UNIVERSIDADE ESTADUAL DE CAMPINAS

Instituto de Biologia

KARINA YANAGUI DE ALMEIDA

GENÔMICA E TRANSCRITÔMICA DE *Saccharum spontaneum* E SEUS HÍBRIDOS: NOVAS PERSPECTIVAS PARA COMPREENDER A FORMAÇÃO DA BIOMASSA LIGNOCELULÓSICA

GENOMICS AND TRANSCRIPTOMICS OF *Saccharum spontaneum* AND ITS HYBRIDS: INSIGHTS TO BETTER UNDERSTAND THE LIGNOCELLULOSIC BIOMASS FORMATION

CAMPINAS

2016

KARINA YANAGUI DE ALMEIDA

GENÔMICA E TRANSCRITÔMICA DE *Saccharum spontaneum* E SEUS HÍBRIDOS: NOVAS PERSPECTIVAS PARA COMPREENDER A FORMAÇÃO DA BIOMASSA LIGNOCELULÓSICA

GENOMICS AND TRANSCRIPTOMICS OF *Saccharum spontaneum* AND ITS HYBRIDS: INSIGHTS TO BETTER UNDERSTAND THE LIGNOCELLULOSIC BIOMASS FORMATION

Tese apresentada ao Instituto de Biologia da Universidade Estadual de Campinas para obtenção do Título de Doutora em Genética e Biologia Molecular, na área de Genética Vegetal e Melhoramento.

Thesis presented to the Institute of Biology of the University of Campinas for the degree of PhD in Genetics and Molecular Biology, in the area of Plant Genetics and Breeding.

ESTE ARQUIVO DIGITAL CORRESPONDE À VERSÃO FINAL DA TESE DEFENDIDA PELA ALUNA KARINA YANAGUI DE ALMEIDA E ORIENTADA PELO PROF. DR. GONÇALO AMARANTE GUIMARÃES PEREIRA.

Orientador: PROF. DR. GONÇALO AMARANTE GUIMARÃES PEREIRA

Co-Orientador: DR. JOSÉ ANTÔNIO BRESSIANI

CAMPINAS

2016

Agência(s) de fomento e nº(s) de processo(s): FAPESP, 2012/05890-1; FAPESP, 2015/17045-2; CNPq, 140776/2012-5

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

Y15g

Yanagui, Karina, 1987-
Genômica e transcritômica de *Saccharum spontaneum* e seus híbridos : novas perspectivas para compreender a formação da biomassa lignocelulósica / Karina Yanagui de Almeida. – Campinas, SP : [s.n.], 2016.

Orientador: Gonçalo Amarante Guimarães Pereira.
Coorientador: José Antônio Bressiani.
Tese (doutorado) – Universidade Estadual de Campinas, Instituto de Biologia.

1. *Saccharum spontaneum*. 2. Cultivos agrícolas energéticos. 3. Genoma de planta. 4. Internódio. 5. Transcriptoma. I. Pereira, Gonçalo Amarante Guimarães,1964-. II. Bressiani, José Antônio. III. Universidade Estadual de Campinas. Instituto de Biologia. IV. Título.

Informações para Biblioteca Digital

Título em outro idioma: Genomics and transcriptomics of *Saccharum spontaneum* and its hybrids : insights to better understand the lignocellullosic biomass formation
Palavras-chave em inglês:
Saccharum spontaneum
Energy crops
Genome, Plant
Internode
Transcriptome
Área de concentração: Genética Vegetal e Melhoramento
Titulação: Doutora em Genética e Biologia Molecular
Banca examinadora:
Gonçalo Amarante Guimarães Pereira [Orientador]
Luiz Filipe Protasio Pereira
Carlos Alberto Labate
Odalys García Cabrera
Jorge Maurício Costa Mondego
Data de defesa: 29-08-2016
Programa de Pós-Graduação: Genética e Biologia Molecular

Campinas, 29 de Agosto de 2016.

COMISSÃO EXAMINADORA

Prof. Dr. Gonçalo Amarante Guimarães Pereira

Prof. Dr. Luiz Filipe Protasio Pereira

Prof. Dr. Carlos Alberto Labate

Dra. Odalys García Cabrera

Prof. Dr. Jorge Maurício Costa Mondego

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

Dedico à minha família e à Priscila Mioto por acreditarem incondicionalmente em mim, pois a maior contribuição para realização dos sonhos é acreditar.
Amo vocês.

AGRADECIMENTOS

À Deus, Força Superior ou outras formas que se expresse.

Aos meus pais e irmãos que são a base da minha formação, a fonte de inspiração e força para continuar minha busca constante pelo conhecimento, sem os quais as conquistas não fariam sentido.

À Priscila Mioto, pelo amor e companheirismo incondicionais, por estar ao meu lado durante toda a jornada onde enfrentamos grandes desafios e superações, por sempre incentivar a busca pelo crescimento e pela excelência nos diversos aspectos das nossas vidas. Essa vitória é nossa.

À Ruth da Silva e ao Nelson C. de Carvalho pelo amor, carinho e por me adotarem na minha segunda família. Aos mais que amigos, Cassito, Paty, Letícia, Flávia, Lucas e Natália pelas incontáveis horas de discussões científicas-filosóficas-culturais, pela força nos momentos mais frágeis e por lembrar que o sucesso só é verdadeiro quando dividido.

Ao meu orientador Gonçalo Amarante Guimarães Pereira pelo incentivo, confiança e liberdade para a investigação das propostas, pelo "manda bala *style*" que me permitiu ir muito além do esperado e superar meus limites. Ao co-orientador José Antônio Bressiani pelas contribuições no desenvolvimento do estudo.

Ao Dr. Eduardo Leal Oliveira Camargo, meu malvado favorito, pelo companheirismo, apoio e orientação durante meu doutorado, por não só acreditar nas ideias e perguntas propostas, mas por trabalhar em equipe para realizá-las. Essa tese não seria possível sem você.

Ao Dr. Piotr A Mieczkowski pela oportunidade de aprendizado na *University of North Carolina,* por ter confiado em meu trabalho e apoiado os grandes experimentos que realizamos.

Ao Dr. Michael G. Hahn por me aceitar em seu excelente time no *Complex Carbohydrate Research Center*, *University of Georgia,* pela confiança e apoio para realizar experimentos extensos em um tempo tão reduzido. Ao Dr. Utku Avci pela amizade, por compartilhar tão espontaneamente a sua experiência e excelência

científica. Ao Dr. Sivakumar Pattathil pelas orientações e apoio para a realização dos estudos propostos.

Ao Dr. Marcelo Falsarella Carazzolle pela orientação e direcionamento durante todo o projeto. Aos bioinformatas Leandro Costa do Nascimento, Sheila Tiemi Nagamatsu e Juliana José pelas contribuições no trabalho.

Aos pesquisadores Dr. Carlos Alberto Labate, Dra. Odalys García Cabrera e Dr. Jorge Mauricio Costa Mondego, por aceitar prontamente o convite para participar da banca examinadora e pelas contribuições para a finalização do trabalho. Ao Dr. Luiz Filipe Protasio Pereira pelo acompanhamento, incentivo, exemplo científico durante toda a minha formação e por participar da banca examinadora.

À Bia Temer pelas inumeráveis conversas científicas-cinematrográficas-existenciais embalas a café e açaí, e pela força durante todo meu doutorado. Ao Doug pelas incomensuráveis contribuições no dia a dia do laboratório, pelo companheirismo, por dividir as expectativas, pressões e desafios desse final de jornada, é nóis! Ao casal científico, Thammy e Leandro, pelo apoio e companheirismo!

Aos companheiros de jogatinas e festas legendárias, Tchela, Desis, Bruninho, Jéssica, Sil, Bruna e ao sempre presente Brunão! Obrigada pelo incentivo e torcida não apenas no doutorado! Aos amigos do Laboratório de Genômica e Expressão, que participaram do meu crescimento científico e do desenvolvimento deste trabalho. Em especial, à Eliane Laranja Dias pelo apoio e incentivo durante todo o doutorado.

À FAPESP pelo financiamento do projeto, processo nº 2012/05890-1 e nº 2015/17045-2, Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP).

Ao CNPq, Conselho Nacional de Desenvolvimento Científico e Tecnológico, pelo financiamento do projeto (processo n. 140776/2012-5).

À Universidade Estadual de Campinas e à Pós Graduação em Genética e Biologia Molecular pela oportunidade de formação em ambiente de excelência científica.

E finalmente a todos que contribuíram para que esta tese se tornasse realidade. Muito obrigada!

RESUMO

Bioetanol é uma proeminente fonte de energia renovável e sustentável que representa uma alternativa consistente à utilização de combustíveis fósseis. O desenvolvimento do bioetanol de segunda geração, uma nova tecnologia para a produção desse biocombustível a partir da parede celular vegetal, demanda a utilização de culturas melhoradas para a produção de biomassa. A cana energia é um híbrido entre espécies de *Saccharum* selecionado durante o programa de melhoramento para apresentar alta produção de biomassa e maior conteúdo de fibras do que de sacarose em sua composição. Para explorar os recursos genômicos e o processo de formação da parede celular nos híbridos de alta biomassa, foram elaborados nesse trabalho i) um *draft* do genoma de *S. spontaneum* e ii) uma caracterização de um internódio de cana energia em alongamento. O *draft* do genoma de *S. spontaneum* contém a maioria dos genes completos da espécie, assim como dados de expressão gênica em diferentes tecidos. O genoma constitui uma base de dados pioneira e importante para ancoragem de dados de transcritoma e para o acesso às sequências de genes (éxons e introns) e regiões reguladoras (promotores) de *Saccharum*. Nesse estudo foram identificadas sequências promotoras de genes com expressão tecido-específica e ubíqua com potencial de aplicação imediata para biotecnologia de *Saccharum*. A caracterização de um internódio em alongamento de cana energia demonstrou que este constitui um ótimo modelo para o estudo do controle transcricional da formação da parede celular, em especial da lignificação dos feixes vasculares. Foi observado que o internódio em alongamento pode representar o desenvolvimento dos quatro primeiros internódios imaturos de cana. Além disso, análises comparativas do internódio em alongamento de cana energia com internódios dos parentais, *S. spontaneum* e cana-de-açúcar, indicaram que o internódio em alongamento pode ajudar a explicar as diferenças entre os híbridos selecionados para produção de açúcar ou biomassa. Assim, nesta tese, foi produzido um atlas genômico e transcritômico de *S. spontaneum* e seu híbrido de cana energia, que servirá de base para entender o processo de formação da parede celular de *Saccharum*.

ABSTRACT

Bioethanol is a prominent renewable and green energy source and presents a consistent alternative to the use of fossil fuels. Development of the second generation bioethanol, a new technology to produce this biofuel from plant cell wall, demands the use of improved crops with increased biomass production. Energy cane is a hybrid between *Saccharum* species selected in breeding programs to present elevated biomass production and higher fiber content than sucrose in its composition. Aiming to explore the genomics resources and the process of cell wall formation in high biomass hybrids, in this work it was elaborated i) a draft genome of *S. spontaneum* and ii) the characterization of an elongating energy cane internode. *S. spontaneum* draft genome contains most of the complete genes of the species, as well as genic expression data of different tissues. The genome comprises a pioneer and important *Saccharum* database for transcriptomic data assembly and for access of *Saccharum* sequences of genes (exons and introns) and regulatory regions (promoters). In this study promoter sequences of genes were identified presenting tissue-specific and ubiquitous expression with potential of immediate application in *Saccharum* biotechnology. Characterization of an energy cane elongating internode demonstrates that it constitutes an excellent model for the study of transcriptional control of cell wall formation, especially of vascular bundle lignification. It was observed that the elongating internode could represent the development of the first four immature internodes. Comparative analysis between elongating internodes of energy cane with internodes from parental genotypes, *S. spontaneum* and sugarcane, indicated that the elongating internode could explain the differences between hybrids selected for sugar or biomass production. In this work, it was produced a genomic and transcriptomic atlas of *S. spontaneum* and its high biomass hybrid that will be useful to understand the cell wall formation l in *Saccharum*.

PREFÁCIO

Esta tese está estruturada em dois capítulos contendo introdução, material e métodos, resultados e discussão. No capítulo 1 descrevemos a construção de um genoma *draft* de *Saccharum spontaneum,* um parental de cana energia que apresenta as características desejáveis para a biomassa lignocelulósica, especialmente alto conteúdo de fibra e rendimento de biomassa. O *draft* é composto pela maioria dos genes (contendo éxons, íntrons, UTRs) e regiões regulatórias (promotores) da espécie. O banco de dados de *S. spontaneum* nesse estudo também inclui transcritomas (RNAseq) de diferentes tecidos. O genoma produzido se mostrou uma ferramenta efetiva para ancorar a montagem de transcritomas e para a identificação de ferramentas biotecnológicas, como os promotores ubíquos e tecido-específicos descritos nesse estudo. A partir desses resultados foram elaboradas as patentes “Promotores constitutivos e tecido-específico de *Saccharum*” e “Método para a montagem de regiões de genes completos de genomas complexos, e uso do mesmo” que se encontram nos anexos, da qual participo como primeira autora na primeira e como co-autora na segunda.

No capítulo 2 investigamos a formação da biomassa de um híbrido pré-comercial de cana energia pela caracterização histológica e transcritômica de um internódio em fase de alongamento. O internódio apresenta um gradiente variando desde a divisão celular até a lignificação completa do feixe vascular, por isso se mostrou um ótimo modelo para controle transcricional da formação da parede celular. Foi possível identificar genes diferencialmente expressos que caracterizam as etapas do desenvolvimento da parede celular (divisão celular, alongamento e lignificação). Assim, o padrão observado nesse internódio em alongamento pode sintetizar o desenvolvimento dos internódios imaturos de cana e fornecer novas perspectivas para a formação da biomassa e as diferenças entre híbridos de *Saccharum* selecionados para diferentes propósitos (produção de açúcar ou biomassa).

Antes dos capítulos é realizada uma introdução geral, seguida da revisão teórica que descreve o gênero *Saccharum* e os novos híbridos de cana energia. Também é destacada a importância dos estudos genômicos e de caracterização da biomassa, bases fundamentais para o entendimento dos híbridos de cana energia, objetivo principal da tese que perpassa ambos os trabalhos.

SUMÁRIO

1 INTRODUÇÃO

O gênero *Saccharum* tem sido explorado nos últimos séculos pelo emprego de seus híbridos de cana-de-açúcar na produção de açúcar e etanol majoritariamente. Com a necessidade de alternativas renováveis para substituição dos combustíveis fósseis e o desenvolvimento de novas tecnologias para a produção de bioetanol a partir de materiais lignocelulósicos (etanol de segunda geração) o gênero *Saccharum* passa a ter uma importância estratégica como fonte de biomassa lignocelulósica através do melhoramento de novos híbridos, a cana energia (TEW; COBILL, 2008; DOS SANTOS et. al., 2016).

Cana energia são híbridos de *Saccharum* selecionados para apresentar alta produtividade e um maior conteúdo de fibras do que de açúcares solúveis em sua composição (MATSUOKA et al., 2014). O melhoramento da cana energia baseia-se em estratégias tradicionais de cruzamentos entre *Saccharum officinarum* (e seus híbridos de cana-de-açúcar) e *Saccharum spontaneum*, uma espécie ancestral do gênero com características desejáveis na biomassa lignocelulósica, como alta capacidade de perfilhamento, alto rendimento de biomassa, maior conteúdo de fibras em sua composição e resistência a estresses bióticos e abióticos (TEW; COBILL, 2008; MATSUOKA et al., 2014). Ao contrário da cana-de-açúcar, que tem sido alvo do melhoramento há algumas décadas, o foco no desenvolvimento da cana energia é recente nos programas de melhoramento (DOS SANTOS et. al., 2016) e existem poucos estudos genômicos publicados sobre *S. spontaneum* e seus híbridos de cana energia.

Nesse contexto, o presente estudo realizou a caracterização dos recursos genéticos de *S. spontaneum* e cana energia pela construção de um “genoma funcional” de *S. spontaneum* e pela análise transcritômica da formação da biomassa em um internódio de cana energia. Essas abordagens permitiram a construção de um banco de dados importante para aprofundar o entendimento sobre os novos híbridos de alta biomassa e para a descoberta de ferramentas de biotecnologia para auxiliar o melhoramento da cana energia.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 ORIGEM E ASPECTOS TAXONÔMICOS DE SACCHARUM SPP.

O gênero *Saccharum* pertence à família Poaceae (gramíneas), subfamília Panicoideae, tribo Andropogoneae e subtribo Saccharinae. Esse gênero é composto por seis principais espécies: *S. spontaneum* (2n = 40-128), *S. officinarum* (2n = 80), *S. robustum* (2n = 60-80), *S. edule* (2n = 60-80), *S. barberi* (2n = 81-124) e *S. sinense* (2n = 111-120) (D'HONT et al., 1998; AMALRAJ &BALASUNDARAM, 2006). O centro de origem e diversidade das espécies *S. officinarum*, *S. robustum e S. edule é o* Sudeste Asiático, principalmente a Ilha de Nova Guiné. As espécies *S. barberi* e *S. sinense* são originárias do Norte da Índia e China. *S. spontaneum* tem como centro de origem a Índia, no entanto, sua distribuição geográfica abrange a região do Mediterrâneo, África e Nova Guiné, se configurando como a espécie com maior diversidade e distribuição geográfica do gênero (D'HONT et al., 1998).

A origem das espécies atuais é controversa pois podem ter ocorrido diversos eventos de hibridação entre espécies do gênero e de outros gêneros próximos, tornando difícil à distinção entre a contribuição genética proveniente de cada um destes cruzamentos (D'HONT et al., 1998; IRVINE, 1999; MING et al., 2001; GRIVET,2004 AMALRAJ &BALASUNDARAM, 2006). Postula-se que *S. spontaneum* e *S. robustum* juntamente com algumas espécies dos gêneros *Erianthus*, *Miscanthus*, *Sclerostachya* e *Narenga*, constituem a base genética que originou as cultivares atuais e, portanto, foram classificadas em um grupo taxonômico 'informal' conhecido como complexo *Saccharum* (HODKINSON, et al. 2002; AMALRAJ &BALASUNDARAM, 2006).

A classificação taxonômica do gênero *Saccharum* tem sido amplamente discutida (D'HONT et al., 1998; IRVINE, 1999; MING et al., 2001; GRIVET,2004 AMALRAJ &BALASUNDARAM, 2006). A despeito de ser aceita a divisão de *Saccharum* em seis espécies, IRVINE (1999) propôs a divisão do complexo apenas em *S. spontaneum* e *S. officinarum* (que englobaria os híbridos e as outras quatro espécies). Essa visão também é reforçada por análises de diversidade genética e estrutura da população atuais (com uma coleção de mais de 1000 acessos do Complexo *Saccharum*), que indicaram a divisão do complexo em três grandes *clusters*: *S. spontaneum* formando o primeiro grupo, *S. officinarum* e os híbridos

modernos de *Saccharum* no segundo grupo, enquanto o terceiro agrupamento é formado, majoritariamente, por gêneros não pertencentes a *Saccharum* (como *Erianthus* e *Miscanthus*) (NAYAK et al., 2014). O cluster contendo predominantemente genótipos de *S. spontaneum* é o que apresenta maior diversidade genética (NAYAK et al., 2014).

Considerando as características fisiológicas e históricas das espécies é possível dividi-las em três grupos: ancestrais, cultivares tradicionais e cultivares modernos. As espécies ancestrais, *S. spontaneum* e *S. robustum*, apresentam baixo conteúdo de açúcar e alto teor de fibras, e são importantes para os programas de melhoramento das cultivares modernas de cana-de-açúcar, e mais recentemente cana energia (MING et al., 2001; GRIVET, 2004; MATSUOKA et al., 2014). As cultivares tradicionais descendem de linhagens de domesticação primária e podem ser divididas em dois grupos: no primeiro estão as cultivares nobres (*S. officinarum*) e no segundo estão inclusas as cultivares do Norte da Índia e da China (*S. barberi* e *S. sinense)*. As canas nobres apresentam alta quantidade de açúcares solúveis no colmo e ainda são utilizadas na agricultura, já as cultivares chinesas e do Norte da Índia apresentam menor conteúdo de açúcar e permanecem restritas a bancos de germoplasma (MING et al., 2001; GRIVET,2004). As cultivares modernas, conhecidas como cana-de-açúcar, são híbridos oriundos de sucessivos cruzamentos entre os cultivares tradicionais e *S. spontaneum* que foram submetidos e extensos ciclos de seleção visando a maior produção de açúcares solúveis (GRIVET, 2004). Mais recentemente outras cultivares modernas denominadas cana energia, selecionadas para a produção de biomassa, estão sendo melhoradas através de sucessivos cruzamentos entre os cultivares tradicionais e espécies ancestrais, notadamente *S. spontaneum*, seguidos de retrocruzamentos com as espécies ancestrais (MATSUOKA et al., 2014).

O número de cromossomos das espécies do gênero *Saccharum* varia entre n=24-128, compreendendo de 8-18 cópias de um conjunto básico de x=10 (*S. officinarum* e *S. robustum*) ou x=8 (*S. spontaneum*) (GRIVET et al., 1996; D'HONT et al., 1998; MING et al., 2001). Por causa das diferenças do número básico de cromossomos de *S. officinarum* e *S. spontaneum* o genoma dos híbridos apresentam uma organização cromossômica complexa e eventos de recombinação (D'HONT et al.1998). A maioria das cultivares modernas de cana-de-açúcar

comumente utilizadas na agricultura contém 75-85% dos cromossomos derivados de *S. officinarum* e 15-25 de% *S. spontaneum* (D'HONT, 2005) (Figura 1).

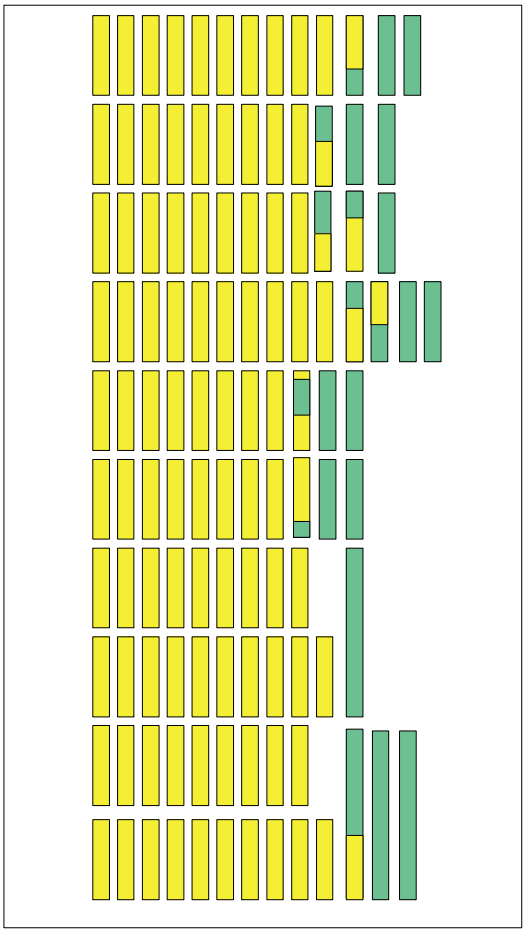


Figura 1- Esquema representativo do genoma de um cultivar moderno de cana-de-açúcar. Os cromossomos são representados por barras, as amarelas correspondendo às regiões genômicas provenientes de *S. officinarum* e as verdes de *S. spontaneum*. . Pela distribuição dos cromossomos é possível observar que os genomas dos híbridos modernos apresentam aneuploidias, poliploidia e recombinações dos cromossomos provenientes de espécies parentais de *Saccharum*. Retirado de Grivet e Arruda, 2001.

O tamanho do genoma (2C) nas espécies de *Saccharum* pode variar de 2,5 a 12,5 Gb. Para algumas espécies como *S. officinarum* essa amplitude de variação é menor (a maioria dos genótipos possui o genoma contendo 6,5-8,5 Gb), já *S. spontaneum* pode ter genótipos com genomas variando de 2,5 a 12,5 Gb (Figura 2) (ZHANG et al., 2012). A alta complexidade do genoma apresentando múltiplas copias, aneuploidia e recombinação de cromossomos torna o sequenciamento do genoma e acesso as informações genéticas de *Saccharum spp*. particularmente difíceis.

