.7.	Abordagens para mineração de dados de expressão gênica.....	163
6.7.1.	Análises de genes significativamente expressos (SE) e diferencialmente (DE).....	163
6.7.2.	Análise de Enriquecimento de Termos.....	166
6.7.3.	Análise de Escore de Atividade das Vias	169
6.7.4.	Análise de Topologia de Vias.....	175
7.	Conclusões e Perspectivas	183
7.1.	Conclusões.....	183
7.2.	Perspectivas.....	184
8.	Referências	186
9.	Anexos	215

1. Introdução

A cana-de-açúcar (*Saccharum spp.*) tem importância mundial na área de energia renovável, incluindo biocombustíveis, bioeletricidade, biopolímeros e açúcar. O estudo e melhoramento da cana-de-açúcar por meio da diversidade genética e de sua adaptação ao ambiente é um grande desafio, principalmente por possuir um genoma complexo (poliploide e aneuploide), com grande variação alélica (Cantarella et. al., 2010), e pela ausência de um genoma sequenciado.

Programas de melhoramento de cana-de-açúcar buscam estratégias para modificar traços físicos, ou fenotípicos, com o objetivo de produzir uma planta com maior acúmulo de sacarose, tolerância a estresses ou maior produção de biomassa. Um dos principais meios para a obtenção de novos cultivares com as características desejadas está na identificação e caracterização de genes alvos.

É possível obter uma melhor compreensão da associação genótipo-fenótipo se há o conhecimento dos sistemas metabólicos que estão envolvidos na expressão dos fenótipos (Bailey et. al., 2002). Estes sistemas envolvem múltiplas reações metabólicas e são agrupadas dentro de módulos funcionalmente distintos chamados vias metabólicas (Hartwell et. al., 1999). Mudanças das enzimas nestas reações poderiam ter um efeito na expressão do fenótipo, sendo essencial descobrir os genes, proteínas e funções que estão envolvidos neste sistema e integrá-los ao estudo do comportamento das vias de sinalização e metabólicas para diferentes características da cana-de-açúcar.

A integração de dados biológicos de diferentes fontes, para estudo e associação entre respostas genéticas e fisiológicas em diferentes condições de estresse e tratamentos integradas às vias metabólicas, tem potencial para revelar novas funções gênicas e fornecer conhecimento sobre o desenvolvimento da cana-de-açúcar e sua adaptação ao ambiente. A integração desses dados biológicos pode ser representado por redes de dados (Gehlenborg et.al.,2010), que são usadas para descrever vários sistemas biológicos, como as Redes Moleculares, Vias Metabólicas e Vias de Sinalização, as quais são avaliadas e/ou inferidas por diferentes tipos de dados e por suas integrações: como expressão gênica, espectrometria de massas, interação proteína-proteína, atividades enzimáticas entre outros. Por meio destas integrações em diferentes condições experimentais, é possível prever comportamentos celulares e identificar vias ou redes contrastantes e/ou específicas, conforme descrito na literatura (Karp et. al. 1994, 2000; Schaefer et. al. 2004; Sharan, et. al., 2006).

A Biologia molecular, aliada às novas ferramentas biotecnológicas, (microarranjo por DNA/cDNA, ChIP-Seq, RNA-seq, etc), tem permitido um amplo estudo do genoma

de organismos complexos, sendo uma das chaves para a descoberta e entendimento da maquinaria celular e da diversidade biológica. Estima-se que com o projeto transcriptoma da cana-de-açúcar (SUCEST), por volta de 90% dos genes tenham sido identificados, dos quais 35% foram anotados por bancos públicos (Vettore et. al., 2003), tendo proporcionado um caminho alternativo para estudos pontuais do sistema biológico da cana-de-açúcar. Organismos modelos ou próximos evolutivamente são uma alternativa para a transferência de informações, proporcionando uma anotação mais completa do transcriptoma.

No Brasil, nosso grupo tem conduzido diversos estudos de expressão gênica em diferentes tratamentos e condições de estresse de cana-de-açúcar, conduzidos por meio de plataformas de microarranjo de cDNA, destacando-se estresse à seca, acúmulo de sacarose e biomassa, deficiência de fósforo, interação por bactérias endofíticas e resposta a fito hormônios, e estudos em altas concentrações de CO₂ (de Souza et. al., 2008). Esses estudos tem produzido listas de genes expressos, sendo identificados em vias de sinalização e metabólicas envolvidas com o acúmulo de sacarose e estresse a seca (Lembke et. al., 2012; Felix et.al., 2009; Papini-Terzi et.al.,2009, 2007), envolvidos em atividades tecido-específico (Papini-Terzi et.al., 2004), e possível participação de defesa a patógenos e resposta a estresses bióticos e abióticos (Rocha et. al., 2007). Apesar de resultados promissores, estas plataformas contêm menos de 15% dos genes estimados em cana-de-açúcar, e com anotações incompletas e ausência de bases de conhecimento, restringem o desenvolvimento e análises para entender grandes sistemas biológicos através de categorias funcionais e vias, ampliando a especificidade, sensibilidade, e relevância das análises.

Um dos primeiros passos para entender como um dado fenótipo está sendo expresso é entender como os genes relacionados a tal fenótipo tem sua expressão regulada. Para isso, faz-se necessário identificar genes e famílias de Fatores de Transcrição, envolvidos em um sistema biológico relacionado a um tecido específico, um fenótipo ou uma condição de estresse. Além disso, transcritos antisense (ou seja, complementares ao mRNA codificador) tem sido identificados como possíveis reguladores da transcrição e envolvidos em diferentes condições de estresse (Matsui et.al., 2008), e a detecção destes transcritos poderá ampliar o entendimento do mecanismo regulatório, relacionado a diferentes fenótipos agrônômicos, como tem sido identificado em Milho (Ma et. al., 2006), Trigo (Coram et. al., 2009) e Arabidopsis (Henz et. al., 2007).

O uso de novas tecnologias geram dados de genoma, transcriptoma, proteoma e metaboloma tem proporcionado novos caminhos para o estudo dos sistemas biológicos.

Estas tecnologias têm produzido um grande volume de dados, porém produzido alta redundância, baixa qualidade, escassez de bases de conhecimento, e anotações incompletas; aliado às limitações de abordagens baseadas puramente em similaridade, tem exigido o desenvolvimento de bancos de dados, aplicação de técnicas e metodologias para analisar e integrar estes dados, tornando-se um desafio computacional.

Deste modo, para permitir o estudo e melhoramento da cana-de-açúcar, com a identificação e interação de genes alvos, categorias funcionais e vias envolvidas em diferentes condições e tratamentos foram propostas novas metodologias, para uma anotação funcional do transcriptoma mais precisa, para os estudos de expressão gênica por larga escala, e o desenvolvimento de uma plataforma baseada em bancos de dados relacionais e de grafos, consolidando inúmeros bancos, com diferentes aplicações para extrair e interpretar a biologia subjacente às medidas moleculares de larga escala e para maior compreensão dos processos biológicos e possíveis modulações do sistema biológico.

2. Motivação

A heterogeneidade e diversidade de dados pode ser um fator limitante em análise de dados biológicos em larga escala, como tem sido discutido em vários estudos e revisões (Lysenko et. al., 2010; Fukushima et. al., 2009; Joyce et. al., 2006; Giardine et. al., 2005; Kitano, 2002). Estudos recentes têm sugerido que para um aumento da produtividade agrícola é necessário um maior entendimento da biologia molecular e da sistêmica das plantas, o qual está vinculado à integração de dados de diferentes "ômicas" (Mochida et. al., 2010). De modo análogo, a predição de anotações funcionais para novos transcriptomas ou transcriptomas parcialmente anotados está limitada à quantidade e qualidade de informação disponível e principalmente da disponibilidade de organismos próximos evolutivamente.

A principal coleção de ESTs (Expressed Sequence Tags) de cana-de-açúcar (SUCEST) obtida a partir de diferentes tecidos e diferentes partes da planta (Vettore et. al., 2001), está agrupada em 43.141 possíveis transcritos únicos (SAS – *Sugarcane Assembled Sequences*), apresenta por volta de 35% de novos transcritos, não identificados em bancos públicos (Vettore et.al., 2003) e menos de 30% anotados funcionalmente. Aliado à ausência de um genoma sequenciado e à falta de um mapeamento das vias metabólicas em cana-de-açúcar, a integração com os sistemas metabólicos, em particular com as vias metabólicas, tem se tornado um grande desafio para estudos mais abrangentes e identificação de vias e genes de interesse da cana-de-açúcar. Alguns trabalhos do nosso grupo têm atualizado a anotação desses transcritos, como o Catálogo SUCAST, com 3.563 SAS envolvidos em vários aspectos de transdução de sinal, e o Catálogo SUCAMET, com cerca de 5.606 SAS envolvidos com metabolismos de carboidratos, tendo proporcionado estudos de expressão gênica por microarranjo em diferentes condições experimentais (Souza et.al., 2001 e Papini-Terzi et.al., 2005, 2007). Alguns experimentos compararam variedades contrastantes para a tolerância ao déficit hídrico e foram realizados diagnósticos moleculares (por microarranjo) e fisiológicos (medidas em campo e laboratório) para diferenciar as variedades.

A partir do transcriptoma do projeto SUCEST, e através do sequenciamento do transcriptoma de diferentes variedades de cana-de-açúcar utilizando tecnologias de nova geração, será possível realizar uma nova montagem *De novo* do transcriptoma e produzir informações novas e mais precisas para melhorar a anotação e categorização funcional, a identificação de novos genes e estudos de domínios de proteínas preditas, além da identificação de homólogos a partir do transcriptoma de outras espécies. Estas informações permitirão identificar uma melhor relação entre as vias metabólicas, as

proteínas e os genes. Uma importante classe de genes a ser estudada e classificada em famílias para as gramíneas são os fatores de transcrição (TFs – *Transcription Factors*) que funcionam como reguladores da transcrição e auxiliam na ativação ou inibição de um gene (Schwechheimer et. al., 1998).

A anotação funcional com as análises por expressão gênica é essencial para o estudo e entendimento de sistemas biológicos. Entre as aplicações possíveis podemos destacar o entendimento de como as células reagem ao um tratamento específico ou condição ambiental, e a identificação de genes responsivos a estes tratamentos e condições. Atualmente existem diversas ferramentas computacionais para atacar inúmeros problemas biológicos, porém muitas destas ferramentas são específicas, desenvolvidas para organismos específicos, ou requerem conhecimentos técnicos específicos e exigem uma longa curva de aprendizado. A anotação funcional permite um melhor aproveitamento do estudo de expressão gênica por microarranjo, porém as plataformas existentes em cana-de-açúcar apresentam cerca de 18% dos genes, e menos de 15% de TFs catalogados (Yilmaz et. al., 2009). Sendo que a regulação da transcrição é o primeiro passo da expressão gênica e pode afetar diferentes “ômicas” (proteômica, metabolômica e fenômica), a análise funcional dos TFs é importante e necessária para esses estudos e para a elucidação das redes funcionais em plantas (Mitsuda et. al. 2009), principalmente pela ausência de uma classificação sistemática para cana-de-açúcar e diversas outras gramíneas. O desenvolvimento de uma plataforma de integração para dados de larga escala, integrado a uma anotação funcional mais precisa e abrangente, uma classificação de famílias de TFs e experimentos de expressão gênica em diferentes condições experimentais, permitirá um maior entendimento do sistema biológico e suas interações na cana-de-açúcar. Entretanto, para as análises em uma nova plataforma será necessária a criação de *pipelines* de análise e novos métodos para anotação funcional, avaliação e integração com estudos de expressão gênica por microarranjos.

A integração de dados heterogêneos inseridos em contextos de conhecimento biológico permite a inferência e identificação de relações biológicas em maior profundidade, porém exige novas abordagens computacionais para a análise simultânea destes dados. As informações atualmente encontram-se dispersas e exigem uma tarefa manual árdua de reunir toda a informação de múltiplas plataformas, que utilizam formatos e anotações inconsistentes. Uma importante estratégia na integração de dados heterogêneos pode ser a abordagem baseada em vias metabólicas, com a conexão do nível molecular com diferentes condições de estresse da planta. No caso da análise de vias metabólicas ou funções específicas relacionadas ao estresse hídrico pode indicar

que certos processos moleculares são responsáveis pelo maior acúmulo de sacarose ou produção de biomassa. Um dos desafios para alcançar esse objetivo implica em não somente decidir quais métodos poderiam decifrar os mecanismos moleculares envolvidos entre estresses e fenótipos, mas envolve também o uso de vias e funções apropriadas a nível molecular para poder capturar informações relevantes, porém, como relacioná-las continua um problema em aberto.

3. Estado da Arte - Fundamentos Biológicos e Computacionais

3.1. A Cana-de-açúcar

A cana-de-açúcar (*Saccharum spp.*) é uma das principais culturas agrícolas do Brasil, e a quinta principal cultura produzida mundialmente; com importância mundial na produção de açúcar, biocombustíveis e como alternativa para produtos em substituição a derivados do petróleo. O Brasil responde por 40% da produção mundial de açúcar bruto (Faostat, 2014), e com a produção de etanol dos Estados Unidos a partir do milho, foram responsáveis por 88% da produção mundial em 2009 (Cheng M. et. al., 2013; Muñoz I et.al., 2014). A cana-de-açúcar possui maior vantagem entre as culturas com potencial energético para produção, tais como elevada produtividade, produção em curtos períodos de desenvolvimento, crescimento sob condições adversas, extração de açúcares diretamente do suco do colmo, ao contrário do amido de milho, que deve ser liquefeito e hidrolisado para a extração dos açúcares, além da baixa necessidade de nutrientes (Waclawovsky A. J. et. al., 2010). Além da importância econômica, a cana-de-açúcar tem impacto significativo na matriz energética nacional, uma vez que seus derivados compreendem cerca de 18% da energia consumida no país (Unica 2012). Com os investimentos para o estudo e melhoramento genético da cana-de-açúcar, espera-se a criação de novas variedades, e o desenvolvimento de novas tecnologias na produção de etanol celulósico (etanol de 2ª geração), levando à maior produtividade no campo e minimizando a expansão de novas áreas plantadas (Buckeridge M. S. et. al., 2012).

Algumas das culturas mais produtivas na agricultura, como cana-de-açúcar (gênero *Saccharum*), sorgo (*Sorghum bicolor*), e milho (*Zea mays*), pertencem à superfamília Panicoideae, uma das linhagens de gramíneas C4 (Laetsch W. M. et. al., 1974; Hatch M.D. et. al., 1976). As plantas C4 apresentam maior fixação de carbono e possuem maiores vantagens em relação às plantas C3, como em condições de seca, alta temperatura, fixação de nitrogênio e limitação de CO₂ (Sage R. et. al., 1999). Dada a importância do metabolismo C4 para produtividade das plantas, já existem projetos que visam alterar o metabolismo das plantas, como o projeto *C4 Rice Project* que busca tornar o arroz, uma planta C3, em uma planta C4, possibilitando aumentar em até 50% a produtividade de grãos, com menor quantidade de água e nutrientes (Von C. S. et. al., 2012; Hibberd J. M. et. al., 2008).

A cana-de-açúcar pertence à família Poaceae, é uma monocotiledônea C4 com capacidade de acumular até 42% da sua biomassa na forma de sacarose (Moore P,

1995). Os cultivares modernos de cana são derivados de genótipos interespecíficos, de cruzamentos entre duas espécies poliploides (Figura 1): *S. officinarum* (2n=80) e *S. spontaneum* (2n = 40-128), com menores contribuições de outras espécies do mesmo gênero, principalmente *S. robustum* (2n=60,80), *S. sinense* (2n = 111-120) e *S. barberi* (2n = 81-124), além de outros gêneros, especialmente *Erianthus* e *Miscanthus*. Os cultivares modernos são derivados de programas de melhoramento realizados em Java e Índia no final da década de 1990, cruzando *S. officinarum* e *S. spontaneum* com objetivo de produzir variedades com características de maior vigor e resistência a estresses bióticos e abióticos, maior capacidade de acúmulo de sacarose e menor teor de fibras (Paterson A. et. al., 2013; Bremer G., 1961; D'Hont A. et. al. 1998; Grivet L. et. al., 2001). Essa complexa organização genômica faz com que cada cruzamento seja um evento imprevisível, devido à existência de múltiplos e distintos alelos para cada *locus* que, por sua vez, podem ocorrer em diferentes doses (Garcia A. A. F. et. al., 2013). Assim, o estudo e melhoramento da cana-de-açúcar por meio da diversidade genética e de sua adaptação ao ambiente é um grande desafio, principalmente pela ausência de um genoma sequenciado, com alta complexidade (poliploide e aneuploide), grande variação alélica, alta taxa de elementos transponíveis (De Setta N. et. al., 2013), e um transcriptoma incompleto e parcialmente anotado (Vettore A. L. et. al., 2003).

Figura 1. Representação diagramática do genoma da cana-de-açúcar moderna. As linhas representam os cromossomos homólogos, as barras em branco os cromossomos herdados de *S. officinarum*, as barras em cinza os cromossomos herdados de *S. spontaneum* e as barras com duas cores as recombinações cromossômicas entre as duas espécies ancestrais.

Com o sequenciamento do genoma da cana-de-açúcar, será possível o entendimento da genética e contribuir para o desenvolvimento de ferramentas moleculares para propósitos de melhoramento, descoberta de genes e marcadores moleculares. Análises do sequenciamento de BACs (*Bacterial Artificial Chromosome*) em duas variedades distintas RB92-5345 e SP80-3280 têm mostrado resultados promissores, como alta colinearidade e sintonia a nível gênico, alta taxa de elementos transponíveis, regiões intergênicas e intrônicas não caracterizadas, e uma expansão do genoma em relação ao genoma de sorgo (Figueira T. R. S. et. al., 2012; De Setta N. et. al., 2013). A obtenção de um genoma referência será essencial para estudos moleculares de redes gênicas, integrados ao transcriptoma e identificação de regiões promotoras e elementos regulatórios, para melhoramento da cana-de-açúcar e desenvolvimento de plantas transgênicas. O sequenciamento do genoma das variedades R570 e SP80-3280 da cana-de-açúcar estão em andamento através de tecnologias de sequenciamento por BACs e WGS (*Whole Genome Shotgun*) (Dal-Bianco M. et. al., 2011).

3.2. Anotação Funcional do Transcriptoma de Cana-de-açúcar

A genômica funcional é uma área da biologia molecular que permite o estudo em ampla escala das funções e interações de genes e proteínas, buscando definir a função de um gene no contexto do genoma de um organismo (Paterson et. al., 2009). Uma das principais características da genômica funcional no estudo da função gênica é a utilização de abordagens e métodos de larga escala. O sequenciamento de novos genomas e transcriptomas de diferentes organismos têm crescido exponencialmente (Figura 2), principalmente com o advento das novas tecnologias de sequenciamento de larga escala, como *454 sequencing*, *Illumina HiSeq*, *Ion Proton*, *PacBio*, porém, atribuir funções aos genes tem se tornado um grande desafio nas pesquisas atuais (Hawkins et. al., 2010).

Figura 2. Evolução do número de sequências de mRNA e genômicas depositadas no GenBank.

A anotação funcional é a ação de caracterizar e anotar um conjunto de genes preditos, obtendo sua provável função biológica, uma regra metabólica ou uma característica estrutural e pode ser separada em várias etapas como descreveremos nos tópicos seguintes. A etapa de anotação permite atribuir um conjunto ou tipo de características para cada produto de um gene, resultando de uma tradução da sequência codificadora, a qual será extremamente útil para estudos específicos de famílias de genes ou genes envolvidos em vias metabólicas específicas.

A anotação e caracterização funcional de novos genomas e genes são feitas de forma automática, sem um estudo experimental direcionado ou por meio de uma curadoria manual, gerando um problema de anotação e na qualidade biológica da informação. As funções de muitos genes e proteínas são ainda desconhecidas, ou parcialmente definidas.

Um dos principais objetivos deste trabalho é a análise funcional do transcriptoma da cana-de-açúcar e a identificação de fatores de transcrição em gramíneas e cana-de-açúcar. Os estudos tem se restringido às gramíneas, incluindo *Arabidopsis thaliana* externa a este grupo por apresentar-se como organismo modelo, a maior parte das gramíneas não possuem genoma completo sequenciado ou estão em fase inicial de anotação dos genes.

Sorghum bicolor é um dos organismos mais próximos evolutivamente de cana-de-açúcar e apresenta um genoma parcialmente sequenciados (Figura 3), e possui cerca de 35% dos genes anotados como hipotéticos ou desconhecidos (Paterson et. al., 2009). Organismos modelo como *Drosophila*, ainda permanecem com cerca de 40% dos genes sem função determinada, mostrando a complexidade na anotação e na necessidade da

integração de diferentes abordagens para estudo e anotação destes transcriptomas (Roy et. al., 2010). A cana-de-açúcar possui relativamente pouca informação de anotação relacionada (cerca de 65% do transcriptoma está anotado como hipotético ou desconhecido) (Nishiyama-Jr et. al., 2013), e com o sequenciamento de outras gramíneas próximas evolutivamente de cana, será possível a atualização desta anotação para uma interpretação e caracterização do transcriptoma de cana-de-açúcar mais completa.

Figura 3. Árvore construída a partir de organismos pré-selecionados, baseado na classificação taxonômica do banco de dados do NCBI (Fderhen, 2012).

Apesar do aumento exponencial de sequências sendo produzidas e depositadas em banco públicos (GenBank –NCBI), têm-se empregado métodos tradicionais de anotação, através da buscas por similaridade, como o software de alinhamento BLAST (Altschul et. Al., 1997). Porém, existem vantagens e desvantagens na utilização desta abordagem. As principais vantagens são a velocidade, alteração dos parâmetros de busca e simplicidade na anotação automática, baseadas em bancos de dados públicos ou particulares previamente anotados. Uma das grandes desvantagens está na propagação de erros de anotação, que são crescentes e problemáticos em bancos de dados biológicos. Este tipo de problema pode ocorrer entre sequências similares que não são ortólogas, como casos de sequências parálogas ou domínios casuais, os quais

não caracterizam tal função ou ortologia a um determinado gene (Friedberg et. al., 2006). Este tipo de erro resulta em um problema em potencial: ciclos de interdependência entre bancos de dados que podem induzir a um ciclo perpétuo de dados imprecisos ou incorretos (Buneman et. al., 2000), e assim, utilizar conhecimento de bancos de dados públicos pode ser perigoso sem o histórico completo de cada propriedade e evidências experimentais.

Tratando-se de organismos não-modelos, como cana-de-açúcar, é preciso cautela no processo de anotação e quais bancos de dados são utilizados como referência. Os organismos não-modelos são organismos que não são estudados extensivamente pela comunidade científica, por razões históricas, ou pela complexidade ou falta de características que os tornem fáceis de serem estudados e manipulados, como a falta de reprodutibilidade em laboratório, longo ciclo de vida e genética complexa, tornando ainda mais lenta a descoberta de função de seus genes. Organismos modelos como *Arabidopsis thaliana*, *Oryza sativa*, *Saccharomyces cerevisiae*, *Drosophila melanogaster* e *Mus musculus*, são organismos amplamente estudados e tiveram seus genomas e transcriptomas sequenciados e anotados manualmente. Apesar do grande volume de dados sendo produzidos e depositados em bancos públicos sua maior parte são de bactérias e organismos modelos distantes evolutivamente da cana-de-açúcar, sendo que 33% dos dados depositados no Uniprot/Swissprot são de Eucariotos e desses, menos de 7% são de plantas (Figura 4).

Figura 4. Distribuição do total de proteínas no Reino Eucariota, correspondendo a 33% do Uniprot¹.

¹ Dados de 26-Nov-2014 do UniProtKB/Swiss-Prot.

A atribuição de uma anotação para um gene, principalmente em organismos não-modelos, pode apresentar uma história complexa, envolvendo múltiplos experimentos biológicos, bancos de dados e inferências computacionais. A confiabilidade de qualquer anotação é dependente do histórico completo deste dado, do genoma de referência e de métodos dos quais foi inferido. A anotação computacional, pode depender de múltiplas evidências experimentais e métodos, de múltiplos genes daquela família, em múltiplas espécies. Esta anotação depende também de evidências estatísticas e computacionais, como similaridade entre sequências e co-expressão, além de dados de expressão gênica por microarranjos, expressão por proteoma, estudos de microRNAs, *Single Nucleotide Polimorphisms* (SNPs), anotações manuais por meio de Ontologias (Uma ontologia é uma explicitação específica de algum tópico. Para nosso propósito, esta é uma representação declarativa e formal que inclui um vocabulário (ou nomes) para referir-se aos termos na área de interesse e as declarações lógicas que descrevem o significado dos termos, como eles estão relacionados entre si, e como podem ou não relacionar-se. Deste modo, o termo ontologias fornece um vocabulário para representar e comunicar conhecimento sobre um conjunto de relacionamentos que existem entre termos nesse vocabulário (*Stanford Knowledge Systems Lab*)), como GO (*Gene Ontology*), e disponibilidade de uma plataforma de integração. Por meio da internet é possível encontrar uma grande variedade de ferramentas, pesquisas e bancos de dados disponíveis para análises de similaridades e predições de genes para inferir suas propriedades, porém a maior parte delas não integram diferentes fontes de dados.

A definição de uma função dentro de um vocabulário ou categorização é um pré-requisito para aprimorar as análises estatísticas e consultas função-dirigida. Um vocabulário controlado e não redundante pode permitir este tipo de análises ou através de declarações que explicitamente instituem uma semântica de relacionamento entre termos em uma ontologia (i.e. hierarquias como "é um" e "parte de"). Deste modo, sistemas como Comissão de Enzimas (EC – *Enzyme Commission*) (Bairoch, 1999), FunCat (Ruepp et. al., 2004) e GO (*Gene Ontology Consortium*, 2004) são importantes recursos para definir funções independente da estrutura ou sequência da proteína.

O estudo do transcriptoma em cana-de-açúcar iniciou-se com a produção de catálogos de ESTs (*Expressed Sequence Tags*) por diferentes grupos de pesquisa em vários países (revisado por Nishiyama-Jr et. al., 2013; de Siqueira et. al., 2013; Manners et. al., 2011). O mais importante desses, o projeto SUCEST (Sugarcane EST Project) (Vettore et. al, 2001; Vettore et. al., 2003) financiado em parceria pela FAPESP e COPERSUCAR, gerou 237.954 ESTs a partir de 26 bibliotecas diferentes de cDNA as quais foram agrupadas em 43.141 transcritos (ou SAS - Sugarcane Assembled

Sequences). A partir do SUCEST, foram gerados catálogos de genes para a utilização em experimentos de microarranjo de cDNA na análise da expressão gênica de cana-de-açúcar em diferentes tecidos, variedades ou submetidas a diferentes tratamentos.

3.3. Detecção de Ortólogos putativos funcionais

A cana-de-açúcar apresenta um genoma complexo, com diversas cópias cromossômicas, aliado à presença de grandes famílias de proteínas e com muitas proteínas parálogas, podendo tornar o processo de detecção de verdadeiros ortólogos ancestrais funcionais, através da similaridade por sequências, um desafio. Algoritmos como HIERANOID (Schreiber et. Al., 2013), que é uma atualização do INPARANOID (Östlund et. al., 2010), OrthoMCL (Chen et. al., 2006) e PHOG (Datta et. al., 2009) têm tentado identificar o ortólogo ancestral a partir de múltiplas proteínas em alinhamentos por similaridade. Estudos através dessas abordagens têm demonstrado que é possível identificar ortólogos funcionais, porém estas abordagens necessitam de curadoria manual, e existem também limites para os quais análises filogenéticas podem ser usadas para transferir função. A importância de um resíduo para uma função de uma proteína usando abordagem filogenética requer o alinhamento de múltiplas sequências ortólogas. Em muitos casos existe um número insuficiente de sequências ortólogas, ou poucas com funções que foram caracterizadas experimentalmente para definí-las como ortólogos funcionais, e é muito difícil medir a confiança de que uma dada função pode ser transferida dado um determinado escore de alinhamento por similaridade. Em geral, os critérios de corte de escore de alinhamento são definidos arbitrariamente e a confiança desconhecida (Karp et. al., 1998). Similaridade estrutural entre proteínas têm se mostrado um método alternativo para inferir função, porém evolução divergente e convergente podem produzir proteínas com mesma função mas com diferentes estruturas, e vice-versa (Hegyi et. al., 1999), necessitando deste modo, de uma curagem manual.

Nesta tese é descrito um método e *pipeline* para dados de larga escala, agregando a análise de ortólogos ancestrais com INPARANOID e HIERANOID para transferir função baseado em sequências de organismos próximos evolutivamente de cana-de-açúcar. As sequências que possuem sua função transferida foram definidas como ortólogos putativos funcionais. Em situações em que um grande número de domínios conservados existe, a probabilidade de que o conjunto de sequências alinhadas compartilhe a mesma função é maior, porém, mesmo com múltiplos domínios conservados, divergências

funcionais podem ser encontradas, dependendo de regras associadas a pressões evolucionárias (Hegyi et. al., 2001).

3.4. Expressão de Transcritos senso-antisenso

Nos últimos anos, a importância da regulação gênica por níveis de RNA, tem mostrado-se uma etapa importante nos processos de transcrição (Khraiwesh et. al., 2012). Os avanços das tecnologias em larga escala e novos métodos em bioinformática, tem permitido recentes descobertas da complexidade de moléculas regulatórias de RNA em eucariotos, incluindo a descoberta de RNAs regulatórios não codificadores (*non-coding RNAs*) ubíquos curtos e funcionais longos (ncRNAs) (Ghildiyal et. al., 2009; Bay et. al., 2014), incluindo microRNAs (miRNAs), RNAs curtos de interferência (siRNAs) e *Piwi-interacting RNAs* (piRNAs).

Os transcritos antisenso natural (NATs - *Natural Antisense transcripts*) são moléculas de RNAs endógenos contendo sequências que são complementares a outros transcritos, surgindo como um novo membro dos RNAs regulatórios. Os NATs são classificados em dois grupos dependendo da sua origem genômica: cis-NATs, que são pares de transcritos formados a partir de fitas opostas do DNA, porém derivados do mesmo *locus* genômico e complementaridade exata, e trans-NATs que são pares de transcritos derivados de regiões genômicas distintas e complementaridade parcial (Lavorgna et. al., 2004) (Figura 5).

Figura 5. Representação dos dois grupos de NATs de acordo com sua origem genômica: cis-NATs e trans-NATs. (A) Os cis-NATs são formados no mesmo locus do respectivo par senso, e (B) trans-NATs e seu respectivo par são formados em locus distintos.

Uma grande variedade de estudos e exemplos têm relatado a ocorrência dos NATs em genomas procariotos e vírus, atuando na regulação da expressão gênica afetando a transcrição do mRNA, processamento e tradução (Wagner et. al., 1994). Estudos recentes mostraram que entre 5% e 10% dos genes NATs em mamíferos e plantas são peças chave em processos fisiológicos, patológicos e ritmos circadianos em plantas (Gutzat et. al, 2012; Lapidot et. al., 2006; Zhang et. al., 2013; Hotta et. al., 2013; Edwards et. al., 2006). Em milho os transcritos antisenso representam cerca de 6.5% em antera e 13.5% em polén (Ma et. al, 2006), e em *Arabidopsis* cerca de 3,7% dos pares de transcritos são cis-NATs. Os estudos e acesso em ampla escala de cis-NATs tem se intensificado principalmente com o advento das novas tecnologias de larga escala, como os sequenciamentos de bibliotecas de cDNA/ESTs (Jen et. al., 2005; Wang et. al., 2005) e métodos computacionais para a predição de cis-NATs. A partir destas bibliotecas de cDNA/ESTs foi possível o desenho de oligonucleotídeos para detecção de transcritos antisenso, para o estudo dos níveis de expressão e co-expressão dos NATs em larga escala em diferentes condições e tratamentos, com maior sensibilidade (Chen et. al., 2012; Lembke et. al., 2012). Nos últimos anos, o advento do sequenciamento em larga escala tem permitido a caracterização sistemática do transcriptoma com alta cobertura e maior acurácia que as abordagens convencionais (Mortazavi et. al., 2008) e, deste modo, novos protocolos para sequenciamento por RNA-seq fita-específica têm sido desenvolvidos para caracterizar e identificar transcritos antisensos, principalmente em cana-de-açúcar (Levin et. al, 2010; Passalacqua et. al., 2012; Nishiyama-Jr et. al., 2014).

A maior parte dos estudos de NATs foca principalmente em cis-NATs, motivada pela facilidade de serem identificados, por apresentarem similaridade exata e serem mais estáveis em estudos por plataformas como microarranjos. Porém, diversos estudos têm sido direcionados aos trans-NATs (Røsok et. al., 2004) identificando cerca de 50% de RNAs dupla fita clonados de tecido normal humano sendo trans-NATs. Em plantas, estudos do genoma de *Arabidopsis*, identificaram 1.320 pares de trans-NATs putativos. Um banco de dados amplo com predições de dados para cis- e trans-NATs, em eucariotos, tem sido desenvolvido envolvendo aproximadamente 2 milhões de pares de NATs, cobrindo 69 espécies de plantas, incluindo espécies modelos e de culturas com genomas anotados/curados disponíveis (Chen et. al., 2012).

3.5. Fatores de Transcrição em Gramíneas e Cana-de-açúcar

A visão inicial de que a informação genética flui do DNA para RNA para proteína, conhecida como Dogma Central (Crick, 1958) tem sido estudada e detalhada com os avanços na ciência ao longo dos anos. A produção do RNA a partir do DNA, definido como transcrição, é um processo altamente coordenado, mediado pela RNA polimerase e por um grande número de proteínas para o reconhecimento do promotor e precisa iniciação da transcrição. O complexo básico para a transcrição é formado pela RNA polimerase, pelos fatores de transcrição gerais (ou básicos) (FTG) e proteínas acessórias (Kornberg et. al., 2007; Maston et. al., 2006). Foram identificados quatro tipos de RNA polimerases em eucariotos: a RNA Pol I transcreve a maioria dos RNAs ribossomais; a RNA Pol II transcreve todas as proteínas de genes codificadores; a RNA Pol III transcreve RNAs de transferência, ribossomais e pequenos RNAs nucleares; e a RNA Pol IV, encontrada em plantas, transcreve pequenos RNAs de interferência, os quais estão envolvidos em processos pós-transcricionais, como silenciamento de genes (Zhang et. al., 2007).

O processo de iniciação da transcrição só é possível com a localização correta do sítio de ligação da RNA Pol II no DNA molde, porém ela necessita de uma variedade de componentes auxiliares, para o exato reconhecimento ao redor do sítio de iniciação da transcrição (*TSS – Transcription Start Site*) e estabilidade do complexo da polimerase, como os TFs: TFIIA, TFIIB, TFIID, TFIIE, TFIIIF e TFIIH (Orphanides et. al., 1997) e proteínas mediadoras, que transduzem a informação desde os elementos promotores distais (i.e., *enhancers*) ao aparato basal, além de outros reguladores transcricionais envolvidos na regulação da transcrição atuando no controle do empacotamento do DNA dentro da cromatina (Kornberg, 2007) (Figura 6).

Os estudos da regulação e transcrição gênica, aliadas ao desenvolvimento da biologia molecular e de novas tecnologias têm permitido a identificação e classificação de novos componentes regulatórios, como os TFs e elementos regulatórios em "cis", os CREs (CREs – CYS-regulatory elements). Estudos recentes tem demonstrado que os TFs se ligam aos CREs, que podem estar localizados a centenas de bases de regiões *downstream* ao TSS, regiões UTR (região não traduzidas – *UnTranslated Regions*), em introns, ou mesmo em exons (ENCODE PROJECT CONSORTIUM, 2007 (Birney et. al., 2007; Lee et. al., 2007; Zhang et. al., 2007)).

A ligação de um TF a um CRE pode modificar o nível transcricional de um gene. Se a ligação de um TF a um CRE aumenta o nível de iniciação da transcrição, este CRE é chamado de *enhancer*. No caso de um TF estar ligado a um CRE e se inibe ou diminui o nível de transcrição, este CRE é chamado de silenciador (Figura 5). E se um CRE bloquear o efeito de enhancers ou silenciadores próximos com um TF ligado é chamado de isolador. Os CREs podem ser únicos ou ocorrer em módulos e estão presentes apenas em regiões promotoras de genes que deveriam ser transcritos em resposta a estímulos específicos. Deste modo, regiões promotoras de genes com padrões similares de CREs poderão apresentar níveis de expressão similares e possivelmente serem regulados pelos mesmos TFs. Os TFs são caracterizados por possuírem um domínio específico de ligação-DNA, contendo resíduos de aminoácidos que se ligam a bases de DNA de elementos com ação em cis, determinando a especificidade da proteína (Latchman, 2005). Para entender a especificidade da ligação-DNA de um TF, é necessário recorrer a uma classificação sistemática dos DBDs, que forneça as sequências específicas das interações proteína-DNA. Uma das primeiras classificações dividiu os DBDs em superclasses e classes, famílias e sub-famílias, de acordo com as estruturas que são conhecidas ou podem ao menos ser hipotéticas por homologia de sequência, definindo cinco principais superclasses (Stegmaier et. al., 2004):

- domínio Básico, caracterizados por uma região rica em resíduos de aminoácidos básicos, com conformação α -hélice que interagem com DNA, um exemplo são os bZIPs compostos por domínio básico + zippers de leucina.

- domínio Hélice-volta-Hélice, é uma particularmente grande e heterogênea família de TFs, e estruturalmente consiste de duas regiões α -hélices dispostas em ângulo reto um com o outro, um exemplo são os Fatores de Choque Térmico, os HSF (*Heat Shock Factor*).

- domínio Zinco-coordenando, caracterizada pela formação de uma estrutura chamada dedo de zinco, responsável pela ligação-DNA a sequência específica, exigindo a presença de Zn^{2+} . Um exemplo é o domínio WRKY que é específico de plantas.

- domínios β -scaffold com pequenos encaixes de contato, não possui uma característica estrutural comum a todos os membros, o que torna esta superclasse mais complexa e mais diversa, porém seu modo geral de interação ligação-DNA, consiste na inserção de um pequeno sulco causando uma torção acentuada da dupla hélice do DNA, um exemplo é o TBP (*TATA-binding protein*)

- outros domínios, compreendem todos aqueles DBDs que podem ser agrupados em famílias por homologia de sequência, mas para os quais não pode ser atribuída nenhuma superclasse devido à ausência de uma informação estrutural conhecida.

Um dos primeiros estudos visando à criação de uma classificação de TFs para eucariotos permitiu classificar e subdividir as famílias de acordo com critérios funcionais e estruturais, baseando-se na presença e tipos de DBDs e sendo amplamente aceito na literatura e utilizado para o estudo, definição de novas famílias em gramíneas (Tabela 1) (Yilmaz et. al., 2009) e outras espécies (Riaño-Pachón et. al., 2007; Davuluri et. al., 2003; Wingender, 1997; Heinemeyer et. al., 1998).

Nesta tese também é descrito um método e *pipeline* em larga escala para identificação e classificação de famílias de fatores de transcrição para gramíneas, focando principalmente em cana-de-açúcar.

Tabela 1. Classificação das famílias de fatores de transcrição dentro das superclasses de DBDs de acordo com suas características.

Superclasse de Domínios	Família de TFs
Domínios Básicos	BZR, bHLH, bZIP, EIL, GeBP, TCP
Domínios Hélice-volta-Hélice	ARR-B, E2F-DP, G2-like, HB, HSF, MYB, MYBrelated, Sigma70-like, zf-HD
Domínio Zinco-coordenando	Alfin-like, C2C2-CO-like, C2C2-Dof, C2C2-GATA, C2C2- C2H2, C3H, CPP, GRF, PLATZ, SBP, SRS, VOZ, WRKY, ZIM, LBD, FAR1-like
Domínios β -scaffold com pequenos encaixes de contato	CCAAT, CSD, GRAS, MADS
Outros Domínios	AP2-EREBP, ARF, ARID, BBR/BPC, CAMTA, DBP, LFY, NAC, PBF-2-like, S1Fa-like, Trihelix, TUB, ABI3VP1, Whirly, mTERF, Ovate, NLP, Orphans

3.6. Análise de Expressão gênica em larga escala

O surgimento de novas tecnologias de larga escala em biologia molecular e geração de novas informações genômicas, possibilitou estudos mais completos sobre a estrutura e função do DNA de um organismo. O perfil do transcriptoma é uma das principais etapas para entender a complexidade dos processos biológicos, direcionando os estudos das sequências genômicas para as proteínas funcionais (Lehesranta et. al., 2005). O nível de expressão gênica é utilizado para estimar o nível de mRNA relativo em espécies através da população de RNA das células e tecidos e para explorar os padrões de transcrição (Hughes et. al., 2001). Existe uma diversidade de tecnologias disponíveis para avaliação dos níveis de mRNA em uma célula, e podem ser divididas em duas principais categorias: baseada em hibridação e sequenciamento. A abordagem baseada em hibridação, utiliza a hibridação de oligonucleotídeos ou fragmentos de sequências que são complementares a alguma parte de uma sequência de um gene expresso. O desenho desses oligonucleotídeos requer conhecimento prévio do transcriptoma em estudo. As principais abordagens baseadas em tecnologias por sequenciamento são RNA-Seq (Wang et. al., 2009), SAGE (*Serial Analysis of gene expression*) (Velculescu et. al., 1995), MPSS (Brenner et. al., 2000). RNA-Seq é considerado o sucessor das tecnologias MPSS (*Massive Paralell Signature Sequencing*), SAGE e plataformas de microarranjos, e tem como objetivo sequenciar todo o transcriptoma expresso em determinada condição experimental, seguido por abordagens de transcriptoma *De novo*, e pelo cálculo da quantidade de *reads* que originaram de cada gene, e utilizando este resultado para estimar o nível de expressão gênica.

Os microarranjos de DNA foram desenvolvidos na década de 1990 e consistem em padrões distintos de sequências de DNA (oligonucleotídeos/fragmentos) depositados (imobilizados) de forma ordenada em dezenas de centenas de pontos em um suporte sólido (como lâmina de vidro) (Schena et. al., 1995). O princípio da técnica é a hibridação competitiva entre uma amostra de cDNA tratada e outra controle, marcadas com diferentes fluoróforos. Os cDNAs são marcados fisicamente com a adição de fluoróforos e, após a hibridação das amostras na lâmina, brilham ao serem excitados por luz de comprimentos de onda específicos. Os cDNAs marcados são hibridados com as amostras contidas no microarranjo, e cada ponto do microarranjo emitirá uma intensidade de fluorescência que indica a quantidade de mRNA hibridado com os fragmentos de DNA presentes em cada ponto. Em seguida são extraídos os níveis de intensidade para cada ponto (fragmento sintetizado) na lâmina em cada uma das

condições e os resultados são utilizados para calcular a razão de expressão para cada fragmento/oligonucleotídeo (i.e. tratamento/controle) (Figura 7).

Figura 7. Esquema de um experimento de microarranjo. Fragmentos de um gene, ou oligonucleotídeos são sintetizados e imobilizados em lâminas de vidro. Os cDNAs das amostras de duas condições experimentais são marcados com os fluoróforos Cy3 e Cy5 e hibridizados. Após a hibridação das amostras na lâmina, um escaner é utilizado para obter imagens monocromáticas, em seguida são extraídos os níveis de intensidade para cada ponto na lâmina em cada uma das condições e os resultados são utilizados para calcular a razão de expressão para cada fragmento/oligonucleotídeo (i.e. tratamento/controle).

Com o crescente estudo de transcriptomas e genomas baseado em plataformas de microarranjo, foram criados grandes repositórios para armazenamento destes dados (Kolesnikov et. al., 2014) e *Gene Expression Omnibus* GEO² do National Center for Biotechnology and Information (NCBI) do Instituto de Saúde Nacional dos Estados Unidos. O depósito dos dados nestes repositórios tornou-se requisito para publicação em várias revistas. Este repositório representa para a comunidade científica um recurso único, pelo qual eles podem utilizar os dados sem nenhum custo para suas pesquisas. Estes sistemas para gerenciamento e armazenamento de microarranjos necessitam de padrões, nomenclaturas e ontologias para gerenciar e compartilhar estes dados de

² GEO: <http://ncbi.nlm.nih.gov/geo/>

forma eficiente, sendo criada pela sociedade de dados de expressão gênica para microarranjos (*MGED – Microarray Gene Expression Data*). Existem vários projetos complementares no MGED, incorporados ao sistema de padronização FGED (The Functional Genomics Data Society):

- O projeto MIAME (Informações mínimas sobre experimentos de microrranjo) tem como objetivo definir as informações que devem sempre ser incluídas nas bases de dados e também fornecer diretrizes para os autores e revisores de manuscritos que descrevem experimentos de microarranjo.
- O projeto MAGE (MicroArranjo e Expressão Gênica) visa proporcionar um padrão para a representação de expressão de dados que facilitaria a troca de informações de microarranjo entre diferentes sistemas de dados. Este formado faz parte do projeto FGED, fornece um formato XML padrão, chamado de *MAGE-ML - MicroArray Mark-up Language*, para reportar dados de microarranjos e informações associadas.
- O projeto MGED-OWG (Grupo de Trabalho de Ontologia) tem como objetivo proporcionar termos normalizados para a anotação de experimentos de microarranjo. Estes termos permitirão consultas de elementos dos experimentos. Além disto, os termos permitirão descrições sem ambiguidade de como os experimentos foram realizados. Os termos serão organizados em formas de ontologias, através de classes e propriedades.

Esta tese seguiu principalmente os padrões MIAME e MAGE para o desenvolvimento do banco de dados e ferramentas para armazenamento local dos microarranjos e para geração de pacotes para submissão ao GEO de forma automatizada.

A análise de dados de microarranjo é um processo complexo e multi-fatorial (Slonim et. al., 2002) e a extração dos níveis de expressão requer o desenvolvimento e aplicação de vários métodos, bem como o desenvolvimento de bancos de dados e programas para integração, manipulação e análises. O processamento de microarranjos são divididos em três fases: Fase inicial (i) para correção de *background*, normalização e procedimentos de análise. Na fase de correção do *background*, a intensidade de fluorescência dos oligonucleotídeos é ajustada para correção de ruídos, então, durante a etapa de normalização, a intensidade dos oligonucleotídeos é normalizada para corrigir diferenças na incorporação e intensidade dos fluoróforos e, finalmente, na fase de análise as intensidades dos oligonucleotídeos são convertidas para níveis de expressão gênica; Fase intermediária (ii), que utiliza os níveis de expressão normalizados para a determinação de genes significativamente expressos (SE) e diferencialmente expressos (DE) e, principalmente, análises de vias comparando níveis de abundância em diferentes

condições experimentais, baseado em análises de agrupamento de genes que apresentam padrões de expressão semelhantes, categorias funcionais mais representadas e nível de atividades de vias (Khatri et. al., 2012); Fase final (iii) quando os resultados destas análises são interpretados e/ou validados experimentalmente.

Deste modo, um dos objetivos desta tese é desenvolver e implementar um *pipeline* automatizado integrado à plataforma para processamento de microarranjos nas fases inicial e intermediária, um método para a identificação de genes DE a partir da razão de expressão e um método mais apropriado para oligonucleotídeos com baixa intensidade, como os oligonucleotídeos antisense, para determinar genes significativamente expressos em cada condição experimental. Foram então desenvolvidas ferramentas de fácil acesso, que permitirão extrair o máximo de informação de genes com baixos e altos níveis de expressão, em diferentes condições experimentais por meio de diferentes abordagens para análise do transcriptoma.

3.7. Abordagens de mineração sobre dados de larga escala

Um dos desafios nas análises de dados oriundos de experimentos de genômica funcional é entender os resultados em um contexto de informações estabelecidas como processos biológicos, vias metabólicas entre outros. A compreensão de quais processos biológicos são significativamente modulados sob uma determinada condição ou tratamento é frequentemente mais informativa que a identificação de um conjunto de moléculas individuais que são diferencialmente expressas, e esta compreensão depende de ferramentas eficientes para integração e interpretação de grandes volumes de dados (Curtis et. al., 2005). Esses resultados são extremamente úteis na identificação de genes que podem atuar em determinada condição experimental ou fenótipo. Porém, através desses resultados, não é possível extrair percepções sistêmicas sobre a biologia subjacente na condição experimental em estudo. Deste modo, o surgimento das tecnologias de estudos de perfil em larga escala, apresenta um novo desafio, que é o de extrair significado destas listas de genes diferencialmente expressas. Uma das abordagens para este desafio é a redução da complexidade das análises, através do desenvolvimento de bases de conhecimento biológico para auxiliar em análises agrupando listas de genes dentro de pequenos grupos. Estas bases de conhecimento podem descrever categorias, ontologias, estruturas, componentes e processos biológicos nos quais cada um dos genes está envolvido, além de localização celular e interação entre genes/proteínas. Um exemplo é a identificação de grupos de genes que possuem uma mesma função e/ou atuam em uma mesma via.

Outra abordagem utiliza as análises de larga escala por meio de níveis de expressão a um nível funcional. Em uma primeira etapa, dezenas de centenas de genes SE ou DE em uma condição experimental podem ser agrupados, baseado em ontologias, categorias funcionais ou vias nas quais estão envolvidos, reduzindo a complexidade do resultado para apenas algumas dezenas de categorias ou vias. Em seguida, a identificação de categorias ou vias que diferem e/ou são reguladas entre duas ou mais condições pode produzir novas hipóteses ou idéias, ao invés de uma simples lista de genes ou proteínas expressas (Glazko et. al., 2009), sendo um de nossos objetivos para o desenvolvimento da plataforma.

Existem alguns métodos de análises de vias e categorias utilizando dados de expressão gênica e que podem ser aplicados a outras tecnologias de larga escala. Um dos objetivos do trabalho é construir bases de conhecimento em cana-de-açúcar, baseado em anotações funcionais e ferramentas integradas à plataforma web CaneRegNet. Esta é uma área que tem evoluído com as tecnologias de larga escala e podem ser divididas em três gerações, *ORA (Over-Representation Analysis)*, *FCS (Functional Class Scoring)* e *PT-Based (Pathway Topology Based)* descritas a seguir (Figura 8) (Khatri et. al., 2012).

Figura 8. Visão geral dos métodos de análises de vias usando dados de expressão gênica como exemplo, mas pode ser aplicado a dados de outras tecnologias de larga escala. O dado gerado por um ou mais experimentos e anotações funcionais (vias) é a entrada para os métodos de análise.

- Abordagens de análises de super-representação (*ORA – Over-Representation Analysis*), surgiram com a necessidade de analisar funcionalmente dados de expressão

gênica, e avançaram no nível das análises de experimentos de microarranjo de genes para conjuntos de genes relacionados a categorias ou vias. Esses métodos agrupam todos os genes que são anotados em algumas categorias funcionais, permitindo ao pesquisador fazer uso de conhecimento previamente acumulado, proporcionando as análises de dados de microarranjo dirigido a bases de conhecimento, como Gene Ontology (Ashburner et. al., 2000), KEGG (Ogata et. al., 1999), ou a categorização funcional de cana-de-açúcar (Lembke et. al., 2012). Esses métodos são conhecidos na literatura como “método da tabela 2x2” (Goeman et. al., 2007), e consistem na busca por uma categoria ou via super-representada em uma lista de genes diferencialmente expresso, ou, equivalentemente, uma super-representação de genes DE entre os genes na categoria ou via. Este processo é repetido para uma lista referência de genes apropriada (i.e., todos os genes medidos em um microarranjo). Em seguida, cada categoria é testada para representação de induzida ou reprimida na lista inicial de genes. Os testes mais utilizados são baseados em hipergeométricas, qui-quadrado ou distribuição binomial, etc. Apesar do grande número de ferramentas e enorme uso, ORA tem diversas limitações. A primeira, é que os testes estatísticos são independentes das medidas de expressão. Isto significa que não são consideradas as intensidades dos oligonucleotídeos, tratando cada gene de forma idêntica e que seriam independentes entre si. ORA tipicamente utiliza listas de genes DE e descarta os outros.

- Abordagens de escore de classes funcionais (*FCS – Functional Class Scoring*), propõem a hipótese de que apesar de grandes mudanças em genes individuais podem ter efeito significativo em uma categoria ou via, genes com pequenas mudanças de forma coordenada em conjuntos de genes funcionalmente relacionados (i.e. vias) podem também apresentar efeitos significativos. A FCS possui diversas vantagens em relação ao ORA. Em geral, os métodos do FCS utilizam todas as medidas moleculares disponíveis para as análises das vias, ao invés de utilizar apenas listas de genes, restritas a critérios de nível de expressão. Ao contrário da ORA que não considera os valores de expressão, esses métodos utilizam esta informação para poder detectar mudanças coordenadas na expressão de genes nas mesmas vias. E outra vantagem é que considerando as mudanças coordenadas na expressão gênica, levam esta informação em consideração para a dependência entre os genes em uma via. Em geral, com algumas exceções (Goeman et. al., 2006), esses tipos de abordagens possuem metodologias semelhantes (Ackermann et. al., 2009), consistindo das seguintes etapas: Primeiro, o cálculo da estatística nível-gene utiliza os valores moleculares de um experimento, calculando o nível de expressão de genes ou proteínas individualmente. Estatísticas no nível-gene incluem correlação de medidas moleculares com fenótipo (Pavlidis et. al., 2004), ANOVA

(Al-Sahrour et. al., 2005), *Z-score* (Kim et. al., 2005), entre outras. Em seguida, realiza-se uma transformação na estatística do nível-gene (valor absoluto, ranqueamento, etc); porém esta estatística nível-gene pode reduzir a quantidade de genes em estudo, devido a esta identificação de genes enriquecidos em um grupo (Ackermann et. al., 2009). Segundo, a estatística do nível-gene em uma categoria é agregada dentro de uma única estatística do nível- categoria. Essas estatísticas podem ser multivariada (Kong et. al., 2006) e identificar interdependência entre os genes, ou podem ser univariada (Tian et. al., 2005) e desprezar a interdependência entre genes. As estatísticas de nível-vias podem incluir estatística Kolmogorov-Smirnov (Subramanian et. al., 2005), soma, média, ou mediana da estatística do nível-gene (Jiang et. al., 2007), entre outros. Independente, o poder do tipo de estatística do nível-vias pode depender da proporção de genes diferencialmente expressos em uma via, o tamanho da via e a taxa de correlação entre os genes na via. Por fim, a avaliação da significância da estatística nível-via. Uma das limitações nessas abordagens, é que elas analisam as vias de forma independente, pois um gene pode estar em mais de uma categoria, significando que as categorias podem ter genes em comum. Consequentemente, em um experimento, podemos ver uma via afetada, ou podemos observar um conjunto de vias sendo afetadas devido ao conjunto de genes comuns. Muitos métodos mudam de expressão gênica para a estatística do nível-gene (p.e., ranqueamento) e descartam a variação do nível de expressão real entre os genes e consequentemente entre as vias, nosso método desenvolvido baseia-se principalmente nos níveis de expressão real de cada gene.

- Abordagens baseadas em Topologia de Vias (*PT-Based – Pathway Topology Based*) utilizam o número e tipo de interações entre produtos de genes, que podem ou não ser parte de um banco de vias. As abordagens ORA e FCS consideram apenas a informação dos genes nas vias para identificar as vias de interesse, ignorando informações adicionais destas bases de conhecimento, como compostos, reações enzimáticas e sinalizações. Independente da interação entre os genes dentro de uma via, ORA e FCS produzirão os mesmos resultados, enquanto que *PT-Based* tem o objetivo de utilizar estas informações adicionais, utilizando, deste modo, a topologia das vias para calcular as estatísticas dos níveis dos genes (Draghici et. al., 2007), ou abordagens calculando Fatores de Impacto analíticos, considerando a estrutura e dinâmica de toda a via e incorporando mudanças nos níveis de expressão (Khatri et. al., 2007). Uma das limitações é a complexidade de generalização dos métodos, e a topologia verdadeira é dependente do tipo de célula devido aos perfis de expressão gênica célula-específica e condição em estudo. Outras limitações incluem a inability de modelar estados

dinâmicos de um sistema e a inability de considerar interações entre vias devido à ausência de informação de relação entre vias.

Com o objetivo de implementar uma estrutura para armazenar as interações das vias metabólicas entre os produtos de genes, dados de expressão gênica em diferentes condições experimentais e informações de anotação funcional, utilizamos como base o Neo4j (<http://neo4j.org>), que é um sistema de gerenciamento de dados em grafos, e uma ferramenta para análises a partir deste sistema.

3.8. Sequenciamento de Nova-Geração

O sequenciamento de nova geração é a tecnologia de ponta aplicada na genômica funcional e predomina atualmente nos estudos de expressão gênica, superando várias desvantagens dos microarranjos (Shendure et. al., 2008). Sequências mais longas são geradas a um menor custo por base em relação a outras tecnologias, como o sequenciamento de Sanger. Além disso, apresenta várias outras vantagens como medidas de grande alcance dinâmico (abrangendo mais de cinco ordens de magnitude), baixo ruído de fundo, baixa quantidade de amostras necessárias, possibilidade de medir o nível de expressão de transcritos e suas isoformas e detectar novos transcritos mesmo na ausência de genoma sequenciado (Wang et. al., 2009).

O corrente avanço nas tecnologias de sequenciamento tem provado ser uma das mais bem sucedidas abordagens atuais, evidenciada na implementação dessa tecnologia em vários produtos comerciais. Esses produtos, definidos na literatura como plataformas de nova geração, ou mais recentemente como plataformas de Sequenciamentos de Segunda Geração, incluindo o *454 Genome Sequencer* (Margulies et. al., 2005) e *Solexa technology* (Fedurco et. al., 2006) e a *SOLiD platform* (Shendure et. al. 2008). Mais recentemente, surgiu o termo Sequenciadores de Terceira Geração com o advento dos sequenciadores de uma molécula (Rusk et. al., 2009). Desses sistemas, os mais interessantes são o *Heliscope Single Molecule Sequencer* (Braslavsky et. al., 2003), a plataforma de sequenciamento *Single Molecule Real Time* (SMRT) da *Pacific Biosciences* (Eid, et al., 2009), e a plataforma *nanoball* da *Complete Genomics* (Drmanac et. al., 2010).

Os avanços nas tecnologias de sequenciamento de DNA permitem novos *insights* a partir das sequências transcritas de genomas e anunciam uma nova era no estudo funcional e estrutural do genoma e na regulação gênica (Graveley et. al., 2008). Eles permitem quantificar a expressão gênica, direcionar como o *splicing* alternativo afeta a função de proteínas, e determinar a função de genes não codificadores, especialmente

novos genes (Graveley et. al., 2011). Sequenciamento de RNA (RNA-seq), sequenciamento de pequenos RNAs (*Small RNA-Seq*) e imunoprecipitação da cromatina seguido por sequenciamento (ChIP-seq) são alguns dos métodos aplicados no sequenciamento de nova geração (Bellingham et. al., 2012; Mortazavi et. al., 2008).

Para conduzir um experimento de RNA-seq são necessários alguns passos descritos a seguir. A primeira etapa no sequenciamento em larga escala é a preparação da amostra. O procedimento depende da tecnologia empregada, mas as características principais são as mesmas: extração do RNA codificador (mRNA), transcrição reversa para cDNA e fragmentação para tamanhos definidos pelo protocolo e amplificação (Wang et. al., 2009). O sequenciamento de RNA-seq *fit* específica está substituindo o método de sequenciamento tradicional como uma abordagem para acessar informações mais precisas do transcriptoma. Dados de *fit* específica fornecem resultados mais detalhados e precisos que os dados não *fit* específica na estimativa dos níveis de expressão de genes *sense* e *antisense* e na montagem de transcritos. Mesmo genomas bem anotados precisam de atualizações e correções que podem ser obtidas usando dados *fit* específica (Sigurgeirsson et. al., 2014).

Um problema para sequenciamento de nova geração é o curto tamanho dos *reads*, podendo resultar em maior fragmentação e *contigs* menores, especialmente em regiões contendo regiões repetitivas, que podem diminuir a qualidade dos *reads*. Para superar esses desafios, algumas plataformas de NGS fornecem protocolos de sequenciamento de *paired-end* (PE), que utilizam fragmentos de até 400pb, e *mate-pair* (MP), que utilizam fragmentos de até 15kb, nas quais ambas as pontas do fragmento de DNA são sequenciadas, em oposição ao sequenciamento *single-end* onde apenas uma ponta é sequenciada (Figura 9). *Reads paired-end* resultam em um alinhamento superior através de regiões contendo seqüências repetitivas e produz *contigs* maiores para sequenciamento *de novo*, permitindo fechar *gaps* em seqüências consenso, resultando em uma cobertura global.

Figura 9. Utilizando a combinação de tamanhos de inserções curtas e longas com sequenciamento *paired-end* resulta em máxima cobertura para montagens *de novo*. *Reads* de inserções longas podem cobrir de forma mais eficiente regiões repetitivas. *Reads* de inserções curtas sequenciados com alta cobertura podem preencher *gaps* deixados pelos *reads* longos.

A plataforma de sequenciamento Roche/454 foi a primeira tecnologia de NGS desenvolvida (Moorthie et. al., 2011; Roche/454 Sequencing Technology³, 2012), porém foi descontinuada em 2013. Este método de sequenciamento é baseado no sequenciamento pelo princípio de pirosequenciamento. O sequenciamento é feito pela síntese da fita de DNA complementar para cada oligonucleotídeo em uma placa. Durante o processo de sequenciamento, as bases de nucleotídeos (A, C, T, G) são adicionadas uma a uma sequencialmente. A fita de DNA é estendida para cada nucleotídeo adicionado, baseado no princípio da complementariedade, adicionado pela DNA polimerase à fita que está sendo sintetizada. A cada incorporação de base ocorre uma reação enzimática em que um íon de fosfato inorgânico é liberado e um sinal de luz é emitido e capturado pela câmera de captura. No caso de sequências repetitivas, vários nucleotídeos são adicionados durante um ciclo e a intensidade da luz emitida torna-se proporcionalmente mais forte. Nucleotídeos não incorporados são descartados e um próximo ciclo de reação ocorre. Esta reação é repetida múltiplas vezes, até que todas as sequências de oligonucleotídeos sejam definidas.

As plataformas de sequenciamento *Illumina* e *Ion Torrent* utilizam uma abordagem de sequenciamento por síntese, baseado em uma pequena diferença no mecanismo de detecção da informação da sequência do nucleotídeo (Illumina Sequencing Technology⁴, 2012; Moorthie et. al., 2011). A plataforma *Ion Torrent PGM*, desempenha um sequenciamento semicondutor que se baseia na detecção de alterações de pH induzida pela liberação de um íon de hidrogênio mediante a incorporação de um nucleotídeo numa cadeia de DNA crescente (Quail et. al., 2012). Em contraste, na plataforma *Illumina* os quatro tipos de nucleotídeos são adicionados ao mesmo tempo. Cada um, porém, é marcado por um fluoróforo diferente e tem um grupo de terminação que previne a extensão da cadeia para mais que um nucleotídeo em um mesmo ciclo. Depois que cada cadeia sobre a placa é estendida por um nucleotídeo, os nucleotídeos não incorporados são removidos e os nucleotídeos incorporados são determinados pela imagem da cor. Em seguida, marcadores fluorescentes e grupos terminadores são removidos e a reação de sequenciamento está pronta para o próximo ciclo, e os ciclos são repetidos até que as sequências sejam determinadas.

