Lucas Eduardo Costa Canesin

Identificação e caracterização de lncRNAs e genes codificadores linhagem-específicos em Andropogoneae: Padrões comuns de evolução de genes emergentes

Identification and characterization of lncRNAs and lineage-specific coding genes in Andropogoneae: Common patterns of evolution of emerging genes

Campinas

03/14

UNIVERSIDADE ESTADUAL DE CAMPINAS

INSTITUTO DE BIOLOGIA

LUCAS EDUARDO COSTA CANESIN

"Identificação e caracterização de lncRNAs e genes codificadores linhagem-específicos em Andropogoneae: Padrões comuns de evolução de genes emergentes"

"Identification and characterization of lncRNAs and lineage-specific coding genes in Andropogoneae: Common patterns of evolution of emerging genes"

Este exemplar corresponde à redação final da DISSERTAÇÃO defendida pelo candidato *LUCAS EDUARDO COSTA CANESIN* e aprovada pela Comissão Julgadora.

Orientador: Dr. Renato Vicentini Dos Santos

DISSERTAÇÃO apresentada ao Instituto de Biologia da UNICAMP para obtenção do Título de Mestre em Genética e Biologia Molecular, na área de Bioinformática.

DISSERTATION submitted to the Instituto de Biologia, UNICAMP for obtaining the title of Master in Genetics and Molecular Biology in the area of Bioinformatics

CAMPINAS,
2014

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

C162i

Canesin, Lucas Eduardo Costa, 1988-
Identificação e caracterização de lncRNAs e genes codificadores linhagem-específicos em Andropogoneae : padrões comuns de evolução de genes emergentes / Lucas Eduardo Costa Canesin. – Campinas, SP : [s.n.], 2014.

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

1. RNA longo não codificante. 2. RNA de plantas. 3. Expressão gênica. 4. Genômica. I. Vicentini, Renato. II. Universidade Estadual de Campinas. Instituto de Biologia. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Identification and characterization lncRNAs and lineage specific coding genes in Andropogoneae : common patterns of evolution of emerging genes
Palavras-chave em inglês:
RNA, long noncoding
RNA, plant
Gene expression
Genomics
Área de concentração: Bioinformática
Titulação: Mestre em Genética e Biologia Molecular
Banca examinadora:
Renato Vicentini dos Santos [Orientador]
Jörg Kobarg
Marcelo Brandão
Data de defesa: 31-03-2014
Programa de Pós-Graduação: Genética e Biologia Molecular

Campinas, **31** de **março** de **2014**

BANCA EXAMINADORA

Dr. Renato Vicentini Dos Santos (orientador) _____
Assinatura

Dr. Jörg Kobarg _____
Assinatura

Dr. Marcelo Mendes Brandão _____
Assinatura

Dr. José Andrés Yunes _____
Assinatura

Dr. Diego Riano Pachon _____
Assinatura

Resumo

Recentemente, a análise de dados de genômica comparativa, buscando elucidar melhor a hipótese nula de modelos evolutivos, i.e. evolução neutra, originou uma nova teoria que eleva o tamanho populacional como principal fator evolutivo. Populações pequenas estão sujeitas a forte influência de deriva genética, o que causa o aumento da entropia do genoma. A complexidade genômica, leia-se conteúdo de sequencias informativas, como genes, é então um subproduto do aumento da entropia e a seleção teria então um papel secundário, sobretudo como moduladora do processo evolutivo. Assumindo este modelo, a emergência e degeneração de transcritos linhagem-específicos estão submetidas primariamente a evolução neutra. A transcrição pervasiva, sobretudo em linhagens germinais, é o agente causal do nascimento de genes e a fixação destes, frente ao reduzido tamanho populacional de eucariotos multicelulares, como as plantas *Saccarum officinarum* e *Sorghum bicolor*, ocorre por deriva genética. A inserção de novos genes, que são inicialmente neutros ou levemente deletérios, em redes funcionais ainda é pouco compreendida. A integração se torna gradativamente mais robusta com a evolução individual destes *loci*. Neste contexto, este estudo buscou identificar genes codificadores e não-codificadores de proteínas de recente emergência em cana-de-açúcar e sorgo a fim de se elucidar a hipótese de que sua arquitetura gênica e integração em redes biológicas apresentam padrões evolutivos comuns. Para isso, realizamos a identificação de lncRNAs de cana a partir de bancos de cDNA, o que permitiu a caracterização da expressão desses transcritos contrastando seis variedades distintas. Em decorrência da disponibilidade do genoma de sorgo, a identificação de genes linhagem-específicos codificadores e não codificadores pode ser resolvida com maior precisão. Pudemos determinar uma correlação entre a sua arquitetura gênica e integração nas redes biológicas e sua idade relativa. Apesar da correlação encontrada, o efeito mais forte observado em transcritos não codificadores revelam outros fatores que devem estar influenciando sua evolução. Levantamos a hipótese de que o evento de tradução possa elevar a eficiência da seleção negativa sobre o transcrito emergente, o que resultaria na taxa de substituição mais acentuado de lincRNAs e maior conservação de genes linhagem-específicos.

Abstract

Recently, comparative genomics studies, aiming to better elucidate the null hypothesis of models of evolution, i. e. the neutral evolution, originate a new theory that elects the population size as the main factor acting in evolution. Small populations are subject to stronger influence of genetic drift, which raises genomic entropy. Genomic complexity, which means the information content in genome, such as genes, is a byproduct of the high entropy levels and selection would then display a secondary role, mainly as a modulator of the evolutionary process. Assuming this model, the emergence and degeneration of lineage-specific transcripts are primarily subject to neutral evolution. The pervasive transcription, especially in germinal cell lines, is the causal agent of birth of genes and their fixation, in face to the reduced population size of multicellular eukaryotes, as *Saccarum officinarum* and *Sorghum bicolor* plant species, is ruled by genetic drift. The integration of new genes, initially neutral or weakly deleterious, in functional networks is still poorly understood. The integration becomes more robust with the individual historical evolutionary path of these loci. In this context, this study aimed identify protein coding and noncoding genes of recent emergence in in sugarcane and sorghum to elucidate the hypothesis that the gene architecture and integration in biological networks display common patterns of evolution. We then identified sugarcane lncRNAs from public cDNA databases that allowed us to characterize the expression of these transcripts in six different contrasting varieties of sugarcane. As Sorghum bicolor genome is available, the identification of lineage-specific coding and noncoding could be done to a higher resolution. We could then determine a correlation between gene architecture and network integration with its relative age. Despite the correlation observed, a stronger effect seen in noncoding transcripts reveals other factors that may be influencing their evolution. We propose the hypothesis that the translation event may increase negative selection efficiency over the emerging transcript, what would result in the stronger replacement of lincRNAs and higher conservation levels of coding lineage-specific genes.

Sumário

Agradecimentos

Agradeço acima de tudo e pela última vez a Deus, a mentira primeira, causa maior de minha paixão à ciência, destruída no processo de descoberta de minha completa ignorância acerca das verdades do mundo.

Agradeço a orientação e a amizade sempre presente do Prof. Dr. Renato Vicentini. O meu crescimento acadêmico seria impossível de outra forma.

Agradeço a Luiz Edurado Del Bem, pela co-orientação e amizade. As colocações sempre precisas foram de grande valia ao desenvolvimento do presente estudo e assim será em iniciativas futuras.

Agradeço a Michel Vincentz, pela colaboração inestimável. As discussões promovidas por ele são alguns dos momentos mais interessantes da minha formação.

Agradeço aos companheiros de Laboratório, que se mostraram sempre dispostos a discussão e contribuíram de forma especial ao bom andamento da dissertação.

Agradeço a meus amigos Diogo, Karina e Priscila. Nosso convívio foi e será sempre um aprendizado enriquecedor. Me sinto feliz por sentir que estas são amizades que persistirão ao espaço-tempo.

Agradeço a minha família, que a despeito de minha omissão em inúmeras ocasiões, sempre me demonstrou amor e paciência.

Por fim, agradeço a FAPESP, pelo apoio técnico e financeiro, imprescindível a conclusão do presente estudo.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

ABA	Ácido Abscísico
BAM	Binary Sequence Alignment/Map
CNC (Rede)	Coding-Noncoding (rede)
EST	Expressed Sequence Tags
FPKM	Fragments Per Kilobase Of Exon Per Million Fragments Mapped
G6P	Glicose-6-fosfato
GEO	Gene Expression Omnibus
GFF	General Feature Format
GO	Gene Ontology
HCCA	Heuristic Cluster Chiseling Algorithm
HRR	Highest Reciprocal Ranking
HXK	Hexoquinase
lincRNA	Long Intergenic Noncoding RNA
lncRNA	Long Noncoding RNA
LSG	Lineage-Specific Gene
MFE	Minimal Free Energy
MS (meio de cultura)	Meio Murashige & Skoog
ncRNA	Noncoding RNA
NGS QC	Next Generation Sequencing Quality Control
OLB	Off-Line Base Caller
ORF	Open Reading Frame
PEG	Polyethylene Glycol
RMT	Random Matrix Theory
RSEM	RNA-Seq by Expectation Maximization
SAM	Sequence Alignment/Map
T6P	Trealose-6-fosfato
TE	Transposable Elements
UTR	Unstranslated Regions

Introdução

Emergência de novos transcritos

Muito se argumentou que inovações genômicas seriam geradas apenas por recombinação dos domínios proteicos já descritos. Como exposto por Jacob (1977), não seriam requeridos novos elementos genéticos em face ao número extensivo de domínios proteicos até então identificados. Supreendentemente, estudos de genômica comparativa revelaram um número significativo de genes codificadores linhagem-específicos nos genomas já sequenciados (Wolf *et al*., 2009; Donoghue *et al*., 2011; Tautz e Domazet-Lošo, 2011; Gibson *et al*., 2013). Não obstante, o sequenciamento de transcriptomas revelou diversas novas unidades de transcrição emergindo de todo o genoma de diversas espécies. Ademais, transcrição pervasiva ocorre em mais de 85% de todo o conteúdo genômico humano (Hangauer *et al*., 2013). Uma proporção considerável dessas novas unidades de transcrição consiste de RNAs não codificadores longos intergênicos (*lincRNAs – long intergenic noncoding RNAs*), que são em sua maioria genes linhagem-específicos (Managadze *et al*., 2013), embora a análise de conservação baseada em similaridade de sequência possa enviesar o número de genes não conservados. Estas novas espécies de RNA indicam que o ruído transcricional desempenha um papel importante na geração de novos elementos genéticos e a própria evolução do genoma.

O modelo mais simples para explicar a emergência de novos transcritos poliadenilados de plantas, e portanto transcritos por *Pol*II, pode ser descrito em termos da deriva genética e consiste no surgimento de um promotor basal, uma sequência de 4-10nt de

extensão (Smale e Kadonaga, 2003), cuja função se dá fundamentalmente devido a propriedades mecânicas estruturais, em detrimento de um mecanismo baseado exclusivamente no reconhecimento de um sítio de sequência específica (Fukue *et al*, 2005), e obviamente um sítio de poliadenilação. A região genômica compreendida entre estes sítios produzirá um transcrito emergente. A ocorrência de ORFs (*open reading frames*) dentro destes dois sítios supracitados irá definir o caminho evolutivo do transcrito, potencialmente como codificador de um novo peptídeo.