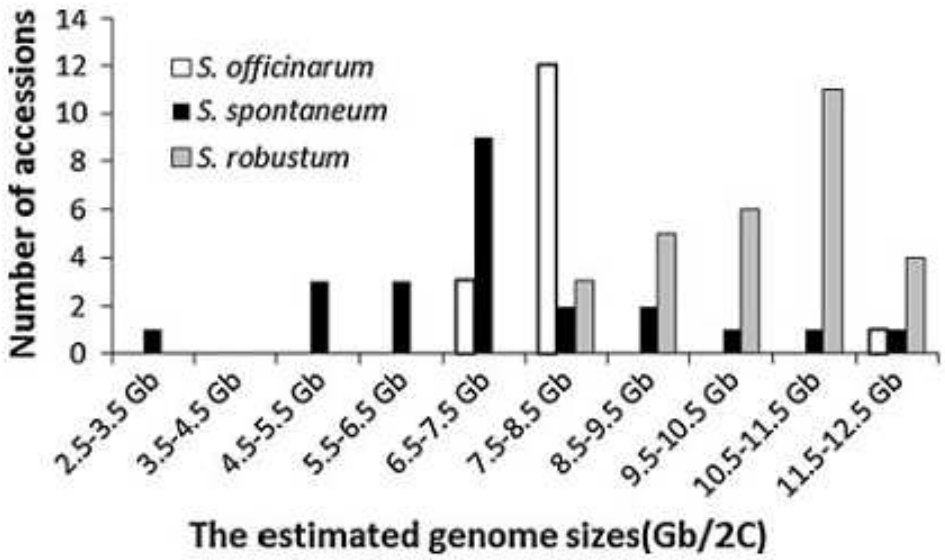


Figura 2: Tamanho estimado do genoma de acessos de *S. spontaneum*, *S. officinarum* e *S. robustum*. Retirado de ZHANG et al. (2012).

As espécies que possuem o genoma sequenciado com relações filogenéticas mais próximas a *Saccharum spp.* são o sorgo (*Sorghum bicolor* ou *Sorghum vulgare*) e o milho (*Zea mays*), ambos pertencendo à tribo Andropogoneae (PATERSON et al. 2009). Sorgo é um modelo de estudo para espécies de *Saccharinae* e outras gramíneas C4, e o arroz é o modelo utilizado para plantas C3. O tempo de divergência entre o sorgo e *Saccharum* é de aproximadamente 8-9 milhões de anos. Após a divergência, as espécies de *Saccharum* sofreram pelo menos mais duas duplicações genômicas completas para alcançar o nível de ploidia atual (JANNOO, N. et al. 2007; PATERSON et al., 2009). Estudos de sintenia identificaram várias regiões genômicas conservadas entre cana-de-açúcar e sorgo (JANNOO, N. et al. 2007; GARSMEUR, O. et al. 2011). A colinearidade entre loci ortólogos com milho e sorgo é importante para auxiliar na localização de importantes genes e grupos de ligação da cana-de-açúcar (GRIVET et al., 1996).

2.2 IMPORTÂNCIA ECONÔMICA DO GÊNERO SACCHARUM

O Brasil é o maior produtor mundial de cana-de-açúcar, seguido de Índia, China, Tailândia e Paquistão respectivamente (FAO, 2013). A cana-de-açúcar e seus derivados são a segunda maior fonte de energia primária da matriz energética nacional, ficando apenas atrás do petróleo (Ministério da Agricultura, Pecuária e Abastecimento, 2009).

No país, a cana ocupa aproximadamente 9 milhões de hectares, equivalente a 2% da área arável (ÚNICA, 2010; FAO, 2013). A produção estimada da cultura para a safra 2016-2017 é de 690,98 milhões de toneladas (CONAB, 2016). A atividade canavieira abrange 62 mil estabelecimentos produtores e sua cadeia produtiva envolve aproximadamente 670 mil trabalhadores (CNA, 2012). A região centro-sul é responsável por 90% da produção nacional e os principais estados produtores são respectivamente São Paulo, Paraná, Bahia e Goiás (ÚNICA, 2012); apenas o estado de São Paulo responde por 68,6% da produção no Centro Sul (NEVES; TROMBIN; CONSOLI, 2011). A produtividade média da cultura é de 85 ton/ha (com variação de 65-120 ton./ha) (Ministério da Agricultura, Pecuária e Abastecimento, 2009).

Majoritariamente a cana-de-açúcar é empregada para produção de açúcar e etanol. A produção estimada de açúcar e etanol para a safra atual é 37,5 milhões e 30 bilhões de litros, respectivamente (CONAB, 2016). Além de ser o maior produtor de cana-de-açúcar, o Brasil também se destaca como grande exportador do açúcar. Nos anos de 2014 e 2015, o rendimento em exportações foi maior que U$ 600 milhões. (CONAB, 2016). O país também abriga o terceiro maior mercado consumidor mundial de açúcar. Além da utilização da cana para produção de açúcar e etanol, a cultura canavieira também é empregada para geração de eletricidade e para a produção de diversos produtos como bebidas (caldo de cana, cachaça, rum), papéis e fármacos (a partir do bagaço), poliestireno, estireno, cetoaldeído, ácido acético, éter e cetona (a partir do etanol) e fertilizantes (a partir da vinhaça e do vinhoto) (DOMINGUES, 2009). Ademais da importância econômica e social já consolidada dos híbridos de *Saccharum* para o país, novas tecnologias para a produção de bioetanol lignocelulósico estão impulsionando o melhoramento de híbridos como fonte de biomassa para viabilizar economicamente a produção desse biocombustível no país (Dos SANTOS, et al., 2016).

2.3 Biocombustíveis e o Etanol de Segunda Geração

Os combustíveis fósseis (petróleo, carvão e gás natural) representam 82% da matriz energética mundial; o petróleo sozinho respondendo por 32% da matriz, se consolidando como principal fonte energética (IEA, 2013). No entanto, preocupações sobre a extensiva exploração das reservas de petróleo com o aumento do custo para o acesso e exploração do combustível, sobre a localização das principais reservas em regiões de instabilidade política gerando flutuações no preço e sobre os efeitos ambientais desencadeados pelo aumento da emissão de gases de efeito estufa com consequente contribuição para o aquecimento global tem motivado a busca de fontes de energias alternativas. As fontes alternativas, como os biocombustíveis, devem ser sustentáveis, capazes de assegurar o suprimento e evitar a instabilidade econômica, minimizar os impactos ambientais e não comprometer a produção de alimentos (IEA, 2008; Banco Central do Brasil, 2012), LEITE; LEAL; CUNHA, 2013).

Os biocombustíveis englobam uma variedade de matérias-primas, tecnologias de conversão e usos. Eles são empregados, em sua maioria, para o transporte e para produção de eletricidade. Os biocombustíveis para o transporte, como o etanol e o biodiesel, figuram entre as fontes de energia renovável com o crescimento mais rápido e promissor atualmente (BNDES & CGEE, 2008). As principais fontes vegetais utilizadas na produção do bioetanol são o milho, nos Estados Unidos, a cana-de-açúcar no Brasil e o trigo e a beterraba nos países da Europa Ocidental, sendo que as culturas do milho e da cana-de-açúcar respondem por 80% do mercado mundial. Na Europa Ocidental e Estados Unidos o custo de produção do etanol é duas e quatro vezes, respectivamente, mais elevado do que no Brasil (HILL et al., 2006, BNDES & CGEE, 2008). A alta produção e a posição de destaque do Brasil no mercado do bioetanol foram asseguradas pela elevada capacidade da cana para fixação do carbono e pelas iniciativas governamentais que disponibilizaram subsídios para viabilizar o início da produção de etanol no país.

O bioetanol produzido a partir de amido e da sacarose é conhecido como etanol de primeira geração (1G), enquanto aquele produzido a partir de material lignocelulósico é denominado etanol de segunda geração (2G) (MATSUOKA; FERRO; ARRUDA, 2009). Uma das desvantagens na produção de bioetanol a partir de amido e sacarose é que estas matérias-primas possuem mercados alternativos

(alimentos, insumos) mais remunerados, dessa forma pode haver conflitos entre a produção de alimentos e combustíveis (BNDES & CGEE, 2008). Além disso, a produção do etanol 1G é geograficamente limitada, já que nem todas as regiões são propícias à produção de culturas para este fim (BNDES & CGEE, 2008).

Nesse contexto diversos esforços estão sendo realizados para produção de bioetanol a partir da hidrólise de materiais lignocelulósicos, pois este poderia então ser produzido em praticamente todas as regiões do mundo a partir de diversas fontes de resíduos orgânicos (BNDES & CGEE, 2008). Essa abordagem se baseia na hidrólise de polissacarídeos (celulose e hemicelulose) da biomassa em açúcares solúveis, que poderão ser fermentados para produção do etanol. Diversas estratégias estão em desenvolvimento para viabilizar a produção do etanol lignocelulósico (DOS SANTOS, et al., 2016). Dentre elas podemos destacar o aperfeiçoamento da tecnologia de hidrólise, a produção de linhagens de microrganismos adaptadas e otimizadas às diferentes etapas do processo, e o desenvolvimento de variedades vegetais comerciais (KARP;SHIELD, 2008, SCHUSTER; CHIN, 2013; DOS SANTOS, et al., 2016). Várias empresas já possuem inclusive plantas em escala industrial produzindo bioetanol de segunda geração, e outras estão com iniciativas em andamento para iniciar a produção (BATISTA, 2013; LANE, 2013; PENNENERGY, 2013; DOS SANTOS, et al., 2016). A primeira planta em escala comercial de etanol 2G iniciou suas operações em 2013 na Itália (Crescentino) com capacidade para produzir 75 milhões de litros/ano de etanol. Em 2014 as plantas de etanol celulósico da POET e DSM (Emmetsburg, IA) e Abengoa's (Seville, Spain), com capacidade para 94,5 e 95 milhões de litros/ano começaram a funcionar. Em 2015 a maior planta foi inaugurada pela DuPont (Iowa, Nevada) com capacidade para produzir 120 milhões de litros de etanol por ano (DOS SANTOS, et al., 2016).

No Brasil, a primeira planta do hemisfério sul começou a operar em 2014 construída pela empresa GranBio (São Miguel dos Campos, Alagoas) com capacidade para 82 milhões de litros/ano. Em 2015, a segunda planta foi inaugurada pela Raizen em Piracicaba visando a produção de 40 milhões de litros/ano e outras iniciativas também estão planejadas ou em construção (DOS SANTOS, et al., 2016). Todas as usinas brasileiras de etanol 2G em operação ou planejadas irão operar com palha, bagaço ou biomassa total de híbridos de *Saccharum* spp. Inicialmente a produção será baseada em cana-de-açúcar, entretanto várias iniciativas já buscam a

produção de uma variedade de *Saccharum* dedicada à produção de bioetanol lignocelulósico, a cana energia. Atualmente, empresas como GranBio e Vignis, e centros de pesquisa como IAC, RIDESA e CTC já possuem um programa de melhoramento focado no desenvolvimento e lançamento de variedades de cana energia (DOS SANTOS, et al., 2016).

2.4 CANA ENERGIA

Cana energia são híbridos de *Saccharum* spp. selecionados para apresentar maior conteúdo de fibra do que de sacarose em sua composição e alta produção de biomassa (MATSUOKA et al., 2014). Comparada à cana-de-açúcar convencional, a cana energia apresenta aproximadamente o dobro de fibras e de rendimento de biomassa (ton/ha) e metade da quantidade de açúcares solúveis (Figura 3) (DOS SANTOS, et al., 2016). Além de produzir grande quantidade de biomassa a baixo custo, a cana energia apresenta outras vantagens, como o baixo requerimento de fertilizantes, a alta resistência a estresses bióticos e abióticos, a alta taxa de propagação e de perfilhamento, a facilidade no transporte e a possibilidade de ser produzida em área marginais atualmente não utilizadas para produção das cultivares tradicionais, o que evitaria o *trade off* com a produção de alimentos (HASSUANI et al. 2005; TEW; COBILL, 2008; SHIELDS; BOOPATHY, 2011; DOS SANTOS, et al., 2016).

Table 1. Comparative Data Between Sugarcane and Energy Cane

CHARACTERISTICS	SUGARCANE	ENERGY CANE
Fibers	17.4%	33%
Soluble sugars	12.6%	5%
Metric tons wet weight/ha	92	180
Metric tons dry weight/ha	16	59
Fertilizer requirements	High	Low
Disease and pest resistance	Low	High
Number of ratoons	5	10
Propagation ratio	1:10	1:100
Tillering	Low	High
Breeding cycle (years)	8–12	4–6

Figura 3: Análise comparativa entre híbridos de cana-de-açúcar e cana energia. A cana energia apresenta aproximadamente metade da sacarose e o dobro de fibras e biomassa que a cana-de-açúcar convencional. Retirado de DOS SANTOS et al. (2016).

O conceito de cana energia surgiu nos anos 70, em Louisiana, quando melhoristas propuseram o desenvolvimento de uma variedade de cana com alto rendimento de biomassa para assegurar a produtividade da cultura que estava em crise na região (MATSUOKA et al., 2014). Alexander e seu grupo observaram que a métrica normalmente utilizada para caracterizar a qualidade das cultivares, a produção de sacarose, não era o melhor parâmetro para avaliar as plantas de *Saccharum*, que tem como característica o alto crescimento e produção de biomassa (MATSUOKA et al., 2014). Entretanto a ideia não foi bem recebida na época, pois essa variedade geraria o dobro de bagaço que as convencionais e demandaria maiores custos para seu transporte, sendo que nessa época o bagaço ainda era um resíduo não aproveitado. Atualmente, com o esforço para viabilização da produção de etanol 2G e o surgimento das plantas em escala comercial, a cana energia se torna uma fonte de biomassa promissora para assegurar a viabilidade econômica do modelo (DOS SANTOS et al. 2016).

A cana energia, além do maior conteúdo de fibra e menor de sacarose, também possui algumas caraterísticas importantes que distinguem das cultivares tradicionais, como sistema radicular vigoroso (Figura 4a), colmos finos e longos, formação de rizomas (Figura 4b e 4c), ciclo de produção (rebrota) de mais de dez anos e grande capacidade de perfilhamento (que aumenta a taxa de multiplicação visto que essa é realizada pelo plantio das gemas presentes nos colmos) (MATSUOKA et al., 2014).

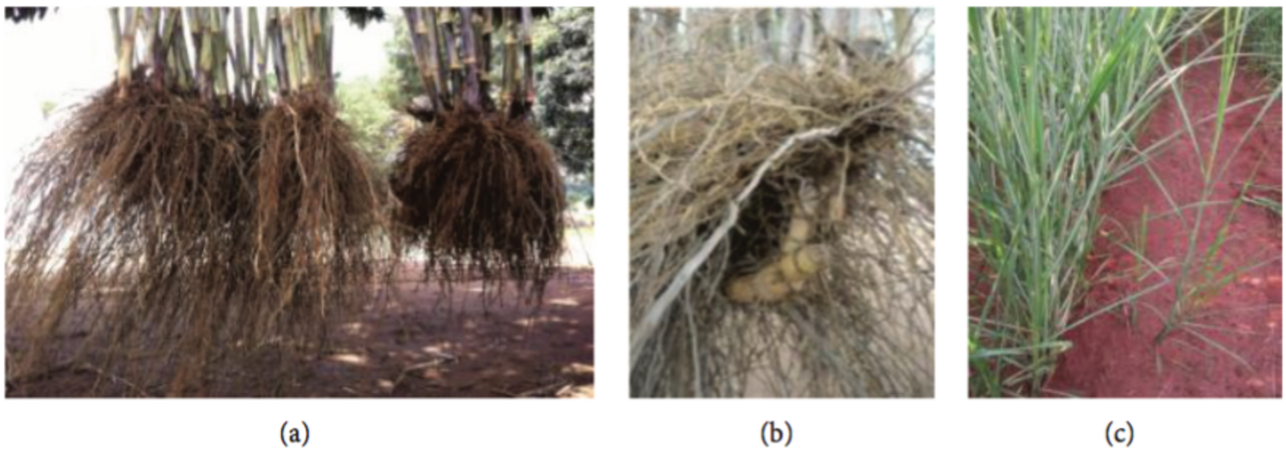


Figura 4: Características fenotípicas da cana energia. a) Sistema radicular de cana energia bem desenvolvido (esquerda) quando comparado à cana-de-açúcar (direita); b) Rizoma de cana energia (normalmente ausente em híbridos de cana-de-açúcar; c) Brotamento de um rizoma de cana energia ao lado da touceira original. Retirado de MATSUOKA et al. (2014).

O melhoramento da cana energia é baseado em cruzamentos intraespecíficos de *Saccharum*, principalmente entre *S. spontaneum* e *S. officinarum* e seus híbridos de cana-de-açúcar. Entretanto, ao contrário do que ocorreu no processo de nobilização da cana-de-açúcar, onde os retrocruzamentos eram feitos com *S. officinarum* para aumentar a capacidade de acumulo de sacarose, para cana energia são realizados retrocruzamentos com *S. spontaneum* para aumentar o conteúdo de fibra e biomassa (TEW; COBILL, 2008; MATSUOKA et al., 2014).

S. spontaneum naturalmente é caracterizado por apresentar alto teor de fibra e menor percentagem de sacarose. Essa espécie é altamente polimórfica, adaptável e apresenta a mais abrangente distribuição geográfica do gênero *Saccharum*. Em certos países (Tailândia, Índia, Filipinas e Indonénia) *S. spontaneum* era classificado como praga por ser um forte competidor com espécies comerciais devido a sua alta capacidade de adaptação ao frio, a seca e outras condições que dificultam o crescimento das plantas. Esta espécie também é altamente vigorosa e resistente a várias doenças (CORDEIRO et al. 2001).

A estratégia de melhoramento tem implicações diretas no conteúdo genético dos híbridos. As cultivares de cana-de-açúcar apresentam aproximadamente 20% do conteúdo do genoma de *S. spontaneum* e 80% de *S. officinarum* (D'HONT, 2005), enquanto espera-se que os híbridos de cana energia apresentem uma contribuição muito maior de *S. spontaneum* no seu genoma (DOS SANTOS et al., 2016), reforçando a importância de estudos para caracterização dos recursos genômicos dessa espécie.

2.5 Sequenciamento de Genomas para Melhoramento de Culturas

O primeiro genoma de planta publicado foi o de *Arabidopsis thaliana* em 2000, deste então diversos genomas vegetais foram produzidos (como arroz, milho, trigo, sorgo, pepino, banana, batata) somando mais de 100 genomas de plantas sequenciados atualmente (DA SILVA et. al., 1993, The Arabidopsis Genome Iniciative, 2000, GOFF et. al. 2002, HUANG et al., 2009, HRIBOVA et al., 2009, PATERSON et. al., 2009, The Potato Genome Sequencing Consortium, 2011; VEECKMAN; RUTTINK; VANDEPOELE, 2016). Inicialmente a montagem de um genoma despendia muito tempo e recursos. No entanto, a consolidação de

tecnologias de sequenciamento *high throughput* tornou o sequenciamento de genomas mais rápido e acessível (MARDIS, 2008).

A exploração dos dados genômicos, contendo todos os genes de um organismo, possibilita a identificação de genes candidatos e o entendimento de como importantes características de interesse agronômico são controladas (MATSUOKA; FERRO; ARRUDA, 2009; EDWARDS & BATLEY, 2010). A detecção de genes candidatos permite a aplicação de estratégias para alterar (aumento/diminuição) a expressão dos mesmos, para transferir essas características entre cultivares ou espécie, ou mesmo, para acelerar o processo de melhoramento tradicional pela seleção assistida por marcadores. Em todos os casos é possível diminuir o tempo e os custos para a obtenção de uma nova cultivar com as características desejáveis. No caso de características com heranças mais complexas, com o envolvimento de vários genes, a montagem do genoma da espécie é o primeiro passo para acessar as informações genéticas e desvendar as relações existentes entre os genes, fornecendo ferramentas para acelerar o melhoramento e a engenharia genética das culturas (EDWARDS & BATLEY, 2010).

Para as plantas poliploides, no entanto, a definição do número de alelos em determinado *locus* ou discriminação entre alelos de um loci parálogo (*loci* derivado de uma duplicação do genoma) é particularmente difícil de revelar e dificulta o processamento correto das informações e a montagem do genoma (GRIVET et al., 1996).

No caso das *Saccharum spp*. esta dificuldade é ainda maior: a poliploidia, a aneuploidia, o número de cromossomos variado e a origem multiespecífica dos híbridos tornam o genoma da cana um dos mais complexos (GUIMARÃES; SILLS; SOBRAL, 1997). Ademais não existem progenitores diploides de *Saccharum* spp. que tornariam o sequenciamento e a montagem do genoma mais acessíveis (GARCIA el al. 2013). Diversos trabalhos foram desenvolvidos para acessar as informações genéticas da cultura: prospecção de marcadores e mapeamento genético, bibliotecas de ESTs representando o transcriptoma da espécie, identificação de QTLs, construção de bibliotecas de BACs e mais recentemente esforços para o sequenciamento do genoma de híbridos de cana-de-açúcar (tanto por estratégias de clonagem envolvendo BACs quanto por sequenciamento *shotgun* de todo o genoma) (AL-JANABI et al., 1993, DA SILVA et al.1993, CORDEIRO; TAYLOR; HENRY, 2000, CORDEIRO et al. 2001, VETTORE et.al. 2003, Sugarcane

Genome Project, FIGUEIRA et al. 2012, DE SETTA et al. 2014; GRATIVOL et al. 2014). Apesar de todos os avanços nos estudos genéticos de *Saccharum spp.* não existe ainda um genoma de referência para o gênero.

Uma abordagem para acessar as informações genômicas de plantas complexas como a cana, é a produção de um "genoma funcional" compreendendo as sequências de genes (éxons e introns). Essa estratégia, adotada no presente estudo, não proporciona uma definição precisa dos alelos como em um genoma completo que contém as regiões gênicas e intergênicas e suas posições nos cromossomos determinadas, mas é suficientemente informativa para permitir o acesso as sequências de genes de interesse e auxiliar na aplicação de outras metodologias, como a montagem de transcritos em larga escala (RNAseq).

Além do acesso às informações genômicas de *S. spontaneum,* com a produção de um genoma *draft* desse importante progenitor do melhoramento de híbridos de alta biomassa de *Saccharum*, este trabalho também buscou caracterizar a dinâmica da formação da parede celular dos novos cultivares de cana energia.

2.6 FORMAÇÃO DA BIOMASSA

A parede celular vegetal é uma matriz complexa e dinâmica de polissacarídeos (celuloses, hemicelulose e pectinas), lignina e glicoproteínas. A percentagem desses polímeros e a interação entre eles na parede celular são influenciados por uma série de fatores como crescimento celular, estágio de desenvolvimento da planta (FRESHOUR et al. 1996), tipo do tecido (De SOUZA et al. 2013), espécie (PATTATHIL et al. 2013), genótipos (De SOUZA et al. 2013) e fatores ambientais (MING et al. 2006) A composição dos polímeros da parede celular impactam diretamente no processo ao qual a biomassa será submetida. Por exemplo, na produção do etanol lignocelulósico, a natureza da biomassa pode influenciar na recalcitrância do material às enzimas hidrolíticas diminuindo a eficiência da hidrólise e o rendimento de etanol (De SOUZA et al. 2015), assim como liberar inibidores do processo de fermentação (JÖNSSON; ALRIKSSON, & NILVEBRANT, 2013).

Para um híbrido de cana-de-açúcar, por exemplo, foi relatado que a parede celular é composta de 28 % de celulose, 58 % de hemiceluloses, 8% de pectinas e 6 % de lignina (De SOUZA et al. 2015). Essas percentagens, assim como

os polímeros que compõe cada classe podem ser variáveis entre genótipos de *Saccharum* spp (MING et al. 2006). Além das diferenças entre os indivíduos, o estágio de desenvolvimento dos tecidos também pode influenciar na variação nos componentes da parede celular (LINGLE, 1997; LINGLE & THOMSON, 2012). Um estudo que caracterizou as principais classes de polímeros (celulose, hemicelulose e lignina) durante o desenvolvimento de internódios de cana-de-açúcar mostrou uma dinâmica na composição da parede durante a maturação e o crescimento dos internódios. Foi observado que durante o desenvolvimento de um internódio existe uma diminuição no conteúdo de hemicelulose (5a), enquanto há um aumento nos açúcares totais, na celulose (5b) e na lignina (5c) ao longo do processo (LINGLE, 1997; LINGLE & THOMSON, 2012). Esse padrão demonstra que entre os internódios imaturos (representados nas semanas de 1-4 nas figuras 5a-5d) há uma variação dos componentes da parede celular maior que a observada entre os internódios maduros, tornando-os especialmente interessantes para a compreensão do processo de formação da biomassa de *Saccharum*.