Uma das principais vantagens das plataformas de sequenciamento em larga escala está nas várias aplicações práticas nos estudos genéticos e genômicos. Para cada

³ <http://454.com/products/technology.asp>

⁴ <http://www.illumina.com/systems.html>

uma destas aplicações existe um conjunto de métodos de análises de acordo com os dados para abordar as questões biológicas subjacentes. Este conjunto de ferramentas de análise varia desde a estimativa da qualidade das bases, montagem dos *reads* em grandes *contigs* (transcriptoma ou genoma), e mapeamento dos *reads* em sequências referência para detectar variação estrutural, avaliar expressão do transcrito, e realizar estudo do perfil da variação de nucleotídeos pontuais (*SNPs* – *Single Nucleotide Polimorphisms*) (Morozova et. al., 2009). A maior parte das tecnologias de sequenciamento de larga escala produz os *reads* brutos no formato FASTQ, cujo escore de qualidade das bases segue o mesmo padrão do *Phred*, mas com carácter ASCII adicionando +64. Para uma revisão completa sobre formato FASTQ e CSFASTQ da ABI Solid (em espaço-cor, não base ou espaço sequência), vejam em (Cock et. al., 2010). O sequenciamento de RNA-Seq para organismos não-modelos são úteis principalmente para montagem *De novo*, além de análises comentadas anteriormente, porém *softwares* de montagens aplicados a sequenciamentos por Sanger não são apropriados, devido ao grande volume de dados e tamanhos das sequências. Para *reads* longos de Sanger, o processo recai na sobreposição dos *reads* para gerar uma sequência consenso mais longa (*contig*). Sequências curtas, porém não se ajustam aos métodos baseados no *layout* de consensos de sobreposição (Zerbino, et. al., 2008). Uma nova metodologia baseada em *reads* curtos foi introduzida, com a adaptação dos grafos de Bruijin para representar e organizar a relação entre os reads usando uma abordagem de caminhos Eulerianos para montar as sequências dos *reads* (Pevzner et. al., 2001), e vários outros grupos tem investigado e produzido novos *softwares* para montagem de *reads* curtos e em grande quantidade. Para a montagem do transcriptoma de cana-de-açúcar com *reads* de sequenciamento de larga escala utilizamos o *software* Trinity (Grabherr et. al., 2011), mas existem diversas alternativas para montagens de transcriptoma e genoma *De novo* (Trapnell et. al., 2010; Simpson et. al., 2009; Shendure et. al., 2008; Bokhari et. al., 2007; Zerbino et. al., 2008).

Alinhamento de *reads* de sequenciamento de larga escala, necessitam de uma estratégia particular, devido ao tamanho, quantidade dos *reads*, acurácia e em termos práticos tempo de execução (Trapnell et. al., 2009). Primeiramente, o uso de tecnologias de sequenciamento de larga escala para descoberta de variantes em genomas completos requer maior acurácia, alinhamento de alta confiança de *reads* curtos para genomas alvo.

Os *reads* de RNA-seq podem ser mapeados em referências com diferentes objetivos; primeiramente, para um genoma completo sequenciado e anotado, com genes preditos, e *reads* mapeados, são usados para calcular níveis de expressão gênica;

no caso de um genoma não anotado, para detectar genes modelos ou inferir novos genes; e, terceiro, um conjunto de genes ou genes condicantes para um genoma desconhecido, o caso de um projeto de sequenciamento e montagem de transcriptoma *De novo*. Existem vários *softwares* para mapeamentos de *reads* curtos, alguns dos primeiros incluem MAQ (Li et. al., 2008a), Mosaik (Stromberg et. al., 2008), SOAP (Li et. al., 2008b), SHRiMP (Rumble et. al., 2009) e Bowtie (Langmead et. al., 2009), com mais recentes atualizações dos algoritmos implementados SOAP2 (Li et. al., 2009a) e o sucessor do Bowtie BWA (Li et. al., 2008a). Esses últimos têm sido atualizados para utilizar o algoritmo BWT, que gerou ganho de velocidade de 30x, para construir uma sequência alvo transformada, tornando-se um dos algoritmos e ferramentas mais utilizadas para mapeamento de *reads* para sequências alvos.

Um particular interesse do uso da tecnologia de sequenciamento de nova geração nesta tese está focado na identificação de novos genes, atualização do catálogo e classificação de fatores de transcrição, desenvolvimento de uma metodologia para identificação de genes completos (*full-length*), análise de transcritos antisenso e aplicação de métodos de análise de expressão em cana-de-açúcar.

3.9. Vias Metabólicas

Os métodos funcionais de anotação deste trabalho têm focado na caracterização das propriedades estrutural ou funcional do gene. Porém, a função biológica na célula é resultado da atividade de diversas vias metabólicas e interações entre moléculas e proteína, e não pode ser atribuído a um simples gene, deste modo foi criado o primeiro banco de vias metabólicas a partir do transcriptoma de cana-de-açúcar e do KEGG.

Uma grande variedade de bancos de dados de redes biológicas tem sido criada. Esses bancos podem ser divididos em bancos de dados de vias e bancos de dados de interações moleculares. Bancos de dados de Vias podem ser classificados como bancos de dados de vias metabólicas e bancos de dados de vias regulatórias, dependendo do foco do conjunto de dados. Esses bancos contêm vias validadas pela literatura ou por outras fontes verificadas por curadores. Existem vários bancos de dados de vias metabólicas, sendo os principais o MetaCyc (Caspi et. al., 2006) e o KEGG PATHWAYS (Kaneshisa et. al., 2000). O MetaCyc possui um banco curado e não redundante de mais de 2.254 vias metabólicas descritas em mais de 2.000 espécies, em sua maioria plantas e microorganismos; as vias metabólicas do MetaCyc são preditas utilizando-se o *Pathway Tools* (Karp et. al., 2002). Os bancos de vias regulatórias como AGRIS (Palaniswamy et. al., 2006), contém informações sobre sequências promotoras de

Arabidopsis, fatores de transcrição da literatura e anotações de genes. O KEGG possui conjuntos de vias metabólicas genéricas e de organismos específicos, e é composta por três principais componentes: espaço genômico (KEGG GENES), o espaço químico (KEGG LIGAND), o espaço da rede (KEGG PATHWAY). O KEGG GENES é uma coleção de catálogo de genes para genomas completos ou parcialmente sequenciados, com informações extraídas de bancos de dados NCBI GenBank e RefSeq (Wheeler et. al., 2007). O KEGG LIGAND é dividido em vários componentes incluindo COMPOUND, GLYCAN, REACTION, ENZYME, entre outros (Kanehisa et. al., 2002). O banco de dados COMPOUND contém estruturas químicas de compostos metabólicos conhecidos, verificados computacionalmente e manualmente. O banco de dados REACTION contém fórmulas de reações enzimáticas. O banco de dados ENZYME contém nomenclatura de enzimas. Cada enzima é identificada por um número EC (*Enzyme Commission*) (Tipton et. al., 2000), que pode ser ligada a outros bancos públicos como UniProt (Bairoch et. al., 2005). Cada reação metabólica na via metabólica do KEGG é ligada com uma entrada no banco REACTION, e as enzimas na reação enzimática podem ser ligadas ao banco no espaço genômico. As vias metabólicas estão classificadas em categorias funcionais: metabolismo, processamento de informação genética, processamento de informação ambiental, processos celulares, sistemas orgânicos e doenças humanas, e uma categoria à parte, relativa a desenvolvimento de drogas. Esse banco de dados é referência na comunidade científica para o mapeamento de proteínas em vias metabólicas (Schmid et. al., 2008; Moriya et. al., 2007; Mao et. al., 2005), e a principal vantagem é sua padronização de nomenclaturas, e integração com outros serviços e bancos de dados disponibilizados pelo KEGG. Outra vantagem é a possibilidade de mapeamentos em larga escala no KEGG, através de ferramentas como o KAAS (*KEGG Automatic Annotation Server*), um sistema web para anotação automática de proteínas, classificando-as em grupos ortólogos e mapeando-as em vias metabólicas (Moriya et. al., 2007).

As vias metabólicas do KEGG são uma coleção de diagramas desenhados manualmente, correspondendo a uma rede conhecida de significância funcional, como no exemplo da representação da via de Glicólise com algumas enzimas identificadas em cana-de-açúcar (Figura 10).

Figura 10. Uma representação parcial de uma via metabólica no KEGG. Os nós são representados por caixas (produtos de genes) e círculos (compostos químicos). Arestas representam interações ou relações moleculares. Em vermelho são os genes de interesse mapeado na Via.

3.10. Plataforma para Integração de dados e estudo do Transcriptoma de Cana-de-açúcar

A implementação de uma plataforma (CaneRegNetDB) e diversos métodos computacionais para integrar experimentos de dados de larga escala e conhecimento biológico de alto-nível da cana-de-açúcar, integrados a dados de outras fontes. O presente projeto centralizará todos os dados do transcriptoma do projeto SUCEST, microarranjos de cDNA SUCAST e SUCAMET e bancos de anotação funcional automática e manual de cana-de-açúcar. Com o surgimento de novas tecnologias como microarranjos de oligonucleotídeos, sequenciamento de larga escala e dados genômicos de outras plantas, torna-se necessário o desenvolvimento de uma plataforma em bioinformática que automatiza todos os processos produzidos para anotação funcional, microarranjos de diferentes plataformas, e para alguns tipos de informações de sequenciamentos de larga escala, tornando a plataforma reutilizável. Porém, o campo de pesquisa da genômica e transcriptômica está constantemente em progresso e novas e mais eficientes técnicas são desenvolvidas e liberadas. A modularidade embutida na plataforma permite a fácil extensão e melhora nos procedimentos de anotação, análise de expressão gênica e integração com outros métodos e aplicativos. Atualmente existe uma grande variedade de ambientes de análise em bioinformática, em diferentes linguagens de programação, incluindo análises complexas com recursos locais (Orvis et. al., 2010; Goecks et. al., 2010), em sistemas remotos através de acesso a serviços

web (Oinn et. al., 2004), ou fazendo uso de sistemas de *grid* distribuídos (Sroka et. al., 2011; Goecks et. al., 2010; Oinn et. al., 2004). Para avaliar diferentes sistemas de fluxo de trabalho, é necessário avaliar as qualidades e fraquezas relativas à implementação destas infraestruturas. A utilização de recursos locais para análises de dados de larga escala tem a vantagem de ter o controle completo sobre os dados e uso de recursos computacionais da máquina dedicados ao projeto. Por outro lado, tem os custos de manutenção, instalação e investimento em pessoal para manter e gerenciar as máquinas. Serviços web, grid e computação em nuvem oferecem uma alternativa atrativa para superar o investimento de capital inicial em *hardware* (Stein, 2002; Hernandez et. al., 2004). Um dos requisitos fundamentais na utilização de recursos remotos para computação é o acesso rápido e barato a bandas de rede para transferência de dados, mas estes requisitos frequentemente impossibilitam o uso de serviços remotos de instituições ou grupos de pesquisa, além de poder apresentar limites no uso dos recursos computacionais.

Correntemente existem três tipos de integração para dados biológicos: integração ligação, integração visual e armazenamento de dados (*data warehouse*) (Davidson et. al., 1995; Stein, 2004). Estas abordagens são diferentes no armazenamento dos dados e no grau de integração. Integração por ligação e visual não armazenam o dado localmente, enquanto a abordagem de armazenamento de dados recupera um subconjunto de dados específico dentro de um banco local, sendo que bancos externos podem ser copiados para dentro do banco local. A abordagem de integração visual utiliza uma interface, que permite receber consultas que são convertidas em consultas a bancos externos, um exemplo é o *DiscoveryLink* (Haas et. al., 2001), enquanto integração ligação somente fornece ligação de informação entre dados de diferentes fontes. Os bancos de armazenamentos não são susceptíveis a fatores externos, como conexão da rede, manutenção do banco, etc. O banco de dados permite otimizar consultas e processos localmente de acordo com o esquema de dados. Curadores podem adicionar anotações a dados existentes. Devido a essas vantagens, esta plataforma utiliza essa estratégia para integração de dados. Porém, há desafios neste tipo de integração de dados. A integração deveria resolver problemas na transformação e integração nos modelos de dados, mapeamento de semânticas e conversão de dados (Davidson et. al., 1995; Hernandez et. al., 2004).

A integração de dados em biologia molecular pode ser definida como *integrômica* (Venkatesh et. al., 2002), e pode ser dividida em duas áreas de pesquisa: 1-conceitual ou integração de dados qualitativa; e 2- algorítmica ou integração de dados quantitativos.

A integração de dados conceitual pode ser definida como a combinação de diferentes bancos de dados, em diferentes formatos, dentro de um esquema global (conceitual). Porém, o grande volume de bancos de dados biológicos tem tornado a integração de relevantes informações biológicas impraticável. Esse tipo de pesquisa busca desenvolver plataformas para recuperar a informação de forma rápida e com o mínimo de intervenção do usuário (Hernandez et. al., 2004).

A integração de dados algorítmico se resume ao uso de diferentes tipos de dados a partir de um *pipeline* de análise de um experimento. Esta integração pode ocorrer em diferentes níveis da análise. Na integração em uma fase inicial, os dados brutos são transformados e combinados em um formato padronizado como entrada para as análises. A fase intermediária ocorre quando resultados das análises são combinadas com outros tipos de informações para análises subsequentes. Em geral, dados brutos de um experimento são combinados com informações adicionais usando métodos matemáticos ou estatísticos para identificar resultados biológicos significativos. No estágio intermediário da integração, a análise de agrupamento é um exemplo deste tipo de integração. A fase final ocorre quando resultados de análises são interpretados e verificados usando informações relevantes ou validadas experimentalmente. Neste contexto, diferentes abordagens para os diferentes estágios de integração na plataforma foram desenvolvidos.

A anotação automática através do método da taxa de identificação bidirecional (*BHR - Bidirectional Hit Rate*) e identificação de genes homólogos podem dar pistas sobre funções gênicas compartilhadas, porém é um estudo que não é possível ser validado facilmente. Nesta tese, foi implementado um método para anotação automática e a análise de identificação de ortólogos para o transcriptoma da cana-de-açúcar e outras plantas. Os dados brutos e as análises são integrados à plataforma.

3.11. Bancos de Dados Relacional e de Grafos

Nas últimas décadas um grande volume de dados biológicos tem sido produzido, especialmente nos últimos anos com as novas tecnologias de larga escala. Os bancos de dados relacionais tem sido a principal ferramenta para modelar e armazenar dados biológicos e seus relacionamentos, permitindo o desenvolvimento de ferramentas para extração destes dados (Bergholz et. al., 2001). Porém, muitos problemas biológicos são difíceis de serem modelados em bancos de dados. Como o conhecimento biológico possui grande diversidade e está sempre em transformação produzindo novos

conhecimentos e requisitando aplicações com questões não convencionais, é preciso criar modelos que possam evoluir rapidamente (Nelson et. al., 2003; Jagadish et. al., 2004).

Nosso objetivo é criar um modelo relacional, com potencial para facilitar a descoberta científica, para poder armazenar a informação de forma estruturada, flexível e que possa ser facilmente expandida para comportar novos dados biológicos, resultados de análises intermediárias nas fases de processamento dos dados e conhecimento biológico agregado de outras fontes.

Um Sistema de Gerenciamento de Bancos de Dados Relacional (SGBD) é uma coleção de programas que permitem armazenar, modificar e extrair informações de um banco relacional possibilitando integridade de dados, escalabilidade, segurança, e concorrência. Bancos relacionais tem suporte na teoria de matemáticas relacionais baseado na teoria de conjuntos e foi concebido por (Codd, 1980). O conjunto de regras para construir uma consulta é conhecido como Linguagem de Consulta. Existem diferentes linguagens para diferentes sistemas, porém a linguagem mais comum é a Linguagem de consulta semi-estruturada (SQL – *structured query language*) (Codd, 1970).

As principais vantagens do uso de SGBDs são:

- Redução da duplicação de dados: melhorando a integridade dos dados.
- Independência dos dados: dados podem ser imaginados como sendo armazenados em tabelas independentemente do armazenamento físico.
- Independência da aplicação: o banco é independente do sistema e programas que estão acessando o mesmo.
- Concorrência: compartilhamento de dado com múltiplos usuários.
- Consultas complexas: consultas simples podem recuperar dados de mais que apenas uma tabela.

Os sistemas metabólicos podem ser modelados como bancos de dados baseados em grafos, que têm sido curados e melhorados constantemente para diferentes organismos, principalmente para plantas. Em particular, a representação de vias metabólicas em grafos (Karp et. al., 2002) permite agregar níveis de expressão de genes, proteínas ou ativação/inibição de enzimas, e realizar complexas operações entre grafos (Karp et. al., 2002). Com o objetivo de criar ambientes de consulta para vias metabólicas, com especificações e ferramentas, os bancos podem ser classificados em três categorias. A primeira corresponde aos bancos de dados referência para vias metabólicas, como KEGG (Kanehisa et. al., 2000), BioCyc e MetaCyc (Caspi et. al., 2008), PlantCyc (<http://plantcyc.org>) e Reactome (Joshi-Tope, et. al., 2005), que

possuem interfaces de consultas sobre as redes bioquímicas. A segunda categoria inclui sistemas biológicos e modelagem de sistemas como o BioSim (Heidtke et. al., 1998), GenSim (Karp, 1993) e PathCase (Elliot et. al., 2008). Um terceiro grupo provê simulações biológicas, modelagens de sistemas e o desenvolvimento de ambientes de consultas para explorar o comportamento das redes bioquímicas, sendo que uma das abordagens é a criação de linguagens declarativas preservando e conservando propriedades estruturais, seus relacionamentos e a representação dos grafos em modelos como o relacional (Tian et. al., 2008), orientação a objetos (Cakmak et. al., 2010), com PL/SQL (Leser, 2005) e XML (Yang et. al., 2008). Essas plataformas modelam as redes de dados em grafos que proporcionam consultas das propriedades estruturais das vias metabólicas (vias, vizinhanças, ciclos, etc), informações de relacionamento entre entidades (inibidores de uma reação em uma via particular). Porém, essas plataformas são parte de sistemas mais complexos ou não estão disponíveis para um reaproveitamento da arquitetura. Deste modo criamos uma estrutura para integração de dados de Vias Metabólicas a partir do KEGG, integrando genes e dados de experimentos de microarranjo para armazenamento em um banco de dados de grafos, através da plataforma de bancos de dados de grafos Neo4J.

Os bancos de dados de grafos (BDG) são uma alternativa aos sistemas de bancos relacionais SBDR. Os bancos relacionais contêm tabelas que são definidas por conjuntos de linhas e colunas. As tabelas podem ser vistas como classes, que contêm diversos objetos e informações, as linhas funcionam como objetos, equanto as colunas correspondem a atributos ou propriedades destes objetos (Rodriguez et. al., 2011). Um das desvantagens de bancos de dados relacionais é a complexidade para representação de estruturas complexas e irregulares como as vias metabólicas, necessitando de a modelagem de diversas tabelas, e exigindo um alto custo para consultas sobre dados com múltiplos relacionamentos, podendo impossibilitar a extração de requisitos semânticos ou interconexões entre diferentes objetos. Na ciência, o grande volume de dados biológicos, envolvendo complexas informações interconectadas, necessita de uma estrutura que permita o armazenamento, recuperação e manipulação destes dados sem perda da semântica, o que torna oneroso utilizar abordagens de SBDR (Hull et. al., 1987). Bancos de dados de grafos sustentam modelos de grafos que permitem para uma persistência direta, o armazenamento de objetos particulares no banco de dados junto com as relações entre eles. Os BDG fornecem um acesso para métodos de consulta que tratam não apenas dos objetos armazenados, mas também da estrutura do grafo em si. A travessia é uma das operações mais conhecidas e empregadas em grafos, e pode ser usada para obter

vizinhos de um objeto específico, isto é, os objetos que estão diretamente relacionados com o objeto inicial.

Apesar das vantagens de ter acesso direto ao grande volume de dados interconectados, existe um grande custo de recursos computacionais devido ao complicado particionamento do banco de dados. Uma das alternativas é o particionamento do grafo entre diferentes máquinas, porém algoritmos para esta operação ainda não foram desenvolvidos com sucesso e o problema continua em aberto (Robinson, 2013).

O grafo é um objeto que contém nós e relacionamentos. Os nós são estruturados pelos relacionamentos e ambos possuem propriedades. A topologia de um grafo G pode ser expressa como $G = (V, E)$, onde V é um conjunto de vértices (nós) e E é um conjunto de arestas (ou relações). Uma aresta conecta dois vértices e ambos precisam existir, i.e., relações pendentes não são permitidas. O relacionamento entre os nós de um grafo podem possuir uma direção, e os grafos podem ser direcionados ou não direcionados. A topologia pode conter mais informação dentro do grafo, o que diferencia os modelos de grafos, alguns exemplos:

- Grafos rotulados: todas as arestas tem um rótulo que define o tipo de aresta. Há também uma variante do grafo rótulo-vértice onde cada vértice tem um rótulo.
- Grafo-Multi: é um grafo rotulado onde múltiplas arestas podem existir entre quaisquer dois vértices, garantido que estas arestas têm diferentes rótulos.
- Grafo atribuído: elementos do grafo podem ter atributos (propriedades) agregados a eles para carregar informação adicional. Propriedades são usualmente expressas como mapas de chaves e seus valores associados.
- Grafo de propriedade: é definido como um atributo grafo-multi direcionado e é um modelo que é implementado pela maioria dos bancos de grafos (Rodrigues et. al., 2010).

A plataforma de banco de dados de grafos empregada é o Neo4j, é escrita na linguagem de programação Java (Naughton et. al., 1996), sendo um projeto de código aberto e sua primeira versão é de 2007. Uma das características do Neo4j é ser um modelo de grafo de propriedades. A API (Application Program Interface) nativa está aberta a diferentes tipos de linguagens, i.e. Java, Python, Ruby, JavaScript, PHP, .NET, etc. A plataforma apresenta como alternativas para consultas ao banco de dados, API e REST nativas expandindo a versatilidade na comunicação com o banco, usando linguagens de consulta SPARQL (Sakr et. al., 2012) ou Cypher (Cypher query language⁵). Neo4j também suporta índices customizados sobre propriedades dos

⁵ <http://docs.neo4j.org/chunked/stable/cypher-query-lang.htm>

elementos utilizando engenharias de indexação externa, empregando o Apache Lucene como engenharia padrão (Apache Lucene search engine library⁶). Neo4j suporta transações ACID (*Atomicity, Consistency, Isolation, Durability*), isto é, obtido mantendo os logs das transações na memória e um gerenciador de trava aplicando travas em qualquer objeto do banco alterado durante uma transação. Na conclusão bem sucedida da operação, as mudanças no log são liberadas para o disco, enquanto que uma reversão de transação significa apenas descarte do log.

A persistência no banco do Neo4j é fornecido por um grafo de armazenamento em plano de fundo, principalmente utilizando arquiteturas de listas adjacentes. Os três principais componentes do grafo, i.e. vértices, arestas e propriedades dos elementos, são armazenados em três arquivos separados. Os vértices são armazenados com um ponteiro para suas arestas e a primeira propriedade. O armazenamento de propriedades é um grupo de listas ligadas, onde há uma lista ligada por vértices. E as arestas são armazenadas com duas listas duplamente ligadas (um para cada extremidade da aresta) com os tipos das arestas e ponteiros para a corrente extremidade da aresta e a primeira propriedade. Todos os ponteiros são expressos na forma de um identificador de objeto inteiro e todas as gravações nos três arquivos têm um tamanho fixo e posição absoluta correspondendo ao identificador do objeto. Deste modo, acessos diretos nos arquivos armazenados podem ser realizados diretamente calculando o *byte* onde a gravação está iniciando. Tais rearranjos podem simplificar a leitura de dados de um armazenamento persistente e permitir uma travessia rápida direta ao grafo (Robinson, 2013).

⁶ <http://lucene.apache.org/core>

4. Objetivos

4.1.1. Objetivos Gerais

O objetivo deste trabalho é desenvolver uma plataforma para estudo da anotação funcional e da expressão gênica do transcriptoma, que permita a identificação de genes alvos para melhoramento da cana-de-açúcar através da definição de vias de sinalização e vias metabólicas associadas a características de interesse, de forma a confrontar as alterações nessas vias com os padrões de expressão gênica. Isto significa desenvolver um sistema que possa ser expandido de forma integrada para agregar novas “ômicas” e ferramentas computacionais para análise da expressão gênica de forma integrada às vias metabólicas e à anotação funcional.

4.1.2. Objetivos Específicos

- Desenvolver metodologias para anotação funcional do transcriptoma da cana-de-açúcar.
- Definir as vias metabólicas de cana-de-açúcar baseando-se nas vias metabólicas do KEGG.
- Implementar métodos para classificação de famílias de Fatores de Transcrição em cana-de-açúcar e gramíneas.
- Desenvolver um *pipeline* para criação e customização de microarranjos de oligonucleotídeos para organismos não modelo a partir do transcriptoma.
- Implementar métodos e desenvolver *pipelines* para análise de dados de microarranjos de cDNA e oligonucleotídeo, para identificação de genes significativamente (SE) e diferencialmente expressos (DE).
- Analisar a presença e expressão de mensagens antisense em cana-de-açúcar.
- Desenvolver uma plataforma para armazenamento e recuperação de dados biológicos através de banco de dado relacional e baseado em grafos, para integração de dados heterogêneos.
- Desenvolver aplicações e métodos de integração, agregadas à plataforma, para mineração de dados biológicos heterogêneos, através do uso de métodos de correlação, agrupamento e análises de enriquecimento.

5. Materiais e Métodos

5.1. Anotação Funcional do Transcriptoma de Cana-de-açúcar

A anotação funcional é a ação de caracterizar um conjunto de genes preditos, atribuindo a eles uma função biológica, uma característica de estrutura ou uma regra funcional ou metabólica. A anotação permite a atribuição ou uma identificação para cada gene resultando da sequência codificadora do gene. Este tipo de informação é extremamente útil para as análises subsequentes baseadas em famílias de genes ou grupos de genes envolvidos em vias metabólicas específicas.

A estrutura de uma das classes de genes, pode ser composta pela sequência codificadora (*CDS*) que contém nucleotídeos traduzidos em proteínas de aminoácidos. A região codificadora inicia com o códon de iniciação, usualmente ATG, e termina geralmente com os códons TAA, TAG ou TGA. As extremidades da região codificadora que são transcritas, mas não traduzidas, são chamadas regiões não traduzidas (*UTR*). Estas regiões *UTR* podem conter elementos regulatórios que controlam a síntese de proteínas. O RNA mensageiro (*mRNA*) (Figura 11) carrega as informações essenciais para a tradução do *mRNA* em proteína. Essas informações estruturais dos genes definem se o gene está completo (*mRNA*). O *mRNA* pode auxiliar na predição de todos os domínios conservados que caracterizam a função de um gene e permite a comparação de forma mais confiável para a identificação de genes homólogos. Além da estrutura é fundamental a anotação através de bancos de dados curados como Uniprot, GO, KEGG, e se disponível bancos públicos de plantas como a plataforma *Pythozome* (Goodstein et. al., 2012). Essas bases servem como referência para transferência da anotação entre espécies, aliadas a outras abordagens como a identificação de domínios que caracterizam e classificam os genes para determinada função ou família funcional. Esta abordagem auxilia na anotação e curagem destes genes, e serve principalmente como referência para o desenho de *chips* de microarranjo de acordo com genes de interesse para cana-de-açúcar.

Em um novo transcriptoma sequenciado, uma das questões mais importantes a ser respondida é a identificação de genes que codificam proteínas e sua estrutura. Porém a identificação da estrutura e genes alelos só é possível, com certo grau de confiabilidade, através de um genoma sequenciado para o organismo em estudo. Essas características são compartilhadas entre todos os genes que codificam proteína, e são utilizadas por métodos *in silico* para construir modelos de genes hipotéticos na predição

de novos genes. As tecnologias de “-ômicas” têm contribuído para a geração e acúmulo de conhecimento como sequenciamento de novos genomas, anotações de genes, proteínas e interações entre moléculas. A disponibilidade dessas informações biológicas em vários organismos, principalmente em plantas (Goodstein et. al., 2012), tem ajudado a melhorar a anotação funcional de novos transcriptomas e genomas, na predição e anotação das estruturas dos genes e suas interações (Shoemaker et. al. 2007).

Figura 11. Estrutura do mRNA, e suas características como Cap 5', sequência codificadora, regiões 5' e 3' UTR não traduzidas e cauda Poli-A.

Os bancos de dados atuais apresentam informações de diversos genomas e transcriptomas de diferentes espécies, geralmente com anotações preditas por similaridade, genes preditos, interações moleculares (Goodstein et. al., 2012; Benson et. al., 2012), redes metabólicas e transdução de sinal (Jeong et. al., 2000; Kanehisa et. al., 2004; Croft et. al., 2010) e redes de interação gênica (Bader et. al., 2003; Gilbert et. al., 2005).

A disponibilidade de novos genomas e transcriptomas sequenciados tem permitido a utilização de métodos de identificação por homologia em nível de sequências, baseado em escores de similaridade. A homologia entre duas proteínas ou sequências de DNA é definida em termos de um ancestral em comum (Koonin et. al., 2005). Duas proteínas podem ter um ancestral em comum devido a um evento de especiação (ortólogo) ou duplicação (parálogo). Sequências homólogas são ortólogas se for inferido que elas descendem de uma sequência de um ancestral comum pelo evento de especiação. Quando uma espécie diverge em duas novas espécies, as cópias de um único gene nessas espécies são conhecidas como gene ortólogos (Fitch et. al., 1970). As sequências homólogas são parálogas se forem criadas por um processo de duplicação dentro de uma mesma espécie, e podem ser encontrados em dois *loci* distintos no genoma.

Como parte deste trabalho, estamos interessados na anotação funcional e estrutural destes genes, iniciando com predição das proteínas e regiões UTR em cana-de-açúcar a partir dos transcritos do projeto SUCEST e do sequenciamento de larga

escala por RNA-Seq. Com as proteínas é possível a identificação de domínios, identificação de genes homólogos entre diferentes gramíneas e a classificação de fatores de transcrição. Neste capítulo, descreveremos os principais métodos para anotação e estratégias para análise e interpretação dos genes a nível funcional e estrutural. Essas informações estão integradas à plataforma *CaneRegNet* para anotação e curadoria do transcriptoma de cana-de-açúcar, aliado à criação de um conjunto de categorias particulares e a classificação de famílias de fatores de transcrição.

Foi desenvolvida uma abordagem que realiza a comparação entre transcriptomas de duas espécies próximas evolutivamente para a transferência de anotação entre genes, pois compartilham em geral grande número de genes ortólogos. Com o grau de conservação entre genes ortólogos, os candidatos podem ser computacionalmente deduzidos por um relacionamento de comparação entre genomas pelo método da taxa de acerto bi-direcional (*BHR - bi-directional hit rate*) (Tatusov et. al., 1997; Nishiyama-Jr et. al., 2014), sendo um dos caminhos mais curtos e com algum critério de confiabilidade na predição de funções. Deste modo, essa abordagem foi aplicada aos transcritos de cana-de-açúcar contra os bancos de dados do *Uniprot*, *Gene Ontology* e *Phytozome* (Sorgo, Milho, Arroz, *Brachypodium*, *Setaria*, *Panicum* e *Arabidopsis*).

Atualmente existem três principais bancos de dados no mundo, sincronizados entre si, para depósito e acesso a sequências, o GenBank nos Estados Unidos (Benson et. al., 2012), o DDBJ no Japão (Tateno et. al., 2002) e o EMBL na Europa (Kulikova et. al., 2004). Um dos bancos de dados referência para transferência de anotação para novos organismos e busca de conhecimento biológico integrado e curado para proteínas é o Uniprot (Magrane et. al., 2001), com uma anotação funcional, precisa e consistente dividida em duas seções: *Uniprot/Swissprot*, com sequências curadas manualmente a partir de anotações automatizadas computacionalmente, e *Uniprot/TrEMBL*, anotadas computacionalmente.

O *Gene Ontology* reúne todos os genes eucariotos através de um vocabulário controlado das regras de proteínas e genes a nível celular, especificando funções biológicas que são compartilhadas e divididas em três ontologias: Processos Biológicos, Funções Moleculares e Componentes Celulares.

Para a identificação de enzimas e proteínas envolvidas em vias metabólicas os estudos foram baseados no banco do KEGG (Kanehisa et. al., 2004).

O Phytozome é um portal de genômica comparativa do departamento de energia dos E.E.U.U o *Joint Genome Institute*, com diversas ferramentas de buscas por

similaridade e informação biológica entre diferentes espécies. O portal disponibiliza as sequências e anotações de genomas de plantas, contendo anotações como PFAM, KOG, KEGG, PANTHER e atribuições GO.

Uma alternativa às metodologias tradicionais para busca por similaridade é o uso de técnicas de buscas por domínios, que são significativamente mais precisas e mais aptas a detectar homólogos mais distantes evolutivamente e a identificar domínios estruturais ou funções conservadas, devido ao poder dos modelos matemáticos subjacentes como HMM (*Hidden Markov Models*) (Eddy et. al., 1998). A principal ideia dessas técnicas é comparar as sequências a modelos estatísticos que descrevem um padrão ou uma família de sequências. Na plataforma *CaneRegNet*, uma análise em larga escala foi realizada com o *software* InterProScan (Quevillon, et. al., 2005), uma ferramenta para identificação de domínios de proteínas, combinando diferentes métodos de reconhecimento em diferentes bancos de dados, como PFAM (Finn et. al., 2008).

O banco de dados PFAM é um banco de famílias de proteínas com domínios estruturais e funcionais em comum. Cada família é representada por alinhamento múltiplo de sequências e modelos de cadeias de Markov, fornecendo uma classificação de clãs, agrupando famílias relacionadas de acordo com a similaridade da sequência ou padrão-HMM. As famílias de domínios PFAM são caracterizadas pelos seguintes tipos: 1) família, proteínas com mesmos domínios; 2) domínios, definindo uma unidade estrutural representada em diferentes famílias; 3) repetições, representando sequências curtas presentes em múltiplas proteínas globulares; 4) motivos, consistindo de sequências curtas fora de proteínas globulares. O *software* utilizado pelo InterProScan para identificação de motivos em novas proteínas baseado em bibliotecas de padrão-HMM foi o HMMER.

Os transcritos produzidos pelo projeto SUCEST e pelo sequenciamento por RNA-Seq foram processados e anotados de acordo com o fluxograma (Figura 12), e as anotações foram utilizadas para seleção dos genes de interesse para o desenho dos oligonucleotídeos de microarranjo. Para a anotação dos transcritos contra diferentes bancos de dados, desenvolvemos um componente seguindo o modelo do KAAS (*KEGG Automatic Annotation Server*), um sistema *web* para anotação automática de proteínas em enzimas e vias metabólicas, classificando-as em grupos ortólogos (Moriya et. al., 2007). Devido à ausência de uma base de dados de proteínas disponível e atualizada para as vias metabólicas do KEGG, realizamos a anotação dos transcritos para

identificação de enzimas, das vias metabólicas e identificação de ortólogos deste grupo com o KAAS.

As abordagens aplicadas à cana-de-açúcar para anotação dos transcritos pode ser dividida em quatro principais etapas:

1 - Anotação dos transcritos com critérios avaliados empiricamente por similaridade com os bancos de dados do *Gene Ontology*, *Phytozome* (Sorgo, Milho, Arroz, *Brachypodium*, *Panicum*, *Setaria* e *Arabidopsis*), *Uniprot/SwissProt* e *NR-NCBI*.

2 - Anotação das vias metabólicas, enzimas e ortólogos de enzimas com o KAAS.

3 - Análise de predição de proteínas, com identificação de domínios e padrões funcionais com PFAM.

4 - Predição de proteínas ortólogas com o *software* Inparanoid.

Os resultados dessas análises foram utilizados em diversos estudos do grupo de pesquisa, como a criação de trinta categorias funcionais curadas manualmente pelo grupo, anotação dos transcritos de cana, principalmente para os transcritos presentes nos microarranjos; no desenvolvimento de uma metodologia para predição de transcritos *full-length*; na identificação e classificação dos fatores de transcrição em cana-de-açúcar e no desenvolvimento de uma estrutura de bancos de dados e interface integrando todo o conhecimento biológico gerado para análises de dados de microarranjo.

Figura 12. Pipeline para anotação do transcriptoma de cana-de-açúcar e identificação de genes ortólogos.

Para as análises de anotação do transcriptoma foram adotados diferentes programas para busca de similaridade como BLAST (BLASTN/TBLASTN/BLASTX) e busca de padrões e domínios com InterProScan (HMMPFAM). De modo complementar foram desenvolvidos *scripts* e APIs (*Application Programming Interface*) em Perl para atualização automática dos bancos de dados descritos acima, *scripts* para processamento dos resultados e critérios estatísticos para avaliação da similaridade para cada *software*.

A etapa inicial de anotação é baseada na busca por similaridade com o programa BLAST, realizando alinhamentos com critério de e-valor $1e-05$ para BLASTX e TBLASTN. A partir do transcriptoma a ser anotado, esse é comparado contra um banco referência (ex: Uniprot) ou transcriptoma referência (ex: espécies do phytozome) em ambas as direções, considerando cada um dos genes no transcriptoma de entrada comparado com todos os genes no transcriptoma referência, e vice-versa.

Os resultados são avaliados com o programa desenvolvido localmente, sendo baseado em dois métodos: melhor escore de alinhamento bi-direcional (*BHR* – *bi-*

directional best Hit Rate) e o melhor escore de alinhamento uni-direcional (*SBH – single-directional best hit Rate*). Todos os genes alinhados pelo BLAST com escore *bit* menor que 30 são descartados da próxima etapa. Esses critérios foram definidos empiricamente, através de múltiplos testes comparativos com os resultados obtidos na página do KAAS com o banco do KEGG. A definição de um critério adicional para anotação, baseado no alinhamento em vice-versa entre dois transcriptomas, é devido às discrepâncias nesses dois resultados. O escore *bit* de um par de genes g1 e g2 de dois transcriptomas T1 e T2, pode apresentar resultados distintos nos alinhamentos em vice-versa, e o melhor escore pode não refletir necessariamente a ordem dos escores rigorosos do *Smith-Waterman* (Smit et. al., 1981). Adaptando a abordagem desenvolvida no KAAS para identificação de genes ortólogos, o BHR pode ser definido como:

$$\text{BHR} = R_f * R_r$$

Aqui, a razão $R = S/S_r$, onde S é o escore de g1 contra g2, e S_r é o escore de g1 contra o melhor gene alinhado no transcriptoma T2 (que pode não ser necessariamente g2). R_f refere-se ao escore do BLAST inicial (T1 contra T2), e R_r refere-se ao escore do BLAST reverso (T2 contra T1). Os genes serão selecionados se apresentarem um $\text{BHR} \geq 0.95$, e $R_f \geq 0.95$ no método SBH.

Deste modo, os transcritos são anotados inicialmente a partir dos bancos do *Gene Ontology* e do *Phytozome*, e para ser atribuída a anotação destes bancos devem ter e-valor $\leq 1e-05$ e um escore *bit* ≥ 30 , e em seguida apresentar um BHR ou $R_f \geq 0.95$. Para os transcritos que não satisfazem estes critérios são em seguida mapeados, na ordem, contra os bancos do Uniprot/Swissprot e NR-NCBI com critério apenas de e-valor, os genes que não são anotados em nenhum destes bancos e não atendem aos critérios estabelecidos são classificados como *no match*. Para manter a proveniência da anotação do *Gene Ontology*, mantivemos a informação relacionada dos tipos de códigos de evidências vinculados às proteínas, fornecendo informações valiosas sobre a qualidade da anotação das proteínas. Os códigos de evidência e categorias utilizadas no GO são listadas na tabela (Tabela 2).

Tabela 2. Códigos de Evidência para a proveniência das anotações do GO definidas pelo Consórcio de Ontologia de Genes (*The Gene Ontology Consortium*).

Código	Tipo de evidência	Tipo de categoria
NAS	Non-traceable Author Statement	Author Statement
TAS	Traceable Author Statement	

IEA	Inferred from Electronic Annotation	Automatically-assigned
IGC	Inferred from Genomic Context	Computational Analysis
ISA	Inferred from Sequence Alignment	
ISM	Inferred from Sequence Model	
ISO	Inferred from Sequence Orthology	
ISS	Inferred from Sequence or Structural Similarity	
RCA	Inferred from Reviewed	
IC	Inferred by Curator	Curator Statement
ND	No biological Data available	
EXP	Inferred from Experiment	Experimental Evidence
IDA	Inferred from Direct Assay	
IEP	Inferred from Expression Pattern	
IGI	Inferred from Genetic Interaction	
IMP	Inferred from Mutant Phenotype	
IPI	Inferred from Physical Interaction	

A segunda etapa de anotação corresponde à identificação das enzimas e vias metabólicas em cana-de-açúcar, e para esta análise foi utilizada a plataforma *on-line* do KAAS, seguindo os mesmos critérios adotados anteriormente para a transferência das anotações, utilizando os bancos de dados de plantas do KEGG. O KAAS apresenta uma etapa adicional que é a identificação de candidatos ortólogos (*KO – KEGG Ortholog*) baseado em seu banco de dados. Os candidatos ortólogos são divididos em grupos de acordo com a anotação do banco KEGG GENES. Finalmente, o escore de atribuição é calculado baseado na verossimilhança e heurística para cada grupo KO. Então, o número do grupo KO com melhor escore é atribuído para a sequência.

O escore de atribuição do grupo de ortólogos é definido de acordo com:

$$S_{ko} = S_h - \log_2(mn) - \log_2\left(\sum_{k=N}^x C_k p^k (1-p)^{x-k}\right)$$

Onde S_h é o maior escore entre todos os candidatos ortólogos no grupo de ortólogos, m e n são os comprimentos das sequências consulta e alvo do BLAST, N é o número de organismos no grupo de ortólogos, x é o número de organismos no grupo original de ortólogos do qual este grupo é derivado, e p é razão do tamanho do grupo ortólogo original *versus* o tamanho total do banco GENES. O segundo termo é

responsável pela normalização do primeiro termo pelo comprimento das sequências, e o terceiro termo é o fator de peso para considerar o número de ortólogos candidatos que são identificados no original.

Apesar de esse método identificar os ortólogos dos transcritos de cana-de-açúcar, essa análise é baseada apenas nas enzimas de plantas depositadas no banco do KEGG restringindo a anotação para transcritos específicos. Essa abordagem, também prediz ortólogos funcionais ancestrais para as enzimas, e não é capaz de identificar genes parálogos em famílias que são conservadas e que podem também apresentar função compartilhada.

Na terceira etapa da anotação, a predição de proteínas para os transcritos de cana-de-açúcar foi realizada com o *software* ESTScan (Iseli et. al., 1999). Inicialmente foram realizadas predições com matriz de *codon-usage* de arroz, porém o sequenciamento e anotação de alguns trechos genômicos de cana-de-açúcar a partir de BACs (de Setta et. al., 2013), permitiu desenvolvermos uma matriz de código genético (*codon-usage*) para cana-de-açúcar com abordagens baseadas no *software* do ESTScan. As matrizes de *codon-usage* são específicas para cada organismo, pois refletem a preferência de códon. A seguir são descritas as etapas para o desenvolvimento da matriz de *codon-usage*, para maiores detalhes (Rielle et. al., 2007).

5.2. Predição de proteínas em Cana-de-açúcar e Identificação de domínios

Para o desenvolvimento de uma matriz de *codon-usage* é necessária a utilização de mRNAs (*full-length*) contendo sequência e anotação genômica, no formato *genbank*. Deste modo, a partir dos BACs anotados foram selecionados 832 mRNAs de acordo com os seguintes critérios:

- 1- Seleção de genes completos com CDS $\geq$ 1000pb.
- 2- Seleção de genes com ao menos 2 exons e 1 intron.
- 3- Seleção de genes com ao menos 100pb nas regiões 5'UTR e 3'UTR.

Foram criados *scripts* para criação de sequências no formato *fasta* contendo regiões 5'UTR+CDS+3'UTR, e criação de arquivos formato *genbank* para cada BAC e cada gene correspondente selecionado na análise.

O desenvolvimento da matriz de *codon-usage* através do ESTScan é baseada em três *scripts*: *extract_mrna*, *prepare_data* e *build_model*.

O primeiro *script*, *extract_mrna*, extrai os dados dos arquivos para cada BAC, contendo os genes pré-selecionados no formato *genbank*. O segundo *script*,

prepare_data, classifica os mRNAs extraídos previamente, separando os mRNAs em dois grupos: dado de treinamento e teste. O dado de treinamento é dividido de acordo com o conteúdo GC dos mRNAs, a redundância é mascarada e as partes codificadoras e não codificadoras dos mRNAs são classificados em dois diferentes arquivos. Finalmente é criada a matriz com o *script build_model*.

Após a geração da matriz, foi utilizado o *script evaluate_model*, para avaliar a eficiência da matriz. O *script* calcula a taxa de falso negativo (sensitividade): a porcentagem de sequências codificadoras classificadas como não-codificadoras, a taxa de falso positivos (especificidade): a porcentagem de sequências não-codificadoras classificadas como codificadoras e a detecção da acurácia dos sítios de início e fim da proteína.

Após a geração da matriz de *codon-usage*, as proteínas foram preditas para os transcritos com o *software* ESTScan, com parâmetros default.

Com o objetivo de obter uma anotação funcional a partir das proteínas preditas dos transcritos utilizamos o *software* InterProScan. O InterproScan corresponde a um *pipeline* que agrega diversas informações de domínios de diferentes bancos de dados e ferramentas, incluindo CATH-Gene3D, PANTHER, Pfam, PIRSF, PRINTS, ProDOM, ProSITE, SMART, SUPERFAMILY e HAMAP, classificando as proteínas dentro de famílias e predizendo domínios e sítios conservados.

Atualmente o InterProScan contém 26.860 entradas, com 18.218 famílias, 7.518 domínios, 277 repetições, 109 sítios de ativação, 72 sítios de ligação, 650 sítios conservados e 16 PTM⁷.

Nosso principal interesse está no resultado da identificação de domínios a partir das proteínas preditas com HMMPFAM, e os parâmetros e critérios de corte foram e-valor menor que 1e-03 e cutoff=--cut_ga, que é o escore mínimo que uma sequência deve alcançar quando construindo um alinhamento completo de uma entrada do PFAM.

A curagem baseada simplesmente na anotação é limitada a informação de uma sequência com uma proteína completa, que pode possuir apenas um alinhamento parcial, sendo limitado a uma funcionalidade comum. Porém, sequências com múltiplas regiões funcionais, mas que não possuem a mesma combinação ou ordem dos domínios poderão não ser identificadas. Deste modo, a agregação de informações referentes à similaridade entre proteínas e identificação de domínios fornece maior suporte para a anotação funcional.

⁷ Banco de dados de Janeiro, 2015.

5.3. Identificação de ortólogos entre Cana-de-açúcar e Gramíneas

A quarta etapa de anotação do transcrito, refere-se ao estudo de homologia para ortólogos entre os transcritos de cana-de-açúcar e as gramíneas (Sorgo, Milho, Arroz, *Brachypodium*, *Setaria* e *Panicum*), incluindo *Arabidopsis* (não gramínea).

O programa InParanoid (Berglund et. al., 2008) utiliza um método que implementa um conjunto de regras heurísticas para unir, deletar e separar grupos de ortólogos preditos. Inicialmente, os principais ortólogos são identificados com pares de proteínas tendo alto escore no BLAST e são usadas como sementes para identificar todos os in-parálogos e out-parálogos para cada espécie. Os parálogos podem ser separados em dois grupos: in-parálogos, pares de parálogos que surgiram após um evento de especiação, e os out-parálogos, pares de parálogos que surgiram antes do evento de especiação. O InParanoid lida com relacionamentos de ortólogos um-para-um e muitos-para-muitos entre dois proteomas, e atribui um valor de confiança para cada grupo de ortólogos e in-parálogos. O valor de confiança corresponde a um escore de similaridade, para mostrar quão distante uma sequência está do ortólogo principal na mesma espécie em uma escala entre 0% a 100%. O valor de 100% é atribuído para ortólogos principais e 0% é atribuído para uma sequência com escore mínimo de similaridade exigido para ser identificado como in-parálogo de um dado grupo. O InParanoid é limitado a comparações pareadas e não permite sobreposição de grupos na presença de proteínas híbridas.

O programa MultiParanoid (Alexeyenko et. al., 2006), surgiu como uma atualização do InParanoid, com o objetivo de construir grupos de proteínas ortólogas multi-espécies entre todas as comparações pareadas com InParanoid. O MultiParanoid pode ser usado somente para poucas espécies, que divergiram aproximadamente em um mesmo ponto de um ancestral comum, caso contrário, a abordagem pode tornar-se inclusiva para out-parálogos.

O programa InParanoid foi usado para identificar ortólogos e in-parálogos em comparações pareadas entre os transcritos traduzidos dos SAS do projeto SUCEST com proteomas de outras quatro plantas: Sorgo, Milho, Arroz, Milho e *Arabidopsis*, com parâmetros específicos (*bootstrap*, matriz para similaridade por BLASTP: BLOSUM62, critério de escore *bit*: 40, sobreposição de sequência e grupo: 0.5). Já o programa MultiParanoid foi utilizado em seguida para agrupar as comparações pareadas com InParanoid entre os grupos ortólogos multi-espécies. Deste modo, um grupo ortólogo é composto de todos os ortólogos e parálogos identificados entre todas as espécies de plantas utilizadas.

Para a identificação de ortólogos e parálogos entre as proteínas preditas de cana-de-açúcar do RNA-Seq, e as proteínas preditas das plantas: Sorgo, Milho, Arroz, *Brachypodium*, *Setaria*, *Panicum* e *Arabidopsis*, foi utilizada uma versão atualizada do InParanoid, o Hieranoid (Schreiber et. al., 2013), pela inferência do relacionamento por homologia pareada entre as proteínas preditas, combinados dentro de grupos de ortólogos, com parâmetros específicos (bootstrap, matriz para similaridade por BLASTP: BLOSUM62, critério de escore *bit*: 40, sobreposição de sequência e grupo: 0.5).

O método Hieranoid utiliza como dado de entrada um multifasta para cada espécie a ser analisada e uma árvore conectando as espécies. A árvore pode ser obtida através do *NCBI taxonomy* (Sayers et. al., 2010) ou pode ser gerada pelo próprio usuário no formato Newick. O método infere ortólogos entre as múltiplas espécies aplicando progressivamente o método pareado InParanoid. Os relacionamentos de ortologia são inferidos nos nós de uma bifurcação na árvore de referência. Utilizando a abordagem progressiva hierárquica, o Hieranoid combina as vantagens de métodos baseados em grafos, que são computacionalmente menos expansivo e baseado em árvores filogenéticas, produzindo grupos hierárquicos estruturados em árvores. O InParanoid, e portanto o Hieranoid, calculam o grafo de ortologia a partir do comprimento total do alinhamento da proteína, e requer um comprimento de sobreposição mínimo de 50%. Isto significa que ortólogos com extensivos rearranjos de domínios podem ser perdidos. Ambos InParanoid e Hieranoid utilizam escore bit do alinhamento como um representante para a comparação de espécies rigorosamente próximas para uma distância evolucionária. O Hieranoid utiliza sequências consensus intermediárias para os ancestrais para o escore de distância entre as proteínas, podendo estimar a distância evolucionária com maior precisão, sendo deste modo, em parte mais precisa que o InParanoid. Ao invés de utilizar todas as sequências a cada análise, um sequência consenso é calculada para cada grupo de ortólogos. O consenso é a sequência de resíduos com a maior frequência de ocorrência para cada alinhamento do grupo de ortólogos.

As anotações para os transcritos de cana-de-açúcar realizadas em cada uma das etapas foram inseridas no banco de dados para anotação manual e integração com dados de expressão gênica e às demais ferramentas de consulta. Baseadas nos resultados obtidos e ferramentas de anotação da plataforma *CaneRegNet* foram estabelecidas trinta categorias funcionais para anotação manual dos transcritos (SAS) de cana-de-açúcar pelo grupo do laboratório. Uma categoria secundária e mais flexível foi definida baseada na estrutura do gene ou famílias gênicas (Anexo 1).

5.4. Sequenciamento e Anotação do transcriptoma de Cana-de-açúcar por RNA-Seq

As plataformas de RNA-Seq utilizadas para o sequenciamento do transcriptoma de cana-de-açúcar foram *Genome Sequencer FLX System* (454 Sequencing) e *Ion PGM Sequencing* (Life Technologies). Para o sequenciamento foram utilizadas amostras de folhas e entrenós imaturo e intermediário de plantas SP80-3280, cultivadas em campo por nove meses, e amostras de folhas de plantas *Saccharum officinarum* e *Saccharum spontaneum*, cultivadas em casa de vegetação por onze meses. Cada amostra corresponde a um *pool* de cinco plantas individuais.

A partir das amostras foram construídas bibliotecas *full-length* para o sequenciamento nas duas plataformas, com objetivo de sequenciar os genes completos expressos no transcriptoma das diferentes espécies. Para as amostras sequenciadas com *Ion PGM Sequencing* foi utilizada a primeira fita de cDNA, para o sequenciamento fita-específica, permitindo a identificação e expressão de transcritos antisenso.

A primeira etapa de análise de sequenciamento em larga escala por RNA-Seq foi a filtragem por qualidade dos *reads*. Os dados brutos do Ion PGM foram trimados com o *software TorrentSuite* (versão 3.6.2) por baixa qualidade da base, remoção de adaptadores e *primers*. Os arquivos FASTQ contendo *reads* de qualidade >Q20 foram gerados usando o *software FastqCreator Ion plug-in* (versão 3.6.0). Os *reads* da plataforma 454 Sequencing foram processados e trimados com o *software Prinseq* (Schmieder et. al., 2011). Primeiro, foi aplicado um filtro de qualidade trimando os *reads* usando escore de qualidade média Phred abaixo de 20, trimando as pontas com N's, poli-A/T's, e removendo *reads* com comprimento inferior a 70pb após trimagem. Em seguida, foi aplicado o método para remover sequências de baixa complexidade e homopolímeros.

A segunda etapa e mais importante corresponde à montagem *De novo* dos *reads* em *contigs* (transcritos). Os *reads* processados do Ion PGM e 454 foram combinados e montados utilizando o *software Trinity* (Grabherr et. al., 2011). O *software Trinity* processa os dados sequencialmente através de três módulos, que permitem a recuperação de transcritos isoformas usando o algoritmo de grafos Bruijn (Wjid et. al., 2012). Para a montagem foi utilizado o parâmetro de tamanho mínimo do transcrito de 200pb, com parâmetros padrão para sobreposição e identidade dos *reads* para estender e montar os transcritos.

Após a montagem, os transcritos foram anotados pelos *pipelines* de anotação funcional do transcriptoma de cana-de-açúcar, descritos anteriormente, baseado na

similaridade com o transcriptoma de 6 espécies de gramíneas (Sorgo, Milho, Arroz, Setaria, Panicum e Brachypodium), incluindo o transcriptoma de cana-de-açúcar do projeto SUCEST, o banco do *Gene Ontology* (GO), o banco do Uniprot/SwissProt e NR-NCBI. As análises e metodologias para identificação dos genes *full-length* e classificação dos Fatores de Transcrição são descritos com maiores detalhes adiante.

Para a análise de nível de abundância dos transcritos e identificação de transcritos específicos por espécie e tecido, foi realizada a busca por similaridade com o *software bowtie2* (Langmead et. al., 2012), entre os *reads* processados e os transcritos montados com o Trinity, utilizando os parâmetros de alinhamento *sensitive-local* e *max penalty for mismatch* = 6.

Os transcritos antisense naturais (NATs – *Natural Antisens Transcripts*) atuam na modulação da expressão de outros genes e corresponde a uma importante área de estudo (Lapidot et. al., 2006; Ma et. al., 2006; Zhou et. al., 2008). Com o sequenciamento fita específica no *Ion PGM*, foi possível o estudo e identificação destes transcritos antisense, na montagem *De novo* do transcriptoma de cana-de-açúcar e realizada uma busca comparativa com os transcritos do projeto SUCEST e das seis gramíneas. Para a análise dos antisensos, os *reads* de cada amostra foram mapeados com o *software bowtie2* com parâmetro *very-sensitive*.

Baseado no arquivo (gff3) de mapeamento dos genes do banco de plantas, nos respectivos genomas, foi extraída a informação da orientação do gene. Se o *read* mapeado no gene estiver nas orientações *forward/forward* ou *reverse/reverse*, este é considerado senso; se a orientação for *reverse/forward* ou *forward/reverse* o *read* é considerado antisense.

O mapeamento dos *reads* contra a montagem *De novo* do transcriptoma forneceu uma medida de abundância dos transcritos e permitiu a comparação entre as espécies e tecidos. Para estimar esta abundância dos transcritos, foi utilizado o *software* RSEM (*RNA-Seq by Expectation Maximization*), que não necessita de um genoma de referência, mas apenas o transcriptoma. O *software* estima a abundância e criação de intervalos de credibilidade baseado no alinhamento dos *reads* nos transcritos referência. Ao invés de utilizar o valor típico de RPKM, o RSEM utiliza uma fração dos transcritos produzidos por uma isoforma ou gene ou múltiplos desse valor por 10^6 para obter transcritos por milhão (*TPM – Transcripts Per Million*). O uso de TPM é preferível, pois é independente do comprimento médio do transcrito expresso, fazendo ele mais comparável entre amostras, resultando em melhores estimativas e apresentando melhores resultados em comparação com IsoEM, Cufflinks, rQuant e versões prévias do RSEM (Li et. al., 2011). Para superar modificações no algoritmo EM (*Expectation*

Maximization) e problemas com o uso de uma referência com genes incompletos ou parciais e *reads* únicos, foram calculadas a média e desvio padrão de cada biblioteca de *reads* a serem incorporadas na análise de expressão, baseados no *script rsem_calculate_expression*. Para *reads* únicos é fortemente recomendado utilizar os parâmetros de distribuição de comprimento dos fragmentos (*--fragment-length-mean* e *--fragment-length-sd*). Para dados de *reads* pareados (*paired-end*), o *RSEM* calcula automaticamente a distribuição do comprimento do fragmento a partir dos dados.

Para comparar os níveis de expressão e identificação dos genes diferencialmente expressos entre diferentes condições experimentais foi utilizado o pacote *edgeR* (Robinson et. al., 2010a), que realiza uma normalização do nível de abundância dos genes entre as diferentes amostras com TMM (*Trimmed-Mean of M-values*) (Robinson et. al., 2010b).

Para estudar os padrões de expressão dos transcritos entre as amostras, foram utilizados os transcritos diferencialmente expressos normalizados obtidos com *edgeR*, que apresentaram diferença de intensidade de expressão entre cada par de amostras de pelo menos 4x mais, com significância estatística corrigida para falsa descoberta de $p=0.001$. Foram realizadas análises de agrupamento hierárquico, para agrupar transcritos com perfil de expressão similar, e amostras que tem perfis de expressão similares de acordo com os transcritos, através de *scripts* em *R-statistics*, utilizando pacotes como *NBclust* (Charrad et. al., 2012) e *clust* (Maechler et. al., 2014).

5.5. Predição de transcritos completos (*full-length*) em Cana-de-açúcar

Os estudos de anotação funcional do transcriptoma estão voltados principalmente para a identificação de função dos transcritos. Porém, para saber se um gene corresponde a uma função putativa é importante ter informações estruturais do gene, como a presença de uma proteína completa, para podermos comparar a presença de domínios e regiões conservadas dessa proteína com genes de outras espécies, fornecendo indicativos de que possuem uma mesma função ou pertencem a uma mesma família gênica. Para a predição e identificação de genes completos a partir de um genoma sequenciado, existem diversas abordagens como *ab-initio* (Bonneau et. al., 2001), alinhamentos de ESTs ou proteínas, genômica comparativa com diferentes *softwares* e abordagens. Entretanto, sem um genoma de referência, baseando-se apenas em transcritos a partir de ESTs, ou mais recentemente dados de RNA-Seq, a predição e identificação de genes é uma tarefa mais complexa, sem uma solução única. Deste modo, a predição e identificação dos genes completos é uma etapa importante

no estudo de transcriptoma de organismos não modelos, para estabelecermos o ORFeoma de cana-de-açúcar, podendo contribuir para a categorização de forma mais completa das proteínas e validação da montagem do genoma em andamento e atualização da base GRASSIUS (Yilmaz et. al., 2009). Essa base de conhecimento web que integra fatores de transcrição e promotores de genes entre as gramíneas, e sua geração também foi parte deste trabalho.

Apresentamos a seguir a metodologia para a identificação de transcritos completos que pode ser empregado para outros organismos, desde que possuam transcriptomas sequenciados e anotados de espécies próximas evolutivamente (Figura 13).

Figura 13. Esquema da metodologia desenvolvida para identificação de transcritos completos (*full-length*) composto de três abordagens distintas, utilizando os genes completos de gramíneas (Sorgo, Milho, Arroz, *Brachypodium*, *Panicum* e Setária). Na Análise I, o CDS deve ser totalmente coberto pelo transcrito com melhor *hit*; na Análise II, o gene é totalmente coberto por dois ou mais transcritos; na Análise III o transcrito é classificado como completo se possuir tamanho médio da CDS > 1000pb e apresentar códon de início e fim.

O ORFeoma refere-se ao conjunto completo de ORFs (*Open Reading Frame*) de um genoma, as ORFs correspondem às CDS (sequências codificantes) dos genes. Para

identificar os transcritos de cana-de-açúcar completos, foram utilizados todos os genes (CDS) completos de seis gramíneas como referência: Sorgo, Milho, Arroz, *Panicum*, Setária e *Brachypodium*. Esses genes contêm o mRNA completo preditos por métodos computacionais baseados no genoma.

Após a seleção dos genes completos das gramíneas, foi realizada a busca por similaridade com BLASTN (critério de e-valor $1e-15$) entre cada um dos transcriptomas de cada gramínea contra os transcritos de cana-de-açúcar, utilizando os transcritos como base de dados. Em seguida, foi calculada a cobertura completa para um gene, baseado em múltiplos *hits* a partir de uma saída de BLAST. Para esse cálculo foi implementado uma função que recebe um conjunto de intervalos não sobrepostos e o novo intervalo a ser adicionado, sendo que o intervalo é a região de similaridade do transcrito com o gene e o início e o final do intervalo correspondem às coordenadas de início e fim no gene. A cada *hit* no gene é passada à função o conjunto de intervalos existente e o novo intervalo. Caso um novo intervalo apresente sobreposição com um ou mais intervalos no conjunto, os intervalos são unidos formando um novo intervalo maior, atualizando o conjunto de intervalos e produzindo ao final um conjunto de intervalos únicos representando a cobertura total do gene em estudo. A seguir um exemplo:

Exemplo:

Seja, o conjunto com os seguintes intervalos: {1..75, 110..215, 300..600}, correspondendo aos três primeiros *hits* de um gene.

Dado um novo *hit*: {45..200}, o mesmo será adicionado ao conjunto existente e teremos:

$$\{1..75, 110..215, 300..600\} + \{45..200\} = \{1..215, 300..600\}$$

Fazendo a junção dos intervalos 1..75 e 110..215, para formar um novo intervalo 1..215.

Este algoritmo é útil principalmente para transcritos que foram montados a partir de projetos de transcriptomas, em que muitos dos genes estão incompletos ou particionados em múltiplos transcritos, devido a questões de baixa cobertura de sequenciamento de algumas regiões, ou por motivos biológicos como variantes de *splicing*.