Diferentes elementos poderão integrar a arquitetura do gene emergente ao longo da história evolutiva deste. Introns e UTRs, estas apenas para genes codificadores, representam um novo espaço mutacional que pode originar novos elementos regulatórios, em especial aqueles com ação em *cis*. Introns podem ainda conferir modularidade ao novo gene. O aumento da complexidade da região promotora pela emergência ou inserção de sítios de ligação de fatores de transcrição irá refletir em uma regulação mais elaborada do transcrito (Kim *et al*, 2009), que inicialmente apresentará forte viés de expressão tecido-específica e então evolui para um padrão de expressão generalizada. Elementos transponíveis desempenham um papel fundamental nesta dinâmica (Kapusta *et al*., 2013), uma vez que sua inserção no genoma pode carregar consigo elementos responsáveis pelo aumento da complexidade da arquitetura do transcrito emergente, especialmente em regiões sujeitas a baixa ou nenhuma seleção negativa, como as regiões intergênicas.

A manutenção desta arquitetura elaborada é diretamente dirigida por seleção purificadora. Mas o nascimento e morte de novos transcritos devem ser governados por evolução neutra, uma vez que apenas níveis muito baixos de seleção purificadora foram detectados nestes loci (Managadze *et al*., 2011). A

despeito das diferenças de intensidade de seleção sobre lincRNAs e genes codificadores de proteínas, foi demonstrado que genes codificadores linhagem-específicos (portanto os genes de emergência mais recente) apresentam também baixa intensidade de seleção purificadora quando comparados aos genes codificadores mais antigos. Considerando o genoma eucariótico como um sistema de elevada entropia (Koonin, 2004), que tipicamente apresenta tamanho populacional pequeno e que então é sujeito a forte efeito de deriva genética (Lynch e Conery, 2003; Lynch, 2007), transcritos emergentes devem ser sujeitos exclusivamente a evolução neutra até sua integração em redes metabólicas, quando seleção poderá eficientemente agir sobre estes novos genes. Isso aponta para uma dinâmica evolutiva comum a genes de emergência recente. Esse processo é governado por padrões emergentes do sistema genômico e representa o aumento de complexidade em decorrência do aumento da entropia no mesmo.

Padrões comuns de evolução

A quebra do paradigma adaptacionista, iniciada por Motoo Kimura e desenvolvido posteriormente por sua então aluna Tomoko Ohta (Kimura, 1968; Ohta, 1973) ainda é subestimada por inúmeros pesquisadores de renome. Isso se dá em grande parte pela incapacidade de aceitação da alteração do *status quo* a que estão acostumados, sobretudo após a publicação por Lynch e Conery de artigo seminal no qual avançam a fronteira da evolução neutra sob uma perspectiva de genômica comparativa e genética de populações. Análises subsequentes têm corroborado seu novo modelo evolutivo, demonstrando que até a multimerização proteica, tida como exemplo pétreo de evolução adaptativa, ocorre por processos estocásticos, i.

e., evolução neutra. Assim, da mesma forma que Dobzhansky (1964) afirma que nada em Biologia faz sentido senão à luz da evolução, Michael Lynch afirma que nada faz sentido em evolução senão a luz de genética de populações.

Em acordo com a teoria entrópica da complexidade genômica (nome proposto por Koonin para os estudos de Lynch e Conery), foi observado que o aumento da complexidade genômica estava relacionado não com processos progressivos de seleção natural que culminaram em redes altamente elaboradas, associadas ao decorrente aumento de *fitness*, mas sim ao tamanho populacional reduzido de eucariotos multicelulares complexos, que chega a ser três ordens de grandeza menor que procariotos. Dessa forma, a complexidade emerge não como fruto de progresso selecionista, mas como uma síndrome genômica, em que o genoma é incapaz de constringir sua inflação pela baixa eficiência da seleção negativa em expurgar eventos de expansão genômica, como eventos de transposição, inserção de DNA exógeno ou duplicações em diversas escalas no genoma. Eugene Koonin observou em analises comparando procariotos, eucariotos unicelulares e eucariotos multicelulares que determinados padrões emergiam em decorrência do sistema genômico. Desta forma, quatro universalidades foram descobertas (Figura 1, adaptada de Koonin, 2011), e nenhuma destas requer que se invoque mecanismos adaptacionistas, sendo propriedades emergentes de sistemas compostos por numerosos elementos que interagem fracamente entre si, em contraste com as fortes interações que mantém sua integridade.

Considerando as universalidades apresentadas acima, o conhecimento sobre os mecanismos de nascimento e morte de genes e principalmente sobre a inclusão destes genes em redes biológicas é de crucial importância para a compreensão dos fatores que

governam estas propriedades emergentes do sistema genômico. Diversos trabalhos tem reportado o surgimento *de novo* de genes codantes a partir de sequências intergênicas em diferentes organismos (Cai et al., 2008; Capra et al., 2010; Carvunis et al., 2012; Knowles e McLysaght, 2009; Zhou et al., 2008). Estes genes novos sofrem pressões evolutivas diferentes das que atuam sobre genes filogeneticamente mais antigos, uma vez que, inicialmente, tem alta probabilidade de não serem funcionais, trazendo nenhum incremento ao fitness ou até mesmo sendo levemente deletérios, devido ao possível gasto energético. Capra et al. (2010) observaram que existe uma forte relação entre a origem de um gene e sua integração com as redes funcionais da célula, sugerindo que genes de emergência recente são co-expressos. Essa relação ainda precisa determinada entre genes emergentes codificadores e não codificadores.

Estas classes de transcritos apresentam características comuns quanto a sua arquitetura gênica e dinâmica genômica. Portanto, sua emergência possivelmente está correlacionada, embora as forças que regem este processo de emergência não são totalmente compreendidas.

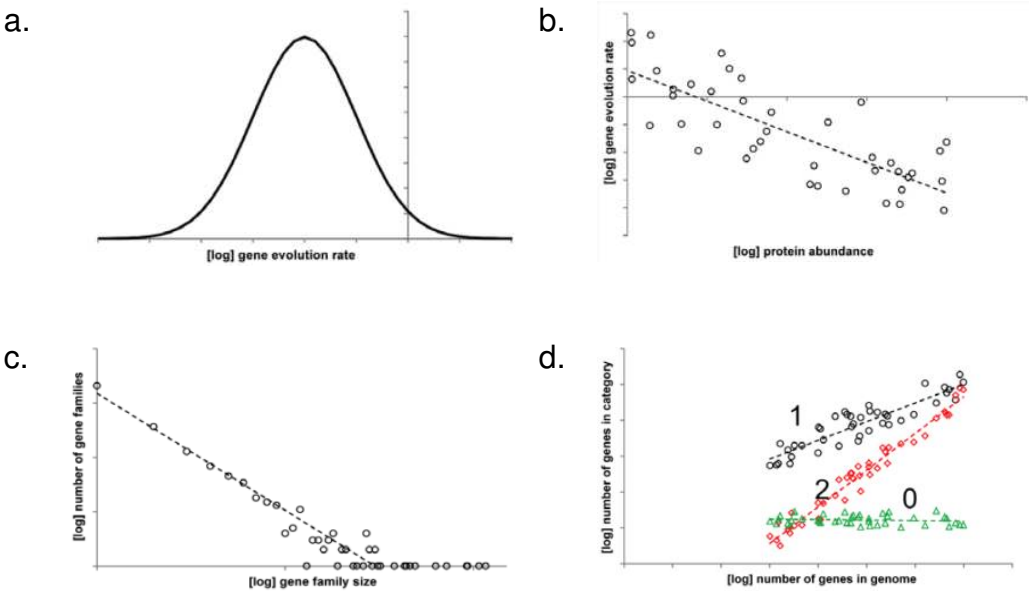


Figura 1. Universalidades genômicas e evolução molecular do fenoma. a. Distribuição log-normal das taxas evolutivas de genes ortólogos (Wolf et al., 2009); b. Anticorrelação entre nível de expressão e a taxa evolutiva de um gene, de forma que genes com acentuada taxa evolutiva são menos expressos e genes que evoluem mais lentamente são mais expressos (Huynen e Nimwegen, 1998; Koonin et al., 2002); c. Distribuição do tipo *power law* do tamanho de famílias de parálogos (Pal et al., 2001; Drummond et al., 2006); e d. Aumento de genes de classes funcionais com o aumento do total de genes (0 – sem dependência, comum a genes *housekeeping*, como elementos do aparato traducional; 1 – dependência linear, característico de enzimas metabólicas; e 2 – dependência quadrática, característico de elementos constituintes de vias de sinalização e transdução de sinal) (Nimwegen, 2003; Molina e Nimwegen, 2009).

Genes codificadores linhagem-específicos (LSG)

LSGs são identificados como genes específicos a determinados grupos taxonômicos, como famílias, espécies ou mesmo para pequenas populações de uma dada espécie. Estes genes são transcritos e traduzidos em taxas mais baixas do que genes conservados com linhagens mais antigas, o que sugere que estejam pouco integrados nas redes metabólicas ou que baixos níveis de expressão sejam suficientes para sua ação. Em *A. thaliana*, estes genes são responsivos a estresses abióticos e bióticos (ver Rutter *et al*., 2012). Trabalhos recentes demonstraram que estes genes são relativamente mais curtos do que genes mais antigos. Sua origem pode estar relacionada a uma sequência duplicada que tenha divergido enormemente das sequências ancestrais que os originaram, ou que sejam sequências novas, que emergiram *de novo* por deriva genética (Wolf *et al*., 2009). Podem ser ainda compostos por pequenas porções de outros genes, que incluem exons, regiões intrônicas e intergênicas e elementos repetitivos, formando estruturas quiméricas ao longo de sua evolução.

Recentemente, foi relatado por (Zhou *et al*., 2008) que 11,9% dos genes se originaram *de novo* de sequências não-codantes no genoma de *Drosophila melanogaster*. Alguns destes genes foram originados por rearranjos linhagem-específicos do genoma ancestral do complexo Drosophila, embora a sintenia não seja mantida entre as espécies que apresentam homólogos. LSGs são transcritos antes mesmo de poderem codificar proteínas (Cai *et al*., 2008; Zhou *et al*., 2008). Silveira *et al*. (2013) identificaram um LSG de *A. thaliana* que apresenta variação epigenética natural. Foi identificado em Poaceae um conjunto de genes de emergência recente conservados apenas dentro deste grupo, que não possuem funções conhecidas e

apresentam alta taxa mutacional. Estes genes apresentam alto conteúdo GC, são mais curtos e são relacionados com elementos *Pack-Mutator-Like* (Campbell *et al*., 2007). Apresentam ainda organização em *clusters* e sequências adjacentes altamente similares. Apesar de haver alta conservação da sequência destes genes em Poaceae, cada espécie alterou o padrão de expressão destes genes. Ao se comparar a sintenia entre arroz e sorgo, a ordem sintênica é mantida bem como a orientação transcricional. Isso pode ser verificado também em todo o grupo Poaceae.

Genes de RNAs não-codificadores longos intergênicos (lincRNAs)

A descoberta de uma parcela significativa de RNAs não-codantes através do sequenciamento de transcriptomas quebrou o paradigma de proteínas como únicos efetores nas redes biológicas. RNAs não-codantes longos intergênicos constituem um grupo distinto dos RNAs não-codantes *housekeeping*, como snRNAs e tRNAs, embora possivelmente consista ainda em um grupo heterogêneo. Estes transcritos com arquitetura gênica complexa são caracterizados como longos (>200nt), apresentam em geral *ORFs* muito pequenas (menores do que 100aa) e/ou com baixa probabilidade de serem codificadoras e, portanto, não funcionais, ou sequer contêm alguma ORF e, similar ao que se observa em genes codificadores, muitos destes transcritos apresentam marcas epigenéticas (Khalil et al., 2009). Diversos exemplos já descritos apresentam estruturas secundárias altamente estáveis, o que deve estar diretamente relacionado a sua função (Kassube et al., 2013), como ribozimas. Também não apresentam identidade com proteínas ou elementos transponíveis.