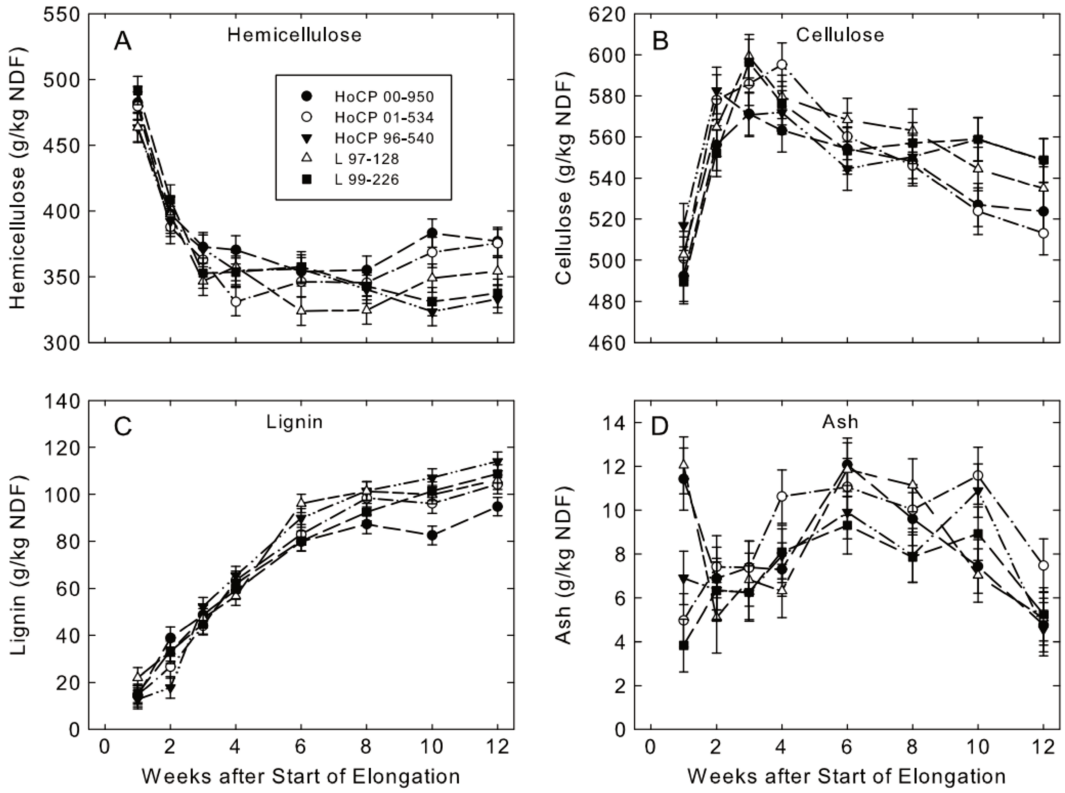


Figura 5. Análise dos componentes da parede celular de cana-de-açúcar durante o desenvolvimento do internódio (mensurado ao longo das semanas após o início do alongamento do internódio). (a) hemicelulose, (b) celulose (c), lignina e (d) cinzas. Retirado de Lingle et al. (2012).

Diversos estudos analisaram as modificações que ocorrem durante o desenvolvimento dos internódios de *Saccharum*. A maioria deles caracterizou o metabolismo da sacarose entre internódios imaturos e maduros, principalmente por ser esse componente o mais desejável no melhoramento de cana-de-açúcar (LINGLE; SMITH, 1991, LINGLE, 1997; CASU et al., 2007, LINGLE; TEW, 2008, PAPINI-TERZI et al., 2009, LINGLE; THOMPSON, 2012). Destes trabalhos, os que focaram na avaliação do metabolismo de sacarose pela expressão dos genes envolvidos na via, utilizaram as técnicas de microarranjo e PCR quantitativo (CASU et al., 2007; PAPINI-TERZI et al., 2009). Recentemente a diferença entre internódios imaturos e maduros de cana-de-açúcar também foi acessada pela caracterização da lignificação nesses tecidos por análises histológicas e de PCR quantitativo de alguns genes envolvidos no metabolismo da lignina (BOTTCHER et al. 2013). No entanto, nenhum estudo sobre a expressão de transcritos em internódios em desenvolvimento de *Saccharum* foi realizado pelo sequenciamento com tecnologias de alto desempenho (RNAseq).

Estudos sobre as mudanças ocorridas durante o desenvolvimento dos internódios também foram realizados em outras espécies de gramíneas como milho (*Zea mays*) e bambu (*Bambusa oldhamii*) e auxiliaram no entendimento das alterações que ocorrem na parede celular dos internódios, fornecendo importantes *insights* sobre a regulação de vias e genes relacionados a características de interesse (JUNG; CASLER, 2006a; JUNG; CASLER, 2006b; SEKHON et al., 2011). Essa compreensão se torna particularmente relevante no contexto do melhoramento de culturas para produção de bioetanol de segunda geração, como a cana energia. Assim, neste estudo nós caracterizamos o transcritoma de um internódio de cana energia em desenvolvimento, utilizando tecnologias de sequenciamento em larga escala, para compreender a formação da biomassa de *Saccharum* spp.

CAPÍTULO 1:

Genoma *draft* de *Saccharum spontaneum*

3.1 INTRODUÇÃO

Saccharum spontaneum é uma espécie ancestral do gênero que possui alto conteúdo de biomassa e baixo de sacarose em sua composição (MING et al., 2001; GRIVET, 2004; MATSUOKA et al., 2014). Por apresentar um sistema radicular bem desenvolvido e uma alta capacidade de perfilhamento, plantas de *S. spontaneum* são fortes competidores, muitas vezes sobrepujando as canas convencionais, por isso são consideradas como pragas em alguns lugares (MATSUOKA et al., 2014). Além disso *S. spontaneum* é bastante resistente a estresses bióticos e abióticos, o que motivou seu uso recorrente no melhoramento de cana-de-açúcar por muitas décadas para introgressão dessas características desejáveis nas variedades tradicionais (MING et al., 2001; GRIVET, 2004).

Com o advento de novas tecnologias para a produção de combustíveis mais sustentáveis, como o etanol lignocelulósico (ou de segunda geração), há uma renovação na importância de *S. spontaneum* que agora passa a ser o principal parental e fonte de caracteres desejáveis para novas variedades de cana voltadas à produção de biomassa, a cana energia (TEW; COBILL, 2008; MATSUOKA et al., 2014; DOS SANTOS et. al., 2016) . Cana energia são híbridos de *Saccharum* spp. com alta produção de biomassa que apresentam um conteúdo maior de fibras do que de sacarose em sua composição (ao contrário da cana-de-açúcar). Essas variedades são uma das fontes mais promissoras de biomassa para viabilizar a produção do etanol lignocelulósico, especialmente para o Brasil (MATSUOKA et al., 2014; DOS SANTOS et. al., 2016). A despeito da importância dos híbridos de *Saccharum*, não existe ainda um genoma disponível para nenhuma das espécies desse gênero.

A consolidação de tecnologias de sequenciamento *high throughput* tornou o sequenciamento de genomas mais rápido e acessível. Atualmente foram sequenciadas mais de 100 plantas que variam desde genomas pequenos (*Utricularia gibba*, 80 Mbp) a genomas enormes, complexos e com muitas regiões repetitivas (*Triticum aestivum*, 17 Gbp) (VEECKMAN; RUTTINK; VANDEPOELE, 2016). Em geral os genomas sequenciados contém os alelos com as regiões gênicas e intergênicas e suas posições nos cromossomos determinadas. No entanto, um catálogo contendo todos os genes do organismo anotados pode ser uma opção para organismos com genomas maiores e mais complexos como

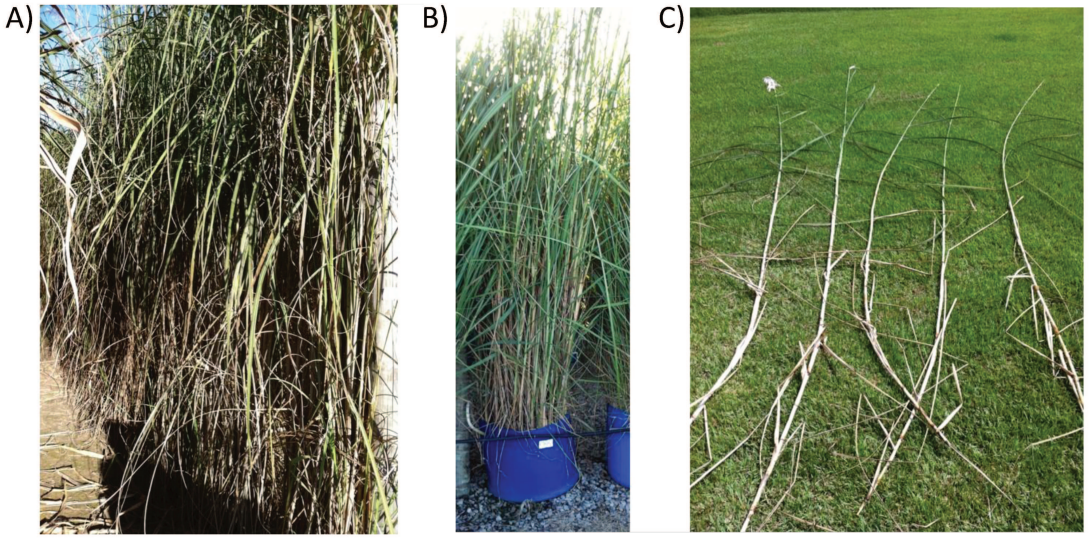
Saccharum ssp., pois permite a realização de estudos de caracterização da funcionalidade dos genes, validação de novos genes, análise de vias bioquímicas e regulatórias, que podem explicar as características estruturais e fisiológicas de interesse da espécie (VEECKMAN; RUTTINK; VANDEPOELE, 2016). Assim, devido a importância de *S. spontaneum* para o desenvolvimento da cana energia e a possibilidade de empregar as tecnologias de alto desempenho para o sequenciamento de genomas, nós produzimos nesse estudo o primeiro genoma *draft* de *Saccharum spontaneum*. A estratégia adotada se baseia no sequenciamento massivo de sequências curtas e no desenvolvimento de metodologias para a montagem dessas sequências com o foco na produção de uma catálogo contendo as regiões codantes do genoma, que possibilitam o acesso as informações básicas da espécie para embasar e direcionar o melhoramento da cana energia.

3.1 MATERIAL E MÉTODOS

3.1.1 Material Vegetal

O genótipo de *S. spontaneum* US85-1008 utilizado nesse estudo foi desenvolvido pelo programa de melhoramento da Louisiana State University (LSU) e o Houma-USDA (Estados Unidos da América) a partir de um cruzamento entre *S. spontaneum* x US60-313 (Figura 6).

Figura 6: US85-1008 no germoplasma da GranBio. Cultivo em sulco de 6m (a) e em vaso (b). Colmos de US85-1008 (c). Imagens gentilmente cedidas pela GranBio.



Essa cultivar, que possui elevado teor de fibra e produtividade, além de alto perfilhamento, integra a base dos cruzamentos do programa de melhoramento da Cana Energia da empresa Granbio e foi gentilmente cedido por esta. Este genótipo, assim como outras espécies de *S. spontaneum,* é classificado como praga por ser um forte competidor com as variedades cultivadas (USDA, ARS). US85-1008 possui colmos conoidais finos, com comprimento médio-longo, alta intensidade de perfilhamento, folhas (limbo) estreitas e touceira de crescimento ereto (Figura 6) . A descrição das médias fenotípicas fornecidas pela empresa encontra-se na tabela 1.

Tabela 1. Médias fenotípicas do genótipo US85-1008, nos Bancos Ativos de Germoplasmas da RIDESA/UFAL (SILVEIRA et al, 2014) e da Biovertis, Brasil.

Tipo	SILVEIRA (2014)	BIOVERTIS (2016)
Idade (meses)	10	12
Peso Médio Colmo (kg)	0,12	0,17
Peso Palha (kg)	0,1	0,06
Peso Palmito(kg)	-	0,05
Diâmetro Colmo (cm)	1,2	1,21
Comprimento do colmo (m)	1,15	2,19
Pureza	72	47
Fibra (%)	14,9	31,46
Pol Cana	6,03	2,83
Açúcar Redutor	1	2
Açúcar Total Recuperável	65,7	30,13

Plantas de US85-1008 foram cultivados em casa de vegetação na Universidade Estadual de Campinas, por aproximadamente um ano. As amostras de folhas para extração de DNA foram coletadas de plantas jovens, enquanto as amostras de tecidos para o sequenciamento de transcritos foram realizadas em plantas maduras (com aproximadamente doze meses).

3.1.2 Quantificação DNA absoluto

Sementes de *Solanum lycopersicum* L. 'Stupicke' utilizadas como padrão interno (2C=2.00 pg – Praça-Fontes et al 2011b) foram cordialmente cedidas pelo Dr. Jaroslav Dolezˇel do *Experimental Institute of Botany*, República Tcheca. As plantas foram cultivadas em casa de vegetação na Universidade Federal de Viçosa (UFV), sob condições ambientais similares (temperatura, umidade, fotoperíodo - intensidade luminosa). Folhas jovens de *S. lycopersicum* 'Stupicke´ foram processadas para citometria de fluxo dentro de duas horas após coleta. Folhas jovens de *S. spontaneum* US85-1008 cultivado em casa de vegetação sob condições ambientais similares foram coletadas e enviadas ao Laboratório de Citogenética e Citometria para as análises de Citometria de Fluxo (CMF). Análises CMF foram realizadas no Laboratório de Citogenética e Citometria, Departamento de Biologia Geral (UFV, Brasil). O tamanho do genoma de *S. spontaneum* foi quantificado a partir de suspensões de núcleos extraídas e coradas de acordo com o procedimento descrito por Carvalho C.R. et. al. (2008). As suspensões foram analisadas em citômetro Partec PAS (*Partec_ GmbH, Munster, Germany*) equipado com uma fonte de laser (488 nm, para valor 2C). Os parâmetros da citometria de fluxo foram calibrados antes de cada quantificação baseadas em análises de CMF do padrão primário (*S. lycopersicum* 'Stupicke') e das amostras (*S. spontaneum*). Três repetições independentes, contando com mais de 10,000 núcleos foram analisadas em cada análise. O valor 2-C do genótipo de *S. spontaneum* foi calculado pela divisão do canal principal do pico de fluorescência G0/G1 do padrão primário pelo canal principal do pico de fluorescência G0/G1 de cada amostra.

Esta etapa foi realizada em colaboração com o Prof. Dr. Carlos Roberto Carvalho e com o estudante Guilherme Mendes Almeida Carvalho do Laboratório de Citogenética e Citometria, Departamento de Biologia Geral (UFV, Brasil).

3.1.3 Extração de DNA Total e Sequenciamento

O DNA total de US85-1008 foi extraído a partir de 100 mg de folha segundo o protocolo a seguir. Ao material pulverizado adicionou-se 700 µL de tampão de extração (100 mM Tris-HCl pH8, 20 mM EDTA, 2% CTAB, 1% PVP 40, 1,4 M NaCl, 0,3% 2-mercaptoetanol). Após misturar vigorosamente as amostras, incubou-se a mistura por exatos 30 minutos a 65ºC. Em seguida, adicionou-se 800 µL de clorofórmio/álcool isoamílico (24:1) às amostras que foram então centrifugadas a 16.000 g por 10 minutos a temperatura ambiente (T.A). Após a centrifugação, foram recuperados 450 µL de sobrenadante, que foi transferido à um novo microtubo contendo 120 µL de solução de extração (5% CTAB, 1,4 M NaCl), onde foram adicionados 570 µL de clorofórmio/álcool isoamílico (24:1). A solução foi centrifugada a 16.000 g por 10 minutos (temperatura ambiente); aproximadamente 400 µL de sobrenadante foram recuperados e transferidos para um novo microtubo, seguido da adição de um volume de isopropanol. Após nova centrifugação (16.000g, 5 min, T.A.), descartou-se o sobrenadante e foi adicionado 1 ml de etanol 70% ao precipitado de DNA. A seguir, as amostras foram centrifugadas (14.000g, 5 min, T.A) e após a evaporação completa do etanol, o precipitado foi solubilizado em 30 µL de água ultrapura. Após a extração, adicionou-se 2 µL de RNAse às amostras que permanecem a 37ºC durante uma hora. A concentração de DNA de cada amostra foi quantificada no equipamento Qubit, utilizando o kit dsDNA BR Assay (Invitrogen™) segundo instruções do fabricante. Todas as amostras apresentaram alta qualidade e rendimento. A montagem das bibliotecas e o sequenciamento foram realizados na *facility Center for Genome Sciences*, da *University of North Carolina*, (*Chapel Hill*, Carolina do Norte, EUA). Foram sequenciadas duas bibliotecas *paired end* de 100pb (com insertos de 180pb - sequências com sobreposição - e 400pb) no *Illumina Hiseq*.

3.1.4 Montagem do genoma *draft* de *S. spontaneum*

Devido à complexidade do genoma poliplóide de *Saccharum spp*, que pode apresentar haplótipos com grandes diferenças genômicas (como rearranjos, inserções e deleções de grandes regiões) a cobertura do sequenciamento não é uniforme, preceito em que se baseiam os montadores de genomas mais utilizados.

Assim uma nova estratégia de montagem foi desenvolvida para *S. spontaneum* pelo doutorando Leandro Costa do Nascimento em seu projeto de doutorado. Em suma, a montagem do genoma de *S. spontaneum* focou apenas em regiões de genes (contendo exons, introns, UTRs e promotores) e utilizou o programa Trinity (GRABHERR et al., 2013) – em geral utilizado para a montagem de dados RNAseqs (que apresentam splicing alternativo, ou seja, cobertura não uniforme no sequenciamento).

No tratamento inicial das sequências de *Illumina*, utilizou-se o programa FASTX-toolkit para realizar a filtragem das sequências de baixa qualidade, a trimagem das sequências e a remoção dos adaptadores. Antes de realizar o alinhamento, cada read de 100 pb é 'clivado' em subreads de 50 pb: 1-50 pb, 25-75 pb, 50-100 pb. Cada subread é mapeado em lócus de três espécies: *Sorghum bicolor* (33,032 gene loci) (PATERSON et al., 2008), *Setaria italica* (35,471 gene loci) (ZHANG et al., 2012) e *Zea mays* (80,713 gene loci) (SCHNABLE et al., 2009) utilizando scripts PERL. Após a clivagem, os subreads são alinhados contra os lócus dessas especies utilizando Bowtie2 (LANGMEAD et al., 2012). Se um dos subreads alinhar com o lócus, os dois reads iniciais de 100 bp (sem clivagem) irão compor o conjunto de sequências utilizadas para montagem daquele lócus. Dessa forma se uma região da sequência apresenta alta similaridade, o restante da sequência fará parte do *contig*, permitindo a montagem dos introns (regiões com similaridade menor). A montagem do conjunto de sequências selecionadas para cada lócus foi realizada com o programa *Trinity* (GRABHERR et al., 2011). Os reads não mapeados nos genes de referência foram então montados *de novo* pelo programa *Trinity*, e são importantes pois irão gerar *contigs* contendo os genes específicos de *S. spontaneum*. A seguir foi realizado o *scaffolding* dos *contigs* com o programa SSPACE2 (BOETZER *et al.*, 2011) utilizando as sequências de DNA e de transcritos de *S. spontaneum* produzidas nesse estudo. O resultado final compõe o genoma *draft* "*Saccharum spontaneum* 1.0". A predição de genes de *S. spontaneum* foi realizada com o programa AUGUSTUS 3.0.1 (STANKE *et al.*, 2006) utilizando dados disponíveis em bancos públicos. Para essa análise foram empregadas as sequência de proteínas de *Sorghum bicolor* (33,012) (PATERSON et al., 2011), *Zea mays* (80,713) (SCHNABLE et al., 2009), *Setaria italica* (35,471) (ZHANG et al., 2012], *Oryza Sativa* (39,049) (OUYANG et al., 2007) e *Arabidopsis thaliana* (27,436) (HUALA et al., 2001) (LAMESCH et al., 2012). Também foram utilizados 358,695 de

transcritos de *Saccharum spontaneum* (produzidos pelo sequenciamento de seis tecidos nesse estudo) e 268,034 transcritos de *Saccharum* spp. (CARDOSO-SILVA et al., 2014; NISHIYAMA Jr. et al., 2014). A anotação funcional foi feita por BLASTp (*e-value cutoff* = 1e-5) contra bancos de dados de proteínas: non-redundant (NR) database do NCBI (National Center for Biotechnology Information), uniref90 e uniref100 (UniProt Reference Clusters – com 90% e 100% de identidade respectivamente), SWISSPROT (SUZEK et al., 2007), PFAM (PUNTA et. al., 2012), KEGG (Kyoto Encyclopedia of Genes and Genomes) (KANEHISA et al., 2000) e Gene Ontology (GO) (ASHBURNER et al., 2000).

Para possibilitar a integração dessas informações e facilitar o acesso às montagens foi desenvolvido um *web site* público (*Sugarcane Database*) no laboratório de bioinformática do LGE. O site contém ferramentas que permitem a busca nos *contigs* anotados usando o nome do gene, palavra chave ou sequências de nucleotídeo/proteína (local BLAST). Toda a etapa de montagem do genoma e transcritos de *S. spontaneum* foi executada pelo doutorando Leandro C. do Nascimento.

3.1.5 Análise das montagens

A avaliação da qualidade de predição do genoma foi feita pela comparação do melhor hit obtido pelo BLASTp contra o banco de dados Uniref com as sequências preditas de *S. spontaneum.* Por essa comparação os *contigs* foram classificados em completos (quando continham a sequência do códon ATG ao STOP), 3’ incompletos (quando continham a sequência do códon ATG mas não o STOP), 5’ incompletos (quando continham a sequência do códon STOP mas não do ATG) e parciais (quando não continham a sequência do códon ATG ou STOP). Para entender a representatividade do resultado frente a outros genomas sequenciados a mesma análise foi realizada com os genomas de sorgo (*S. bicolor*) e milho (*Zea mays*) já publicados. Também foi realizada uma comparação da montagem de *S. spontaneum* com o SUCEST (VETTORE *et al.*, 2001; VETTORE *et al.*, 2003) usando a ferramenta BLASTn (com e-value de 1e-10). Outra estratégia para acessar a representatividade do genoma *draft* foi o mapeamento das sequências de proteínas de *S. spontaneum* no Mapman usando a ferramenta

Mercartor (LOHSE M., et al, 2013). A mesma análise foi feita com as proteínas de outras gramíneas sequenciadas (*Oryza sativa*, *Zea mays* e *Sorghum bicolor*) e com a montagem *ab initio* de transcritos de cana energia (Descrita no capítulo 2).

Para acessar a representatividade da montagem em vias específicas foi realizada uma busca manual em genomas anotados e em bancos de mapas metabólicos como KEGG e o *PlantCyc.* Baseado em vias metabólicas já descritas, foram utilizadas sequências de proteínas de sorgo, milho, arroz, *Arabidopsis thaliana* ou *Populus trichocarpa* de cada etapa da via para buscar por tblastn os *contigs* correspondentes de *S. spontaneum* e avaliar se este continha o gene completo para a proteína de interesse. Os resultados foram anotados em tabelas do *Excel* contendo: nome do gene, EC Number, sequência e organismo da proteína utilizada na busca, identificação do *contig* de *S. spontaneum* localizado na busca, e-Value, tamanho do *contig* (pb), tamanho do gene de referência (aa), início e final do gene no *contig* (pb), início e final da proteína (aa), e se o gene estava completo em *S. spontaneum*.

3.1.6 Extração de RNA e Sequenciamento de RNA

A extração de RNA total de 6 tecidos (folha, colmo, raiz, nó, gemas e meristema) de US85-1008 foi realizada segundo o protocolo de Zeng e Yang (2002) modificado por PROVOST et al.(2007). Todos os tecidos foram extraídos em triplicata biológica, totalizando 18 amostras. A montagem das bibliotecas foi realizada a partir de 1 ug de RNA total, utilizando o protocolo 'Truseq Stranded mRNA Sample Preparation Guide' (Illumina). A análise de qualidade e quantificação das bibliotecas foi realizada por eletroforese em gel na plataforma LabChip® GX. Para cada tecido foram sequenciadas uma réplica com bibliotecas *paired end* de 100pb (referência para montagem) e duas bibliotecas *single end* de 100 pb no *Illumina Hiseq*. A montagem das bibliotecas e o sequenciamento foi realizado na *facility Center for Genome Sciences*, da *University of North Carolina*, (*Chapel Hill*, Carolina do Norte, EUA durante o período de doutoramento sanduíche sob supervisão do Dr. Piotr A Mieczkowski.