Após o cálculo da cobertura das CDS dos genes e transcritos, os candidatos a genes completos são selecionados apenas se os genes apresentarem cobertura total da CDS > 95% e identidade > 85% e se o transcrito mapeado possuir cobertura >50%,

parâmetros definidos arbitrariamente. A predição dos genes completos é conduzida por três análises, descritas abaixo.

As primeiras duas análises utilizam como referência o alinhamento por BLASTN entre os transcritos de cana-de-açúcar e as CDSs dos genes completos das gramíneas, com os parâmetros de e-valor menor que 1e-15, desabilitando filtro por sequências de baixa complexidade e repetitivas. Na Análise I, o transcrito é classificado como completo se cobre completamente o gene consultado, de acordo com os critérios adotados acima, sendo que os respectivos genes e transcritos são desconsiderados para a análise seguinte. Na Análise II, são selecionados os genes completos de gramíneas se dois ou mais transcritos de cana-de-açúcar cobrem o gene consultado, seguindo os mesmos critérios e desconsiderando os transcritos para a análise seguinte. Nas duas análises, os genes foram selecionados pela similaridade com cada espécie de acordo com a ordem (Sorgo, Milho, *Panicum*, *Setaria*, Arroz e *Brachypodium*), baseado nas espécies com maior número de transcritos com *hits* (Anexo II), ou seja, os transcritos identificados como completos na comparação com Sorgo, não foram utilizados na análise com Milho, e assim por diante. Na Análise III, calculamos o tamanho médio das CDS (1.030,85 Tabela 3) para os genes completos das seis gramíneas. Os *contigs* não classificados nas duas análises anteriores são selecionados como genes completos putativos se apresentarem tamanho da CDS > 1.030,85 e a CDS incluir códons de início e fim, abordagem similar a outra já descrita na literatura (Liu et. al., 2012).

Tabela 3. Número de genes, completos e tamanho médio da CDS em cada espécie do Phytozome v9.0.

Espécies	Número de Genes	Número de genes completos	Tamanho médio CDS
<i>Brachypodium distachyon</i>	31.029	7.441	1.117,50
<i>Oryza sativa</i>	49.061	18.119	1.160,00
<i>Panicum virgatum</i>	70.071	14.062	843,5
<i>Sorghum bicolor</i>	39.441	27.533	1.200,35
<i>Setaria italica</i>	40.599	14.007	911
<i>Zea mays</i>	63.540	26.353	952,75
Tamanho médio			1.030,85

Os transcritos completos preditos em cana-de-açúcar foram comparados com transcritos de outras gramíneas, quanto à sua composição estrutural para identificação e estimativa de tamanhos das regiões CDS e UTR. A predição dos parálogos através do software Hieranoid, permitiu a identificação dos genes completos únicos, agrupando possivelmente alelos, variantes e parálogos de um dado gene.

Os resultados produzidos para os transcritos de cana-de-açúcar foram incorporados à plataforma *CaneRegNet* e ao banco de dados.

5.6. Identificação e Classificação de Fatores de Transcrição em Gramíneas e Cana-de-açúcar

Para estabelecer a arquitetura de redes regulatórias para estudo e identificação de genes envolvidos em diferentes condições e tratamentos, é necessário inicialmente obter as informações dos Fatores de Transcrição (TFs), seus genes alvos no genoma e correspondentes sítios de ligação. Os TFs são os principais elementos para o entendimento dos estudos de redes regulatórias e nas interações entre TFs e ativação ou inibição da atividade de genes e níveis de expressão. A criação de um catálogo de famílias de TFs em cana-de-açúcar e gramíneas é essencial para futuramente construir e estudar redes regulatórias em cana-de-açúcar e comparativamente em outras gramíneas.

Parte deste trabalho foi realizado no estágio no laboratório do Dr. Erich Grotewold da Universidade de Ohio, USA, com o desenvolvimento de uma base de conhecimento (GRASSIUS) que integra informações de TFs e promotores de genes entre gramíneas (Yilmaz et. al., 2009). Neste trabalho foi construído um *pipeline*, para catalogação das regras de domínios e classificação dos TFs para as gramíneas (Cana, Sorgo, Milho e sub-espécies de Arroz), fornecendo uma primeira catalogação e classificação de TFs completa para as gramíneas (Anexo III).

A criação do banco de TFs em gramíneas resultou na criação de diversas etapas para a classificação dos TFs (Figura 14). As informações de genes e anotações foram obtidas a partir de bancos de sequências de genomas, do projeto SUCEST para cana-de-açúcar, e de outros bancos com alguns níveis de informação de TFs de plantas, incluindo PlnTFDB (Riaño-Pachón et. al., 2007), PlantTFDB (Guo et. al., 2008), e DBD (Kummerfeld et. al., 2006).

Figura 14. Diagrama de fluxo descrevendo os passos envolvidos na geração dos Fatores de transcrição.

Baseado nas plataformas PlnTFDB, PlantTFDB, DBD e AGRIS (Davuluri et. al., 2006; Palaniswamy et. al., 2006) foi definido o número de famílias de TFs, com atualização da classificação de TFs dentro de famílias baseadas em regras específicas descrevendo quais domínios são necessários e quais são proibidos ou são permitidos para classificar uma proteína em uma família particular, excluindo famílias de TFs dos outros bancos que podem estar envolvidos na regulação da transcrição e não possuem domínios de ligação ao DNA.

No projeto SUCAST (Souza et. al., 2001), baseado nos transcritos do projeto SUCEST, os primeiros 654 TFs foram identificados por similaridade com TFs conhecidos e por análises de domínios PFAM. As sequências de nucleotídeos e proteínas preditas foram integradas ao banco de TFs conhecidos e agrupados por espécies, resultando em um conjunto não redundante. Para a eliminação de sequências redundantes foi utilizado o módulo Perl "Digest::MD5" consistindo de uma sequência de 32 dígitos hexadecimais que identificam inequivocamente cada sequência de cada espécie. Na segunda etapa, foram realizadas buscas por similaridade com BLAST (BLASTP/BLASTX), para eliminar redundância dentro de cada espécie. As proteínas foram consideradas duplicadas se tivessem cobertura $\geq 99\%$ e identidade $\geq 99\%$, e o alinhamento da sequência consultada ocorresse em até 3 resíduos do início do códon. Se todas as condições fossem satisfeitas, a proteína mais longa era mantida e as proteínas eliminadas eram classificadas como idênticas ou variantes de *splicing*. As proteínas identificadas como

TFs, sem domínio PFAM conhecido ou contendo somente domínios SUPERFAMILY, foram classificadas como “Órfãs”.

Baseado nas informações disponíveis no AGRIS e nos outros bancos de TFs, criamos uma lista de domínios PFAM curada manualmente, que foi usada para gerar a base de dados contendo todas as sequências para cada domínio catalogado no banco do PFAM. Cada domínio PFAM pode ser representado por uma média de 67 sequências de domínios de diferentes espécies, e esta foi a referência para a abordagem para identificação de novos TFs nos respectivos bancos genômicos. Isso foi estabelecido pela coleta de todas as sequências de proteínas preditas de *Oryza sativa indica* (Gramene, <http://www.gramene.org/>) e *japonica* (TIGR5, <http://rice.plantbiology.msu.edu/>), *Sorghum bicolor* (JGI, <http://genome.jgi-psf.org/Sorbi1/>), *Zea mays* (<http://maizesequence.org>) e sequências de nucleotídeos de *Saccharum hybrids* (SUCEST, <http://sucest-fun.org>). Para as três primeiras espécies, foi utilizado o BLASTP, e BLASTX para cana-de-açúcar, para o alinhamento contra o banco de dados de sequências de domínios de TFs, com critério de e-valor 1e-05 e 1e-15, gerando uma coleção completa das proteínas preditas para cada espécie.

A próxima etapa envolveu a busca de domínios com InterProScan contra o banco de domínios PFAM, mantendo apenas os resultados com maior significância, baseado no e-valor 1e-03 e descartando todos os falsos positivos da primeira etapa por similaridade. Finalmente, com a identificação de domínios PFAM para cada uma das sequências, estabelecemos as regras para identificação e classificação dos TFs, conforme as sequências de domínios de ligação-DNA (Luscombe et. al., 2000). Os scripts e regras de domínios podem ser obtidos diretamente no GRASSIUS⁸.

A abordagem desenvolvida foi atualizada para acrescentar novas famílias de TFs e empregada recentemente para a classificação das famílias de TFs dos transcritos de cana-de-açúcar obtidos a partir do RNA-Seq.

5.7. Estudo do Transcriptoma de Cana-de-açúcar por microarranjo e sequenciamento em larga escala

5.7.1. Pipeline para desenho e validação de oligonucleotídeos senso e antisenso

⁸ <http://www.grassius.org>

O projeto SUCEST, permitiu a criação do projeto SUCAST e SUCAMET, a partir dos quais foram criadas lâminas de microarranjo de cDNA representando apenas 4% e 10%, respectivamente do transcriptoma de cana-de-açúcar. Entretanto, para o estudo dos níveis de expressão em um organismo, é crucial possuir uma cobertura ampla do maior número possível de genes, com uma plataforma de microarranjo com alta especificidade e sensibilidade. Adicionalmente, a qualidade de um microarranjo deve garantir reprodutibilidade dos resultados para garantir sua significância estatística. A escolha da estratégia de seleção dos oligonucleotídeos para a plataforma de microarranjo depende do objetivo do experimento biológico, podendo representar uma grande cobertura do transcriptoma buscando novas descobertas, ou hipóteses direcionadas onde os microarranjos são desenhados de acordo com os genes relevantes para um problema biológico importante, sendo uma de nossas etapas para seleção destes genes de interesse. Tomiuk (Tomiuk et. al., 2001) discutem vários critérios e detalhes para uma seleção de *probes* adequada para um microarranjo.

Existem algumas ferramentas para a customização do desenho de oligonucleotídeos para plataformas de micrarranjo, como OligoArray (Rouillard et. al., 2003), porém poucas se baseiam na orientação e permitem adaptação e modificações, e não foram encontradas ferramentas que permitam o desenho de oligonucleotídeos na orientação senso e antisenso para um mesmo gene de forma automática e com critérios específicos. Deste modo, foi desenvolvido neste trabalho o *CaneOligoDesign*, uma estratégia para seleção e customização de oligonucleotídeos de microarranjo, de acordo com critérios específicos e recomendados pela *Agilent Technologies* (Hughes et. al., 2001), incluindo os seguintes passos (Figura 15).

Figura 15. Estratégia do *CaneOligoDesign*, para desenho de oligonucleotídeos e validação, composto por cinco 4 etapas: 1) Seleção dos SAS de interesse pela anotação funcional, e Identificação dos SAS codificadores, não codificadores e da orientação; 2) Desenho de todos os oligonucleotídeos senso e antisenso para cada SAS selecionado, descartando oligonucleotídeos que não atendam aos critérios de hibridação; 3) Seleção dos oligonucleotídeos únicos de acordo com os critérios de especificidade; 4) Classificação dos SAS pelos tipos de oligonucleotídeo com orientação (senso, e o par senso e antisenso).

Os transcritos (SAS) do projeto SUCEST foram utilizados como base para o desenho e seleção dos oligonucleotídeos e criação do microarranjo customizado de oligonucleotídeos de 60-mer no formato 4x44k, sintetizados pela empresa *Agilent Technologies*.

A primeira etapa trata da seleção dos genes com interesse na descoberta e possibilidade de criação de novas hipóteses. Foram selecionados inicialmente todos os transcritos de cana-de-açúcar, resultando em 29.689 SAS, descartando transcritos *singletons* (compostos por apenas um EST) e selecionando transcritos com funções relacionadas a estresse hídrico e acúmulo de sacarose, incluindo os SAS dos microarranjos SUCAST e SUCAMET, de acordo com a anotação funcional do

transcriptoma. Para o desenho dos oligonucleotídeos senso e antisenso é necessário conhecer a orientação dos SAS, porém, existe pouca informação confiável já que os mesmos foram montados a partir de ESTs e a cana-de-açúcar não possui o genoma sequenciado.

Assim, foi recuperada a informação da busca por similaridade entre os SAS e os bancos de proteínas por BLASTX e foi atribuída a orientação aos SAS de acordo com critério de *e-valor* menor que $1e-05$ e com os três primeiros *hits* apresentando mesma orientação. Para os SAS sem similaridade com os bancos de proteínas, foi atribuída orientação pelas predições com ESTScan, e por fim, a orientação foi atribuída de acordo com a maior frequência da orientação dos ESTs sequenciados que compõe os SAS, de acordo com a identificação da orientação dos ESTs sequenciados, sendo as extensões dos ESTs .g (*forward*, orientação 5'→3') e .b (*reverse*, orientação 3'→5').

Após a identificação da orientação dos transcritos, na terceira etapa, as informações da sequência dos transcritos e orientação foram passadas ao programa para o desenho dos oligonucleotídeos senso e antisenso, de acordo com o esquema do programa *CaneOligoDesign* (Figura 16). O algoritmo considera a região 3' como início do SAS e desenha todos os oligonucleotídeos possíveis ao longo do gene nas orientações senso e antisenso. São selecionados preferencialmente dois oligonucleotídeos senso, um próximo à posição 50pb e outro a 350pb em relação ao 3' do transcrito e um antisenso ao mesmo transcrito próximo à posição 50pb do seu respectivo 5', sendo que o início dos oligonucleotídeos selecionados precisam ter idealmente uma distância de 40pb ao redor desta posição pré-determinada. A escolha de um oligonucleotídeo mais próximo à região 3' de um gene pode reduzir a probabilidade de hibridação cruzada por causa da divergência da sequência desta região ser tipicamente maior (Wang et. al., 2003), e raramente afetada por eventos de *splicing* alternativo. A escolha pelo desenho de oligonucleotídeos na porção central da região codificadora possibilita uma anotação mais confiável, porém pode apresentar alto grau de similaridade com outros genes, e para eliminarmos a possibilidade de hibridação cruzada, a etapa final verifica a especificidade do oligonucleotídeo selecionado.

Figura 16. Esquema do desenho dos oligonucleotídeos pelo programa *CaneOligoDesign*.

A quarta etapa do processo, corresponde à seleção dos oligonucleotídeos desenhados de acordo com os critérios de especificidade e sensibilidade estabelecidos a seguir:

- tamanho: 60mer
- orientação: 5' -> 3'
- IUPAC: ACTG
- %G: <50%
- %GC: >= 35% && <= 55%
- Tm: >= 68 && <= 76
- Homopolímero: <= 6
- Localização: 3'

Os oligonucleotídeos desenhados têm tamanho de 60-*mer*, e a orientação dos oligonucleotídeos senso e antisenso foram desenhados a partir dos transcritos. Para serem sintetizados na lâmina de microarranjo precisam estar na orientação 5' -> 3' e seguir o padrão *IUPAC*. A sequência do oligonucleotídeo precisa possuir uma frequência de G abaixo de 50% e uma frequência de GC entre 35% e 55%, e o fabricante recomenda que a frequência de GC esteja em torno de 40%. Embora a *Agilent Technologies* recomende uma *Tm* (*Melting Temperature*) próxima a 80°C para os oligonucleotídeos (Hughes et. al., 2001), utilizamos a frequência de *Tm* de oligonucleotídeos de uma das plataformas proprietárias da empresa, através da fórmula mostrada a seguir. A análise demonstrou que para 91% das sondas comerciais *Agilent*, o *Tm* variava entre 68°C e 76°C. Deste modo, para o desenho dos nossos oligonucleotídeos, utilizamos a mesma fórmula e permitindo uma variação na taxa de *Tm* entre 68°C e 76°C.

Segue, a fórmula para o cálculo da *Tm* para cada oligonucleotídeo:

$$T_m = 64,9 + 41 * \left(\frac{nG + nC - 16,4}{\text{tamanho(nt)}} \right),$$

onde *nG* representam o número de bases *G*, e *nC* o número de bases de *C*, e *nt* os nucleotídeos.

A última etapa corresponde à especificidade do oligonucleotídeo. Diversos estudos para avaliação de hibridação cruzada em lâminas de microarranjo (Hughes et. al., 2001; Wren et. al., 2002), tem mostrado que sequências com menos de 80% de identidade, possuem menos de 20% de hibridação cruzada e que sequências com menos de 70% de similaridade diminuem muito mais a chance de hibridação cruzada. A avaliação da especificidade e busca por alvos alternativos foi obtida por homologia usando BLASTN, com e-valor menor que $1e-15$ entre os oligonucleotídeos selecionados e os ESTs de cana-de-açúcar do Projeto SUCEST. O primeiro *hit* do oligonucleotídeo deveria ter 100% de identidade com o respectivo transcrito e os *hits* seguintes um *score bit* menor que o *cutoff* = 42.1, correspondendo à cobertura do oligonucleotídeo menor que 21pb sem *gaps* (35% de cobertura). Análises estatísticas das frequências do *score bit* e cobertura para o 2º. *hit* do oligonucleotídeo contra os ESTs foram efetuadas para avaliar os critérios adotados (Figuras 17 e 18). Para avaliar o nível de redundância dos oligonucleotídeos selecionados, foram mapeados por BLASTN com critérios de e-valor menor que $1e-15$ e com 100% de cobertura e 100% de identidade, contra o transcriptoma de Sorgo, e foram identificados apenas 30 oligonucleotídeos que apresentaram similaridade em mais de um gene de Sorgo, indicando que o *pipeline* foi capaz de desenhar oligonucleotídeos com alto grau de especificidade.

Figura 17. Representação gráfica da frequência da análise do *Score bit* do segundo *hit* dos oligonucleotídeos mapeados contra os ESTs decana-de-açúcar por BLASTN. A linha vertical vermelha indica a posição do *Score bit* 41.

Figura 18. Representação gráfica da frequência da análise da cobertura do segundo *hit* dos oligonucleotídeos mapeados contra os ESTs de cana-de-açúcar, através do BLASTN.

A lâmina de microarranjo customizada possui formato 4x44k, consistindo de 4 arranjos separados em uma lâmina, com 45.220 oligonucleotídeos cada arranjo. O microarranjo é composto por oligonucleotídeos senso e antisenso desenhados pelo nosso *pipeline*, representando 14.522 SAS, cerca de 40% do transcriptoma de Cana-de-açúcar. O microarranjo customizado é composto por 1.217 Controles da *Agilent*, responsáveis pela correção de sinal de fundo, estimativa do sinal dos pontos hibridados e ajuste do grid para identificação de todos os pontos, e 21.901 oligonucleotídeos duplicados representando os SAS e divididos em oligonucleotídeos senso (13.660 na posição 50pb, 861 na posição 350pb) e antisenso (7.380 na posição 50pb).

5.7.2. Processamento e Quantificação de sinais de fluorescência de microarranjos

O desenho experimental é a primeira etapa para estudos com microarranjos. Se os experimentos não são desenhados apropriadamente, os métodos de análises não serão aptos a obter resultados conclusivos. Porém, precisa haver um equilíbrio entre o alto custo para cada microarranjo e a necessidade de réplicas suficientes para fortalecer os resultados e diminuir a taxa de falsos positivos. A customização de lâminas contendo réplicas de genes e hibridação de réplicas biológicas são essenciais para controlar o ruído introduzido pela localização do oligonucleotídeo na lâmina, pela variação na incorporação dos fluoróforos (marcadores) e para diminuir a variabilidade biológica.

Os métodos empregados e desenvolvidos para pré-processamento e análise do nível de expressão gênica em cana-de-açúcar são baseados nas lâminas customizadas

de cDNA (SUCAST, SUCAMET) e de oligonucleotídeos (*CaneRegNet*), esta última desenvolvida como parte deste trabalho. Todas as lâminas são baseadas em sistemas de duas cores (hibridação competitiva), onde duas amostras distintas são marcadas com um fluoróforo diferente cada e hibridadas juntas.

O desenho experimental adotado para os experimentos de microarranjo do laboratório é composto pela comparação entre uma condição tratada e uma condição controle, hibridadas na mesma lâmina, com a hibridação de uma réplica biológica, com a troca dos marcadores entre as amostras para auxiliar no efeito da variação do sinal entre cada marcador e possibilitar cruzamento dos resultados entre as réplicas biológicas para seleção de candidatos em comum.

5.7.3. Ajuste de Sinal e correção do sinal de fundo

O pré-processamento da imagem de um experimento de microarranjo é uma etapa fundamental para análises computacionais e estatísticas posteriores. Nessa fase são realizados vários processamentos como, extração da intensidade de fluorescência, subtração do ruído de fundo (*background*), filtragem dos elementos com sinal não informativo e a normalização dos canais. Através destes pré-processamentos serão definidos os parâmetros para determinar os genes significativamente expressos em uma determinada condição experimental.

As lâminas de microarranjo de cDNA (SUCAST e SUCAMET) são pré-processadas para a extração e processamento das intensidades de sinal de cada ponto com o *software Array Vision 6.0 (Image Research)*. O método de quantificação da intensidade da fluorescência dos pontos é baseado na medida *Median Artifact Removed Density (MARMDens)* que consiste no cálculo de uma medida de densidade excluindo os pixels que apresentem sinal acima ou abaixo de 4 desvios absolutos da mediana (MAD), auxiliando na exclusão de artefatos da hibridação. A intensidade de fluorescência de cada ponto corresponde ao valor de *MARMDens* subtraindo a mediana do sinal de fundo.

Para as lâminas de microarranjos de oligonucleotídeos esses pré-processamentos foram realizados através do *software Feature Extraction*, versão 10.7 (*Agilent Technologies*) e por *scripts* em Perl e R-statistics (<http://www.r-project.org/>). Foi utilizado o protocolo GE1_107_Sep09 para a identificação dos oligonucleotídeos na lâmina, definição do grid correto das regiões marginais, para reduzir erros sistemáticos originados a partir do viés da marcação pelos fluoróforos, características morfológicas irregulares, hibridação cruzada e diferenças de concentrações das amostras (Delenstarr et. al., 2001).

Para a extração do ruído de fundo, produzindo a intensidade corrigida, não é utilizada a correção pela subtração direta do *background* e ao contrário, é aplicado o algoritmo de eliminação de tendências espaciais (*spatial detrend*), que estima e remove o ruído devido a um gradiente sistemático sobre a lâmina (maiores detalhes no Tutorial⁹). A eliminação de tendências espaciais do fundo consiste de duas etapas: calcular um ajuste da superfície e subtrair este ajuste dos pontos (*spots*). Esse processo é feito independentemente para cada canal vermelho (Cy5) e verde (Cy3) após a subtração do fundo e antes do ajuste de global do fundo.

O cálculo do ajuste da superfície estima o sinal superficial (*foreground*) do dado, medindo a intensidade do ponto após a subtração do sinal médio do mesmo e realiza o ajuste da superfície para o dado selecionado. A principal vantagem desse método é a remoção de variações na intensidade da superfície para diferentes regiões da lâmina. Pela realização da eliminação de tendências espaciais e subtração do ajuste da superfície através da superfície, o dado torna-se mais consistente e reprodutível. Uma correção adicional ao fundo é a aplicação do algoritmo de eliminação de tendências multiplicativas (*MultDetr - multiplicative detrending*), desenhado para compensar variações lineares leves se o processamento não for homogêneo entre as lâminas, resultando em concentrações distintas entre as reações químicas, como por exemplo, entre os cantos e o centro, conhecido como efeito “domo”. O gradiente multiplicativo é removido dividindo o sinal pelo gradiente observado.

Uma etapa adicionada ao protocolo do *software Feature Extraction*, é a análise para testar a significância dos pontos, ou seja, determinar se o ponto não faz parte do ruído do fundo (Delenstarr et. al., 2012; Kafsack et. al., 2012).

A identificação da significância dos pontos é uma etapa importante nas implementações e análises desta tese. Esta informação foi empregada principalmente nas análises dos níveis de expressão dos oligonucleotídeos antisense, que são muito mais baixos comparados aos níveis dos oligonucleotídeos senso e nas análises de enriquecimento e vias para cada condições experimental (Lemke et. al., 2012). Para determinar se a intensidade do ponto é significativa comparado à intensidade do fundo, ou apenas ruído de fundo, a análise é baseada em dois testes *IsWellAboveBG* (“é bem acima do sinal de fundo”) e *IsPosAndSign* (“é positivo e significativo”), os quais são baseados no modelo de erro da *Agilent*, utilizando o erro aditivo como estimativa para o ruído do sinal de fundo.

⁹ http://www.chem.agilent.com/library/usermanuals/public/g4460-90039_featureextraction_user.pdf

A estimativa do erro do fundo para o teste da significância é baseada no modelo de erro do *Feature Extraction*, e o erro aditivo é utilizado para estimar o ruído do sinal de fundo. Os testes *IsPosAndSignif* ("é Positivo E Significante") e *IsWellAboveBG* ("é Bem Acima do Fundo") são baseados no desvio padrão do fundo, com o cálculo do erro aditivo para estimar o ruído de fundo, e na distribuição de probabilidade Gaussiana. A significância da intensidade do sinal comparada com a intensidade do fundo é calculada usando o erro aditivo do modelo de cálculo do erro da *Agilent*. O teste da significância do modelo de erro utiliza a probabilidade Gaussiana para verificar se o sinal é maior que 0 com um erro aditivo; o teste utiliza o sinal do ponto e do erro (assume-se que a distribuição do sinal de fundo está ao redor de 0 com um erro), com um p-valor = 0,01 ($\text{ErroAditivo}/2.6$). Se a probabilidade é menor que ou igual a p-valor=0,01, o sinal é marcado como positivo e significativo ($\text{IsPosAndSign} = 1$), caso contrário é marcado como não significativo ($\text{IsPosAndSign} = 0$). Para determinar se o sinal corrigido está bem acima do sinal de fundo (*IsWellAboveBG*), o sinal é comparado com o ruído do sinal de fundo, com o erro aditivo multiplicado pelo desvio padrão do sinal de fundo, e se o sinal corrigido é maior ele é marcado como bem acima ($\text{IsWellAboveBG} = 1$), caso contrário, é marcado como não acima ($\text{IsWellAboveBG} = 0$). Deste modo, após os cálculos de correção do sinal e testes com o cálculo do fundo e do erro aditivo, se o sinal corrigido for marcado como $\text{IsWellAboveBG}=1$ e $\text{IsPosAndSign}=1$, será definido como significativo e bem acima do sinal do fundo, indicando que o sinal não é ruído e representa o nível de expressão real de um gene. Essas medidas são armazenadas no arquivo de saída do *software* e servem como entrada para a submissão do experimento na plataforma *CaneRegNet*.

5.7.4. Normalização de microarranjos

Após o cálculo e correção do sinal, é realizada a normalização. A normalização dos pontos é utilizada para correção de viés devido às variações na incorporação dos fluoróforos e intensidade da fluorescência entre os marcadores Cy3 e Cy5 (Quackenbush, 2001; Yang et. al., 2002), funcionando como um processo para estimar com maior precisão a razão entre as intensidades. Duas normalizações foram adotadas para as lâminas customizadas, uma normalização Linear e posteriormente uma normalização por LOWESS (*Locally Weighted polynomial regression*). O método de normalização Linear por fluoróforo assume que o viés dos sinais entre os marcadores (canais) é independente da intensidade e, deste modo, utiliza uma abordagem global para a normalização dos marcadores. O fator de normalização linear é calculado por

canal e é definida a média geométrica da intensidade do sinal dos sinais para normalização igual a 1000:

$$FatorDeNormalizaçãoLinear = \frac{1000}{10^{\left(\frac{1}{n} \sum_{i=1}^n \log X_i\right)}}$$

Onde X é o sinal corrigido do ponto e n é o número de pontos utilizado para a normalização.

Geralmente, as intensidades do fluoróforo Cy5 (vermelho) tende a ser maior que as intensidades do Cy3 (verde), mas a magnitude da diferença geralmente depende da intensidade global (Davis, et. al., 2006). As razões para variação entre os marcadores podem ser devido à diferença de incorporação dentro das amostras, pela diferença na resposta de emissão para a excitação do laser em diferentes frequências, diferenças na leitura do foto multímetro e diferenças nas várias áreas da lâmina, resultando em variações de foco.

O método empregado para normalização e eliminação de variações entre marcadores é o método de ajuste não paramétrico LOWESS (Cleveland et. al., 1979; Cleveland et. al., 1988). Uma das grandes vantagens do método é que não é necessário especificar um tipo particular de função a ser usada como modelo. O método assume que o viés do fluoróforo é dependente da intensidade e deste modo realiza uma abordagem local para normalização do fluoróforo. O fator de normalização do fluoróforo é calculado ajustando a curva de regressão linear localmente ponderada para os pontos escolhidos para normalização. A quantidade de viés por marcador é determinada a partir da curva para a intensidade do ponto. Cada ponto recebe um fator de normalização de fluoróforo por LOWESS diferente por canal. O método LOWESS corrige o log da razão do dado, de modo que a tendência central após a normalização por fluoróforo situa-se ao longo de zero para todas as faixas de intensidade, assumindo um número igual para oligonucleotídeos induzidos e reprimidos em qualquer faixa de sinal. O fator de normalização de fluoróforo por LOWESS é derivado para cada canal pelos procedimentos descrito a seguir¹⁰:

a – Uma curva de regressão linear é ajustada para o gráfico de M vs A , onde $M(\text{eixo } x) = \text{Log}(\text{Cy5/Cy3})$ e $A(\text{eixo } y) = \frac{1}{2} * \text{Log}(\text{Cy5} * \text{Cy3})$. Cy5 e Cy3 são os sinais corrigidos. Essa curva LOWESS é ajustada através da tendência central de gráfico de M vs A é definida como M_{fit} e é uma função de A .

¹⁰ http://www.chem.agilent.com/Library/usermanuals/Public/G4460-90026_FE_Reference.pdf

b – O passo de normalização do fluoróforo transforma o dado, de modo que a tendência central de *Mfit* para cada valor de A é deslocado para ser igual a zero.

c – Após o fator de correção ser definido para todos os pontos ele é dividido uniformemente sobre os canais Cy5 e Cy3. Os novos sinais após correção Cy5' e Cy3' são obtidos a partir da transformação dos valores de Cy5 e Cy3 originais:

$$\text{Cy5}' = \text{Cy5} * (10^{MFit/2}) \text{ e } \text{Cy3}' = \text{Cy3} * (10^{MFit/2})$$

d – Se o log da razão original situa-se exatamente ao longo da linha de ajuste *Mfit*, a nova razão por log é ajustada para zero:

$$\text{Se } \log(\text{Cy5}/\text{Cy3}) = MFit, \text{ então } \log(\text{Cy5}) = \log(\text{Cy3}) + MFit$$

$$\text{ou } \log(\text{Cy5}' * 10^{MFit/2}) = \log(\text{Cy3}' * 10^{-MFit/2}) + MFit$$

$$\text{ou } \log(\text{Cy5}') + MFit/2 = \log(\text{Cy3}') - MFit/2 + MFit$$

$$\text{ou } \log(\text{Cy5}'/\text{Cy3}') = 0$$

e – O fator de normalização do fluoróforo por LOWESS para Cy5 é $1/(10^{M'/2})$ e para Cy3 é $10^{M'/2}$.

Após a normalização por LOWESS de cada canal a intensidade normalizada é salva nas colunas denominadas: *gProcessedSignal* (sinal processado Cy3) e *rProcessedSignal* (sinal processado Cy5). Esses valores são utilizados pelo *pipeline* de processamento das lâminas de microarranjo da plataforma, para os cálculos de expressão dos genes significativamente expressos e diferencialmente expressos, análises subsequentes e consultas.

Existem alguns métodos de análises de vias e categorias utilizando dados de expressão gênica e que podem ser aplicados a outras tecnologias de larga escala. Um dos objetivos desta tese foi construir bases de conhecimento em cana-de-açúcar, baseado em anotações funcionais e ferramentas integradas à plataforma *CaneRegNet*.

5.7.5. Identificação de genes significativamente expressos

A aplicação de métodos computacionais e estatísticos para análise de dados biológicos é muito mais exploratória do que uma solução estatística. As melhores conclusões analíticas são feitas com a contribuição do conhecimento biológico dos pesquisadores, aliado a bancos de dados com anotações funcionais e algoritmos computacionais e análises de enriquecimento derivados de métodos estatísticos.

Os dados transformados, resultantes de uma etapa inicial para processamento das lâminas de microarranjo, são reportados como níveis de expressão para cada gene,

como razão de expressão ou como logaritmo da razão de expressão. A razão de expressão é simplesmente o valor normalizado do nível de expressão para um particular gene entre uma condição tratada dividida por uma condição controle. Esses dados são utilizados para diferentes tipos de análise, como identificação de genes significativamente expressos (SE), diferencialmente expressos (DE), análises funcionais de enriquecimento, análises de vias significativas estatisticamente entre diferentes condições experimentais ou entre múltiplas condições experimentais (Khatri et. al., 2012; Huang et. al., 2009). Para ampliar o ganho no entendimento biológico, é necessário integrar às análises a anotação funcional dos genes ou grupos de genes. Esses tipos de análises ampliam a probabilidade dos pesquisadores de identificar processos biológicos mais pertinentes aos seus estudos. Todos os métodos discutidos aqui são independentes da tecnologia usada para obter os dados, e foram implementados e integrados à plataforma *CaneRegNet* para estudo do transcriptoma de cana-de-açúcar.

Uma variedade de técnicas pode ser usada para identificar genes com expressão alterada em uma ou mais condições experimentais. Neste trabalho foram utilizadas e integradas à plataforma *CaneRegNet*, duas abordagens distintas para a seleção de genes expressos.

A primeira abordagem é baseada no critério de que somente aqueles genes que podem ser detectados acima dos níveis de sinal de fundo, os genes SE, deveriam ser considerados para análises posteriores, e esta abordagem tem sido empregada e validada em diversos estudos (Lembke et. al., 2012; Delenstarr et. al., 2001 Kafsack et. al., 2012; Peterson et. al., 2005). Se o nível de expressão de um gene é baixo e não pode ser distinguido do sinal de fundo em qualquer uma das amostras, a identificação de genes DE (induzidos ou reprimidos), baseado no *fold change*, deveria confirmar se os genes possuem um nível de expressão significativo entre as amostras. Porém, uma preocupação é que após a etapa de normalização em experimentos por microarranjo, a informação do nível de expressão absoluta é perdida. Além disso, medir o nível absoluto do sinal de fundo é difícil, pois existe variação na hibridação em diferentes áreas da lâmina, áreas com DNA podem apresentar nível maior, podem variar de oligonucleotídeo para oligonucleotídeo, depender do conteúdo G-C, e pelas diversas medidas para normalizar o sinal de fundo. Deste modo, a maioria dos pontos na lâmina possui um sinal detectado em uma faixa próxima do sinal de fundo e eliminar os genes em uma análise inicial, pode proporcionar uma grande perda de informação.

Em nossa primeira abordagem para a identificação de genes SE, adotamos um dos métodos incorporados pela *Agilent Technologies* em suas lâminas de microarranjo, como descrito anteriormente, para detectar com maior precisão o sinal de fundo e permitir a aplicação de medidas estatísticas e de erro para determinar se um gene é significativamente expresso e com sinal acima do sinal de fundo. A implementação dessa abordagem na plataforma *CaneRegNet* permitiu estudo de genes significativamente expressos, que poderiam não ser identificados por análises de razão de expressão. Esses resultados são utilizados pelas ferramentas desenvolvidas para diversas análises da plataforma, entre elas a análise de enriquecimento de termos, e a análise de escore de atividade das vias entre diferentes condições experimentais e estudos específicos (Bossers et. al., 2010). A identificação de genes SE mostrou-se eficiente, principalmente para o estudo dos genes antisenso, por apresentarem um nível de expressão muito mais baixo em relação aos genes expressos na orientação senso, sendo em sua maior parte descartados em análises para identificação de genes DE (Figura 19 e 20).

Figura 19. Correlação entre os sinais corrigidos dos fluoróforos Cy3 e Cy5 para os oligonucleotídeos senso (azul), antisenso (vermelho) e controles da lâmina (amarelo).

Figura 20. Gráfico do logaritmo da razão de expressão dos oligonucleotídeos senso (verde) e antisenso (laranja) vs logaritmo da média dos sinais. A linha vermelha delimita o intervalo de credibilidade com $p=90\%$, identificando os genes induzidos/reprimidos pelo método Htself adaptado.

5.7.6. Identificação de genes diferencialmente expressos

Para determinar se um gene é diferencialmente expresso ou não, baseado na razão de expressão é necessário determinar um corte ou *cutoff* para a razão de expressão, o que corresponde matematicamente a testar a hipótese nula H_0 : “o gene não possui variação diferencial entre as duas amostras”. Os testes de hipóteses estatísticas (Box et. al., 1969; Montgomery et. al., 1983) permitem responder questões particulares relacionadas a uma ou mais populações com base em amostras randômicas destas populações, e existem diversos modelos empregados para estudos da variação da expressão gênica (Vêncio et.al., 2005; Montgomery et. al., 1983).

A segunda abordagem adotada e integrada à plataforma *CaneRegNet*, corresponde à análise para identificação de genes DE. Para definir um gene DE, podemos adotar um critério de corte na medida do nível de expressão ou utilizar testes estatísticos (Quackenbush et. al., 2001; Chen et. al., 1997; Pan et. al., 2002). Para essas análises foi empregado o método *HTSelf* (Vêncio et. al., 2005), que é aplicado a uma lâmina comparando as duas amostras, marcadas com diferentes fluoróforos, onde $M = \log_2(\text{Cy3}/\text{Cy5})$, $A = \log_2(\text{Cy3})/2 + \log_2(\text{Cy5})/2$. É possível notar que as observações de M se distribuem em torno de $M = 0$, e M aumenta para os menores valores de A, ou seja, existe uma dependência com a intensidade média. A curva vermelha é o *cutoff* dinâmico de 90% de credibilidade (Figura 20).

O método *HTSelf* propõe uma alternativa para o problema de escassez de réplicas biológicas e técnicas para um experimento de microarranjo. São utilizadas duas abordagens pelo método *HTSelf*, a determinação de valores críticos (*cutoff*) dependentes da intensidade derivados de um banco de dados normativo, conhecidos por experimentos *self-self*, e o uso de métodos não-paramétricos (Vêncio et. al., 2005).

A abordagem para definir o *cutoff* baseia-se em estratégias experimentais, como o uso de bancos de dados normativos compostos de hibridações *self-self*. Experimentos *self-self* são experimentos normativos (controle) onde o mesmo material biológico é marcado com ambos os fluoróforos Cy3 e Cy5 e são hibridados juntos na mesma lâmina. Essa abordagem determina o *cutoff* dependente da intensidade para classificar os genes como diferencialmente expressos. O método *HTSelf* utiliza um banco de dados normativo de experimentos *self-self* para definir a função densidade de probabilidade (fdp) nula e procurar pelos extremos (*outliers*). Para levar em conta a dependência da variabilidade com a intensidade, a fdp nula é estimada usando-se um processo de janela-deslizante que se desloca por toda a faixa de intensidades. O algoritmo utiliza uma técnica não paramétrica, a Estimativa de Densidade Kernel, que estima a função densidade de probabilidade de uma variável empiricamente, com base em amostras observadas desta variável (Silverman, 1986), uma técnica que não necessita assumir um modelo, mas necessita de parâmetros para determinar o início, fim, tamanho da janela e tamanho do passo a ser percorrido na faixa da intensidade do gráfico. Uma vez definido o *cutoff*, é produzida uma curva (vermelha) sobre a faixa de intensidades (Figura 18), definindo um limiar de corte para os genes com razão de expressão induzidos e reprimidos, e o mesmo é aplicado a um experimento com duas amostras uma tratada e um controle, essa última é utilizada para a hibridação *self-self* para a definição da curva de *cutoff*.

Os desenhos experimentais adotados pelo nosso grupo e no *pipeline* de análise de experimentos de microarranjo são baseados na hibridação de um experimento e uma ou duas réplicas biológicas, correspondendo à hibridação de uma amostra tratada contra um controle, e uma réplica biológica com inversão da marcação dos fluoróforos entre as amostras. Como são realizados muitos experimentos independentes, com diferentes condições experimentais e geralmente são restritos a poucas réplicas biológicas, o custo para uma hibridação adicional de uma lâmina *self-self*, para cada experimento seria impraticável. Deste modo, adotamos o método *HTself*, para as análises de identificação de genes diferencialmente expressos, porém o método para a determinação do *cutoff*, é aplicado diretamente sobre uma faixa de intensidades (M x A) obtida através do próprio experimento entre uma amostra tratada e um controle, e em seguida

classificando os genes em DE induzidos, reprimidos e não diferenciais, adotando critério adicionais para determinar se o gene é DE em uma condição experimental, se:

- possui pelo menos 70% das réplicas para um mesmo gene dentro da lâmina com mesmo perfil de expressão (induzido, ou reprimido);
- estão presentes em todas as réplicas biológicas.

Os estudos com lâminas de microarranjo por cDNA e oligonucleotídeo tem validado essa abordagem e alcançado uma taxa de validação experimental de 85% (Lembke et. al., 2012; Hotta et. al., 2013; de Souza et. al., 2008).

5.7.7. Análise de Co-expressão entre condições experimentais

A correlação entre duas variáveis M e N é a medida cujos valores variam juntos sistematicamente (Box et. al., 1969). O coeficiente de correlação de Pearson ou coeficiente de correlação produto-momento é calculado a partir de duas amostras M e N como a razão entre suas covariâncias e o produto de seus desvios padrão, indicando a medida cuja relação entre as variáveis M e N é linear. O coeficiente de correlação de Pearson mede o grau de correlação linear entre as duas variáveis quantitativas. É um índice adimensional com valores situados entre -1 e 1 inclusive, que reflete a intensidade de uma relação linear entre dois conjuntos de dados. O coeficiente obtido a partir das amostras (s) é uma estimativa da verdadeira correlação (ρ) da população normal bivariada (M, N). A hipótese nula $\rho = 0$ pode ser testada com a estatística:

$$t = \frac{s * \sqrt{n - 2}}{\sqrt{1 - s^2}}$$

A qual segue a distribuição de t-estudante com (n - 2) graus de liberdade. O p-valor desse teste é a probabilidade de que o coeficiente de correlação de Pearson da amostra (s) seria ao menos tão extremo quanto observado se a hipótese nula é verdadeira ($p = 0$).

A plataforma *CaneRegNet* permite o estudo de coexpressão de genes entre múltiplas condições experimentais de microarranjo de diferentes estudos. Para a análise utilizando o coeficiente de correlação de Pearson, todos os genes que são comuns a todas as condições experimentais e que possuem valor de expressão não nulo na consulta são identificados e usados para a análise. De acordo com o número de experimentos selecionados é adotado um critério de corte para o coeficiente, baseado na tabela estatística de valores críticos (Peaeson et. al., 1976). Os resultados são

exibidos em formato de tabela, integrando informações do nome do gene, anotações funcionais, valores de expressão em cada condição experimental e uma coluna com o nível de correlação para avaliação da coexpressão para cada gene entre as diferentes condições experimentais. Após a seleção dos genes é realizada uma etapa de normalização dos genes pela centralização da mediana, sendo possível adotar outros métodos como centralização da média e *Z-score*, para que os resultados possam ser analisados por métodos de agrupamento. A lista obtida de genes que podem ser vistos como co-expressos entre diferentes condições experimentais é automaticamente cruzada com a lista de genes DE pelo método *HTSelf* adaptado, e produz uma informação adicional para identificar os genes co-expressos entre as diferentes condições e classificados como DE.

5.7.8. Análise de Enriquecimento de Termos

As análises para detecção de genes SE ou DE produzem listas de genes, as quais são difíceis de serem analisadas e não trazem informação relevante sem conhecimento biológico agregado. Uma importante abordagem de estudo são as análises de dados de expressão gênica incorporando anotações funcionais dos genes. Essas abordagens avaliam estatisticamente a fração de genes em uma particular categoria funcional ou via metabólica encontrada entre um conjunto de genes com expressão diferencial alterada, permitindo estudar os processos biológicos como um todo e não pontualmente. Os bancos ou vocabulários de anotação adotados nas implementações deste trabalho, baseados na anotação funcional do transcriptoma de cana-de-açúcar são:

- Gene Ontology,
- Vias Metabólicas e Enzimas do KEGG (*Kyoto Encyclopedia of genes and genomes*) e
- Categorização Funcional, construída no laboratório contendo uma lista de trinta categorias funcionais em cana-de-açúcar (Anexo I).

A abordagem implementada na plataforma *CaneRegNet* para estudo do transcriptoma por microarranjo foi o *GeneMerge* (Castillo-Davis et. al., 2003), tendo como principais vantagens: 1- ser código aberto e facilmente integrável a qualquer plataforma ou *pipeline* de análise de dados de expressão por microarranjo ou RNA-Seq; 2- facilidade para incorporação de bancos de dados de categorizações funcionais públicas ou de organismos não modelo particulares. Essa abordagem realiza as análises,

baseando-se no tamanho populacional (número de genes) representado na lâmina de microarranjo e não no tamanho total do transcriptoma, e permite detectar pequenas, porém consistentes mudanças em conjuntos de genes funcionalmente relacionados. Mais especificamente o método busca encontrar um conjunto de genes, que são mais frequentemente DE, do que o resto do transcriptoma, ou da população de um microarranjo.

O *GeneMerge* retorna uma informação descritiva a respeito do gene sob investigação e é baseado em escores de ranqueamento em relação a super-representações potenciais das categorias em um dado conjunto de genes. Para a execução do algoritmo são necessários 4 arquivos de entrada:

- 1 – a lista de genes (SE ou DE), ou conjunto de estudo particular.
- 2 – a lista da população de genes, de acordo com o conjunto de estudo (microarranjo).
- 3 – um arquivo de associação por ID (identificação) entre os genes e as categorias funcionais.
- 4 – um arquivo com a descrição das categorias funcionais.

Para os estudos de microarranjos, por exemplo, o conjunto de estudo pode ser a lista dos genes induzidos em uma determinada condição.

O arquivo de saída é gerado em formato texto delimitado por tabulações, que pode ser facilmente aberto em planilhas ou ser processado por *scripts* para geração de tabelas customizadas e gráficos, como é realizado na plataforma *CaneRegNet*. A tabela contém as categorias funcionais, associadas a cada um dos genes no estudo e escores do ranqueamento para as categorias super-representadas.

O escore de ranqueamento para as categorias super-representadas é obtido utilizando-se a função hipergeométrica. A distribuição hipergeométrica descreve a probabilidade discreta de selecionar r itens de um tipo em uma amostra de tamanho k de uma população de tamanho n , onde p é igual à proporção dos itens do tipo n na população, e a amostragem é sem substituição (Sokal et. al., 1995)

$$P(r|n, p, k) = \frac{\binom{pn}{r} \binom{(1-p)n}{k-r}}{\binom{n}{k}}$$

Deste modo, a função hipergeométrica fornece uma quantificação do nível de uma “surpresa” em encontrar a super-representação para um item específico em uma certa amostra de tamanho k amostrada de uma população maior, de tamanho n . A

função hipergeométrica fornece a probabilidade exata da amostragem de r genes com uma categoria particular de uma amostra de tamanho k de uma população de tamanho n dado que a categoria existe na fração p no conjunto populacional de genes. Como o *GeneMerge* acessa a super-representação para todas as categorias dentro de um dado conjunto de genes de estudo, uma correção é necessária para contabilizar para super-representação que invariavelmente ocorrerá por acaso quando múltiplos testes são realizados. Para esta correção é aplicado o método de correção de Bonferroni (Bland et. al., 1995), com uma correção baseado no número de termos examinado em cada análise.

O escore de super-representação para categorias representadas no conjunto de estudo por apenas um gene, não são calculados. Os escores corrigidos e não corrigidos são adicionados no arquivo de saída.

5.7.9. Análise de escore de classes funcionais

Os métodos de análise de enriquecimento de categorias em determinados experimentos fornecem informações relacionadas com as categorias mais representadas em uma determinada condição experimental, baseada na lista de genes SE ou DE, ou uma lista particular em estudo. Porém, esses tipos de métodos não levam em consideração os níveis de expressão de cada gene, e grandes variações nos níveis de expressão de alguns genes podem ter efeitos significativos em uma via metabólica, mas níveis de expressão fracos. Entretanto, se coordenados em conjuntos de genes funcionalmente relacionados podem ter um efeito muito mais significativo (Khatri et. al., 2012).

Dentro das abordagens da classe FCS (Functional Class Scoring), foi desenvolvido um *pipeline* integrado à plataforma *CaneRegNet*, para comparar a atividade das vias metabólicas em diferentes condições experimentais de microarranjo ou RNA-Seq. Utilizando os níveis de expressão gênica para cada condição experimental, o programa calcula o escore de atividade das vias metabólicas, representando as potenciais atividades das vias e permitindo suas comparações entre as diferentes condições. O *pipeline* permite a análise comparativa entre duas ou múltiplas condições experimentais aplicando diferentes métodos estatísticos utilizando o pacote *R-statistics* (Ripley, 2001).

Para a análise comparativa entre duas condições experimentais aplicamos o teste não-paramétrico de Wilcoxon rank-sum ou Mann-Whitney, considerando amostras não pareadas (Wilcoxon et. al., 1970), seguido por agrupamento das vias com quantil

(Hyndman et. al., 1996), divididas em grupos de (0, 5, 25, 50, 75, 95 e 100%). Para a análise entre múltiplas condições é realizada a análise de agrupamento hierárquico (Draghici et. al., 2003), seguido pela determinação do número relevante de grupos com o pacote NbClust (Charrad et. al., 2012). Uma das vantagens do método de agrupamento hierárquico é que ele é um método não supervisionado e, não necessita definir inicialmente o número de grupos, tem o objetivo de agrupar objetos similares em grupos, construindo dendogramas para representar a hierarquia dos grupos, fornecendo informações de possíveis relações entre os grupos. Para realizar o corte no dendograma para obter um resultado por grupos, utilizamos o pacote NbClust com distância euclidiana para estimar a matriz de dissimilaridade e o método *average* para a análise de agrupamento, porém a ferramenta permite a seleção de vários outros métodos como ligação simples, ligação completa, ligação média, *wards*, etc. A distinção entre eles depende de como a distância entre os grupos é calculada e da dispersão das intensidades. Também é realizada uma análise de correlação de Spearman entre as amostras (Kotlyar et. al., 2002), possibilitando um estudo do nível de correlação entre as condições experimentais baseado no escore das vias.

O algoritmo desenvolvido para estimar o escore de atividade das vias pode ser dividido em cinco etapas:

- 1 – Seleção do arquivo, que associa os genes e as respectivas vias metabólicas, gerado a partir da anotação funcional.
- 2 – Todas as vias associadas a um gene no conjunto de estudo são selecionadas.
- 3 – Cada via selecionada recebe um escore baseado na abundância relativa do nível de expressão do gene para o qual está relacionado.
- 4 – É calculada a porcentagem de genes identificados em cada via, baseado no número total de genes associados a cada via e o total de genes na via.
- 5 – Calcula-se a soma dos escores para cada uma das vias, para obter o escore total da via, normalizando pela proporção de genes detectados em cada via e o total de genes da via. O escore normalizado de cada via representa o nível de atividade na condição de estudo.

Assim, o escore de atividade da via é calculado por:

$$PA(P) = \frac{(g1 + g2 + g3 + \dots + gN)}{M} * N$$

Onde, g_j é a intensidade do gene j identificado na via P , N é o número total de genes detectados na via P , M é o total de genes que fazem parte da via P .

No estudo da atividade das vias metabólicas, se uma via metabólica está mais ativa em certa condição experimental, o nível de expressão médio dos genes nessa via

tem a probabilidade de ser relativamente maior. Para uma variação no cálculo do escore devido à diferença entre o número total de genes em cada via, normalizamos o escore de atividade da via pela porcentagem de genes detectados em cada via, em relação ao total de genes dessa via. Como o nível de expressão dos genes passou por várias normalizações e não é absoluto, podemos apenas prever que a via A é mais ativa que a via B, e temos maior interesse nas vias que apresentam perfil de atividade mais parecido ou mais discrepante entre diferentes condições experimentais. A comparação de atividade das vias entre duas ou mais condições experimentais pode fornecer grande suporte para novas hipóteses.

5.7.10. Análise de vias metabólicas em grafos

Neste desenvolvimento foi empregado o Neo4j, um sistema de gerenciamento de dados de grafos, código aberto e completamente desenvolvido em Java (Angles et. al., 2012). Esta estrutura tem o objetivo de fornecer métodos de consultas que não tratam apenas dos objetos armazenados, mas também da estrutura do grafo. O exemplo mais tradicional é a operação de travessia, sendo a forma mais simples que pode ser usada para obter os vizinhos de um objeto em particular armazenado no banco.

A partir das informações armazenadas no banco de dados relacional, descrevendo as vias metabólicas do KEGG, foi desenvolvido um script Perl para extrair as informações, ou obter maiores informações diretamente do site do KEGG via REST API e produzir um arquivo no formato Cypher, para armazenamento das vias, enzimas, compostos, reações, genes homólogos e níveis de expressão em experimentos de microarranjos. O Cypher é uma linguagem declarativa de grafos que permite consultas eficientes e expressivas, e atualizar o banco de dados de grafos, para inserção das vias metabólicas e suas relações, porém o Neo4j oferece alternativas para inserção e atualização como no caso de grandes volumes de dados por arquivos em formato tabular, via Java REST API. Abaixo alguns modelos para a criação dos nós e arestas com múltiplos atributos, para inserção no banco de grafos:

- Modelo para criação de nós:
 - o **CREATE (nó1:Label { atributo1:'3.1.1.4', atributo2:'phospholipase', atributo3:'R01317'})**
- Modelo para criação de aresta de relacionamento entre enzima e composto:
 - o **CREATE (nó1)-[:TIPO_RELACIONAMENTO { atributoA:'R01317', atributoB:'r', atributoC: 'reaction' }]->(nó2)**

Há um grande número de operações que podem ser aplicadas a sistemas de bancos de grafos e em diferentes níveis de complexidade e outros aspectos. As relações entre nós são principalmente exploradas por operações de travessia, seja por uma travessia simples fornecendo os vizinhos de um nó, ou operações complexas explorando grandes partes e caminhos do grafo.

As ferramentas para consulta ao banco de grafos foram desenvolvidas para a integração de dados biológicos heterogêneos permitindo a travessia entre diferentes fontes de dados. Elas permitem travessia através de vias metabólicas, vias de transdução de sinal e perfis moleculares como medidas de expressão gênica, podendo ser facilmente adaptadas para medidas de proteômica e inclusive medidas de metabolômica. As ferramentas permitem a visualização gráfica dos caminhos ou nós recuperados na consulta através das vias metabólicas e vias de sinalização do KEGG, pela utilização do KEGG REST-API de forma dinâmica.

As ferramentas de análise no banco de grafos, foram implementadas em linguagem de programação Java, para consultas e travessias nas vias metabólicas, integrando dados de expressão gênica por microrarranjo às proteínas (enzimas), definindo condições de consulta, como direção dos caminhos no grafo, tipos de relacionamentos entre os nós e identificação dos níveis de expressão em condições experimentais específicas. As travessias de dados utilizam nós de proteínas (enzimas), com os respectivos nós genes relacionados pelo tipo de relacionamento *homolog*, como nós centrais para a construção das redes metabólicas integradas. Mais especificamente, em ordem para construir uma rede integrada consistindo de vários tipos de arestas, os tipos de arestas *reaction_with*, são unidos através dos compostos (substratos/produtos) produzidos pela interação da enzima com o substrato (composto) em uma ordem específica da travessia produzindo um produto (novo composto), que poderá ser o substrato de uma outra enzima e assim sucessivamente, conectando toda a rede de interações.

A fim de estudar a topologia de vias metabólicas em um nível de módulos funcionais e para facilitar as análises e interpretações de perfis de dados moleculares como dados de expressão gênica via conectividade de redes integradas, descrevemos abaixo as principais operações empregadas nas ferramentas de consulta desenvolvidas e refletem alguns dos requisitos de consulta no sistema de banco de dados de grafos:

- Criar Índices. Indexando um nó pela propriedade chave. Esta operação é realizada antes da inserção dos elementos no banco de grafos.

- O *framework* de travessia da API Java para o Neo4j consiste de algumas interfaces com adição de nós e relacionamentos: *TraversalDescription*, *Evaluator*, *Traverser* e *Uniqueness*, são os principais. A interface *Path* tem um propósito especial nas travessias, pois é utilizado para representar uma posição no grafo quando avaliando esta posição. Além disso, a interface *PathExpander* é central para as travessias. Existe também um conjunto de interfaces para usos avançados, quando é necessário um controle explícito da ordem de travessia: *BranchSelector*, *BranchOrderingPolicy* e *TraversalBranch*.

- Travessia. Uma busca simples em largura ou profundidade até a profundidade de N a partir de um nó inicial. Os tipos de objetos retornados podem ser divididos em: nó, relacionamento, caminho contendo representações completas dos nós de início e fim, caminho completo contendo representação completa de todos os nós e relacionamentos. Para decidir como o grafo deve ser percorrido, a ordem de visita pode ser em profundidade ou largura. A direção dos tipos de relacionamentos podem ser: entrada, saída ou todos. A singularidade pode ser de vários tipos: *none*, *node_global*, *relationship_global*, *node_path*, *relationship_path*, *node_recent*, *relationship_recent*¹¹. Podem ser definidas medidas de corte para que a travessia seja interrompida, e o nível máximo de profundidade.

- Caminho mais curto. Esta operação calcula o caminho mais curto entre dois nós, à distância máxima de 15 relacionamentos. Podem ser definidos uma única ligação, o nó inicial, o relacionamento de ligação e o nó final. Características descrevendo o relacionamento como tipo de relacionamento e direções são todos usados na busca pelo caminho mais curto. Pode ser marcado um caminho como opcional.

- Encontrar nós vizinhos. A operação de travessia mais primitiva, para encontrar o vizinho mais próximos a partir de um nó inicial, definindo o tipo de relacionamento.

A importância para a comparação entre duas condições experimentais é baseada pelo número de caminhos associados com a enzima selecionada, o comprimento médio e o número médio de nós do tipo enzima ativos. Foram utilizadas estas medidas para determinar a importância de uma enzima, de acordo com estudos prévios em doenças indicando que estão associados ao grau (número de interações) e comprimento médio dos caminhos ativos são características significantes para determinar genes que causam doenças (Aoki et. al., 2007; Xu et. al., 2006; Wei et. al., 2006).

¹¹ <http://neo4j.com/docs/stable/javadocs/org/neo4j/graphdb/traversal/Uniqueness.html>

Uma das ferramentas de análise desenvolvidas baseia-se na seleção de uma via metabólica, uma enzima de interesse e duas condições experimentais de microarranjo, para a identificação dos caminhos em qualquer direção passando pela enzima, os caminhos a partir da enzima e os caminhos chegando na enzima na via de interesse, determinando o total de caminhos contendo a enzima de interesse, o nível de expressão médio e a porcentagem de nós (genes) ativos em cada experimento para cada caminho recuperado, com a visualização gráfica da via do KEGG com as enzimas identificadas.

A partir da enzima de interesse, são recuperados todos os caminhos com o método de travessia do Neo4j Java API: *traversalDescription()*, com uma busca em profundidade, controlando o comportamento da travessia pelos parâmetros: máximo de profundidade de 20, excluindo posição inicial para retorno, por relacionamentos do tipo *reaction_with* (entre nós com rótulos *Enzyme* e *Compound*), com as direções em relação à enzima: *incoming* (chegando), *outgoing* (saindo), *both* (ambas as direções), com singularidade *node_global* (nenhum nó pode ser visitado mais que uma vez).

Para cada caminho recuperado, será feita a busca para recuperar para cada nó (enzima), os níveis de expressão dos genes (enzimas) para cada experimento de interesse, através do método criado para buscar todos os nós vizinhos ao nó enzima baseado no método *traversalDescription()*, com relacionamento do tipo *Experiment*, direção *incoming* e rótulo do nó *Experiment* e *name* igual ao nome do experimento pré-determinado.

Ao final é retornado o número total de caminhos para a enzima de interesse, em cada tipo de direção (*both*, *incoming*, *outgoing*), com o tamanho de cada caminho baseado nos nós com rótulo *Enzyme*, e retornando o caminho mais longo, sendo impresso textualmente e representado visualmente pelas vias do KEGG, e o total de nós do tipo *Enzyme* ativos, para cada experimento selecionado, com o nível de expressão médio para os nós com nível de expressão para cada experimento.

A segunda abordagem para análise no banco de grafos corresponde à busca do caminho mais curto entre duas enzimas, para cada direção (*incoming*, *outgoing*, *both*), em uma via metabólica de interesse, comparando os caminhos entre duas condições experimentais.

A busca do caminho mais curto é realizado através do método *shortestPath()* da classe *GraphAlgoFactory*, composto por uma grande diversidade de algoritmos de grafos, um algoritmo que retorna o caminho mais curto entre dois nós (enzimas), utilizando os parâmetros de comprimento máximo em profundidade do caminho, e o

método *PathExpander()* funcionando como um expensor de relacionamento, um meio flexível para definir os tipos de relacionamentos e direção para um caminho. Nesta implementação são definidos o tipo de relacionamento *reaction_with* e direção *both*. Outras alternativas de busca implementadas são os métodos de busca de todos os caminhos disponíveis entre dois nós, o método *Dijkstra* que retorna o caminho de menor custo entre dois nós, baseado no custo de cada relacionamento.

Com o caminho mais curto retornado para a consulta entre duas enzimas é realizada a busca para cada nó do tipo *Enzyme*, pelos respectivos níveis de expressão dos genes para cada experimento pré-determinado, através do método criado para buscar todos os nós vizinhos ao nó enzima baseado no método *traversalDescription()*, com relacionamento do tipo *Experiment*, direção *incoming*, rótulo do nó *Experiment* e *name* igual ao nome do experimento.

Ao final da ferramenta serão recuperados os caminhos mais curtos para cada direção (*both*, *incoming*, *outgoing*) entre as duas enzimas selecionadas, o total de nós do tipo enzima para o caminho, com o total de nós ativos e a média do nível de expressão para cada condição experimental, representados textualmente e visualmente pelas vias do KEGG.

5.8. Plataforma CaneRegNet para Integração do Transcriptoma de Cana-de-açúcar

Nos capítulos anteriores, foram descritos os principais métodos computacionais e metodologias para a predição e anotação funcional do transcriptoma de cana-de-açúcar, para as análises individuais e comparativas dos níveis de expressão gênica por microarranjo entre diferentes condições experimentais, em um estágio inicial. A anotação desses dados é apenas a primeira etapa no estudo do transcriptoma, necessitando de inspeção, análise e validação experimental. Este estágio inicial de anotação e classificação de genes em categorias e funções pode proporcionar a descoberta de novas hipóteses, na integração deste conhecimento biológico às análises de expressão gênica por microarranjo. Deste modo, em um estágio intermediário de estudo do transcriptoma é essencial o acesso às anotações funcionais e alternativas para análise e integração com dados de microarranjo para curadoria e validação.

Neste capítulo, estaremos descrevendo a estrutura de armazenamento e organização da informação biológica do transcriptoma da cana-de-açúcar. O

procedimento de análise do transcriptoma em larga escala requer uma solução eficiente, flexível e escalonado para facilitar o acesso a esses dados biológicos, à quantificação e normalização, e prospecção por métodos computacionais e estatísticos, de uma maneira rápida e interativa, integrando diferentes tipos de dados de diferentes fontes. A ideia fundamental é prover ao pesquisador uma plataforma em que o usuário possa realizar diferentes análises por um sistema modular, como o processamento de experimentos de microarranjo de forma automatizada.