Entretanto, a maioria dos esforços realizados sobre lincRNAs focaram primariamente na descrição abrangente de catálogos desses genes, e suas funções são em sua maioria desconhecidas (ver Ulitsky e Bartel, 2013). Apesar da maioria dos avanços acerca da função de lincRNAs ter sido conduzida em modelos de mamíferos, mecanismos similares foram identificados em plantas (Zhu e Wang, 2012). Estes mecanismos incluem a inibição da atividade de miRNAs (Franco-Zorrilla et al., 2007), em que o lncRNA IPS1 sequestra miRNAs ao mimetizar o alvo destes; modificação de estados da cromatina, e então regulação epigenética (Heo et al., 2013), que envolve o recrutamento de complexos proteicos que atuam sobre a cromatina em sítios específicos; e relocalização de proteínas (Bardou et al., 2011; Campalans et al., 2004). A despeito da diversidade de funções já identificadas, estes mecanismos claramente mostram um papel na regulação gênica fina para lincRNAs.

O maior esforço de caracterização de lincRNAs em plantas foi conduzido em *Arabidopsis thaliana*, no qual foram combinadas diferentes plataformas de análise de expressão de todo o genoma (Liu et al., 2012). Foram identificados 6.500 lincRNAs, apesar de mais de 6000 potenciais lincRNAs terem sido removidos da análise por estarem envolvidos com elementos repetitivos , sendo que Kapusta et al., (2013) demonstraram que a emergência de lincRNAs está relacionada com eventos ancestrais de transposição. Este conjunto demonstrou forte viés de expressão tecido e *stress* específica e níveis muito baixos de conservação. Também foi possível determinar o envolvimento conjunto dos genes SE, CBP20 e CBP80 na biogênese e processamento dos lincRNAs. Em outro esforço de identificação de *ncRNAs* em *Zea mays*, utilizando bibliotecas de ESTs, identificou ~650 lincRNAs (Boerner & McGinnis, 2012). Como a maioria de ncRNAs identificados consistia em lncRNAs envolvidos

na biogênese de sRNAs, este resultado mostra possível viés deste tipo de banco de dados.

Para genes codificadores de proteínas, observa-se que genes funcionalmente importantes apresentariam níveis de conservação evolutiva proporcionais, em decorrência da ação de seleção purificadora sobre esses loci. Porém, mesmo lincRNAs conservados em linhagens distantes como o *Xist* (Duret et al., 2006), em mamíferos, apresentam elevada taxa de mutação, especialmente quando comparados com genes codificadores com a mesma distribuição taxonômica. Não obstante, Managadze et al. (2011) demonstraram que lincRNAs estão sujeitos a menores pressões de seleção purificadora do que genes codificadores. Desse modo, isso se reflete no padrão linhagem-específico característico de *loci* de lincRNAs e sugere elevada taxa de substituição destes transcritos. Ainda estão em aberto questões fundamentais acerca do regime evolutivo de lincRNAs, as quais poderão ser sanadas quando da identificação de catálogos consistentes desses genes e investigação profunda de mecanismos de ação.

Resposta gênica a sinalização por açúcares e ABA

Os açúcares constituem a fonte primária de energia e fornecem *backbones* de carbono às células vegetais. Por conseguinte, ocorreu ao longo do curso evolutivo a cooptação deste nutriente pelo organismo como modulador de redes metabólicas. Isso se reflete no papel de destaque da glicose desempenhando essa função regulatória (Forde, 2002; Gutiérrez et al., 2007; Rolland et al., 2006; Rook et al., 2006). A sacarose é o principal produto da fotossíntese e é degrada em passos subsequentes a sua mobilização

pelos vasos condutores das plantas. Um dos produtos de sua degradação é a própria glicose, o que revela a integração da via fotossintética e os processos de sinalização glicose-dependentes, envolvidos com diversos aspectos do desenvolvimento e crescimento vegetal. O balanço da disponibilidade de açúcares realizado ativamente pelas plantas é de fundamental importância para manutenção das redes biológicas. Mecanismos diferentes são responsáveis pelo metabolismo de açúcares nos tecidos fonte e dreno e em resposta a condições ambientais desfavoráveis, bióticas ou abióticas. Tecidos fonte produzem trioses-fosfato pelo ciclo de Calvin. Estes metabólitos intermediários podem ser temporariamente imobilizados na forma de amido no próprio plastídio; remobilizados para tecidos dreno, atuando também como importante *input* de sinalização nestes tecidos, durante o período escuro, ou pode ser exportado para o citosol. No segundo caso, pode ainda ser utilizado na via glicolítica ou ser convertido a sacarose, a ser estocada no vacúolo ou mobilizada para tecidos dreno. O açúcar mobilizado é capturado pelo tecido dreno por transporte simplástico (via plasmodesmas) ou transporte apoplástico. Invertases então degradam a sacarose a glicose ou glicose-6P. Organismos expostos a situação de privação energética podem ainda degradar celulose constituinte da parede celular como fonte primária de açúcares. Neste contexto, hexoquinases (HXKs) apresentam papel fundamental de sensores conservados evolutivamente de glicose (Figura 2, adaptada de Rolland et al., 2006). Sua função regulatória é desacoplada da função catabólica. Proteínas transmembrana também participam da via de sinalização como receptores sensíveis ao nível extracelular de açúcares (Figura 2c).

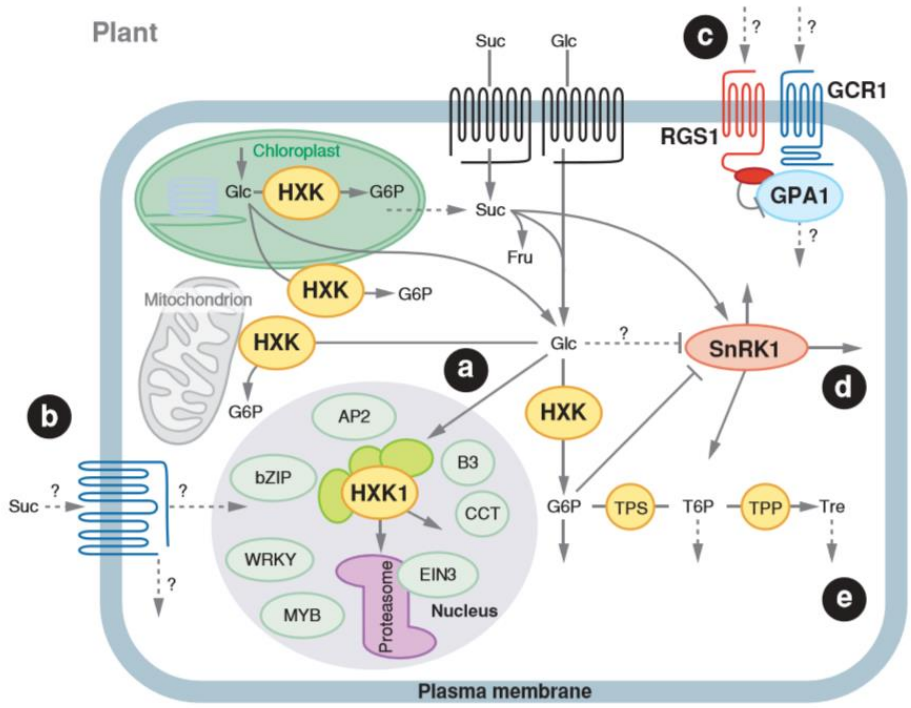


Figura 2. Modelo de mecanismos de detecção de açúcares em plantas. a. O sensor de glicose HXK1 é identificado sobretudo na mitocôndria embora possa integrar complexos de alto peso molecular no núcleo, onde regula o fator de transcrição EIN3. HXK pode ainda ser encontrada em plastídeos ou no citosol. b. Sacarose e outros dissacarídeos são detectados na membrana plasmática, possivelmente por transportadores de uma mesma família gênica. c. Sinalização via proteína G (RG1 e GPA1) está envolvida na transdução do sinal de glicose. d. Proteínas SnRK1 desempenham um papel importante na sinalização por açúcares e por inanição, embora a importância da regulação dessas proteínas por sacarose (*Suc*) e G6P ainda é pouco conhecida. e. Trealose e T6P apresentam importantes papéis na regulação, subsequentes a SnRK1.

Vias de sinalização por açúcares permeiam todos os estágios de desenvolvimento da planta, incluindo germinação de sementes, crescimento vegetativo, diferenciação de tecidos

reprodutivos e senescência. A regulação do metabolismo de açúcares é portanto altamente elaborada, sobretudo devido a interação de sinais endógenos e exógenos característicos de um organismo séssil e produtor de açúcares, o que leva a integração destas vias às vias de sinalização hormonal desses organismos, como o ABA (Figura 3, adaptada de Rolland et al., 2006). O *crosstalk* entre vias de sinalização de glicose e ABA foi demonstrado por diversos estudos, em especial acerca dos fatores de transcrição do tipo *bZIP*, o que revela a importância de regulação a nível transcricional para vias sinérgicas de glicose e ABA. Genes de quinases e de fatores de transcrição identificados como intermediários na via de sinalização sinérgica de açúcares e ABA em cana-de-açúcar apresentam resposta conservada com *Arabidopsis thaliana* (Papini-Terzi et al., 2009), o que sugere que o *crosstalk* evoluiu anteriormente à divergência dessas duas linhagens.

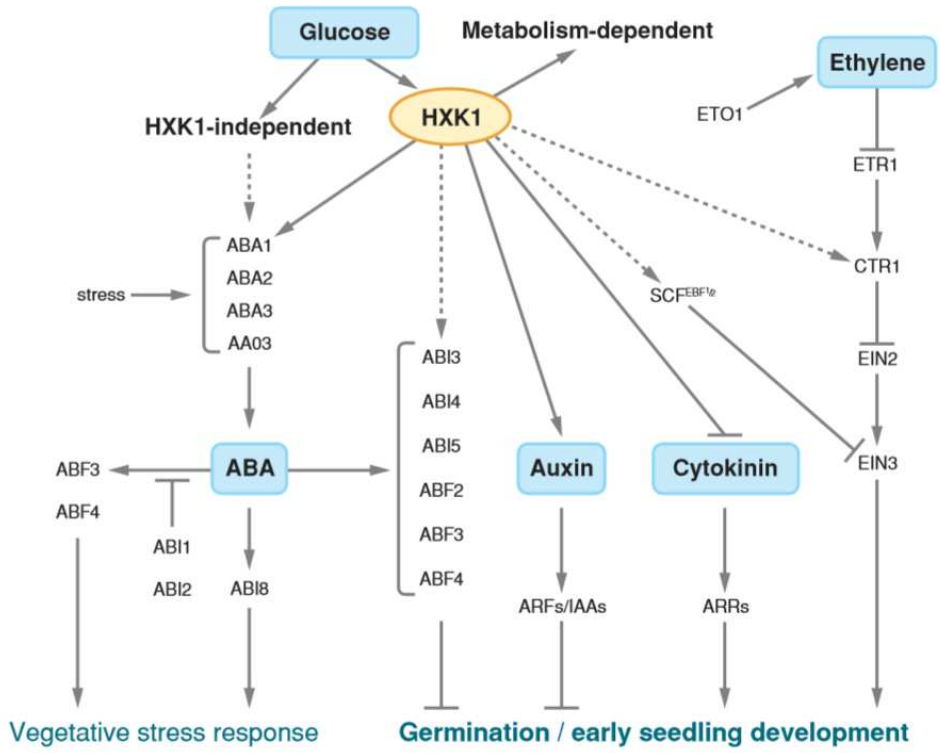


Figura 3. Modelo de interações genéticas entre sinalização por açúcares e fitormônios. Sinalização por glicose mediada por HXK1 envolve um incremento de ABA e induz a amplificação do sinal pelo fitormônio. A vias de sinalização por glicose antagoniza, via HXK, a via de sinalização por etileno, ao promover a degradação do fator de transcrição EIN3 (Yanagisawa et al., 2003). A sinalização por HXK interage positivamente e negativamente com vias de sinalização por auxina e por citocinina, respectivamente. Assim, o metabólito glicose apresenta comportamento similar ao de fitormônios, em plantas fotossintetizantes, o que revela seu papel central na modulação do desenvolvimento vegetal.