3.1.7 Identificação de genes tecido específicos e regiões promotoras

Para determinação do grupo de genes contendo regiões promotoras de interesse no genoma de *S. spontaneum*, foi realizado uma busca por todos os genes que continham o inicio do gene na posição correta (ATG inicial) e pelo menos 1000 bp *upstream* ao gene. A partir da lista de 11.767 genes selecionados por conter a região promotora (início do gene na posição correta e pelo menos 1000 bp upstream) e dos dados de expressão gênica de cada transcrito, foram geradas duas listas de genes definidas por conterem genes (i) constitutivos ou (ii) tecido específicos. Para o cálculo de genes tecido específicos foi exigido que o gene tivesse valores de RPKM acima de 10 entre as réplicas biológicas de um tecido específico e zero nos outros. A lista de genes constitutivos foi gerada através do cálculo da variância de cada gene, isto é, desvio padrão dividido pela média, dos valores de RPKM de um gene entre todos os tecidos, sendo que o gene foi considerado constitutivo quando essa variância ficou abaixo de 15%.

3.2 RESULTADOS

3.2.1 Determinação do tamanho do genoma de *S. spontaneum*

O tamanho do genoma e o número de cromossomos dos indivíduos de *Saccharum* spp. é conhecido apenas para poucos genótipos (ZHANG et al., 2012). Para a maioria dos estudos e programas de melhoramento de *Saccharum* spp. essas informações são desconhecidas, inclusive para o genótipo empregado nesse projeto. Assim, uma etapa importante do estudo foi a determinação do conteúdo de DNA total (2C) do cultivar de *S. spontaneum* (US85-1008) por citometria de fluxo. Como padrão interno foi utilizado *Solanum lycopersicum* L. 'Stupicke' (2C=2.00 pg).

A partir da comparação com o padrão interno, foi possível estimar o conteúdo de DNA total em uma célula diploide (2C) de US85-1008 em 9,32pg. Como 1 pg corresponde a 978 Mb, o tamanho estimado do genoma de US85-1008 é de 9,11 Gb (Figura 7 e tabela 2), o coeficiente de variação entre as réplicas foi de 0,04%.

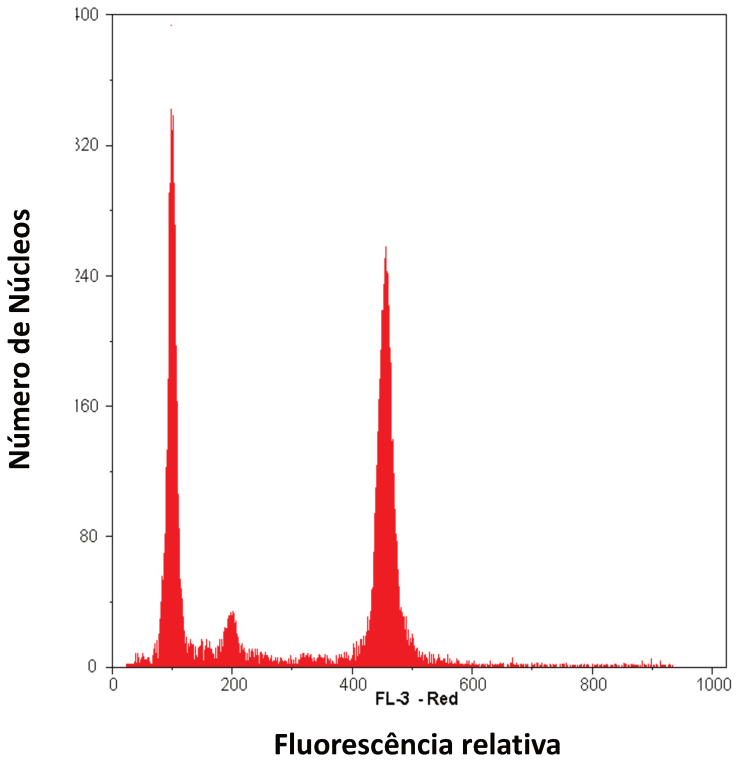


Figura 7. Histograma de intensidade de fluorescência por citometria de fluxo de núcleos G0/G1 de *Solanum lycopersicum* (padrão) *x S. spontaneum* (US85-1008).

Tabela 2. Tamanho estimado de genoma de *Solanum lycopersicum* (padrão) e de genótipos de *S. spontaneum* (US-85-1008).

Tipo	R1	R2	R3	2 DNA (pg) (média)
Padrão (Solanum lycopersicum)	2.00	2.00	2.00	2.00
S. spontaneum (US85 - 1008)	9.27	9.33	9.35	9,32

Genótipos de *S. spontaneum* com conteúdo 2C quantificado por CMF apresentaram tamanho de genoma variando de 2,5 a 12,5 Gb (ZHANG et al., 2012), logo, a cultivar avaliada nesse estudo corrobora esses resultados e se encontra entre os maiores genomas da espécie. Esta etapa foi realizada em colaboração com o Prof. Dr. Carlos Roberto Carvalho e com o estudante Guilherme Mendes Almeida Carvalho do Laboratório de Citogenética e Citometria, Departamento de Biologia Geral (UFV, Brasil).

3.2.2 Sequenciamento genômico de *S. spontaneum* e Montagem do genoma *draft*

O sequenciamento de US85-1008 produziu 114 milhões de *reads paired-end* com inserto de 175 bp e 185 milhões de *reads paired-end* com inserto de 400 bp, totalizando aproximadamente 600 milhões de sequências (60Gb) que foram utilizadas para a montagem do genoma (tabela 3). A cobertura estimada do sequenciamento é de aproximadamente 6x o tamanho do genoma monoplóide determinado para os genótipos. Essa cobertura é considerada baixa mesmo para organismos com genomas menores e menos complexos como leveduras, evidenciando o desafio e a importância do *draft* produzido.

Tabela 3. Dados de sequenciamento de *S. spontaneum*

Tamanho do inserto (bp)	Tamanho do *read* (bp)	Número de *reads*	Total sequenciado
175	100	114,691,902 (x2)	23 Gb
400	100	185,045,504 (x2)	37 Gb

Para estabelecer o *pipeline* e os padrões para as montagens, seria ideal utilizar um genótipo que tivesse o transcriptoma bem caracterizado, como por exemplo, o genótipo SP80-3280, um híbrido comercial que tem a maior representatividade de ESTs: 135,534 sequências, que correspondem a menos da metade dos ESTs disponíveis. Como o foco do trabalho é o desenvolvimento da Cana Energia e as informações de transcritos são insuficientes, optou-se por utilizar US85-1008 para estabelecer os parâmetros do *pipeline* e futuramente realizar a montagem dos outros genótipos. O resultado da montagem realizada pelo aluno Leandro Costa do Nascimento encontra-se resumido na tabela 4.

Tabela 4. Sumário do genoma *draft* de *S. spontaneum* (US85-1008).

	Montagem do genoma de *S. spontaneum*
Número de contigs >= 1,000bp	**79,690**
Número de contigs >= 2,000bp	**9.098**
Maior contig (bp)	**53,214**
Média dos tamanhos dos contigs (bp)	**1,597**
N50 (bp)	**2,223**

O número total de contigs >=1000 bp gerados no genoma draft foi de 79,690, aproximadamente 2 contigs por *locus*. O maior contig da montagem possui mais de 53 kb e foram produzidos mais de 9 mil contigs >=2000 bp. O tamanho médio dos contigs foi de 1,6 kb e o N50 de 2,2 kb (o N50 indica que 50 % do total de pares de base no genoma esteja contida em tamanhos => ao N50) (VEECKMAN; RUTTINK; VANDEPOELE, 2016).

Para integrar os dados gerados foi criado o ‘Sugarcane Database’ (http://bioinfo03.ibi.unicamp.br/sugarcane/) (Figura 8), onde é possível visualizar o alinhamento das montagens no genoma de referência, realizar buscas pelo nome do gene, palavra chave ou sequências de nucleotídeo/proteína (local BLAST) e baixar todos os conjuntos de dados (montagens ou dados brutos).

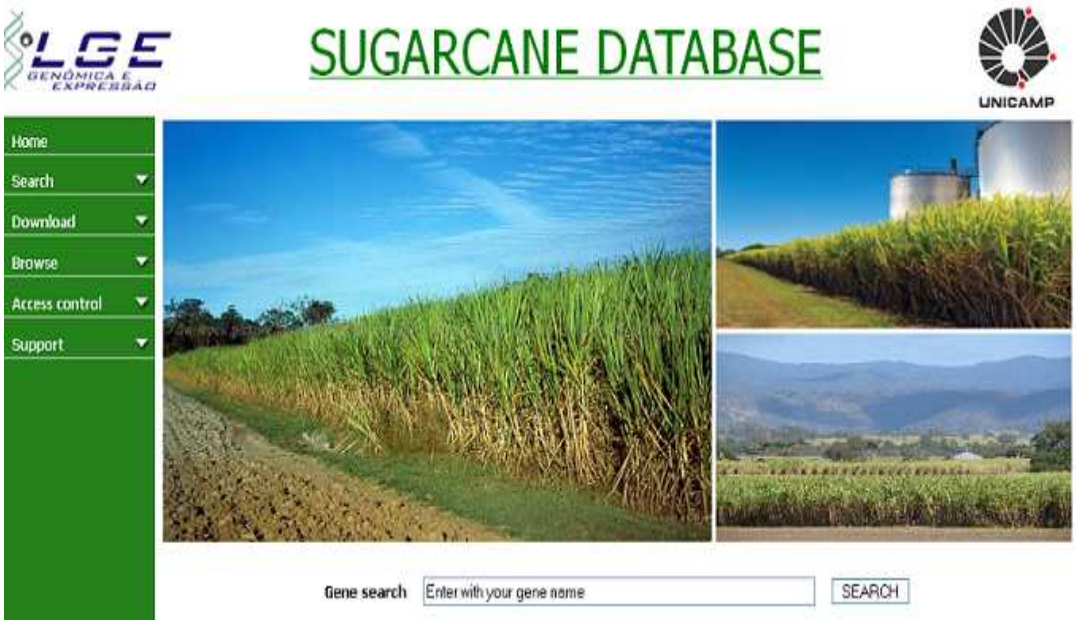


Figura 8: *Sugarcane Database*: banco que permite o armazenamento e integração de dados brutos e *contigs*/unigenes produzidos pelas montagens.

3.2.3 Análise da montagem do genoma de *S. spontaneum*

Para avaliar se os genes estavam completos na montagem foi realizado um BLASTp da predição de genes dos genomas de *S. spontaneum, Sorghum bicolor e Zea mays* contra o banco de dados UniRef90 e os contigs foram classificados em completos (quando continham a sequência do códon ATG ao STOP), 3' incompletos (quando continham a sequência do códon ATG, mas não o STOP), 5' incompletos (quando continham a sequência do códon STOP, mas não do ATG) e parciais (quando não continham a sequência do códon ATG ou STOP) (Figura 9). Aproximadamente 30% dos genes de *S. spontaneum* produzidos na montagem estão completos, número menor que o de genes completos em sorgo - 40%, mas superior ao de milho, que possui mais de 20% dos genes completos. O número de contigs 3' e 5' incompletos de *S. spontaneum* ficou próximo ao de sorgo e o milho apresenta menos *contigs* incompletos que ambos. O número de contigs no hits para a montagem foi > 20%, aproximadamente o mesmo resultado obtido para as proteínas de sorgo, enquanto para milho foi observado que mais de 40% dos *contigs* são identificados como *no hits*. Essa análise demonstra a qualidade da montagem de *S. spontaneum*, que apresenta resultados próximos aos do genoma de sorgo, uma espécie diploide, cujo genoma é referência para os estudos com

gramíneas. Os resultados também indicam que o draft de *S. spontaneum* possui *contigs* mais completos que a montagem do genoma publicado de milho.

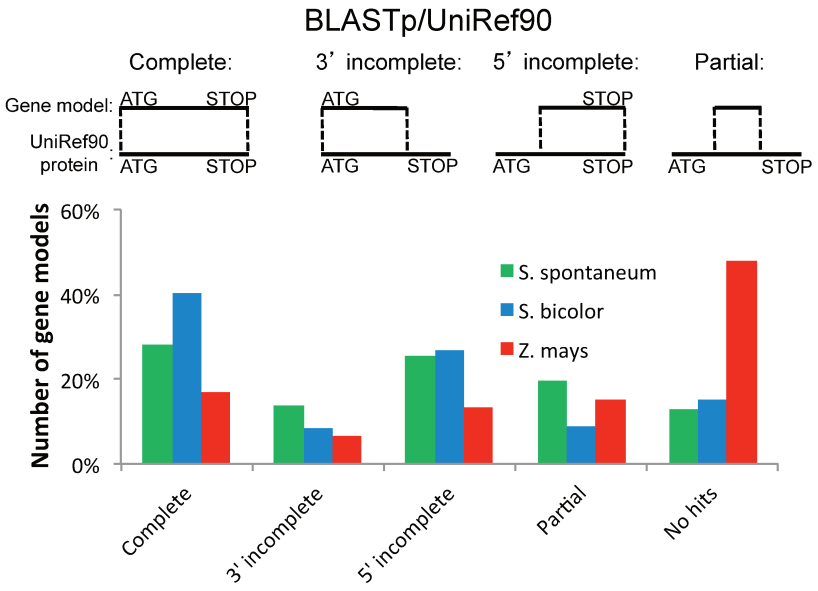


Figura 9. Resultados do BLASTp da predição de genes dos genomas de *S. spontaneum, Sorghum bicolor e Zea mays* contra o banco de dados Uniref. Foram considerados genes completos os contigs que apresentaram ATG and STOP códon, incompletos 3' os que apresentaram apenas o ATG, incompletos 5' os que continham apenas o STOP e parcial os *contigs* que não possuiam os códons START ou STOP em sua sequência.

Assim, embora "*Saccharum spontaneum* 1.0" represente uma montagem bastante fragmentada (N50 = 2,223 bp) que cobre somente 30% do genoma monoplóide, a comparação da representatividade de genes completos, contendo o ATG ou STOP ou parcialmente completos, com sorgo e milho sugere que o *draft* produzido cobre a maior parte das regiões codantes (já que a montagem foi baseada em sequências de *loci* de outras plantas).

A representatividade das regiões codantes também foi validada ao mapear as predições dos genomas de *S. spontaneum, Oryza sativa*, *Zea mays* e *Sorghum bicolor* nos mapas metabólicos do *software* Mapman usando a ferramenta Mercator (Figura 10). Todos os genomas de gramíneas possuem uma parte significativa das proteínas preditas que não são identificadas em bancos de dados e não são mapeadas em vias metabólicas conhecidas. Para *O. sativa,* o primeiro genoma de gramínea sequenciado e, *S. bicolor*, um genoma de referência no estudo de gramíneas, essa porcentagem de genes não identificados é de 47,54% e 43,30 %

respectivamente. O *draft* de *S. spontaneum* apresentou 52,77 % de proteínas não classificadas, número próximo ao de *Z. mays*, 54,39 %. Para destacar a significância da predição do genoma *draft*, a predição de uma montagem *ab initio* de um genótipo de *Saccharum* spp, a cana energia (descrita no capítulo 2) também foi mapeada nas vias metabólicas e como resultado, 71,13 % das proteínas preditas não foram classificadas (quase 20% a mais que *S. spontaneum*). Dessa forma é possível observar como a utilização do genoma *draft* pode auxiliar na caracterização dos genes de *Saccharum* spp. e ser utilizado como referência para as montagens de transcritos expressos.

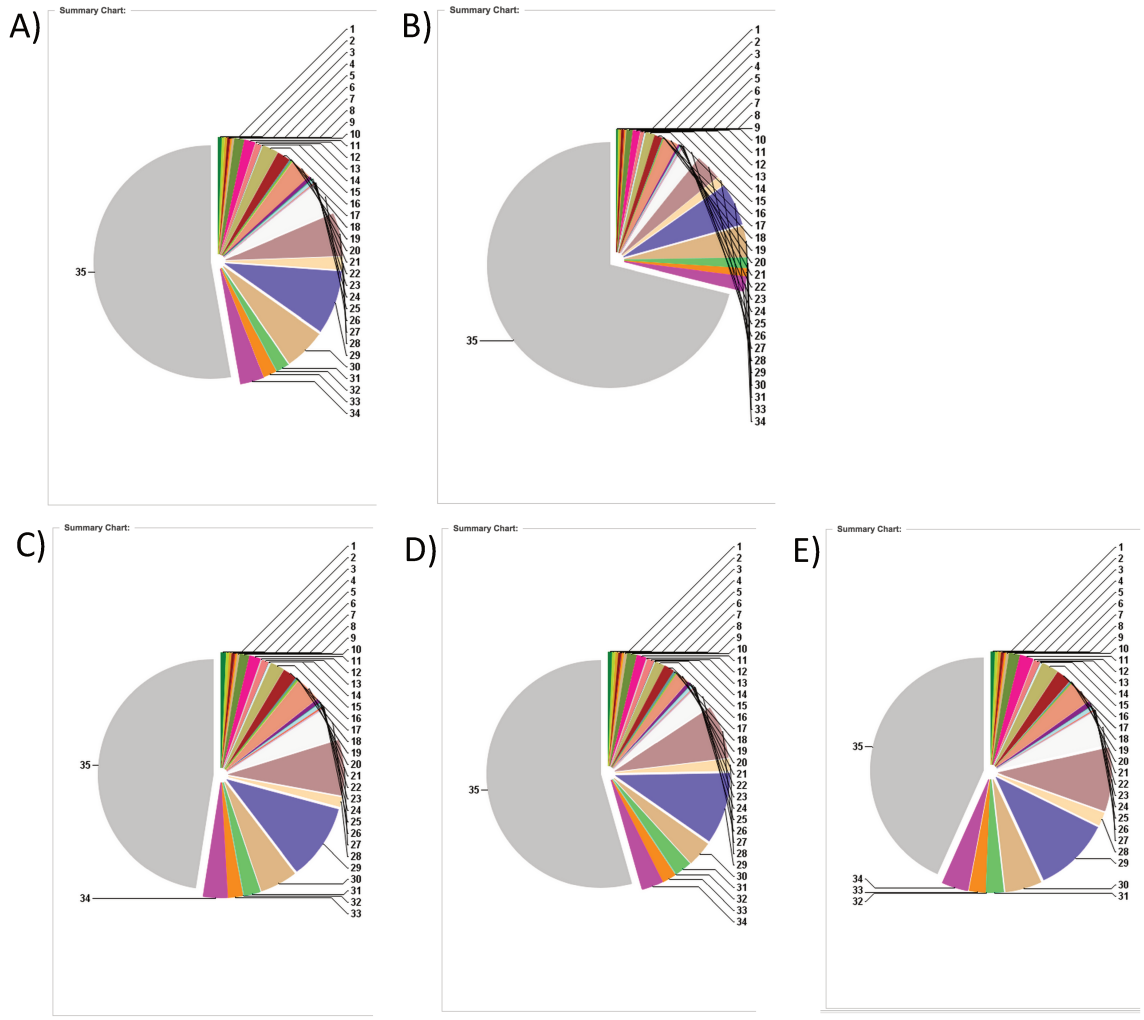


Figura 10. Mapa das principais classes metabólicas A) do genoma *draft* de *S. spontaneum*, B) da montagem *ab initio* de cana energia C) do genoma de *Oryza sativa*, D) do genoma de *Zea mays* e E) do genoma de *Sorghum bicolor*. O mapeamento foi realizado utilizando a ferramenta Mercator. A região em cinza representa a porcentagem de *no hits* do mapeamento, destacando que o *draft* de *S. spontaneum* permitiu a identificação de um número de classes e no hits similar aos genomas publicados para as outras espécies e superior ao da montagem *ab initio*.

Outra abordagem para avaliar a representatividade e potencial aplicação da montagem para estudos genômicos, foi a curadoria manual de vias especificas. Assim, a partir de informações de genomas anotados e de bancos de vias metabólicas (KEGG e Plantcyc) foram selecionadas sequências de proteínas de espécies próximas para cada enzima das vias de interesse (como as vias de metabolização da sacarose e de biossíntese da lignina). Em seguida, as proteínas foram comparadas com os *contigs de S. spontaneum*, por tBLASTn, e os resultados foram anotados em tabelas de *Excel* conforme descrito anteriormente. Aproximadamente 80% das proteínas analisadas apresentaram pelo menos um *contig* de US85-1008 contendo o gene completo (incluindo todos os éxons e os íntrons) (Figura 11 e 12).

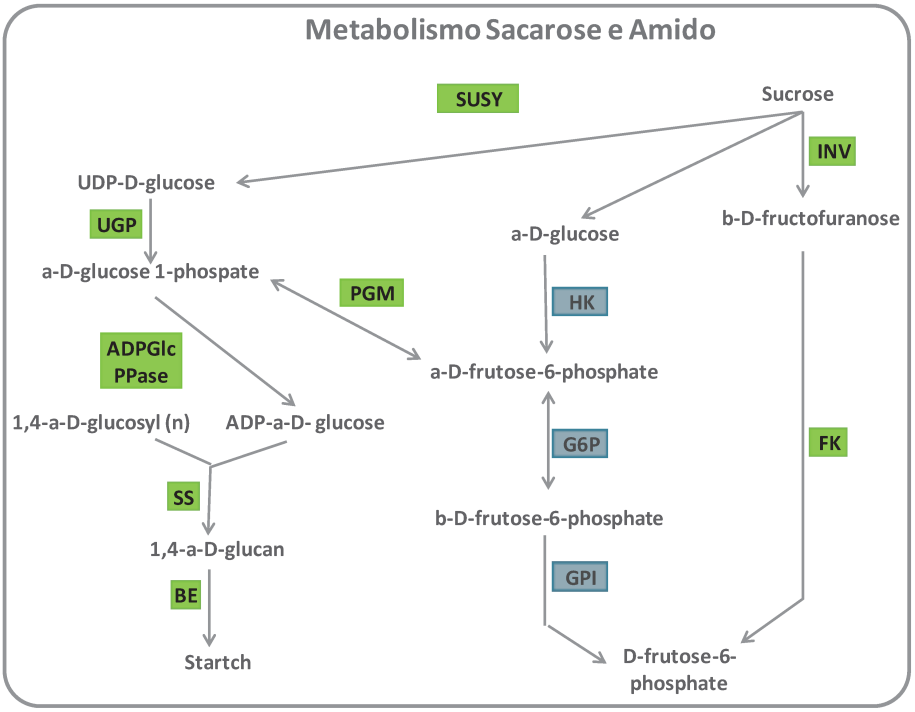


Figura 11: Análise dos genes envolvidos no metabolismo central de sacarose e amido. A maioria dos genes da via estão representadas em apenas um *contig* de US85-1008 (verde). Alguns genes não possuem todos os aminoácidos da proteína representados no *contig* ou apresentavam a proteína 'quebrada' em dois *contigs* (azul). SUSY, Sucrose Synthase; UGP, UDP glucose pyrophosphorylase; INV, Invertase; HK, hexokinase; G6P, G6P-1-epimerase; ADPGlc PPase, Glucose-1-phosphate adenylyltransferase; PGM, Phosphoglucomutase-1; GPI, Glucose-6-phosphate isomerase; BE, Glycogen branching enzyme; FK, fructokinase.