Os sistemas biológicos podem ser descritos em vários níveis de abstração, que inclui moléculas biológicas (p.e., RNA, proteínas, metabólitos, etc), interações biomoleculares (p.e., reações metabólicas), vias (p.e, cascatas ou conjuntos de interações coordenadas em ordem para realizara funções biológicas), atividade gênica (p.e., expressão gênica em determinadas condições. A plataforma *CaneRegNet* (Nishiyama et. al., 2010) foi desenvolvida para o armazenamento e integração destes dados biológicos para a integração de dados biológicos de anotação e classificação funcional, de lâminas de microarranjo e estudos de níveis de expressão por microarranjo e sequenciamento de larga escala, estudos de plantas transgênicas e medidas fisiológicas, visando o estudo da genômica funcional (Figura 21) (Manners et. al., 2011; Casu et. al., 2010; Menossi et. al.,2007).

Figura 21. Diagrama da Plataforma *CaneRegNet* e Abstração dos Dados, baseado em seis tópicos principais (Categorização de genes, Anotação Funcional, Estudos de perfil de expressão, Sequenciamento, Fisiologia e Plantas Transgênicas).

5.8.1. Estrutura do Banco de dados Relacional

A plataforma *CaneRegNet* está estruturada em um banco de dados relacional MySQL, desenvolvida para armazenar e integrar informações do transcriptoma da cana-de-açúcar, de outras espécies e outros bancos de dados como Uniprot/SwissProt, NR-NCBI, Gene Ontology, relacionados por meio de *Accessions* que são únicos para cada proteína. A estrutura de banco de dados desenhada armazena desde informações simples como a anotação de cada gene, até estruturas mais complexas para gerenciamento de toda plataforma e dos dados de microarranjo. Uma das principais tarefas da plataforma é o gerenciamento dos dados de todos os estágios de processamento do microarranjo desde a informação básica da amostra e desenho experimental até cálculos de medidas de expressão e relacionamento com vias metabólicas.

O banco de dados tem uma topologia de estrela, significando que todas as tabelas referem-se a um objeto central, sendo neste caso o gene. O uso de Sistemas de Gerenciamento de Bancos de Dados (SGBD) fornece suporte para armazenamento de grandes volumes de dados, um eficiente acesso através de linguagens de consulta, suporta a execução atômica e independente de transações concorrentes suporta armazenamento de grandes bancos de dados, suporta a execução simultânea auxiliando nas submissões de múltiplas lâminas de microarranjo, com vários estágios de processamento e armazenamento, além de acessos simultâneos aos dados sendo atualizados e acessados por meio de ferramentas pelos múltiplos usuários do sistema.

O diagrama de entidade relacionamento é apresentado de forma resumida, contendo apenas as entidades (Figura 22), mas foi criado um dicionário de dados do banco de dados (Anexo IV) contendo todos os detalhes e descrições das entidades, atributos e regras do modelo. Este banco de banco visa a descrever e representar dados relativos aos diversos conceitos biológicos independentes.

- O módulo "*Probe Set*" é composto pelos componentes: ESTs, transcritos, proteínas e oligonucleotídeos. Conceitualmente este módulo é o centralizador para conexão entre a maioria dos módulos;

- O módulo "*Gene Categorization*" corresponde às categorizações de famílias de proteínas quinases, fosfatases e de parede celular e classificação das famílias de fatores de transcrição em gramíneas.

- O módulo "*Functional Annotation*" corresponde às anotações automáticas, curagem manual e análises provenientes do: *Swissprot/Uniprot*, *Gene Ontology*, KEGG, bancos de gramíneas, InterProScan e proteínas homólogas, onde os genes são associados aos correspondentes termos GO, Enzimas, vias ou anotações específicas de outros bancos.

- O módulo "*Physiology*" é composto por diversas medidas fisiológicas extraídas da cana-de-açúcar em campo, relacionadas a amostras com estudo da expressão gênica por microarranjo.

- O módulo "*Expression Profile*" é dividido em quatro partes lógicas de dados: desenho do arranjo, amostra, experimento (hibridização) de microarranjo e medidas de expressão (intensidade de expressão por canal "**probe_intensity**", razão de expressão por hibridização "**probe_statistics**" e razão de expressão por cruzamento entre réplicas técnicas/biológicas "**probe_expression**").

- O módulo "*Metabolic Pathways*" representa abstratamente os grafos das vias metabólicas e suas interações, através de: enzimas, compostos (substrato e produto) e reações enzimáticas, relacionado aos módulos *Probe set* e *Expression Profile*.

Para padronizar e aperfeiçoar o compartilhamento das informações biológicas e metodológicas para experimentos de microarranjo e manter o histórico dos experimentos, foram estabelecidas algumas padronizações e infraestruturas para armazenamento da informação e compartilhamento entre a comunidade científica (Quackenbush J, 2004; Brazma et. al., 2003; Edgar et. al., 2002; Ball et. al., 2005). O MIAME (*Minimum Information About a Microarray Experiment*) descreve a informação mínima sobre um experimento de microarranjo que é necessário para interpretar os resultados de um experimento inequivocamente e potencialmente reproduzir o experimento e análises (Ball et. al., 2004; Brazma, 2009). Deste modo, foi desenhada uma estrutura de dados comum, flexível e robusta para compartilhamento de dados de microarranjo baseado em expressão gênica (Whetzel et. al., 2006). A modelagem da

estrutura do *CaneRegNetDB* foi baseada nestes princípios, sendo adaptada às necessidades e especificidades dos experimentos de microarranjo para cana-de-açúcar e do laboratório de Transdução de Sinal (IQ-USP). Revistas de artigos científicos e comunidade científica passaram a exigir antes da publicação de artigos, a submissão destes dados de microarranjo no formato MIAME ou similar, para submissão prévia dos dados de microarranjo em um repositório público. O NCBI-GEO (Barrett et. al., 2007), tornou-se uma das referências em depósito de experimentos de microarranjo, adotando novo formato mais simples SOFT (*Simple Omnibus Format in Text*) baseado no MIAME. O SOFT é um formato desenhado para submissões de grandes volumes de dados e é baseado em três entidades: *Plataform*, contendo todas as informações e descrição relacionada ao desenho do microarranjo; *Sample*, responsável pela descrição das amostras, condições experimentais para o experimento de microarranjo e detalhes do protocolo de hibridação; e *Series*, para descrever o estudo contendo múltiplas *Samples*, sumarizando os objetivos e interesses do estudo e fornecendo detalhes específicos sobre o estudo.

Foi implementado na plataforma *CaneRegNet* uma interface seguindo os padrões MIAME e SOFT para extração das informações do banco de dados *CaneRegNetDB* relacionadas aos experimentos de microarranjo, agrupando-os em *Series* e gerando os arquivos em formato de submissão para o NCBI-GEO. Cada etapa da geração do arquivo de submissão é baseado na recuperação da informação e geração de formulários para o usuário alterar as informações armazenadas previamente ou validar o conteúdo para preparação do pacote final para submissão.

5.8.2. Controle de Acesso da Plataforma *CaneRegNet*

Para restringir o acesso de usuários não autorizados em um ambiente multiusuários e multi-projetos foi criado um modelo no banco de dados para armazenamento e gerenciamento de usuários, projetos e dados vinculados (Figura 23).

Figura 23. Diagrama de Entidade Relacionamento representando o controle de acesso da plataforma CaneRegNet.

Os conceitos de Autorização e Autenticação são distintos (Fischer-Hübner et. al., 2003) e em nossa plataforma a autenticação é um pré-requisito para a autorização, devido às restrições de acesso aos projetos e informações armazenadas. O método de autenticação é baseado no cadastro do usuário com um nome de usuário e uma senha particular. O sistema de autorização possui três tipos de nível de acesso: 0 – usuário comum, 1 – coordenador de projetos; 2 – administrador, e é responsável pelo controle de acesso de leitura, escrita, inserir, criar e apagar, para módulos e dados (Tabela 4). O modelo de banco de dados criado (Figura 23), permite ter um controle amplo da estrutura do *CaneRegNetDB* e genérico o suficiente para permitir a criação e gerenciamento de novos módulos. A lógica implementada na plataforma *CaneRegNet* possibilita ao sistema gerenciar o acesso de leitura e escrita a cada uma das ferramentas e páginas de acesso *dinâmico* às informações, controlar a permissão para inserção, atualização e deleção de anotações funcionais e experimentos de microarranjo, por meio da entidade *dbaccess*. Para o acesso aos experimentos de microarranjo, e aos resultados de processamentos armazenados no *CaneRegNetDB*, o sistema possui um controle refinado através das entidades *dbaccess* e *table_tuple_access*, disponibilizando todas as informações existentes de acordo com o nível de acesso do usuário.

Tabela 4. Lista dos principais módulos criados no CaneRegNet e referência para o gerenciamento de toda a informação.

Módulo	Descrição
project	Contém informações sobre o projeto e responsáveis. Este projeto estará vinculado aos experimentos, resultados e análises de microarranjo.
sample_provider	Contém informações referentes ao responsável pelas amostras biológicas.
sample	Contém detalhes e informações biológicas específicas sobre a amostra.
experiment	Contém detalhes e informações biológicas específicas sobre o experimento de microarranjo.
cross_ref	Contém a descrição e informações sobre análise de cruzamento de experimentos de microarranjo e métodos de análise.
categorization	Contém informações e anotação funcional curada por usuários.
probe_set	Contém informações a respeito dos genes e estudos relacionados.
transgenic_plant	Contém informações referentes à construção de plantas transgênicas.
rt_pcr	Contém informações referentes ao qRT-PCR.
transcript_annotation	Contém informações e anotação funcional curada por usuários.
Assembly	Contém informações sobre as montagens de transcriptoma e genoma.
oligo_probe	Contém informações referentes aos oligonucleotídeos desenhados e aqueles utilizados na lamina de microarranjo.

5.8.3. Estrutura do Banco de dados de Grafos

O banco de dados relacional representa as informações biológicas como coleções de partes individuais de um sistema biológico mais complexo. Para representar e estudar fenômenos biológicos como um todo, como as vias metabólicas, foi desenvolvida uma estrutura em bancos de grafos, baseado no sistema gerenciador de grafos Neo4j.

Esta estrutura permite o armazenamento da estrutura real das vias metabólicas. O modelo implementado expressa as vias como um grafo, armazenando nós e relacionamentos conectando-os, incorporando propriedades biológicas em ambas as representações (Figura 24). O tipo de grafo empregado nesta implementação é uma combinação de grafos direcionados, rotulados e com atributos. Este tipo de grafo é também conhecido como “grafo de propriedades” e permite a representação de vértices e arestas rotulados, e atributo de dados (ou propriedades) para ambos. Nosso modelo de grafos pode ser definido como $G = (N, A, L, f_N, f_A, R)$ em que N é o conjunto de nós, representando as vias metabólicas, as enzimas, os genes homólogos e os experimentos de microarranjo, A é o conjunto de arestas e L é o conjunto de rótulos dos nós e arestas. Os rótulos tem a forma (atributo, valor). As duas funções f_N e f_A , são funções de atribuições dos nós ($f_N: N \rightarrow L$) e a função das arestas ($f_A: A \rightarrow L$), respectivamente. As regras para determinar as combinações entre os nós e arestas são obtidas a partir dos requisitos do conhecimento biológico e empregadas nos programas desenvolvidos para as consultas.

A estrutura do grafo é composta pelos nós, relacionamentos e atributos a seguir:

- nó **Via metabólica**, representando toda a rede composta de enzimas e compostos, atributos: *code*, *name*.
- nó **Enzima**, representando as enzimas presentes na Via metabólica, atributos: *code*, *name*, *reaction*.
- nó **Composto**, representando os substratos e produtos na Via metabólica, atributos: *code*, *type*: p (produto) e s (substrato).
- nó **Gene**, representando os genes homólogos para cada Enzima, atributo: *name*.
- nó **Experimento**, representando os experimentos de microarranjo com valor de expressão, relacionado com a Enzima, atributos: *name*, *value* (valor de expressão).
- aresta **PART_OF**, representando o relacionamento entre a Enzima e a Via Metabólica, atributo: *role enzyme*.
- aresta **REACTION_WITH**, representando o relacionamento entre a Enzima e os Compostos, atributos: *code*, *role*: reaction e *direction* (i: irreversível, r: reversível).
- aresta **HOMOLOG**, representando o relacionamento entre a Enzima e os Genes homólogos, atributo: *role*, *homolog*.

- aresta **EXPERIMENT**, representando o relacionamento entre a enzima e o experimento de microarranjo, atributo: *name*.

Figura 24. Representação geral das vias metabólicas, integradas a genes e experimentos de microarranjo no banco de grafos.

5.8.4. Implementação e Arquitetura da Plataforma CaneRegNet

A arquitetura central do nosso sistema de processamento, integração e visualização de dados, chamado *CaneRegNet* é composta por três camadas; de Visualização, Controle e Modelo (Figura 25), implementadas em Java e Perl scripts. Para o gerenciamento de vários conjuntos de experimentos e estudos relacionados, de seus meta-dados relacionados e grupos de aplicações específicas, adotou-se um conceito de agrupamento de toda informação dentro de projetos. Cada projeto consiste de um conjunto de experimentos e informações importadas e relacionadas às amostras e condições experimentais originais, aliada aos novos dados e relatórios produzidos em estágios intermediários de análise.

O servidor de aplicações da plataforma *CaneRegNet* é composto por uma camada de Servlets que controlam a lógica da aplicação web e administram a informação, uma camada de páginas JSP que exibem a requisição do usuário e uma camada de classes Java que disponibilizam as funcionalidades implementadas como a recuperação de informação do banco de dados.

Figura 25. Arquitetura da plataforma *CaneRegNet*.

As arquiteturas de aplicações web baseiam-se geralmente no conjunto de padrões de software MVC (*Model, View, Controller*), pois é flexível à implementação de novos requerimentos e alterações em banco de dados, permitindo simplificar todas as fases de desenvolvimento de aplicações web (Bryan Basham et. al., 2008; Eric Freeman et. al., 2004), entre elas evitar a redundância do código e facilitar a manutenção. Porém, no caso de banco de dados que estão em constante desenvolvimento, com um modelo que pode sofrer alterações com maior frequência, e com entidades que podem mudar de significado, surgem problemas na manutenção desses arcabouços (*frameworks*) que utilizam estes padrões, pois a camada do *Modelo dos arcabouços* está fortemente relacionada a um banco de dados por meio de classes para cada entidade, o que pode proporcionar numa desvantagem quando o banco de dados tende a mudar regularmente.

Dessa forma, a plataforma de aplicações web *CaneRegNet* tem sido desenvolvida, considerando-se o padrão MVC, porém com uma diferença importante em relação a outros arcabouços existentes. Essa diferença corresponde à camada do Modelo, que conecta e armazena temporariamente a informação do banco de dados. No caso, são

utilizadas classes com listas dinâmicas, que são robustas a alterações no banco de dados, e ao mesmo tempo simples na sua manutenção. A estrutura da plataforma da aplicação web pode ser observada na Figura 25, e consiste de três camadas principais:

1- A camada de Visualização (*package jsp*) é responsável pela apresentação e por receber os dados de entrada do usuário. Ela é composta de páginas JSP para cada ferramenta, scripts JQuery, scripts JavaScript e arquivos com estilo CSS. As páginas JSP (Bryan et. al., 2008) são construídas dinamicamente, baseado no tipo de consulta do usuário e para cada ferramenta. Cada ferramenta web é controlada por seu próprio Servlet, pois funcionalidades e tipos de dados variam para cada ferramenta. Como as ferramentas web precisam exibir grandes quantidades de informação, implementou-se a funcionalidade de paginação, por requisições do tipo AJAX (*Asynchronous Javascript and XML*) e também usou-se a linguagem JavaScript e tags DIV para ocultar e exibir informação. As aplicações desenvolvidas estão todas dentro dessa camada e as páginas seguem o mesmo padrão da arquitetura representada na Figura 26. As aplicações desenvolvidas nesta tese podem ser agrupadas por:

- Aplicações de Gerenciamento, Controle de Acesso, Submissão e Atualização;
- Aplicações de Análise de expressão gênica;
- Aplicações de Busca do Transcriptoma e Anotação Funcional.

2- A camada de Controle (*package controller*), composta pelos Servlets, controla a lógica da aplicação web e gerencia o fluxo de informação. Ela extrai da solicitação do usuário os dados de entrada e faz a interpretação de seu significado para as classes Java, obrigando as classes Java a se atualizarem e disponibilizar o novo estado para as páginas JSP. Esta camada possui um mapeamento da informação e possui um alto nível de compreensão do fluxo de dados sem entrar em detalhes. Foi desenvolvido um Servlet para cada ferramenta, permitindo um controle individual e simplificado de cada ferramenta.

3- A camada de Modelo (*package common*), composta por classes Java, disponibiliza as funcionalidades implementadas para cada ferramenta, entre elas a recuperação da informação do banco de dados, criação/validação de arquivos em formato texto e XML, processamento de dados, criação de imagens dinâmicas, geração de estatísticas, gerenciamento da conexão com o banco de dados, execução de processos em segundo plano usando multitarefas (*Threads*), entre outras funcionalidades. Como boa prática de programação, estruturamos as classes com intuito

de evitar a redundância de código. Cada classe agrupa um conjunto de métodos relacionados e que são apresentados a seguir:

- A classe *SessionLogin* é o gerenciador inicial da aplicação e administra o *login* do usuário, a criação de novas seções e o controle de seções ativas. Usa a classe *DBConnectionPoolingListener*, que conecta com o *pool* de conexões do MySQL.

- A classe *database* disponibiliza métodos para recuperar informação do banco de dados *CaneRegNetDB* e dos bancos de dados da ferramenta GBrowse. São utilizadas listas dinâmicas do tipo *String* como estrutura de armazenamento temporário de dados, pela sua flexibilidade e simplicidade, são úteis quando a quantidade de colunas e tipos de dados nas tabelas do banco de dados tende a sofrer alterações.

- A classe *model* dispõe de métodos para recuperar dados invocando a classe *database*, nesta classe realiza-se um processamento avançado nos dados, segundo os requisitos específicos de cada aplicação, construindo resultados dinâmicos para serem disponibilizados nas páginas JSP.

- A classe *utils* disponibiliza métodos para a criação de figuras estatísticas utilizando a biblioteca *jfreechart* (Gilbert et. al., 2002), para formatar números em notação científica ou ponto flutuante, cálculos estatísticos, leitura e validação de arquivos em formato texto, envio de e-mail, descompactação de arquivos em formata zip e tar, criação *logs* personalizados no formato XML, criação de imagens dinâmicas para aplicações como: "*Protein Domain*", "*Alignment Results*", "*Probe Intensity*", "*Probe Expression*", "*Probe Statistics*", "*Term Enrichment*" baseado nos dados do banco *CaneRegNetDB*. Além de fornecer métodos para conversão de dados do banco *CaneRegNetDB*, referente aos catálogos de anotações funcionais em arquivos para o formato XML.

- A classe *parserHTML*, utiliza as classes JSoup de código aberto, e disponibiliza o método para acessar páginas *web* externa como páginas gráficas do KEGG, para extrair uma imagem de interesse e armazená-la localmente para visualização dentro da plataforma *CaneRegNet*.

Figura 26. Representação detalhada da arquitetura da plataforma *CaneRegNet*, exemplificada por uma aplicação de busca de *reads*, mas segue o mesmo padrão para outras buscas mudando apenas as páginas JSP.

5.8.5.Interface e Estrutura de Integração

Todos os resultados, métodos e metodologias descritos nesta tese foram formatados e armazenados no *CaneRegNetDB* ou implementados na forma de ferramentas na plataforma *CaneRegNet*. A plataforma fornece uma infraestrutura extensiva e flexível para a anotação do transcriptoma, pré-processamento de experimentos de microarranjo e, principalmente, para consultas e análises de expressão gênica em vários níveis da informação, possibilitando análises de prospecção de forma automática e eficiente na integração entre as informações.

A plataforma possui diversas funcionalidades e tem como objetivo fornecer ao usuário uma estrutura que está centrada em três principais componentes: 1- acesso às informações biológicas do transcriptoma; 2- armazenamento e processamento de experimentos de microarranjo; 3- ferramentas de análise e integração de dados de microarranjo e anotações funcionais.

A interface do *CaneRegNet* está organizada em 10 seções (Abas) para representação e gerenciamento da informação relacionada à cana-de-açúcar, e o acesso é restrito e gerenciado por um controle de acesso. O acesso à área restrita é efetuada pelo *Servlet SessionLogin*, que se conecta ao banco de dados e utiliza as classes *model.java* e *database.java* para recupera as informações do usuário relacionadas aos níveis de acesso aos dados da plataforma. A classe *database.java* realiza a abertura ou reutilização de uma conexão do *pool* de conexões e, em seguida, cria um objeto de ligação com o método *createStatement* e executa-se uma consulta usando o método *executeQuery*. Os dados retornados pelo método *executeQuery* são salvos em uma estrutura de lista dinâmica para ser utilizada pela interface e em seguida fecham-se as conexões ativas, devolvendo a conexão ao *pool* de conexões. O sistema utiliza os *Servlets SessionController* e *SessionContent* e requisições do tipo AJAX para montar e exibir a página e seus conteúdos. Todas as páginas da interface seguem o mesmo padrão de consultas e requisições. A página da interface de entrada composta pelas 10 seções:

- Home,
- Cane Genome,
- Cane Transcriptome,
- Cane Gene Expression,

- *Cane Physiology,*
- *Functional Annotation,*
- *Cane Transgenic Plant,*
- *Generic Tools,*
- *Publications,*
- *Admin.*

Cada uma das seções possuem diversas ferramentas e resultados relacionados com os estudos da seção particular, sendo que todas as informações são centralizadas em algum nível ao módulo *probe_set*, que contém a informação dos genes. As ferramentas desenvolvidas para consultas aos dados do banco e vinculados aos genes, seguem um mesmo padrão de utilização das *Servlets* e Classes implementadas. O fluxograma a seguir (Figura 27), ilustra o comportamento da aplicação *web* para uma busca simples de um gene, considerando a interação entre os arquivos JSP, os *Servlets*, as bibliotecas e classes Java desenvolvidas. Após a consulta de um gene pelo nome, se o mesmo for encontrado, será criada uma nova página composta por 7 seções:

- *Summary*: sumário com informações essenciais sobre o gene.
- *Genome*: informações genômicas relacionadas ao gene.
- *Transcriptome*: informações gerais do gene, como anotação funcional, resultados de diferentes análises de expressão por microarranjo para os experimentos acessíveis ao usuário.
- *Proteome*: informações das proteínas preditas de genes, e anotação funcional.
- *Gene Network*: para estudos futuros de redes de regulação gênica.
- *Phenotype*: informações sobre plantas transgênicas e medidas de fisiologia relacionadas a este gene.
- *Pathway*: informações do gene relacionadas à anotação funcional com KEGG, como enzimas, compostos, reações enzimáticas e exibição da figura da via metabólica.

As seções são representadas no fluxograma pela chamada dos métodos do pacote JSP. A página inicial apresenta um sumário sobre o gene (aba *Summary*). O acesso às demais seções (abas) contendo informações específicas para o gene de interesse serão

exibidas dinamicamente por requisições do tipo AJAX ao serem clicadas, para exibir todas as informações relacionadas presentes no *CaneRegNetDB*.

Figura 27. Fluxograma de busca de um gene, mostrando as etapas de consulta e processamento da informação e montagem da estrutura de exibição da nova seção.

A interface desenvolvida para submissão e processamento de dados de microarranjo segue o mesmo padrão da estrutura exemplificada (Figura 27). As interfaces são baseadas em formulários incorporando funcionalidades de validação dos campos obrigatórios, segundo padrões estabelecidos pelo GEO-NCBI, MIAME e características específicas relacionadas à lâmina de microarranjo. Para a construção dos formulários são construídas listas dinâmicas com os atributos e informações recuperadas do *CaneRegNetDB*, seguindo a lógica de controle de acesso para verificar se o usuário possui acesso as informações relacionadas ao experimento sendo criado, atualizado ou deletado.

A interface para gerenciamento e criação dos experimentos de microarranjo é composta dos principais módulos:

1 – Criação de uma plataforma de microarranjo, descrevendo as informações do microarranjo, da composição do microarranjo e dos protocolos de hibridação utilizados, todos armazenados em diferentes entidades do *CaneRegNetDB*.

2 – Criação de forma independente e sequencial do projeto, amostras e experimento. O formulário de submissão do experimento disponibiliza as informações anteriores para serem vinculadas, solicita a definição do critério de credibilidade (p-valor) utilizado na análise de expressão gênica, através do método *HTSelf* adaptado e o arquivo de saída dos *softwares FeatureExtraction* ou *ArrayVision*. Se o arquivo for inserido na criação do experimento a etapa intermediária de processamento da lâmina de microarranjo será automaticamente iniciada executando *scripts* em *Perl* e *R-statistics* e, ao final, será enviado um *email* ao usuário de sucesso ou problemas identificados, e os resultados armazenados no *CaneRegNetDB*. Se o arquivo não for inserido, a descrição do experimento pode ser salva e o processamento executado pela interface de atualização do experimento, quando o arquivo estiver disponível.

3 – Criação do cruzamento entre réplicas biológicas/técnicas. Esta etapa tem o objetivo de realizar um cruzamento entre os resultados obtidos pelo processamento na etapa anterior para múltiplas condições experimentais, para identificar genes diferencialmente expressos comuns às condições experimentais selecionadas, produzindo novos resultados que serão incorporados ao *CaneRegNetDB*.

Diversos métodos descritos nesta tese foram implementados e integrados à plataforma *CaneRegNet*. Alguns destes serviços são empacotados em *pipeline* baseado em *Java*, *Perl script* e *R-statistics*, para as análises de microarranjo, integradas à

anotação funcional. O processo para as consultas e análises pode ser visualizado no fluxograma abaixo (Figura 28).

Figura 28. Fluxograma representando a lógica das aplicações para aquisição do conhecimento através da integração de dados de diferentes fontes. Aplicações para consulta de dados do Transcriptoma, Anotação Funcional e Análises por dados de microarranjo.

6. Resultados e Discussão

6.1. Anotação Funcional do transcriptoma de Cana-de-açúcar

Nesta seção, apresentamos e discutimos os primeiros resultados obtidos para anotação funcional do transcriptoma de cana-de-açúcar com as estratégias descritas anteriormente, para a construção de uma base de conhecimento. As análises foram realizadas sobre o conjunto de transcritos do projeto SUCEST, representando um total de 43.141 SAS e o segundo conjunto representa o sequenciamento por RNA-Seq de bibliotecas *full-length* dos tecidos de folha das espécies de cana *S. officinarum* e *S. spontaneum* e também de folha, entrenó maduro e intermediário do híbrido comercial SP80-3280. Este último foi escolhido, pois aproximadamente metade das bibliotecas do projeto SUCEST é dessa variedade e pelo sequenciamento do genoma da cana-de-açúcar (<http://sugarcanegenome.org>) também utilizar esta mesma variedade.

Para a obtenção dos transcritos a partir do sequenciamento por *RNA-Seq*, os *reads* gerados a partir do sequenciamento pelas plataformas *454 Sequencing* e *Ion PGM* foram pré-processados e em seguida montados com o *software* Trinity, gerando um total de 195.765 transcritos (Tabela 5).

Tabela 5. Dados do sequenciamento por *RNA-Seq* e da montagem com o *software* Trinity.

Descrição	Ion PGM	454
Corridas	5	3
Reads	29.260.184	2.403.750
Tamanho total dos reads	5,69 Gb	0,97 Gb
Média	194 bases	404 bases
Mediana	197 bases	436 bases
Read mais longo	645 bases	861 bases
Read mais curto	8 bases	50 bases
Montagem com Trinity		
Número de transcritos	195.765	
Tamanho total dos transcritos	133.946.191 bases	
transcrito mais longo	8.854 bases	
transcrito mais curto	201 bases	
Número de transcritos > 1Kb	39.266 (20,1%)	
Número de transcritos > 2Kb	8.052 (4,1%)	

Número de transcritos > 3Kb	1.748 (0,9%)
Número de transcritos > 5Kb	154 (0,1%)
Média	684 bases
Mediana	462 bases

Os reads de cada amostra foram mapeados contra os 195.765 transcritos utilizando o *software bowtie2*, permitindo identificar os transcritos pela especificidade da expressão entre as amostras, ou seja, transcritos tecido ou espécie específicos. Esse mapeamento permitiu identificar um total de 195,406 (99,82%) dos transcritos, sendo possível identificar com alta identidade *reads* em praticamente todos os transcritos, exceção para 359 transcritos. Esse grande número de transcritos pode estar relacionado com o grande número de alelos, parálogos e possíveis variantes de *splicing* em cana-de-açúcar por ser um genoma poliploide e aneuploide. Com o mapeamento dos *reads* de cada amostra nos transcritos, podemos verificar que 59% dos transcritos possuem pelo menos um *read* sendo representado por cada amostra (linha azul na Tabela 6), indicando que são comuns entre as diferentes espécies e tecidos. Apenas 4.415 (2,25%) dos transcritos são expressos especificamente em SP80-3280 indicando que a maior parte dos genes é herdada de espécies ancestrais. Um número de transcritos significativo (28.449 ou 15,2%) foi identificado apenas em *S. spontaneum* e SP80-3280 em quanto uma quantidade bem menor foi expressa exclusivamente em SP80-3280 e *S. officinarum* (5.864 ou 3%). Apenas 1.425 (0,73%) dos transcritos foram específicos para os genótipos ancestrais e o dobro de transcritos (4.756 ou 2,4%) apresentou expressão apenas no híbrido SP80-3280 (linhas cinzas). Dos 3.086 transcritos presentes nos genótipos ancestrais, 1.425 são expressos nos dois ancestrais e 1.343 são expressos somente em *S. spontaneum* (linha laranja) e apenas 318 em *S. officinarum* (linha verde).

Tabela 6. Representação do mapeamento dos *reads* de cada amostra nos transcritos, com o total de transcritos presentes em cada grupo. As linhas cinza representam transcritos específicos em SP80-3280; Amarelo, genótipos ancestrais; Laranja, *S. spontaneum*; Verde, *S. officinarum*.

Folha			Entrenó 1	Entrenó 5		
<i>S.officinarum</i>	<i>S. spontaneum</i>	SP80-3280	SP80-3280	SP80-3280	#transcritos	%transcritos
X	X	X	X	X	115.462	59,08826
X	X	X		X	16.471	8,429117
	X	X	X	X	10.205	5,22246
X	X		X	X	9.098	4,655947
	X	X		X	6.153	3,148829
X	X			X	5.663	2,898069

	X			X	4.368	2,235346
	X		X	X	4.156	2,126854
X	X	X	X		2.951	1,510189
X	X	X			2.427	1,242029
X		X	X	X	1.868	0,955958
	X	X			1.521	0,778379
X	X				1.425	0,729251
	X				1.343	0,687287
X		X		X	1.312	0,671423
	X	X	X		1.218	0,623318
X	X		X		1.179	0,603359
		X		X	1.060	0,54246
		X	X	X	1.053	0,538878
				X	1.029	0,526596
X				X	994	0,508684
X			X	X	844	0,431921
	X		X		828	0,423733
			X	X	787	0,402751
X		X			355	0,181673
		X			341	0,174508
X					318	0,162738
X		X	X		294	0,150456
			X		249	0,127427
		X	X		237	0,121286
X			X		197	0,100816

6.1.1. Anotação Funcional por Gramíneas

A anotação do transcriptoma de cana-de-açúcar é composta por diversas etapas e inicialmente os transcritos foram mapeados contra os bancos de genes preditos das gramíneas milho (*Zea mays*), sorgo (*Sorghum bicolor*), arroz (*Oryza sativa*) que já possuem genoma sequenciado e em estágio avançado de fechamento, e *Setaria* (*Setaria italica*), *Panicum* (*Panicum virgatum*), *Brachypodium* (*Brachypodium distachyon*) com genomas parciais, além dos SAS do SUCEST, do *Gene Ontology* e do *KEGG* (Tabela 7). Os bancos de dados de gramíneas, provenientes do projeto *Phytozome*, foram anotados por sistemas de anotação automatizada, baseados no próprio genoma da espécie, através da predição dos genes e atribuição de anotação baseado em bancos públicos como *Uniprot*, *Gene Ontolog*, *KEGG* e predição de domínios.

Tabela 7. Número total de genes por espécie e em cada banco de dados utilizado

Espécies	# Genes	Com Anotação	Sem Anotação	Unknown
<i>B. distachyon</i>	31.029	24.431	5.504	1.094

O.sativa	49.061	31.427	16.129	1.505
P. virgatum	70.071	49.379	18.327	2.365
S. bicolor v2.1	39.441	27.136	11.052	1.253
S. italica	40.599	27.379	11.901	1.319
Z. mays	63.540	44.887	16.793	1.860

As comparações entre os transcritos e os bancos de dados foram realizadas com buscar por similaridade com BLAST, com e-valor 1e-05 e foram processadas pelo programa desenvolvido localmente baseado nos métodos descritos na seção 4.1, de melhor *hit* uni-direcional (BLASTX) e o melhor *hit* bi-direcional, com alinhamentos em ambas as direções por BLASTX e TBLASTN, considerando cada gene no transcriptoma A como consulta comparada contra todos os genes no transcriptoma B e vice-versa. Uma abordagem de alinhamento bi-direcional é um método estatisticamente significativo e possibilita a inferência de potencial ortologia funcional entre os genes e seus correspondentes em um organismo modelo. Este relacionamento putativo torna-se a base para a transferência da anotação de um gene em um organismo modelo ou melhor anotado para um novo gene.

Comparando os resultados obtidos com a abordagem uni-direcional (Tabela 8), baseado no alinhamento direto e critérios de escore bit > 30, e-valor 1e-05 e SBH > 0.95, foi possível identificar homologia de até 84,26% dos transcritos do SUCEST e de 64,47% dos transcritos do RNA-Seq com outras espécies e bancos públicos, sendo que os alinhamentos selecionados com estes critérios apresentaram uma cobertura média dos transcritos de 65% (Figura 29).

Tabela 8. Números absolutos e porcentagens do total de hits entre os transcritos de cana-de-açúcar do SUCEST e RNA-Seq contra os bancos das 6 gramíneas, *Gene Ontology* e *KEGG*, através do método de alinhamento Uni-direcional e Bi-direcional, com e-valor 1e-05, escore bit > 30, SBH > 0.95 ou BHR > 0.95.

Epécies	Alinhamento Uni-direcional		Alinhamento Bi-direcional (BHR)	
	#SUCEST (%)	#RNA-Seq (%)	#SUCEST (%)	#RNA-Seq (%)
S. bicolor	36.352 (8,26%)	126.213 (64,47%)	18.366 (42,57%)	28.170 (14,39%)
Z. may	32.572 (75,5%)	106.247 (54,27%)	16.818 (38,98%)	27.525 (14,06%)
S. italica	32.260 (74,78%)	103.981 (53,12%)	16.587 (38,45%)	14.875 (7,6%)
P. virgatum	32.171 (74,57%)	103.833 (53,04%)	17.570 (40,73%)	30.872 (15,77%)
O. sativa	29.363 (68,06%)	89.911 (45,93%)	14.985 (34,73%)	26.102 (13,33%)
B. Distachyon	29.053 (67,34%)	88.298 (45,10%)	14.659 (33,98%)	24.951 (12,75%)
Total transcritos	37.552 (87,04%)	136.894 (69,93%)	23.149 (53,66%)	46.511 (23,76%)
Gene Ontology	29.605 (68,62%)	78.273 (39,98%)	10.152 (23,53%)	19.637 (10%)

KEGG	-	-	2.584 (6%)	14.020 (7%)
------	---	---	------------	-------------

Figura 29. Representação da porcentagem de genes pelo alinhamento contra os transcritos, utilizando o algoritmo para calcular a cobertura total de todos os hits em um mesmo gene, distribuídos em faixas de porcentagem de cobertura.

A diferença entre a porcentagem de transcritos identificados com o SUCEST e *RNA-Seq* é provavelmente devido ao grande número de transcritos montados com o *RNA-Seq* (cerca de 195 mil). Esses números apresentam uma visão do total de genes homólogos entre cana-de-açúcar e outras gramíneas, sendo que o número é maior à medida que a espécie é mais próxima de cana-de-açúcar. Por exemplo, cerca de 90% dos genes de sorgo (cerca de 27 mil dos 29 mil genes preditos), espécie analisada mais próxima evolutivamente de cana-de-açúcar, apresenta hit com cerca de 62% dos transcritos do *RNA-Seq* (Figura 30), números semelhantes ao alinhamento dos transcritos contra o SUCEST. Os ~7% do SUCEST que não possuem hit contra o *RNA-Seq* podem ser transcritos específicos de algum tecido, como raiz, gema lateral, flor que foram usados nas bibliotecas de sequenciamento do projeto SUCEST, mas não no *RNA-Seq*. Reciprocamente, apenas 60% dos transcritos de *RNA-Seq* estão presentes no SUCEST, indicando que aproximadamente 78 mil transcritos são sequências novas em cana-de-açúcar, uma das razões pode estar relacionada com o baixo número de *reads* (cerca de 5%) provenientes de tecido de folha no SUCEST.

Figura 30. Representação da porcentagem de genes (gramíneas) e transcritos (RNA-Seq) identificados nos alinhamentos entre as gramíneas e transcritos do *RNA-Seq*, com critério de e-value $1e-20$.

Pela comparação do método uni-direcional com o método bi-direcional podemos notar uma diminuição significativa no número de transcritos identificados e candidatos para transferência da anotação a partir dos transcritos de gramíneas e bancos de dados. Para que as informações de anotação por ambos os métodos não sejam perdidas, todos os resultados são armazenados no banco de dados *CaneRegNetDB* e são acessíveis pelas ferramentas de anotação e consulta dos transcritos. O resultado da abordagem bi-direcional para o transcriptoma do SUCEST apresentou um filtro em média de 50% dos transcritos, porém com os transcritos do *RNA-Seq* foram filtrados cerca de 70% dos transcritos. Essa grande diminuição pode estar relacionada com grande número de transcritos obtidos com o *RNA-Seq*, com genes quebrados em diversos pedaços (transcritos) ou devido ao grande número de alelos ou variantes de *splicing* que podem estar presente no transcriptoma, influenciando no cálculo do melhor *hit*. Entretanto, apesar do grande número de transcritos filtrados pela anotação, ainda foram selecionados cerca de 30 mil transcritos por *RNA-Seq*, o que na verdade corresponde a cerca de 90% do transcriptoma de cana-de-açúcar, dando fortes indícios de que os transcritos restantes podem ser partes, homólogos ou variantes de *splicing*. Sem um genoma referência é praticamente impossível determinar a origem destes transcritos, porém com o sequenciamento do genoma da cana-de-açúcar em andamento estes transcritos serão analisados a nível estrutural, possibilitando identificá-los e classificá-los como alelos, parálogos ou variantes de *splicing*, bem como determinar a expressão de cada alelo de um gene e sua taxa de *splicing*.

Para a transferência da anotação dos transcritos do SUCEST e *RNA-Seq* a partir das gramíneas, utilizamos a ordem de atribuição de anotação das espécies de acordo com o maior número de *hits* (Figura 30) e maior taxa de cobertura com os transcritos

de cana-de-açúcar (Figura 29), sendo portanto, na seguinte ordem: sorgo, milho, *Panicum*, *Stária*, arroz e *Brachypodium*. Como apresentado na tabela 7, sorgo possui cerca de 67% do seu transcriptoma anotado, com 3% anotado como desconhecido (*unknown*) e 30% sem anotação definida, seguindo um padrão semelhante para as demais gramíneas.

Como resultado da aplicação do método bi-direcional para a comparação entre as gramíneas e os transcritos da cana-de-açúcar, foram anotados 18.366 (42%) dos transcritos do SUCEST, representados em 5.895 categorias. As quatro principais categorias são representadas por 4.186 proteínas desconhecidas (*unknown*), 284 proteínas quinases, 209 proteína superfamília RING/U-box, 198 proteínas *Tetratricopeptide repeat (TPR)-like* e 188 proteínas NB-ARC, em anexo (Anexo VIII) apresentamos a lista das categorias que apresentam pelo menos 50 transcritos baseado na anotação das gramíneas. Com a mesma abordagem aplicada aos transcritos de *RNA-Seq*, foram anotados 28.170 transcritos (24%), representados em 6.272 categorias, tendo as mesmas categorias mais frequentes identificadas no SUCEST, com 10.342 proteínas *unknown*, 536 proteínas quinases, 505 NB-ARC e 443 *Tetratricopeptide repeat (TPR)-like*, em anexo (Anexo IX) apresentamos a lista das categorias que apresentam pelo menos 50 transcritos baseado na anotação das gramíneas.

A comparação dos dados de *RNA-Seq* com as sequências presentes no chip de microarranjo *Agilent Technologies* mostrou que apenas 435 sondas do *chip* não foram identificadas pelo *RNA-Seq*.

6.1.2. Anotação por Gene Ontology

Utilizando a mesma metodologia para anotação dos transcritos a partir do banco de dados do GO, foram anotados 10.152 (23%) transcritos do SUCEST, representados em 4.566 categorias. Com a mesma abordagem para os transcritos do *RNA-Seq* foram identificados 19.637 (10%) transcritos com anotação no GO, representados em 4.875 categorias, sendo as categorias mais bem representadas por 6.458 proteínas de núcleo, 4.666 proteínas de processos biológicos, 4.248 proteínas com função molecular, 3.596 proteínas de cloroplasto, 2.784 proteínas do citoplasma, entre outras. Baseado nos códigos de evidência das proteínas do Gene *Ontology*, agregamos a informação do histórico da proteína identificada pelo método bi-direcional aos transcritos para aumentar a credibilidade da anotação (Tabela 9), sendo que uma proteína pode ter múltiplos códigos.

Tabela 9. Número de transcritos anotados pelo banco do *Gene Ontology*, pela abordagem bi-direcional e agrupados nas categorias de código de evidência.

Código	Tipo de evidência	# transcritos SUCEST	# transcritos RNA-Seq
NAS	<i>Non-traceable Author Statement</i>	397	746
TAS	<i>Traceable Author Statement</i>	1.569	2.652
IEA	<i>Inferred from Electronic Annotation</i>	2.039	3.069
IGC	<i>Inferred from Genomic Context</i>	0	2.379
ISA	<i>Inferred from Sequence Alignment</i>	1	0
ISM	<i>Inferred from Sequence Model</i>	7.379	14.424
ISO	<i>Inferred from Sequence Orthology</i>	164	239
ISS	<i>Inferred from Sequence or Structural Similarity</i>	9.316	17.964
RCA	<i>Inferred from Reviewed</i>	18.873	34.750
IC	<i>Inferred by Curator</i>	212	0
ND	<i>No biological Data available</i>	5.038	9.589
EXP	<i>Inferred from Experiment</i>	2	1
IDA	<i>Inferred from Direct Assay</i>	13.436	23.623
IEP	<i>Inferred from Expression Pattern</i>	1.772	2.999
IGI	<i>Inferred from Genetic Interaction</i>	1.126	2.379
IMP	<i>Inferred from Mutant Phenotype</i>	4.929	9.725
IPI	<i>Inferred from Physical Interaction</i>	1.218	2.206

Nos transcritos de *RNA-Seq* anotados em categorias por GO é possível identificar várias categorias relacionadas a respostas a estresses (*response to wounding, response to cold, defense response to bacteria, defense response to fungus, response to water deprivation, response to salt stress*) e categorias ligadas à transdução de sinal (*kinase activity, regulation of transcription, positive regulation of transcription, DNA binding, sequence specific DNA-binding*). Nas figuras 31, 32 e 33 estão representadas as dez ontologias mais abundantes em relação à anotação dos transcritos do SUCEST e *RNA-Seq*, em todos os casos as dez ontologias mais abundantes são praticamente as mesmas

encontradas no SUCEST e RNA-Seq, mudando apenas duas ou três posições na ordem de abundância.

Figura 31. Distribuição das Ontologias de Processos Biológicos mais frequentes no SUCEST e RNA-Seq

Figura 32. Distribuição das Ontologias de Funções Moleculares mais frequentes no SUCEST e RNA-Seq.

Figura 33. Distribuição das Ontologias de Componentes Celulares mais frequentes no SUCEST e *RNA-Seq*.

Deste modo, cerca de (24%) dos transcritos do SUCEST e 10% dos transcritos do *RNA-Seq* foram anotados com ao menos uma categoria do GO. A figura 34, descreve a distribuição das anotações do GO entre as ontologias, mostrando que 10.152 transcritos do SUCEST e 12.481 transcritos são anotados de termos provenientes das três ontologias, e aproximadamente 28% estão associados com categorias associadas a pelo menos duas ontologias. As anotações são muito úteis para estudos e caracterizações de genes específicos e identificação de genes de interesse, principalmente para experimentos de microarranjo e análises para extrair e coletar conjuntos de genes com características funcionais específicas.

(A)

(B)

Figura 34. Diagrama de Venn da distribuição dos transcritos anotados do SUCEST (A) e *RNA-Seq* (B) com as anotações por GO entre as três diferentes ontologias.

Para separar as anotações primárias obtidas a partir da abordagem uni-direcional, do método bi-direcional e anotações resultantes do *Gene Ontology*, os dados foram armazenados em entidades distintas do banco de dados como *similarity* e *transcript_annotation*. Isto inclui informações de identificação das proteínas, identificação de acesso do *Uniprot*, propriedades, código de evidência, origem e data dos dados. Deste modo, anotações posteriores poderão utilizar estas informações de forma integrada ao consultar um determinado transcrito.

Os transcritos anotados pelas gramíneas e pelo GO foram agregados, resultando na anotação de 9.579 transcritos do SUCEST anotados por ambas as *databases*, com 576 anotados somente pelo GO e 13.570 somente pelas gramíneas. Para os transcritos anotados do *RNA-Seq*, foram identificados 16.708 transcritos anotados pelos bancos de gramíneas e GO, 29.803 anotados apenas pelo banco de gramíneas e 2.931 anotados apenas pelo *Gene Ontology*, estes resultados indicam que cerca de 3 mil transcritos podem não ter sido selecionados pelo banco de gramíneas por causa dos critérios de estringência ou por serem proteínas específicas de cana-de-açúcar.

6.1.3. Anotação Funcional por Vias Metabólicas e Enzimas (KEGG)

A associação dos transcritos do SUCEST e *RNA-Seq* com enzimas e vias metabólicas foi realizada através do banco de dados do KEGG, utilizando a ferramenta online KAAS (*KEGG Automatic Annotation Server*), devido à restrição de acesso às sequências de proteínas depositadas no KEGG. Esta ferramenta permite a anotação funcional por comparação por BLAST, contra um banco curado de genes do KEGG. Para o mapeamento contra o banco do KEGG, foram utilizadas como referência os genes curados de todas as plantas monocotiledôneas (Anexo VII), como o transcriptoma do SUCEST e *RNA-Seq* representam provavelmente mais de 90% dos transcritos de cana-de-açúcar, o método de melhor *hit* bi-direcional é mais indicado para comparação contra os transcriptomas completos curados no KEGG, e para determinar o número de ortólogos. Para estas análises foram considerados os critérios de BHR > 0.95 e *escore bit* > 30.

Com o mapeamento dos transcritos no KEGG, foram identificadas 123 vias metabólicas para os transcritos do SUCEST e 126 vias com os transcritos obtidos pelo

RNA-Seq (Tabela 10), as três vias presentes apenas no RNA-Seq são *FoxO signaling pathway*, *Indole alkaloid biosynthesis* e *Monoterpenoid biosynthesis*. No RNA-Seq e no SUCEST foram identificadas 1.329 e 924 enzimas, respectivamente. O software KAAS atribuiu 1.527 KOs aos 2.584 transcritos do SUCEST, porém com uma média de 4 transcritos por KO, enquanto aos transcritos do RNA-Seq 2.932 KOs foram atribuídos aos 14.020 transcritos, porém com uma média de 11 transcritos por KO, indicando que podem haver muitos transcritos alelos, variantes de *splicing* ou partes de um mesmo gene.

Tabela 10. Número total de Vias, enzimas, KEGG ortólogos e transcritos do SUCEST e RNA-Seq identificados com o software KAAS.

	#SUCEST	#RNA-Seq
KEGG	2.584 (6%)	14.020 (7%)
kEGG Ortólogos (KO)	1527	2932
KEGG Enzimas	924	1329
KEGG Vias	123	126

As vias metabólicas mais representadas no transcriptoma foram *Metabolic Pathways* com 7.385 transcritos do RNA-Seq e 1.593 do SUCEST, e *Biosynthesis of secundar metabolites* com 2.761 transcritos do RNA-Seq e 631 do SUCEST. Essas vias foram removidas da Figura 35 para melhor representação das vias mais representadas por cada transcriptoma. Na tabela 11, apresentamos as enzimas mais representadas nos transcriptoma do SUCEST e RNA-Seq.

Figura 35. Vias metabólicas do KEGG mais representadas com os transcritos anotados do SUCEST e RNA-Seq pelo KAAS, for removidas desta figura as Vias *Metabolic Pathways* e *Biosynthesis of secundar metabolism*.

Tabela 11. As sete Enzimas mais representadas por transcritos do SUCEST e RNA-Seq

Enzimas	#SUCEST	#RNA-Seq
pfkA, PFK; 6-phosphofructokinase 1 [EC:2.7.1.11]	32	240
UGT; glucuronosyltransferase [EC:2.4.1.17]	90	210
RFA1, RPA1, rpa; replication factor A1	64	180
HK; hexokinase [EC:2.7.1.1]	32	152
peroxidase [EC:1.11.1.7]	76	148
AOC3, AOC2, tynA; primary-amine oxidase [EC:1.4.3.21]	8	144
E2.3.1.9, atoB; acetyl-CoA C-acetyltransferase [EC:2.3.1.9]	42	140

Com estes mapeamentos concluímos a primeira versão das vias metabólicas em cana-de-açúcar baseado no banco do KEGG. As vias metabólicas e componentes foram modelados e implementados no banco de dados relacional compondo um módulo adicional, para ser integrado à plataforma *CaneRegNet*. Adicionalmente, estes dados no banco de dados relacional são convertidos para armazenamento das Vias Metabólicas e seus componentes, em um *banco de dados de grafos* para realizar análises diretamente no grafo ou rede de nós para cada via metabólica, integrada aos dados de expressão gênica.

Com essa informação de vias metabólicas, verificamos que o chip de microarranjo contendo sondas para detecção de 14.522 diferentes transcritos de cana-de-açúcar (Chip Agilent) apresenta 1613 transcritos (11% do total do chip) identificados pelo mapeamento de vias metabólicas, o que corresponde a 62% dos transcritos identificados em vias metabólicas. Assim, podemos observar que a lamina de microarranjo possui uma boa cobertura das vias metabólicas de cana-de-açúcar.

Considerando os mapeamentos pelo método bi-direcional contra os bancos de dados das 6 gramíneas, *Gene Ontology* e *KEGG*, ao final foram anotados 23.149 transcritos do SUCEST. O total de transcritos identificados pelas duas abordagens foi de 37.552 transcritos, os 5.589 restantes foram mapeados contra os bancos de dados do *Uniprot* e NR-NCBI, com critério de e-valor 1e-05 e sendo identificados 1.552 transcritos, resultando em 4.037 transcritos específicos de cana-de-açúcar (i.e., não identificados nos bancos analisados). Em relação aos transcritos de *RNA-Seq* foram anotados 49.442 transcritos pelo método bi-direcional contra os bancos descritos acima,

e foram identificados pelas duas abordagens 152.922 transcritos. Os 42.753 transcritos restantes foram mapeados contra os bancos de dados do *Uniprot* e NR-NCBI, com critério de e-valor 1e-05 e foram identificados 7.081 transcritos, resultando em 35.672 transcritos específicos de cana-de-açúcar.

6.1.4. Predição de proteínas e Identificação de domínios PFAM

Para determinar se os transcritos são *full-length*, primeiramente foi utilizado o *software* ESTScan para a criação de uma matriz de *codon-usage* para cana-de-açúcar, e em seguida esta matriz foi utilizada para a predição das proteínas para os transcritos do SUCEST e RNA-Seq. A predição das proteínas a partir do 43.141 SAS, gerou um total de 39.021 proteínas preditas, e baseado apenas nas proteínas preditas, foram identificadas 6.679 (15.5%) dos transcritos apresentando códon de iniciação (*Start*) e códon de terminação (*Stop*) (Tabela 12). A predição de proteínas a partir dos 195.765 transcritos obtidos com o *RNA-Seq*, gerou um total de 187.935 predições de proteínas, sendo identificado um total de 24.259 proteínas com códon de iniciação e terminação. Estes resultados serviram como parte da metodologia para predição dos genes *full-length* em cana-de-açúcar.

Tabela 12. Número total de proteínas preditas com a matriz de *codon-usage* de cana-de-açúcar para os transcriptoma do SUCEST e RNA-Seq.

	SUCEST	RNA-Seq
Total transcritos	43.141	195.765
Total proteínas preditas	39.021	187.935
Total proteínas sem predição	4.120	7.830
Códon início e fim	6.679	24.259
Códon início	7.109	30.941
Códon fim	1.419	6.5521
sem códon início e fim	11.043	67.214
Domínios PFAM	18.658	57.525
tamanho médio	625	566
menor proteína	51	51
maior proteína	5.214	8.856

As proteínas preditas foram analisadas pela presença de domínios conservados e 18.658 transcritos do SUCEST e 57.525 transcritos do *RNA-Seq* apresentaram algum

domínio do banco de dados do Pfam (Punta et al. 2011), sendo que as categorias de domínios mais representativas são mostradas nas figuras 36 e 37. Apesar de algumas diferenças entre os 10 domínios mais abundantes (Figuras 36 e 37), todas aparecem com grande representação entre os dois transcriptoma. De acordo com a descrição do próprio Pfam, os domínios PPR, WD40 e LRR, são de função desconhecida ou presentes em uma grande gama de proteínas não relacionadas. Por outro lado, em concordância com a categorização do *Gene Ontology* (Figuras 31, 32 e 33), é possível observar domínios relacionados às categorias de transdução de sinal, principalmente quinases (Pkinase e Pkinase_Tyr) e de resposta a estresses, como o domínio NB-ARC, o qual está relacionado a respostas a doenças (van Ooijen et al. 2008). O NB-ARC foi identificado em maior abundância no *RNA-Seq*, e comparativamente entre os transcritos específicos dos genótipos foram encontrados com maior frequência em *S. spontaneum*, sendo a espécie responsável por introduzir tolerância a estresse os híbridos de cana-de-açúcar (Paterson et. al., 2013), sugerindo seu envolvimento com este fenótipo. Por último, observamos o domínio citocromo p450 (p450), pertencente a uma superfamília de proteínas e responsável por catalisar a oxigenação de uma grande variedade de compostos, presente em enzimas como C4H e F5H da via de biossíntese de lignina, bem como em diversas outras enzimas de inúmeras vias, como, por exemplo, em enzimas do metabolismo de lipídeos.

Figura 36. O gráfico representa as proporções dos 7 domínios mais abundantes no PFAM.

Figura 37. O gráfico representa as proporções dos 7 domínios mais abundantes no PFAM.

6.2. Identificação de proteínas Ortólogas entre Cana-de-açúcar e Gramíneas

Para a identificação de genes homólogos com outras espécies foram utilizados os *softwares Inparanoid* e *MultiParanoid* para identificar as proteínas ortólogas e parálogas de cana-de-açúcar a partir das proteínas preditas do SUCEST com Sorgo, Milho, Arroz e *Arabidopsis* (Tabela 13). Para a identificação de ortólogos para as proteínas preditas de *RNA-Seq* foi utilizado o *Hieranoid*, uma nova versão do MultiParanoid com as espécies de Sorgo, Milho, Arroz, *Brachypodium*, *Setaria*, *Panicum* e *Arabidopsis*, como descrito na metodologia.

Tabela 13. Número de genes ortólogos putativos entre cana, gramíneas e *Arabidopsis*, baseado no InParanoid (comparação pareada) e MultiParanoid (grupos de ortólogos entre múltiplas espécies usando a saída do InParanoid)

	<i>A. Thaliana</i>	<i>Z. Mays</i>	<i>O. Sativa</i>	<i>S. Bicolor</i> v1.4	SUCEST
# proteínas	35.386	106.046	51.258	36.338	39.021
# ortólogos	17.299	32.881	21.041	16.332	17.176
# grupos	5.820	9.847	8.135	8.503	11.645

Foram formados 18.611 grupos de ortólogos, nos quais estão distribuídas 129.108 proteínas das diferentes espécies comparadas. Deste total, 11.645 grupos de ortólogos estão representados por pelo menos uma proteína de cana-de-açúcar de um

total de 17.176 (43%) proteínas. Destes grupos de ortólogos foram identificadas 8.083 proteínas parálogas em cana-de-açúcar, agrupadas em 2.554 grupos, com uma média de 3,16 proteínas por grupo. Este agrupamento por ortologia permitiu classificar 16.332 (44,9%) proteínas de Sorgo, 32.881 (31%) proteínas de Milho, 21.041 (41%) proteínas de Arroz e 17.299 (48,9%) proteínas de *Arabidopsis* (Figura 38). Dentre todos os grupos de ortólogos, foram identificados 5.201 grupos de ortólogos representado por proteínas de todas as espécies. A análise entre as proteínas parálogas em cana-de-açúcar e a análise de predição de domínios, identificou que em média 93,4% das proteínas dentro de um mesmo grupo possuem ao menos um domínio em comum, reforçando a estrringência e robustez do método de agrupamento por ortólogos e parálogos.

Figura 38. Distribuição das proteínas de cana-de-açúcar em grupos ortólogos compartilhados com outras espécies. Os valores se referem ao total de grupos de ortólogos e de proteínas identificadas como ortólogos com cana-de-açúcar, e parálogos dentro da própria espécie.

Os grupos de ortólogos foram validados pela comparação da anotação manual de 9 genes associados com metabolismo de sacarose (Hotta et. al., 2013). Dos 178 genes anotados que foram agrupados em 28 grupos de ortólogos, somente 2 foram falso positivos e não correspondiam a anotação da enzima modelo. Adicionalmente 31 SAS incompletos que não foram classificados em nenhum dos grupos, foram anotados manualmente como parte de um dos grupos de ortólogos, como parálogos putativos. Entre as outras espécies, 17 genes não foram classificados em nenhum grupo e também poderiam ser incluídos em um dos grupos estudados. Estes grupos de ortólogos permitiram a associação dos oligonucleotídeos presentes nas lâminas de microarranjo em cada um dos clusters de ortólogos. Esta análise de grupos de ortólogos entre as espécies associada à lâmina de microarranjo permitiu identificar pares comuns entre lâminas das diferentes espécies, eliminando um viés na análise de cada lâmina. Por

exemplo, 246 grupos de ortólogos presentes em ambos os bancos de cana-de-açúcar e milho, foram associados a experimentos de microarranjo para estudo de ritmo circadiano (Hotta et. al., 2013). Destes, 114 (46,3%) foram considerados circadianos no banco de cana-de-açúcar e 29 (11,8%) foram considerados circadianos no banco de milho. No total, a lâmina de microarranjo de cana-de-açúcar apresentou 5.490 grupos de ortólogos, 2.027 dos quais (36,9%) foram considerados circadianos, uma proporção muito maior que as encontradas em outros experimentos de ritmo circadiano por microarranjo em outras plantas (Hotta et. al, 2013). Também comparamos se o período dos grupos de ortólogos foi conservado em diferentes espécies. A comparação pareada usando grupos de ortólogos circadianos que eram comuns a um par de lâminas mostrou que o maior coeficiente de determinação (r) foi entre os dois conjuntos de dados de *Arabidopsis* ($r=0.72$). Entre as comparações utilizando os grupos de ortólogos de cana-de-açúcar, o maior coeficiente de determinação foi entre cana-de-açúcar e Milho ($r=0.67$) (Hotta et. al, 2013).

Para permitir uma estudo mais abrangente de cada grupo de ortólogos e parálogos identificados entre as gramíneas foi realizada a análise filogenética com o software clustalw (Thompson et. al., 2002) em duas etapas a primeira para realizar o alinhamento (Figura 39), com os parâmetros: -align, -bootstrap=500, -type=protein -matrix=BLOSUM, -iteration=alignment -numiter=500 e em seguida calcular a árvore filogenética, com os parâmetros: -tree -bootstrap=500 -clustering=NJ -outputtree=phylip, para a geração das imagens das árvores foi utilizado o software newick (Junier et. al., 2010), com a linha de comando: `nw_display -s -S -v 30 -b 'opacity:5' -w 900 -i 'font-size:5' -l 'font-family:serif;font-style:italic' -d 'stroke-width:2;stroke:blue' group1.ph 1> group1.svg 2>>group1.result`, e finalmente convertendo para formato png com o software *convert* (Figura 40). Abaixo apresentamos o resultados para o alinhamento de um dos grupos de ortólogos/parálogos :

```

LOC_0s07g46670.3      ---MSAMAMSSSLNPNAAPIPAAFRRVEDFSPWWELVKTTAWFRDHWFRQHQMYEDAAA
LOC_0s07g46670.2      ---MSAMAMSSSLNPNAAPIPAAFRRVEDFSPWWELVKTTAWFRDHWFRQHQMYEDAAA
SCBFSB1046H07.g       -----MASSSLNPNAPLFIIPAAYRQVEDFSPWWELVKTTAWFRDHWFRQHQHQAAYD
SCRULB1060G11.g       -----MASSSLNPNAPLFIIPAAYRQVEDFSPWWELVKTTAWFRDHWFRQHQHQAAYD
SCRURT2009E09.g       -----MASSSLNPNAPLFIIPAAYRQVEDFSPWWELVKTTAWFRDHWFRQHQHQAACD
SCJFRZ2005G11.g       -----MASSSLNPNAPLFIIPAAYRQVEDFSPWWELVNTTAWFRDHWFRQHQHQAAYD
SCCCAD1004E07.g       -----MASSSLNPNAPLFIIPAAYRQVEDFSPWWELVKTTAWFREHWFRQHQHQAACD
SCCCRZ1002H05.g       -----MASSSLNPNAPLFIIPAAYRQVEDFSPWWELVKTTAWFREHWFRQHQHQAACD
Sb02g041980.1        MSTVAAMASSSLNPNAPLFIIPAAYRQVEDFSPWWELVKTTAWFRDHWFRQHQHQAAYD
GRMZM2G327692_T01    MSAVAAMATSSSLNPNAPLFIIPAAYRQVEEFSPWWELVNTTAWFRDHWFRQHQHQAAYA
LOC_0s03g23010.5      ---MSATAVSSSLNPDAPLFIIPAARQVEDFSPQWDLVKTTAWFRDHWFRQHQHQAAD
SCUTCL6035C09.g       --MSAVAVSSSLNPDAPLFIIPAALLQVEDFSPQWDLVTTGWFRDHWFRQHQHQAAD
AT4G14270.2          MTATIDRRSLTNPNAPVDFPVEFREVEDFSPKWWDLVTTSKWFRDFWLSANSENEFLGG
AT4G14270.1          MTATIDRRSLTNPNAPVDFPVEFREVEDFSPKWWDLVTTSKWFRDFWLSANSENEFLGG
                        *:***.*.*.*.  .***:***:***:***:***:***:***:***:***:***:***:
LOC_0s07g46670.3      EEDH--HRQVFDLDAALLPDDSLDLDLMDVTDLDFYTPD--PPVHHKPTTTPTLDAVL
LOC_0s07g46670.2      EEDH--HRQVFDLDAALLPDDSLDLDLMDVTDLDFYTPD--PPVHHKPTTTPTLDAVL
SCBFSB1046H07.g       AAA---FSTDDDFDVAALLPDD---SVDLLDLDLDFYAPDVHQVAKAPALLPGYDL
SCRULB1060G11.g       AAA---FSTDDDFDVAALLPDD---SVDLLDLDLDFYAPDVHQVAKAPALLPGYDL
SCRURT2009E09.g       AA-----DDFDVAALLPDDSDVLLDMDVTDLDFYAPDVHQVAKAPALLPGYDL
SCJFRZ2005G11.g       TAFSTADADDDFDVAALLPDDSIDLLDMDVTDLDFYAPDVHQVAKAPALLPGYDL
SCCCAD1004E07.g       AAA-----LLPDDSIDLFDHQVAK--PAPALLPGYDL
SCCCRZ1002H05.g       AAA-----LLPDDSIDLDQVAKAPAPALLPGYDL
Sb02g041980.1        AAFA--ADAADDFDVAALLPDDSDVLLDMDVTDLDFYAPDVHHQAKAPAPGYDL
GRMZM2G327692_T01    AAAD-----AELDVAALLPDD---SVDLLDLDLDFYAPDVHHHNAKAPALLPGYDL
LOC_0s03g23010.5      SLALH-----DAAADDDDLGLLPDDAFDDDDDDDLFFDQTHNL
SCUTCL6035C09.g       QLDAG-----GLLADDEDLFFTRPSFKKAPPRPVKPKDPAGGRP
AT4G14270.2          DDFS-----VMEEFEEMIASSDGGSMADTVT--DVASYLKML
AT4G14270.1          DDFS-----VMEEFEEMIASSDGGSMADTVTEADVASYLKML
                        :
LOC_0s07g46670.3      RALSLSNPRLGGAGMVR-----HAEKPAQFVAAGGAARHVIHQPR-----
LOC_0s07g46670.2      RALSLSNPRLGGAGMVR-----HAEKPAQFVAAGGAARHVIHQPR-----
SCBFSB1046H07.g       RALSLSSPRAAAVVAAPSRAQQQ--RFADKPAYHAGAIHQPR-----
SCRULB1060G11.g       RALSLSSPRAAAVVAAPSRAQQQ--RFADKPAYHAGARGAIHQPR-----
SCRURT2009E09.g       RALSLSSPRAAAVVAAPSRAQQQQRLADKPAYHAGAIHQPR-----
SCJFRZ2005G11.g       RALSLSSPRAAAVVAAPSRAQQQ--RFADKPAYHAGARGAARRAIHQPR-----
SCCCAD1004E07.g       RALSLSFPRPPPPRAQ-----QQQKFSOKP-----
SCCCRZ1002H05.g       RALSLSSPR--AAAQ-----QQQRFADKPAYHAAPSTSLARRALLLLLLPPWLLFL
Sb02g041980.1        RALSLSSPRAAATVVAAP--SPQQQRYADKPAYHAGGRGAARRAIHQPR-----
GRMZM2G327692_T01    RALSLSSPRAAAPSPSP--LLRAQQQRHADRPAHHARGAARRVIIHQPR-----
LOC_0s03g23010.5      LVDPQPPAALKTGTD-----FK-----
SCUTCL6035C09.g       LNIAESTKEKIYRSKVSSCSPKYNQKKYMNPNFNCRRNHIIYQPR-----
AT4G14270.2          LNIAESTKEKIYRSKVSSCSPKYNQKKYMNPNFNCRRNHIIYQPR-----
AT4G14270.1          LNIAESTKEKIYRSKVSSCSPKYNQKKYMNPNFNCRRNHIIYQPR-----

```

Figura 39. Alinhamento entre genes ortólogos e parálogos com Clustalw.