Redes moleculares de interação e resposta

A formação de extensos bancos de dados com informações sobre diversas formas de interação entre os componentes celulares possibilita a criação de modelos computacionais holísticos, que permitem que sejam observados padrões emergentes e o próprio comportamento do sistema em inúmeras situações. Essas redes biológicas constituem o cerne da Biologia de Sistemas e a análise de suas propriedades, como as interconexões formadas e o efeito de perturbações no modelo, possibilita a compreensão de processos importantes de um dado organismo (Ideker et al., 2001a; Ideker et al., 2001b).

Dentre os diversos tipos de abordagens sistêmicas disponíveis para análise de redes, a rede de regulação CNC (Coding-Noncoding) é o método mais utilizado para identificar circuitos celulares (Luo et al., 2007; Zhang & Horvath, 2005). Este método permite a identificação de módulos gênicos pertencentes às redes analisadas, uma vez que genes de um mesmo módulo são altamente correlacionados. Foi relatado por Luo et al. (2007) que a abordagem baseada na Teoria de Matriz Randômica (RMT – Random Matrix Theory) é eficiente na identificação de redes de co-expressão gênica, uma vez que permitem a distinção entre propriedades não-randômicas, específicas ao sistema, do ruído aleatório.

Liao et al. (2011) realizou a predição funcional em larga escala de lncRNAs através da construção de redes de co-expressão CNC de alta qualidade e precisão. A caracterização funcional foi obtida pela utilização de três métodos (co-expressão e co-localização, método *Hub-based* e análise de módulos de rede), que se mostraram coerentes e complementares, aumentando a confiabilidade da predição. Ainda há muitas questões importantes, mas sem respostas,

sobre o papel regulatório dos RNAs não codantes, assim como sobre o surgimento e função dos genes linhagem-específicos.

O conhecimento sobre os mecanismos de emergência destes genes e principalmente sobre a inclusão destes genes em redes biológicas é de crucial importância para a compreensão do desenvolvimento vegetal, como, por exemplo, a regulação do metabolismo de sacarose em cana-de-açúcar.

O genoma de sorgo e o transcriptoma de cana

Nos anos subsequentes a publicação do genoma de *A. thaliana*, inúmeros projetos de sequenciamento de genomas de Viridiplantae foram executados, como os de *Oryza sativa* ssp. *japonica* e ssp. *indica, Glycine max* e *Vitis vinifera* (Goff et al., 2002; Jaillon et al., 2007; Schmutz et al., 2010; Yu et al., 2002), motivados por interesses agronômicos ou por serem plantas-modelo, como *Solanum lycopersicum* (Tomato Consortium, 2012), muito utilizado como modelo para o desenvolvimento de frutos climatéricos, e sorgo (*S. bicolor*) (Paterson et al., 2009), que representa um modelo para o desenvolvimento de gramíneas. Neste contexto, ainda está em execução o sequenciamento do genoma da cana-de-açúcar (*S. officinarum*, www.sugarcanegenome.org), que tem grande destaque na produção de biocombustíveis, uma vez que o principal produto metabólico explorado na cultura ainda é a sacarose, utilizada para a produção de etanol, embora também seja apropriada para produção de biomassa e, assim, constitui um modelo também para a produção de etanol de segunda geração. O maior desafio no sequenciamento da cana-de-açúcar é a complexidade estrutural que seu genoma apresenta, sendo poliplóide e aneuplóide, apresentando geralmente 8

a 10 homólogos, além do tamanho do genoma haplóide (~1Gb) (D'Hont et al., 1996).

Com genoma diplóide, de arquitetura relativamente simples, *S. bicolor* se apresenta como modelo ideal para análise genômica de gramíneas, em especial plantas C4, como a cana-de-açúcar. A publicação de seu genoma, a segunda gramínea a ser sequenciada, se configurou como acréscimo significativo na compreensão da arquitetura genômica do grupo mais recente dentro de Viridiplantae. Segundo análise realizada por Paterson et al. (2009), o genoma de ~730Mb apresenta aproximadamente 62% de sua totalidade na forma de heterocromatina e a expansão de seu genoma desde a divergência entre *O. sativa* e *S. bicolor* é principalmente devida a atividade de retrotransposons LTR (família '*gypsie*-like' – 55%). A distribuição de elementos transponíveis através do genoma demonstra preponderância em regiões pobres em genes codificadores, possivelmente devido a efeitos colaterais de seleção sobre estes loci. Transposons representam apenas ~7.5% do genoma, sendo a família 'CACTA-like' a mais frequente (5% do genoma). Muitos membros desta família são resquícios não autônomos dos elementos originais e carregam fragmentos de genes não relacionados a transposição em substituição as sequências contendo genes relacionados a transposição. De 34496 modelos gênicos, apenas 27640 foram determinados como genes codificadores genuínos. Destes, 57,8% apresentam blocos sintênicos conservados com *O. sativa*. Outros 5197 modelos gênicos apresentam tamanho reduzido (<150aa), são mais divergentes que genes de arroz e possuem funções desconhecidas.

Visando a manipulação genética mais precisa de plantas como plantas-modelos e sobretudo de cultivares, o entendimento da regulação fina dos genomas e a evolução dos genomas é de suma

importância. Assim, existe grande interesse em novas classes de moléculas como lncRNAs e genes linhagem-específicos, que são pouco conhecidos, mas desempenham papéis importantes em diversos estágios do desenvolvimento vegetal.

Objetivo geral

Avaliar os padrões de arquitetura de RNAs de emergência recente, codificadores e não codificadores, e as interações de expressão gênica existentes entre estes transcritos, responsivos a vias específicas de sinalização.

Objetivos específicos

- Identificar e caracterizar lncRNAs de cana-de-açúcar que apresentem evidências de importância biológica, a partir de banco público de ESTs;
- Identificar e caracterizar o conjunto de genes de lincRNAs de sorgo quanto a sua arquitetura, estrutura secundária, regulação por sRNAs e resposta a vias de sinalização por açúcares e ABA e a *stress* hídrico;
- Identificar e caracterizar o conjunto de genes linhagem-específicos de sorgo quanto a sua arquitetura e resposta a vias de sinalização por açúcares e ABA e a *stress* hídrico;
- Avaliar a inserção de LSGs e lincRNAs em redes de regulação gênica;
- Determinar padrões de evolução comuns a genes de emergência recente.

Capítulo 1 – Identificação de lncRNAs de cana-de-açúcar em bancos de dados de EST

Introdução

A cana-de-açúcar representa um mercado em franca expansão. A exploração da sacarose obtida de seu processo ainda é o principal metabólito usado para obtenção de etanol, embora novas variedades estejam sendo desenvolvidas almejando a exploração do etanol de segunda geração, obtido a partir da biomassa da planta. A complexidade de seu genoma representa uma barreira aos programas de melhoramento assistido. O maior desafio no sequenciamento do genoma da cana-de-açúcar é a complexidade estrutural que seu genoma apresenta, sendo poliplóide e aneuplóide, apresentando geralmente 8 a 10 cópias homólogas, além do tamanho do genoma haplóide (1Gb) (D'Hont et al, 1996), constituído em grande parte por sequencias repetitivas. Como ainda carecemos da sequência do genoma, uma alternativa explorada por grupos de pesquisa é a análise do transcriptoma de *S. officinarum*, inicialmente utilizando-se de tecnologias de ESTs e mais recentemente através de sequenciamento de larga escala.

Uma vez que a maioria de RNAs não codificadores longos foram descobertos recentemente e carecem majoritariamente de estudos funcionais, os elementos funcionais envolvidos nas vias metabólicas de interesse melhor descritos e compreendidos são genes codificadores de proteínas. Esforços de identificação do transcriptoma codificador indicam que genoma haplóide deve conter 40-45 mil genes (Cardoso-Silva et al., 2014). A identificação precisa destes genes permite que se encontre alvos de estudos funcionais que conduzam a elucidação das redes metabólicas de cana. Porém, a importância de ncRNAs é aceita como fundamental a diversos processos biológicos. Os miRNAs, por exemplo, são elementos envolvidos na regulação a nível pós-transcricional que regulam

finamente uma gama expressiva de genes e culminam na modulação de tais redes metabólicas. Não obstante, estudos recentes demonstraram que RNAs longos não codificadores, uma classe identificada principalmente em decorrência do advento do sequenciamento em larga escala, podem apresentar papel funcional importante em processos metabólicos fundamentais ao desenvolvimento da planta, sobretudo como intermediários em vias de regulação.

Dessa forma, buscamos identificar e caracterizar lncRNAs de cana que apresentem indícios de importância biológica, disponíveis em bancos públicos. Estes transcritos são os principais candidatos a estudos funcionais subsequentes.

Material e Métodos

Mineração de lncRNAs

Desenvolvemos uma estratégia *in silico* para minerar lncRNAs a partir de bancos de dados de EST cDNA. O conjunto de 121.342 unigenes de *Saccharum officinarum*, disponível *online* pelo projeto *Gene Index* (http://compbio.dfci.harvard.edu/tgi/), foi alinhado a um banco local composto por todas as proteínas preditas de Viridiplantae, disponíveis no banco de dados Phytozome (http://www.phytozome.net/), e com o banco nr, do banco de dados GenBank (http://www.ncbi.nlm.nih.gov/genbank/ - Benson et al., 2013) (BLASTx, e-value < 1e-5). Sequências que não alinharam a nenhuma proteína foram então mapeadas no genoma de *Sorghum bicolor* (Phytozome, v1.4) com o *software* sim4 (Ogasawara e Morishita,

2003), exigindo 70% de cobertura e permitindo introns de até 15kb. Identificamos possíveis genes codificadores não anotados através da determinação de ORFs maiores do que 100aa, utilizando o *software* ESTScan (Iseli et al., 1999). Sequências que não apresentam identidade com proteínas preditas, mas são mapeadas no genoma de sorgo e apresentam nenhuma ORF ou ORFs pequenas constituem o conjunto de lncRNAs de cana-de-açúcar representados no projeto *Gene Index*. Em seguida, removemos transcritos menores do que 200bp, critério utilizado para remover sequências de *housekeeping* ncRNAs. Ademais, retiramos sequências de RNAs ribossomais e pri-miRNA (identificados com BLASTn com *e-value* ≤ 1e-5 contra banco de dados de rRNAs de plantas e contra as sequências de pri-miRNAs de gramíneas disponíveis no banco de dados MiRBase). Por fim, retiramos sequências mascaradas com o *software* RepeatMasker, usando parâmetros *default*, alimentado com os bancos de referência de *repeats* Repbase v18.08 (Jurka e Kapitonov, 2005) e *TIGR Plant Repeat Database* (Ouyang & Buell, 2004).