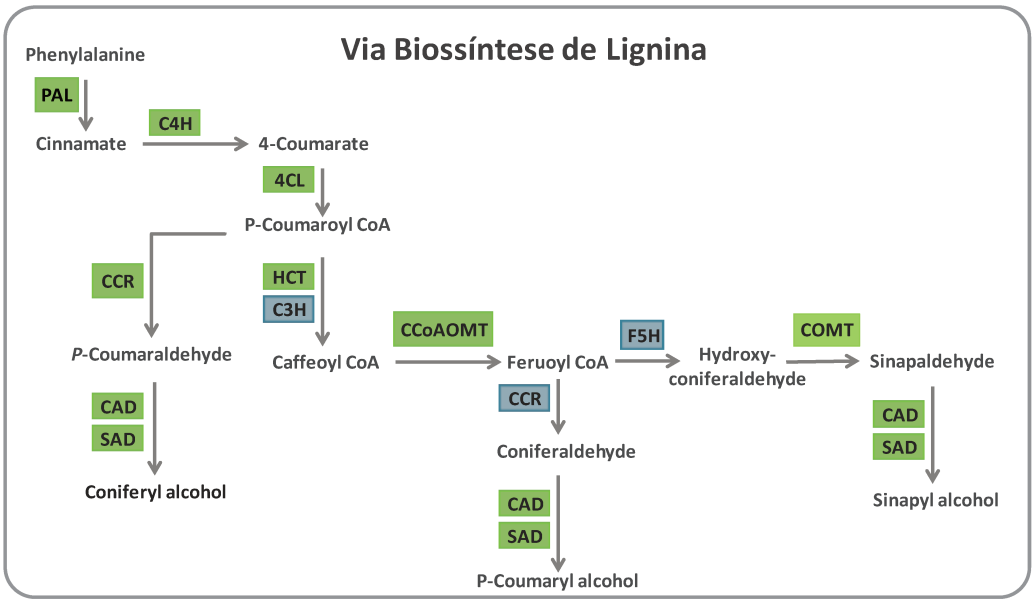


Figura 12: Análise dos genes da via de biossíntese de lignina. A maioria dos genes da via estão representados em apenas um *contig* de US85-1008 (verde). Alguns genes não possuem todos os aminoácidos da proteína representados no *contig* ou apresentavam a proteína 'quebrada' em dois *contigs* (azul). PAL, phenylalanine ammonia-lyase; C4H, cinnamate 4-hydroxylase; 4CL, 4-coumarate-CoA ligase; CCR, cinnamyl CoA reductase; CAD, cinnamyl alcohol dehydrogenase; SAD, sinapyl alcohol dehydrogenase; HCT, hydroxycinnamoyl CoA: shikimate transferase; C3H, p-coumarate 3-hydroxylase; CCoAOMT, caffeoyl CoAO-methyltransferase; F5H, ferulate 5-hydroxylase; COMT, caffeic acid O-methyltransferase; CAD, cinnamyl alcohol dehydrogenase; SAD, sinapyl alcohol dehydrogenase.

Algumas proteínas estavam representadas por mais de um *contig*, indicando que ocorreu uma quebra na montagem. Nesse caso os *contigs* correspondentes representavam a maior parte da proteína. Assim é possível observar que para a maioria dos genes o *draft* contém a região codante em um único *contig*.

3.2.4 Sequenciamento de RNA em larga escala e genes tecido-específicos

No total foram produzidas e sequenciadas 18 bibliotecas de mRNAseq de seis tecidos de *S. spontaneum* (folha, raiz, internódio, nó, meristema apical e gema lateral) que geraram mais de 160 Gb. A montagem das sequências produziu 358,695 transcritos com tamanho médio de contig de 753 bp. Foram anotados 84,802 transcritos contra o banco de dados Uniref90 e 46,892 contra o Swiss Prot. Essas bibliotecas foram importantes para melhorar a acurácia das

predições de *S. spontaneum* apresentadas anteriormente e também para obter informações sobre a dinâmica da expressão em diferentes tecidos.

Ao analisar a expressão dos genes preditos, foi possível identificar 1572 contigs com expressão específica em um dos tecidos (Figura 13). Destes 643 são expressos exclusivamente em folha, 754 em raiz, 101 no nó, 31 na gema lateral, 28 no meristema e 15 em colmo. A análise de enriquecimento de GOs (*Gene Ontology*) para os genes tecido específicos demonstrou que apenas folha e raiz apresentam genes enriquecidos para categorias de processos biológicos (Anexo 1). Para folha os processos enriquecidos estão em sua maioria envolvidos nas vias fotossintéticas, na biogênese de ribossomos, montagem e organização de proteínas e no metabolismo celular, de lipídeos, carboidratos e metabólitos secundários. Na raiz apenas dois processos mostraram enriquecimento nas análises de GO: biogênese/organização de parede celular e senescência.

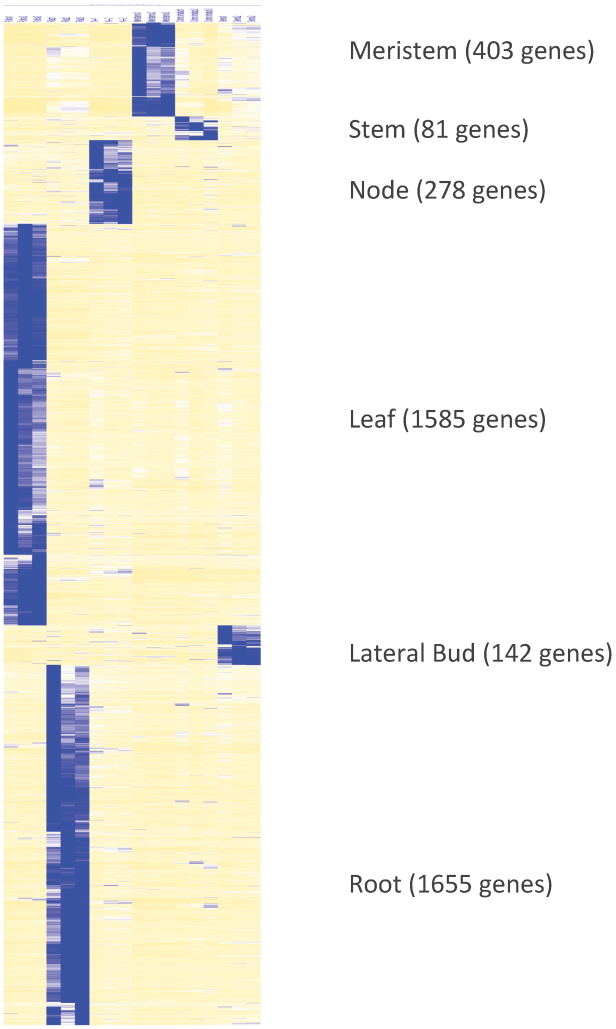


Figura 13: Expressão de genes tecido-específico em folha, raiz, nó, internódio, meristema apical e gema lateral de *S. spontaneum*.

3.2.5 Identificação de regiões promotoras

O genoma de *S. spontaneum* produzido no trabalho contém majoritariamente as regiões codantes do genoma. Além de conter os íntrons, éxons e UTRs, muitos dos *contigs* produzidos também apresentam as regiões promotoras dos genes. Para identificação das regiões promotoras, foram selecionados todos os *contigs* preditos contendo uma região de no minimo 1000 bp antes do códon de iniciação da transcrição (ATG). Assim foram selecionados 11.767 genes preditos contendo a potencial região promotora. Para selecionar os promotores de maior interesse para aplicacões biotecnológicas foram utilizadas duas estratégias: a seleção de genes com expressão ubíqua nos tecidos analisados e de genes expressos predominantemente em um dos dos tecidos. Assim foram identificados 17 promotores tecido-específicos (6 de folha, 5 de raiz, 2 de internódio, 1 de nó e 3 de meristema) e 29 constitutivos.

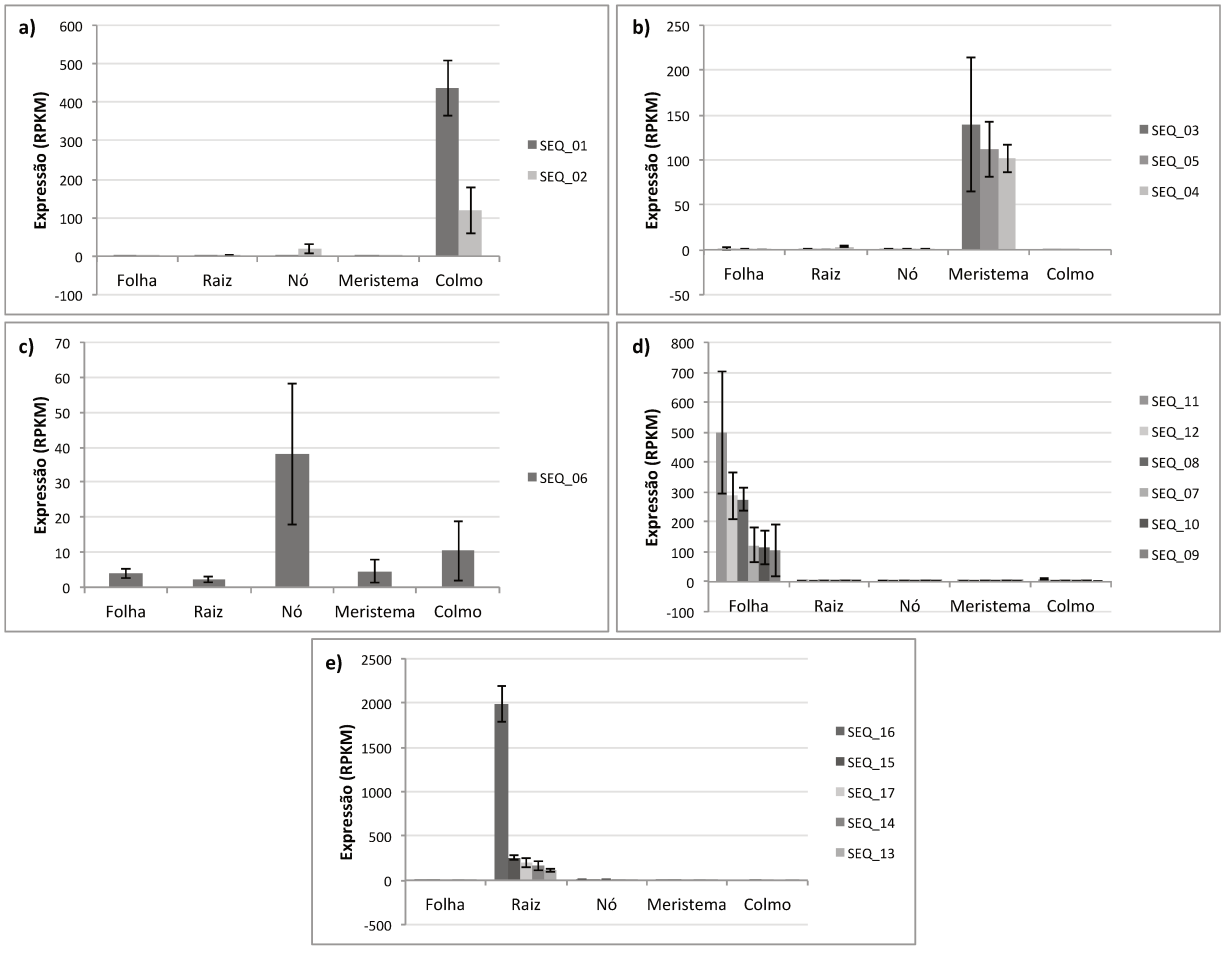


Figura 14 - Expressão dos genes tecido específicos contendo a região promotora nos contigs de *S. spontaneum*.

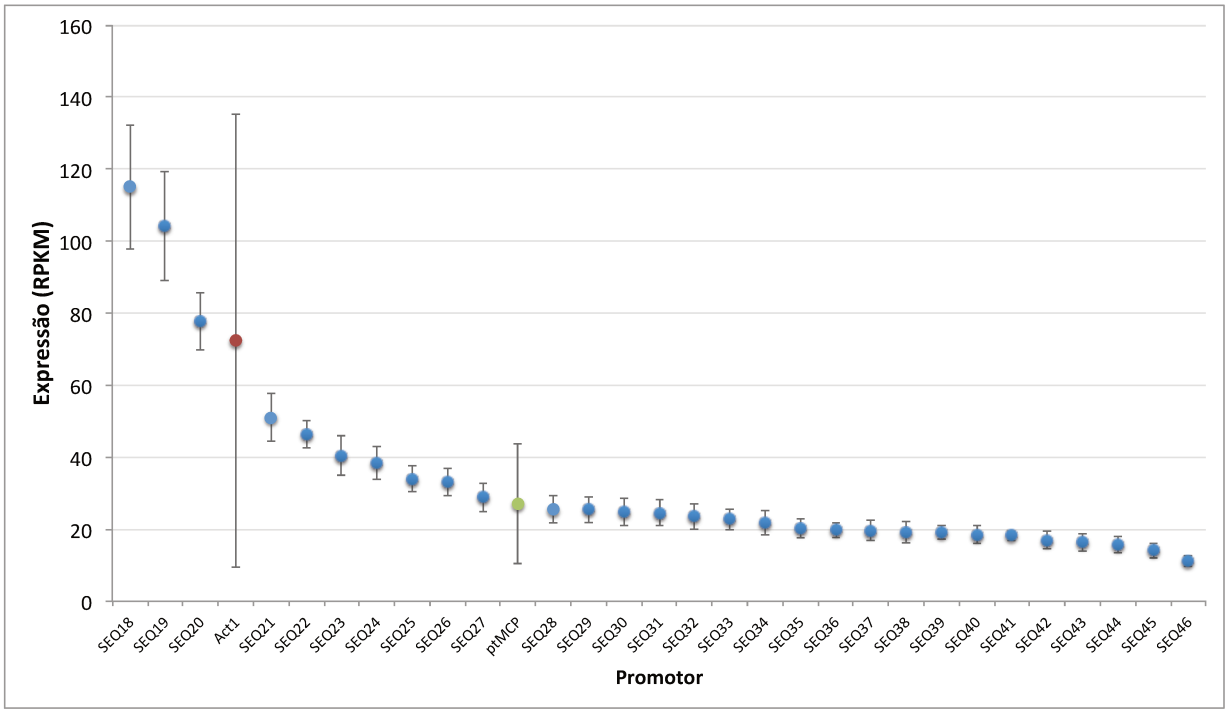


Figura 15 - Expressão dos genes de expressão ubíqua contendo a região promotora nos contigs de *S. spontaneum*. Em vermelho está destacado a expressão de um gene constitutivo de expressão forte descrito na literatura e em verde a de um gene com expressão moderada.

O isolamento e validação dos promotores identificados está em andamento e foi produzida uma patente que está em tramitação junto a Agência de Inovação da Unicamp (Anexo II).

4 Conclusões

Nesse primeiro estudo realizou-se a construção do primeiro *draft* do genoma além de um catálogo de genes expressos em diferentes tecidos de *S. spontaneum*. A despeito da montagem fragmentada produzida no *draft,* as regiões codantes estão bem representadas e permitem a identificação e caracterização da maior parte dos genes da espécie. Foi possível identificar potencias regiões promotoras de genes ubíquos e tecido-específicos, com potencial de aplicação imediata na engenharia genética de *Saccharum*.

CAPÍTULO 2:

An elongating internode from energy cane as a model to better understand the biosynthesis of lignocellulosic biomass and patterned deposition through *Saccharum* hybrids

SUMMARY

Second-generation biofuel and/or biochemical industries shall rely on dedicated crops for its economic sustainability. The ideal feedstock needs to produce biomass in high amounts and quality, a characteristic that is directly linked to cell wall composition. Energy cane is a high-biomass dedicated crop, selected during *Saccharum* breeding program to fit particular industrial needs of 2G bioethanol production. To investigate cell wall deposition and composition during biomass development, we selected an elongating internode of a pre-commercial energy cane hybrid. To define this process we divided the internode in five sections and performed a detailed RNAseq and cell wall characterization. The histological analyses by phoroglucinol-HCl stain revealed a remarkable gradient that range from cell division and protoxylem lignification to the internode maturation and complete vascular bundle lignification. This was also addressed by RNAseq analysis that revealed more than 9.000 differentially expressed genes between the internode sections. Gene ontology analyses showed enriched categories in each section. In agreement to the hystochemical analysis, the expression profile of the lignification genes showed the same gradient pattern with high expression in sections with complete lignified vascular bundles. Gene expression analysis revealed promising candidates for transcriptional regulation of energy cane lignification.

INTRODUCTION

Traditional *Saccharum* hybrids are well known as C4 plants capable of accumulate large amounts of sucrose in their parenchyma cells. Also, as C4 plants, these hybrids are capable to produce, by their high photosynthetic efficiency, massive quantities of lignocellulosic biomass. The high biomass production feature is still not exploited intensively due to the lack of information on the main mechanisms underlying cell wall formation in *Saccharum* and other species of monocotyledons (DOS SANTOS et al., 2016).

Saccharum plants, as do other species of grass, present an interesting developmental pattern in each stalk, displaying a gradient ranging from immature (top four internodes) to mature ones (bottom internodes) with high capacity of

storage of soluble sugars (JACOBSEN ET AL, 1992; LINGLE; THOMSON, 2012). Throughout the internode's development, cell wall composition changes and its dynamics are very relevant in context of biomass-dedicated crops, such as energy cane.

During sugarcane internode elongation an increase in total sugars, cellulose and lignin was observed, while the hemicellulose decreased (LINGLE; SMITH, 1991, LINGLE, 1997; LINGLE; THOMSON, 2012). As the internode ceases elongation (by the fourth internode), a maturation process begins and sucrose starts to be stored in parenchyma cells (LINGLE; SMITH, 1991, LINGLE, 1997; CASU et al., 2007, LINGLE; TEW, 2008, PAPINI-TERZI et al., 2009, LINGLE; THOMPSON, 2012). Older internodes (bottom) contain the highest amount of stored sucrose. Dynamics of cell wall components and sucrose storage, regulated by the plant's carbon flux, are a key feature to understand the biomass formation. Previous studies of *Saccharum* internode development have focused only in physiological aspects of its process LINGLE; SMITH, 1991, LINGLE, 1997; LINGLE; THOMSON, 2012). In this study, we performed a molecular characterization of lignocellulosic biomass formation. Here we present the first RNAseq of energy cane, a biomass dedicated crop, in which lignocellulose biomass formation and pattern were investigated.

EXPERIMENTAL PROCEDURES

Plant Material and Sampling

A pre-commercial energy cane (*Saccharum* hybrids) (Supplementary Material – Figure 1) obtained from the crossing between *Saccharum spontaneum* (IN82-84) and sugarcane (NCo310), kindly provided by the GranBio's breeding program was used in this study. Data provided by GranBio described this energy cane as presenting high fiber and biomass content and an elevated capacity of tillering. Plants with high biomass genotype were grown in greenhouse for 12 months. The internode position was determined according with Kujiper model (KUJIPER, 1915). An elongating internode (3^{rd}) of one-year-old plants was harvested and divided equally in 5 sections (A-E). This internode was chosen due to the higher gradient of lignification between the basal and apical extremities, a feature observed during the experiments with energy cane and also described in literature. Each section (in biological triplicates) was submitted to histological and transcript expression analysis.

Anatomical and Histological Analysis

Stem transverse sections (80 µm thick) were obtained with a Sapphire knife on an automatic vibrating blade microtome (Leica VT 1000S). Sections were stained with the Weisner reagent (phloroglucinol-HCl) and immediately observed under bright-field microscopy (DM IRBE, Leica) coupled with a CCD camera (DFC 300 FX, Leica). Lignin and phenolic compounds were also observed by auto florescence under UV-light.

RNA extraction and sequencing

After harvested and sectioned, samples were immediately frozen in liquid nitrogen and stored at -80^{o}C. Total RNA of three replicates of each section (A-E) was extracted following the protocol described by ZENG E YANG (2002) modified by PROVOST et al. (2007). Quality control and quantification of the RNA samples were performed using a Caliper LabChip XT microcapillary gel electrophoresis and a Qubit fluorometric measurement (Life Technologies), respectively. Sequencing library preparations were generated using the TruSeq Stranded mRNA Sample Preparation Kit (Illumina), from 1ug of total RNA, according with manufacture's instructions. Quality control of the libraries was performed as described above. Each library was sequenced using Illumina HiSeq 2000 (HS) to generate 50 bp paired–end reads.

Transcriptome assembly and annotation

FastQC (BIOINFORMATICS B, 2011) and SortMeRNA (KOPYLOVA; NOE; TOUZET, 2012) software were used to verify the base quality of reads and percentage of ribosomal RNA for each library, respectively. The filtered reads were submitted for *ab-initio* transcriptome assembly using Trinity software (BIOINFORMATICS B, 2011). This analysis produced a large number of transcript sequences that were filtered in order to reduce the false positive using two criteria: 1. the presence of protein-coding regions bigger than 200 bp and 2. the expression quantification, calculated by TPM (transcript per million) metric, bigger than 1.0, in at least one library. The identification of protein-coding regions was performed by Transdecoder software (BIOINFORMATICS B, 2011) using hexamer frequency analysis. The expression analysis was performed using RSEM software (KOPYLOVA; NOE; TOUZET, 2012) by aligning all reads for each library against the

transcript sequences that produced four different results: TPM by transcript, TPM by gene (locus), read counts by transcript and read counts by gene locus.

The filtered transcriptome assembly (reference assembly) was annotated by Blast2GO software version 2.8 (HAAS, 2012) using Non-Redundant protein (NCBI/NR), InterPro and Gene Ontology (GO) databases. The same program was also used to group datasets in GO according to the biological process. The complete bioinformatic pipeline used for this work is described in Supplementary - Figure 2.

Differential expression analysis

The read count matrix for genes was used to identify differentially expressed genes (DEGs) by merging two different statistical softwares, edgeR (LANGMEAD, 2010) and DESeq (LI;DEWEY, 2011), considering the threshold values of adjusted p-value <= 0.05 and |fold change| >= 2. The lists of DE genes in each analysis (AxB, AxC, AxD, AxE) were separated into UP and DOWN regulated and subjected to GO enrichment analysis to identify significantly enriched GO slim terms (Plant GO slim) using Blast2GO software and a p-value ≤ 0.05.

Functional Analysis

Gene expression was estimated using FPKM (fragments per kilobase of exon per million fragments) values. Differentially expressed genes were defined by present FDR >0.01 and -2 < fold-change >2. Hierarchical clustering of DEGs was performed using the Expander 7 software based on DEGs Z-score. To assign specific functional plant categories to energy cane genes we used Mercartor (an automated large scale tool of functional annotation of plant data) (LOHSE et al.,2013).

RESULTS AND DISCUSSION

Energy cane hybrid

Data provided by GranBio described the energy cane genotype as presenting high fiber and biomass content and an elevated capacity of tillering. Detailed information conceded about this pre commercial hybrid and its parent's lineages are summarized in table 1.

Table 1. Characterization of energy cane and its progenitors: sugarcane and *S. spontaneum.* Energy cane hybrids present features of heterosis phenomenon, such as an increase in fiber and stalk length, resulting in hybrids with superior biomass characteristics when compared to parents.

Features	Energy	Sugarcane	*S. spontaneum*
Age (months)	12	12	12
Fiber (%)	38,64	14,76	30,46
Stalk lenght (m)	3,1	2,25	2,34
Stalk diameter (cm)	1,29	2,28	0,81
Stalk mediam weight	0,32	0,71	0,11
Leaf weight (kg)	0,08	0,09	0,03
Tops weight (kg)	0,07	0,14	0,05
Purity (%)	77,76	98,72	53,15
Pol (%)	12,67	15,4	3,31
Reducing sugars (%)	1,11	0,51	1,22
Brix (%)	16,3	15,6	-
Tillering	High	Low-medium	High

As expected in an energy cane genotype, the fiber content observed is higher than *S. spontaneum*and sugarcane. This hybrid presents higher capacity of tillering than it's progenitor, *S. spontaneum*. Energy cane stalks observed are longer (3,1m) than other genotypes (2,25m and 2,34m). Fiber content and stalk length features highlighted the heterosis phenomenon observed in energy cane hybrids. Some features of the *Saccharum* hybrid are intermediary, when compared to the parental genotypes, such as stalk diameter, stalk weight and juice purity. It's interesting to note that percentage of pol and brix are close to sugarcane content, while reducing sugars (fructose and glucose) are higher in energy cane and *S. spontaneum* than sugarcane. These results place the *Saccharum* hybrid utilized in our study as an example of energy cane.

Histological characterization of each section defines the developmental stage of internode elongation with magnification of vascular bundle lignification

Sugarcane presents a specific lignification gradient in stalks. The first internode (apex) contain lignin only in proxylem region, while the second internode presents lignification in proto and metaxylem cells. The third internode presents the previous pattern but also is lignified at inner sclerenchyma sheath and hypodermis. In the fourth internode lignin is also observed in outer sclerenchyma sheath and some sieve tubes in central bundles. After the fifth internode, the elongation process ceases and

the internodes are classified as mature. Mature internodes also present lignification in storage parenchyma, epidermis and sieve tubes in peripheral bundles (JACOBSEN ET AL, 1992; LINGLE, 2012).

In our study we characterized the immature internode presenting the highest gradient of lignification from the basis to apex region. To better understand this gradient we divided equally the third internode in five sections, A to E, where A is the apex section and E is the bottom section of the internode (Figure 1). Anatomical features were analyzed microscopically using phloroglucinol-HCL staining (red color) to visualize lignification patterns (Figure 1, B-M) and auto florescence under UV-light to determine the presence of phenolic compounds (light blue) (Figure 1, N-S).