Figura 40. Árvore filogenética para os genes ortólogos e parálogos construída com Clustalw.

Para a identificação de genes ortólogos entre as proteínas preditas por *RNA-Seq* das gramíneas (Milho, Arroz, Sorgo, *Brachypodium*, *Panicum*, *Setaria*) e *Arabidopsis* foi utilizado o *software* Hieranoid. Como descrito na metodologia, o método infere ortologia entre múltiplas espécies aplicando progressivamente o método pareado do InParanoid, estas inferências são realizadas em cada nó de bifurcação da árvore guia das espécies (Figura 3). As oito espécies consideradas nesta análise estão organizadas nos seguintes nós (*Saccharinae*, *Andropogoneae*, *Paniceae*, *Panicoideae*, *BEP_clade*, *Poaceae* e *Mesangiospermae*), seguindo esta ordem para a análise de pareamento entre os pares (Tabela 14). A cada nó é gerado um conjunto de sequências consenso para ser utilizado na comparação com a outra espécie ou grupo em um nó superior, as sequências não agrupadas são utilizadas junto com as sequências consenso na comparação seguinte.

Tabela 14. Número de genes ortólogos putativos entre cana, gramíneas e *Arabidopsis*, dentro de cada nó da árvore utilizando o Hieranoid.

	RNA-Seq	<i>S. bicolor</i>	<i>Z. mays</i>	<i>S. italica</i>	<i>P. virgatum</i>	<i>O. sativa</i>	<i>B. distachyon</i>	<i>A. thaliana</i>	Saccharinae	Andropogoneae	Paniceae	Panicoideae	BEPclade	Poaceae
# proteínas	187935	39441	63540	40599	70071	49061	31029	35386	12858	23353	21858	33099	17382	35362
Saccharinae	22371	17647	-	-	-	-	-	-	-	-	-	-	-	-
Andropogoneae	4970	9693	29045	-	-	-	-	-	10659	-	-	-	-	-
Paniceae	-	-	-	25902	32904	-	-	-	-	-	-	-	-	-
BEPclade	-	-	-	-	-	24167	21451	-	-	-	-	-	-	-
Panicoideae	4460	3206	5296	3275	9098	-	-	-	1464	17323	18690	-	-	-
Poaceae	1454	682	2488	1132	3519	5747	3013	-	151	1019	1184	18982	16358	-
Mesangiospermae	1922	723	2614	723	2970	1507	1053	21300	83	276	264	995	325	12271
# ortólogos	35177 (19%)	31692 (80%)	39443 (62%)	31032 (76%)	48482 (70%)	31421 (64%)	25517 (82%)	21300 (62%)	12357	18618	20138	19977	16683	12271
# grupos	20444	24685	28921	26163	33826	22798	20652	11423	22847	26641	26768	23523	22899	11423

Foram formados 20.444 grupos de ortólogos com cana-de-açúcar, sendo que 35.177 (18,7%) proteínas estão representadas em pelo menos um grupo entre as diferentes espécies. Destes grupos de ortólogos foram identificadas 22.733 proteínas parálogas em cana-de-açúcar, agrupadas em 8.005 grupos, com uma média de 2,8 proteínas por grupo. Este agrupamento por ortologia permitiu classificar 20.455 (51,8%) proteínas de Sorgo, 26.219 (41,2%) proteínas de Milho, 19.368 (39,5%) proteínas de Arroz, 28.517 (40,7%) proteínas de *Panicum*, 18308 (45,1%) proteínas de *Setaria*, 16.333 (52,6%) proteínas de *Brachypodium* e 15.651 (44,2%) proteínas de *Arabidopsis* (Figura 38). Dentre todos os grupos de ortólogos foram identificados 5.201 grupos de ortólogos representado por proteínas de todas as espécies.

A análise entre as proteínas parálogos em cana-de-açúcar e a análise de predição de domínios identificou que em média 92,5% das proteínas pertencentes a um mesmo grupo do Hieranoid possuem ao menos um domínio em comum, reforçando a estrringência e robustez do método de agrupamento por ortólogos e parálogos. Este resultado permite adicionar mais um nível no critério de transferência de anotação para as proteínas de cana-de-açúcar identificadas.

6.3. Predição de genes *Full-Length*

Para determinar se os transcritos montados a partir do sequenciamento de *RNA-Seq* são *full-length*, foi realizada a predição das CDS para estes transcritos, com a identificação das ORFs e os códons de iniciação e terminação. Nesta análise foram utilizados os genes CDS *full-length* de seis gramíneas como referência: Sorgo, Milho, *Panicum*, *Setaria*, Arroz e *Brachypodium*, seus respectivos genes contêm o mRNA completo (5' UTR + CDS + 3' UTR) (Tabela 15).

Tabela 15. Número de genes, genes *full-length**, genes com 5' e 3' UTR e tamanho médio da CDS em cada espécie.

Espécies	# Genes	# <i>Full-Length</i>	Média CDS	# Genes 3' UTR	# Genes 5' UTR
<i>B. distachyon</i>	31.029	7.441	1.280,32	22.485	11.601
<i>O.sativa</i>	49.061	18.119	1.081,58	34.116	34.584
<i>P. virgatum</i>	70.071	14.062	1.024,05	3.412	3.419
<i>S. bicolor</i>	39.441	27.533	1.208	37.262	36.813
<i>S. italica</i>	40.599	14.007	1.134,48	27.022	21.421
<i>Z. mays</i>	63.540	26.353	997,89	58.046	56.442

Desenvolvemos uma abordagem, como descrito na metodologia, para a identificação de transcritos *full-length* em cana-de-açúcar, ou outros organismos, desde que tenham disponíveis transcriptoma ou genomas de espécies próximas evolutivamente. A abordagem é composta por três etapas. As duas primeiras análises são baseadas no alinhamento entre os transcritos de cana e as CDSs dos genes *full-length* das espécies por BLASTN, com parâmetros: *e-value* < 1e-15, *max_target_seqs*: 10000, *outfmt*: 6, *dust*: NO, *num_threads*: 20. O resultado do alinhamento por BLASTN foi processado por um *script* (Anexo XIII) para calcular a cobertura total do gene *full-length*, selecionando apenas os transcritos com cobertura >50% e o respectivo gene com cobertura >95% e identidade > 80%. Deste modo, na análise I, baseada na cobertura total do gene pelo transcrito do 1º hit, o transcrito foi considerado *full-length*, somente se o mesmo cobre completamente o primeiro hit da base de genes *full-length*. Foram identificados 9.960 (9.410 únicos) transcritos *full-length* em cana-de-açúcar e os genes e transcritos identificados nesta primeira etapa da análise foram descartados das análises posteriores.

Na análise II, baseada na cobertura total do gene por dois ou mais transcritos, foram selecionados os genes *full-length* de cada espécie e se dois ou mais transcritos cobrirem o respectivo gene, baseado nos mesmos critérios adotados na análise I, os genes e respectivos transcritos são selecionados. Em ambas as análises, os genes de cada espécie foram utilizados em uma ordem específica para a busca por similaridade (Sorgo, Milho, Panicum, Setaria, Oryza e Brachypodium), baseada nas espécies com maior número de *hits* com os transcritos (Anexo V). Os transcritos identificados como *full-length* em comparação com Sorgo não foram utilizados na análise com os genes de Milho, e aqueles identificados como *full-length* na comparação com Milho, não foram utilizados na análise com *Panicum* e assim por diante. Esta análise identificou 26.384 transcritos representando 4.027 genes de gramíneas (3.952 transcritos únicos), com uma média de 6,55 transcritos por gene *full-length*.

Na análise III, baseada no tamanho médio das CDSs e presença de códon de iniciação e terminação, foi feito o cálculo do tamanho médio da CDS (10.030,85pb, Tabela 16) para os genes *full-length* das seis gramíneas e *contigs* não classificados nas duas análises anteriores foram selecionados como genes *full-length* putativos, se o tamanho da respectiva CDS for maior que 1.030,85 e a CDS for completa (incluindo códons de iniciação e terminação), abordagem similar por Liu e colaboradores (Liu et. al., 2012). Esta análise produziu um total de 4.243 transcritos que atendem aos critérios da análise III.

A integração da predição dos transcritos *full-length*, com a predição de famílias de genes e parálogos apresentados previamente, permitiu a identificação de transcritos *full-length* unicamente representados pelo agrupamento de alelos, variantes de *splicing* e parálogos para um dado gene.

Tabela 16. Número de transcritos *full-length* identificados em cada análise.

			Análise I		Análise II		Análise III
			1o Hit contra CDS dos genes FL		Múltiplos hits contra CDS dos genes		CDS transcrito > média das CDS dos genes FL
			Cob >=95% and Ident >= 80%		Cob >=95% and Ident >= 80%		CDS > 1030,85
Espécies	#genes	#Full-Length	#genes	#transcritos	#Genes	#transcritos	#transcritos
<i>Brachypodium distachyon</i>	31.029	7.441	2.043	1.797	23	113	-
<i>Oryza sativa</i>	49.061	18.119	3.281	2.791	53	351	-
<i>Panicum virgatum</i>	70.071	14.062	5.599	4.114	131	645	-
<i>Sorghum bicolor v2.1</i>	39.441	27.533	7.348	6.515	3.391	22.916	-
<i>Setaria italica</i>	40.599	14.007	3.834	3.393	74	478	-
<i>Zea mays</i>	63.540	26.353	7.809	5.656	355	1.881	-
Total	293.741	107.515	29.914	24.266	4.027	26.384	16.053
Total Únicos por Parálogo			29.364	9.410	3.952		12.365
Total Únicos	293.741	107.515	29.364	9.410	3.952	26.384	4201
Análises I + II + III							
Total genes gramíneas		33.941					
Total FL cana únicos		9.410+3.952+ 4.201 = 17.563					

Como resultado dessas três análises, nós identificamos 40.587 transcritos representando 17.563 genes *full-length* únicos (Tabela 16). Se assumimos o tamanho estimado do transcriptoma de cana-de-açúcar em 33 mil genes (Yilmaz et. al., 2009), nós identificamos 53.2% de todos os genes *full-length* em cana-de-açúcar. Este número de transcritos *full-length* únicos é muito maior que reportado para bibliotecas *full-length* de outras plantas, como poplar (3.990) (Ralph et. al., 2008), tomate (11.502) (Aoki et. al., 2010) e *Brachypodium* (10.513) (Mochida et. al., 2013), mesmo a estimativa do

número total de genes no genoma é similar entre estas plantas (34 mil para tomate, 31 mil para *Brachypodium* e 33 mil para cana-de-açúcar). Notavelmente, a identificação de genes *full-length* em cana-de-açúcar foi conduzida pela montagem *de novo* neste estudo, em contraste com as bibliotecas *full-length* citadas anteriormente, as quais utilizaram um *draft* do genoma para facilitar a montagem. Entre todos os 40.587 transcritos identificados nas análises de *full-length* (Tabela 16), 40.407 mostraram similaridade contra os bancos de dados do Phytozome (Tabela 16). Proteínas desconhecidas, domínios NB-ARC e proteínas quinases foram as categorias mais frequentes (Figura 41), resultados similares aos obtidos para todos os transcritos anotados.

Figura 41. Anotação Funcional dos transcritos *full-length* (40.407). O gráfico mostra as categorias mais frequentes da anotação do Phytozome. Foram identificadas 5.038 categorias.

O comprimento das ORFs de cana-de-açúcar e transcritos seguem a distribuição dos genes *full-length* de outras gramíneas, especialmente quando considerados os comprimentos de ORFs acima de 1.500pb (Figura 42A) e 2.000pb para transcritos (Figura 42B). Cerca de 40% das ORFs de cana-de-açúcar apresentam uma faixa de comprimento de 1.000 a 1.500 pb (Figura 42B). Além da análise para identificação de transcritos *full-length*, considerando inicialmente apenas a cobertura da CDS como mencionado acima, realizamos a estimativa e comparação do comprimento das regiões 5' e 3' UTR (Figura 42C), reforçando a evidência de que os transcritos são realmente

sequências *full-length*, ou ao menos, correspondem a uma CDS completa com partes de regiões UTR.

Figura 42. Distribuição do comprimento das ORFs e cana-de-açúcar, transcritos *full-length* (FLs), e UTRs. Os gráficos representam a comparação da distribuição (barra cinza) do tamanho das ORFs (A) e transcritos *full-length* (B) com outras gramíneas (linhas coloridas). A distribuição do comprimento das UTRs 5' e 3' (C, barras preto e branco) dos transcritos *full-length* em cana-de-açúcar são apresentados também.

A anotação funcional dos transcritos de cana-de-açúcar por RNA-Seq foram baseadas nos bancos de dados do *Gene Ontology*, *Kyoto Encyclopedia of Genes and Genomes* (KEGG), PFAM e Phytozome (v9.0), como descrito anteriormente. Foram identificados 78.273 (40%) dos transcritos em termos GO. As categorias mais representativas foram "*nucleus*", "*protein binding*", e "*regulation of transcription, DNA dependent*". Comparado a anotação do transcriptoma de cana-de-açúcar por Cardoso-Silva e colaboradores (Cardoso-Silva et. al., 2014) (14.983 genes únicos anotados com termos GO), um número maior de transcritos foi anotado pelo banco do GO, 24.085 transcritos, mesmo utilizando os mesmos critérios como descritos no estudo anterior (Cardoso-Silva et. al., 2014) (i.e., comprimento mínimo do *contig* e excluindo isoformas).

O ORFeoma descrito neste trabalho fornecerá informações relevantes relacionadas a expressão alélica e ancestral, especialmente aquelas relacionadas a repostas a estresse, devido à identificação de muitos genes nesta categoria. O ORFeoma servirá de guia para a montagem e validação do genoma de cana-de-açúcar, para unir, ordenar e orientar fragmentos genômicos (Mortazavi et. al., 2011; Xue et. al., 2013).

6.4. Identificação e Classificação de Famílias de Fatores de Transcrição

Enquanto muitas proteínas participam na regulação da expressão gênica, nos limitamos aqui à definição dos TFs a proteínas que contêm motivos estruturais característicos, i.e., domínios de ligação ao DNA, o qual está envolvido no reconhecimento de uma sequência de DNA curta (usualmente 4-8 pares de base). Baseado nesta estrutura de domínio de ligação ao DNA, os TFs foram classificados dentro de um número variável de diferentes famílias (usualmente 40-60), e em plantas, de 5-7% de todas as proteínas codificando genes correspondem a TFs com estas características (Riechmann et. al., 2000a e 2000b).

Para a identificação de TFs em cana-de-açúcar e gramíneas foi desenvolvido um *pipeline* dividido em várias etapas como descrito na metodologia para o catálogo e classificação de famílias de TFs com base em domínios conservados para cada família (Yilmaz et. al., 2009), agregado à criação de uma plataforma (GRASSIUS) para armazenamento e gerenciamento de todas as estas informações, como o catálogo de regiões promotoras e sítio de ligação, vinculadas aos respectivos TFs, além da criação do ORFeoma de TFs de gramíneas. Esta abordagem foi aplicada aos genomas de Sorgo, Milho, Arroz e ao transcriptoma do SUCEST. Mais recentemente esta mesma abordagem foi empregada para o catálogo e classificação dos TFs a partir dos transcritos de RNA-Seq, atualizando o banco de TFs de cana-de-açúcar (Nishiyama et. al., 2014).

Para estimar o número mínimo de TFs em milho, sorgo e cana-de-açúcar, os tamanhos dos genomas respectivos foi sobreposto com a regressão (Figura 43, linhas pontilhadas em vermelho). Estas análises sugerem que um total de 3.470, 2.137 e 1.977 TFs podem ser esperados para milho, sorgo e cana-de-açúcar, respectivamente (Tabela 17). Esta análise não inclui TFs correspondendo às famílias a serem identificadas de frações do genoma que ainda são desconhecidos. Interessantemente, análises similares em genomas não-plantas (fungos e animais), uma tendência similar é esperada (Figura 43, pontos azuis). O ajuste da regressão ($r = 0.74$, ponto preto, Figura 43) foi significativamente melhorado combinando os pontos de plantas e não plantas ($r = 0.82$, dado não mostrado), sugerindo que plantas e animais tem um relacionamento similar entre número de genes e número de TFs.

Figura 43. Estimativa do número de TFs em genomas de gramíneas. Correlação entre o número de TFs e o número total de genes em genomas foi baseado nos genomas completamente anotados de *Arabidopsis*, arroz, poplar e *Chlamydomonas* (pontos pretos). A curva de regressão linear ($r = 0.87$) foi usada para estimar o número de TFs preditos de milho, sorgo, e cana-de-açúcar (Tabela 17)

Tabela 17. Número de TFs em gramíneas. O número esperado de TFs foi estimado da Figura 43.

	<i>Z. mays</i>	<i>S. bicolor</i>	<i>S. hybrids</i>
Tamanho Genoma (MB)	2.5	700	900
Número de Genes	59.000	36.338	33.62
Número de TFs esperado	3.470	2.137	1.997
TFs no GRASSIUS	3.337	2.448	1.647

Quando todos os TFs no GrassTFDB são organizados em espécies e famílias, diferenças interessantes tornam-se evidentes (http://grassius.org/tf_summary.html). Por exemplo, apesar de milho possuir um grande número de genes, o número de TFs em milho em todas as famílias não é o mais representativo. Milho, tem significativamente mais TFs em famílias como Homeobox, AP2-EREBP, bZIP, E2F-DP, MYB, SBP do que em outras gramíneas. Interessantemente, cana-de-açúcar possui muito mais membros nas famílias C3H, C2C2-YABBY, TUB e ZIM, mesmo com um transcriptoma parcial. Estas tendências podem refletir a expansão ou contração de famílias individuais em uma taxonomia particular. A amplificação de reguladores R2R3-MYB durante a radiação das gramíneas (Dia et. al., 2003) fornece um possível mecanismo para a expansão das famílias. A contração do número de genes nestas

famílias poderia estar associada com perda de genes (Bennetzen, 2007) ou perda de domínios, um fenômeno que tem sido reportado para domínios MYB (Braun et. al., 2001). A significativa diferença no tamanho de famílias particulares de TFs entre cana-de-açúcar e Sorgo é potencialmente interessante, desde que as regiões codificadoras para os genomas respectivos tem mostrado ser 94,5% idênticos, altamente sintênicos e colineares (Jannoo et. al., 2007) .

Com todas as estimativas a respeito do número esperado de TFs que o GRASSIUS deveria conter, foi iniciado a geração do GrassTFDB. Neste sentido, foram obtidos todos os TFs de plantas de PlnTFDB (Riano et. al., 2007), PlantTFDB (Guo et. al., 2008) e DBD (Kummerfeld et. al., 2006) e uma lista compreensiva e não redundante de TFs foi gerada. TFs não identificados anteriormente foram buscados através dos genomas mais atualizados e identificados por domínios PFAM. TFs preditos em cada espécie foram agrupados em 59 famílias, excluindo famílias que são na verdade fatores de regulação da transcrição e não possuem sítio de ligação ao DNA, formando uma nova classe de famílias contendo apenas TFs e com a atualização dos critérios de regras de domínios para cada família (Anexo III), cruzando com as regras e famílias dos bancos anteriores do AtTFDB e AGRIS (Davuluri et. al., 2003; Palaniswamy et. al., 2006). O número de TFs em cana-de-açúcar está bem abaixo do esperado no GrassTFDB. Para Arroz, Milho e Sorgo, o número esperado de TFs está bem próximo, e para Sorgo está um pouco acima do esperado, isto pode ocorrer devido a variantes de *splicing*, genes duplicados e alelos.

Com o intuito de atualizar o banco de dados de TFs de cana-de-açúcar a mesma metodologia foi aplicada os transcritos de *RNA-Seq*. O trabalho resultou na identificação de 3.399 (2.651 únicos) transcritos pertencendo a 49 famílias de TFs (Anexos III e XIV). Uma vez que os genes modelos estimados indicam que cerca de 2.000 TFs são codificados no genoma de cana-de-açúcar, a criação do ORFeoma de cana-de-açúcar a partir do *RNA-Seq* identificou uma boa proporção de TFs variantes e alelos (Nishiyama et. al., 2014). As famílias mais abundantes são "*far red-impaired response*" (FAR1-lke), "*MYB-related*", e "*DNA binding protein phosphatase*" (DBP) e "NAC". Interessantemente, no transcriptoma dos genótipos ancestrais sequenciados por *RNA-Seq*, não foram identificados candidatos para as famílias CCAAT-DR1, CCAAT-HAP5, FLO/FLY, Trihelix e VOZ (Figura 44, Anexo XV).

Figura 44. Tabela com a distribuição do número de TFs por família identificados no SUCEST (linhas verdes), no RNA-Seq (linhas azuis), e os TFs *full-length* (linhas vermelhas).

Dentre os 17.563 transcritos *full-length* únicos, nós identificamos 1.705 TFs *full-length*, sendo que destes, pela análise de parálogos identificamos 937 TFs únicos (Anexo XIV), deste modo, foram identificados 46,8% (937 de 2.000) de todos os TFs únicos em cana-de-açúcar. Estes resultados são promissores e facilitarão a caracterização funcional de TFs em cana-de-açúcar, e permitirão a atualização do GRassTFDB e TFORFeome da plataforma GRASSIUS, como podemos ver na Tabela 18, comparativamente com outros bancos de TFs de plantas. Análises futuras destes TFs poderão ampliar o conhecimento dos relacionamentos entre a expressão dos TFs e respostas específicas a tratos de interesse, particularmente por que foram amostrados genótipos de *Saccharum*, que exibem uma grande variação de características incluindo tolerância a estresse, conteúdo de açúcar, produtividade e conteúdo de fibra (*S. spontaneum* e *S. officinarum* são altamente diferentes para essas características).

Tabela 18. Comparação do GrassTFDB e TFs obtidos pelo RNA-Seq com outros bancos disponíveis de TFs.

	<i>Z. mays</i>	<i>O. sativa</i> (japonica)	<i>O. sativa</i> (indica)	<i>S. bicolor</i>	<i>S. species</i> SUCEST (RNA-Seq FL)	<i>Brachypodium</i> <i>distachyon</i>
GrassTFDB	2.630	1.772	1.836	1.759	1.354 (1.705)	1.529
PlnTFDB (79)	3608	2722	2.393	2.231	-	-
PlantTFDB(56)	625	2408	1.891	2198	672	1751
DBD (>150)	2.252	2.114	-	1.452	-	-

Um dos objetivos desta tese foi a identificação de um conjunto completo de TFs putativos em cana-de-açúcar e as gramíneas Sogo, Milho, Arroz (sub-espécies indica e japonica), que são codificados por genomas totalmente sequenciados ou anotados. O primeiro objetivo foi a catalogação das proteínas das gramíneas e predição das proteínas de cana-de-açúcar a partir do transcriptoma para a identificação e classificação de famílias de TFs. Pela divulgação do GRASSIUS, foi possível disponibilizar um primeiro conjunto curado e com famílias de TFs específicas para gramíneas com um conjunto completo putativo de famílias de TFs, com Milho, Sorgo, Arroz (indica e japonica) e cana-de-açúcar (Yilmaz et. al., 2009). Com este trabalho nós contribuímos ativamente como a anotação do transcriptoma de cana-de-açúcar e integramos com as outras gramíneas. Um ponto adicional foi a atualização e definição de domínios modelos (perfil-HMMs do banco do PFAM) para identificação de conjuntos de TFs e classificação em famílias.

6.5. Análise de Expressão senso e antisenso em Cana-de-açúcar

6.5.1. Expressão de pares senso-antisenso por microarranjo

As plataformas de microarranjo de cana atuais cobrem menos que 15% do transcriptoma da cana, não medem transcritos antisenso, não têm alta sensibilidade e especificidade (Nishiyama et. al., 2011), dificultando o estudo das redes de interação gênica e o comportamento do sistema em resposta a diferentes tratamentos.

Foi criada uma nova plataforma de microarranjo, através do *pipeline* descrito na metodologia, com alta especificidade e sensibilidade, permitindo medir padrões de expressão em transcritos senso e antisenso, possibilitando o estudo mais refinado do comportamento das redes de regulação e da maquinaria celular (Bossers et al., 2010).

A lâmina contém 21.901 oligonucleotídeos senso e antisenso duplicados, representando 14.522 SAS divididos em oligonucleotídeos senso (13.660 na posição 50pb, 861 na posição 350pb) e antisenso (7.380 na posição 50pb).

A reanotação e criação de um catálogo funcional dos transcritos contidos na lâmina de microarranjo permitiu classificar as categorias mais abundantes como proteínas desconhecidas "*unknown*" com 10.027 SAS das quais, 5.543 foram confirmadas pela análise de anotação funcional baseado no método bi-direcional, se considerado o método uni-direcional apenas 4.071 SAS são realmente proteínas desconhecidas, as demais possuem alguma categorização secundária. Além da categoria de desconhecidos, as categorias mais abundantes são: "*no match*" (342 SAS),

Metabolismo de Carboidratos (332), proteínas de parede celular (377), proteínas de metabolismo (329), proteínas de estresse (326), e fatores de transcrição (348). Alguns fatores de transcrição foram re-anotados para categorias mais específicas além de outros apresentados (Figura 45 e Anexo XVI).

Figura 45. Distribuição das categorias mais representativas na lâmina de microarranjo separada em transcritos senso e antisenso.

Os métodos para identificação de genes DE em nosso laboratório são baseados na razão de expressão (tratado/controle) entre dois marcadores (Cy3, Cy5) através do método *HTSelf* adaptado. Porém, este método é mais estrigente e dificultam o estudo de transcritos com baixo nível de expressão ou estudo dos antisenso por também apresentarem em geral níveis de expressão muito baixa. Deste modo empregamos o método para seleção dos genes SE, estudando os níveis de expressão por tratamento ou condição individualmente.

Na comparação entre a distribuição dos sinais entre os oligonucleotídeos senso e antisenso, podemos verificar que os oligonucleotídeos antisenso apresentam uma distribuição não homogênea concentrando-se em uma região de mais baixa intensidade na região entre 6 e 8, ao contrário dos oligonucleotídeos senso que apresentam uma distribuição das intensidades mais homogênea variando entre 6 e 13 (Figuras 46 e 47), esta metodologia ampliará o estudo da atividade transcripcional, na tentativa de esboçar as interações gênicas e integrá-las as atividades dos fatores de transcrição e das vias metabólicas.

Figura 46. Gráfico da densidade da distribuição da intensidade dos sinais dos oligonucleotídeos senso abaixo e acima do *background*.

Figura 47. Gráfico da densidade da distribuição da intensidade dos sinais dos oligonucleotídeos antisense abaixo e acima do *background*.

Análises através de experimentos em condições de estresse por seca em série temporal com três tempos (24h, 72h e 120h) sem irrigação, foram conduzidas para avaliar e validar os níveis de expressão do par senso e antisense na lâmina de microarranjo desenhada (Lembke et. al., 2012). A análise de SE identificou 11.780 oligonucleotídeos com sinal acima do sinal de fundo, em ao menos uma das seis condições experimentais e 7.973 oligonucleotídeos em todas as condições experimentais, tratado e controle em (24h, 72h e 120h). Assim, foram identificados como SE, 75% dos oligonucleotídeos senso e 11,9% dos antisense. Interessantemente o número de oligonucleotídeos SE na condição controle e tratado no tempo 24h são iguais, um pouco maior na condição de estresse no tempo 72h e quatro vezes maior na condição de estresse em 120h em relação à condição controle (Lembke et. al., 2012).

Para comparar o perfil de expressão entre o transcrito senso e antisense, selecionamos somente os transcritos com o par senso e antisense na lâmina. Utilizando o sinal de intensidade da razão entre as condições tratado/controle, classificando em os

pares em induzido, reprimido e não diferencial, baseado na metodologia de SE. Foi possível observar que o perfil de expressão dos pares senso e antisenso em geral é muito similar nas três condições experimentais da série temporal. Porém, a maioria dos oligonucleotídeos induzidos ou reprimidos possuem seu antisenso respectivo não diferencial (Figura 48, barras riscadas). As observações seguintes, em relação aos pares (senso e antisenso), mostram que se o oligonucleotídeo senso está induzido ou reprimido o antisenso respectivo também apresenta o mesmo perfil, e a minoria dos pares apresenta perfil de expressão contrastantes, ou seja, oligonucleotídeo senso induzido e antisenso respectivo reprimido (Figura 48).

Figura 48. Padrão de expressão dos pares de oligonucleotídeos senso e antisenso com sinal acima do sinal de fundo. (A) O padrão de expressão dos pares separados pelo tempo sem irrigação. O eixo Y indica o número de oligonucleotídeos, o eixo X indica o padrão de expressão do senso e antisenso. Os tons de cinza indicam o padrão de expressão entre os pares.

Quando analisado o perfil de expressão entre os pares senso e antisenso que foram identificados como DE pelo método adaptado *HTSelf*, não pudemos observar os pares com padrão de expressão oposta e no tempo 24h sem irrigação todos os pares foram classificados como não diferenciais (Figura 49), devido à alta estrigência utilizada do método, relacionada ao baixo nível de expressão dos oligonucleotídeos antisenso.

Figura 49. Padrão de expressão dos pares senso e antisenso identificados pelo método adaptado HTSelf. O eixo X indica o padrão de expressão do oligonucleotídeo antisenso e o eixo Y o número de oligonucleotídeos. As cores na legenda indicam o padrão de expressão dos oligonucleotídeos senso do par.

Os estudos conduzidos em Lembke et. al., 2012, mostraram que a atividade dos pares (senso e antisenso) possui padrões específicos de expressão, indicando que podem estar envolvidos em algum nível da regulação gênica, como evidenciado por hipóteses em outros trabalhos descritos neste estudo.

Deste modo, a lâmina desenhada com oligonucleotídeos senso e antisenso, fornecem uma boa sensibilidade e cobertura do transcriptoma de cana-de-açúcar para a identificação de proporções representativas dos transcritos antisenso e pares de transcritos senso antisenso.

6.5.2. Estudo de transcritos antisenso por *RNA-Seq*

O sequenciamento de *RNA-Seq* feita específica permitiu o estudo e identificação de transcritos antisenso em cana-de-açúcar e comparação com os SAS e as outras seis gramíneas utilizadas nos estudos anteriores. Os *reads* de cada genótipo foram mapeados contra os transcritos dos SUCEST e das seis gramíneas para a presença de antisenso, utilizando o software de alinhamento Bowtie com os parâmetros:

verysensitive-local: (-D 20 -R 3 -N 0 -L 20 -i S,1,0.50); p: 32 (número de processadores); comprimento mínimo de cobertura do read >100 bp.

Os transcritos antisense naturais (NATs) afetam vários processos biológicos. Os NATs modulam a expressão de seus respectivos transcritos através de vários mecanismos (Lapidot et. al., 2006). A expressão de NATs tem sido estudada em diversas plantas, como milho (Ma et. al., 2006), arroz (Zhou et. al., 2008), *Arabidopsis* (Henz et. al., 2007; Wang et. al., 2014) e *Brassica rapa* (Yu et. al., 2013). Em cana-de-açúcar, a expressão de NAT tem sido identificada para respostas a seca (Lembke et. al., 2012) e é modulada de certo modo em ritmo circadiano (Hotta et. al., 2013).

Como cana-de-açúcar não possui genoma referência com genes modelos preditos, a identificação de antisense foi realizada pelo mapeamento dos *reads* contra os genes modelos de outras gramíneas utilizando os *reads* do sequenciamento pelo Ion PGM. Nós detectamos NATs em todas as amostras (Figura 50). A expressão dos transcritos de folha apresentou maior nível de expressão, o que é particularmente evidente quando a análise foi conduzida contra as demais espécies (Figura 50).

Figura 50. Número de *reads* identificados como NATs, baseado na orientação do alinhamento com as gramíneas.

Esses dados sugerem que alguns processo em folhas podem ser mais dependentes ou afetados por mecanismos de antisense do que em entrenós. Além disso,

não foi identificado diferença no nível de expressão total entre os genótipos ancestrais e variedades comerciais híbridas (SUCEST) (Figura 50), sugerindo que a característica foi conservada entre as espécies *Saccharum*, apresentando uma informação interessante para fortalecer a comparação dos níveis de expressão de antisenso com estudo de expressão gênica por microarranjo entre diferentes genótipos já existentes.

Estudos recentes tem identificado NATs em aproximadamente 70% dos *locus* para codificação de proteínas em *Arabidopsis*, sugerindo que NATs são muito mais amplos que previamente postulado (Wang et. al., 2014). Em nossas análises, o mapeamento de *reads* identificou um total de 28.884 transcritos antisenso (14,7%). Analisando estes transcritos com as análises I e II para identificação e transcritos *full-length*, encontramos apenas 46 (0,16% dos transcritos NATs) transcritos *full-length* em cana-de-açúcar sobrepondo totalmente seus respectivos transcritos senso (Nishiyama et. al., 2014). A diferença nos números de NATs (28.844) e NATs *full-length* (46), deve ser interpretada com cuidado, pois nossa metodologia para sequenciamento baseou-se no enriquecimento de RNAs poliadenilados e há evidências de que NATs tendem a ser não poliadenilados (Kiyosawa et. al., 2005; Conley et. al., 2012). Deste modo, os resultados podem sugerir que NATs em cana-de-açúcar, particularmente NATs *full-length*, tendem a ser não poliadenilados ou raramente apresentam cauda poli-A.

Dentre os *reads* alinhados em cada espécie, a grande maioria é de transcritos senso e em algumas espécies uma pequena porção são NATs (Figura 51). Sorgo é a espécie analisada evolutivamente de cana-de-açúcar foi a que apresentou menor taxa de *reads* antisenso identificados, porém como esperado foi a espécie que apresentou maior taxa de similaridade por *reads* senso, até mais que o SUCEST (40 e 27% do total de *reads* alinhados em sorgo e no SUCEST, respectivamente), evidenciando a importância de sorgo para análises em cana-de-açúcar.

Figura 51. Porcentagem dos *reads* senso e NATs alinhados contra cada espécie em relação ao total de *reads* sequenciados no PGM (29 milhões de *reads*).

O cruzamento dos 28.844 transcritos NATs com a anotação funcional pelo método bi-direcional atribui anotação apenas para 2.258 (8%) transcritos NAT, sendo as funções mais representativas: NB-ARC, RING/U-Box, proteína quinase, proteínas de repetição PPR e TPR, 2-oxoglutarate, hidrolases.

6.6. Plataforma CaneRegNet para Estudo e Análise do Transcriptoma de Cana-de-açúcar

Nós desenvolvemos uma plataforma de bioinformática para estudo do transcriptoma de cana-de-açúcar baseado na anotação funcional, incorporando semanticamente camadas baseadas em ontologias e anotações, permitindo novos serviços para a descoberta por dados. Esta plataforma permite o acesso a vários algoritmos e metodologias desenvolvidas neste trabalho para estudo da anotação funcional e de experimentos de microarranjo, e implementação de ferramentas incorporadas à plataforma para análises integradas entre dados de expressão gênica, catálogos e classificações de funções gênicas (Figura 52). Uma das principais funcionalidades da plataforma está relacionada ao sistema de submissão e processamento automatizado de experimentos de expressão gênica por microarranjo. Uma grande variedade de plataformas para integração de dados biológicos tem sido desenvolvidas (Flymine - Lyne et.al., 2007, SMD - Demeter et.al., 2007 e KaPPA-View4 - Sakurai et.al., 2011), porém a maioria não possui código fonte aberto, ou apresenta uma estrutura específica para um determinado organismo, ou um nível de complexidade e indexação da informação que torna impossível o reaproveitamento da plataforma e códigos para desenvolvimento ou adaptação para novos sistemas.

A arquitetura é composta de aplicações web contruídas em três camadas baseada em tecnologias Java Servlets e Java Pages com o *container* web integrado Tomcat (<http://jakarta.apache.org/tomcat/>), acessível via ambientes web. Esta plataforma foi construída para cana-de-açúcar, mas apresenta uma estrutura genérica para o armazenamento e gerenciamento para estudo de transcriptomas de organismos não-modelos ou projetos de sequenciamento de transcriptoma, que possuem uma taxa de identificação de genes com alta similaridade, por bancos como Uniprot ou *RefSeq* muito pequena, comportando vários níveis de anotação e categorização particular.

A entidade central do CaneRegNetDB é o “*probe_set*” (Anexos VI). Todos os tipos de dados armazenados no banco de dados herdam propriedades desta entidade. Índices de busca foram criados para os atributos do *probe_set* (*id*, *probe_type_id*, *probe_name*). Com a criação de um único ponto de referência, todas as buscas podem ser realizadas em todos os tipos de dados ao mesmo tempo, aumentando a eficiência da busca. A entidade *transcript* no *CaneRegNetDB* é responsável pelo armazenamento dos transcriptomas e atualmente, contém transcritos dos transcriptoma de Cana-de-Açúcar (SUCEST e RNA-Seq), Sorgo, Milho, Arroz, *Setaria*, *Panicum*, *Brachypodium* e *Arabidopsis*. Cada experimento de microarranjo e anotação está associado com um transcrito através de uma chave estrangeira que relaciona a anotação ou os resultados do experimento ao transcrito. Esta informação está acessível ao usuário, através das ferramentas de busca relacionando estas múltiplas entidades no *CaneRegNetDB*, assim como a busca por palavras-chave e expressões regulares (parte de nomes) dos transcritos e anotações, exibindo o transcrito e todas as informações relacionadas que compartilham o mesmo *id* do transcrito.

Figura 52. Visão geral da plataforma de integração para estudo do transcriptoma de cana-de-açúcar.

Essa combinação fornece uma boa performance, estabilidade e disponibilização de um software livre. A apresentação lógica segue o padrão MVC (*Model View Controller*), com uma diferença na camada de modelo, que conecta e armazena temporariamente a informação do banco de dados. Correntemente, o MySQL é utilizado como a base de dados do projeto.

Para fornecer uma aplicação com um sistema de gerenciamento de usuários e projetos, foi desenvolvido um sistema sofisticado baseado em uma estrutura no banco de dados, *Servlets* e páginas JSP, como descrito na metodologia, para o controle de acesso de usuário, e de aplicações baseadas nos níveis de acesso dos usuários para os módulos do sistema. O gerenciamento e administração do sistema pode ser realizado através do ambiente web para todas as aplicações.

A plataforma foi desenvolvida em linguagem de programação Java, com APIs Perl e compreende os principais componentes:

- *Bancos de dados de Armazenamento*: um banco de dados relacional é criado para consistentemente armazenar todos os dados anotados, dados de níveis de expressão por diferentes abordagens e resultados intermediários de análises do transcriptoma. Este banco tem uma topologia em estrela centrada no gene.

- *Anotação Funcional*: esta etapa atribui funções biológicas, regras metabólicas ou características estruturais para os transcritos. A anotação funcional é principalmente baseada na abordagem de similaridade, através do mapeamento contra bancos de outras espécies, assumindo que regiões altamente conservadas mantêm a mesma função ou regras. Grande variedade de bancos estão disponíveis para inferir propriedades para os genes, incluindo bancos de proteínas (NCBI, UniProt, Gene Ontology), domínios de proteínas (PFAM, SMART, Prosite, DBD), e mapas metabólicos (KEGG).

- *Processamento de Microarranjos*: A análise de expressão gênica por plataformas de microarranjos possui diversas fases de processamento, como descritos anteriormente. Um sistema automatizado, com controle restrito de acesso, regras e etapas para submissão, análises intermediárias, possibilita extrair informações de diferentes condições experimentais e disponibilizar de forma automática os dados brutos e processados, através das ferramentas da plataforma que estão integradas à anotação funcional.

- *Interface web integrativa*: desenvolvimento de uma plataforma flexível com interface gráfica, que automatize todas as abordagens necessárias para uma anotação completa e estudo do transcriptoma por níveis de expressão, de forma integrada, com capacidade para agregar dados genômicos, tornando a plataforma útil para projetos

futuros. A plataforma é especializada em análises de enriquecimento de termos, nível de atividade de vias, agrupamento hierárquico, integrando microarranjos de diferentes plataformas e condições experimentais e dados de vias, anotações e categorizações funcionais.

Parte deste trabalho está relacionado ao desenvolvimento de diversas aplicações para análise de expressão gênica integrada a anotação funcional do transcriptoma de cana-de-açúcar. Com o objetivo de rastrear, gerenciar e documentar as aplicações desenvolvidas na plataforma, foi mantido um registro de todas as aplicações. As aplicações envolvem desenvolvimentos em linguagem Java, scripts Perl e R e tem sido utilizadas para uma variedade de objetivos no gerenciamento de dados, processamentos e análises. Uma lista contendo as principais aplicações é descrita na Tabela 19. As aplicações para análise de expressão gênica permitem a seleção de múltiplos experimentos e filtros para seleção de genes de interesse, categorias ou anotações funcionais por expressão regular e critérios de corte pelo nível de expressão gênica.

Tabela 19. Lista de algumas aplicações desenvolvidos para o gerenciamento de dados, processamento e análises, separados por módulos.

Aplicações	Descrição
Gerenciamento, Controle de Acesso, Submissão e Atualização	
<i>appRegister</i>	Aplicativo responsável pelo registro e atualização de usuários.
<i>appControl</i>	Aplicativo responsável pelo gerenciamento de usuário, permissões de acesso de leitura, escrita e visualização de informações.
<i>appDentry</i>	Aplicativo responsável pela entrada de dados de microarranjo e anotação.
<i>appInsertAgilentPlatform</i>	Aplicativo para submissão de lâminas de microarranjo.
<i>appUploadRawDataExperiment</i>	Aplicativo responsável pela submissão de dados extras vinculados a um experimento por microarranjo.
<i>appGenerateFormats</i>	Aplicativo para criação de arquivos formato <i>Platform</i> , <i>Series</i> e <i>Sample</i> .
<i>appUpdateSASAnnotation</i>	Aplicativo para recuperação da anotação do SAS.
<i>appPublic</i>	Aplicativo para submissão e registro de novas publicações e vinculação de material suplementar.
<i>appPublication</i>	Aplicativo de busca e acesso a publicações locais.

Ferramentas de Busca e Anotação Funcional	
<i>appSearch</i>	Aplicativo para busca de genes diferencialmente expressos, por seleção de múltiplos experimentos, por função e por expressão regular.
<i>appSearchSASAlignment</i>	Aplicativo para exibição gráfica e da estrutura da montagem do transcrito.
<i>appSearchCatalogue</i>	Aplicativo para busca de catálogos do transcriptoma.
<i>appSearchSASvsHomologs</i>	Aplicativo para busca de genes ortólogos e parálogos em gramíneas.
<i>appSearchGeneByBlast</i>	Aplicativo para busca de resultados de alinhamentos com método uni-direcional.
<i>appSearchSASCode</i>	Aplicativo para busca de descrição dos SAS.
<i>appSearchSASseqs</i>	Aplicativo para busca de sequências fasta.
<i>appSearchPathWay</i>	Aplicativo para consulta de gene presentes em vias metabólicas.
Análises de Expressão gênica	
<i>appMicroarrayDesignInfo</i>	Aplicativo para consulta da composição da lâmina de microarranjo.
<i>appSummary</i>	Aplicativo para geração de estatísticas de experimentos de microarranjo depositados e analisados.
<i>appSearchExperiment</i>	Aplicativo para busca de experimentos de microarranjo e visualização de descrição.
<i>appCrossingExperiment</i>	Aplicativo para cruzamento entre duas condições experimentais.
<i>appCrossExperiment</i>	Aplicativo para cruzamento de múltiplas condições experimentais.
<i>appSearchSignificantlyExpressedGenes</i>	Aplicativo para identificação de genes significativamente expressos em múltiplas condições.
<i>appSearchGene</i>	Aplicativo para seleção de genes diferencialmente expressos.
<i>appTissueMatrix</i>	Aplicativo para análise matriz de expressão por tecidos.
<i>appSamplesByCorrelation</i>	Aplicativo para análise de correlação entre diferentes condições experimentais.
<i>appMicroarrayDesignInfo</i>	Aplicativo para consulta dos elementos da lâmina de microarranjo.
<i>appSamplesForClust</i>	Aplicativo para seleção de múltiplos experimentos e saída em formato para análises por agrupamento.

<i>appVennDiagram</i>	Aplicativo para criação de múltiplos grupos, selecionando N experimentos e geração de resultado em diagrama de Venn.
<i>appGeneMergeTermEnrichment</i>	Aplicativo para análise de enriquecimento de termos, para uma lista de genes de interesse.
<i>appSearchPotentiallyPathways</i>	Aplicativo para Comparação entre duas ou mais condições experimentais para cálculo do escore de atividade das vias metabólicas.
Análise de fisiologia, Transgenia e Validação gênica	
<i>appSearchPhysiology</i>	Aplicativo para consulta de experimentos para obter resultados de medidas fisiológicas e bioquímicas.
<i>appPhysiology</i>	Aplicativo para submissão, atualização e deleção de experimentos de fisiologia.
<i>appSearchRTPCR</i>	Aplicativo para consulta de análises e resultados de genes para validação de nível de expressão gênica por qRT-PCR
<i>appsearchTrasngenicPlant</i>	Aplicativo para consulta de análises e resultados de plantas transgênica baseado em estudos de superexpressão e silenciamento de genes.
Análise genômica	
<i>appUpdateContigs</i>	Aplicativo para atualização de <i>contigs</i> genômicos.
<i>appSearchMotifs</i>	Aplicativo para busca de motivos em regiões promotoras.
<i>appSearchSynteny</i>	Aplicativo para análise sintenia entre regiões genômicas.

A interface de consulta inicial é construída baseada em 10 principais sessões para representação e gerenciamento da informação relacionada à cana-de-açúcar (Figura 53), e é composta das seguintes sessões:

- *Home*, página de busca geral no banco de dados por expressão regular. Uma consulta SQL por expressão regular é construída, realizando uma busca em entidades específicas relacionadas aos transcritos, experimentos de microarranjo, anotações funcionais e publicações de artigos científicos.

- *Cane Genome*, a seção apresenta links para acesso direto a resultados do genoma, como os softwares GBrowse e GBrowseSyn, e ferramenta de consulta a partir do nome do transcrito ou do contig genômico.

- *Cane Transcriptome*, a seção possui ferramentas desenvolvidas para consulta dos resultados de anotações funcionais, análises de ortologias, visualização da

montagem dos transcritos, além de buscas por expressão regular para o transcriptoma armazenado no banco de dados.

- *Cane Gene Expression*, a seção apresenta todas as ferramentas desenvolvidas para prospecção e análise de experimentos por expressão gênica por microarranjo.

- *Cane Physiology*, a seção apresenta as ferramentas de inserção, atualização e deleção de medidas de experimentos de fisiologia e ferramenta de consulta e cruzamento entre diferentes condições experimentais.

- *Cane Functional Annotation*, a seção possui ferramentas para atualização da anotação e categorização gênica, busca por similaridade de genes com diferentes fontes de dados, busca por catálogos, famílias de TFs e categorizações funcionais, e ferramentas para busca de genes e desenho de vias metabólicas.

- *Cane Transgenic Plant*, esta seção possibilita a busca de experimentos de plantas transgênicas e relacionamento com os genes utilizados e presentes no banco de dados.

- *Generic Tools*, corresponde à inclusão de ferramentas externas adaptadas para análises de dados restritos, internos ao grupo, como BLAST.

- *Publications*, a seção possui ferramenta de inserção, deleção e consulta de trabalhos publicados pelo grupo e material suplementar presente no trabalho, como dados de expressão, análises de agrupamentos e figuras de validações. A seção também possui um aplicativo para geração de arquivos para microarranjo em formato GEO para submissão ao GEO-NCBI.

- *Admin*, esta seção é responsável pelo gerenciamento e controle de acesso dos usuários cadastrados no sistema e aplicações para o cadastro, submissão e inicialização de análises de novos experimentos de microarranjo, assim como submissão de novas plataformas de microarranjo.

Figura 53. Consulta e Visualização para transcritos através da interface CaneRegNet.

A plataforma é baseada em diversos módulos como descrito na metodologia, todos integrados e interligados pela entidade *probe_set* e abaixo apresentamos os principais resumidamente com suas funcionalidades.

- Módulo *Transcriptome*

- *RegExp Transcript Search*
 - Busca do transcrito, oligonucleotídeo, EST ou anotação, pelo nome ou parte do nome. A ferramenta retorna uma lista de transcritos com um link com descrição detalhada, contendo todas as informações armazenadas no *CaneRegNetDB* com o transcrito buscado.
- *Search Gene vs Homologs*
 - Busca pelo nome(s) do(s) transcrito(s) ou selecionando apenas o grupo de espécies, retornará a lista de transcritos e genes ortólogos e parálogos entre as espécies.
- *Search SAS/EST/Oligo seqs*
 - Busca pelo nome do(s) transcrito(s), EST ou oligonucleotídeo, possibilitando filtrar por *full-length*, para recuperar o fasta de nucleotídeo e proteína.
- *Search Genes by Blast*
 - Busca pelo nome do(s) transcrito(s) ou anotação e seleção do(s) banco(s) de mapeamento(s) para recuperação da anotação do(s) transcrito(s).
- *Search SAS Alignment Viewer*
 - Busca do transcrito e Visualização da montagem com ESTs.

- Módulo *CaneGeneExpression*

- *Crossing replicates*
 - Esta ferramenta permite selecionar individualmente duas amostras hibridadas e analisadas na plataforma *CaneRegNet*. A saída apresenta o nome do gene, categorias funcionais e porcentagem de induzidos, reprimidos e não diferenciados comparativamente em cada amostra.
- *Crossing experiment (raw ratio)*
 - Ferramenta para recuperação dos genes DE pelo método HTSelf adaptado, analisados pelo cruzamento com a(s) replica(s) biológica(s). Inicialmente é(são)

selecionado(s) os experimento(s) de interesse. A saída exibe o perfil de expressão média entre o experimento e a(s) réplica(s).

- *Crossing experiments by regular expression*
 - o Ferramenta para recuperar lista de genes DE, permitindo filtro com expressão regular por, nome do transcrito, anotação ou categoria *functional*, porcentagem de transcritos replicados na lâmina com mesmo perfil de expressão, selecionar por “induzidos e reprimidos” ou “não diferenciados”, exibição de valor de razão de expressão para cada réplica ou média entre elas, para retornar a lista de transcritos com perfil de expressão em cada condição experimental na forma de tabela comparando todos os experimentos selecionados, com informações as categorias e anotações.
- *Cross-group experiments by Venn diagram*
 - o Ferramenta desenvolvida com o princípio de modelo de diagramas de Venn, mais corretamente método de Euler. Permite definir o número de grupos, e cada grupo pode incorporar um grande número de experimentos. A ferramenta utiliza a informação dos genes DE, filtrando pela porcentagem de transcritos dentro da lâmina e retornando uma tabela com grupos e número de transcritos em cada grupo (induzido, reprimido, não diferencial), permitindo exibir a lista dos transcritos, com suas anotações e perfil de expressão para cada grupo dinamicamente.
- *Crossing experiments by correlation*
 - o Ferramenta para análise de transcritos com Co-Expressão entre diferentes condições experimentais, baseado no método de correlação de Pearson para cálculo da significância do nível de correlação entre três ou mais experimentos, com implementação de uma normalização adicional por *mean centering*, *median centering* ou *Z-score* para eliminar possíveis variações entre hibridações e realizando um cruzamento do resultado com o perfil de expressão dos transcritos em cada experimento, permitindo avaliar os mais correlacionados e selecionados como DE pelo HTSelf adaptado.
- *Search significant expressed genes*
 - o Método desenvolvido para análise da expressão gênica pelo nível de intensidade por amostra hibridada, um gene é expresso se seu nível de intensidade está significativamente acima do sinal de fundo, cruzando com o resultado da(s) réplica(s) diminuindo a taxa de falsos positivos.
- *Term enrichment analysis*

- Ferramenta para analisar o enriquecimento de termos, para uma lista de genes de interesse (DE, SE, etc), através da integração à plataforma da ferramenta *GeneMerge* adaptada ao nosso transcriptoma. A ferramenta permite a busca por enriquecimento nos bancos: Categoria Funcional local, Gene Ontology (base única e base dividida por ontologia: Função Molecular, Processos Biológicos e Componentes Celulares), Vias Metabólicas do KEGG, Enzimas do KEGG.
- *Microarray Design Info*
 - Busca por informações relacionadas ao desenho da plataforma de microarranjo.
- *Search experiment description.*
 - Busca por informações para os experimentos hibridados.
- *Search Potentially Activated Pathways/Terms*
 - Ferramenta para comparar atividade de vias metabólicas a partir de expressão gênica para identificação de diferenças entre condições experimentais. A ferramenta utiliza os genes SE, com os níveis de expressão de cada gene, para cada experimento. Após o cálculo de atividade de cada via, se forem selecionados dois experimentos será realizada uma análise de teste não paramétrico e não pareado de Wilcoxon Rank-Sum, e se forem selecionados mais de dois experimentos um agrupamento hierárquico por Pearson, mas com possibilidade de escolha de outros métodos e análise de correlação entre as amostras para comparar a atividade das vias entre as diferentes condições e agrupá-las em grupos de padrões.

- Módulo *Functional Annotation*

- Update Transcript Annotation
 - Ferramenta para atualização da anotação dos transcritos de cana-de-açúcar.
- Update Transcript Annotation
 - Inserir ou atualizar automaticamente anotação de múltiplos transcritos por arquivo (formato por tabulação).
- Search SAS Annotation
 - Ferramenta para busca de anotação e categorias funcionais para um ou mais transcritos.
- Search by Catalogue
 - Ferramenta para busca de transcritos pelo nome ou por catálogo funcionais.
- Search Metabolic PathWay

- Ferramenta para busca de Vias metabólicas e informações para um ou mais transcritos.

- Módulo *Administrator*

- *Access Control*
 - Ferramenta para gerenciar o acesso aos módulos e dados.
- *Data Entry*
 - Ferramenta para criar, atualizar, deletar e realizar análises de dados de microarranjo.
- *Upload Raw-Data Experiments*
 - Ferramenta para importar arquivos grandes relacionados a dados de experimentos de microarranjo.
- *Insert Agilent Platform tool*
 - Ferramenta para importar novas plataformas Agilent.

6.6.1.Principais Interfaces de Consulta e Visualização

A seguir apresentamos a página de consulta para um transcrito de interesse. A página é construída de acordo com a integração das várias camadas da arquitetura e acesso ao banco de dados como descrito na metodologia, inicialmente é apresentada a página de “Summary” contendo informações básicas sobre o transcrito, com abas adicionais para cada tipo de análise ou informação vinculada ao transcrito no *CaneRegNetDB* (Figura 54).

Milton yutaka Nishiyama Junior!

Home

Logout

RegExp Search

SCC

transcript

Search

CST - Chip Signal Transduction | **CME** - Chip Metabolism | **CRN** - Chip Cane Regnet | **qRT** - qRT-PCR | **TrP** - Transgenic Plant

Results: 1

name	probe_name	probe_purpose	probe_type	detail
SCCCI	SCC(	CME,CRN	transcript	Details

Summary of (SCCCRZ1004F05.g)

Genome

Transcriptome

Proteome

Gene Network

Phenotype

Pathway

Name SCCCI
Length 1385
Organism -
SoGI Gene Index TC11

☐ Homologs

Homolog Groups

- ☐ Group 905: Saccharum s. vs Arabidopsis
- ☐ Group 2052: Saccharum s. vs Oryza s.
- ☐ Group 2700: Saccharum s. vs Zea m.
- ☐ Group 2928: Saccharum s. vs Sorghum
- ☐ Group 4193: Saccharum s. vs Sorghum

☐ Alleles

☐ ESTs

☐ Organism

☐ Pathways

☐ Synonyms

☐ Sequence

☒ Proteins

☐ Publication

Summary of (SCC

Genome

Transcriptome

Proteome

Gene Network

Phenotype

Pathway

☒ Proteins

☐ Domains

Domains

☐ details

☐ Homologs

☐ PathWay

PathWay

- ☐ Amino Acid Metabolism
- ☐ Carbohydrate Metabolism
- ☐ Metabolism

Figura 54. Busca e Visualização para transcritos através da interface do *CaneRegNet*. A página inicial permite a busca por expressão regular, do nome do transcritos, parte dele, por anotações, ESTs ou oligonucleotídeos, enviando as consultas ao bancos de dados à camada de controle. A página exibida é estruturada como descrito na metodologia para apresentar os resultados de cada módulo vinculado ao transcrito e ao clicar em uma aba diferente as informações são exibidas dinamicamente.

A interface de consulta de Vias Metabólicas permite recuperar, para um ou mais transcritos as Vias Metabólicas, Enzimas e Reações com as quais está relacionado pela análise com o KAAS, exibindo dinamicamente via REST a figura da Via Metabólica e a marcação da(s) enzima(s) presente(s) (Figura 55).

Figura 55. Visualização da busca para uma lista de transcritos e página de resultado com as informações presentes no *CaneRegNetDB* e a apresentação da figura da Via Metabólica com os transcritos representados graficamente, através das enzimas.

Uma das abas mais importantes "Transcriptome" sumariza toda a informação relacionada ao transcrito, recuperando a informação dos níveis de expressão em todos os experimentos depositados no *CaneRegNetDB*, acessíveis ao usuário. São exibidos gráficos e respectivas tabelas com experimentos e níveis de expressão (Figura 56).

- ☐ Clones (ESTs)
- ☐ Probe Sets
- ☐ qPCR
- ☐ SOM
- ☐ Literature
- ☐ Blast Alignments
- ☐ Northern Digital
- ☐ Assembly Alignment
- ☐ Gene Ontology

Categorization

Category

function_category	Signal Transduction
structural_category	transcription factor
category	Transcription Factor
sub_category1	AP2/EREBP
sub_category2	unknown [Zea mays]
sub_category3	ethylene-responsive transcription f

Probe Intensity

Probe Intensity

CaneRegNet

Probe Intensity values - Max value 60000

Probe Statistics

Probe Statistics

CaneRegNet

Probe Statistics - values log²

Probe Intensity

Probe Statistics

Probe Expression

Probe Expression

CaneRegNet

Probe Expression - values log²

Figura 56. Página de Visualização para todas as informações para o transcrito consultado. Os subitens podem ser abertos para informações detalhadas, como níveis de expressão em todos os experimentos de forma gráfica e tabela, dos três níveis de análise de expressão: *probe_intensity* com níveis de expressão normalizados, *probe_statistics* com níveis de razão (tratado/controle) de expressão, e *probe_expression* contendo transcritos DE presentes em todas as réplicas, e exibidos de acordo com o nível de acesso do usuário.

6.6.2.Interface para Submissão e Processamento de lâminas de Microarranjo

Um dos objetivos iniciais do desenvolvimento do *CaneRegNet* e *CaneRegNetDB* foi o gerenciamento, armazenamento e processamento de experimentos por microarranjo e ferramentas de consulta e cruzamentos entre as diversas condições experimentais, permitindo que diversas instituições e grupos de pesquisa criem seus próprios projetos e experimentos, com gerenciamento e recuperação da informação restritos. As análises de dados armazenados poderiam ser facilitados pela incorporação na plataforma de aplicações externas e aplicações desenvolvidas localmente. Deste modo, foi elaborado um fluxo de trabalho para automatizar todo o processamento e análise inicial de experimentos de microarranjo (Figura 57), com parâmetros pré-definido e padronizados para reproduções e com um histórico completo do experimento. Controles de qualidade tem sido introduzidos ao longo do fluxograma em posições críticas para garantir a qualidade do dado.

Um processo similar baseado neste mesmo fluxograma corresponde à criação de experimentos híbridos, permitindo selecionar amostras hibridadas em lâminas distintas para criação de um novo experimento híbrido.

As etapas iniciais não descritas no fluxograma correspondem a:

- Criação de uma plataforma de microarranjo, utilizando as entidades do *CaneRegNetDB*: *array*, *array_design*, *hybridization* e *hybridization_protocol*.
- Criação de um projeto, o qual concentrará o acesso de todas as amostras, experimentos e análises por expressão gênica, entidades: *project*.
- Criação das amostras, as quais serão utilizadas pelo fluxograma principal para criação dos experimentos e são vinculadas ao projeto específico, entidades: *sample* e *sample_provider*.