Nós utilizamos então três fontes de evidência de funcionalidade biológica para essas sequências: I. presença de estrutura secundária estável, determinada através do *folding* das sequências alvo (RNAfold, *Vienna package*), normalizadas pelo índice de *Z-score*, como descrito por Clote et al. (2005); II. mapeamento de sRNAs; e III. similaridade com sequências de EST de *S. bicolor* (BLASTn, e-value < 1e-5).

O conjunto de lncRNAs identificado foi alinhado contra transcritos montados por Cardoso-Silva et al. (2014), conjunto de potenciais RNAs não-codificadores. Realizamos então a quantificação da expressão de genes de lncRNAs representados na amostragem por RNA-seq em diferentes genótipos de cana-de-açúcar, como descrito abaixo.

Material vegetal e extração de RNA

Seis genótipos foram incluídos neste estudo. IACSP96-3046 e IACSP95-3018 são os parentais de uma população de mapeamento do Programa de Melhoramento de Cana-de-açúcar do IAC/Apta. IACSP95-3018 é um clone promissor que também foi usado como parental no programa de melhoramento. IACSP93-3046 é uma variedade que exibe bom perfilhamento (Mancini et al., 2012) e resistência a ferrugem (Landell et al., 2005). SP81-32506RB925345 e SP80-32806 e RB835486 são os parentais de duas populações de mapeamento do Programa de Melhoramento de Cana-de-açúcar da UFSCar, o qual compõe o RIDESA.

Folhas na terceira posição (McCormick et al., 2006) foram coletadas de uma planta por genótipo e congelada imediatamente. A extração de RNA total foi realizada utilizando um protocolo modificado (Kistner e Matamoros, 2005). A integridade e quantidade do RNA isolado foram determinadas usando um 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA/EUA). Quantidades iguais de RNA de alta qualidade de cada genótipo foram utilizadas para síntese de cDNA.

Construção de biblioteca de mRNAs

Bibliotecas *paired-end* de mRNA foram geradas a partir de 4mg de RNA total, em acordo com a instrução do produtor (Illumina Inc., San Diego, CA/EUA). A qualidade da biblioteca foi determinada novamente com um 2100 Bioanalyzer (Agilent Technologies). A amplificação em *cluster* foi realizada com TruSeq PE Cluster Kit e um cBot (Illumina), e cada amostra foi sequenciada em uma lane

separada de GAIIx, utilizando o kit TruSeq SBS 36 Cyle (Illumina). Os reads gerados tinham 72bp.

Os dados brutos gerados pelo sequenciamento Illumina foram convertidos do formato BCL para qSeq utilizando o *software Off-line Basecaller*, v1.9.4 (OLB). Os arquivos foram então convertidos ao formato FastQ, o qual contém sequências de 72bp de comprimento, utilizando *scripts* desenvolvidos em nosso laboratório. Sequências de baixa qualidade foram removidas utilizando o *toolkit* NGS QC (Patel e Jain, 2012). Estas sequências incluíam reads com bases ambíguas, reads menores de 70bp e reads com uma qualidade *Phred* abaixo de 20. Todos os reads gerados foram depositados no banco de dados do National Center for Biotechnology Information (NCBI) e podem ser encontrados sob o número de acesso SRA073690.

Análise de expressão

Para encontrar a contribuição genotípica de cada transcrito, os *reads* de cada biblioteca foram mapeados contra todos os transcritos montados identificados como lncRNAs utilizando o alinhador *Bowtie* (Langmead et al., 2009). Os arquivos BAM gerado pelo *Bowtie* foram então usados para estimar a abundância a nível de transcrito para cada biblioteca usando o *software RSEM* (RNA-Seq by Expectation Maximization) (Li e Dewey, 2011).

Resultados e Discussão

Mineração de dados de EST: identificação de lncRNAs putativos

Do conjunto inicial de 121.342 unigenes recuperados, 23.529 sequências não apresentaram similaridade com nenhuma proteína descrita. Os demais unigenes foram então mapeados no genoma de *S. bicolor*, resultando 4.476 sequências mapeadas. Destas, 1.884 não apresentam ORF funcional e representam o conjunto de ncRNAs de cana-de-açúcar. Nós retiramos então aqueles ncRNAs menores do que 200bp, removendo a maioria das classes de pequenos RNAs, como snoRNAs e tRNAs. Como os precursores de miRNAs podem ser longos, nós alinhamos os unigenes a pre-miRNAs de plantas depositados no banco de dados miRBase (http://www.mirbase.org/) e identificamos 10 precursores. Por fim, nós utilizamos o *software RepeatMasker* para identificar de elementos transponíveis no conjunto candidato com os bancos de dados de elementos transponíveis *Plant Repeat Database* (http://plantrepeats.plantbiology.msu.edu/) e *RepBase* (http://www.girinst.org/repbase/). Identificamos 227 elementos repetitivos dentre os unigenes restantes. O conjunto candidato de lncRNAs de cana-de-açúcar disponível em bancos públicos consiste em 1473 transcritos (Figura 4). A distribuição de tamanho dos lncRNAs mostra grande número de transcritos menores do que 500bp (Figura 5). Uma vez que o tamanho médio identificado em catálogos de lncRNAs é de ~800kb, esse enriquecimento de transcritos menores pode refletir transcritos incompletos. A baixa expressão desta classe de transcritos pode responder pela incompletude destes.

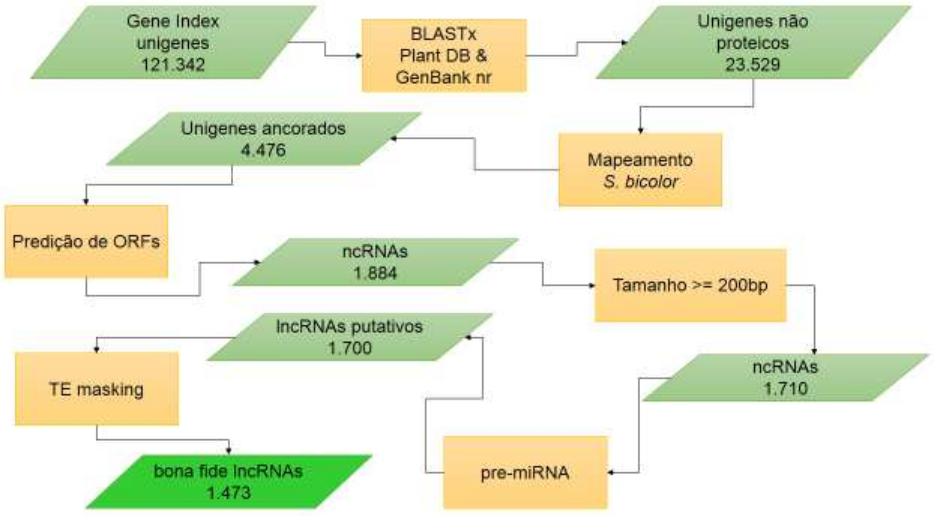


Figura 4. *Pipeline* proposto para prospecção de lncRNAs em ESTs de *S. officinarum*. Os valores contidos nas caixas indicam o conjunto de EST que passaram pelos filtros aplicados (caixas amarelas)

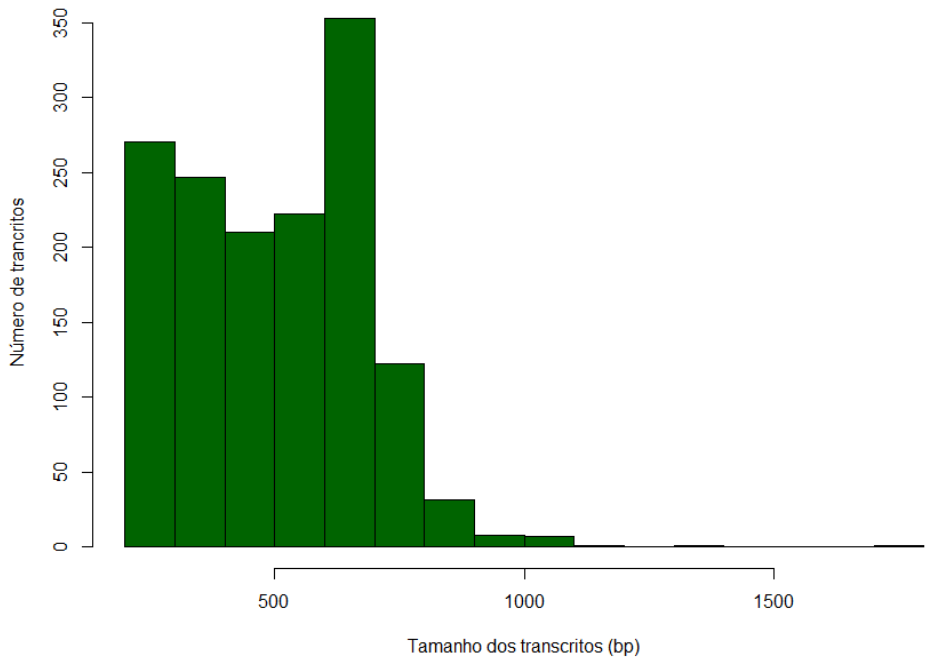


Figura 5. Distribuição de tamanho do conjunto de lncRNAs de cana-de-açúcar

Nós buscamos então evidências que atribuíssem atividade biológica a estes transcritos, através de três fontes. Determinamos que 14,5% dos lncRNAs identificados tem sRNAs mapeados na sua sequência. Apesar do número de siRNAs e miRNAs não diferir no conjunto total (Figura 6a), a distribuição de mapeamento por transcrito mostra um mosaico de alvos exclusivos de siRNAs variando a alvos exclusivos de miRNAs, o que demonstra diferentes comportamentos biológicos destas sequências: alvos exclusivos de siRNAs são silenciados de forma similar a TEs e alvos exclusivos de miRNAs são regulados finamente, o que denota importância funcional. Identificamos 59,6% similares a EST de sorgo, e portanto tem expressão conservada em Andropogoneae. Não obstante, 26,9% possuem estrutura secundária estável. A Tabela 1 resume 10 lncRNAs com maior probabilidade de possuírem estrutura secundária estável. Ao todo, 1.110 lncRNAs (75,3%) contem ao menos uma evidência de importância biológica (Figura 7).