Basal section of internode presents several visible nuclei in parenchyma cells (Figure 1.C, 1.H, 1.M), characteristic of meristematic tissues. This section comprises the basal intercalary meristem (dark green) of energy cane's internode (Figure 1. A, section E) where numerous cells are actively dividing and is formed mainly by primary cell walls (lignin is visible only in proxylem). In section D (Figure 1.G and 1.L) is possible observe the absence of visible nuclei, lignification of protoxylem cells and the beginning of secondary cell walls growth (with lignin deposition) in metaxylem, inner sclerenchyma sheath and xylem parenchyma cells, a process that is followed in the section C (Figure 1.F and 1.K). This lignification pattern is stronger in sections A and B (Figure 1.B, 1.D, 1.E, 1.I and 1.J) where is possible to observe the presence of lignified secondary cell walls in all vascular bundle (protoxylem, metaxylem, inner sclerenchyma sheath and xylem parenchyma cells) as expected in mature internodes.

Phenolic compounds distribution was observed by autoflorescence under UV-light and presented a similar pattern than phloroglucinol-HCL staining. Results showed phloem and proxylem cells with high auto florescence in section E (Figure 1.P and 1.S) while section C (Figure 1.O and 1. R) and A (Figure 1.N and 1.Q) present this fluorescence in vascular bundle tissues - protoxylem, metaxylem, phloem, inner and outer sclerenchyma sheath and xylem parenchyma cells as observed in mature internodes. Anatomical analysis of the lignin and phenolic compounds pattern showed a gradient that comprehends all the variations described in immature internodes (1^{st} - 4^{th}) (JACOBSEN ET AL, 1992) in this unique internode, highlighting its potential as a model to understand biomass formation of *Saccharum*.

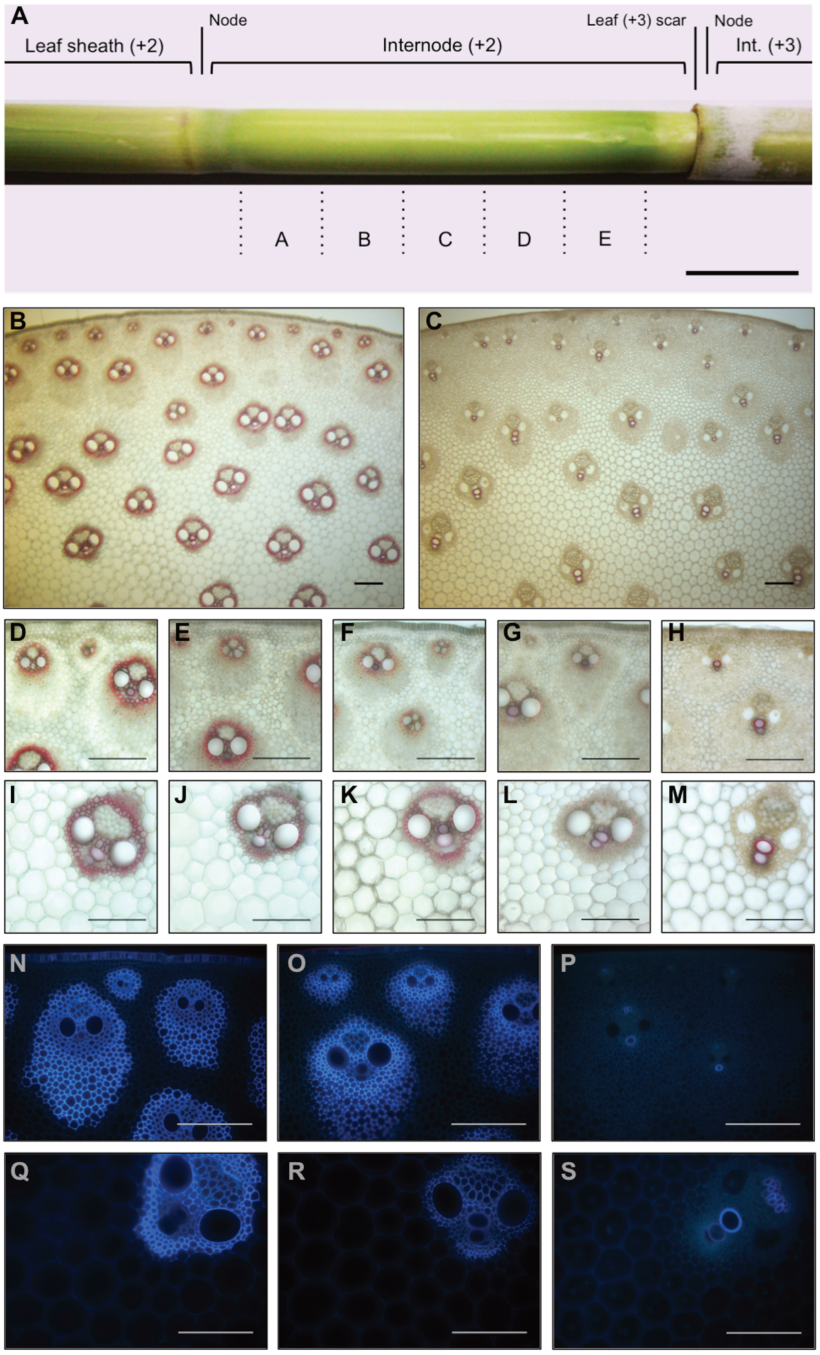


Figure 1. Anatomical characterization of energy cane's elongating internode. From A to M, sections were stained in phoroglucionol-HCl; from N to S, phenolic compounds distribution was observed by autoflorescence under UV-light. **A)** Elongating internode divided in five sections (A-E), were the basis of internode (section E) is composed by basal intercalary meristem and the apex contain fully lignified vascular bundles (section A). **B)** Top section of third internode showing lignified vascular bundles. **C)** Internode bottom section with bundles lignified only around the protoxylem region. **D-M**) Detailed gradient of internode's in the sections (E-A). Section E contains visible nuclei indicating a meristematic activity (H and M) that is absent in other sections. Lignification was observed only in protoxylem of section E (H-M) and strongly in vascular bundle (protoxylem, metaxylem, phloem, xylem parenchyma and sclerenchyma cells) of section A and B. Sections C and D present a weak lignification pattern in vascular bundle tissues. **N-S)** Detailed gradient of internode's phenolic compounds deposition in sections (A, C and E).

De novo assembly of energy cane transcriptome

For each internode section (A, B, C D and E) three biological replicates were sequenced, generating around 15 millions of 50 bp paired-end reads. Based on quality analysis it was possible to determine that a read *trimming s*tep was not required. The ribosomal analysis showed a low percentage of ribosomal sequences (less than 2%, except for two libraries), as exhibit in Supplementary Material - Table 1, highlighting libraries quality. These results allow the pipeline application to all datasets (Supplementary Material – Figure 2).

The *ab-initio* transcriptome assembly produced 264444 transcript sequences larger than 200 bp. The application of a filtering pipeline (reference assembly) was reduced to 52836 protein-coding transcripts (see method - Supplementary Material – Figure 1). Final transcripts list was annotated against three databases, NR, sorghum and swissprot, generating 38419, 36173 and 23776 annotated transcripts, respectively. A Principal component analysis (PCA) was performed to evaluate global transcript expression of sections (Figure 2). PCA graphic shows that major variance among sections is observed in axis x (94,16%), highlighting the differential expression profile from cell division zone (E) and elongation zone (D). Axis X also shows the similarity among the maturation zones (A, B, C). A minor variation in axis y (2%) is an indicative to distinguish subtle differences among the apical lignified sections (A and B) and a slight transition between the elongation and maturation zones (C).

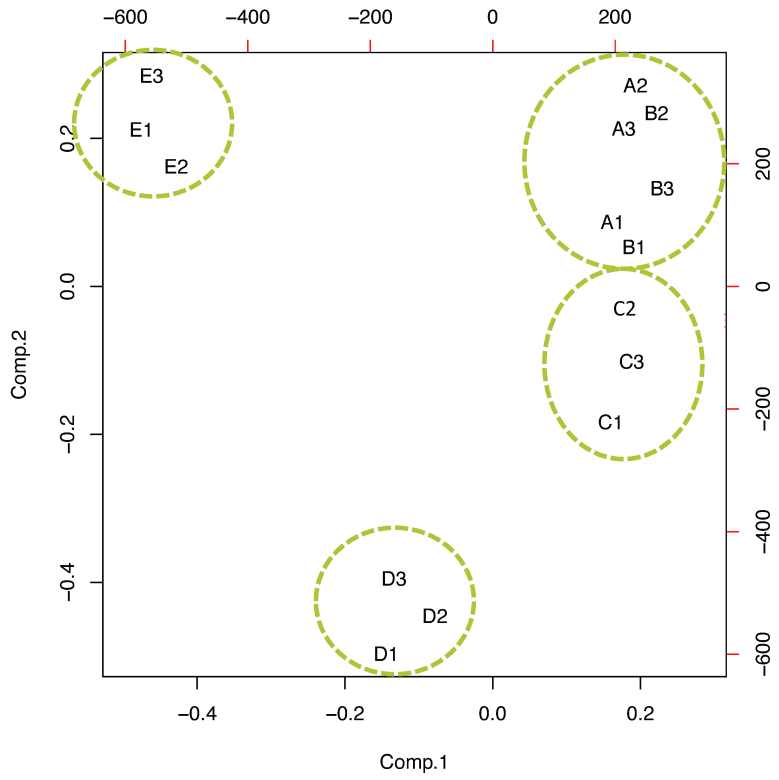


Figure 2. Principal component analysis (PCA) displaying global transcript expression profiles variation among the sections. Major variance among sections is observed in axis x (94,16%) highlighting the specificity of cell division (E) and elongation zones (D), as the similarity among the maturation zones (A, B, C). A minor variation in axis y (2%) is an indicative to distinguish subtle differences among the apical sections (A and B) and a slight transition between the elongation (D) and maturation zones (C).

Global transcription profile of elongating internode sections

In our study, 9468 genes were identified as differently expressed. Differentially expressed genes (DGEs) between sections (p-value <= 0.05 and fold change >= 2) were defined by DESeq2 and edgeR software package (Supplementary Material – Figure 3). In order to obtain a global understanding of DEGs grouping and function we performed a hierarchical clustering. We observed that clustering analysis reveals a gradient of gene expression ranging from cell division to lignification. We divided this gradient in three zones: DZ - division zone (green), EZ – elongation zone (grey) and MZ – maturation zone (red) (Figure 3). Division zone comprises internode E section, sections C and D form elongation zone, and, finally maturation zone includes sections A, B and C.

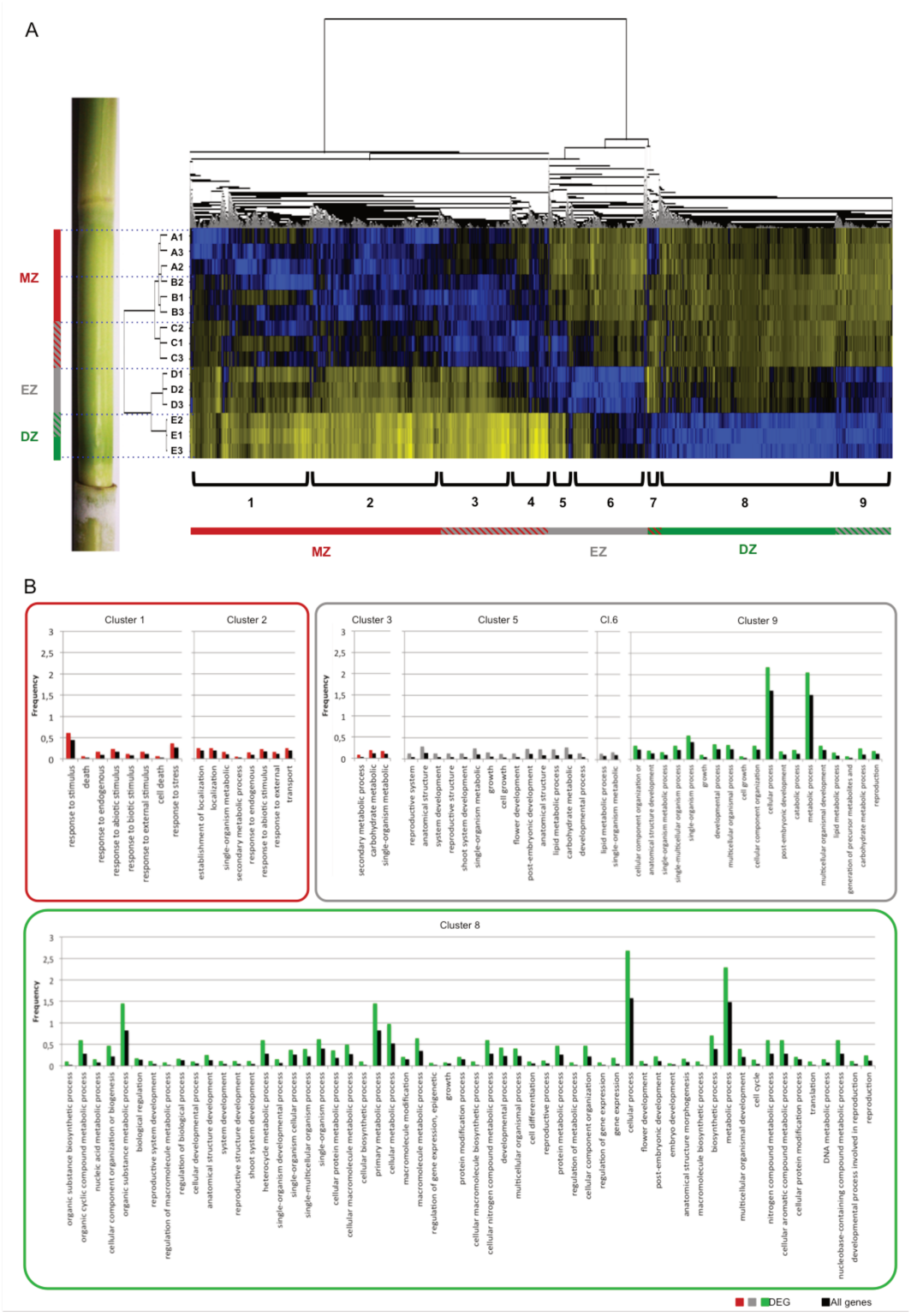


Figure 3. Hierarchical clustering of differentially expressed genes (DEGs). A) Internode gradient was divided in: DZ – division zone (green), EZ – elongation zone (grey) and MZ – maturation zone (red). Genes with higher expression are represented in blue and with lower expression in yellow. **B)** DEGs were divided in nine major clusters that were attributed to maturation zone showing enrichment in secondary metabolic process and lignification of cell wall (clusters 1 and 2), elongation zone with cell expansion, increment in production of cell wall components and beginning of lignin deposition (clusters 3, 5, and 6) and division zone under process of cell division and production of cell components (cluster 8 and 9).

We observed nine major clusters in the clustering results (1-9), which were classified in the internode gradient zones (Figure 3). Each cluster was submitted to GO annotation, using previous NR annotation, in order to define which GO terms show a statistically over (under) - representation in the comparison (enrichment). Results of enrichment GO analysis were grouped into the development zones (MZ, EZ and DZ) (Figure 3).

Maturation zone includes clusters 1 and 2. Cluster 1 showed enrichment in process involved in cell death and response to stimulus (endogenous, external, stress, biotic and abiotic). Cluster 2 was enriched in secondary metabolic process, transport and response to stimulus (endogenous, external, stress and biotic). MZ in histological analysis is characterized by a stronger lignification process than EZ and DZ, which could be related with the enrichment in secondary metabolic process.

Elongation zone comprises clusters 3, 5 and 6. Enriched categories for EZ are related to carbohydrate, lipid and secondary metabolic processes, and also with cell growth and development. EZ also present an interface with DZ on cluster 9, that was characterized by presenting enrichment in process involved in cell growth and development, generation of precursor metabolites, anatomical structure morphogenesis, metabolism of carbohydrate, lipid and secondary metabolites and cellular compound organization.

EZ is under elongation process, with cell expansion, increment in production of cell wall components and lignin deposition. This zone presents an interface with MZ and DZ, highlighted by the overlap of enriched categories. As an example, both MZ and EZ present higher expression of secondary metabolic process (related to lignin metabolism), while EZ and DZ presents enrichment to cell growth and development process.

Division zone is characterized by an intense process of cell division which is reflected in the enriched GOs for this zone (cluster 8) as cell cycle, cellular component organization, cellular developmental process, cellular protein biogenesis and modification, development, nucleic acid process, protein biogenesis and modification, cell differentiation and anatomical structure morphogenesis.

Clustering analysis of DEGs allowed a division of the internode in three major zones of maturation, elongation and division. Expression profile of these zones presents enrichment in some metabolic categories that reflected the biological process characteristic of the internode region.

Transcription profiles of genes involved in cell wall formation during the internode elongation

Cellulose synthases genes (CesA) observed in this study were divided in two groups, one involved in cell wall deposition in primary cell walls (Figure 4.a) and the second related to cellulose synthesis in secondary cell wall (Figure 4.b). Genes involved in primary walls, in general, showed lower expression than genes of secondary walls. Although, in section E, a zone of cell division, the expression of CesA genes involved in primary cell walls was higher than secondary ones. Genes from group 2 increased the transcript abundance from very low levels in section E, to high levels in elongating sections D and C, and started to decrease in maturation sections B and A. Similar patterns of CesA genes were observed in a maize elongating internode (ZHANG et al. 2014). Other genes involved in cellulose deposition, glycosyl hydrolase 9B7 (GH987) and KORRIGAN (KOR) displayed a similar pattern of CesA group 2 (Figure 4.d), with higher expression in the elongation transition zone and maturation process (section C).

Glycosyltransferase genes (GTs) have been related with cell wall biosynthesis, more specifically, members of GT43 family of genes that were involved in glucurono-arabinoxylans (GAX) synthesis showed high expression mainly in sections C and B (Figure 4.c) related to elongation and maturation process. In maize this family also presented the higher peak in elongating and transition zones (ZHANG et al. 2014). Other genes involved in GAX metabolism by directing sugars nucleotides to hemicellulose precursors, are UGDH (UDP-glucose dehydrogenase) and UXS (UDP-Xyl synthase).

UGDH catalyzes the irreversible oxidation of UDP-glucose to UDP-glucuronate, resulting in a unidirectional flow of UDP-glucuronate (UDP-GlcA) into a group of sugar nucleotides used in hemicelluloses and pectins biosynthesis (GIBEAUT AND CARPITA, 1994; LABATE et al. 2010). UXS is responsible for catalyzing the irreversible decarboxylation of UDP- GlcA to form UDP-Xyl, that results in commiting carbon to the formation of pentose sugars, the main component of GAX polymers. UGDH and UXS showed similar patterns with higher gene expression in the elongation zone (sections C and D) (Figure 4.h).

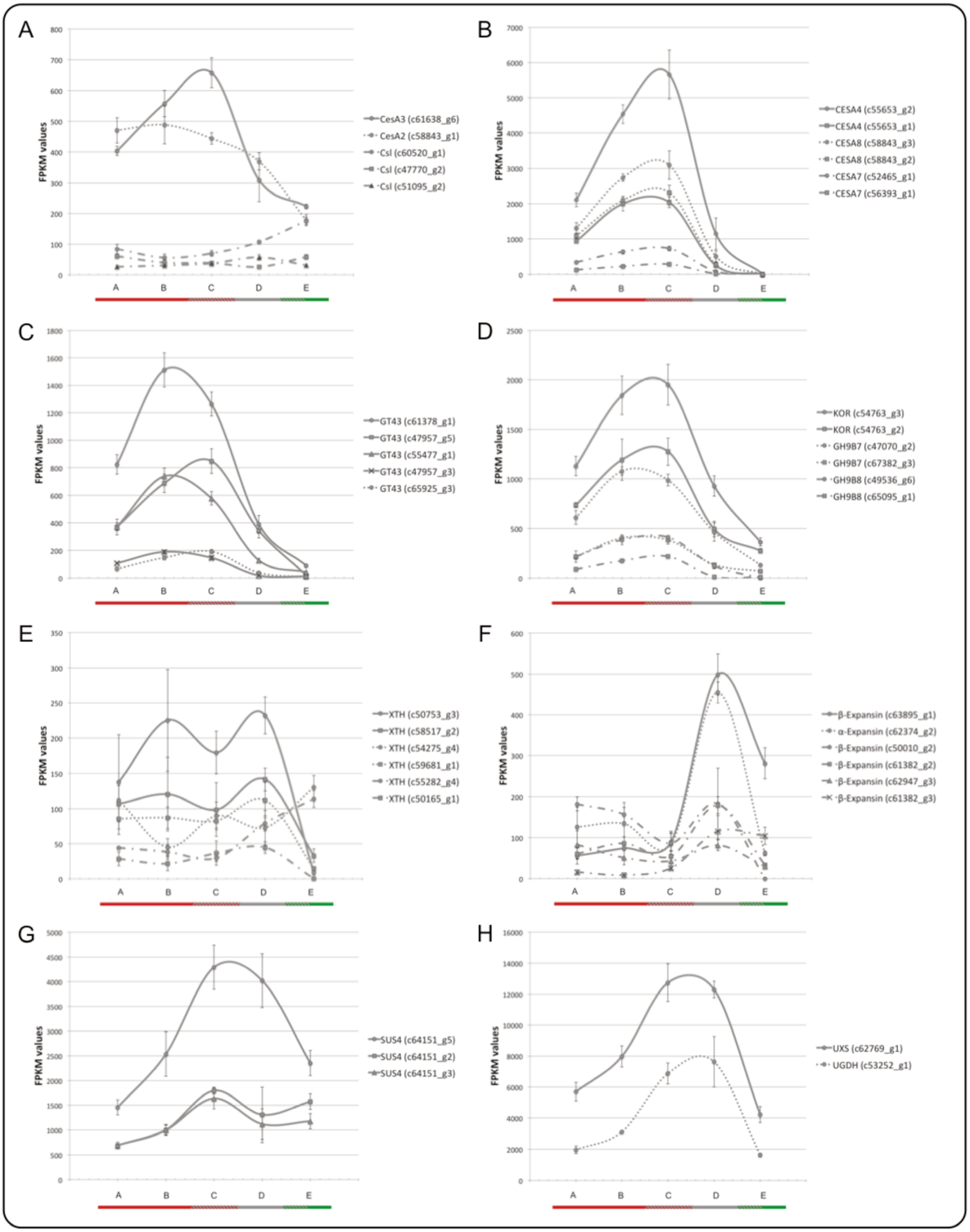


Figure 4. Transcript levels of genes involved in **a)** cellulose synthesis of primary cell wall (CesA3, CesA2 and Csl), **b)** cellulose metabolism of secondary cell wall (CesA4, CesA7 and CesA8), **c)** GAX synthesis (GT43) **d)** cellulose deposition (KOR and GH987) **e)** remodeling of xyloglucans (XTH), **f)** cell wall extension (α and β expansins) **g)** sucrose metabolism (SUS4) **h)** GAX synthesis (UGDH and UXS).

Xyloglucanendotransglycosylase/hydrolase (XTH) enzymes were expected to play a role in the remodeling of cell wall structures, in particularly xyloglucans. As in maize (ZHANG et al. 2014), these genes did not display an expression pattern during the internode elongantion (Figure 4.e).

Expansins are involved in cell wall extension. These enzymes showed high expression in elongation zone (D), agreeing with its function as loosening agents during cell elongation (Figure 4.f).

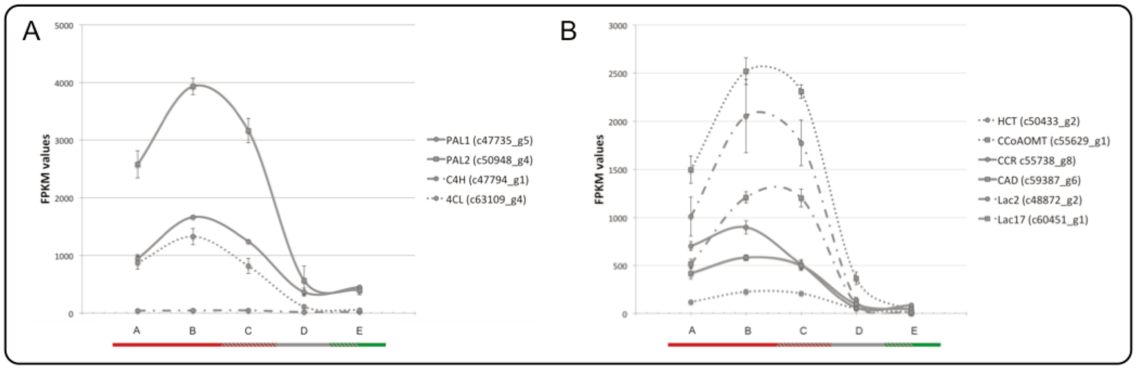


Figure 5. Transcript levels of genes involved in **a)** phenylpropanoids metabolism and **b)** lignin metabolism.