As etapas para submissão e análise de um experimento de microarranjo são (Figura 57):

- 1- O usuário pode selecionar entre Inserir, Atualizar, Duplicar e Apagar um experimento e aparecerão ao longo do processo apenas os projetos e/ou experimentos ao qual o usuário possui acesso.
- 2- Ao selecionar o projeto aparecerá um formulário para novo experimento (ou lista de experimentos para atualização, aos quais o usuário possui acesso).
- 3- Preenchimento do formulário do experimento, como seleção da plataforma de microarranjo, as duas amostras hibridadas, e opcionalmente o arquivo do *Feature Extraction*. O sistema faz validação de todos os campos obrigatórios por *JavaScript*, com uma confirmação final antes da submissão.
- 4- Após a submissão inicial, se for selecionado duplicar um experimento, serão duplicadas as informações do experimento, modificando apenas parâmetros estatísticos para a re-análise, necessário para utilização de outras abordagens de análise. Se for inserido um novo experimento, ou o arquivo for inserido posteriormente à criação do experimento, será realizado o controle de qualidade do arquivo de entrada e se não atender aos padrões esperados, o experimento ficará com *status* de inválido e o usuário alertado com os problemas identificados. Caso o arquivo não seja inserido o experimento finalizará nesta etapa.
- 5- Após a validação do arquivo é iniciada a etapa de análise através de scripts Perl e R, com o método HTSelf adaptado, com verificação e validação ao final da análise e submissão dos resultados no *CaneRegNetDB*, inserindo os resultados nas entidades: *probe_intensity* e *probe_statistics*. No *probe_statistics*, são armazenadas principalmente a mediana da razão de expressão e quantos transcritos replicados na lâmina foram identificados como induzidos, reprimidos e não diferenciados.
- 6- Após a validação de inserção dos dados no *CaneRegNetDB* é enviado um *email* ao usuário de finalização com sucesso e disponibilização do experimentos nas ferramentas de análise por expressão gênica.

Figura 57. Fluxograma de Submissão de um experimento de microarranjo. E criação de experimentos híbridos.

6.6.3.Interface para Cruzamento entre réplicas de experimento por microarranjo

Uma etapa importante após a análise para identificação dos genes DE é o cruzamento entre réplicas biológicas/técnicas. Esta ferramenta tem o objetivo de permitir ao usuário selecionar múltiplos experimentos, correspondendo a todas as réplicas para experimento analisado e extrair a lista de genes DE em comum entre todas as réplicas. A lâmina de microarranjo pode possuir múltiplos genes replicados e na análise da razão de expressão do experimento são definidos o perfil de cada réplica do gene em uma mesma amostra, determinando quantos são induzidos (*up*), reprimidos (*down*) e não diferenciais (*inside*).

As etapas para criação de cruzamentos entre experimentos de microarranjo são (Figura 58):

1 - O usuário pode selecionar entre Inserir, Atualizar, e Apagar um cruzamento e aparecerão ao longo do processo apenas os projetos e/ou cruzamentos ao qual o usuário possui acesso.

2 - Selecionando um cruzamento será aberto o formulário e os experimentos disponíveis ao usuário para preenchimento das informações relacionadas ao cruzamento, e seleção dos experimentos desejados

3 - Após o preenchimento e seleção dos experimentos é realizada a validação e as informações são salvas no *CaneRegNetDB*.

4 - Automaticamente é inicializado o cruzamento entre as réplicas extraíndo as informações de cada experimento a partir da entidade *probe_statistics* e na análise entre as réplicas experimentais são calculados os totais de *tup*, *tdown*, *tinside* para cada gene, e as regras a seguir determinam o perfil de expressão do gene e a razão de expressão média entre as diferentes condições:

Se $tup \geq (tdown + tinside)/2$ Então é definido como induzido

Senão Se $tdown \geq (tup + tinside)/2$ Então é definido como reprimido

Senão é definido como não diferenciado

Os valores de expressão médio e porcentagem para cada gene como induzidos, reprimidos e não diferenciados são armazenados no banco de dados, enviado email para o usuário e disponibilização dos resultados para análises nas ferramentas da plataforma, caso contrário será enviado email ao usuário de falha da análise e problema identificados.

Figura 58. Fluxograma do processo de cruzamento entre múltiplos experimentos a partir da razão de expressão obtida pelo método HTSelf adaptado.

6.7. Abordagens para mineração de dados de expressão gênica

6.7.1. Análises de genes significativamente expressos (SE) e diferencialmente (DE)

Um dos objetivos deste trabalho foi transformar análises complexas, realizadas manualmente, dentro de uma plataforma automatizada que permita a realização de

análises de dados de expressão gênica e esteja integrada às anotações funcionais e informações armazenadas no *CaneRegnetDB*. Esses desenvolvimentos permitem a compreensão de quais processos biológicos são significativamente modulados sob uma determinada condição ou tratamento, sendo muito mais informativo do que a identificação de listas de genes significativamente e diferencialmente expressos em determinadas condições experimentais. Para a compreensão desses processos de forma automatizada e acessível ao usuário, desenvolvemos e implementamos diversas abordagens para a integração e interpretação destes dados, as principais são as análises de enriquecimento de termos ou categorias e análise do escore de atividade das vias metabólicas, as quais podem utilizar como entrada resultados gerados a partir das análises para identificação de genes SE e DE e recuperados diretamente pela classe *Utils*.

A estratégia desenvolvida para estes aplicativos está apresentada na Figura 28 da metodologia. Os aplicativos de análise de genes SE e DE permitem a seleção de experimentos de microarranjo, de acordo com o nível de acesso do usuário (Figuras 59A e 60A), definindo parâmetros e filtros de acordo com o aplicativo selecionado e exibindo uma tabela comparativa (Figuras 59B e 60B) com informações dos genes, anotações e níveis de expressão em cada experimento selecionado. Ao final das análises, os resultados são exibidos na mesma página de consulta exibindo diferentes resultados de acordo com a abordagem escolhida.

Os aplicativos para análise de genes SE e DE correspondem às análises iniciais para estudos de expressão gênica, permitindo a identificação de número de genes, níveis de expressão, total de genes induzidos e reprimidos entre diferentes condições experimentais e identificação de funções de interesse, muito similar a abordagens existentes para análises de dados por expressão gênica, porém com um adicional por apresentar um anotação curada, geração de estatísticas diretas sobre a lista final de diagramas de Venn para cada análise ao final do relatório exibido, e lista dos genes mais presentes entre todas as condições experimentais.

Estes aplicativos e análises de genes SE e DE tem sido utilizadas em estudos para comparação de séries temporais em condições de estresse hídrico (Lembke et. al., 2012), tendo apresentado uma taxa de 75% dos oligonucleotídeos identificados como SE e 11.9% para os antisense em relação à lâmina (Tabela 20A), e separado para cada condição experimental (Tabela 20B), foram selecionados alguns candidatos para validação por qRT-PCR, com validação de 85%. Outros estudos também utilizaram este aplicativo para ritmo circadiano na cana-de-açúcar (Lembke et. al., 2012; Hotta et. al.,

2013), apresentando um nível de validação de 88% dos genes significativamente expressos com ritmo na cana-de-açúcar.

Tabela 20. Número de oligonucleotídeos senso e antisenso identificados como SE nos experimentos por estresse hídrico, com total em relação à lâmina (A) e por condição experimental (B).

(A) Oligonucleotídeos SE expressos			
	Total na lâmina	SE	%
Senso	14522	10904	75
Antisenso	7380	876	11.9
(B) Oligonucleotídeos SE na série temporal			
24h controle	10030	609	
24h tratado	10110	609	
72h controle	9745	470	
72h tratado	9689	503	
120h controle	8814	286	
120h tratado	7611	833	

(A)

Search Significant Expressed Genes

Description: The goal of this tool is to provide the user a list of significant expressed genes, by selecting experiments or crossing. The output is shown in a table format containing the gene name, annotation data and the selected experiments values.

Select one or more experiments:

A1V2C.Pool - pool of V2 irrigated for 24h, 72h, 120h X pool of V2 irrigated for 24h, 72h, 120h
A1V2E6_0.80 - Drought 120h variety SP90-1638 (susceptible) p=0.90 cutoff Self-bars
A1V2E5_sw - Drought 72h variety SP90-1638 (susceptible) p=0.90 cutoff Self-bars
A1V2E5 - Drought 72h variety SP90-1638 (susceptible) p=0.90 cutoff Self-bars
A1V2C.Pool_sw - pool of V2 irrigated for 24h, 72h, 120h X pool of V2 irrigated for 24h, 72h, 120h - biological replicate
A1V2E6_sw_0.80 - Drought 120h variety SP90-1638 (susceptible) p=0.90 cutoff Self-bars
A1V2E6_sw - Drought 120h variety SP90-1638 (susceptible) p=0.90 cutoff Self-bars
A1V2E6 - Drought 120h variety SP90-1638 (susceptible) p=0.90 cutoff Self-bars
pKanPP2C.pool18.I.L1 - Cy3 treatment, Cy5 control
PP2Csil.18A-04.I.vs.18A-01.I_sw - Comparison between pKanPP2C.18A-04.I.L3 and pKanPP2C.18A-01.I.L3 and dye swap from PP2Csil.18A-04.I.vs.18A-01.I
PP2Csil.11-29.I.L1_swt - null
PP2Csil.4B-09.D.vs.4B-09.I - Comparison between pKan.4B-09.D.L1 and pKan.4B-09.I.L3
PP2Csil.18A-04.D.vs.4B-09.D - Comparison between pKanPP2C.18A-04.D.L1 and pKan.4B-09.D.L1
pKanPP2C.11-29.I.L1 -
PP2Csil.18A-04.I.vs.6AC-3A.I_sw - Comparison between pKanPP2C.18A-04.I.L3 and pKan.6AC-3A.I.L3. Biological replicate and dye swap from PP2Csil.18A-04.I.vs.6AC-3A.I.L3
PP2Csil.11-29.I.L1 - null
PP2Csil.18A-04.D.vs.18A-01.I_sw - Comparison between pKanPP2C.18A-04.D.L1 and pKanPP2C.18A-01.I.L3
A1V2E3_sw - Drought 24h variety SP90-1638 (susceptible) p=0.90 cutoff bars of Pool
pKanPP2C.pool11.I.L1 - Cy3 control, Cy5 treatment
PP2Csil.18A-04.I.vs.18A-01.I - Comparison between pKanPP2C.18A-04.I.L3 and pKanPP2C.18A-01.I.L3

Search by:

Enter a subject annotation (Regular Expression):

Download output file
[Results in file](#)
[Results in file empty expression replaced by zero](#)

Type of result:

Filter by expression:

(B)

Significant Expressed Genes - Analysis Result

Index	transcript_name	probe_name	function_category	structural_category	category	sub_category1	A1V2E5	A1V2E5_sw	A1V2E6	A1V2E6_sw
							log2 (cy5/cy3)	log2 (cy5/cy3)	log2 (cy5/cy3)	log2 (cy5/cy3)
1	SCSF	GS1_SS_	Signal Transduction	kinase/phosphatase	Receptors	receptor Ser/Thr kinase	-0.7652	1.2558	-0.0844	-1.5348
2	SCVP	GS1_SS_	Signal Transduction	kinase/phosphatase		Purple acid phosphatase; Zinc(II) purple acid phosphatase;	1.2455	0.6582	2.4498	1.5709
3	SCEQ	GS1_SS_	Signal Transduction		Transcription Factor	AP2/EREBP	0.2854	1.3672	0.9293	0.5300
4	SCSF	GS1_SS_	Signal Transduction	kinase	Receptors	Receptor Ser/Thr kinase	-2.3625		-0.8088	
5	ISCCO	GS1 AS	Signal Transduction	transcription factor		NAC-domain containing protein	0.2709			

Figura 59. Apresentação da interface de entrada (A) e saída (B) do aplicativo para seleção de genes SE.

(A)

Search Genes in DB

Select one or more experiments:

A1V2E3_0.80 - Drought 24h variety SP90-1638 (susceptible) p=0.80 cutoff bars of Pool
A1V2E5_0.80 - Drought 72h variety SP90-1638 (susceptible) p=0.80 cutoff Self-bars
A1V2E6_0.80 - Drought 120h variety SP90-1638 (susceptible) p=0.80 cutoff Self-bars

Search by: RegExp annotation
Enter a subject annotation (Regular Expression):

Select type of expression: up/down
Show ratio information: mean log2ratio
Percent:

transcript_name	probe_name	function_category	structural_category	category	sub_category1	sub_category2	A1V2E3_0.80	A1V2E5_0.80	A1V2E6_0.80
SC	GS1_SS_10653_13533	Cell wall metabolism	Lignin metabolism	Cell wall	dirigent-like protein		-1.3788		
SCI	GS1_SS_25899_16947	Cell wall metabolism	Lignin metabolism	Cell wall	dirigent-like protein			3.2119	4.2530
SC	GS1_SS_16580_15645	Cell wall metabolism	Lignin metabolism	Cell wall	dirigent-like protein			-2.6191	-3.5735

(B)

Experiments for exclusive Groups:

experiment	total	up	down	inside
A1V2E3_0.80	1	0	1	0
A1V2E5_0.80 + A1V2E6_0.80	2	2	2	0

Figura 60. Apresentação da interface da entrada (A) e saída (B) do aplicativo para seleção de genes DE.

6.7.2. Análise de Enriquecimento de Termos

Esta abordagem busca determinar se um grupo de genes expressos por microarranjo em um determinado tratamento exibe um conjunto de funções conhecidas e relacionadas e quais são mais representativas. Este método de classificação de grupos gênicos fornece indícios de quais grupos são responsivos ou não a determinadas condições biológicas ou tratamentos.

Diversas abordagens tem sido desenvolvidas para análises de enriquecimento, porém a maior parte são aplicativos para serem executados via web, não permitem a criação de bancos de dados particulares ou rodam em aplicativos que não permitem

serem integrados a plataformas web (GenMAPP, Dahquist et. al., 2002; GOSTat, Beißbarth etl al., 2004; agriGO, Du et. al., 2010, entre outros), por suprir essas deficiências encontradas nos aplicativos anteriores e por utilizar um método estatístico apropriado para a análise de enriquecimento de termos, utilizando distribuições hipergeométricas e correção de escore por Bonferroni, decidimos pela utilização do GenMerge (Castillo D et. al., 2003). Deste modo, aplicativo desenvolvido e integrado à plataforma utiliza o algoritmo do GeneMerge para análise de enriquecimento de termos e utiliza como termos, categorias funcionais de cana-de-açúcar, vias metabólicas, anotação de enzimas e ontologias do *Gene Ontology*. A etapa inicial da análise corresponde ao upload de uma lista de genes, contendo ou não uma coluna adicional informando os perfis de expressão respectivos (*up*, *down*, *inside*), parâmetros adicionais para exibição do número mínimo de representantes em cada categoria e filtro por categorias *unknown* e *no match* e o banco de anotações desejado (Figura 61).

Esse aplicativo tem sido empregado em diversos estudos, como na análise de série temporal para experimentos em 24h, 72h, e 120h de estresse por seca (Lembke et. al., 2012), aplicado a bancos de Ontologias por GO e ao banco de vias metabólicas (Figura 62), e ao banco de categorias funcionais de cana-de-açúcar, tendo sido observado um enriquecimento de categorias funcionais distintas em cada ponto e identificada uma progressão de eventos, através de funções relacionadas ao estresse hídrico, durante a evolução do estresse (Lembke et. al, 2012).

Term Enrichment Analysis	
Choose the Database for the Enrichment Term Analysis:*	CaneRegNet Microarray - KEGG Pathways
Upload the list of genes:*	<p>The file must have 1 or 2 columns:</p> <p>gene example of data: SCCCAD1001A01.g SCCCAD1004A01.c SCCCAD1005781.g</p> <p>or</p> <p>gene up / down example of data: SCCCAD1001A01.g up SCCCAD1004A01.c down SCCCAD1005781.g down</p> <p>Browse... resultA1V2E6.txt</p>
Show only the more representative data, above:	1
Columns to show in image:	Pop_fraction & Study_fraction
Filter Categories:	☐ Unknown ☐ No Match
Submit	

Figura 61. Interface para análise de enriquecimento de termos, recebe como entrada um arquivo com uma lista de genes para a análise, seleção do banco de dados para o estudo de enriquecimento, número de representantes por categoria a ser exibida e filtros adicionais.

Figura 62. Interface de saída do aplicativo para análise de enriquecimento de termos. Apresentando o gráfico do escore de enriquecimento de cada categoria (em azul), o gráfico da representação populacional e do conjunto de estudo e a tabela com os escores e genes de cada termo.

A análise de enriquecimento de termos pode ser empregada nas diferentes categorizações criadas neste projeto e pelo grupo do laboratório. Deste modo, baseado nos experimentos de série temporal em condições de estresse, as análises com os genes DE enriquecidos em famílias de TFs permitiu identificar TFs que podem estar respondendo ao estresse por seca em pontos específicos nas condições de estresse ao longo do tempo (Figura 63). Estudo recente com a gramínea *Boehmeria nivea* L. Gaud (Liu et. al., 2013) identificou por microarranjo e sequenciamento de larga escala e

validação por qRT-PCR alguns TFs (NAC, HD-Zip, bHLH, DOF, C2H2L, MYB, AP2, ARF, GRAS e WRKY) responsivos ao estresse por seca. Em nosso estudo de caso baseado em pontos experimentais de estresse por seca (24h, 72h e 120h), identificamos todos os TFs também enriquecidos em nossa análise, como bHLH, AP2, NAC, bZIP, MYB e WRKY aparecendo no primeiro ponto de 24h, os demais aparecendo no ponto de 72h, à exceção de GRASS e ARF que aparecem apenas no ponto de 120h.

Estudos com estas mesmas condições experimentais foram conduzidos utilizando as categorias funcionais desenvolvidas pelo grupo e foram evidenciadas relações entre os tempos de estresse e as categorias enriquecidas em cada ponto, indicando o enriquecimento de categorias relacionadas ao estresse (Lembke et. al., 2012). Estudos de ritmo circadiano em cana-de-açúcar por séries temporais também tem empregado a mesma ferramenta e banco de categorias e também identificado relação entre as categorias mais enriquecidas com os pontos ao longo do tempo (Hotta et. al., 2013), evidenciado a importância dos resultados obtidos pela abordagem de enriquecimento de termos com o GeneMerge.

Figura 63. Comparação do escore de enriquecimento das famílias de TFs entre os três pontos experimentais para estresse por seca (24h, 72h e 120h) sem irrigação.

6.7.3. Análise de Escore de Atividade das Vias

Baseado nas abordagens FCS, desenvolvemos e implementamos um método integrado à plataforma *CaneRegNet*, para cálculo do escore de atividade das vias metabólicas para cada condição experimental, utilizando os níveis de expressão dos genes (Nishiyama-Jr et. al., 2014). O método é dividido em duas análises: uma análise comparativa com teste de ranqueamento de Wilcoxon da atividade das vias entre duas condições experimentais, e outra análise para comparação entre múltiplas condições

por agrupamento hierárquico das vias metabólicas e correlação de Spearman entre as condições experimentais baseado no escore das vias.

Há duas alternativas para as análises de escore de atividade das vias, pela seleção do projeto e em seguida os experimentos disponíveis no *CaneRegnetDB*, de acordo com o nível de acesso do usuário, ou pelo *upload* de um arquivo em formato tabela, com a primeira coluna contendo o nome dos genes e as colunas seguintes, os níveis de expressão dos respectivos genes em cada condição experimental; são disponibilizados parâmetros e filtros similares aos aplicativos para análises de SE e DE, além da escolha da versão da via metabólica com os genes identificados nas vias metabólicas (Figura 64). Os níveis de expressão dos genes em cada experimento selecionado podem ser originados a partir da análise de seleção de genes DE ou pela análise de genes SE, como discutido anteriormente e levantamento com experimentos depositados no *CaneRegNetDB* em média >65% (9.400) genes apresentam expressão significativa (SE), enquanto em média menos de 8% (1.200) são identificados como DE, fazendo grande diferença neste tipo de abordagem devido a representatividade de genes em cada via metabólica e, principalmente, para o cálculo do nível de atividade ou abundância média de expressão para cada via metabólica.

Search Potentially Activated Pathways/Terms

Description: The goal of this tool is to provide the user a list of significant expressed genes, by selecting experiments or crossing. The output is shown in a table format containing the gene name, annotation data and the selected experiments with **log2 values**.

Select one or more projects:

CaneRegNet COMT
CaneRegNet Ethylene
CaneRegNet Polysaccharide synthesis in culm
CaneRegNet1 Ancestral
CaneRegNet1 Brix
CaneRegNet1 Circadian

Expressed Genes by: Crossing

(A)

Search

Search Potentially Activated Pathways/Terms

Description: The goal of this tool is to provide the user a list of significant expressed genes, by selecting experiments or crossing. The output is shown in a table format containing the gene name, annotation data and the selected experiments values.

Select one or more experiments:

(B)

A1CR2C1L1_0.96 - HB vs LB - Leaf +1 - 7 months (Progenies from the cross 2, SP70-1143 x RB925211) p=0.96
A1CR2C2PhotoL1_0.90 - HPvsLP-Leaf+1-7 months (Genotypes Cross II: 153vs238)
A1CR2C2PhotoL1_0.90 - HPvsLP-Leaf+1-7 months (Genotypes Cross II: 153vs238)
A1CR3C1F1R1_BBxAB_0.96 - HB vs LB - Leaf+1 - 7 months (Progenies from the cross 3, SP80-3280 x RB855156) p=0.96
A1CR3C1In1_0.90 - HB vs LB - Internode 1 - 7 months (Progenies from the cross 3, SP80-3280 x RB855156) p=0.90
A1CR3C1L1_0.90 - HB vs LB - Leaf+1 - 7 months (Progenies from the cross 3, SP80-3280 x RB855156) p=0.90
A1CR3C2PhotoL1_0.90 - HPvsLP-Leaf+1-7 months (genotype 321vs318 from cross 3: SP80-3280xRB855156)
A1CR4C1In1_0.90 - HB vs LB - Internode 1 - 7 months (Progenies from the cross 4, RB855002 x RB855035) p=0.90
A1CR4C1L1_0.90 - HB vs LB - Leaf+1 - 7 months (Progenies from the cross 4, RB855002 x RB855035) p=0.90
A1CR4C1L1_0.96 - HB vs LB - Leaf+1 - 7 months (Progenies from the cross 4, RB855002 x RB855035) p=0.96
A1CR4C2PhotoL1_0.90 - LPxHP - Leaf+1-7 months - genotypes from crossing 4
A1CRT3C1In1_0.90 - HB vs LB - Internode 1 - 7 months (Progenies from the cross T3) p=0.90
A1CRT3C1L1_0.90 - HB vs LB - leaf +1 - 7 months (Progenies from the cross T3) p=0.90
A1CRT3C1L1_0.96 - HB vs LB - leaf +1 - 7 months (Progenies from the cross T3) p=0.96
A1CRT3C2PhotoL1_0.90 - HPvsLP-Leaf+1-7 months (Genotypes Cross T3: 710vs714)
GA (gibberellin - 0.5mg/L) - Evaluation of differential gene expression of sugarcane seedlings with gibberellins application
SnRK1 superexpression - differential expression of genes from SnRK1 sugarcane superexpression
SnRK1_SI_Exp x Control WT - Differential expression of genes from SnRK1 sugarcane silencing
SnRK1_SI_Exp x Control WT_2 - Differential expression of genes from SnRK1 sugarcane silencing

Search by: RegExp annotation

Type of result: Raw Intensity

Filter by expression: Show genes in at least one experiment

Upload your file:

The input file must have this pattern:

transcript_name	experiment	01	exper
transcript_name	Officix7515_In1.c	Officix	
SCQGSB1083B12.g	9.7632	9	
SCUTAD1033B04.b	6.1361	6	
SCUTAD1034G07.g	10.7420	1	

An example:

Browse... No file selected.

Submit a file

Search Potentially Activated Pathways/Terms

Description: The goal of this tool is to provide the user a list of significant expressed genes, by selecting experiments or crossing. The output is shown in a table format containing the gene name, annotation data and the selected experiments values.

Select gene category input: KEGG File SugarcaneSAsvsKEGG_KO_5SpeciesScore30

Select hierarchical clustering method: average

Enter cut tree threshold: 15

Select correlation type: spearman

Execute

(C)

Figura 64. Interface para análise comparativa de escore de atividade das vias entre diferentes condições experimentais. Inicialmente é escolhida a forma de seleção dos experimentos (A), por arquivo ou por projeto, em seguida são selecionados dois ou mais experimentos disponíveis ao usuário no *CaneRegNetDB* (B) e o banco de vias desejado, com o método para agrupamento hierárquico das vias e o método para correlação entre as condições experimentais (C).

O aplicativo para análise de atividade das vias entre duas condições experimentais utiliza o método de teste de Wilcoxon não pareado, considerando que as amostras são independentes.

Estamos interessados em testar se as distribuições das amostras são iguais em localização, ou seja, estamos interessados em saber se uma população tende a ter valores maiores do que a outra, ou se elas têm a mesma mediana. Em seguida calculamos os grupos *quantile*, definindo sete grupos de vias correspondentes às probabilidades (0, 5, 25, 50, 75, 95, 100), a partir da razão entre as duas condições experimentais. Sendo exibido ao final um gráfico da razão entre as duas condições experimentais com o *p-value* calculado pelo teste de Wilcoxon, neste caso a hipótese nula é que a distribuição entre as duas amostras diferem por um deslocamento da localização de m , e a hipótese alternativa é de que as amostras diferem de algum outro deslocamento de localização.

Para o estudo de caso foram comparados os experimentos A1V2E3_24h_tratado e A1V2E6_120h_tratado, e pelo teste de Wilcoxon o *p-value* = 3.7e-12 indicou a uma grande diferença entre as amostras escolhidas, validando a hipótese alternativa, com podemos verificar no gráfico (Figura 65).

Figura 65. Representação do log2 da razão do escore de atividade entre as duas condições experimentais, com o *p-value* do teste de Wilcoxon.

A análise por *quantile* determinou os valores para cada intervalo: (0, 5, 25, 50, 75, 95, 100) ~ (-3.1881154, -0.3975466, 0.4823275, 1.2656205, 1.9758057, 3.1893627, 5.5828529), com o seguinte total de vias em cada intervalo (1, 25, 31, 31, 25, 1), este resultado exibe um arquivo com a lista das vias em cada intervalo possibilitando identificar as vias específicas para cada grupo do *quantile*.

O aplicativo para análise de atividade das vias entre múltiplas condições experimentais foi aplicado aos experimentos utilizados como estudo de caso anteriormente, correspondendo aos pontos experimentais de 24h, 72h e 120h, sendo separados por amostras controle (A1V2E3_0.80c, A1V2E5_0.80c, A1V2E6_0.80c) e tratadas (A1V2E3_0.80t, A1V2E5_0.80t, A1V2E6_0.80t). A análise é inicializada pelo aplicativo da plataforma e realiza a chamada de scripts em Perl e R para realizar o cálculo do escore para cada via em cada condição experimental, gerando um arquivo com o nome da via em cada linha e em cada coluna o experimento com o respectivo escore calculado. Este arquivo com os escores será utilizado como entrada para o script em R para as análises de agrupamento hierárquico, e para a análise de correlação de Spearman entre as condições experimentais.

A análise sobre os experimentos selecionados produzirá inicialmente um *heatmap* do agrupamento hierárquico com a criação de dendogramas para agrupamento das vias de acordo com a atividade relativa e o dendograma de agrupamento entre as condições experimentais (Figura 66). Baseado no dendograma das amostras, podemos verificar a separação destas duas amostras (A1V2E5_0.80t e A1V2E6_0.80t) das demais, correspondendo às amostras no ponto de maior estresse hídrico, após 72h e 120h respectivamente, sugerindo uma maior variação no nível de atividade das vias em relação às outras amostras, sugerindo que podem estar relacionadas à resposta ao estresse (Figura 64, grupo em vermelho na base da figura). Um grupo contendo as vias de *Photosynthesis*, *Photosynthesis – antenna proteins*, *Pentose phosphate*, *Fructose and Mannose metabolismo*, *Carbon fixation*, *Diterpenoid biosynthesis* e *Porhyrin and chlorophyll metabolismo*, sugere uma diminuição de atividade destas vias, provavelmente devido a resposta ao estresse hídrico, diminuição nos níveis de atividade das vias de *Photosynthesis* e *Carbon fixation* para experimentos em condições de estresse hídrico tem sido relatados (Cramer et. al., 2013; Tezara et. al., 1999). As amostras no ponto inicial medido 24h (A1V2E3_0.80c, A1V2E3_0.80t) também foram agrupadas separadamente, e apresentaram um nível de atividade geral mais baixo na maioria das vias, sugerindo que esta primeira medida não identificou variação entre o tratamento e o controle do mesmo ponto. Esses resultados são condizentes com a

análise de enriquecimento de vias que também não identificou vias enriquecidas para este primeiro ponto de 24h. Um terceiro grupo corresponde às amostras controle (A1V2E5_0.80c, A1V2E6_0.80c) que não sofreram estresse hídrico, apresentam um padrão de atividade diferenciado e foram agrupadas separadamente das demais amostras, mostrando um nível de atividade um pouco diferente das amostras tratadas correspondentes. Para cada grupo de vias metabólicas do agrupamento hierárquico, são gerados os gráficos dos grupos de vias com o nível de atividade relativa em cada ponto experimental (Figura 67), e são disponibilizados arquivos texto para download contendo as vias e os respectivos escores de atividade em cada ponto experimental. O aplicativo exibe o *heatmap* da correlação de Spearman entre as amostras (Figura 68), e este gráfico revela um alto nível de correlação entre as A1V2E3_0.80c, A1V2E3_0.80t, A1V2E5_0.80c e A1V2E5_0.80, e uma alta correlação entre as duas condições em pontos de maior estresse hídrico A1V2E5_0.80t, A1V2E6_0.80t, sugerindo um padrão na resposta ao estresse ao longo do tempo.

Figura 67. Grupos de vias extraídas do agrupamento hierárquico usando *R-statistics*. Linhas cinzas indicam vias individuais, linha verde nível de atividade média por grupo e no topo o total de vias identificadas no cluster.

Figura 68. Heatmap mostrando a matriz de correlação de Spearman hierarquicamente agrupada resultado da comparação entre os níveis de atividade das vias para cada par de amostras.

6.7.4. Análise de Topologia de Vias

Os sistemas metabólicos podem ser modelados como bancos de dados de grafos, e têm sido curados e melhorados constantemente para diferentes organismos, principalmente para plantas. Em particular, a representação de vias metabólicas em grafos (Karp P.D. and Paley M.S., 1995), permite agregar níveis de expressão de genes,

proteínas ou ativação/inibição de enzimas, e realizar complexas operações entre grafos (Karp P.D. et. al., 2000). Com o objetivo de criar ambientes de consulta para vias metabólicas, com especificações e ferramentas, os bancos podem ser classificados em três categorias. A primeira categoria corresponde aos bancos de dados referência para vias metabólicas, como KEGG (Kanehisa et.al., 2000), BioCyc e MetaCyc (Caspi R et. al., 2008), PlantCyc (<http://plantcyc.org>) e Reactome (Joshi-Tope G et al, 2005), que possuem interfaces de consultas sobre as redes bioquímicas. A segunda categoria inclui sistemas biológicos e modelagem de sistemas como o BioSim (Heidtke KR, 1998), GenSim (Karp PD, 1993) e PathCase (Elliott et al. 2008). Um terceiro grupo provê simulações biológicas, modelagens de sistemas e o desenvolvimento de ambientes de consultas para explorar o comportamento das redes bioquímicas, sendo que uma das abordagens é a criação de linguagens declarativas preservando e conservando propriedades estruturais, seus relacionamentos e a representação dos grafos em modelos de banco de dado relacional (Periscope/GQ - Tian Y., 2008), de orientação a objeto (MQL - Cakmak, 2008) sendo parte do *framework* PathCase, com PL/SQL (PQL - Leser U, 2005) e o XML (bcnQL - Yang, 2008). Essas plataformas modelam as redes em grafos que proporcionam consultas das propriedades estruturais das vias metabólicas (vias, vizinhanças, ciclos, etc), informações de relacionamento entre entidades (inibidores de uma reação em uma via particular), porém as plataformas bcnQL e Periscope não possuem código fonte disponíveis e não possuem uma plataforma para testes, e o MQL corresponde a uma linguagem desenvolvida para o *framework* PathCase, sendo uma ferramenta interessante, porém não possui código aberto, e não possui estrutura para incorporar novos dados através do website.

Devido à indisponibilidade destas plataformas e ferramentas, e restrições devido à modelagem adequada a organismos ou perguntas específicas, que permitam a re-utilização e adaptação de novas funcionalidades para análises de dados específicos e que possam ser incorporadas à plataforma web CaneRegNet, iniciamos o desenvolvimento de uma nova estrutura baseada em um banco de dados de grafos, através do banco de dados de grafos Neo4j. O Neo4j é código aberto e está em constante desenvolvimento, contendo diversos pacotes com métodos para consultas e análises baseados em estruturas de grafos. A estrutura desenvolvida no banco de grafos é baseada em redes de vias metabólicas e tem o objetivo de permitir a adição de novos nós e relacionamentos para diferentes ômicas, possibilitando a utilização de métodos e funções existentes para a criação de novas ferramentas de análises de dados de perfis de expressão por transcriptômica, proteômica e metaboloma ampliando o alcance de integração e análise baseado em vias metabólicas. O banco de grafos pode ser acessado

via consultas através de programas em diversas linguagens de programação como Java, Perl, Python entre outras através de APIs nativas em Java ou para clientes REST-Neo4j, além de possuir uma interface para consultas com a linguagem declarativa Cypher, inspirada em SQL para descrever os padrões nos grafos. Esta linguagem permite selecionar, consultar, atualizar, inserir e deletar sem a necessidade de descrever exatamente como fazer, além de poder ser usada por meio de APIs.

Como exemplo, foram realizadas consultas com as ferramentas desenvolvidas, utilizando como referência a Via Metabólica de Amido e Teor de Sacarose, a qual está envolvida com o estresse por seca (Yao et. al., 2013), tendo sido confirmada pela análise comparativa do escore de atividade das vias, entre os experimentos de estresse por seca de microarranjo A1V2E3 (24 horas sem irrigação), A1V2E5 (72 horas sem irrigação) e A1V2E6 (120 horas sem irrigação), apresentando maior diferença de atividade da via entre os experimentos A1V2E3 e A1V2E6 (Figura 69).

Selecionamos a enzima phosphoglucomutase EC:5.4.2.2, para a análise de identificação do número de caminhos, caminho mais longo e número de enzimas ativas nos caminhos para os experimentos selecionados. Considerando os caminhos em todas as direções passando pela enzima 5.4.2.2, foram encontrados 65 possíveis caminhos, contabilizando caminhos internos aos mais longos, sendo que o caminho mais longo apresentou 11 enzimas, o experimento A1V2E3 identificou apenas 1 enzima ativa, enquanto o experimento A1V2E6 identificou 4 enzimas ativas, indicando que após 120 horas de estresse o número de genes ativos aumentou, indicando uma possível relação com o nível de estresse da planta. O resultado exibido será o caminho considerando todas as direções, entre parênteses os números das enzimas e entre colchetes os compostos, e a representação deste caminho na via metabólica do KEGG (Figura 70).

(5.4.2.2)--[C00103]--(3.2.1.21)--[C00185]--(3.2.1.4)--[C00760]--(2.4.1.12)--
[C00029]--(2.7.7.9)--[C00718]--(2.4.1.21)--[C00498]--(3.6.1.21)--[C00394]--
(2.7.7.34)--[C00208]--(2.4.1.25)--[C00031]--(2.4.1.10)--[C00089]--(3.2.1.26)

Figura 69. Representação do escore de atividade das vias nas três condições experimentais de estresse por seca.

7. Conclusões e Perspectivas

7.1. Conclusões

As tecnologias de larga escala em biologia molecular tem evoluído e produzido grandes volumes de dados, porém há uma escassez de bases de conhecimento, e anotações incompletas, proporcionando desafios para anotação e metodologias para interpretação e integração destes dados. Deste modo, os tipos de abordagens para integração computacional descritas nesta tese tornam-se extremamente importantes. O estudo do transcriptoma da cana-de-açúcar tem sido foco principal das análises e desenvolvimentos desta tese. Este trabalho contribuiu significativamente nas áreas de anotação funcional, integração, visualização e mineração de dados de expressão gênica, a partir de dados obtidos por plataformas de microarranjo, sequenciamento de larga escala e dados da literatura. Apresentamos o banco de dados CaneRegNetDB, que é o banco de dados referência para cana-de-açúcar, consolidando inúmeros bancos integrados à plataforma CaneRegNet que permite o gerenciamento, armazenamento, processamento, recuperação, análises e visualização, de experimentos de microarranjo, integrados a conhecimentos biológicos, com diferentes aplicações para mineração de dados para extrair e interpretar a biologia subjacente às medidas moleculares de larga escala, através da integração de bancos de anotações funcionais e estudos por expressão gênica, para busca de genes, categorias funcionais e vias metabólicas estatisticamente significantes em condições experimentais de interesse.

Esta tese foi motivada inicialmente pela necessidade da atualização da anotação do transcriptoma de cana-de-açúcar, para a criação da plataforma de microarranjo para medir níveis de expressão de genes senso e antisense de interesse, cobrindo 40% do transcriptoma. A melhora na qualidade e especificidade da anotação dos genes, atribuindo uma anotação única a partir de bancos de diferentes gramíneas, por *Gene Ontology* e *KEGG*, foi alcançada através da estratégia de novas pipelines e abordagens apresentados no Capítulo 3, os quais permitiram ampliar a integração de dados pelas metodologias desenvolvidas, como o catálogo de vias metabólicas para cana-de-açúcar, o qual possibilitou a criação de um banco de dados grafos e ferramentas para estudo da expressão gênica, através da topologia das vias. Foi possível a identificação de genes ortólogos e parálogos entre cana-de-açúcar e gramíneas, uma metodologia para a construção do ORFeoma de cana-de-açúcar, e uma estratégia para a identificação e classificação do primeiro banco completo putativo das famílias de TFs em gramíneas. A grande limitação com abordagens baseadas em similaridade de sequências, ocorre

devido à baixa similaridade, ausência de organismos próximos evolutivamente e principalmente ausência de estudos experimentais com caracterização funcional destes genes, necessitando de curadoria manual.

Nesta tese apresentamos o banco de dados CaneRegNetDB e a plataforma CaneRegNet, baseados em uma lógica abrangente para gerenciamento e controle de acesso por módulos e dados para fácil gerenciamento de múltiplos grupos de pesquisa, projetos e disponibilização dos dados. A plataforma permite o gerenciamento, a submissão, o processamento e análises de forma automatizada para estudo da expressão gênica, identificando e armazenando resultados no CaneRegNetDB, através de pacotes e codificação de programas em scripts Java, Perl e R. A criação de bases de conhecimento em cana-de-açúcar permitiu a implementação e desenvolvimento de métodos integrados à plataforma CaneRegNet para estudo dos níveis de expressão gênica por métodos de larga escala, para as análises de categorias e vias dirigidas a bases de conhecimento. Como o método para enriquecimento de termos e o método para cálculo do escore de atividades das vias metabólicas integrado a métodos de agrupamento hierárquico e correlação, possibilitando maior compreensão dos processos biológicos e identificação de possíveis modulações de acordo como a condição experimental ou tratamento.

O resultado é o desenvolvimento de metodologias para anotação de transcriptoma de cana-de-açúcar, e uma plataforma robusta para mineração de dados, integrada a um banco de dados relacional, com uma variedade de ferramentas e APIs em Java, Perl e R servindo como recursos de bioinformática para projetos similares na anotação e no estudo de expressão gênica por tecnologias de larga escala.

7.2. Perspectivas

Esperamos que o banco de dados e a plataforma continuem em pleno desenvolvimento e atualização para tornar-se não apenas a referência para estudos do transcriptoma de cana-de-açúcar, mas também para os estudos do genoma da cana-de-açúcar, como já funciona atualmente para alguns resultados preliminares do sequenciamento do genoma. E para que os dados disponíveis na plataforma CaneRegNet, possam ser combinados com as sequências genômicas, para identificar e entender como os motivos regulatórios controlam a expressão gênica com estudos disponíveis no CaneRegNetDB. A integração destes dados com os experimentos no laboratório de ChIP-Seq possibilitarão a identificação dos genes e regiões genômicas reguladas pelos fatores de transcrição classificados no CaneRegNetDB. A atualização da

plataforma CaneRegNet e do banco de dados CaneRegNetDB para integrar dados de novas ômicas como proteoma e metaboloma, e dados de expressão gênica por sequenciamento de larga escala, aliada a atualização e criação novas aplicações web, o CaneRegNet ganhará enorme visibilidade e capacidade para geração de novas bases do conhecimento para os estudos da cana-de-açúcar.

Será importante o aprimoramento dos métodos existentes com a incorporação de novos métodos de classificação mais robustos utilizando as bases do conhecimento construídas e outros métodos para análise de expressão gênica como inferência de redes (Hurley et al., 2012; Fujita et al., 2014).

Alguns desenvolvimentos futuros visam a abordagens para a integração das vias metabólicas, com dados de expressão por transcriptoma, proteoma ou metaboloma para permitir análises topológicas e funcionais, definindo novas hipóteses topológicas integradas a novos esquemas de cálculos de escore de enriquecimento.

8. Referências

- Ackermann M, Strimmer K (2009) A general modular framework for gene set enrichment analysis. *BMC Bioinformatics* 10: 47.
- Alexeyenko A, Tamas I, Liu G, Sonnhammer EL (2006) Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics* 22: e9-15.
- Al-Shahrour F, Diaz-Uriarte R, Dopazo J (2005) Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information. *Bioinformatics* 21: 2988–2993.
- Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. "Basic local alignment search tool." *Journal of molecular biology* 215, no. 3 (1990): 403-410.
- Angles Renzo. Say hi to GraphDB-Bench. 2012. (<http://dcc.utalca.cl/~rangles/files/gdm2012.pdf>)
- Aoki K., Ogata Y. & Shibata D. (2007) Approaches for extracting practical information from gene co-expression networks in plant biology. *Plant and Cell Physiology* 48, 381–390.
- Aoki K, Yano K, Suzuki A, Kawamura S, Sakurai N, et al. (2010) Large-scale analysis of full-length cDNAs from the tomato (*Solanum lycopersicum*) cultivar Micro-Tom, a reference system for the Solanaceae genomics. *BMC Genomics* 11: 210.
- Ashburner, Michael, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis et al. "Gene Ontology: tool for the unification of biology." *Nature genetics* 25, no. 1 (2000): 25-29.
- Bader, Gary D., Doron Betel, and Christopher WV Hogue. "BIND: the biomolecular interaction network database." *Nucleic acids research* 31, no. 1 (2003): 248-250.
- Bai, Youhuang, Xiaozhuan Dai, Andrew P. Harrison, and Ming Chen. "RNA regulatory networks in animals and plants: a long noncoding RNA perspective." *Briefings in functional genomics* 2014: elu017).
- Bairoch, Amos, Rolf Apweiler, Cathy H. Wu, Winona C. Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger et al. "The universal protein resource (UniProt)." *Nucleic acids research* 33, no. suppl 1 (2005): D154-D159.

Bairoch, Amos. "The ENZYME data bank in 1999." *Nucleic acids research* 27, no. 1 (1999): 310-311.

Ball, Catherine A., Ihab AB Awad, Janos Demeter, Jeremy Gollub, Joan M. Hebert, Tina Hernandez-Boussard, Heng Jin et al. "The Stanford Microarray Database accommodates additional microarray platforms and data formats." *Nucleic Acids Research* 33, no. suppl 1 (2005): D580-D582.

Ball, Catherine, Alvis Brazma, Helen Causton, Steve Chervitz, Ron Edgar, Pascal Hingamp, John C. Matese et al. "An open letter on microarray data from the MGED Society." *Microbiology (Reading, England)* 150, no. Pt 11 (2004): 3522.

Barrett, Tanya, Dennis B. Troup, Stephen E. Wilhite, Pierre Ledoux, Dmitry Rudnev, Carlos Evangelista, Irene F. Kim, Alexandra Soboleva, Maxim Tomashevsky, and Ron Edgar. "NCBI GEO: mining tens of millions of expression profiles—database and tools update." *Nucleic acids research* 35, no. suppl 1 (2007): D760-D765.

Bateman, Alex, Lachlan Coin, Richard Durbin, Robert D. Finn, Volker Hollich, Sam Griffiths-Jones, Ajay Khanna et al. "The Pfam protein families database." *Nucleic acids research* 32, no. suppl 1 (2004): D138-D141.

Bellingham, Shayne A., Bradley M. Coleman, and Andrew F. Hill. "Small RNA deep sequencing reveals a distinct miRNA signature released in exosomes from prion-infected neuronal cells." *Nucleic acids research* (2012): gks832.

Bennetzen JL (2007) Patterns in grass genome evolution. *Curr Opin Plant Biol* 10: 176-181

Benson, Dennis A., Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and Eric W. Sayers. "GenBank." *Nucleic acids research* (2012): gks1195.

Bergholz, Andre, Stephan Heymann, Jörg A. Schenk, and Johann Christoph Freytag. "Biological sequences integrated: a relational database approach." *Acta biotheoretica* 49, no. 3 (2001): 145-159.

Berglund AC, Sjolund E, Ostlund G, Sonnhammer EL (2008) InParanoid 6: eukaryotic ortholog clusters with inparalogs. *Nucleic Acids Res* 36: D263–266. 42.

Berrar, Daniel P., Werner Dubitzky, and Martin Granzow, eds. *A practical approach to microarray data analysis*. Boston, MA: Kluwer Academic Publishers, 2003.

Birney, Ewan, John A. Stamatoyannopoulos, Anindya Dutta, Roderic Guigó, Thomas R. Gingeras, Elliott H. Margulies, Zhiping Weng et al. "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project." *Nature* 447, no. 7146 (2007): 799-816.

Bland, J. Martin, and Douglas G. Altman. "Multiple significance tests: the Bonferroni method." *Bmj* 310, no. 6973 (1995): 170.

Bokhari, Jiang, Z., Tang, H., Ventura, M., Cardone, M. F., Marques-Bonet, T., She, X., Pevzner, P. A. and Eichler, E. E. (2007) Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution *Nat Genet* 39, 11, 1361–8. 28.

Bonneau, Richard, and David Baker. "Ab initio protein structure prediction: progress and prospects." *Annual review of biophysics and biomolecular structure* 30, no. 1 (2001): 173-189.

Bossers K, Ylstra B, Brakenhoff RH, Smeets SJ, Verhaagen J, van de Wiel MA, 2010. Intensity-based analysis of dual-color gene expression data as an alternative to ratio-based analysis to enhance reproducibility. *BMC Genomics* 11:112. doi:10.1186/1471-2164-11-112.

Box, G. E. P., Hunter, W., and Hunter, S. (1969). *Statistics for experimenters*. John Wiley.

Box, George EP, William Gordon Hunter, and J. Stuart Hunter. "Statistics for experimenters." (1978): 144.

Braslavsky, I., Hebert, B., Kartalov, E. and Quake, S. R. (2003) Sequence information can be obtained from single DNA molecules. *Proc Natl Acad Sci U S A* 100 , 7, 3960–3964 ISSN 0027-8424 (Print); 0027-8424 (Linking). 4, 12.

Braun EL, Grotewold E (2001) Fungal Zuotin proteins evolved from MIDA1-like factors by lineage-specific loss of MYB domains. *Mol. Biol. & Evol.* 18: 1401- 1412

Brazma, Alvis, Helen Parkinson, Ugis Sarkans, Mohammadreza Shojatalab, Jaak Vilo, Niran Abeygunawardena, Ele Holloway et al. "ArrayExpress—a public repository for microarray gene expression data at the EBI." *Nucleic acids research* 31, no. 1 (2003): 68-71.

Brazma, Alvis. "Minimum information about a microarray experiment (MIAME)—successes, failures, challenges." *The Scientific World Journal* 9 (2009): 420-423.

Bremer, G. "Problems in breeding and cytology of sugar cane." *Euphytica* 10, no. 1 (1961): 59-78.

Brenner, Sydney, Maria Johnson, John Bridgham, George Golda, David H. Lloyd, Davida Johnson, Shujun Luo et al. "Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays." *Nature biotechnology* 18, no. 6 (2000): 630-634.

Bryan Basham, Kathy Sierra, Bert Bates. *Head First Servlets & JSP*. Second Edition O'Reilly Media Inc. 2008.

Bryan Basham, Kathy Sierra, Bert Bates. *Head First Servlets & JSP*. Second Edition O'Reilly Media Inc. 2008.

Buckeridge, Marcos S., Amanda P. Souza, Rebecca A. Arundale, Kristina J. Anderson-Teixeira, and Evan DeLucia. "Ethanol from sugarcane in Brazil: a 'midway' strategy for increasing ethanol production while maximizing environmental benefits." *GCB Bioenergy* 4, no. 2 (2012): 119-126.

Buneman, Peter, Sanjeev Khanna, and Wang-Chiew Tan. "Data provenance: Some basic issues." In *FST TCS 2000: Foundations of software technology and theoretical computer science*, pp. 87-93. Springer Berlin Heidelberg, 2000.

Cakmak, Ali, GULTEKIN OZSOYOGLU, and Richard W. Hanson. "Querying metabolism under different physiological constraints." *Journal of bioinformatics and computational biology* 8, no. 02 (2010): 247-293.

CANTARELLA, H. ; Buckeridge, Marcos S. ; Van Sluys M-A ; SOUZA, A. P. ; Garcia, Antônio A. F. ; Nishiyama-Junior, M. Y. ; MACIEL-FILHO, R. ; CRUZ, C. H. B. ; BONOS, S. ; SOUZA, G. M. . Sugarcane: the most efficient crop for biofuel production. In: Taylor & Francis Group. (Org.). Chittaranjan Kole; Chandrashekhar P. Joshi; David R. Shonnard. Boca Rotan: Taylor & Francis Group, 2010, v. , p. -.

Cardoso-Silva CB, Costa EA, Mancini MC, Balsalobre TW, Canesin LE, et al. (2014) De novo assembly and transcriptome analysis of contrasting sugarcane varieties. *PLoS One* 9: e88462

Carolina Gimiliani, Milton Yutaka Nishiyama Jr, Paloma Mieko Sato, Rodrigo Fandino de Andrade, and Glaucia Mendes Souza. "Identification of sense and antisense transcripts

regulated by drought in sugarcane." *Plant molecular biology* 79, no. 4-5 (2012): 461-477.

Caspi R, Foerster H, Fulcher CA, et al., 2008. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res* 36:D623-31.

Caspi, Ron, Hartmut Foerster, Carol A. Fulcher, Rebecca Hopkinson, John Ingraham, Pallavi Kaipa, Markus Krummenacker et al. "MetaCyc: a multiorganism database of metabolic pathways and enzymes." *Nucleic acids research* 34, no. suppl 1 (2006): D511-D516.

Castillo-Davis, Cristian I., and Daniel L. Hartl. "GeneMerge—post-genomic analysis, data mining, and hypothesis testing." *Bioinformatics* 19, no. 7 (2003): 891-892.

Casu, Rosanne, Carlos Takeshi Hotta, Glaucia Mendes Souza, Robert J. Henry, and Chittaranjan Kole. "Functional genomics: transcriptomics of sugarcane-current status and future prospects." *Genetics, Genomics and Breeding of Sugarcane* (2010): 167-191.

Charrad, Malika, Nadia Ghazzali, Véronique Boiteau, and Azam Niknafs. "NbClust Package: finding the relevant number of clusters in a dataset." In *UseR!* 2012.

Chen Y, Dougherty E R, Bittner M. Ratio-based decisions and the qualitative analysis of cDNA microarray images. *J. Biomed. Opt.*, 2:364–374, 1997.

Chen, Dijun, Chunhui Yuan, Jian Zhang, Zhao Zhang, Lin Bai, Yijun Meng, Ling-Ling Chen, and Ming Chen. "PlantNATsDB: a comprehensive database of plant natural antisense transcripts." *Nucleic acids research* 40, no. D1 (2012): D1187-D1193..

Chen, Feng, Aaron J. Mackey, Christian J. Stoeckert, and David S. Roos. "OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups." *Nucleic acids research* 34, no. suppl 1 (2006): D363-D368.

Chen, Ming-Hsu, Prabhjot Kaur, Bruce Dien, Frederick Below, Michael L. Vincent, and Vijay Singh. "Use of tropical maize for bioethanol production." *World Journal of Microbiology and Biotechnology* 29, no. 8 (2013): 1509-1515.

Cleveland, William S. "Robust locally weighted regression and smoothing scatterplots." *Journal of the American statistical association* 74, no. 368 (1979): 829-836.

Cleveland, William S., and Susan J. Devlin. "Locally weighted regression: an approach to regression analysis by local fitting." *Journal of the American Statistical Association* 83, no. 403 (1988): 596-610.

Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L. and Rice, P. M. (2010) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res* 38, 6, 1767–1771 ISSN 1362-4962 (Electronic); 0305-1048 (Linking). 26.

Codd, Edgar F. "A relational model of data for large shared data banks." *Communications of the ACM* 13, no. 6 (1970): 377-387.

Codd, Edgar F. "Data models in database management." In *ACM Sigmod Record*, vol. 11, no. 2, pp. 112-114. ACM, 1980.

Conley AB, Jordan IK (2012) Epigenetic regulation of human cis-natural antisense transcripts. *Nucleic Acids Res* 40: 1438–1445.

Coram TE, Settles ML, Chen X (2009) Large-scale analysis of antisense transcription in wheat using the Affymetrix GeneChip wheat genome array. *BMC Genomics* 10:253

Cramer, Grant R., Steve C. Van Sluyter, Daniel W. Hopper, Dana Pascovici, Tim Keighley, and Paul A. Haynes. "Proteomic analysis indicates massive changes in metabolism prior to the inhibition of growth and photosynthesis of grapevine (*Vitis vinifera* L.) in response to water deficit." *BMC plant biology* 13, no. 1 (2013): 49.

Crick, F.H.C. 1958. On protein synthesis. *Symp Soc Exp Biol* 12:138–163.

Croft, David, Gavin O'Kelly, Guanming Wu, Robin Haw, Marc Gillespie, Lisa Matthews, Michael Caudy et al. "Reactome: a database of reactions, pathways and biological processes." *Nucleic acids research* (2010): gkq1018.

Curtis, R. Keira, Matej Orešič, and Antonio Vidal-Puig. "Pathways to the analysis of microarray data." *Trends in biotechnology* 23, no. 8 (2005): 429-435.

Dal-Bianco, Maximiller, Monalisa Sampaio Carneiro, Carlos Takeshi Hotta, Roberto Giacomini Chapola, Hermann Paulo Hoffmann, Antonio Augusto Franco Garcia, and Glaucia Mendes Souza. "Sugarcane improvement: how far can we go?." *Current opinion in biotechnology* 23, no. 2 (2012): 265-270.

- Datta, Ruchira S., Christopher Meacham, Bushra Samad, Christoph Neyer, and Kimmen Sjölander. "Berkeley PHOG: PhyloFacts orthology group prediction web server." *Nucleic acids research* (2009): gkp373.
- Davidson, Susan B., Chris Overton, and Peter Buneman. "Challenges in integrating biological data sources." *Journal of Computational Biology* 2, no. 4 (1995): 557-572.
- Davis, J. Wade. "Statistics for Microarrays: Design, Analysis and Inference." *Journal of the American Statistical Association* 101, no. 474 (2006).
- Davuluri, Ramana V., Hao Sun, Saranyan K. Palaniswamy, Nicole Matthews, Carlos Molina, Mike Kurtz, and Erich Grotewold. "AGRIS: Arabidopsis gene regulatory information server, an information resource of Arabidopsis cis-regulatory elements and transcription factors." *BMC bioinformatics* 4, no. 1 (2003): 25.
- De Setta, Nathalia, Cláudia B. Monteiro-Vitorello, Cushla J. Metcalfe, Guilherme Marcelo Q. Cruz, Luiz E. Del Bem, Renato Vicentini, Fábio Tebaldi S. Nogueira et al. "Building the sugarcane genome for biotechnology and identifying evolutionary trends." *BMC genomics* 15, no. 1 (2014): 540.
- de Siqueira Ferreira S, Nishiyama M, Paterson A, Souza G (2013). Biofuel and energy crops: high-yield Saccharinae take center stage in the post-genomics era. *Genome Biology* 14(6): 210.
- De Souza, Amanda Pereira, Marilia Gaspar, Emerson Alves Da Silva, Eugênio César Ulian, Alessandro Jaquiel Wacławovsky, RENATO VICENTINI DOS SANTOS, MARCELO MENOSSEI TEIXEIRA, GLAUCIA MENDES SOUZA, and MARCOS SILVEIRA BUCKERIDGE. "Elevated CO₂ increases photosynthesis, biomass and productivity, and modifies gene expression in sugarcane." *Plant, cell & environment* 31, no. 8 (2008): 1116-1127.
- Delenstarr G, Cattell H, Chen C, Dorsel A, Kincaid R, Nguyen K, Sampas N, Schidel S, Shannon K, Tu A, Wolber P: Estimation of the confidence limits of oligo nucleotide array-based measurements of differential expression.
- Delenstarr, Glenda, Herb Cattell, Chao Chen, Andreas N. Dorsel, Robert H. Kincaid, Khanh Nguyen, Nicholas M. Sampas et al. "Estimation of the confidence limits of oligonucleotide-array-based measurements of differential expression." In *BIOS 2001 The International Symposium on Biomedical Optics*, pp. 120-131. International Society for Optics and Photonics, 2001.

Delenstarr, Glenda, Herb Cattell, Scott Connell, Andreas Dorsel, Robert H. Kincaid, Khanh Nguyen, Nick Sampas et al. "Estimation of the confidence limits of oligonucleotide microarray-based measurements of differential expression; Kafsack, Björn FC, Heather J. Painter, and Manuel Llinás. "New Agilent platform DNA microarrays for transcriptome analysis of *Plasmodium falciparum* and *Plasmodium berghei* for the malaria research community." *Malaria journal* 11, no. 1 (2012): 187.

D'Hont, Angélique, David Ison, Karine Alix, Catherine Roux, and Jean Christophe Glaszmann. "Determination of basic chromosome numbers in the genus *Saccharum* by physical mapping of ribosomal RNA genes." *Genome* 41, no. 2 (1998): 221-225.

Dias AP, Braun EL, McMullen MD, Grotewold E (2003) Recently duplicated maize R2R3 Myb genes provide evidence for distinct mechanisms of evolutionary divergence after duplication. *Plant Physiol.* 131: 610-620

Draghici S, Khatri P, Tarca AL, Amin K, Done A, et al. (2007) A systems biology approach for pathway level analysis. *Genome Res* 17: 1537–1545. 39.

Drmanac, Radoje, Andrew B. Sparks, Matthew J. Callow, Aaron L. Halpern, Norman L. Burns, Bahram G. Kermani, Paolo Carnevali et al. "Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays." *Science* 327, no. 5961 (2010): 78-81.

Eddy, Sean R. "Profile hidden Markov models." *Bioinformatics* 14, no. 9 (1998): 755-763.

Edgar, Ron, Michael Domrachev, and Alex E. Lash. "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository." *Nucleic acids research* 30, no. 1 (2002): 207-210.

Edwards, Kieron D., Paul E. Anderson, Anthony Hall, Neeraj S. Salathia, James CW Locke, James R. Lynn, Martin Straume, James Q. Smith, and Andrew J. Millar. "FLOWERING LOCUS C mediates natural variation in the high-temperature response of the *Arabidopsis* circadian clock." *The Plant Cell Online* 18, no. 3 (2006): 639-650).

Eid, John, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso et al. "Real-time DNA sequencing from single polymerase molecules." *Science* 323, no. 5910 (2009): 133-138.

Elliott, B., Kirac, M., Cakmak, A. et al. PathCase: Pathways Database System. *Bioinformatics* 24(21): 2526-2533. 2008.

Eric Freeman, Elisabeth Freeman, Kathy Sierra, Bert Bates. *Head First Design Patterns*. First Edition O'Reilly Media Inc. 2004.

Federhen, Scott. "The NCBI taxonomy database." *Nucleic acids research* 40, no. D1 (2012): D136-D143.

Fedurco, Milan, Anthony Romieu, Scott Williams, Isabelle Lawrence, and Gerardo Turcatti. "BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies." *Nucleic acids research* 34, no. 3 (2006): e22-e22.

Figueira, Thais Rezende S., Vagner Okura, Felipe Rodrigues da Silva, Marcio Jose da Silva, Dave Kudrna, Jetty SS Ammiraju, Jayson Talag, Rod Wing, and Paulo Arruda. "A BAC library of the SP80-3280 sugarcane variety (*saccharum* sp.) and its inferred microsynteny with the sorghum genome." *BMC research notes* 5, no. 1 (2012): 185.

Fischer-Hübner, Simone. *IT-security and privacy: design and use of privacy-enhancing security mechanisms*. Springer-Verlag, 2001; Dieter Zeller. *Design and Development of a User Managment System for Molecular Biology Database Systems*. Master's thesis, TU-Graz, 2003.

Fitch WM (June 1970). "Distinguishing homólogous from analogous proteins". *Systematic Zoology* 19 (2): 99–113. doi:10.2307/2412448. PMID 5449325.

Friedberg, Iddo. "Automated protein function prediction—the genomic challenge." *Briefings in bioinformatics* 7, no. 3 (2006): 225-242.

Fukushima, Atsushi, Miyako Kusano, Henning Redestig, Masanori Arita, and Kazuki Saito. "Integrated omics approaches in plant systems biology." *Current opinion in chemical biology* 13, no. 5 (2009): 532-538.

Fujita, André, and Satoru Miyano. "A Tutorial to Identify Nonlinear Associations in Gene Expression Time Series Data." In *Transcription Factor Regulatory Networks*, pp. 87-95. Springer New York, 2014.

Garcia, Antonio AF, Marcelo Mollinari, Thiago G. Marconi, Oliver R. Serang, Renato R. Silva, Maria LC Vieira, Renato Vicentini et al. "SNP genotyping allows an in-depth characterisation of the genome of sugarcane and other complex autopolyploids." *Scientific reports* 3, 2103..

Gene Ontology Consortium. "The Gene Ontology (GO) database and informatics resource." *Nucleic acids research* 32, no. suppl 1 (2004): D258-D261.

Ghildiyal, Megha, and Phillip D. Zamore. "Small silencing RNAs: an expanding universe." *Nature Reviews Genetics* 10, no. 2 (2009): 94-108. .

Giardine, Belinda, Cathy Riemer, Ross C. Hardison, Richard Burhans, Laura Elnitski, Prachi Shah, Yi Zhang et al. "Galaxy: a platform for interactive large-scale genome analysis." *Genome research* 15, no. 10 (2005): 1451-1455.

Gilbert, David. "The jfreechart class library." *Developer Guide. Object Refinery* (2002).

Gilbert, Don. "Biomolecular interaction network database." *Briefings in bioinformatics* 6, no. 2 (2005): 194-198.

Glazko G, Emmert-Streib F (2009) Unite and conquer: univariate and multivariate approaches for finding differentially expressed gene sets. *Bioinformatics* 25: 2348–2354.