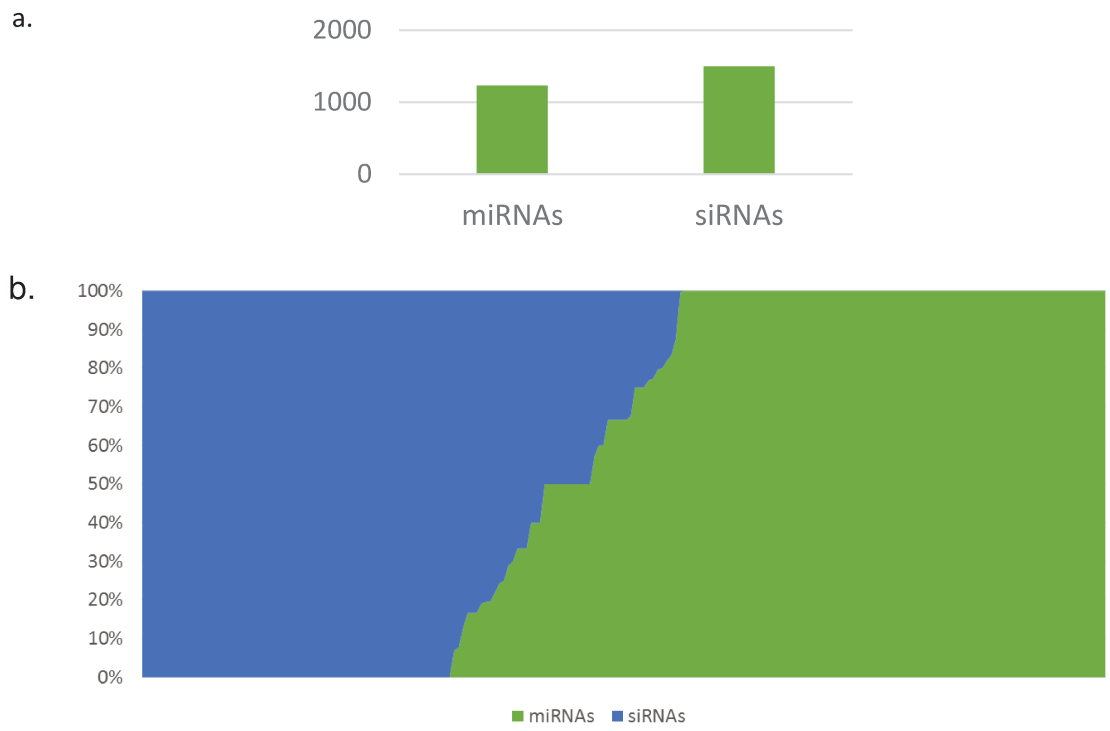


Figura 6. Mapeamento de bibliotecas de sRNAs. a. Distribuição global de mapeamento por classe de sRNA. b. Cobertura por classe de sRNAs sobre a extensão de 214 lncRNAs.

Tabela 1. lncRNAs com forte evidência de estruturação

Unigene ID	Start position	End position	Z-score
CA089842	30	274	10.02
CA298455	30	236	7.55
CA165541	60	301	5.92
CA261731	301	450	5.54
CA216577	295	390	5.24
CA270696	30	201	5.18
CA206277	90	228	5.02
CA230563	60	296	4.97
CA214248	90	301	4.93
TC137595	90	301	4.83

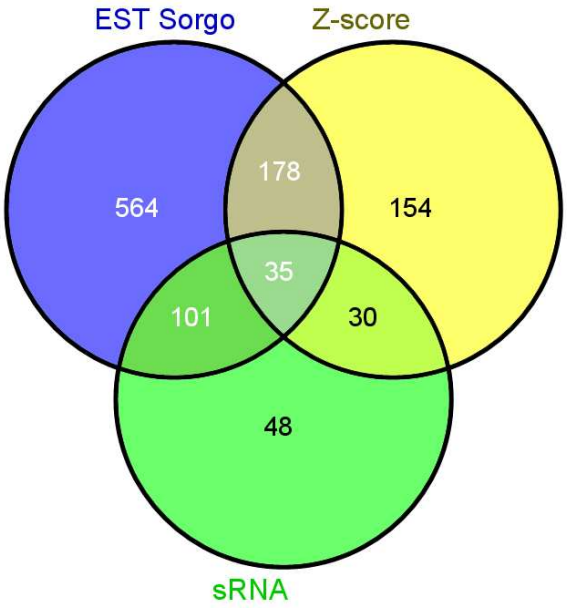


Figura 7. Sequências de EST com evidências indiretas de importância biológica, combinando estrutura secundária estável, conservação com *S. bicolor* e mapeamento de sRNAs.

Nós comparamos este conjunto inclusivo (1.110 sequências) com 18.910 transcritos montados por Cardoso-Silva et al. (2014), que não possuíam similaridade com proteínas de plantas. Nós encontramos 358 lncRNAs representados no conjunto de transcritos montados. Destes, 42% apresentaram estrutura secundária estável e 40% apresentaram similaridade com EST de sorgo. Nenhum dos unigenes representados no RNA-seq continha sRNAs mapeados a sua sequência. Por fim, comparamos o perfil de expressão dos lncRNAs nos diferentes genótipos. Surpreendentemente, lncRNAs majoritariamente apresentaram expressão genótipo-específica (Figura 8), o que indica caminhos evolutivos distintos para um mesmo transcrito mesmo em linhagens muito próximas entre si. Uma análise de clusterização hierárquica revelou um padrão de separação entre

os genótipos de programas de melhoramento distintos. Este resultado está em acordo com a observação de que variedades de um mesmo programa de melhoramento contenham a mesma base genética.

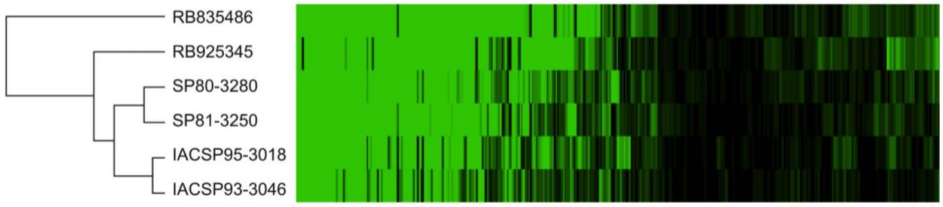


Figura 8. Clusterização hierárquica de 358 lncRNAs putativos de cana-de-açúcar. O padrão de expressão observado permitiu a identificação dos genótipos baseados na sua habilidade de armazenar açúcares e em acordo com os cruzamentos biparentais envolvidos nas diferentes populações de mapeamento

Nós observamos que lncRNAs de plantas apresentam elevada variação intraespecífica nos perfis de expressão. Este padrão pode decorrer da fraca inserção destes transcritos, de emergência evolutiva recente, nas redes metabólicas, de forma que seu perfil de expressão ainda é pouco regulado e apresenta comportamento randômico. Diversos trabalhos relataram que estes transcritos exibem padrões de expressão específicos a tecidos ou tipos celulares (Derrien et al., 2012; Guo et al., 2013; Hangauer et al., 2013; Liu et al., 2012).

Conclusões

A despeito da limitação imposta a bancos de EST como fonte completa do transcriptoma funcional de espécies, pudemos identificar um conjunto significativo de genes de lncRNAs representados. O *pipeline* proposto mostrou-se eficiente na identificação de alvos de interesse a estudos funcionais, representando 75% do conjunto identificado. Os indícios de regulação via sRNAs e o alto nível de estruturação de 65 transcritos sugerem genes de relevância às redes biológicas. Este estudo adiciona informações relevantes acerca da dinâmica evolutiva desses transcritos e elenca alvos de interesse para estudos futuros. A variação do perfil de expressão de uma porção significativa de lncRNAs respalda a hipótese de que estes transcritos estão fracamente inseridos em redes biológicas. Assim, transcritos de emergência recente nos mostram que os genes adquirem a capacidade de serem transcritos e são posteriormente cooptados a redes biológicas. O regime neutro que governa sua evolução é refletido na variação observada na expressão de lncRNAs em um mesmo tecido de variedades diferentes de uma espécie. A evidência de fraca pressão de seleção sobre lncRNAs corrobora essa hipótese e a extende a toda a classe de lncRNAs, majoritariamente linhagem-específicos.

Capítulo 2 – Prospecção e caracterização de transcritos de emergência recente de *Sorghum bicolor* – Padrões evolutivos comuns a genes linhagem-específicos

Introdução

O paradigma estabelecido sobre o surgimento de novos genes exclusivamente por duplicações genômicas não corresponde à dinâmica genômica revelada por tecnologia de sequenciamento em larga escala. A descoberta da extensão da transcrição por quase todo o genoma de eucariotos multicelulares, espaço rico no até então chamado DNA lixo, revelou uma gama de transcritos desconhecidos, sobretudo RNAs não codificadores. Trouxe a luz também a existência de genes específicos a cada linhagem estudada. Certamente, isto indica que eventos espúrios de transcrição, o ruído transcricional, desempenha um papel importante na geração de novas informações genômicas. Ademais, estes genes linhagem-específicos podem estar correlacionados com eventos de especiação.

lincRNAs são os modelos gênicos ideais para estudo da emergência de novos genes não codificadores, devido a sua posição no genoma. Porém, a maioria dos estudos realizados utilizou modelos animais, de forma que apenas *Arabidopsis thaliana* e *Zea Mays* foram analisadas em larga escala quanto ao conteúdo de lncRNAs, dos quais apenas uma fração é intergênica. Estes genes apresentam papéis funcionais diversos, embora majoritariamente atuem como reguladores.

Apesar de catalógos de LSGs serem mais fáceis de identificar, muitos desses ainda não estão anotados e são genes preditos computacionalmente quando da anotação de cada genoma. Dessa forma, o enriquecimento de papéis específicos é desconhecido. Porém, a importância óbvia desses genes no processo de especiação sugere funções importantes as diferentes linhagens evolutivas.

Sob o modelo proposto pela teoria entrópica da complexidade genômica, processos estocásticos resultantes de propriedades emergentes do sistema genômico podem governar a evolução de genes, codificadores ou não. Estes padrões podem ser evidenciados na arquitetura dos genes e na sua integração em redes biológicas. Neste capítulo, investigamos se padrões dessa natureza podem ser observados em genes do transcriptoma de sorgo.

Material e Métodos

Filoestratigrafia do transcriptoma codificador de *S. bicolor*

As 33.032 proteínas representativas de cada *loci* preditas para os genes anotados no genoma de *S. bicolor* v2.1 (Phytozome - Goodstein *et al.*, 2011) foram alinhadas a sequências proteicas de espécies representativas de grupos taxonômicos ancestrais da filogenia de Viridiplantae (Tabela 2). As sequências das classes taxonômicas 'Procariotos' (Bacteria + Archea) e 'Metazoa/Fungi' são todas as sequências de proteínas anotadas de genomas referência do banco de dados Uniprot (Uniprot Consortium, 2014). As proteínas de Viridiplantae foram obtidas do banco Phytozome e foram subdivididas de acordo com a taxonomia apresentada pelo próprio *Phytozome*. Utilizamos a estratégia de *best reciprocal hits* de BLASTp (e-value <= 1e-5) para determinar os ortólogos de cada proteína de sorgo dentro do grupo Viridiplantae. Para os dois grupos mais basais, foi utilizado o método de *reciprocal hit*, como proposto por Quint et al. (2012). As sequências foram ranqueadas segundo a classe mais antiga na qual se detectou um ortólogo.

Tabela 2 Genomas utilizados para datação relativa do transcriptoma codificador de S. bicolor

Classe Taxonômica	Genomas de Referência
Procariotos	3839
Metazoa/Fungi	2759
Viridiplantae	1
Embryophyta	1
Tracheophyta	1
Magnoliophyta	27
Poaceae	2
Panicoideae	2
Andropogoneae	1
LSG	1

Sorghum bicolor RNA-seq

Sementes de *Sorghum bicolor* cv 'Rio' foram esterilizadas em 5mL de etanol 70% por 5 minutos seguidos de 20 minutos de imersão em solução de hipoclorito de sódio 40%. As sementes foram então lavadas com água e dispostas em placas de petri contendo papel filtro embebido em água para germinação. Nós excisamos cotilédones de plântulas de ~7 dias de idade e os transferimos para 20mL de meio MS meia força, sem a adição de açúcares. A solução foi mantida sob agitação e luz contínua por 24 horas. Em seguida, conduzimos 4 tratamentos: solução de glicose (3%); solução de sacarose (5.7%), solução de manitol (3%) – usado como controle osmótico; e ABA (10 µM). Os controles negativos foram transferidos para solução de meio MS meia-força, sem adição de açúcares. Cada tratamento foi conduzido com 3 réplicas contendo 7-10 plântulas. RNA total (PoliA+)

foi extraído com 'RNeasy Mini Kit' (Qiagen) e a integridade foi determinada com 'Agilent 2100 Bioanalyzer'. As bibliotecas de RNA foram preparadas como recomendado pelas especificações do Illumina GA® e submetidas então a corridas de 47bp. As réplicas foram devidamente indexadas para multiplex. *Reads* de baixa qualidade e menores do que 35bp foram descartados.