Genes involved in phenylpropanoids metabolism, precursors of lignin and genes related to lignin metabolism showed a similar expression pattern (Figure 5.a and 5.b). These genes present very low expression in division zone (section E) and progressively increased its expression in subsequent sections, zones of elongation and maturation, with a peak of transcript levels in sections B and C. Its pattern is corroborated by the histological characterization of lignification gradient ranging from E (less lignified) to A (vascular bundle lignification).

Energy cane and its parental lineages (*S. spontaneum* and sugarcane)

Since the elongating internode of energy cane displayed an interesting gradient, with regions characterized by presenting more lignified cell walls, and cell wall composition is an important and distinctive feature of sugarcanes and high biomass genotypes as *S. spontaneum* and energy cane, we performed a PCA analysis comparing energy cane internode sections (A-E) and a third internode (not

sectioned) RNAseq from *S. spontaneum* and sugarcane (energy cane progenitors) as showed in Figure 6 (data produced in other experiments of the group). It's possible to observe that major component to separate the samples is responsible by 91,49% of the differences between then and divide the samples in three groups: the first comprising sugarcane closely of sections D and E, the second with the maturation zone of energy cane (A, B and C) and third with *S. spontaneum* replicates. It is interesting to note that *S. spontaneum* is closely related to group 2 (MZ) rather than group 3 (sugarcane + DZ and EZ). This result could indicate that the pattern observed in an elongating internode (ranging from less lignified to lignified cell walls, especially in vascular bundle sheet) may help to understand the difference between genotypes selected to storage high amounts of soluble sugars in culm and another ones selected to present high fiber and biomass content.

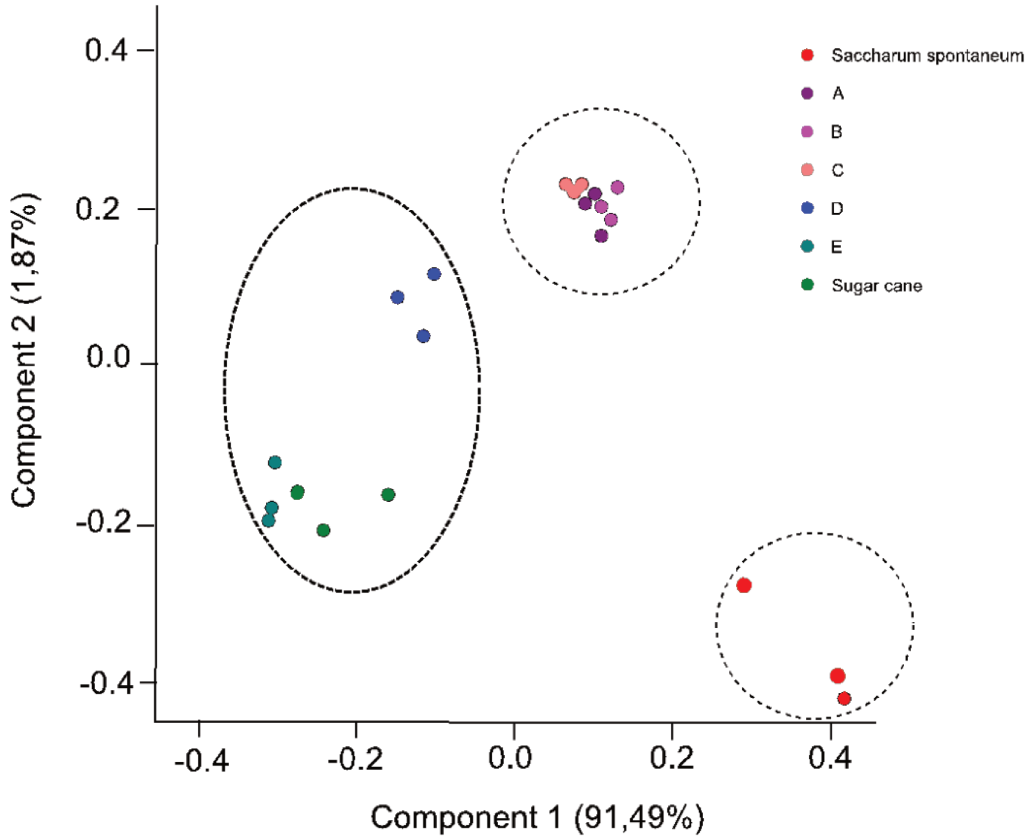


Figure 7. Principal component analisys (PCA) displaying global transcript expression profiles variation of five sections (A-E) of energy cane elongating internode in comparison with an elongating internode of the progenitors genotypes: *Saccharum spont*aneum (red) and sugarcane (green). This graph highlighted sugarcane and *Saccharum spontaneum* variation when compared with the sections of energy cane internode, showing that sugarcane internode (green) is closer to less lignified sections D and E (blue and lightblue), while *S. spontaneum* internode (red) is closely of lignified sections A, B and C (purple, pink, light pink).

GENERAL CONCLUSIONS

In order to better understand the lignocellulosic biomass of energy cane, we performed a characterization of its formation, using internode in elongation as a model.

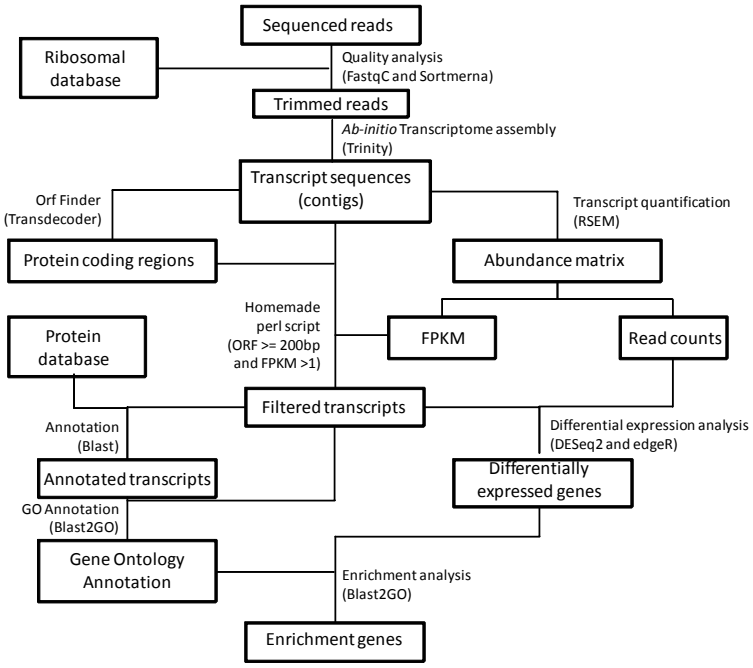
Our histological analysis showed a distinctive gradient of lignification on this internode, from cell division to lignification. As a sign of cell division, it was possible to observe that internode basal region (section E) presents visible nuclei. RNAseq analysis of five internode sections confirmed this pattern with the differentially categories of expressed genes, resulting to a higher lignin biosynthesis genes expression in the apical internodes (sections A and B).

Analysis of gene expression of specific pathways involved in cellulose, hemicellulose, lignin and sucrose metabolism showed specific patterns for each component's metabolism. These results indicate that the elongating internode dynamics could synthetize, in one internode, the development pattern of all immature internodes of *Saccharum*, constituting an excellent model for biomass formation of the genera.

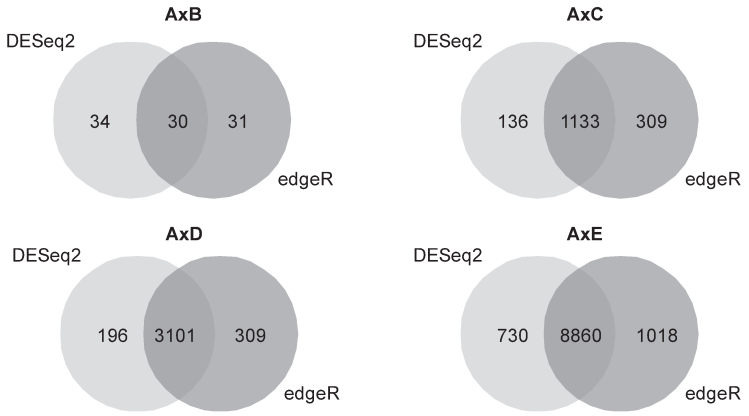
We also observed in this study that the pattern observed in an elongating internode (ranging from less lignified to lignified cell walls, especially in vascular bundle sheet) may help to understand the difference between hybrids selected to distinct purposes, as storage high amounts of soluble sugars in culm or present high fiber and biomass content.

SUPPLEMENTARY MATERIAL

Supplementary- Figure 1- Pre- commercial energy cane developed by GranBio.



Supplementary - Figure 2. Transcriptome pipeline. It is a representation of the entire pipeline to identify differential expressed genes between different groups.



Supplementary - Figure 3. DESeq2 and edgeR comparison. Differential expressed genes in a pairwise comparison with DESeq2 and edgeR using A as common reference.

Supplementary - Table 1. Sequencing quality analysis. Sequencing results for the libraries in the first column (triplicate of each section A-E). The second column show the quantity of reads that were generated and the third one shows the percentage of ribosomal reads in this libraries.

Libraries	Reads processed	Ribosomal reads
A1	19737315	1.08%
A2	10883232	1.93%
A3	15869799	4.21%
B1	10475107	1.23%
B2	10750170	1.98%
B3	13775442	6.69%
C1	15470843	1.33%
C2	15727753	1.47%
C3	12986005	1.54%
D1	11853797	1.62%
D2	17128850	1.44%
D3	14611299	0.80%
E1	19226585	1.84%
E2	18043767	1.49%
E3	13422750	2.14%

5 CONCLUSÕES

- O primeiro *draft* do genoma de *S. spontaneum* foi concluído e contém a maior parte dos genes completos da espécie.

- O draft é uma base de dados pioneira e importante para ancoragem de dados do transcritoma, informações de sequências completas de genes expressos e regiões reguladoras.

- As sequências promotoras, tecido-específico e ubíquas, identificadas no estudo possuem aplicação imediata para biotecnologia de *Saccharum*.

- O internódio em alongamento de cana energia analisado constitui um ótimo modelo para controle transcricional da formação da parede celular, em especial a lignificação dos feixes vasculares.

- O padrão de desenvolvimento do internódio em alongamento pode sintetizar, em um único internódio, o desenvolvimento dos internódios imaturos de cana.

- Análises comparativas das seções do internódio em alongamento de cana energia com internódios inteiros de *S. spontaneum* e cana-de-açúcar, indicam que o internódio em alongamento pode explicar as diferenças entre os híbridos selecionados para diferentes propósitos (acumulo de açúcares solúveis no colmo ou alto conteúdo de fibra e alto rendimento de biomassa).

Assim, durante o projeto foi produzido um atlas genômico, transcritômico e glicômido de *S. spontaneum* e seu híbrido de cana energia, que servirá de base para entender os processos de formação e arquitetura fina da parede celular de *Saccharum*.

6 REFERÊNCIAS BIBLIOGRÁFICAS

AL-JANABI, S.M.et. al. A genetic linkage map of Saccharum spontaneum (L.) 'SES 208'. **Genetics**, v. 134, p. 1249-1260, 1993.

ALTSCHUL, S. F., et al. Gapped Blast and PSI-Blast: A new generation of protein database search programs. **Nucleic Acids Research** ,v. 25, p. 3389–3402, 1997.

AMALRAJ, V.A.; BALASUNDARAM, N. On the taxonomy of the member s of '*Saccharum* complex'. **Genetic Resources and Crop Evolution**, v.53, p.35-41, 2006.

ARGOUT, X. et al. The genome of *Theobroma cacao*. **Nature Genetics**, v.43, n.2, p. 101-108, fev. 2011

ASHBURNE, M. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, **Nature Genetics**, p. 25-29, 200.

BIOINFORMATICS B. FASTQC: a quality control tool for high throughput sequence data. Cambridge, UK: Babraham Institute; 2011.

BATEMAN, A., *et al.* The Pfam Protein Families Database. **Nucl. Acids Res**., 30(1), p. 276-280, 2002.

BATISTA, F. Etanol de 2ª geração ganha escala comercial, Associação da Indústria de Cogeração de Energia, 2013. Disponível em: < http://www.cogen.com.br/noticia_print.asp?id_noticia=12011>. Acesso em 29 set.2013

BAUDET, C.; DIAS, Z. New EST Trimming Strategy in: Brazilian Symposium on Bioinformatics. **Lecture Notes in Bioinformatics**, Berlin, Germany, 2005. Springer: Verlag. v. 3594 pp. 206-209, 2005.

BNES & CGEE. Bioetanol de cana-de-açúcar : energia para o desenvolvimento sustentável. Rio de Janeiro : BNDES, 2008. 316 p.

BOTTCHER A., et al. Lignification in Sugarcane: Biochemical Characterization, Gene Discovery, and Expression Analysis in Two Genotypes Contrasting for Lignin Content. **Plant Physiology**, Vol. 163, pp. 1539–1557, December 2013.

BRASIL – Banco Central do Brasil. **Indicadores Econômicos Consolidados**, 2012. Disponivel em< http://www.bcb.gov.br> Acesso em 26 mar. 2012.

BRASIL. Ministério da Agricultura, Pecuária e Abastecimento. **Anuário Estatístico da Agroenergia. Brasília**, DF, 2009. 160 p. Disponível em < http://www.agricultura.gov.br> Acesso em 23 mar. 2012

BROOKES, A. J. The essence of SNPs. **Gene,** v.234, p.177-186, Jul. 1999

CARVALHO C.R. ET al. Genome size, base composition and karyotype of Jatropha curcas L., an important biofuel plant. **Plant Sci** 174:613–617, 2008.

CASU, R.E. et al. dentification of transcripts associated with cell wall metabolism and development in the stem of sugarcane by Affymetrix GeneChip Sugarcane Genome Array expression profiling. **Funct Integr Genomics**, v.7, p.153–167, 2007

CNA - Confederação da Agricultura e Pecuária do Brasil. **Análise do PIB das cadeias produtivas do algodão, cana-de-açúcar, soja, pecuária de corte e de leite no Brasil**. Brasília, DF, 2012. 68 p.

CONAB - Companhia Nacional de Abastecimento. Acomp. safra bras. cana, v. 3 - Safra 2016/17, n. 1 - Primeiro levantamento, Brasília, p. 1-66, abr. 2016.

CORDEIRO, G.M.; TAYLOR, G.O.; HENRY, R.J. Characterization of microsatellite markers from sugarcane (*Saccharum* spp.), a highly polyploid species. **Plant Science** 155: 161-168, 2000.

CORDEIRO, G.M. et al. Microsatellite markers from sugarcane (*Saccharum* spp) ESTs cross transferable to Erianthus and Sorghum. **Plant Science** 160: 1115-1123, 2001

DOMINGUES, D.S. **SURE e Garapa: caracterização molecular e distribuição de dois retrotransposons com LTR em cana-de-açúcar.** 2004. Tese (Doutorado em Biotecnologia) – Biotecnologia, Universidade de São Paulo, São Paulo, 2009.

DA SILVA, J.A.G. et al. RFLP linkage map of Saccharum spontaneum. **Genome** 36: 782-791, 1993.

DAL-BIANCO, M. et al. Sugarcane improvement: how far can we go? **Current Opinion in Biotechnology**, v.23, p.265–270, 2012.

DE SETTA, N .et al. Building the sugarcane genome for biotechnology and identifying evolutionary trends. **BMC Genomics** 15:540, 2014.

DE SOUZA, A. P., LEITE, D. C. C., PATTATHIL, S., HAHN, M. G. & BUCKERIDGE, M. S. Composition and Structure of Sugarcane Cell Wall Polysaccharides: Implications for Second-Generation Bioethanol Production**. Bioenergy Res**. 6, 564–579, 2013.

DE SOUZA, A. P. et al. How cell wall complexity influences saccharification efficiency in Miscanthus sinensis. **J. Exp. Bot**., 2015.

DOS SANTOS, L.V. et al., Industrial Biotechnology., v.12(1), p.40-57. Feb. 2016
D'HONT, A. et al. Determination of basic chromosome numbers inthe genus *Saccharum* by physical mapping of ribosomal RNA genes. Genome, v.41, p. 221-225,1998.

D'HONT, A. Unraveling the genome structure of polyploids using FISH and

GISH;examples of sugarcane and banana.Cytogenetic Genome Research. v.109, p. 27-33, 2005.

EDWARDS, D.; BATLEY, J. Plant genome sequencing: applications for crop improvement. **Plant Biotechnology Journal** v. 8 , p. 2–9, 2010.

FAO - FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS. **Top producion – sugar cane**. 2013. Disponível em: <http://faostat3.fao.org/browse/Q/QC/E>. Acesso em: 23 jul. 2016

FASTX-Toolkit. FASTQ/A short-reads pre-processing tools. Disponível em: < http://hannonlab.cshl.edu/fastx_toolkit/> Acesso em 04 mai. 2013.

FERREIRA et al. Biofuel and energy crops: high-yield Saccharinae take center stage in the post-genomics era, **Biofuel and energy crops**, 14:210, 2013.

FIGUEIRA et al. A BAC library of the SP80-3280 sugarcane variety (*Saccharum* sp.) and its inferred microsynteny with the sorghum genome, **BMC Research Notes**, 5:185, 2012.

FRESHOUR, G. et al. Developmental and Tissue-Specific Structural Alterations of the Cell-Wall Polysaccharides of *Arabidopsis thaliana* Roots. **Plant physiology** 110, 1996.

GARCIA, A., et al. SNP genotyping allows an indepth characterization of the genome of sugarcane and other complex autopolyploids. **Science Reports** 3: 3399, 2013.

GASMEUR et. al.,High homologous genes conservation despite extreme autopolyploid redundancy in sugarcane. **New Phytologist**, 189(2), p.629-642, 2011.

GOFF, S.A. et al. A draft sequence of the rice genome (*Oryza sativa* L. ssp japonica). **Science**, v. 296 ,p. 92-100, 2002.

GRABHERR, M.G. et. al. Full-length transcriptome assembly from RNA-seq data without a reference genome. **Nat Biotechnol,** 15; 29(7), p.644-652, 2011.

GRATIVOL, C. et al., Sugarcane genome sequencing by methylation filtration provides tools for genomic research in the genus *Saccharum*. **Plant J** 79:162---172, 2014.

GRIVET, L. et al. RFLP Mapping in Cultivated Sugarcane (*Saccharum* spp.): Genome Organization in a Highly Polyploid and Aneuploid Interspecific Hybrid. **Genetics**, v.142, p. 987- 1000, mar.1996.

GRIVET, L.; ARRUDA, P. Sugarcane genomics: depicting the complex genome of an important tropical crop. **Current Opinion in Plant Biology**, n.5, p.122–127, dez. 2001.

GRIVET, L.et al. A review of recent molecular genetics evidence for sugarcane evolution and domestication. **Ethnobotany Research & Applications**, n.2, p.9-17, 2004.

GUIMARÃES, C.T., SILLS, G.R., SOBRAL, B.W.S. Comparative mapping of Andropogoneae:*Saccharum* L. (sugarcane) and its relation to sorghum and maize. **Proceedings of the National Academy of Sciences of the United States of America.** V.94, p.14261-14266, dez. 1997.

GUPTA, P.K.; RUSTGI, S. Molecular markers from the transcribed/expressed region of the genome in higher plants. **Functional e integrative genomics**. v.4, p.139-162, abr. 2004.

IBGE – Instituto Brasileiro de Geografia e Estatística. Indicadores IBGE. Estatística da Produção Agrícola. 2012. Disponível em < http://www.ibge.gov.br > Acesso em 23 mar. 2012

HAAS, B.J., et al., De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc, 2013. 8(8): p. 1494-512.

HASSUANI , S. J. et al. **Biomass power generation: sugar cane bagasse and trash.** Piracicaba: PNUD-CTC, Série Caminhos para Sustentabilidade, 2005.

HILL, J. et al. Environmental, economic, and energetic costs and benefits of biodiesel and ethanol biofuels. PNAS, v. 103, 11206 –11210, jul. 2006.

HODKINSON T. R. et. al. *Saccharum* and related genera (Saccharinae, Andropogoneae, Poaceae) based on DNA sequences from ITS nuclear ribosomal DNA and plastid trnL intron and trnL-F intergenic spacers. **Journal of plant research**, v.115, p. 381–92, 2002.

HRIBOVA, E. et al. Analysis of genome structure and organization in banana (Musa acuminata) using 454 sequencing. **Plant and Animal Genomes,** XVII, 2009.

HUANG, S. et al. Recent developments of the cucumber genome initiative-an international effort to unlock the genetic potential of an orphan crop using novel genomic technology. **Plant and Animal Genomes,** XVII, 2009.

HUANG, X.; MADAN, A. CAP3: A DNA Sequence Assembly Program. **Genome Research,** 9, p.868-877, 1999

IEA - International Energy Agency, World Energy Outlook 2008, 2008.

IEA - International Energy Agency, World Energy Outlook 2012, 2012.

IEA - International Energy Agency, Key World Energy STATISTICS, 2013.

IRVINE, J.E. Saccharum species as horticu ltural classes. **Theoretical and Applied Genetics**, v.98, n.2, p.186-94, 1999.

JACOBSEN, K. R., FISHER, D. G., MARETZKI, A. AND MOORE, P. H. Developmental Changes in the Anatomy of the Sugarcane Stem in Relation to Phloem Unloading and Sucrose Storage. Botanica Acta, 105: 70–80, 1992

JANNOO, N. et al. Orthologous comparison in a gene-rich region among grasses reveals stability in the sugarcane polyploid genome. **Plant J**. 50, p. 574–585, 2007.

JÖNSSON, L. J., ALRIKSSON, B. & NILVEBRANT, N. O. Bioconversion of lignocellulose: inhibitors and detoxification. **Biotechnol Biofuels** 6, 16, 2013.

JORDÃO JUNIOR, H. **Desenvolvimento de um sistema basado em marcadores moleculares de DNA tipo microssatélites para identificação de variedades de cana-de-açúcar**,2009. Dissertação (Mestrado em Biologia Funcional e Molecular) – Universidade Estadual de Campinas , Campinas, 2009.

JUNG, H.G., CASLER, M.D. Maize Stem Tissues: Cell Wall Concentration and Composition during Development, **Crop Science,** v.46, p. 1793–1800, 2006a.

JUNG, H.G., CASLER, M.D. Maize Stem Tissues: Impact of Development on Cell Wall Degradability, **Crop Science,** v.46, p.1801–1809, 2006b.

KANEHISA, M.; GOTO, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. **Nucl. Acids Res.**, 28 (1), p.27-30, 2000.

KARP, A.; SHIELD, I Bioenergy from plants and the sustainable yield challenge, New Phytologist, v.179: p.15–32, 2008.

KOSKI, L.B. et. al. AutoFACT: An Automatic Functional Annotation and Classification Tool. **BMC Bioinformatics**, 6:151, 2005.

KUIJPER J. DeGroei van Bladschijf, Bladscheede em Stengel van het suikerriet. Arch Suikerind Ned Indië 1915;23:528–556.

LABATE M. T. V., et al. Cloning and endogenous expression of a Eucalyptus grandis UDP-glucose dehydrogenase cDNA. **Genetics and Molecular Biology.** 33(4):686-695, 2010.

LANGMEAD, B., Aligning short sequencing reads with Bowtie. Curr Protoc Bioinformatics, Chapter 11: p. Unit 11 7, 2010.

LEITE, R.C.C.; LEAL, M.R.L.V.; CUNHA, M.P. A Guerra entre Petróleo e Etanol. Interesse Nacional, n. 22, p.55-63, jul-set 2013.

LI, B. AND C.N. DEWEY, RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. Bmc Bioinformatics, 2011. 12: p. 323.

LI, H.; DURBIN, R. Fast and accurate short read alignment with Burrows-Wheeler Transform. **Bioinformatics**, v. 25, p.1754-1760, 2009.

LINGLE, S.E.; SMITH, C. S. Sucrose Metabolism Related to Growth and Ripening in Sugarcane Internodes, CropSci, v.31, p.172-177, 1991.