Goecks, Jeremy, Anton Nekrutenko, and James Taylor. "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences." *Genome Biol* 11, no. 8 (2010): R86.

Goeman JJ, Buhlmann P (2007) Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* 23: 980–987.

Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC (2004) A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* 20: 93–99; Kong SW, Pu WT, Park PJ (2006) A multivariate approach for integrating genomewide expression data and biological knowledge. *Bioinformatics* 22: 2373–2380.

Goodstein, David M., Shengqiang Shu, Russell Howson, Rochak Neupane, Richard D. Hayes, Joni Fazo, Therese Mitros et al. "Phytozome: a comparative platform for green plant genomics." *Nucleic acids research* 40, no. D1 (2012): D1178-D1186.

Grabherr, Manfred G., Brian J. Haas, Moran Yassour, Joshua Z. Levin, Dawn A. Thompson, Ido Amit, Xian Adiconis et al. "Full-length transcriptome assembly from RNA-Seq data without a reference genome." *Nature biotechnology* 29, no. 7 (2011): 644-652.

Graveley, Brenton R. "Molecular biology: power sequencing." *Nature* 453, no. 7199 (2008): 1197-1198.

Graveley, Brenton R., Angela N. Brooks, Joseph W. Carlson, Michael O. Duff, Jane M. Landolin, Li Yang, Carlo G. Artieri et al. "The developmental transcriptome of *Drosophila melanogaster*." *Nature* 471, no. 7339 (2011): 473-479.

Grivet L, Arruda P: Sugarcane genomics: depicting the complex genome of an important tropical crop. *Curr Opin Plant Biol* 2001, 5:122–127.

Guo, An-Yuan, Xin Chen, Ge Gao, He Zhang, Qi-Hui Zhu, Xiao-Chuan Liu, Ying-Fu Zhong, Xiaocheng Gu, Kun He, and Jingchu Luo. "PlantTFDB: a comprehensive plant transcription factor database." *Nucleic Acids Research* 36, no. suppl 1 (2008): D966-D969.

Gutzat, Ruben, and Ortrun Mittelsten Scheid. "Epigenetic responses to stress: triple defense?." *Current opinion in plant biology* 15, no. 5, 2012: 568-573.

Haas, Laura M., Peter M. Schwarz, Prasad Kodali, Elon Kotlar, Julia E. Rice, and William C. Swope. "DiscoveryLink: A system for integrated access to life sciences data sources." *IBM systems Journal* 40, no. 2 (2001): 489-511.

Hatch, M. D., and T. Kagawa. "Photosynthetic activities of isolated bundle sheath cells in relation to differing mechanisms of C4 pathway photosynthesis." *Archives of biochemistry and biophysics* 175.1 (1976): 39-53.

Hawkins, R.D., Hon, G.C., and Ren, B. (2010). Next-generation genomics: an integrative approach. *Nat Rev Genet* 11, 476-486.

Hegyí, Hedi, and Mark Gerstein. "Annotation transfer for genomics: measuring functional divergence in multi-domain proteins." *Genome research* 11, no. 10 (2001): 1632-1640.

Hegyí, Hedi, and Mark Gerstein. "The relationship between protein structure and function: a comprehensive survey with application to the yeast genome." *Journal of molecular biology* 288, no. 1 (1999): 147-164.

Heidtke KR, Schulze-Kremer S, 1998. BioSim — A new qualitative simulation environment for molecular biology, in *Proc of 6th IntConfIntellSystMolBiol*, pp. 85–94.

Heinemeyer, T., Edgar Wingender, Ingmar Reuter, Henning Hermjakob, Alexander E. Kel, O. V. Kel, Elena V. Ignatieva et al. "Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL." *Nucleic acids research* 26, no. 1 (1998): 362-367.

- Henz SR, Cumbie JS, Kasschau KD, Lohmann JU, Carrington JC, et al. (2007). Distinct expression patterns of natural antisense transcripts in Arabidopsis. *Plant Physiol* 144: 1247–1255.
- Hernandez, Thomas, and Subbarao Kambhampati. "Integration of biological sources: current systems and challenges ahead." *ACM SIGmod Record* 33, no. 3 (2004): 51-60.
- Hibberd, J. M.; Sheehy, J. E.; Langdale, J. A. (2008). "Using C4 photosynthesis to increase the yield of rice—rationale and feasibility". *Current Opinion in Plant Biology* 11 (2): 228–231. doi:10.1016/j.pbi.2007.11.002.
- Hotta, Carlos Takeshi, Milton Yutaka Nishiyama Jr, and Glaucia Mendes Souza. "Circadian Rhythms of Sense and Antisense Transcription in Sugarcane, a Highly Polyploid Crop." *PloS one* 8, no. 8 (2013): e71847.
- Huang, Da Wei, Brad T. Sherman, and Richard A. Lempicki. "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists." *Nucleic acids research* 37, no. 1 (2009): 1-13.
- Hughes, Timothy R., and Daniel D. Shoemaker. "DNA microarrays for expression profiling." *Current opinion in chemical biology* 5, no. 1 (2001): 21-25.
- Hughes, TR, Mao, M, Jones, AR, Burchard, J, Marton, MJ, Shannon, KW, Lefkowitz, SM, Ziman, M, Schelter, JM, Meyer, MR, Kobayashi, S, Davis, C, Dai, H, He, YD, Stephanians, SB, Cavet, G, Walker, WL, West, A, Coffey, E, Shoemaker, DD, Stoughton, R, Blanchard, AP, Friend, SH & Linsley, PS (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer, *Nature Biotechnology*, 19, 342-347.
- Hull, R. AND King, R. 1987. Semantic database modeling: Survey, applications, and research issues. *ACM Comput. Surv.* 19, 3, 201-260.
- Hurley, Daniel, Hiromitsu Araki, Yoshinori Tamada, Ben Dunmore, Deborah Sanders, Sally Humphreys, Muna Affara et al. "Gene network inference and visualization tools for biologists: application to new human transcriptome datasets." *Nucleic acids research* 40, no. 6 (2012): 2377-2398.
- Hyndman, R. J. and Fan, Y. (1996) Sample quantiles in statistical packages, *American Statistician* *50*, 361-365.

Iseli, Christian, C. Victor Jongeneel, and Philipp Bucher. "ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences." In ISMB, vol. 99, pp. 138-148. 1999.

J D Wren, A Kulkarni, J Joslin, R A Butow, and H R Garner. Cross-hybridization on PCRspotted microarrays. IEEE Engineering in Medicine and Biology Magazine , 21(2):71{5, April 2002.

Jagadish, H. V., and Frank Olken. "Database management for life sciences research." ACM SIGMOD Record 33, no. 2 (2004): 15-20.

Jen, Chih-Hung, Ioannis Michalopoulos, David R. Westhead, and Peter Meyer. "Natural antisense transcripts with coding capacity in Arabidopsis may have a regulatory role that is not linked to double-stranded RNA degradation." Genome biology 6, no. 6 (2005): R51.

Jeong, Hawoong, Bálint Tombor, Réka Albert, Zoltan N. Oltvai, and A-L. Barabási. "The large-scale organization of metabolic networks." Nature 407, no. 6804 (2000): 651-654.

Jiang Z, Gentleman R (2007) Extensions to gene set enrichment. Bioinformatics 23: 306–313.

Jorgensen, Richard A. "Plant Genetics and Genomics: Crops and Models."

Joshi-Tope, G., Marc Gillespie, Imre Vastrik, Peter D'Eustachio, Esther Schmidt, Bernard de Bono, Bijay Jassal et al. "Reactome: a knowledgebase of biological pathways." Nucleic acids research 33, no. suppl 1 (2005): D428-D432.

Joyce, Andrew R., and Bernhard Ø. Palsson. "The model organism as a system: integrating'omics' data sets." Nature Reviews Molecular Cell Biology 7, no. 3 (2006): 198-210.

Kafsack, Björn FC, Heather J. Painter, and Manuel Llinás. "New Agilent platform DNA microarrays for transcriptome analysis of Plasmodium falciparum and Plasmodium berghei for the malaria research community." Malaria journal 11, no. 1 (2012): 187).

Kanehisa, Minoru, and Susumu Goto. "KEGG: kyoto encyclopedia of genes and genomes." Nucleic acids research 28, no. 1 (2000): 27-30; Ogata, Hiroyuki, Susumu Goto, Kazushige Sato, Wataru Fujibuchi, Hidemasa Bono, and Minoru Kanehisa. "KEGG: Kyoto encyclopedia of genes and genomes." Nucleic acids research 27, no. 1 (1999): 29-34.

Kanehisa, Minoru, Susumu Goto, Shuichi Kawashima, and Akihiro Nakaya. "The KEGG databases at GenomeNet." *Nucleic acids research* 30, no. 1 (2002): 42-46.

Kanehisa, Minoru, Susumu Goto, Shuichi Kawashima, Yasushi Okuno, and Masahiro Hattori. "The KEGG resource for deciphering the genome." *Nucleic acids research* 32, no. suppl 1 (2004): D277-D280.

Karp PD, A qualitative biochemistry and its application to the tryptophan operon, in *Artificial Intelligence and Molecular Biology*, pp. 289–324, 1993.

Karp, Peter D. "What we do not know about sequence analysis and sequence databases." *Bioinformatics* 14, no. 9 (1998): 753-754).

Karp, Peter D., Suzanne Paley, and Pedro Romero. "The pathway tools software." *Bioinformatics* 18, no. suppl 1 (2002): S225-S232.

Khatri P, Draghici S, Tarca AL, Hassan SS, Romero R (2007) A system biology approach for the steady-state analysis of gene signaling networks. *Proc 12th Iberoamerican Congress on Pattern Recognition, CIARP 2007*; Valparaiso, Chile.

Khatri, Purvesh, Marina Sirota, and Atul J. Butte. "Ten years of pathway analysis: current approaches and outstanding challenges." *PLoS computational biology* 8, no. 2 (2012): e1002375.

Khraiwesh, Basel, Jian-Kang Zhu, and Jianhua Zhu. "Role of miRNAs and siRNAs in biotic and abiotic stress responses of plants." *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* 1819, no. 2 (2012): 137-148.

Kim SY, Volsky DJ (2005) PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics* 6: 144.

Kitano, Hiroaki. "Computational systems biology." *Nature* 420, no. 6912 (2002): 206-210.

Kolesnikov, Nikolay, Emma Hastings, Maria Keays, Olga Melnichuk, Y. Amy Tang, Eleanor Williams, Mirosław Dylag et al. "ArrayExpress update—simplifying data submissions." *Nucleic acids research* (2014): gku1057.

Kong SW, Pu WT, Park PJ (2006) A multivariate approach for integrating genomewide expression data and biological knowledge. *Bioinformatics* 22: 2373–2380.

Koonin EV (2005). "Orthologs, paralogs, and evolutionary genomics". *Annual Review of Genetics* 39: 309–38. doi:10.1146/annurev.genet.39.073003.114725. PMID 16285863.

Kornberg , R. D., 2007 The molecular basis of eukaryotic transcription. Proc Natl Acad Sci U S A 104: 12955–12961; LATCHMAN , D. S., 2005 Gene Regulation . BIOS Advanced Text. Taylor & Francis Group, fifth edition.

Kornberg, Roger D. "The molecular basis of eukaryotic transcription." Proceedings of the National Academy of Sciences 104, no. 32 (2007): 12955-12961.

Kotlyar, Max, Stefanie Fuhrman, Alan Ableson, and Roland Somogyi. "Spearman correlation identifies statistically significant gene expression clusters in spinal cord development and injury." Neurochemical research 27, no. 10 (2002): 1133-1140.

Kulikova, Tamara, Philippe Aldebert, Nicola Althorpe, Wendy Baker, Kirsty Bates, Paul Browne, Alexandra van den Broek et al. "The EMBL nucleotide sequence database." Nucleic acids research 32, no. suppl 1 (2004): D27-D30.

Kummerfeld, Sarah K., and Sarah A. Teichmann. "DBD: a transcription factor prediction database." Nucleic acids research 34, no. suppl 1 (2006): D74-D81.

Kiyosawa H, Mise N, Iwase S, Hayashizaki Y, Abe K (2005) Disclosing hidden transcripts: mouse natural sense-antisense transcripts tend to be poly(A) negative and nuclear localized. Genome Res 15(4):463–474

Laetsch, W. M. "The C4 syndrome: a structural analysis." Annual Review of Plant Physiology 25, no. 1 (1974): 27-52..

Langmead, B., Trapnell, C., Pop, M. and Salzberg, S. L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10, 3, R25 ISSN 1465-6914 (Electronic); 1465-6906 (Linking). 30, 31, 32, 33, 89, 90.

Langmead, Ben, and Steven L. Salzberg. "Fast gapped-read alignment with Bowtie 2." Nature methods 9, no. 4 (2012): 357-359.

Lapidot,M. and Pilpel,Y. (2006) Genome-wide natural antisense transcription: coupling its regulation to its different regulatory mechanisms. EMBO Rep., 7, 1216–1222.

Latchman , D. S., 2005 Gene Regulation . BIOS Advanced Text. Taylor & Francis Group, fifth edition.

Lavorgna,G., Dahary,D., Lehner,B., Sorek,R., Sanderson,C.M. and Casari,G. (2004) In search of antisense. Trends Biochem. Sci., 29, 88–94.

Lee , J., K. HE , V. STOLC , H. LEE , P. FIGUEROA , et al. , 2007 Analysis of transcription factor HY5 genomic binding sites revealed its hierarchical role in light regulation of development. *Plant Cell* 19 : 731–749; .

Lehesranta, Satu J., Howard V. Davies, Louise VT Shepherd, Naoise Nunan, Jim W. McNicol, Seppo Auriola, Kaisa M. Koistinen, Soile Suomalainen, Harri I. Kokko, and Sirpa O. Kärenlampi. "Comparison of tuber proteomes of potato varieties, landraces, and genetically modified lines." *Plant Physiology* 138, no. 3 (2005): 1690-1699.

Lembke, Carolina Gimiliani, Milton Yutaka Nishiyama Jr, Paloma Mieko Sato, Rodrigo Fandino de Andrade, and Glaucia Mendes Souza. "Identification of sense and antisense transcripts regulated by drought in sugarcane." *Plant molecular biology* 79, no. 4-5 (2012): 461-477.

Leser U, 2005. A query language for biological networks, ECCB/JBI 2005, *Bioinformatics* 21(Suppl 2):ii33–ii39.

Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, Gnirke A, Regev A. 2010. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods* 7: 709–715.

Li, Bo and Colin N. Dewey. "RSEM: Accurate Transcript Quantification from RNA-Seq Data With or Without a Reference Genome." *BMC Bioinformatics*. 2011.

Li, H. and Durbin, R. (2009b) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 14, 1754–1760 ISSN 1367-4811 (Electronic); 1367-4803 (Linking). 30, 31, 33, 87, 89.

Li, H., Ruan, J. and Durbin, R. (2008a) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 18, 11, 1851–1858 ISSN 1088-9051 (Print); 1088-9051 (Linking).30, 33, 90.

Li, R., Li, Y., Kristiansen, K. and Wang, J. (2008b) SOAP: short oligonucleotide alignment program. *Bioinformatics* 24, 5, 713–714 ISSN 1367-4811 (Electronic); 1367-4803 (Linking). 30, 31.

Li, R., Yu, C., Li, Y., Lam, T.-W., Yiu, S.-M., Kristiansen, K. and Wang, J. (2009a) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* 25, 15, 1966–1967 ISSN 1367-4811 (Electronic); 1367-4803 (Linking). 30, 31, 33.

- Liu, Touming, Siyuan Zhu, Qingming Tang, Yongting Yu, and Shouwei Tang. "Identification of drought stress-responsive transcription factors in ramie (*Boehmeria nivea* L. Gaud)." *BMC plant biology* 13, no. 1 (2013): 130
- Liu S, Zhang Y, Zhou Z, Waldbieser G, Sun F, et al. (2012) Efficient assembly and annotation of the transcriptome of catfish by RNA-Seq analysis of a doubled haploid homozygote. *BMC Genomics* 13: 595.
- Luscombe, Nicholas M., Susan E. Austin, Helen M. Berman, and Janet M. Thornton. "An overview of the structures of protein-DNA complexes." *Genome Biol* 1, no. 1 (2000): 1-37.
- Lysenko, Atem, Matthew Morritt Hindle, Jan Taubert, Mansoor Saqi, and Christopher John Rawlings. "Data integration for plant genomics—exemplars from the integration of *Arabidopsis thaliana* databases." *Briefings in bioinformatics* 10, no. 6 (2009): 676-693.
- M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–29, May 2000.
- M. Kanehisa and S. Goto., 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27-30.
- Ma J, Morrow DJ, Fernandes J, Walbot V (2006) Comparative profiling of the sense and antisense transcriptome of maize lines. *Genome Biol* 7(3):R22.
- Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K.(2014). cluster: Cluster Analysis Basics and Extensions. R package version 1.15.3.
- Magrane, Michele, and UniProt Consortium. "UniProt Knowledgebase: a hub of integrated protein data." *Database* 2011 (2011): bar009.
- Manners J & Casu R (2011). Transcriptome Analysis and Functional Genomics of Sugarcane. *Tropical Plant Biology* 4(1): 9-21.
- Manners, John M., and Rosanne E. Casu. "Transcriptome analysis and functional genomics of sugarcane." *Tropical Plant Biology* 4, no. 1 (2011): 9-21.

Mao, Xizeng, Tao Cai, John G. Olyarchuk, and Liping Wei. "Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary." *Bioinformatics* 21, no. 19 (2005): 3787-3793.

Margulies, Marcel, Michael Egholm, William E. Altman, Said Attiya, Joel S. Bader, Lisa A. Bembien, Jan Berka et al. "Genome sequencing in microfabricated high-density picolitre reactors." *Nature* 437, no. 7057 (2005): 376-380.

Maston, Glenn A., Sara K. Evans, and Michael R. Green. "Transcriptional regulatory elements in the human genome." *Annu. Rev. Genomics Hum. Genet.* 7 (2006): 29-59.

Menossi, M., M. C. Silva-Filho, M. Vincentz, M-A. Van-Sluys, and G. M. Souza. "Sugarcane functional genomics: gene discovery for agronomic trait development." *International Journal of Plant Genomics* 2008 (2007).

Minoru Kanehisa, Susumu Goto, Masahiro Hattori, Kiyoko F Aoki-Kinoshita, Masumi Itoh, Shuichi Kawashima, Toshiaki Katayama, Michihiro Araki, and Mika Hirakawa. From genomics to chemical genomics: new developments in kegg. *Nucleic Acids Res*, 34(Database issue):D354–D357, Jan 2006.

Mitsuda, Nobutaka, and Masaru Ohme-Takagi. "Functional analysis of transcription factors in *Arabidopsis*." *Plant and Cell Physiology* 50, no. 7 (2009): 1232-1248. MOCHIDA, K. AND SHINOZAKI, K. (2010). Genomics and bioinformatics resources for crop improvement, *Plant and Cell Physiology* 51(4): 497–523.

Mochida K, Uehara-Yamaguchi Y, Takahashi F, Yoshida T, Sakurai T, et al. (2013) Large-scale collection and analysis of full-length cDNAs from *Brachypodium distachyon* and integration with Pooidae sequence resources. *PLoS One* 8: e75265

Montgomery, D. C. (1983). *Design and analysis of experiments*. Wiley.

Moore, P. H. "Temporal and spatial regulation of sucrose accumulation in the sugarcane stem." *Functional Plant Biology* 22, no. 4 (1995): 661-679..

Moorthie, Sowmiya, Christopher J. Mattocks, and Caroline F. Wright. "Review of massively parallel DNA sequencing technologies." *The HUGO journal* 5, no. 1-4 (2011): 1-12.

Moriya, Yuki, Masumi Itoh, Shujiro Okuda, Akiyasu C. Yoshizawa, and Minoru Kanehisa. "KAAS: an automatic genome annotation and pathway reconstruction server." *Nucleic acids research* 35, no. suppl 2 (2007): W182-W185.

Mortazavi, Ali, Brian A. Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. "Mapping and quantifying mammalian transcriptomes by RNA-Seq." *Nature methods* 5, no. 7 (2008b): 621-628.

Mortazavi A, Schwarz EM, Williams B, Schaeffer L, Antoshechkin I, et al. (2010) Scaffolding a *Caenorhabditis* nematode genome with RNA-seq. *Genome Research* 20: 1740–1747.

Muñoz, Ivan, Karin Flury, Niels Jungbluth, Giles Rigarlsford, Llorenç Milà i Canals, and Henry King. "Life cycle assessment of bio-based ethanol produced from different agricultural feedstocks." *The International Journal of Life Cycle Assessment* 19, no. 1 (2014): 109-119.

Naughton, Patrick, and Herbert Schildt. *Java: the complete reference*. Osborne/McGraw-Hill, 1996.

Nelson, Melanie R., Stephanie J. Reisinger, and Stephen G. Henry. "Designing databases to store biological information." *Biosilico* 1, no. 4 (2003): 134-142.

Nishiyama Jr, Milton Yutaka, Savio Siqueira Ferreira, Pei-Zhong Tang, Scott Becker, Antje Pörtner-Taliana, and Glaucia Mendes Souza. "Full-Length Enriched cDNA Libraries and ORFeome Analysis of Sugarcane Hybrid and Ancestor Genotypes." *PloS one* 9, no. 9 (2014): e107351 .

Nishiyama-Jr, M. Y., F. F. R. Vicente, C. G. Lembke, P. M. Sato, M. L. Dal-Bianco, R. A. Fandino, C. T. Hotta, and G. M. Souza. "The SUCEST-FUN Regulatory Network Database: designing and energy grass." *International Society of Sugarcane Technologists, Mexico* (2010).

Nishiyama-Jr MY, Vicente F, Sato PM, Ferreira SS, Feltus FA, Souza GM (2013). Transcriptome Analysis in the *Saccharinae*. *Genomics of the Saccharinae*. A. Paterson, Springer. 11: 121-140.

O. Morozova, M. Hirst, and M. A. Marra. Applications of new sequencing technologies for transcriptome analysis. *Annu Rev Genomics Hum Genet*, 10:135{151, 2009.

Ogata, Hiroyuki, Susumu Goto, Kazushige Sato, Wataru Fujibuchi, Hidemasa Bono, and Minoru Kanehisa. "KEGG: Kyoto encyclopedia of genes and genomes." *Nucleic acids research* 27, no. 1 (1999): 29-34.

Oinn, Tom, Matthew Addis, Justin Ferris, Darren Marvin, Martin Senger, Mark Greenwood, Tim Carver et al. "Taverna: a tool for the composition and enactment of bioinformatics workflows." *Bioinformatics* 20, no. 17 (2004): 3045-3054.

ORPHANIDES , G., and D. REINBERG , 2002 A unified theory of gene expression. *Cell* 108 : 439–451 .

Orphanides G, Lagrange T, Reinberg D. 1996. The general transcription factors of RNA polymerase II. *Genes Dev.* 10:2657–83; Latchman, David S. "Transcription factors: an overview." *The international journal of biochemistry & cell biology* 29, no. 12 (1997): 1305-1312.

Orvis, Joshua, Jonathan Crabtree, Kevin Galens, Aaron Gussman, Jason M. Inman, Eduardo Lee, Sreenath Nampally et al. "Ergatis: a web interface and scalable software system for bioinformatics workflows." *Bioinformatics* 26, no. 12 (2010): 1488-1492; .

Östlund, Gabriel, Thomas Schmitt, Kristoffer Forslund, Tina Köstler, David N. Messina, Sanjit Roopra, Oliver Frings, and Erik LL Sonnhammer. "InParanoid 7: new algorithms and tools for eukaryotic orthology analysis." *Nucleic acids research* 38, no. suppl 1 (2010): D196-D203.

Ø. Røsok and M. Sioud, "Systematic identification of sense-antisense transcripts in mammalian cells," *Nature Biotechnology*, vol. 22, no. 1, pp. 104–108, 2004.

Palaniswamy K, James S, Sun H, Lamb R, Davuluri RV, Grotewold E (2006) AGRIS and AtRegNet: A platform to link cis-regulatory elements and transcription factors into regulatory networks. *Plant Physiol* 140: 818–829.

Palaniswamy, Saranyan K., Stephen James, Hao Sun, Rebecca S. Lamb, Ramana V. Davuluri, and Erich Grotewold. "AGRIS and AtRegNet. a platform to link cis-regulatory elements and transcription factors into regulatory networks." *Plant physiology* 140, no. 3 (2006): 818-829.

PanW. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, 18:546–554, 2002.

PanW. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, 18:546–554, 2002.

Papini-terzi F.S., J.M. Felix, F.R. Rocha, A.J. Waclawovsky, E.C. Ulian, S.M. Chabregas, M.C. Falco, M.Y. Nishiyama-Jr, R.Z.N Vêncio, R. Vicentini, M. Menossi and G.M. Souza. The Sucest-Fun Project: Identifying Genes that Regulate Sucrose Content in Sugarcane Plants. Proc. Int. Soc. Sugar Cane Technol., Vol. 26, 2007.

Papini-Terzi, F. S., F. R. Rocha, R. Z. Vencio, K. C. Oliveira, M. Felix Jde, R. Vicentini, S. Rocha Cde, A. C. Simoes, E. C. Ulian, S. M. di Mauro, A. M. da Silva, C. A. Pereira, M. Menossi, and G. M. Souza. 2005. Transcription profiling of signal transduction-related genes in sugarcane tissues. DNA Res 12 (1):27-38.

Passalacqua KD, Varadarajan A, Weist C, Ondov BD, Byrd B, Read TD, Bergman NH. 2012. Strand-specific RNA-seq reveals ordered patterns of sense and antisense transcription in *Bacillus anthracis*. PLoS ONE 7: e43350.

Paterson, Andrew H., John E. Bowers, Rémy Bruggmann, Inna Dubchak, Jane Grimwood, Heidrun Gundlach, Georg Haberer et al. "The *Sorghum bicolor* genome and the diversification of grasses." *Nature* 457, no. 7229 (2009): 551-556.

Paterson AH, Moore PH, Tew TL (2013) The Gene Pool of *Saccharum* species and their Improvement. In: Paterson A, editor. *Genomics of the Saccharinae*: Springer. pp. 43–72

Pavlidis P, Qin J, Arango V, Mann J, Sibille E (2004) Using the Gene Ontology for microarray data mining: A comparison of methods and application to age effects in human prefrontal cortex. *Neurochem Res* 29: 1213–1222.

Peaeson, E., and H. Haetlet. "Biometrika tables for statisticians." Biometrika Trust (1976).

Petersen, David, G. V. R. Chandramouli, Joel Geoghegan, Joanne Hilburn, Jonathon Paarlberg, Chang H. Kim, David Munroe et al. "Three microarray platforms: an analysis of their concordance in profiling gene expression." *Bmc Genomics* 6, no. 1 (2005): 63].

Pevzner, P. A., Tang, H. and Waterman, M. S. (2001) An Eulerian path approach to DNA fragment assembly Proc Natl Acad Sci U S A 98, 17, 9748–53. 28.

Quackenbush J. Computational analysis of microarray data. *Nat Rev Genet*, 2:418–427, 2001.

Quackenbush J. Data standards for 'omic' science. *Nat Biotechnol*, 22:613–614, 2004.

Quail MA, Smith M, Coupland P et al. (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genom* 13:341.

Quevillon, Emmanuel, Ville Silventoinen, Sharmila Pillai, Nicola Harte, N. Mulder, Rolf Apweiler, and Rodrigo Lopez. "InterProScan: protein domains identifier." *Nucleic acids research* 33, no. suppl 2 (2005): W116-W120.

Ralph SG, Chun HJE, Cooper D, Kirkpatrick R, Kolosova N, et al. (2008) Analysis of 4,664 high-quality sequence-finished poplar full-length cDNA clones and their utility for the discovery of genes responding to insect feeding. *BMC Genomics* 9: 57.

Riaño-Pachón, Diego M., Slobodan Ruzicic, Ingo Dreyer, and Bernd Mueller-Roeber. "PlnTFDB: an integrative plant transcription factor database." *BMC bioinformatics* 8, no. 1 (2007): 42.

Riechmann JL, Heard J, Martin G, Reuber L, Jiang C, Keddie J, Adam L, Pineda O, Ratcliffe OJ, Samaha RR, Creelman R, Pilgrim M, Broun P, Zhang JZ, Ghandehari D, Sherman BK, Yu G (2000a) Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. *Science* 290: 2105-2110.

Riechmann JL, Ratcliffe OJ (2000b) A genomic perspective on plant transcription factors. *Curr Op Plant Biol* 3: 423-434.

Rielle, Ludivine. "To develop matrices for ESTScan." (2007).

Ripley, Brian D. "The R project in statistical computing." *MSOR Connections* 1, no. 1 (2001): 23-25.

Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome biology*. 2010b; 11:R25.

Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." *Bioinformatics* 26, no. 1 (2010a): 139-140.

Robinson, Webber, Eifrem. *Graph Databases*. 2013, O'Reilly Media. <http://graphdatabases.com/>.

Rocha, F. R.; F. S. Papini-Terzi; M. Y. Nishiyama Jr.; R. Z. Vencio; R. Vicentini; R. D. Duarte; V. E. de Rosa Jr.; F. Vinagre; C. Barsalobres; A. H. Medeiros; F. A. Rodrigues; E. C. Ulian; S. M. Zingaretti; J. A. Galbiatti; R. S. Almeida; A. V. Figueira; A. S. Hemerly;

M. C. Silva-Filho; M. Menossi and G. M. Souza. 2007. Signal transduction-related responses to phytohormones and environmental challenges in sugarcane. *BMC Genomics* 8:71.

Rodriguez, M.A., Neubauer, P., "The Graph Traversal Pattern ," *Graph Data Management: Techniques and Applications*, eds. S. Sakr, E. Pardede, IGI Global, ISBN:9781613500538, August 2011.

Rodriguez, Neubauer. Constructions from dots and lines. *Bulletin of the American Society for Information Science and Technology*, 2010, Pages 35–41.

Rouillard, Jean-Marie, Michael Zuker, and Erdogan Gulari. "OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach." *Nucleic acids research* 31, no. 12 (2003): 3057-3062.

Roy, S., Ernst, J., Kharchenko, P.V., Kheradpour, P., Negre, N., Eaton, M.L., Landolin, J.M., Bristow, C.A., Ma, L., Lin, M.F., et al. (2010). Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* 330, 1787-1797.

Ruepp, Andreas, Alfred Zollner, Dieter Maier, Kaj Albermann, Jean Hani, Martin Mokrejs, Igor Tetko et al. "The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes." *Nucleic acids research* 32, no. 18 (2004): 5539-5545.

Rumble, S. M., Lacroute, P., Dalca, A. V., Fiume, M., Sidow, A. and Brudno, M. (2009) SHRiMP: accurate mapping of short color-space reads. *PLoS Comput Biol* 5, 5, e1000386 ISSN 1553-7358 (Electronic). 30.

Rusk, Nicole. "Cheap third-generation sequencing." *Nature Methods* 6, no. 4 (2009): 244-244.

S Tomiuk and K Hofmann. Microarray probe selection strategies. *Briefings in Bioinformatics*,2(4):329{40, December 2001.

S. Draghici. *Data Analysis Tools for DNA Microarrays*. Chapman Hall/CRC, 3edition, 6 2003.

Sage, Rowan; Russell Monson (1999). "7". *C4 Plant Biology*. pp. 228–229.

Sakr, Sherif, Sameh Elnikety, and Yuxiong He. "G-SPARQL: a hybrid engine for querying large attributed graphs." In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 335-344. ACM, 2012.

Sayers, E. W., Barrett, T., Benson, D. A., Bolton, E., Bryant, S. H., Canese, K. et al. (2010). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 38, D5–D16.

Schena, Mark, Dari Shalon, Ronald W. Davis, and Patrick O. Brown. "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." *Science* 270, no. 5235 (1995): 467-470.

Schmid, Ralf, and Mark L. Blaxter. "annot8r: GO, EC and KEGG annotation of EST datasets." *BMC bioinformatics* 9, no. 1 (2008): 180.

Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27: 863–864.

Schreiber, Fabian, and Erik LL Sonnhammer. "Hieranoid: hierarchical orthology inference." *Journal of molecular biology* 425, no. 11 (2013): 2072-2081.

Schwechheimer, Claus, and Michael Bevan. "The regulation of transcription factor activity in plants." *Trends in Plant Science* 3, no. 10 (1998): 378-383.

Shendure, J. and Ji, H. (2008) Next-generation DNA sequencing. *Nat Biotechnol* 26, 10, 1135–1145 ISSN 1546-1696 (Electronic); 1087-0156 (Linking). 8, 13.

Shendure, Jay, and Hanlee Ji. "Next-generation DNA sequencing." *Nature biotechnology* 26, no. 10 (2008): 1135-1145..

Shoemaker, B. A. and Panchenko, A. R. (2007). Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. *PLoS Comput Biol* , 3(4):e43.

Sigurgeirsson, Benjamín, Olof Emanuelsson, and Joakim Lundeberg. "Analysis of stranded information using an automated procedure for strand specific RNA sequencing." *BMC genomics* 15, no. 1 (2014): 631.

Silverman BW (1986) *Density Estimation*. Chapman and Hall, London, UK.

Simpson, Jared T., Kim Wong, Shaun D. Jackman, Jacqueline E. Schein, Steven JM Jones, and Inanç Birol. "ABYSS: a parallel assembler for short read sequence data." *Genome research* 19, no. 6 (2009): 1117-1123.

- Slonim, Donna K. "From patterns to pathways: gene expression data analysis comes of age." *Nature genetics* 32 (2002): 502-508.
- Smith, T. and Waterman, M. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, 147, 195–197.
- Sokal R.R., F.J. Rohlf. 1995. *Biometry: The Principles and Practice of Statistics in Biological Research*. W.H. Freeman and Co., New York, Third Edition.
- Souza AP ; GASPAR, A. ; Silva, E.A. ; ULIAN, E. C. ; Waclawovsky AJ ; Nishiyama-Junior, M. Y. ; VICENTINI, R. ; MENOSSI, M. ; SOUZA, G. M. ; Buckeridge M . Elevated CO₂ induces increases in photosynthesis, biomass, productivity and modifies gene expression in sugarcane. *Plant, Cell and Environment*, v. 31, p. 1116-1127, 2008.
- Souza, G.M., Simoes, A.C.Q., Oliveira, K.C., Garay, H.M., Fiorini, L.C., Gomes, F.S., Nishiyama-Junior, M.Y., da Silva, A.M. (2001). The sugarcane signal transduction (SUCAST) catalogue: prospecting signal transduction in sugarcane. *Genet. Mol. Biol.* 24:25-34.
- Sroka, Jacek, Łukasz Bieniasz-Krzywiec, Szymon Gwóźdź, Dariusz Leniowski, Jakub Łacki, Mateusz Markowski, Claudio Avignone-Rossa, Michael E. Bushell, John Joe McFadden, and Andrzej M. Kierzek. "Acorn: A grid computing system for constraint based modeling and visualization of the genome scale metabolic reaction networks via a web interface." *BMC bioinformatics* 12, no. 1 (2011): 196.
- Stegmaier , P., A. E. KEL , and E. WINGENDER , 2004 Systematic DNA-binding domain classification of transcription factors. *Genome Inform* 15 : 276–286.
- Stein L. (2002) Creating a bioinformatics nation *Nature* 417 , 119–120. 36.
- Stein, Lincoln D. "Integrating biological databases." *Nature Reviews Genetics* 4, no. 5 (2003): 337-345.
- Stromberg, M. and Marth, G. (2008) MOSAIK: A reference-guided assembler for next-generation sequence data Manuscript in preparation. 30, 31, 85, 86.
- Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102: 15545–15550.

Tateno, Yoshio, Tadashi Imanishi, Satoru Miyazaki, Kaoru Fukami-Kobayashi, Naruya Saitou, Hideaki Sugawara, and Takashi Gojobori. "DNA Data Bank of Japan (DDBJ) for genome scale research in life science." *Nucleic acids research* 30, no. 1 (2002): 27-30.

Tatusov, R., Koonin, E. and Lipman, D. (1997) A genomic perspective on protein families. *Science*, 278, 631-637; Moriya, Yuki, Masumi Itoh, Shujiro Okuda, Akiyasu C. Yoshizawa, and Minoru Kanehisa. "KAAS: an automatic genome annotation and pathway reconstruction server." *Nucleic acids research* 35, no. suppl 2 (2007): W182-W185.

Tezara, W., V. J. Mitchell, S. D. Driscoll, and D. W. Lawlor. "Water stress inhibits plant photosynthesis by decreasing coupling factor and ATP." *Nature* 401, no. 6756 (1999): 914-917.

Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, et al. (2005) Discovering statistically significant pathways in expression profiling studies. *Proc Natl Acad Sci U S A* 102: 13544-13549.

Tian, Yuanyuan, Jignesh M. Patel, Viji Nair, Sebastian Martini, and Matthias Kretzler. "Periscope/gq: a graph querying toolkit." *Proceedings of the VLDB Endowment* 1, no. 2 (2008): 1404-1407.

Tipton, Keith, and Sinéad Boyce. "History of the enzyme nomenclature system." *Bioinformatics* 16, no. 1 (2000): 34-40.

Trapnell, C. and Salzberg, S. L. (2009) How to map billions of short reads onto genomes. *Nat Biotechnol* 27, 5, 455-457 ISSN 1546-1696 (Electronic); 1087-0156 (Linking). 29, 104.

Trapnell, Cole, Brian A. Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J. van Baren, Steven L. Salzberg, Barbara J. Wold, and Lior Pachter. "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation." *Nature biotechnology* 28, no. 5 (2010): 511-515.

Velculescu, Victor E., Lin Zhang, Bert Vogelstein, and Kenneth W. Kinzler. "Serial analysis of gene expression." *Science* 270, no. 5235 (1995): 484-487.

Vêncio, Ricardo ZN, and Tie Koide. "HTself: self-self based statistical test for low replication microarray studies." *DNA Research* 12, no. 3 (2005): 211-214.

Venkatesh, T. V., and Harry B. Harlow. "Integromics: challenges in data integration." *Genome Biol* 3, no. 8 (2002): 1-3..

Vettore, André L., Felipe R. da Silva, Edson L. Kemper, and Paulo Arruda. "The libraries that made SUCEST." *Genetics and Molecular Biology* 24, no. 1-4 (2001): 1-7.

Vettore, André L., Felipe R. da Silva, Edson L. Kemper, Glaucia M. Souza, Aline M. Da Silva, Maria Inês T. Ferro, Flavio Henrique-Silva et al. "Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane." *Genome research* 13, no. 12 (2003): 2725-2735.

Von Caemmerer, S.; Quick, W. P.; Furbank, R. T. (2012). "The Development of C4 Rice: Current Progress and Future Challenges". *Science* 336 (6089): 1671–1672. doi:10.1126/science.1220177.

W.S. Cleveland and S.J. Devlin. Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association*, 83(403):596_610, 1988.

W.S. Cleveland. Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association*, 74(368):829_836, 1979.

Waclawovsky, Alessandro J., Paloma M. Sato, Carolina G. Lembke, Paul H. Moore, and Glaucia M. Souza. "Sugarcane for bioenergy production: an assessment of yield and regulation of sucrose content." *Plant Biotechnology Journal* 8, no. 3 (2010): 263-276

Wagner, E.G. & Simons, R.W. Antisense RNA control in bacteria, phages, and plasmids. *Annu. Rev. Microbiol.* 48, 713–742 (1994).

Wajid, B. & Serpedin, E. (2012). Review of general algorithmic features for genome assemblers for next generation sequencers. *Genomics, Proteomics & Bioinformatics*, 10, 58{73. 26.

Wang H, Chung PJ, Liu J, Jang IC, Kean MJ, et al. (2014) Genome-wide identification of long noncoding natural antisense transcripts and their responses to light in *Arabidopsis*. *Genome Research*.

Wang XJ, Gaasterland T, Chua NH. 2005. Genome-wide prediction and identification of cis-natural antisense transcripts in *Arabidopsis thaliana*. *Genome Biol* 6: R30.

Wang, Zhong, Mark Gerstein, and Michael Snyder. "RNA-Seq: a revolutionary tool for transcriptomics." *Nature Reviews Genetics* 10, no. 1 (2009): 57-63.

Wei H., Persson S., Mehta T., Srinivasasainagendra V., Chen L., Page G.P., Somerville C. & Loraine A. (2006) Transcriptional coordination of the metabolic network in *Arabidopsis*. *Plant Physiology* 142, 762–774.

Wheeler, David L., Tanya Barrett, Dennis A. Benson, Stephen H. Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M. Church et al. "Database resources of the national center for biotechnology information." *Nucleic acids research* 35, no. suppl 1 (2007): D5-D12.

Whetzel, Patricia L., Helen Parkinson, Helen C. Causton, Liju Fan, Jennifer Fostel, Gilberto Fragoso, Laurence Game et al. "The MGED Ontology: a resource for semantics-based description of microarray experiments." *Bioinformatics* 22, no. 7 (2006): 866-873.

Wilcoxon, Frank, S. K. Katti, and Roberta A. Wilcox. "Critical values and probability levels for the Wilcoxon rank sum test and the Wilcoxon signed rank test." *Selected tables in mathematical statistics* 1 (1970): 171-259.

Wingender, E. "Classification scheme of eukaryotic transcription factors." *Molecular Biology* 31, no. 4 (1997): 483-497.

Wren, Jonathan D., Amit Kulkarni, John Joslin, Ronald A. Butow, and R. G. Harold. "Cross-hybridization on PCR-spotted microarrays." *Engineering in Medicine and Biology Magazine, IEEE* 21, no. 2 (2002): 71-75.

Xu, Jianzhen, and Yongjin Li. "Discovering disease-genes by topological features in human protein–protein interaction network." *Bioinformatics* 22, no. 22 (2006): 2800-2805.

X Wang and B Seed. Selection of oligonucleotide probes for protein coding sequences. *Bioinformatics* , 19(7):796–802, May 2003.

Yang H, Sunderraman R, Tian H, 2008. bcnQL: A query language for biochemical networks, *IEEE International Conference on Bioinformatics and Biomedicine*.

YANG, Y.H.; DUDOIT, S.; LUU, P.; LIN, D.M.; PENG, V.; NGAI, J.; SPEED, T.P. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*, v. 30, n. 4, p. e15, 2002.

Yao, Lu-Ming, Biao Wang, Lin-Jing Cheng, and Tian-Long Wu. "Identification of key drought stress-related genes in the hyacinth bean." *PloS one* 8, no. 3 (2013): e58108.

- Yilmaz, Alper, Milton Y. Nishiyama, Bernardo Garcia Fuentes, Glaucia Mendes Souza, Daniel Janies, John Gray, and Erich Grotewold. "GRASSIUS: a platform for comparative regulatory genomics across the grasses." *Plant physiology* 149, no. 1 (2009): 171-180.
- Yu X, Yang J, Li X, Liu X, Sun C, et al. (2013) Global analysis of cis-natural antisense transcripts and their heat-responsive nat-siRNAs in *Brassica rapa*. *BMC Plant Biol* 13: 208.
- Zerbino, D. R. and Birney, E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 18, 5, 821–829 ISSN 1088-9051 (Print). 28, 29, 46, 56, 64, 84, 142.
- Zhang, Zhengdong D., Alberto Paccanaro, Yutao Fu, Sherman Weissman, Zhiping Weng, Joseph Chang, Michael Snyder, and Mark B. Gerstein. "Statistical analysis of the genomic distribution and correlation of regulatory elements in the ENCODE regions." *Genome research* 17, no. 6 (2007): 787-797.
- Zhang, Xiaoming, Yifan Lii, Zhigang Wu, Anton Polishko, Huiming Zhang, Viswanathan Chinnusamy, Stefano Lonardi, Jian-Kang Zhu, Renyi Liu, and Hailing Jin. "Mechanisms of small RNA generation from cis-NATs in response to environmental and developmental cues." *Molecular plant* 6, no. 3 (2013): 704-715.
- Zhang, Xiaoyu, Ian R. Henderson, Cheng Lu, Pamela J. Green, and Steven E. Jacobsen. "Role of RNA polymerase IV in plant small RNA metabolism." *Proceedings of the National Academy of Sciences* 104, no. 11 (2007): 4536-4541.
- Zhou X, Sunkar R, Jin H, Zhu JK, Zhang W (2008) Genome-wide identification and analysis of small RNAs originated from natural antisense transcripts in *Oryza sativa*. *Genome Research* 19: 70–78.

9. Anexos

Anexo I. Tabela de Categorização Funcional para Transcriptoma de Cana-de-açúcar

Categoria Funcional	Categoria Estrutural	Descrição
1. Carbohydrate metabolism		
	Glycolysis	
	Krebs Cycle	
	Gluconeogenesis	
	Pentose Phosphate pathway	
	Synthesis and degradation of starch	
	Fructose/Mannose/Galactose	
	Calvin Cycle	
	Inositol metabolism	
2. Light Harvesting	Photochemical reaction centers (PSI e PSII)	
3. Oxidative phosphorylation		
4. Lipid metabolism		
	Fatty acid biosynthesis	
	Fatty acid elongation in mitochondria	
	Fatty acid metabolism	
	Synthesis and degradation of ketone bodies	
	Steroid hormone biosynthesis	
	Glycerolipid metabolism	
	Glycerophospholipid metabolism	
	Ether lipid metabolism	
	Sphingolipid metabolism	
	Arachidonic acid metabolism	
	alpha-Linolenic acid metabolism	
	Biosynthesis of unsaturated fatty acids	
5. Nucleotide metabolism		
	Purine metabolism	
	Pyrimidine metabolism	
6. AminoAcid and nitrogen metabolism		
	Alanine, aspartate and glutamate metabolism	
	Glycine, serine and threonine metabolism	
	Cysteine and methionine metabolism	
	Valine, leucine and isoleucine degradation	
	Valine, leucine and isoleucine biosynthesis	
	Lysine biosynthesis	
	Lysine degradation	
	Arginine and proline metabolism	

	Histidine metabolism	
	Tyrosine metabolism	
	Phenylalanine metabolism	
	Tryptophan metabolism	
	Phenylalanine, tyrosine and tryptophan biosynthesis	
	beta-Alanine metabolism	
	Taurine and hypotaurine metabolism	
	Selenoamino acid metabolism	
	Cyanoamino acid metabolism	
	Glutathione metabolism	
	Nitrogen assimilation pathway	
7. Cell wall metabolism		
	Polysaccharide biosynthesis	
	Glycosylphosphatidylinositol(GPI)-anchor biosynthesis	
	Glycosphingolipid biosynthesis - globo series	
	Glycosphingolipid biosynthesis - ganglio series	
	Other glycan degradation	
	Lignin metabolism	
	Suberin metabolism	
	Phenylpropanoid biosynthesis	
	Structural proteins of the cell wall	
	Cell wall growth/extension	
	cell wall differentiation	
8. Vitamin and Cofactor metabolism		
	Thiamine metabolism	
	Riboflavin metabolism	
	Vitamin B6 metabolism	
	Nicotinate and nicotinamide metabolism	
	Pantothenate and CoA biosynthesis	
	Biotin metabolism	
	Lipoic acid metabolism	
	Folate biosynthesis	
	One carbon pool by folate	
	Retinol metabolism	
	Ubiquinone and other terpenoid-quinone biosynthesis	
9. Phorphyrin and chlorophyll metabolism		
	phorphyrin and chlorophyll metabolism	
10. Hormone biosynthesis		
	Auxin synthesis and degradation	tryptophane metabolism
	ABA synthesis and degradation	carotenoid byosynthesis
	Ethylene synthesis and degradation	
	Cytokinin synthesis and degradation	zeatin biosynthesis

	Brassinosteroid synthesis and degradation	Brassinosteroid biosynthesis
	Gibberellin synthesis and degradation	diterpenoid biosynthesis
	Salicylic acid synthesis and degradation	
	Methyl jasmonate	
11. Carotenoid Metabolism		
	carotenoid biosynthesis	
12. Other terpenoids metabolism		
	Terpenoid backbone biosynthesis	
	Monoterpenoid biosynthesis	
	Sesquiterpenoid biosynthesis	
	Limonene and pinene degradation	
	Polyketide sugar unit biosynthesis	
13. Flavonoid and anthocyanin metabolism		
	Isoflavonoid biosynthesis	
	Anthocyanin biosynthesis	
	Flavonoid biosynthesis	
14. Biosynthesis of other Secondary Metabolites		
	Stilbenoid, diarylheptanoid and gingerol biosynthesis	
	Flavone and flavonol biosynthesis	
	Indole alkaloid biosynthesis	
	Isoquinoline alkaloid biosynthesis	
	Tropane, piperidine and pyridine alkaloid biosynthesis	
	Acridone alkaloid biosynthesis	
	Caffeine metabolism	
	Betalain biosynthesis	
	Glucosinolate biosynthesis	
	Benzoxazinoid biosynthesis	
15. Sulphur metabolism		
16. DNA metabolism		
	Replication/repair	
	Histones and associated proteins	
	Methylation and acetylation	
	Telomere maintenance	
17. RNA metabolism		
	Transcription	
	RNA binding proteins	
18. Protein Metabolism		

	Translation	
	Chaperones	
	Degradation	
	Modification	
19. Signal Transduction	Biotic stress signal transduction	ABA
		Drought
		Cold
		Light
		Heat
	Abiotic stress signal transduction	Jasmonic Acid
		Salicylates
	Internal signal transduction	Auxin
		Cytokinin
		Ethylene
		Brassinosteroid
	Signal transduction - general	Some Kinases and phosphates
		Some Transcription Factors
		Calcium signaling
20. Cell cycle		
	Meiosis	
	Mitosis	
	Cell cycle control	
21. Circadian Clock		
22. Development		
	Differentiation	
	Speciation	
23. Transport		
	Simporters	
	Antiporters	
	Aquaporin	
	Membrane transporters	
	pumps	
	channels	
	Carriers (catalyze the movement of inorganic ions and simple organic solutes across membranes)	
24. Senescence and cell death		
25. Cytoskeleton and vesicle trafficking		
	Secretory pathways	

	Endocytic pathway	
	Protein sorting	
26. Pathogen Resistance		
27. Redox Metabolism		
28. Transposable element-related		
	transposase	
	reverse transcriptase	
29. Others		
30. Unknown		

Anexo II. Número de transcritos únicos baseados no 1º hit contra os genes full-length de gramíneas com cobertura >95% e identidade >80%. Espécies: Sb: *S. bicolor*; Zm: *Z. mays*; Pv: *P. virgatum*; Si: *S. italica*; Os: *O. sativa*; Bd: *B. distachyon*.

Grupos de Espécies	Número de transcritos únicos por grupo de espécies
Zm	1225
Sb	1150
Pv	939
Sb + Zm	814
Sb + Si + Zm	375
Bd + Os + Pv + Sb + Si + Zm	373
Pv + Sb + Zm	367
Os + Pv + Sb + Si + Zm	326
Pv + Sb + Si + Zm	324
Si	279
Pv + Sb	278
Os + Sb + Si + Zm	218
Sb + Si	211
Os + Sb + Zm	201
Os	191
Pv + Sb + Si	161
Pv + Zm	161
Os + Pv + Sb + Zm	147

Bd + Os + Sb + Si + Zm	138
Os + Sb	136
Bd + Pv + Sb + Si + Zm	124
Bd + Os + Sb + Zm	120
Bd + Os + Pv + Sb + Zm	117
Os + Pv + Sb + Si	105
Bd + Os + Pv + Sb + Si	99
Bd + Sb + Zm	91
Os + Sb + Si	89
Bd + Pv + Sb + Zm	85
Si + Zm	78
Os + Pv + Sb	77
Bd + Sb + Si + Zm	73
Bd + Sb	72
Bd	64
Pv + Si	62
Os + Zm	52
Bd + Os + Sb	47
Bd + Os + Sb + Si	42
Bd + Os + Pv + Sb	41
Bd + Pv + Sb	39
Bd + Pv + Sb + Si	38
Bd + Sb + Si	37
Pv + Si + Zm	35
Os + Pv	34
Os + Si + Zm	31
Os + Pv + Si + Zm	31
Bd + Os + Pv + Si + Zm	29
Os + Pv + Zm	26
Os + Si	23
Bd + Zm	22
Os + Pv + Si	17
Bd + Os + Pv + Zm	17
Bd + Pv + Si + Zm	16
Bd + Pv	16
Bd + Os	15
Bd + Os + Si + Zm	13
Bd + Os + Si	12
Bd + Os + Pv + Si	11

Bd + Si	10
Bd + Pv + Zm	10
Bd + Os + Zm	10
Bd + Si + Zm	7
Bd + Pv + Si	6
Bd + Os + Pv	3

Anexo III. Comparativo da lista de Famílias de TFs nos Bancos de Dados GRASSTFDB, PantTFDB, PlnTFDB e Regras domínios permitidos, proibidos e possíveis para a classificação das famílias de TFs.

GRASSTFDB	PlantTFDB	PlnTFDB	Domínios	Proibidos	PFAMID domínios	Pode possuir domínio	PFAMID proibidos
AB13/VP1	AB13/VP1	ABI3VP1	B3	AP2,Auxin_resp	PF02362		PF00847,PF06507
Alfin-like	Alfin-like	Alfin-like	Alfin_like	DDT,Homeobox,JmjC,JmjN	Alfin-like	PHD, SNF2_N, zf-C2H2, zf-C5HC2, zf-CCCH,zf-TAZ	PF02791,PF00046,PF02373,PF02375
AP2/EREBP	AP2/EREBP	AP2-EREBP	AP2	B3,SNF2_N	PF00847	B3(PF02362),AUX/IAA(PF00176)	
ARF	ARF	ARF	Auxin_resp, AUX_IAA, B3		PF06507	SNF2_N(PF02309),B3(PF02362)	
ARID	ARID	ARID	ARID		PF01388	HMG_box,JmjC,JmjN,Myb_DNA-binding,PHD	
ARR-B	GARP-ARR-B	ARR-B	G2_like,Myb_DNA-binding, Response_reg	CCT	G2-like,PF00249,PF00072		PF06203
BBR-BPC	BBR-BPC	BBR/BPC	DUF1004		PF06217		
bHLH	bHLH	bHLH	HLH		PF00010		
bZIP	bZIP	bZIP	bZIP_1, bZIP_2	HLH	PF00170,PF07716		PF00010
BZR	BES1	BES1	DUF822		PF05687		
C2C2-CO-like	C2C2-CO-like	C2C2-CO-like	CCT, zf-B_box		PF06203,PF00643		
C2C2-Dof	C2C2-Dof	C2C2-Dof	zf-Dof		PF02701		
C2C2-GATA	C2C2-GATA	C2C2-GATA	GATA		PF00320	CCT,FAR1,ZIM	
C2C2-YABBY	C2C2-YABBY	C2C2-YABBY	YABBY		PF04690	HMG_Box	
C2H2	C2H2	C2H2	zf-C2H2	JmjC,JmjN,PHD	PF00096	FHA	PF02373,PF02375,PF00628
C3H	C3H	C3H	zf-CCCH	PHD, SNF2_N, zf-C2H2	PF00642		PF00628,PF00176,PF00096
CAMTA	CAMTA	CAMTA	CG-1,IQ		PF03859,PF00612	Ank	

CCAAT-Dr1	BBR-BPC	CCAAT	CCAAT-Dr1		PF00808		
CCAAT-HAP2	CCAAT-HAP2	CCAAT	CBFB_NFYA,CBFD_NFY B_HMF,NF-YC		PF02045		
CCAAT-HAP3	CCAAT-HAP3	CCAAT	CBFB_NFYA,CBFD_NFY B_HMF,NF-YC		PF00808		
CCAAT-HAP5	CCAAT-HAP5	CCAAT	CBFB_NFYA,CBFD_NFY B_HMF,NF-YC		PF00808		
CPP	CPP	CPP	CXC	PcG	PF03638		PF00856
CSD		CSD	CSD, zf-CCHC		PF00313, PF00098		PF01388,PF04690
DBP		DBP	DNC, PP2C		DNC,PF00481		
E2F/DP	E2F/DP	E2F-DP	E2F_TDP		PF02319		
EIL	EIL	EIL	EIN3		PF04873		
FAR1-like			PF03101,PF10551,PF04 434				
FLO/LFY	LFY	LFY	FLO_LFY		PF01698		
G2-like	GARP-G2-like	G2-like	G2-like		G2-like		
GeBP	GeBP	GeBP	DUF573		PF04504		
GRAS	GRAS	GRAS	GRAS		PF03514		
GRF	GRF	GRF	GRF		GRF		
Homeobox	HB	HB	Homeobox,KNOX1,KNO X2		PF00046	ELK,HALZ,HD-ZIP_N,PHD	
HRT	HRT-like	HRT	HRT		HRT		PD718761,PD185086
HSF	HSF	HSF	HSF_DNA-binding		PF00447		
LBD	AS2		DUF260		PF03195		
MADS	MADS	MADS	SRF-TF		PF00319	K-box	
mTERF			PF02536				
MYB	MYB	MYB	Myb_DNA-binding	ARID, Response_reg, SNF2_N	PF00249		PF01388, PF00072, PF00176

MYB-related	MYB-related	MYB-related	Myb_DNA-binding	ARID, G2-like, Response_reg, SNF2_N	PF00249	zf-CCHC	G2-like,PF01388, PF00072, PF00176
NAC	NAC	NAC	NAM		PF02365		
NLP	Nin-Like	RWP-RK	RWP-RK		PF02042		
Ovate			PF04844				
PLATZ	PLATZ	PLATZ	PLATZ		PF04640		
S1Fa-like	S1Fa-like	S1Fa-like	S1FA		PF04689		
SBP	SBP	SBP	SBP		PF03110		
SRS	SRS	SRS	DUF702		PF05142		
TCP	TCP	TCP	TCP		PF03634		
Trihelix	Trihelix	Trihelix	trihelix		trihelix		
TUB	TLP	TUB	Tub		PF01167		
VOZ-9	VOZ	VOZ	VOZ		VOZ		
Whirly	PBF-2- like/Whirly	PBF-2-like	Whirly		PF08536		
WRKY	WRKY	WRKY	WRKY		PF03106	zf-BED	
zf-HD	zf-HD	zf-HD	ZF-HD_dimer		PF04770		
ZIM	ZIM	ZIM	Zim	GATA	PF06200		PF00320
	AUX/IAA	AUX/IAA	AUX_IAA	Auxin_resp	PF02309		PF06507
	FHA	FHA	FHA	zf-C2H2	PF00498	PP2C	PF00096
	JUMONJI	Jumonji	JmjC,JmjN, zf-C2H2, zf- C5HC2		PF02373,PF02375,PF0009 6,PF02928		
	MBF1	MBF1	MBF1		PF08523		
	PHD	PHD	PHD	DDT, Homeobox,JmjC,JmjN	PF00628	HMG_box,SET	PF02791,PF00046,PF02373, PF02375
					CCAAT-Dr1		
	HMG	HMG	HMG_Box	ARID, YABBY	PF00505		
	LIM	LIM	LIM		PF00412	zf-CCHC	PF00628, PF00096

	LUG	LUG	LUFS		LUFS		PF00628
	PcG	SET	SET	PHD, zf-C2H2	PF00856		
	TAZ	TAZ	zf-TAZ	PHD	PF02135		
	ULT	ULT	ULT		ULT		
	NZZ	NOZZLE	NOZZLE		PF08744		
		bHSH	TF_AP-2		PF03299	PHD	
		DDT	DDT		PF02791		
		Pseudo ARR-B	CCT, Response_reg		PF06203, PF00072		
		RB	RB_B		PF01857		
		SAP	STER_AP		STER_AP		
		Sigma70-like	Sigma70_r2,Sigma70_r3, Sigma70_r4		PF04542,PF04539,PF0454 5	Myb_DNA-binding	PF00847,PF00628
		SNF2	SNF2_N	AP2,PHD	PF00176		
	GIF		SSXT		PF05030		

Anexo IV. Dicionário do banco de de Dados CaneRegNetDB

Database				
Database Description	CaneRegNetDB - Sugarcane Database for Regulatory Networks			
Database Name	sucestfundb_dvlp			
Tables				
1 array				
Table Description	Information about the used array, technical details about the platform			
Column Name	Data Type	Not Null	Default Value	Comment
id	int(10)	YES		Unique identifier for each array. MySQL sequence generated number

label	varchar(64)	YES		Name of the chemical used to label the sample, e. g., fluorescent labelling with cyanine fluorochromes (Cy3/Cy5) or radioactive tracing with radioisotopes
manufacturer	varchar(64)	NO		Name of the company which produces the arrays, e. g., Agilent, Affymetrix (or simply Name of the manufacturer of arrays)
platform	varchar(64)	NO		Type of platform the array is based on, such as spotted or in situ synthesized
surface	varchar(64)	NO		type of surface where the array elements are bound, e. g., glass or nylon membrane
identification_code	varchar(64)	NO		laboratory general code to identify the array
description	text	NO		general description about the array platform
update_date	timestamp(14)	YES	CURRENT_TIMESTAMP	Date the array information was performed/updated
2 array_design				
Table Description		Contains Microarray chip cDNA and Probe Set information. Describes which probe and probe type in a probe set a particular genome		
Column Name	Data Type	Not Null	Default Value	Comment
id	int(10)	YES		Unique identifier for each array design. MySQL sequence generated number
array_id	int(10)	YES		Unique identifier for each array. Link to the array table.
probe_name	varchar(255)	YES		original probe name in the design of the array. In the cases of cDNA clones after the sequencing the name can change and will be identified in the correct probe
bad_spot	int(2)	NO		probe identification name. The name can be linked to the entities: oligo_probe and transcript represented by unique names
feature_num	int(5)	NO		Index of feature location in Agilent arrays or positional information for features in sectors from manufactured arrays
feature_row	int(5)	NO		Row of feature within the entire array, increasing from top to bottom. 1-based counting
feature_column	int(5)	NO		Column of feature within the entire array, increasing from left to right. 1-based counting
plate	int(50)	NO		unique identifier for each plate 96/384 wells
position	varchar(5)	NO		positional information for probes in the source plate
correct_probe_name	varchar(255)	NO		correct probe name after the re-sequencing of cDNA clones or any other reason to change the original name

unknown_probe	varchar(255)	NO		identification number added to the string "unknown" to create a new name to identify the probes that couldn't be identified by sequencing or other reason.
quality_pcr	varchar(10)	NO		value assigning the polymerase chain reaction quality control
quality_chromatogram	varchar(10)	NO		value assigning the chromatogram quality control
sequencing_status	enum('confirmed','not confirmed','not available','changed')	NO	not available	string category which indicate whether the probe was confirmed, not confirmed or not available after sequencing
description	text	NO		text containing detailed description about the array design
file_name	varchar(255)	NO		name of the file representing the array design
file_path	varchar(255)	NO		directory pathway to access the array design file
update_date	timestamp(14)	YES	CURRENT_TIMESTAMP	date record was inserted/last updated
user_update	varchar(64)	NO		person who last updated the record
molecule	enum('cDNA','DNA','Protein')	NO	cDNA	categories used to identify of which consists the probes synthesized in the array design, composed by cDNA, DNA or Protein
type	varchar(12)	NO		attribute to identify the type of array, if it will be miRNA, oligo cDNA, oligo DNA or Protein
3 categorization				
Table Description		Table contains tag annotation and keywords annotation linked to the transcript annotation table, setting relevant transcript information		
Column Name	Data Type	Not Null	Default Value	Comment
id	int(10)	YES		Unique identifier for each annotation tag related to each category. MySQL sequence generated number
transcript_annotation_id	int(10)	YES		Unique identifier for each category annotation. Link to the transcript_annotation table.
annotation	text	NO		general description about the transcript tag characterization
tag	varchar(255)	NO		general tag transcript classification, related to transcript function and/or structural
level	int(10)	NO		tag sub category level

4 cross_experiment				
Table Description	Linking table between the tables experiment and cross, representing the number of experiments with replicates are used to the search for differentially expressed probes			
Column Name	Data Type	Not Null	Default Value	Comment
id	int(10)	YES		Unique identifier for a relation between the tables experiment and cross. MySQL generated sequence number
experiment_id	int(10)	YES		Non-unique identifier references for experiment table
cross_ref_id	int(10)	YES		Non-unique identifier references for cross table
5 cross_ref				
Table Description	Table describing the experiments used to be correlated, methods and parameters applied in the definition of the differentially expressed probes. Usually are considered two experiments (one biological replicate)			
Column Name	Data Type	Not Null	Default Value	Comment
id	int(10)	YES		Unique identifier for each cross. MySQL sequence generated number
name	varchar(64)	YES		Unique and short name of the experiment crossing
type_name	varchar(32)	NO		Type of experiment set (technical replicate, biological replicate)
code	varchar(100)	NO		identification code for the experiments cross
description	text	NO		general description of the experiment
status	enum('1','0')	NO	0	condition if use the crossing results '1' or discard it '0'
analysis_method	varchar(255)	NO		method used in the analysis (HTSelf, Rank Product, SAM, etc)
parameters	varchar(255)	NO		parameters used to make the normalization, analysis and cross the experiments
probability	varchar(32)	NO		credibility value used to define the cutoff curve
update_date	timestamp(14)	YES	CURRENT_TIMESTAMP	date the cross was created/updated
visibility	enum('Public','Private','Group')	NO	Private	conditional permission to use the cross results
6 dbinfo				

Table Description	Table reference describing the comparisons between two databases for any type of program (blast, interproscan, cross_match, blat, etc)			
Column Name	Data Type	Not Null	Default Value	Comment
id	int(10)	YES		Unique identifier for a result dbinfo
program	varchar(50)	YES		software used in the comparison of two databases
query_base	varchar(255)	YES		Non-unique name of query data set
subject_base	varchar(255)	YES		Non-unique name of target data set, the reference
parameters	varchar(128)	NO		Description of software parameters used to compare two datasets
description	text	NO		general description about the run
filename	varchar(255)	YES		pathway and file name for the generated result
probe_type	varchar(64)	NO		This attribute must contain the exact name of used probe entity (oligo_probe, transcript, etc). The probe_type is used in the construction of SQL queries to search the probe name in the other entities (oligo_probe, transcript)
update_date	varchar(128)	NO		Date the run analyze was performed
7 experiment				
Table Description	Informations about the experiments used in the microarray analysis. Definition of the channels labelling cy3, cy5, the control in the treatment, the analysis path and the permissions to access the experiment			
Column Name	Data Type	Not Null	Default Value	Comment
id	int(10)	YES		Unique identifier for each experiment. MySQL sequence generated number
hybridization_protocol_id	int(10)	YES		Unique identifier for each hybridization_protocol. Link to the hybridization_protocol table.
project_id	int(10)	YES		Unique identifier for each array_design. Link to the array_design table.
array_id	int(10)	YES		Unique identifier for each project. Link to the project table.
sample_id	int(10)	NO		Unique identifier for each sample. Link to the sample table.
control_id	int(10)	NO		Unique identifier for the control sample. Link to the sample table.
expt_name	varchar(64)	YES		experiment unique name
description	text	NO		full text description

hybridization_file_name	varchar(64)	NO		hybridization unique file name
hybridization_path	varchar(64)	NO		localization hybridization results pathway
analysis_path	varchar(128)	NO		localization analysis pathway
replicate_type	varchar(64)	YES		identification type of replicate experiment as biological, techincal, dye swap, etc
cy3_label	varchar(64)	YES		sample cy3 marked identification (control)name
cy5_label	varchar(64)	YES		sample cy5 marked identification (control)name
control	varchar(64)	YES		sample name used as a control
expt_date	varchar(64)	NO		experiment realization date
code	varchar(100)	NO		experiment identification code, general identification
update_date	timestamp(14)	YES	CURRENT_TIMESTAMP	date the experiment was created/updated
visibility	enum('Public','Private','Group')	NO		conditional permission to use the experiment info
8 hybridization_protocol				
Table Description		Information about the protocol details and conditions used in the microarray slide hybridization		
Column Name	Data Type	Not Null	Default Value	Comment
id	int(10)	YES		Unique identifier for each hybridization_protocol. MySQL sequence generated number
scanparam	varchar(255)	NO		parameters used to scan images
quantity_target	varchar(32)	NO		molecule quantity used in the hybridization
labeling_protocol	text	NO		A brief description of the protocol describing any methods/techniques and the experimental applications of the protocol.
solution	varchar(255)	NO		description solution used in protocol
temperature	int(11)	NO		temperature used for hybridization
volume	int(11)	NO		hybridization final volume
washing_procedure	varchar(255)	NO		description washing procedure

extraction	text	NO		software used to extract the image
description	text	NO		hybridization general description
file_path	varchar(255)	NO		localization pathway of the result files
9 oligo_probe				
Table Description	Table describing the oligo designed from cDNA sequences or DNA genomes and oligo conditions			
Column Name	Data Type	Not Null	Default Value	Comment
id	int(10)	YES		Unique identifier for each oligo_probe. MySQL sequence generated number
probe_name	varchar(64)	YES		Unique name to identify the oligo probe designed from DNA or cDNA sequences
seq	text	NO		A biologically significant sequence of base pairs corresponding to the oligo probe
molecule	varchar(5)	NO		molecule type that was used to generate the oligo probe cDNA or DNA
prediction	varchar(255)	NO		tools, methods and parameters used to predict the oligo
description	text	NO		text containing detailed description about the oligo probe
orientation	enum('sense','antisense')	NO	sense	oligo orientation design. Description if the oligo is sense or antisense in relation to the cDNA or DNA sequence region
tm	float	NO		Time Melting temperature based on the oligo sequence calculated by : $tm = 69 + 41 * ((\text{NumberG} + \text{NumberC} - 16.4) / \text{SeqLength})$. For Agilent oligos the time melting conditions are $TM \geq 68$ e $TM \leq 76$
start_pos	int(11)	NO		designed oligo start position in relation to the reference
end_pos	int(11)	NO		designed oligo end position in relation to the reference
reference	varchar(64)	NO		sequence name used as reference in the oligo design
p_g	float	NO		percentage of GC nucleotide in the designed oligo. Default %GC $\geq 35\%$ and %GC $\leq 50\%$
p_c	float	NO		percentage of C nucleotide in the designed oligo.
p_gc	float	NO		percentage of G nucleotide in the designed oligo. Default %G $< 50\%$
homopolymer	enum('1','0')	NO	0	bit value, definition if the oligo is classified as homopolymer as 1 or not 0
organism_id	int(5)	YES		Unique identifier for each organism. Link to the organism table.
transcript_id	int(10)	NO		Identifier relating the transcript to their oligo_probes. Link to the transcript table.