Montagem e fusão de transcritos de *S. bicolor*

O conjunto de transcritos utilizado para mineração de lincRNAs de S. bicolor foi montado com a reunião de três conjuntos de dados, sendo: i. RNA-seq de plântulas de ~7 dias, descrito acima; ii. RNA-seq de plântulas de ~9 dias, realizado e descrito por Dugas et al., (2011); e iii. Transcriptoma referência disponibilizado no banco Phytozome (*S. bicolor* v2.1), montado a partir de 2,5B *reads* de RNA-seq e conjunto de 209.835 ESTs de *S. bicolor*.

Os *reads* de alta qualidade gerados pelos dois grupos (Tabela 3) foram reunidos em um único conjunto e mapeados no genoma de *S. bicolor* v2.1 com o *software* STAR v2.3 (Dobin et al., 2013), que identifica também sítios de *splicing*. O mapeamento foi utilizado para alimentar o *software* Cufflinks v2.1.1 (Trapnell *et al*., 2010), o qual monta os transcritos baseado no nível de expressão dado pelo mapeamento somado a presença de sítios de *splicing* no locus em análise. Estes transcritos montados foram anotados em um arquivo GFF. Os arquivos de anotação provenientes dos experimentos de RNA-seq e dos transcritos públicos montados e depositados no banco de dados Phytozome foram fundidos com o software *gffread*, disponível como acessório no pacote Cufflinks.

Tabela 3. Caracteríscas e sinais testados nos experimentos de RNA-seq utilizados para montagem dos transcritos de *S. bicolor*

	BR Dataset	US Dataset (Dugas et al, 2011)
Plant Material	7 days old	9 days old
Tissues	Cotyledons	Root and shoot
Treatments	10μM ABA 167mM Glucose 167mM Sucrose 167mM Mannitol (osmotic control) → 2h	20μM ABA 20% PEG-8000 → 27h

Mineração de lincRNAs de *S. bicolor*

Uma vez unificado o transcriptoma de *S. bicolor*, aplicamos uma estratégia para a identificação de lincRNAs de sorgo. Determinamos 4 critérios para identificação desta classe de transcritos, como descrito a seguir:

I. Foram retiradas sequências de genes sobrepostos a genes anotados ou que incidem em regiões adjacentes (500bp à montante e 2kb à jusante ao gene codificador); genes similares (BLASTx, e-value <= 1e-20) ao banco de proteínas de plantas revisado (Uniprot/Swissprot); e sequências com probabilidade de ser codificadora acima de 30%, determinada com o *software* de aprendizado de máquina CPAT v1.2. Esse software foi treinado com sequências codificadoras de proteínas dos transcritos primários do transcriptoma referência (Phytozome) e as

sequências não codificadoras de *Arabidopsis thaliana* (~6.500 lincRNAs) e de *Zea mays* (~650 lncRNAs) (Liu et al., 2012; Boerner e McGinnis, 2012).

II. O segundo critério é a remoção de sequências de transcritos menores do que 200bp, critério utilizado em estudos de identificação de lincRNAs por remover sequências de *housekeeping* ncRNAs.

III. Remoção de sequências de RNAs ribossomais (identificados com BLASTn com e-value ≤ 1e-20 contra um banco de dados de rRNAs de plantas) e pri-miRNA foram também excluídas. Três bibliotecas públicas de sRNA-seq de sorgo (Zhai et al., 2011) foram utilizadas para alimentar o pacote sRNAbench v12/13.

IV. Remoção de sequências que mapeiam mais de uma vez no genoma de sorgo e sequências mascaradas com o software RepeatMasker, alimentado com os bancos de referência de *repeats* Repbase v18.08 (Jurka e Kapitonov, 2005) e *TIGR Plant Repeat Database* (Ouyang & Buell, 2004).

Conservação de lincRNAs de *S. bicolor*

A análise de conservação de lincRNAs de sorgo foi conduzida de forma a identificar primeiramente lincRNAs sintênicos com *Zea mays*, *Setaria italica* e *Oryza sativa* (Phytozome). Os transcritos primários de cada loci de lincRNAs de sorgo foram mapeados no genoma dessas espécies com software BLAT (Kent, 2002), com os parâmetros *default*. Em seguida utilizamos o software MCScanX para identificar blocos sintênicos de pelo menos 5

ortólogos. S*cripts* desenvolvidos no laboratório foram utilizados para identificar lincRNAs que estavam inseridos nestes blocos sintênicos e apresentavam mapeamento na referida posição sintênica dos genomas comparados. A cobertura de alinhamento dos lincRNAs sintênicos foi determinada como *cutoff* para a determinação de lincRNAs conservados.

Caracterização da arquitetura gênica de LSGs e lincRNAs

A fim de determinar a correlação evolutiva dos transcritos codificadores e lincRNAs linhagem-específicos, e que portanto tem idade de emergência próxima entre si, nós analisamos características da arquitetura gênica desses conjuntos. As características acessadas são o tamanho do transcrito, número de exons, tamanho dos introns, níveis máximos de expressão, conteúdo GC, e número de eventos ancestrais de transposição nas adjacências dos transcritos. O nível máximo de expressão foi determinado com o software Cuffdiff, normalizado na forma de FPKM. Eventos ancestrais de transposição estão disponíveis no banco Phytozome como *track* do gBrowser de *S. bicolor* v2.1 e foram identificadas com o software RepeatMasker.

Análise de expressão diferencial

Nós utilizamos a anotação integrada obtida como descrito acima (ver *Montagem e fusão de transcritos de* S. bicolor) para realizar a contagem dos reads mapeados para cada locus identificado, utilizando os arquivos de mapeamento SAM e o *software*

HTSeq (Anders et al., 2014). Os arquivos de contagem foram então utilizados para análise de resposta diferencial entre as bibliotecas de tratamentos e de tecidos distintos. Nós utilizamos o método de normalização *voom* (Law et al., 2014), implementado na linguagem R, e subsequente análise com *pipeline* descrito no manual do pacote *limma*. Loci com p-valor ajustado pelo método *empirical bayes* menor ou igual a 0,01 foram considerados diferencialmente expressos.

Construção de redes de co-expressão *coding-noncoding* (CNC)

Determinamos os valores absolutos de expressão (FPKM – Cuffdiff), normalizados na escala logarítmica na base 2, para todos os transcritos, a partir dos dados de RNA-seq. Em seguida calculamos o coeficiente de correlação de Pearson para cada dupla de genes. As correlações foram então ranqueadas pelo índice de correlação (R^2). Pares ranqueados reciprocamente entre os dez mais correlatos (HRR máximo de 10) (Mutwil et al., 2010) foram considerados coexpressos.

Anotação funcional de lincRNAs

Os lincRNAs com no mínimo 10 parceiros codantes foram anotados através da análise de enriquecimento de categorias de GO atribuídas a estes parceiros, como descrito por Guo et al. (2013). Após determinar enriquecimento de termos GO dos parceiros, estes termos são atribuídos ao lincRNA analisado.

Resultados e Discussão

Identificação da filoestratigrafia de *S. bicolor*

A determinação da idade relativa de genes de uma dada espécie permite a avaliação do papel destes genes em processos importantes a sua linhagem, como os genes que surgiram após a conquista do meio terrestre por plantas superiores. Como apresentado na Figura 9, identificamos ~19% dos genes preditos no genoma revisado de *S. bicolor* como genes linhagem-específicos. Esta classe é maior do que todas as demais classes, havendo ainda enriquecimento também para as duas classes mais antigas (Procariotos e Metazoários/Fungos). Espera-se que genes existentes desde a separação das linhagens de procariotos e metazoários/fungos sejam componentes de vias metabólicas basais. Estudos realizados com *Arabidopsis thaliana* mostram que ~4,5% do genoma corresponde a genes linhagem-específicos. Como observado na Figura 9, 15% dos genes surgiram entre a divergência do ancestral comum a Tracheophyta, representado por *Sellaginella moellendorfii* e a divergência do ancestral comum a Magnoliophyta, as angiospermas. Esse enriquecimento reflete a surgimento de estruturas como a flor, que possibilitou a colonização e a diversificação deste grupo de plantas. Porém, espera-se que o sequenciamento de novos representantes dos grupos mais basais de angiospermas redistribuirá uma porção significativa os genes atribuídos a Magnoliophyta.

O contraste observado entre a tendência de aumento proporcional do número de LSGs ao tempo de divergência e o enriquecimento de LSGs em sorgo sugere elevada taxa de substituição destes genes (Figura 10). Ao considerarmos a taxa de

emergência *de novo* de transcritos como sendo constante em função da taxa de mutação, o aparente gargalo apresentado por grupos de divergência recente (Andropogoneae, Panicoideae e Poaceae) indica que poucos genes de emergência recente são mantidos evolutivamente quando há divergência entre populações. Este comportamento referente a emergência e morte de genes pode ser observado em outros trabalhos recentes (Gibson et al., 2013; Tautz e Domazet-Lošo, 2011; Zhao et al., 2014).

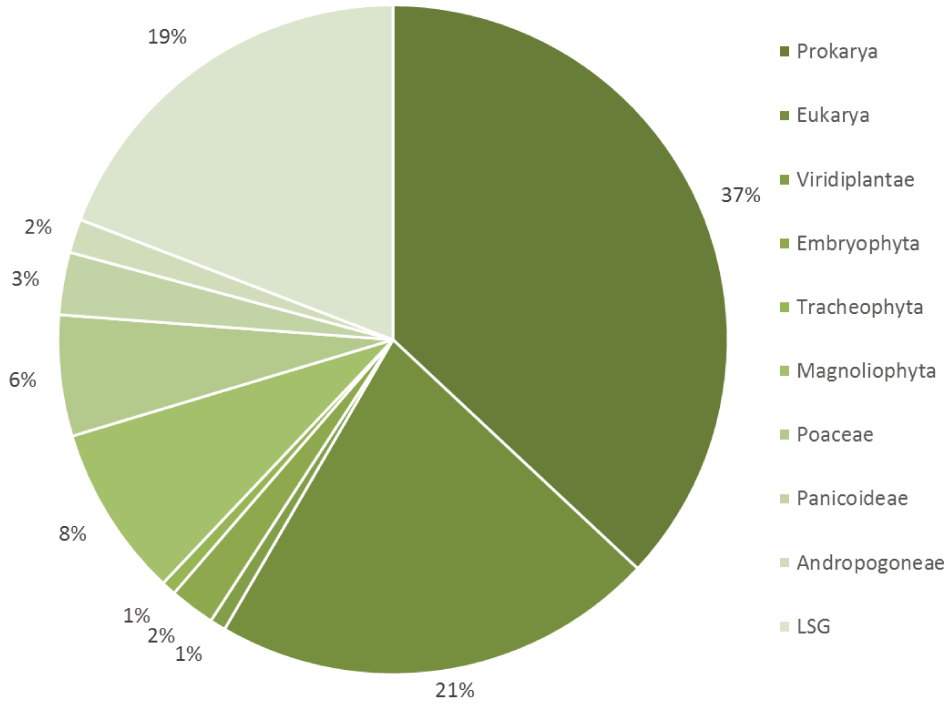


Figura 9. Distribuição dos genes de sorgo por classe de idade.