LINGLE, S.E Seasonal Internode Development and Sugar Metabolism in Sugarcane, **CropSci**, v.37, p.1222-1227, 1997.

LINGLE, S.E; TEW, T. L. A Comparison of Growth and Sucrose Metabolism in Sugarcane Germplasm from Louisiana and Hawaii,
Crop Science, v. 48, p. 1155-1163, 2008.

LINGLE, S.E.; THOMSON, J.L. Sugarcane Internode Composition During Crop Development. **Bioenerg. Res**., v.5, p.168–178, 2012.

LOGEMANN J, SCHELL J, WILLMITZER L:Improved method for the isolation of RNA from plant tissues.**Anal Biochem**,163:16-20, 1987.

LOHSE M., et al. Mercator: a fast and simple web server for genome scale functional annotation of plant sequence data. **Plant Cell Environ**, nov. 2013.

LUO et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. **GigaScience** 1:18, 2012.

MARDIS, E. R. Next-Generation DNA Sequencing Methods. **Annual Review of Genomics and Human Genetics**, v. 9, p. 387-402, jun. 2008.

MATSUOKA, S.; FERRO, J. ; ARRUDA, P. The Brazilian experience of sugarcane ethanol industry. **In Vitro Cell.Dev.**Biol.—Plant v. 45, p.372– 381, 2009.

MATSUOKA, S. et al. Energy Cane: Its Concept, Development, Characteristics, and Prospects. Adv. Bot. 2014, 1–13 (2014).

MENOSSI, M. et al. Sugarcane functional genomics:gene discovery for agronomic trait development. **International Journal of Plant Genomics**. v. 2008,11 p. 2008.

MING, R. et al. QTL Analysis in a Complex Autopolyploid:Genetic Control of Sugar Content in Sugarcane. **Genome research**, v.11, p. 2075-84, 2001.

MING, R. et al. Sugarcane Improvement through Breeding and Biotechnology. Plant Breeding Reviews 118, 2006.

MING, R., ET AL. **Sugarcane Improvement through Breeding and Biotechnology**, in Plant Breeding Reviews, Volume 27 (ed J. Janick), UK. 2010

MORIYA, Y. et al. KAAS: an automatic genome annotation and pathway reconstruction server. **Nucleic Acids Res**. 35, 182-185, 2007.

KAYAK, S. N. et. al. Promoting Utilization of Saccharum spp. Genetic Resources through Genetic Diversity Analysis and Core Collection Construction. PLoS ONE 9(10): e110856, 2014

NCBI - National Center for Biotechnology Information. National Library of Medicine, National Health Institutes. Pubmed. Washington, dc, 2004. Disponível em: < http://www.ncbi.nlm.nih.gov/> Acesso em: 20 jun. 2012.

NEVES, M.F.; TROMBIN, V.G.; CONSOLI, M. **The Sugar-Energy Map of Brazil**. In: Eduardo L. Leão de Sousa e Isaias de Carvalho Macedo. (Org.). Ethanol and Bioeletricity: Sugarcane in the Future of the Energy Matrix. 1 ed. São Paulo: Luc Projetos de Comunicação, v. 1, p. 14-43, 2011.

PAPANI-TERZI, F. S. et al. Sugarcane genes associatedwith sucrose content, **BMC Genomics**, 20:210, 2009

PATERSON A.H. et. al. The *Sorghum bicolor* genome and the diversification of grasses. **Nature**. v. 457:551, 2009.

PATTATHIL, S., HAHN, M. G., DALE, B. E. & CHUNDAWAT, S. P. S. Insights into plant cell wall structure, architecture, and integrity using glycome profiling of native and AFEXTM-pre-treated biomass**. J. Exp. Bot.** 1–16, 2015.

POPPER, Z. A. Evolution and diversity of green plant cell walls. Curr. Opin. **Plant Biol**. 11, 286–292, 2008.

PUNTA, M. et al.The Pfam protein families database, **Nucleic Acids Research, 2012**

RAJANDEEP S. S. Genome-wide atlas of transcription during maize development, The Plant Journal, v. 66, p. 553–563, 2011

SANTOS et al, Second-Generation Ethanol: The Need is Becoming a Reality. **Industrial Biotechnology,** p. 40-57, fev, 2016.

SILVEIRA, L.C. I. da; Melhoramento genético da cana-de-açúcar para obtenção de cana energia. Tese (Doutorado em Agronomia), Universidade Federal do Paraná, Curitiba, 2014.

SHIELDS, S. ; BOOPATHY, R. Ethanol production from lignocellulosic biomass of energy cane. **International Biodeterioration & Biodegradation**, v. 65, 142 -146, 2011.

SLATER, G.S.C.; BIRNEY, E. Automated generation of heuristics for biological sequence comparison. **BMC Bioinformatics** 6:31, 2005

STAJICH JE, et.al. The bioperl toolkit: Perl modules for the life sciences. **Genome Res**
12, p.1611–1618, 2002.

Sugarcane Genome Project. Disponível em: < http://sugarcanegenome.org/ >. Acesso 15 mar. 2012.

SUMAN, A. et al. Sequence-related amplified polymorphism (SRAP) markers for assessing genetic relationships and diversity in sugarcane germplasm collections. **Plant Genetic Resources: Characterization and Utilization,** p. 222 -231, 2008.

SUZEK, B.E., et. al. Uniref: comprehensive and non-redundant UniProt reference clusters. **Bioinformatics**, 23 (10), p.1282-1288, 2007.

TEW, T.L.; COBILL, R.M. Genetic Improvement of Sugarcane (Saccharum spp.) as an Energy Crop, Genetic Improvement of Bioenergy Crops, p. 249-272, 2008.

The Arabidopsis Genome Iniciative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. **Nature**, v.408 , p. 796–815, 2000.

The Potato Genome Sequencing Consortium. Genome sequence and analysis of the tuber crop potato. **Nature**, 475, p. 189-195, jul. 2011

TOMKINS, J.P. et al. A bacterial artificial chromosome library for sugarcane. **Theor Appl Genet**, 99(3–4), p.419–424, 1999

ÚNICA – União da Indústria de Cana de Açúcar. **Setor Sucroenergético**. Disponível em: < http://www.unica.com.br}> . Acesso em: 24 mar. 2012

USDA, ARS, National Genetic Resources Program, Germplasm Resources Information Network (GRIN). Disponível em <http://www.ars-grin.gov/npgs/>. Acesso em 02 mai. 2012.

_____. **Produção Brasileira de Cana-de-açúcar, Açúcar e Etanol**, 2012. Disponível em < http://www.agricultura.gov.br> Acesso em 23 mar. 2012

VEECKMAN, E.; RUTTINK, T.; VANDEPOELE, K, Are We There Yet? Reliably Estimating the Completeness of Plant Genome Sequences. The Plant Cell, ago 2016

VETTORE A.L. et. al. Analysis and functional annotation of an expressed sequenc e tag collection for tropical crop sugarcane**. Genome R**

YU, J. et al. A draft sequence of the rice genome (Oryza sativa L.spp Indica), Science, 296, 79, 2002.

ZHANG, J., et. al. Genome size variation in three Saccharum species. **Euphytic**a, v.185, p. 511-519, 2012

ZHANG, Q. et al. Spatial gradients in cell wall composition and transcriptional profiles along elongating maize internodes. **BMC Plant Biology**, 14:27, 2014.

ANEXO I - Enriquecimento da categoria de processos de GOs (Gene Ontology) de genes tecido-específico de folha e raiz.

	GO-ID	Term	Category	FDR	P-Value	#Test	#Ref	#notAnnotTest	#notAnnotRef	Over/Under
Folha	GO:0015979	photosynthesis	P	5,39E-93	1,65E-95	120	357	523	30368	over
	GO:0006091	generation of precursor metabolites and energy	P	2,47E-49	1,51E-51	109	799	534	29926	over
	GO:0051186	cofactor metabolic process	P	2,99E-15	2,28E-17	53	602	590	30123	over
	GO:0006790	sulfur compound metabolic process	P	2,39E-11	2,56E-13	42	502	601	30223	over
	GO:0042254	ribosome biogenesis	P	3,53E-10	4,85E-12	33	349	610	30376	over
	GO:0016491	oxidoreductase activity	F	9,32E-10	1,42E-11	113	2773	530	27952	over
	GO:0022613	ribonucleoprotein complex biogenesis	P	1,81E-09	3,03E-11	33	376	610	30349	over
	GO:0061024	membrane organization	P	3,74E-06	8,56E-08	28	399	615	30326	over
	GO:0005975	carbohydrate metabolic process	P	1,67E-05	4,33E-07	77	2010	566	28715	over
	GO:0071822	protein complex subunit organization	P	4,49E-05	1,39E-06	32	575	611	30150	over
	GO:0070271	protein complex biogenesis	P	4,49E-05	1,39E-06	32	575	611	30150	over
	GO:0006461	protein complex assembly	P	4,49E-05	1,39E-06	32	575	611	30150	over
	GO:0044085	cellular component biogenesis	P	6,49E-05	2,28E-06	47	1056	596	29669	over
	GO:0065003	macromolecular complex assembly	P	2,19E-04	8,37E-06	32	630	611	30095	over
	GO:0043933	macromolecular complex subunit organization	P	2,19E-04	8,37E-06	32	630	611	30095	over
	GO:0065008	regulation of biological quality	P	3,59E-04	1,48E-05	27	499	616	30226	over
	GO:0042592	homeostatic process	P	3,59E-04	1,48E-05	27	499	616	30226	over
	GO:0044281	small molecule metabolic process	P	5,38E-04	2,38E-05	109	3507	534	27218	over
	GO:0022607	cellular component assembly	P	6,34E-04	2,90E-05	36	801	607	29924	over
	GO:0044710	single-organism metabolic process	P	1,66E-03	7,86E-05	129	4439	514	26286	over
	GO:0003824	catalytic activity	F	1,05E-02	5,60E-04	311	12858	332	17867	over
	GO:0019748	secondary metabolic process	P	2,60E-02	1,43E-03	20	439	623	30286	over
	GO:0006629	lipid metabolic process	P	4,23E-02	2,39E-03	56	1793	587	28932	over
	GO:0044237	cellular metabolic process	P	4,53E-02	2,63E-03	251	10338	392	20387	over
Raiz	GO:0071554	cell wall organization or biogenesis	P	1,73E-05	2,91E-07	35	521	719	30093	over
	GO:0007568	aging	P	8,21E-03	1,63E-04	10	91	744	30523	over

ANEXO II - Patente em tramitação na INOVA (Agência Unicamp de Inovação), em fase de realização dos experimentos para comprovar a funcionalidade dos promotores para depósito do pedido junto ao INPI

CAMPO DA INVENÇÂO

A invenção está relacionada ao campo de genômica funcional de plantas, biologia molecular, biotecnologia e engenharia genética vegetal, e em particular, à regulação seletiva (espacial/intensidade) e modulação da expressão de um gene em plantas. A invenção descreve novas sequências nucleotídicas isoladas de *Saccharum spontaneum* contendo regiões promotoras de tecido específicos com intensidades fortes ou moderados de expressão e os métodos para sua utilização. Os promotores descritos são úteis em cassetes de expressão e vetores de expressão para a transformação genética de mono e dicotiledôneas.

FUNDAMENTOS DA INVENÇÂO

O gênero Saccharum é composto por seis espécies: *S. spontaneum*, *S. officinarum*, *S. robustum*, *S. edule*, *S. barberi* e *S. sinense* [1]. A despeito dessas espécies não serem cultivadas comercialmente, os híbridos interespecíficos do gênero (*Saccharum spp*) e de espécies de gêneros próximos como *Erianthus*, *Ripidium*, *Miscanthus*, *Narenga*, *Sclerostachya*, dão origem a variedades comerciais cana-de-açúcar e cana energia, originadas por programas de melhoramento genético e importantes fontes de matéria prima. A cana-de-açúcar é empregada majoritariamente para produção de açúcar e etanol de 1^{o} geração (atual tecnologia), e no Brasil seus derivados constituem a principal fonte renovável da matriz primária do país [2]. Cana energia são cultivares selecionados para apresentar maior conteúdo de fibra do que sacarose em sua composição [3]. Esses cultivares são proeminentes fontes de biomassa para diversos processos, como a produção do etanol de 2^{o} geração (lignocelulósico) bem como produção de bioquímicos e energia elétrica por co-geração [4].

ANEXO III – Pedido de patente (BR 10 2016 019505 5) depositado no INPI (Instituto Nacional da Propriedade Industrial).

< Uso exclusivo do INPI >

INPI INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL PROTOCOLO GERAL
marlenegf 018160003357
24/08/2016 13:44 DESP
BR 10 2016 019505 5

Espaço reservado para o protocolo

Espaço reservado para o código QR

INPI INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL

INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL
Sistema de Gestão da Qualidade
Diretoria de Patentes

DIRPA	Tipo de Documento: Formulário	DIRPA	Página: 1/3
Título do Documento: Depósito de Pedido de Patente		Código: FQ001	Versão: 01
		Procedimento: DIRPA-PQ006	

Ao Instituto Nacional da Propriedade Industrial:
O requerente solicita a concessão de um privilégio na natureza e nas condições abaixo indicadas:

1. Depositante (71):

1.1 Nome: UNIVERSIDADE ESTADUAL DE CAMPINAS - UNICAMP

1.2 Qualificação: PESSOA JURÍDICA DE DIREITO PÚBLICO, AUTARQUIA ESTADUAL

1.3 CNPJ/CPF: 46.068.425/0001-33

1.4 Endereço Completo: CIDADE UNIVERSITÁRIA "ZEFERINO VAZ"

1.5 CEP: 13083-970

1.6 Telefone: 19 3521-5015 1.7 Fax: 19 3521-5210

1.8 E-mail: patentes@inova.unicamp.br

☒ continua em folha anexa

2. Natureza: ☒ Invenção ☐ Modelo de Utilidade ☐ Certificado de Adição

3. Título da Invenção ou Modelo de Utilidade (54):

"MÉTODO PARA A MONTAGEM DE REGIÕES DE GENES COMPLETOS DE GENOMAS COMPLEXOS, E USO DO MESMO"

☐ continua em folha anexa

4. Pedido de Divisão: **do pedido Nº** **Data de Depósito:**

5. Prioridade: ☐ Interna (66) ☐ Unionista (30)

O depositante reivindica a(s) seguinte(s):

País ou Organização do depósito	Número do depósito (se disponível)	Data de depósito

☐ continua em folha anexa

6. Inventor (72):

☐ Assinale aqui se o(s) mesmo(s) requer(em) a não divulgação de seus nome(s), neste caso não preencher os campos abaixo.

6.1 Nome: LEANDRO COSTA DO NASCIMENTO

6.2 Qualificação: BRAS,SOLT,PESQUIS.

6.3 CPF: 349.141.848-85

6.4 Endereço Completo: RUA LIMA, 130, VILA JOANA, EM JUNDIAÍ - SP

6.5 CEP: 13216-020

6.6 Telefone: 19 3521.6651 6.7 FAX:

6.8 E-mail: l.costa.nascimento@gmail.com

☒ continua em folha anexa

7. Declaração de divulgação anterior não prejudicial. ☐
Artigo 12 da LPI – periodo de graça.
Informe no item 11.13 os documentos anexados, se houver.

8. Declaração na forma do item 3.2 da Instrução Normativa nº 17/2013:

☐ Declaro que os dados fornecidos no presente formulário são idênticos ao da certidão de depósito ou documento equivalente do pedido cuja prioridade está sendo reivindicada.

9. Procurador (74):

9.1 Nome: FERNANDA LAVRAS COSTALLAT SILVADO

9.2 CNPJ/CPF: 295.166.068-57 9.3 API/OAB: 210.899

9.4 Endereço Completo: PROCURADORIA GERAL DA UNICAMP, EM CAMPINAS - SP

9.5 CEP: 13083-970

9.6 Telefone: 19 3521-4771 9.7 FAX: 19 3521-4944

9.8 E-mail: proc-geral@pg.unicamp.br

☐ continua em folha anexa

10. Listagem de sequências biológicas.
Informe nos itens 11.9 ao 11.12 os documentos anexados, se houver. ☐

INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL
Sistema de Gestão da Qualidade
Diretoria de Patentes

DIRPA	Tipo de Documento: Formulário	DIRPA	Página: 3/3
Título do Documento: Depósito de Pedido de Patente		Código: FQ001	Versão: 01
		Procedimento: DIRPA-PQ006	

11. Documentos Anexados:

(Assinale e indique também o número de folhas):
(Deverá ser indicado o número total de somente uma das vias de cada documento).

		Documentos Anexados	folhas
☒	11.1	Guia de Recolhimento da União (GRU).	1
☒	11.2	Procuração.	1
☐	11.3	Documentos de Prioridade.	
☐	11.4	Documento de contrato de trabalho.	
☒	11.5	Relatório descritivo.	25
☒	11.6	Reivindicações.	2
☒	11.7	Desenho(s) (se houver). Sugestão de figura a ser publicada com o resumo: nº, 1 por melhor representar a invenção (sujeito à avaliação do INPI).	6
☒	11.8	Resumo.	1
☐	11.9	Listagem de sequências em arquivo eletrônico: _____nº de CDs ou DVDs (original e cópia).	
☐	11.10	Código de controle alfanumérico no formato de código de barras referente às listagem de sequências.	
☐	11.11	Listagem de sequências em formato impresso.	
☐	11.12	Listagem de sequências - Declaração de acordo com a Resolução INPI nº 70/2013.	
☐	11.13	Outros (especificar)	

12. Total de folhas anexadas: 36 **fls.**

13. Declaro, sob as penas da Lei que todas as informações acima prestadas são completas e verdadeiras.

CAMPINAS, SP, EM 24.08.2016
Local e Data

Assinatura e Carimbo
FERNANDA LAVRAS COSTALLAT SILVADO
Procuradora de Universidade Subchefe
Matrícula nº 309279
OAB/SP nº 210.899

1 Continuação dos dados do depositante/interessado:

1.2 Qualificação: UNIVERSIDADE ESTADUAL DE CAMPINAS – UNICAMP, pessoa jurídica de direito público, autarquia estadual devidamente inscrita no CNPJ sob nº 46.068.425/0001-33 e isenta de inscrição estadual.

1.4 Endereço completo: Cidade Universitária "Zeferino Vaz" – Distrito de Barão Geraldo, em Campinas – SP – CEP 13083-970

6. Dados dos outros três inventores:

6.1 Nome: **MARCELO FALSARELLA CARAZZOLLE**
6.2 Qualificação: brasileiro, casado, pesquisador
6.3 CPF Nº 284.648.398-12
6.4 Endereço completo: Rua Luiz Bissoto, 240, Jd. Santa Rosa, em Valinhos - SP
6.5 CEP: 13275-110
6.6 Telefone: (19) 3521.6651/99162.1974
6.7 FAX: (19)
6.8 E-Mail: mcarazzo@lge.ibi.unicamp.br

6.1 Nome: **KARINA YANAGUI**
6.2 Qualificação: brasileira, solteira, estudante de pós graduação
6.3 CPF Nº 059.942.059-66
6.4 Endereço completo: Rua Piolim, 515, Jd. Boa Esperança, em Campinas - SP
6.5 CEP: 13091-510
6.6 Telefone: (19) 3521.6237
6.7 FAX: (19)
6.8 E-Mail: karinayanagui@lge.ibi.unicamp.br

6.1 Nome: **GONÇALO AMARANTE GUIMARÃES PEREIRA**
6.2 Qualificação: brasileiro, casado, prof. universitário
6.3 CPF Nº 289.870.395-87
6.4 Endereço completo: Rua Dr. Lauro Pimentel, 323, Cidade Universitária, em Campinas - SP
6.5 CEP: 13083-250
6.6 Telefone: (19) 3521.6237/9922.34316
6.7 FAX: (19)
6.8 E-Mail: goncalo@unicamp.br

MÉTODO PARA A MONTAGEM DE REGIÕES DE GENES COMPLETOS DE GENOMAS COMPLEXOS, E USO DO MESMO

CAMPO DA INVENÇÃO

[1] A presente invenção se insere nos campos da computação e bioinformática, em particular, à montagem de genomas de organismos complexos (poliploides e heterozigotos) e descreve um novo método para a montagem direcionada para regiões de genes, com foco na montagem de genes completos (incluindo exons, introns, UTRs e promotores), a partir de um sequenciamento de DNA de baixa cobertura (de 1x a 200x, mas preferencialmente 1x a 10x).

[2] O referido método é aplicável para a identificação de genes em qualquer genoma complexo, principalmente no caso de plantas poliploides e heterozigotas, que possuam um ou mais organismos filogeneticamente próximos com genomas completos ou parcialmente sequenciados.

FUNDAMENTOS DA INVENÇÃO

[3] A obtenção de um genoma de referência de um determinado organismo, incluindo os genes e a expressão destes genes em diversos tecidos e condições, é de extrema importância para diversas áreas de aplicações que vão desde a medicina, onde possibilita o desenvolvimento de novos medicamentos, quanto para a agricultura, onde possibilita o desenvolvimento de espécies transgênicas melhores adaptadas as condições de manejo e clima.

[4] Os sequenciadores de DNA são capazes de ler fragmentos de sequências que, na média, variam de tamanho entre 100 e 10.000 nucleotídeos dependendo do tipo de sequenciador utilizado. Genomas de organismos simples, como bactérias, possuem milhões de nucleotídeos e genomas

ANEXO IV – Declaração referente a Bioética e Biossegurança

UNICAMP

Of. CIBio/IB 35/2010

Cidade Universitária "Zeferino Vaz", 16 de agosto de 2011.

Prof. Dr. MARCELO BROCCHI
Chefe do Departamento Genética, Evolução e Bioagantes
Instituto de Biologia - UNICAMP

Prezado Professor:

Informamos que o projeto abaixo relacionado, envolvendo OGM do tipo I, sob responsabilidade do **Prof. Dr. GONÇALO A. G. PEREIRA**, protocolado sob o número **2011/03**, foi aprovado pela **CIBio-IB/UNICAMP**, em reunião sua 55ª. ordinária (15/08/2011) para ser desenvolvido nas dependências do Departamento Genética, Evolução e Bioagantes, Laboratório de Genômica e Expressão:

No. Projeto (data da aprovação)	Data de recepção	Nome do Projeto	Prazo para envio de relatório à CIBio
2011/03 (15/08/2011)	19/07/2011	Genômica e Biotecnologia, sub-projetos: 1) 2010/01 - Projeto Gene Discovery em Eucalipto; 2) Rotas Verdes para o Propeno, 3) Modificação de linhagens industriais de Saccharomyces cerevisiae para o aumento da produtividade e floculação condicional., 4) Projeto genoma de Crinipellis perniciosa, fungo causador da doença vassoura-de-bruxa do cacau, 5) Cultivo de microalgas para produção de cadeias carbônicas lipídicas, e 6) Transformação genética de cana-de-açúcar com o gene do inibidor de ripsina de inga laurina e análise da toxidade das plantas transgênicas sobre o desenvolvimento biológico de Diatraea saccharalis	Fevereiro/2012

Recomendamos que sejam observadas as instruções normativas referentes transporte e contenção da OGMs, disponíveis na webpage da CTNBio <www.ctnbio.gov.br>.

Atenciosamente,

Marcelo Lancellotti
Presidente da CIBio/IB-UNICAMP

C/C.: Prof. Dr. Gonçalo A. G. Pereira

CIbio/IB-Unicamp
Comissão Interna de Biossegurança
Instituto de Biologia - Unicamp
Caixa Postal 6109 - 13083-970 Campinas SP
Tel.: (19) 3521-6359 – e-mail: comisib@unicamp.br

ANEXO V - Termo de autorização

Profa. Dra. Rachel Meneguello
Presidente
Comissão Central de Pós-Graduação
Declaração

As cópias de artigos de minha autoria ou de minha co-autoria, já publicados ou submetidos para publicação em revistas científicas ou anais de congressos sujeitos a arbitragem, que constam da minha Dissertação/Tese de Mestrado/Doutorado, intitulada **Genômica e Transcritômica de Saccharum spontaneum e seus híbridos: novas perspectivas para compreender a formação da biomassa lignocelulósica** , não infringem os dispositivos da Lei n.° 9.610/98, nem o direito autoral de qualquer editora.

Campinas, 19 de julho de 2016

Assinatura : Karina Yanagui de Almeida
Nome do(a) autor(a): **Karina Yanagui de Almeida**
RG n.° 95614647

Assinatura :
Nome do(a) orientador(a): **Goncalo Amarante Guimarães Pereira**
RG n.° 1713878