10 organism				
Table Description	The table describing the plat species and taxonomy			
Column Name	Data Type	Not Null	Default Value	Comment
id	int(5)	YES		Unique identifier for each organism. MySQL sequence generated number
name	varchar(255)	YES		Plant species
common_name	varchar(255)	NO		Common/usual name
code	varchar(64)	NO		used code to identify the plant species
taxonomy	text	NO		A text describing the plant taxonomy
11 probe_expression				
Table Description	Linking table between tables cross and probe_set sumarizing the ratio and procentage of expression for the probes in each experiment and all experiments when used more than 1 biological replicate			
Column Name	Data Type	Not Null	Default Value	Comment
id	int(10)	YES		Unique identifier for each probe_expression. MySQL sequence generated number
cross_ref_id	int(10)	YES		Unique identifier for each cross. Link to the cross table.
probe_set_id	int(10)	YES		Unique identifier for each probe_set. Link to the probe_set table.
ratio_expt1	float	NO		log (base 2) median ratio intensities for the probe in experiment 1
ratio_expt2	float	NO		log (base 2) median ratio intensities for the probe in experiment 2
mean_ratio_expts	float	NO		log (base 2) median ratio intensities for the probe in all used experiments
percent_expt1	float	NO		Percentage of probes replicates above the cutoff based on specific method (HTSelf, Rank Product, SAM, etc) in experiment 1
percent_expt2	float	NO		Percentage of probes replicates above the cutoff based on specific method (HTSelf, Rank Product, SAM, etc) in experiment 2
mean_percent_expts	float	NO		Percentage of probes replicates above the cutoff based on specific method (HTSelf, Rank Product, SAM, etc) in all used experiments
expr_type	enum('up','down','in side')	NO	inside	categories which indicate whether probe is significantly induced (up-regulated), repressed (down-regulated) or not significantly expressed (inside)
12 probe_intensity				

Table Description	Linking table between tables experiment and probe_set describing the ratio and intensity expression for each probes in each experiment			
Column Name	Data Type	Not Null	Default Value	Comment
id	int(10)	YES		Unique identifier for each probe_intensity. MySQL sequence generated number
experiment_id	int(10)	YES		Unique identifier for each experiment. Link to the experiment table.
probe_set_id	int(10)	YES		Unique identifier for each probe_set. Link to the probe_set table.
array_side	enum('L','R','U')	NO	U	Characters indicating the probe intensity from the left or right side of the array for manufactured arrays, for Agilent arrays there is only one side "U"
cy3_intensity	float	NO		normalized spot pixel intensity at channel 1 (cy3), for manufactured array the channels are not normalized
cy5_intensity	float	NO		normalized spot pixel intensity at channel 2 (cy5), for manufactured array the channels are not normalized
log_intensity	float	NO		log (base 2) of the spot pixel intensity between channels cy3 and cy5 without normalization
log_ratio	float	NO		log (base 2) of the spot pixel intensity between channels cy3 and cy5 normalized
13 probe_set				
Table Description	Linking table between the original probes (transcripts, oligos, miRNAs, etc) summarizing the probes used in some experiment analyze as Microarray, Transgenic Plant, RT-PCR and the functional annotations detailed in categories			
Column Name	Data Type	Not Null	Default Value	Comment
id	int(10)	YES		Unique identifier for each probe_set. MySQL sequence generated number
probe_type_id	int(10)	YES		Unique identifier for each probe set. Link to the probe tables (transcript, oligo_probe, etc).
description	text	NO		description of probe design and selection
probe_type	varchar(64)	YES		This attribute must contain the exact name of used probe entity (oligo_probe, transcript, etc). The probe_type is used in the construction of SQL queries to search the probe name in the other entities (oligo_probe, transcript)
update_date	timestamp(14)	YES	CURRENT_TIMESTAMP	Date the probe_set was performed/updated

14 probe_statistics				
Table Description	Table containing summarized statistics for the probes in each experiment, indication the probe regulation as (up, down or inside) and the intensities			
Column Name	Data Type	Not Null	Default Value	Comment
id	int(10)	YES		Unique identifier for each probe_statistics. MySQL sequence generated number
cross_ref_id	int(10)	YES		Unique identifier for each cross. Link to the cross table.
probe_set_id	int(10)	YES		Unique identifier for each probe_set. Link to the probe_set table.
total	smallint(6)	NO		Total number of the same probe_set_id represented in the array originally
total_used	smallint(6)	NO		Total number of the same probe_set_id whose intensity was quantified and considered in the statistics computation
infinity	smallint(6)	NO		total number of the probe_set_id whose intensity is saturated and the software didn't define the real intensity, labelling "+Inf"
minus_infinity	smallint(6)	NO		total number of the probe_set_id whose intensity is saturated and the software didn't define the real intensity, labelling "-Inf"
median	float	NO		Sum of the arithmetic median log (base 2) intensities for each replicate probe ratios
mean	float	NO		Sum of the arithmetic mean log (base 2) intensities for each replicate probe ratios
stdev	float	NO		Standard deviation of the pixel log (base 2) intensity ratios for the replicate probes
up	smallint(6)	NO		Total number of probe replicates in the same array whose probes are set as up-regulated
down	smallint(6)	NO		Total number of probe replicates in the same array whose probes are set as down-regulated
inside	smallint(6)	NO		Total number of probe replicates in the same array whose probes are set as without differential expression
p_up	float	NO		Probability value (confidence level) that probe is set as up-regulated
p_down	float	NO		Probability value (confidence level) that probe is set as down-regulated

p_inside	float	NO		Probability value (confidence level) that probe is set as inside or not differentially expressed
15 project				
Table Description	Contain informations about the Project including the responsible for the project and the funding. The project is related with other experiments, results and analysis, like entities from Microarray			
Column Name	Data Type	Not Null	Default Value	Comment
id	int(10)	YES		Unique identifier for each project. MySQL sequence generated number
name	varchar(64)	YES		Name of the project to be related to the analysis and results in the database
responsible	varchar(64)	YES		the Responsible/coordinator for the project
funding	varchar(255)	NO		Funding agency/company to the development of the project
class	varchar(45)	YES		Project Description
description	text	NO		General Name of the project
16 rt_pcr				
Table Description	Table describing the RT-PCR methods for the detection of changes in gene expression, the primers, parameters and conditions in the experiment			
Column Name	Data Type	Not Null	Default Value	Comment
id	int(10)	YES		Unique identifier for each rt_pcr. MySQL sequence generated number
purpose	varchar(64)	YES		goals to realization of the experiment
applications	varchar(64)	YES		traditional and new applications ex:(genotyping, pathogen detection, quality control and assay validation, DNA damage measurement, drug therapy efficacy, array verification, quantitation of gene expression ,etc)
pforward	varchar(64)	NO		forward orientation primer name

preverse	varchar(64)	NO		reverse orientation primer name
seq_pforward	varchar(255)	NO		forward primer sequence
seq_preverse	varchar(255)	NO		reverse primer sequence
source	varchar(128)	YES		sample tissue and procedence information about the material
status	varchar(64)	YES		completed, in progress, not started or problems occurring
location	varchar(128)	YES		localization of primer in freezer/box/etc
exps_in_analysis	text	NO		A text describing the characteristic in analysis (Ex.: Brix, drought) and general features of the experiment
description	text	NO		A text discribing the plant variety/species, tissues and gene familiy in analysis
visibility	enum('Public','Private','Group')	NO	Private	conditional permission to use the rt_pcr results
update_date	timestamp(14)	YES	CURRENT_TIMESTAMP	Date the rt_pcr information was performed
17 rt_pcr_analysis				
Table Description		Table describing the RT-PCR results and purposes of the experiment, linked to the table of rt_pcr		
Column Name	Data Type	Not Null	Default Value	Comment
id	int(10)	YES		Unique identifier for each rt_pcr_analysis. MySQL sequence generated number
rt_pcr_id	int(10)	YES		Unique identifier for each rt_pcr. Link to the rt_pcr table.
experiment	varchar(128)	NO		Experimental characteristics in analysis
behavior	varchar(255)	NO		Results of gene expression in the RT-PCR
expected	varchar(255)	NO		Results expected based on plant characteristics or previous results
notification	text	NO		A text describing the specific context of the experiment and general notes
18 rt_pcr_has_probe_set				
Table Description		linking table between the tables rt_pcr and probe_set, describing if more than one probe is used in the rt_pcr analyze		

Column Name	Data Type	Not Null	Default Value	Comment
id	int(10)	YES		Unique identifier for a relation between the tables rt_pcr and probe_set. MySQL generated sequence number
rt_pcr_id	int(10)	YES		Non-unique identifier references for rt_pcr table
probe_set_id	int(10)	YES		Non-unique identifier references for probe_set table
19 sample				
Table Description	This table associate the experiment provider and analysis, containing plant characterization, enviromental conditions, purpose to the use of the sample and the plant specie (organism)			
Column Name	Data Type	Not Null	Default Value	Comment
id	int(10)	YES		Unique identifier for each sample. MySQL sequence generated number
sample_provider_id	int(10)	YES		Unique identifier for each sample_provider. Link to the sample_provider table.
name	varchar(64)	YES		sample unique name
variety	varchar(64)	NO		plant variety name/code
control	varchar(255)	NO		sample description treated/not treated
treatment	varchar(255)	NO		description type of treatment
tissue	varchar(64)	NO		sample tissue
molecule	enum('RNA','DNA','P rotein')	NO	RNA	kind of molecule used for a experiment (DNA/RNA)
harvesting	varchar(255)	NO		condition time
growth	varchar(255)	NO		description about the soil conditions
humidity	varchar(255)	NO		measure of the average daily humidity
light	varchar(255)	NO		measures of light conditions
nutrients	varchar(255)	NO		measures about soil nutrients
pH	varchar(32)	NO		measure soil ph
sugar_measure	varchar(32)	NO		sugar measure description
temperature	varchar(32)	NO		enviromental temperature
vitamins	varchar(32)	NO		vitamins description

developmental_stage	varchar(255)	NO		description sample developmental stage
purpose	text	YES		description experimental purpose
collect_date	varchar(128)	NO		Date sample collection time
code	varchar(64)	NO		sample identification code
plant_material	varchar(255)	NO		plant material type collected
20 sample_provider				
Table Description	The table describe the sample provider, including institution, address and responsible			
Column Name	Data Type	Not Null	Default Value	Comment
id	int(10)	YES		Unique identifier for each sample_provider. MySQL sequence generated number
name	varchar(64)	YES		Sample short name
institution	varchar(64)	YES		Sample provider's Institution's name
responsible	varchar(64)	NO		Person responsible for providing the sample
email	varchar(64)	NO		sample provider's email
city	varchar(64)	NO		Sample provider's city
city_state	varchar(64)	NO		Sample provider's state
address	varchar(255)	NO		Sample provider's address
21 seq_reads				
Table Description	Table of all EST sequences in the project. Relation to the transcript table connecting the EST sequences with your related transcript (contig/cluster)			
Column Name	Data Type	Not Null	Default Value	Comment
id	int(10)	YES		Unique identifier for each EST read. MySQL sequence generated number
probe_name	varchar(64)	YES		Unique name of EST sequence read
plate_position	char(3)	NO		Plate Position 96-well
longest	enum('1','0')	NO	0	bit value, definition if the EST sequence is classified as longest as 1 or not 0
trimmed_seq	text	NO		A text describing the plant taxonomy

transcript_id	int(10)	YES		Unique identifier for each EST sequence. Link to the transcript table
22 similarity				
Table Description	Table contain general similarity softwares results, the results can admit any kind of results from different softwares (BLAST, BLAT, InterProScan, cross_match, etc)			
Column Name	Data Type	Not Null	Default Value	Comment
id	int(10)	YES		Unique identifier for a similarity. MySQL generated sequence number
dbinfo_id	int(10)	YES		Unique identifier for each dbinfo. Link to the dbinfo table.
query	varchar(255)	YES		Query sequence name for the match comparison
query_annotation	text	NO		predicted/curated functional annotation
query_length	int(20)	NO		query sequence length
query_start	int(11)	NO		start hsp position of query alignment
query_end	int(11)	NO		end hsp position of query alignment
subject	varchar(255)	YES		Target sequence for match comparison
subject_annotation	text	NO		predicted/curated functional annotation
subject_length	int(20)	NO		target sequence length
subject_start	int(11)	NO		start hsp position of subject alignment
subject_end	int(11)	NO		end hsp position of subject alignment
hit_order_number	int(5)	YES		the hit order number, starting from the best hit, highest score
evaluate	double	NO		E Value (Expect Value) describes the likelihood that a sequence with a similar score will occur in the database by chance.
score	float	YES		The alignment score is the sum of the scores specified for each of the aligned pairs of letters, and letters with nulls, in the alignment
score_bits	float	NO		A log-scaled version of a score. Normalized score
alignment_length	int(11)	NO		total alignment between the query and target sequence
alignment_identity	float	NO		total identity between the query and target sequence.
positives	float	NO		number of positive base alignments between the query and target sequence.
frame	varchar(10)	NO		orientation of the translated ORF
gaps	int(10)	NO		representation of the number of gaps in the alignment

byte_offset	int(10)	NO		sets FILEHANDLE's position. Indicating the alignment position in the big file to seek and get the information faster
23 transcript				
Table Description	Table describing the gene transcripts and the organism, including the nucleotide sequence and the predicted protein. This table can include alleles, paralogs and non-coding genes			
Column Name	Data Type	Not Null	Default Value	Comment
id	int(10)	YES		Unique identifier for each transcript. MySQL sequence generated number
probe_name	varchar(64)	YES		Unique name of the transcript probe
seq	text	NO		partial/total nucleotide sequence
full_length	enum('NC','N','C','-')	NO	-	description of the transcription, NC: contain N-terminal and C-terminal or N: contain only N-terminal or C: contain only C-terminal
cdstatus_oct2004	varchar(128)	NO		crop design status of use in 2004
cdstatus	varchar(128)	NO		crop design status of use
mw	int(11)	NO		the protein molecular weight
iep	double	NO		Calculate the isoelectric point of a protein
tmhmm	varchar(255)	NO		prediction of transmembrane helices in proteins
protein	text	NO		partial/total nucleotide sequence
code_number	varchar(20)	NO		internal numerical code
lab_code	varchar(20)	NO		identification code of the laboratory
organism_id	int(5)	YES		sequence orientation based on blastx alignment and ESTs in the cluster
seq_orientation	varchar(2)	NO		Unique identifier for each organism. Link to the organism table.
24 transcript_annotation				
Table Description	table reserved to the transcript annotation and categorization, identification of functional and structural main categories			
Column Name	Data Type	Not Null	Default Value	Comment

id	int(10)	YES		Unique identifier for each transcript annotation. MySQL sequence generated number
transcript_id	int(10)	NO		Unique identifier for each transcript. Link to the transcript table
function_category	varchar(255)	NO		manual annotation and categorization describing the gene function, based on Gene Ontology (GO), Metabolic Pathways (KEGG), Uniprot (Swissprot) and NR-NCBI
structural_category	varchar(255)	NO		manual annotation and categorization describing the gene structural function, based on Gene Ontology (GO), Metabolic Pathways (KEGG), Uniprot (Swissprot) and NR-NCBI
category	varchar(255)	NO		Major gene functional category. Manual partially curated
sub_category1	varchar(255)	NO		automatically generated categorization based on Uniprot Database
sub_category2	varchar(255)	NO		automatically generated categorization based on NR-NCBI Database. NR: Non-Redundant database composed by proteins and predicted proteins
sub_category3	varchar(255)	NO		automatically generated categorization based on GO (Gene Ontology) and KEGG (Kyoto Encyclopedia of Genes and Genomes) Database
description	text	NO		description of probe design and selection
update_date	timestamp(14)	YES	CURRENT_TIMESTAMP	Date the probe_set was performed/updated
evidence_code	varchar(128)	NO		<p>Annotators may find the evidence code decision tree a useful aid in selecting the correct evidence code for an annotation</p> <ul style="list-style-type: none"> # Experimental Evidence Codes <ul style="list-style-type: none"> # EXP: Inferred from Experiment # IDA: Inferred from Direct Assay # IPI: Inferred from Physical Interaction # IMP: Inferred from Mutant Phenotype # IGI: Inferred from Genetic Interaction # IEP: Inferred from Expression Pattern # Computational Analysis Evidence Codes <ul style="list-style-type: none"> # ISS: Inferred from Sequence or Structural Similarity # ISO: Inferred from Sequence Orthology # ISA: Inferred from Sequence Alignment # ISM: Inferred from Sequence Model # IGC: Inferred from Genomic Context # RCA: inferred from Reviewed Computational Analysis # Author Statement Evidence Codes <ul style="list-style-type: none"> # TAS: Traceable Author Statement # NAS: Non-traceable Author Statement # Curator Statement Evidence Codes <ul style="list-style-type: none"> # IC: Inferred by Curator # ND: No biological Data available

				# Automatically-assigned Evidence Codes # IEA: Inferred from Electronic Annotation # Obsolete Evidence Codes # NR: Not Recorded
responsible	varchar(255)	YES		the responsible for the transcript annotation
quality_control	enum('0','1','2','3','4','5')	YES	0	the annotation quality by the user
note	text	NO		general annotation and relevant information about the transcript and tag identification related to the categories
25 transgenic_plant				
Table Description		Table describing all informations related to the transgenic plant in construction, the single or multiple genes used, the primer sequences, expected phenotypes, methods of transformation, and purposes as siRNA, silencing, super expression, etc. The Accession permission to the construction and results are controlled		
Column Name	Data Type	Not Null	Default Value	Comment
id	int(10)	YES		Unique identifier for each transgenic_plant. MySQL sequence generated number
purpose	varchar(64)	YES		experiment goal (superexpression, silencing, miRNA, siRNA, etc)
pforward	varchar(64)	NO		forward primer name
preverse	varchar(64)	NO		reverse primer name
seq_pforward	varchar(255)	NO		forward primer sequence
seq_preverse	varchar(255)	NO		reverse primer sequence
CTC_id	varchar(64)	NO		plant identification at Centro de Tecnologia Canavieira
annotation	text	NO		A text description or annotation describing the specific context of the experiment
functional_annotation	varchar(255)	NO		functional category
preferential_tissue	varchar(255)	NO		plant tissue in which the gene is going to be used
responsible	varchar(128)	YES		Researcher responsible for making the gene construction
expected_phenotype	varchar(255)	NO		expected phenotype of the transgenic plant
plant_variety	varchar(255)	NO		plant variety name

transformation_method	text	NO		method used to insert gene construction in the cells
explant	varchar(255)	NO		plant tissue or organ collected from a plant that will be used to initiate in vitro culture
gene_expression_data	text	NO		gene expression data obtained in microarray or real time PCR experiments
plant_status	varchar(255)	YES		stage and place (i.e. laboratory or greenhouse) of plant regeneration and plant development
CTNBio_status	varchar(255)	NO		status of CTNBio analysis/liberation
notification	text	NO		A text describing the specific context of the experiment and general notes
visibility	enum('Public','Private','Group')	NO	Private	conditional permission to use the transgenic_plant results
update_date	timestamp(14)	YES	CURRENT_TIMESTAMP	username and date the experiment was inserted/updated
26 transgenic_plant_analysis				
Table Description	Table describing the Transgenic Plant results, plant and environmental conditions, purposes of the experiment, linked to the table of Transgenic_Plant			
Column Name	Data Type	Not Null	Default Value	Comment
id	int(10)	YES		Unique identifier for each transgenic_plant_analysis. MySQL sequence generated number
transgenic_plant_id	int(10)	YES		Unique identifier for each transgenic_plant. Link to the transgenic_plant table.
type_clone	varchar(100)	NO		DNA sequence inserted in the vector used for transformation
insert_sequence	text	NO		produced nucleotide insert sequence in the cloning vector
insert_length	varchar(64)	NO		length of the sequence inserted in the vector
vector	varchar(128)	NO		vector used to clone the DNA sequence of interest
clone_site	varchar(128)	NO		site of the vector where the DNA sequence was inserted
clone_site_orientation	varchar(64)	NO		clone site strand
transforming_characterization	text	NO		method used to confirm the plant transformation
phenotype_characterization	text	NO		methods used to characterize transgenic plants phenotype
agronomic_characterization	text	NO		agronomic traits characterized
pforward	varchar(64)	NO		forward primer name used to analyse transgenic plants
preverse	varchar(64)	NO		reverse primer name used to analyse transgenic plants

seq_pforward	varchar(255)	NO		forward primer sequence used to analyse transgenic plants
seq_preverse	varchar(255)	NO		reverse primer sequence used to analyse transgenic plants
selection_process	varchar(255)	NO		selection agent or method used to select transgenic plants in vitro
sucrose_content	varchar(255)	NO		measurement plant sucrose accumulation
drought_tolerance	varchar(255)	NO		measures and information for tolerance to drought stress
notification	text	NO		general description for the transgenic plant in production
update_date	varchar(255)	NO		Date the transgenic_plant analyze was performed
updateFile	varchar(255)	NO		pathway and file name for the generated result
27 <i>transgenic_plant_has_probe_set</i>				
Table Description	linking table between the tables transgenic_plant and probe_set, describing if more than one probe is used in the transgenic_plant analyze			
Column Name	Data Type	Not Null	Default Value	Comment
id	int(10)	YES		Unique identifier for a relation between the tables transgenic_plant and probe_set. MySQL generated sequence number
transgenic_plant_id	int(10)	YES		Non-unique identifier references for transgenic_plant table
probe_set_id	int(10)	YES		Non-unique identifier references for probe_set table

Anexo V. Modelo de Entidade Relacionamento do Banco de Dados Relacional CaneRegNetDB para as informações estruturais e funcionais dos genes

Anexo VI. Modelo de Entidade Relacionamento do Banco de dados relacional CaneRegNetDB dedicada ao gerenciamento, submissão e análise de experimentos e lâminas de microarranjo.

Anexo VII. Tabela com o total de genes curados no KEGG para todas as espécies de plantas monocotiledôneas.

IDKEGG	código	Espécie	# genes
T01015	osa	Oryza sativa japonica (Japanese rice) (RefSeq)	28453
T02163	dosa	Oryza sativa japonica (Japanese rice) (RAPDB)	35679
T02995	obr	Oryza brachyantha (malo sina)	20502
T01717	bdi	Brachypodium distachyon	23460
T01086	sbi	Sorghum bicolor (sorghum)	33002
T01088	zma	Zea mays (maize)	42101
T02818	sita	Setaria italica (foxtail millet)	29355
T03274	pda	Phoenix dactylifera (date palm)	26456
T03447	mus	Musa acuminata (wild Malaysian banana)	30716

Anexo VIII – Tabela Transferência Anotação transcritos SUCEST SAS a partir das gramíneas

# genes	Anotação
4186	unknown
284	Protein kinase superfamily protein
209	RING/U-box superfamily protein
192	Tetratricopeptide repeat (TPR)-like superfamily protein
188	NB-ARC domain-containing disease resistance protein
177	alpha/beta-Hydrolases superfamily protein
169	P-loop containing nucleoside triphosphate hydrolases superfamily protein
129	Transducin/WD40 repeat-like superfamily protein
122	ARM repeat superfamily protein
109	NAD(P)-binding Rossmann-fold superfamily protein
109	Leucine-rich repeat protein kinase family protein

108	S-adenosyl-L-methionine-dependent methyltransferases superfamily protein
107	Pentatricopeptide repeat (PPR) superfamily protein
104	RNA-binding (RRM/RBD/RNP motifs) family protein
101	basic helix-loop-helix (bHLH) DNA-binding superfamily protein
100	Major facilitator superfamily protein
91	F-box family protein
83	Peroxidase superfamily protein
75	Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin superfamily protein
74	Histone superfamily protein
70	Eukaryotic aspartyl protease family protein
70	2-oxoglutarate (2OG) and Fe(II)-dependent oxygenase superfamily protein
68	F-box/RNI-like superfamily protein
63	GDSL-like Lipase/Acylhydrolase superfamily protein
62	UDP-Glycosyltransferase superfamily protein
57	Heavy metal transport/detoxification superfamily protein
55	RING/FYVE/PHD zinc finger superfamily protein
55	Homeodomain-like superfamily protein
50	Protein phosphatase 2C family protein

Anexo IX – Tabela Transferência Anotação transcritos RNA-Seq Trinity a partir das gramíneas

# genes	Anotação
10342	unknown
536	Protein kinase superfamily protein
505	NB-ARC domain-containing disease resistance protein
443	Tetratricopeptide repeat (TPR)-like superfamily protein
406	RING/U-box superfamily protein
388	alpha/beta-Hydrolases superfamily protein
315	Pentatricopeptide repeat (PPR) superfamily protein
277	P-loop containing nucleoside triphosphate hydrolases superfamily protein

274	Leucine-rich repeat protein kinase family protein
265	ARM repeat superfamily protein
245	Transducin/WD40 repeat-like superfamily protein
216	basic helix-loop-helix (bHLH) DNA-binding superfamily protein
212	RNA-binding (RRM/RBD/RNP motifs) family protein
210	Major facilitator superfamily protein
207	S-adenosyl-L-methionine-dependent methyltransferases superfamily protein
199	NAD(P)-binding Rossmann-fold superfamily protein
155	F-box family protein
151	F-box/RNI-like superfamily protein
150	UDP-Glycosyltransferase superfamily protein
148	2-oxoglutarate (2OG) and Fe(II)-dependent oxygenase superfamily protein
145	Heavy metal transport/detoxification superfamily protein
136	Protein of unknown function (DUF594)
128	RING/FYVE/PHD zinc finger superfamily protein
122	Homeodomain-like superfamily protein
121	Chaperone DnaJ-domain superfamily protein
107	LRR and NB-ARC domains-containing disease resistance protein
105	Protein phosphatase 2C family protein
104	hydroxyproline-rich glycoprotein family protein
101	Peroxidase superfamily protein
100	Ankyrin repeat family protein
97	Integrase-type DNA-binding superfamily protein
96	GDSL-like Lipase/Acylhydrolase superfamily protein
91	Eukaryotic aspartyl protease family protein
89	Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin superfamily protein
85	Thioredoxin superfamily protein
85	S-locus lectin protein kinase family protein
76	Haloacid dehalogenase-like hydrolase (HAD) superfamily protein
75	Zinc finger C-x8-C-x5-C-x3-H type family protein
74	HXXXX-type acyl-transferase family protein
73	RNI-like superfamily protein
73	MATE efflux family protein

73	Calcium-binding EF-hand family protein
72	Galactose oxidase/kelch repeat superfamily protein
71	Mitochondrial substrate carrier family protein
71	Calcium-dependent lipid-binding (CaLB domain) family protein
69	O-fucosyltransferase family protein
67	FAR1-related sequence 5
66	Pentatricopeptide repeat (PPR-like) superfamily protein
65	transducin family protein / WD-40 repeat family protein
65	Sec14p-like phosphatidylinositol transfer family protein
65	Protein of unknown function (DUF581)
65	Nucleotide-diphospho-sugar transferases superfamily protein
63	heat shock transcription factor A6B
63	F-box and associated interaction domains-containing protein
63	Cysteine proteinases superfamily protein
63	Concanavalin A-like lectin protein kinase family protein
61	Late embryogenesis abundant (LEA) hydroxyproline-rich glycoprotein family
61	DNAJ heat shock N-terminal domain-containing protein
60	Polynucleotidyl transferase, ribonuclease H-like superfamily protein
60	Leucine-rich repeat (LRR) family protein
59	Pectin lyase-like superfamily protein
59	NAD(P)-linked oxidoreductase superfamily protein
57	nodulin MtN21 /EamA-like transporter family protein
56	GRAS family transcription factor
56	Disease resistance protein (CC-NBS-LRR class) family
55	DNA binding
55	AP2/B3-like transcriptional factor family protein
54	O-Glycosyl hydrolases family 17 protein
53	Leucine-rich receptor-like protein kinase family protein
53	F-box/RNI-like/FBD-like domains-containing protein
53	Aldolase-type TIM barrel family protein
51	SAUR-like auxin-responsive protein family
50	HSP20-like chaperones superfamily protein

Anexo X – Tabela transferência Anotação transcritos SUCEST com Vias Metabólicas KEGG – KAAS

# genes	Vías Metabólicas
1593	Metabolic pathways
631	Biosynthesis of secondary metabolites
202	Biosynthesis of amino acids
194	Purine metabolism
187	Spliceosome
183	Protein processing in endoplasmic reticulum
179	Carbon metabolism
158	Oxidative phosphorylation
154	RNA transport
147	Pyrimidine metabolism
128	Phagosome
125	Starch and sucrose metabolism
117	Ubiquitin mediated proteolysis
114	Plant hormone signal transduction
97	mRNA surveillance pathway
96	Ribosome biogenesis in eukaryotes
96	Amino sugar and nucleotide sugar metabolism
88	Endocytosis
85	Glycolysis / Gluconeogenesis
84	Nucleotide excision repair
77	Glycerophospholipid metabolism
75	Glutathione metabolism
74	Peroxisome
72	DNA replication
69	Pyruvate metabolism
69	Plant-pathogen interaction
69	N-Glycan biosynthesis
67	Homologous recombination
62	Porphyrin and chlorophyll metabolism
61	Fatty acid metabolism
61	Cysteine and methionine metabolism
60	Mismatch repair
59	Arginine and proline metabolism

57	RNA polymerase
56	Protein export
56	Photosynthesis
56	Glycerolipid metabolism
56	Basal transcription factors
53	Carbon fixation in photosynthetic organisms
52	2-Oxocarboxylic acid metabolism
50	Base excision repair
50	Aminoacyl-tRNA biosynthesis
49	Citrate cycle (TCA cycle)
47	Phenylpropanoid biosynthesis
47	Phenylalanine metabolism
47	Glyoxylate and dicarboxylate metabolism
47	Fructose and mannose metabolism
47	Ascorbate and aldarate metabolism
45	Glycine, serine and threonine metabolism
44	Phosphatidylinositol signaling system
43	Alanine, aspartate and glutamate metabolism
42	Pentose phosphate pathway
42	Pentose and glucuronate interconversions
41	Terpenoid backbone biosynthesis
41	Glycosylphosphatidylinositol(GPI)-anchor biosynthesis
40	Inositol phosphate metabolism
40	Fatty acid degradation
39	Tyrosine metabolism
37	SNARE interactions in vesicular transport
37	Phenylalanine, tyrosine and tryptophan biosynthesis
36	Valine, leucine and isoleucine degradation
35	ABC transporters
33	Propanoate metabolism
32	Sulfur metabolism
32	Galactose metabolism
32	Butanoate metabolism
28	Sphingolipid metabolism

27	Pantothenate and CoA biosynthesis
27	Fatty acid biosynthesis
26	Nitrogen metabolism
25	Lysine degradation
25	alpha-Linolenic acid metabolism
24	Tryptophan metabolism
24	Regulation of autophagy
24	Nicotinate and nicotinamide metabolism
24	Circadian rhythm - plant
23	Ubiquinone and other terpenoid-quinone biosynthesis
23	Other glycan degradation
23	Biosynthesis of unsaturated fatty acids
23	Arachidonic acid metabolism
22	Stilbenoid, diarylheptanoid and gingerol biosynthesis
22	Biotin metabolism
21	Selenocompound metabolism
21	beta-Alanine metabolism
20	Lysine biosynthesis
20	Ether lipid metabolism
20	Cyanoamino acid metabolism
19	Folate biosynthesis
18	Valine, leucine and isoleucine biosynthesis
18	Sulfur relay system
18	Histidine metabolism
18	Flavonoid biosynthesis
17	Limonene and pinene degradation
16	Steroid biosynthesis
16	Glycosaminoglycan biosynthesis - heparan sulfate / heparin
15	Thiamine metabolism
15	Riboflavin metabolism
15	Photosynthesis - antenna proteins
15	One carbon pool by folate
14	Glycosaminoglycan degradation

14	Carotenoid biosynthesis
13	Taurine and hypotaurine metabolism
12	Linoleic acid metabolism
12	Isoquinoline alkaloid biosynthesis
12	Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate
12	Fatty acid elongation
11	Degradation of aromatic compounds
10	Tropane, piperidine and pyridine alkaloid biosynthesis
10	Other types of O-glycan biosynthesis
10	Diterpenoid biosynthesis
9	Synthesis and degradation of ketone bodies
9	Non-homologous end-joining
9	Glycosphingolipid biosynthesis - globo series
9	C5-Branched dibasic acid metabolism
8	Glycosaminoglycan biosynthesis - keratan sulfate
7	Vitamin B6 metabolism
7	Brassinosteroid biosynthesis
6	Cutin, suberine and wax biosynthesis
5	Zeatin biosynthesis
5	Glycosphingolipid biosynthesis - ganglio series
4	Sesquiterpenoid and triterpenoid biosynthesis
3	Lipoic acid metabolism
2	Flavone and flavonol biosynthesis
2	Caffeine metabolism
2	Betalain biosynthesis
1	Glucosinolate biosynthesis

Anexo XI – Tabela transferência Anotação transcritos RNA-Seq com Vias Metabólicas KEGG – KAAS

# genes	Vias metabólicas
7385	Metabolic pathways
2761	Biosynthesis of secondary metabolites
976	Purine metabolism

957	Biosynthesis of amino acids
811	Carbon metabolism
725	RNA transport
700	Pyrimidine metabolism
608	Spliceosome
573	Starch and sucrose metabolism
543	Protein processing in endoplasmic reticulum
447	Oxidative phosphorylation
428	Ubiquitin mediated proteolysis
419	mRNA surveillance pathway
406	Glycerophospholipid metabolism
402	Amino sugar and nucleotide sugar metabolism
379	Ribosome biogenesis in eukaryotes
376	Phagosome
375	Endocytosis
361	Glycolysis / Gluconeogenesis
356	N-Glycan biosynthesis
354	Nucleotide excision repair
334	DNA replication
321	Arginine and proline metabolism
312	Pyruvate metabolism
304	Aminoacyl-tRNA biosynthesis
303	Cysteine and methionine metabolism
302	Plant hormone signal transduction
293	Glutathione metabolism
287	Peroxisome
281	Glycine, serine and threonine metabolism
278	Porphyrin and chlorophyll metabolism
266	Alanine, aspartate and glutamate metabolism
261	Mismatch repair
259	Homologous recombination
239	2-Oxocarboxylic acid metabolism
236	Glycerolipid metabolism
230	Fructose and mannose metabolism

228	Citrate cycle (TCA cycle)
228	Base excision repair
223	Fatty acid metabolism
221	ABC transporters
219	Galactose metabolism
211	Inositol phosphate metabolism
210	Nicotinate and nicotinamide metabolism
208	Pentose phosphate pathway
207	Phosphatidylinositol signaling system
205	FoxO signaling pathway
204	Plant-pathogen interaction
204	Pentose and glucuronate interconversions
195	Carbon fixation in photosynthetic organisms
189	Glyoxylate and dicarboxylate metabolism
186	Protein export
185	Basal transcription factors
180	Tyrosine metabolism
179	Terpenoid backbone biosynthesis
176	Dilated cardiomyopathy
173	RNA polymerase
173	Pantothenate and CoA biosynthesis
168	Glycosylphosphatidylinositol(GPI)-anchor biosynthesis
163	Propanoate metabolism
157	Phenylalanine, tyrosine and tryptophan biosynthesis
156	Phenylalanine metabolism
141	Sphingolipid metabolism
136	Photosynthesis
135	One carbon pool by folate
133	Valine, leucine and isoleucine degradation
132	Phenylpropanoid biosynthesis
131	Sulfur metabolism
129	Butanoate metabolism
125	Histidine metabolism
121	Fatty acid degradation

120	Selenocompound metabolism
119	Nitrogen metabolism
118	Lysine degradation
117	Ascorbate and aldarate metabolism
116	Other types of O-glycan biosynthesis
111	Riboflavin metabolism
109	Glycosaminoglycan biosynthesis - heparan sulfate / hep
107	SNARE interactions in vesicular transport
104	Folate biosynthesis
103	Cyanoamino acid metabolism
101	Steroid biosynthesis
101	beta-Alanine metabolism
100	Fatty acid biosynthesis
99	Ubiquinone and other terpenoid-quinone biosynthesis
99	Ether lipid metabolism
98	Tryptophan metabolism
97	Lysine biosynthesis
92	Other glycan degradation
91	Biosynthesis of unsaturated fatty acids
90	alpha-Linolenic acid metabolism
86	Taurine and hypotaurine metabolism
85	Non-homologous end-joining
84	Regulation of autophagy
83	Valine, leucine and isoleucine biosynthesis
83	Arachidonic acid metabolism
81	Sulfur relay system
77	Glycosaminoglycan degradation
74	Circadian rhythm - plant
73	Thiamine metabolism
72	Carotenoid biosynthesis
68	Glycosaminoglycan biosynthesis - chondroitin sulfate /
67	Tropane, piperidine and pyridine alkaloid biosynthesis
63	Limonene and pinene degradation
62	Stilbenoid, diarylheptanoid and gingerol biosynthesis

59	Biotin metabolism
56	Isoquinoline alkaloid biosynthesis
51	Fatty acid elongation
51	Degradation of aromatic compounds
48	Photosynthesis - antenna proteins
47	Glycosphingolipid biosynthesis - globo series
45	Linoleic acid metabolism
40	Glycosaminoglycan biosynthesis - keratan sulfate
35	Zeatin biosynthesis
35	Vitamin B6 metabolism
34	Flavonoid biosynthesis
33	C5-Branched dibasic acid metabolism
28	Synthesis and degradation of ketone bodies
28	Cutin, suberine and wax biosynthesis
26	Glycosphingolipid biosynthesis - ganglio series
25	Flavone and flavonol biosynthesis
23	Sesquiterpenoid and triterpenoid biosynthesis
19	Diterpenoid biosynthesis
13	Lipoic acid metabolism
12	Caffeine metabolism
11	Indole alkaloid biosynthesis
11	Brassinosteroid biosynthesis
10	Betalain biosynthesis
4	Monoterpenoid biosynthesis
3	Glucosinolate biosynthesis
1	D-Arginine and D-ornithine metabolism
1	Benzoxazinoid biosynthesis

Anexo XII – Script para Calcular Escore do melhor hit a partir de um alinhamento entre dois transcriptomas bi-direcional (BHR)

```
#!/usr/bin/perl -w
```

```
use strict;
```

```
#Select Hits with BHR > 0.95
```

```
#Script to read forward and reverse alignment Transcriptome A vs Transcriptome B
```

```
#Calculate the BHR and print only the hits with BHR > 0.95
```

```
if(@ARGV != 4) {
    die "Usage: <%BHR> <prefix transcript_name> <Blast Genome AvsB output>
    <Reverser Blast BvsA output>\n";
}
my $pbhr = shift;
my $pfx = shift;
my $fbblast1 = shift;
my %hblast1 = ();
my @aline = ();

#GeneA
#1 - GeneB scorebit 15 AvsB
#2 - GeneC scorebit 14 AvsC
#3 - GeneD scorebit 12 AvsD
#GeneB
#1 - GeneD scorebit 13 BvsD
#2 - GeneE scorebit 11 BvsE
#3 - GeneA scorebit 11 BvsA
#BHR = (DBA against DBB) * (DBB against DBA)
#BHR = (Savsb/Savs1stb) * (Sbvsa/Sbvs1sta)
#BHR = (15/15) * (11/13) = 1-0.84 = 0.16

my %hrepeat = ();
my %hblast1st = ();
my $aux = "";
my $id = 1;
open(FB,$fbblast1);
while(<FB>) {
    chomp;
    #Blast output format:
    #qseqid qlen qframe sseqid slen sframe nident pident length mismatch gapopen
    qstart qend sstart send qcovs qcovhsp evaluate bitscore score
    push(@aline, split(/\t/, $_));
```

```

$id = 1 if($aux ne $aline[0]);
if(defined $hrepeat{"$aline[0]"}{"$aline[3]"}) {
    undef @aline;
    next;
}
$blast1st{"$aline[0]"} = "$aline[3]" if($id == 1);
$blast1{"$aline[0]"}{"$id"}{"$aline[3]"} = $aline[19];
$hrepeat{"$aline[0]"}{"$aline[3]"}++;

$aux = "$aline[0]";
$id++ if($aux eq $aline[0]);
undef @aline;
}
close(FB);

while(<>) {
    chomp;
    push(@aline, split(/\t/, $_));
    $id = 1 if($aux ne $aline[0]);

    if(defined $hrepeat{"$aline[0]"}{"$aline[3]"}) {
        undef @aline;
        next;
    }
    $blast1st{"$aline[0]"} = "$aline[3]" if($id == 1);
    $blast1{"$aline[0]"}{"$id"}{"$aline[3]"} = $aline[19];
    $hrepeat{"$aline[0]"}{"$aline[3]"}++;

    $aux = "$aline[0]";
    $id++ if($aux eq $aline[0]);
    undef @aline;
}
close(FB);

my $v = 0;
my $sa = 0;

```

```

my $sa1 = 0;
my $sb = 0;
my $sb1 = 0;
my $bhr = 0;
my $pos = 1;
my $g1st = "";
my $c1st = "";
my $gv1 = 0;
my $cv1 = 0;
my $g = "";
my $gv = 0;

%hrepeat = ();
foreach my $c (grep /$pfx/, keys %hblast1) { #Contig 1st Hit A vs B

    foreach my $pc (1..scalar(keys %{ $hblast1{$c} })) {

        foreach my $g (keys %{ $hblast1{$c}{$pc} }) {

            next if(!defined $hblast1st{$g});

            if($pc eq 1) {
                ($g1st,$gv1) = each %{ $hblast1{$c}{$pc} };
                $c1st = $hblast1st{$g1st};
            }
            else {
                #Get Corrent Contig vs GeneHit by Pos
                ($g,$gv) = each %{ $hblast1{$c}{$pc} };
                $c1st = $hblast1st{$g};
            }

            #Get HitOrderNumber for pair $g1st vs $c - $c may not be the 1sthit in reverse
alignment
            my $pos = 0;
            if($c eq $c1st) {
                $pos = $pc;
            }

```

```

else {
    foreach my $pg ( 2..scalar(keys %{$hblast1{$g}}) ) {
        while( my ($cx, $v) = each %{$hblast1{$g}{$pg}} ) {
            $pos = $pg and last if($cx eq $c);
        }
    }
}

next if(!$pos);

#Get Score contig vs gene - 1stHit
$sa1 = (!defined $hblast1{$c}{"1"}{$g1st}) ? 0:$hblast1{$c}{"1"}{$g1st};

#Get Score contig vs gene - CurrentHit
$sa = (!defined $hblast1{$c}{$pc}{$g}) ? 0:$hblast1{$c}{$pc}{$g};

#Get Score gene vs contig - HitOrderNumber $pos
$sb = (!defined $hblast1{$g}{$pos}{$c}) ? 0:$hblast1{$g}{$pos}{$c};

#Get Score gene vs first hit contig
$sb1 = (!defined $hblast1{$g}{"1"}{$c1st}) ? 0:$hblast1{$g}{"1"}{$c1st};

if($sa1 && $sb1) {
    $bhr = $sa/$sa1 * $sb/$sb1;
}
else {
    $bhr = 0;
}

#BHR = (Savsb/Savs1stb) * (Sbvsa/Sbvs1sta)
if(($bhr >= $pbhr) && (!exists $hrepeat{$c})) {
    print "$c\t$g\t$pc\t$bhr\n";
    $hrepeat{$c}++;
}
}
}
}

```

Anexo XIII – script processar output do Blast e gerar cobertura total do gene consulta ou referência utilizando algoritmo para calcular cobertura completa com todos os *hits*

```
#!/usr/bin/perl
#The script read a Blast output and generate the total coverage adding/concatenating
intervals and return the whole coverage
#for the choosen reference (query or subject)
#Parameters:
#1 - Use for the calcule and print only first hit or all hits
#2 – List of full-length genes identified in the previous analyzed species
#2 - Choose the reference to calculate the full coverage
#3 - Filter to print only if >%QueryCoverage
#4 - Filter to print only if >%Coverage
#5 - Filter to print only if >%Identity
#6 - Blast output
#Get Blastn Output m6 custom format and print the statistics for the Query or Subject
#blast_hits_m6.4.pl      f      q      Sugarcane.1stHitCovHigher99vsSpecies.txt
454PGMTrinityAll_vs_Sbicolor.full-length.blastn
use strict;
use vars qw(%hacc);

if(@ARGV != 7) {
    print "Usage : $0 <(f)irst/(a)ll hits/hsp> <(e)xclude the full-length in the file list/(n)o
exclude> <Coverage Query/Subject (q/s)> <Print if > PercCov> <Print if > PercIdent>
<list full-length contigs> <blast_output> \n";
    exit 0;
}
#BLASTOUTPUT PATTERN
#qseqid qframe qlen sseqid slen sframe nident pident length mismatch gapopen qstart
qend sstart send qcovs qcovhsp evaluate bitscore score

my $hit      = shift;
my $excludel = shift;
my $reference = shift;
my $percC    = shift;
```

```

my $percI      = shift;
my $filefl     = shift;
my ($refid,$auxid) = ($reference eq 'q') ? (0,3):(3,0);

my ($blast, %dataset, %hsp, $hsp_order_number,
    @b_dataset,$b_dataset,@b_hsps,$b_hsp,$identities,$positives,$status_hit_id);
my ($spa, $spi, $sum_alignment_length, $sum_percent_identities, $al, $pi, $sps) = 0;
my @afields    = ();
my %hhitsbygene = ();
my %hsubject    = ();
my $aux        = "";

my %hgeneftl    = ();
open(FFL,"$filefl");
while(<FFL>) {
    $hgeneftl{$1}++ if(/\\S+\\t+(\\S+)/);
}
close(FFL);

#exit;
my %hquery_hit = ();
my %hhits = ();
while(<>) {
    #qseqid qlen qframe sseqid slen sframe nident pident length mismatch gapopen
    qstart qend sstart send qcovs qcovhsp evalule bitscore score
    @afields = split(/\\t/,$_);

    #Exclude from coverage the Contigs already analyzed and Classified as Full-Length
    #To cover the whole gene from other species
    next if($hgeneftl{$afields[$auxid]} && $excludefl eq 'e');

    #Next if choose only 1st Hit
    next if($hhits{"$afields[$refid]"} && $hit eq 'f');

    if($refid) { #Subject Hits

```



```

    add_interval("$afields[$refid]",$afields[13],$afields[14]);
    $hsubject{"$afields[$refid]"}{'length'} = $afields[4];
    #print "1a - $afields[$refid]\t$afields[13]\t$afields[14]\n";
}
else { #Query Hits
    add_interval("$afields[$refid]",$afields[11],$afields[12]);
    $hsubject{"$afields[$refid]"}{'length'} = $afields[1];
    #print "1b - $afields[$refid]\t$afields[$auxid]\t$afields[11]\t$afields[12]\n";
    $hquery_hit{"$afields[$refid]"}{"$afields[$auxid]"}
    =
    ["$afields[11],$afields[12]"];
    $hhits{"$afields[$refid]"}++;
}
#if reference=query {query}{sum}{subject}
#if reference=subject {subject}{sum}{query}
$hsubject{"$afields[$refid]"}{'sum'}{"$afields[$auxid]"}++;

push @{$hsubject{"$afields[$refid]"}{'pidentity'}},$afields[7];
#print
"1>$afields[0]\t$afields[$refid]\t$afields[4]\t$afields[7]\t$afields[14]\t$afields[15]\n"
;
}

```

```

my $total = 0;
my $pcov = 0;
my $pimean = 0;
my %hpcovfreq = ();
my $sum = 0;

```

```

foreach my $gene (keys %hacc) {
    $total = 0;
    $pimean = 0;
    $sum_alignment_length = 0;

```

```

    for(my $idc1= scalar(@{$hacc{"$gene"}})-1;$idc1>=0;$idc1=$idc1-2) {
        my ($s,$e) = (${$hacc{"$gene"}}[$idc1-1],${$hacc{"$gene"}}[$idc1]);
        $sum_alignment_length += ($e-$s);
    }

```

```

}

foreach(@{$hsubject{"$gene"}{'pidentity'}}) {
    $total += $_;
}

if(scalar(@{$hsubject{"$gene"}{'pidentity'}}) > 1) {
    $pimean = $total/($hsubject{"$gene"}{'sum'});
}
else {
    $pimean = $total;
}

$pcov = (($sum_alignment_length+1)/$hsubject{$gene}{'length'})*100;

if( $pcov > 99 ) {
    $hpcovfreq{10}++;
}else {
    $hpcovfreq{int($pcov/10)}++;
}

$sum = keys %{$hsubject{$gene}{'sum'}};
$hhitsbygene{$sum}++;

if($pcov >= $percC && $pimean >= $percI) {
#   if($pcov >= $percC) {
        printf
"Query\t$gene\t$hsubject{$gene}{'length'}\t$sum\t%3.3f\t%3.3f\t[@{$hacc{"$gene"}{''}}]\n", $pcov, $pimean;
        foreach my $h (keys %{$hquery_hit{$gene}}) {
            if(exists $hgene{$h} && $exclude{$h} eq 'e') {
                print "S\t*$h\t$hquery_hit{$gene}{$h}\n"; #This print will not work
because it was removed before
            }
            #enter in this loop (with the next command)
        }
        else { #This result will be printed only if the FL gene is covered >=99%
            print "Subject\t$h\t$hquery_hit{$gene}{$h}\n";
        }
    }
}

```

```

    }
}
$sum_alignment_length = 0;
}

my $sum = 0;
print "#hit_pct_cov_bin\tcount_in_bin\t>bin_below\n";
foreach my $pcov (sort {$b <=> $a} keys %hpcovfreq) {
    $sum += $hpcovfreq{$pcov};
    printf "${pcov}0\t$hpcovfreq{$pcov}\t$sum\n";
}

```

```

#Print the Number of Genes by Number of hits
print "#number_of_genes\tnumber_of_hits\n";
foreach my $hits (sort {$a <=> $b} keys %hhitsbygene) {
    printf "$hits\t$hhitsbygene{$hits}\n";
}

```

#Sub-Routines###

```

sub add_interval {
    my ($acc, $inicio, $fim)=@_;
    my (@conjunto,$i_inicio, $i_fim, $i, @r);

    #inicializa hash para gene caso nao exista
    if(!exists $hacc{"$acc"}) {
        $hacc{"$acc"}=();
    }
    else {
        push @conjunto,@{$hacc{"$acc"}};
    }

    ($inicio, $fim) = sort {$a <=> $b} ($inicio, $fim);

```

```

for($i_inicio=0;
    $i_inicio<scalar(@conjunto) && $inicio >= $conjunto[$i_inicio];

```

```

    $i_inicio++)
{}
if($i_inicio && !($i_inicio % 2) && $inicio <= $conjunto[$i_inicio-1]+1) {
    $i_inicio--;
}

for($i_fim=$i_inicio;
    $i_fim<scalar(@conjunto) && $fim > $conjunto[$i_fim];
    $i_fim++)
{}
if($i_fim<scalar(@conjunto) && !($i_fim % 2) && $fim >= $conjunto[$i_fim]-1) {
    $i_fim++;
}

for($i=0; $i<$i_inicio; $i++) {
    push @r, $conjunto[$i];
}

push @r, $inicio if !($i_inicio%2);
push @r, $fim if !($i_fim%2);

for($i=$i_fim; $i<scalar(@conjunto); $i++) {
    push @r, $conjunto[$i];
}

@conjunto=@r;

#hash recebe calculo das coberturas para o gene
$hacc{"$acc"} = [@conjunto];
# print "3 - $acc\t@conjunto\n";
}

```

Anexo XIV – Número de Fatores de Transcrição e Famílias nos transcriptomas do SUCEST e RNA-Seq.

Famílias TFs	TFs RNA-Seq	TFs FL RNA-Seq	TFs SUCEST
ABI3VP1	101	40	79
Alfin-like	97	67	25
AP2-EREBP	172	77	125
ARF	55	47	49
ARID	15	8	8
ARR-B	56	33	24
BBR/BPC	3	2	5
bHLH	171	73	92
bZIP	189	91	85
BZR	7	5	11
C2C2-CO-like	80	43	39
C2C2-Dof	35	14	27
C2C2-GATA	55	20	16
C2C2-YABBY	6	3	18
C2H2	18	9	80
C3H	106	61	135
CAMTA	13	13	8
CCAAT-DR1	0	0	5
CCAAT-HAP2	22	9	6
CCAAT-HAP3	32	14	5
CCAAT-HAP5	0	0	19
CPP	14	9	7
CSD	144	62	23
DBP	234	129	91
E2F/DP	14	4	7
EIL	11	8	9
FAR1-like	339	157	0
FLO/FLY	0	0	1
G2-like	8	3	38
GeBP	21	10	9
GRAS	169	68	77

GRF	40	20	5
Homeobox	84	57	83
HSF	91	50	24
LBD	23	9	16
MADS	35	13	41
mTERF	56	24	0
MYB	46	25	125
MYB-related	261	133	49
NAC	166	90	89
NLP	14	8	11
Ovate	25	4	0
PLATZ	26	6	7
S1Fa-like	1	1	3
SBP	42	21	19
SRS	18	5	3
TCP	19	12	4
TUB	31	21	18
Whirly	4	3	1
WRKY	178	102	50
zf-HD	11	2	6
ZIM	41	20	50
Total	3399	1705	1727

Anexo XV – Representação do número de fatores de transcrição únicos por família para cada espécie

Família TF	Maize (tfome)	Sorghum	S. hybrids (RNA-SeqFL)	S. hybrids (SUCEST)	Rice	Brachypodium
ABI3-VP1	51	60	40	57	55	42
Alfin-like	19	13	67	11	9	10
AP2-EREBP	212	156	77	109	164	141
ARF	38	27	47	48	27	24
ARID	10	7	8	8	6	8
ARR-B	8	10	33	3	6	7
BBR/BPC	4	5	2	4	4	3

bHLH	175	143	73	63	135	124
bZIP	128	91	91	74	92	85
BZR	10	8	5	11	6	7
C2C2-CO-like	14	9	43	2	8	10
C2C2-Dof	46	29	14	25	30	27
C2C2-GATA	36	27	20	13	23	21
C2C2-YABBY	13	8	3	18	8	8
C2H2	10	7	9	12	9	5
C3H	54	44	61	104	46	44
CAMTA	5	6	13	0	6	7
CCAAT	0	32	0	35	0	40
CCAAT-DR1	17	0	0	0	1	0
CCAAT-HAP2	16	9	9	0	11	0
CCAAT-HAP3	4	0	14	0	12	0
CCAAT-HAP5	18	0	0	0	22	0
CPP	13	8	9	13	11	9
CSD	4	0	62	0	0	0
DBP	4	0	129	0	0	0
E2F-DP	19	10	4	7	9	8
EIL	9	7	8	8	9	6
FAR1-like	15	0	157	0	0	0
FLO/LFY	2	0	0	0	0	0
G2-like	59	41	3	40	44	45
GeBP	21	15	10	6	13	14
GRAS	86	76	68	71	60	48
GRF	15	10	20	3	12	11
Homeobox	133	83	57	66	95	82
HRT	0	1		0	1	1
HSF	29	24	50	16	25	24
LBD	44	0	9	0	0	0
MADS	77	76	13	36	70	55
mTERF	30	0	24	0	0	0
MYB	167	113	25	48	114	94
MYB-related	116	80	133	86	71	54
NAC	134	123	90	85	144	100
NLP	17	13	8	11	13	15
Orphans	339	177	0	99	185	163
OVATE	43	0	4	0	0	0

PLATZ	15	0	6	0	0	0
Rcd1-like	10	0	0	0	0	0
S1Fa-like	2	0	1	0	0	0
SBP	32	18	21	16	19	18
SHI/STY (SRS)	9	0	5	0	0	0
Sigma70-like	9	0	0	0	0	0
TCP	44	28	12	3	22	21
Trihelix	41	19	0	6	17	15
TUB	15	13	21	29	15	11
VOZ	5	2	0	4	2	2
WHIRLY	2	3	3	1	2	2
WRKY	125	94	102	48	103	82
ZF-HD	21	15	2	6	15	15
ZIM	36	19	20	49	21	21

Anexo XVI - Categorias Representativas da lâmina de microarranjo

Categorias	#genes
Carbohydrate metabolism	1539
Protein metabolism	1201
Transcription Factor	1186
no match	1137
Protein kinases	1011
Receptors	988
Nucleic acid metabolism	906
Stress response	887
Cell cycle	723
Pathogenicity	695
Amino acid metabolism	554
Transporters	486

Lipid, fatty-acid and isoprenoid metabolism	438
Development	420
Secondary metabolism	414
Bioenergetics	270
Stress	268
Unknown protein	268
Small GTPases	181
Ubiquitination	179
Hormone biosynthesis	168