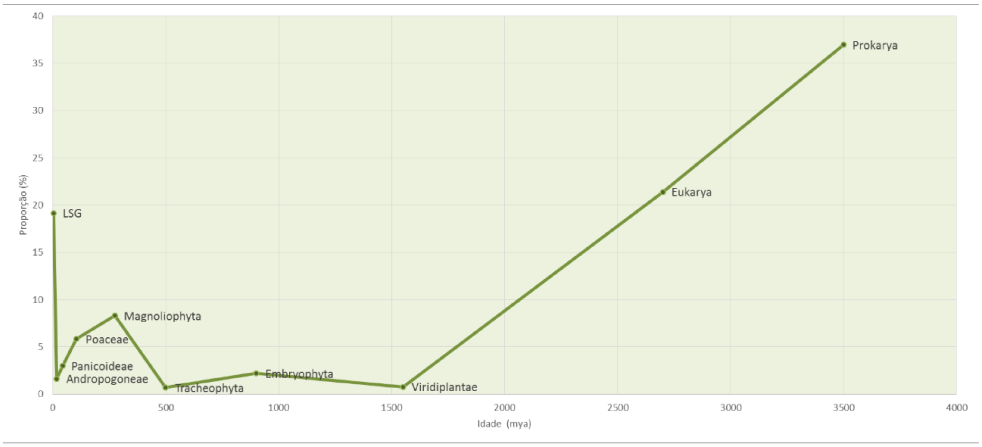


Figura 10. Distribuição do transcriptoma de sorgo em função do tempo de divergência de classe de idade.

Identificação de lincRNAs de *S. bicolor*

Nosso *pipeline* desenvolvido para identificação de lincRNAs resultou em um conjunto exclusivo de 432 *loci* de lincRNAs, representando apenas 0,88% do transcriptoma de sorgo. Encontramos ~41.000 potenciais sequências codificadoras que representam o próprio conjunto de genes anotados, mas também transcritos intrônicos ou antisenso a genes codificadores, representados nas amostras utilizadas. Apesar destes transcritos possivelmente desempenharem papel importante na dinâmica genômica, sua evolução deve estar correlata a genes sobrepostos a sua sequência. A busca por *housekeeping ncRNAs* revelou ainda ~1000 transcritos identificados como pri-miRNAs. Este grupo representa um conjunto alvo interessante, uma vez que encerra possíveis reguladores ainda não descritos para *S. bicolor*. De forma similar, uma porção significativa dos transcritos filtrados como TE não representam grupos anotados nos bancos públicos identificados. Apesar da duplicação de lincRNAs ser obviamente esperada,

transcritos que não foram mascarados mas apresentavam número expressivo de *hits* no próprio genoma de *S. bicolor* são elementos transponíveis desconhecidos em plantas. Apesar destes dois grupos de transcritos serem de grande interesse a estudos funcionais, sua caracterização foge ao escopo deste trabalho e será concluída em esforços futuros. O chamado lincRNoma de *H. sapiens*, que representa o catalogo melhor descrito de lincRNAs, contem 8 mil transcritos (Cabili et al., 2011). Outros trabalhos com modelos animais também apontam para um catálogo de lincRNAs com 10–15 mil genes. Nesse contexto, apesar de catálogos robustos de lincRNAs em plantas inexistirem, o conjunto de 432 genes de lincRNAs possivelmente representa uma fração do conjunto completo de lincRNAs de sorgo. Estes genes são altamente regulados, específicos a determinados tecidos e sinais, o que implica que a disponibilização de novos dados de transcriptômica de sorgo, abrangendo uma gama maior de situações, certamente permitirá desvendar novos genes dessa classe.

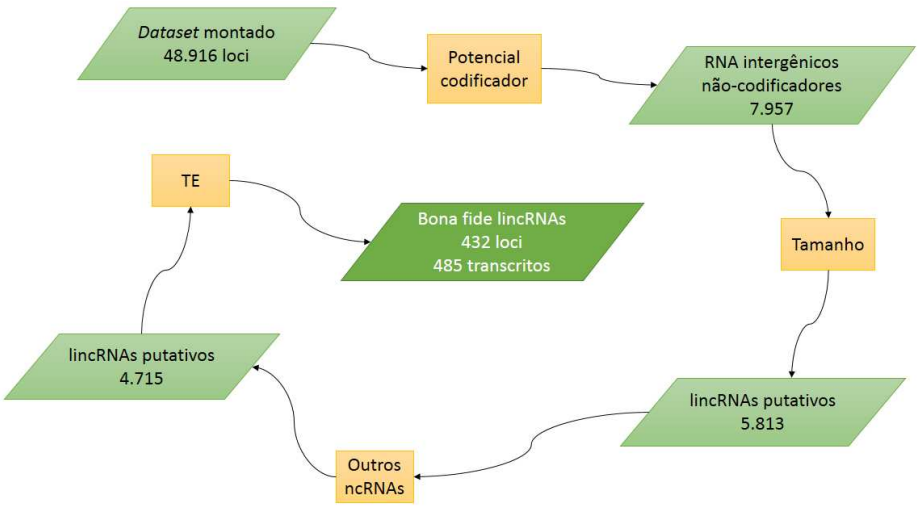


Figura 11. *Pipeline* de mineração de lincRNAs de *S. bicolor*

Caracterização dos lincRNAs de *S. bicolor*

A funcionalidade de genes de lincRNAs ainda é controversa, havendo grande resistência na aceitação destes transcritos como integrantes importantes dos sistemas biológicos. Uma vez que o evento de transcrição não garante nenhuma relevância funcional a estes transcritos, buscamos adicionar informações que contribuam para desvendar o possível papel deste conjunto de lincRNAs identificados. Utilizamos bibliotecas de sRNAs depositadas no banco *GEO Datasets* para investigar o possível envolvimento destes transcritos com vias de regulação por meio de pequenos RNAs (Figura 12). Independentemente do tecido analisado, encontramos pouca variação de lincRNAs alvos, porém, é possível observar mudanças na representatividade de classe de sRNA em relação ao tecido considerado, como o enriquecimento de miRNAs na biblioteca extraída do estolão de sorgo, atingindo aproximadamente o dobro da contagem de miRNAs da biblioteca extraída de flores. Não obstante, analisamos o nível de estruturação dos lincRNAs através de *z-scores* da MFE de janelas de 30 a 300bp dos transcritos (Figura 13). Seffens e Digby (1999) demonstraram que mRNAs não apresentam MFE menor do que conjuntos randômicos com a mesma composição e comprimento, o que está de acordo com o pressuposto de que mRNAs não apresentam estrutura secundária funcional. Assim, observamos que 38,4% do conjunto de lincRNAs apresenta níveis de MFE pelo menos dois desvios padrão da média de transcritos randômicos, indicando a existência de estrutura funcional nesses transcritos. A Tabela 4 apresenta os vinte lincRNAs com maior evidência estatística de presença de estrutura secundária estável.

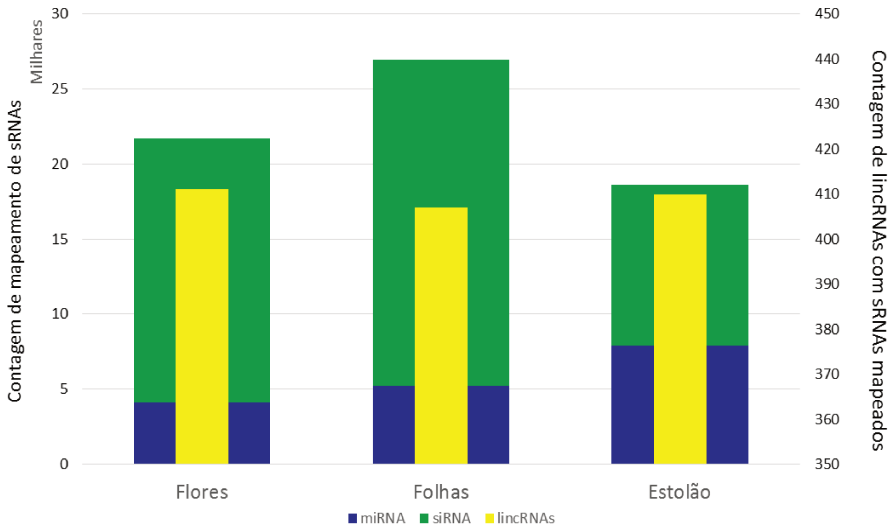


Figura 12. Mapeamento de pequenos RNAs de *S. bicolor* sobre os transcritos primários de cada locus de lincRNAs identificado

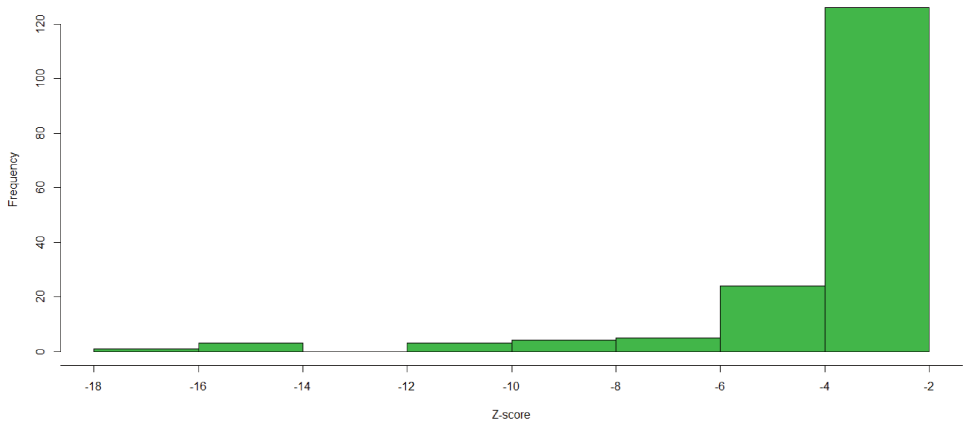


Figura 13. Distribuição de *z-scores* de lincRNAs. A despeito da maioria dos transcritos estruturados apresentarem valores intermediários (-4 < *z-score* <= -2), identificamos transcritos com estruturas muito estáveis.

Tabela 4. Ranqueamento de lincRNAs por *z-score*.

lincRNA	Start position	End position	Z-score
asmbl_6569.sorghumv21_pasa5	90	301	-17.55
asmbl_2013.sorghumv21_pasa0	60	261	-14.09
asmbl_1389.sorghumv21_pasa1	30	215	-14.01
asmbl_972.sorghumv21_pasa1	90	253	-11.37
asmbl_435.sorghumv21_pasa6	261	510	-11.27
asmbl_1766.sorghumv21_pasa6	30	155	-11.23
asmbl_718.sorghumv21_pasa9	0	96	-9.81
asmbl_1693.sorghumv21_pasa9	30	261	-9.25
asmbl_2816.sorghumv21_pasa5	90	281	-8.85
asmbl_776.sorghumv21_pasa8	246	780	-8.06
asmbl_1724.sorghumv21_pasa5	30	266	-7.42
asmbl_3177.sorghumv21_pasa8	30	301	-7.18
asmbl_574.sorghumv21_pasa4	301	510	-6.92
asmbl_541.sorghumv21_pasa7	180	301	-6.42
asmbl_9387.sorghumv21_pasa0	90	301	-6.04
asmbl_7474.sorghumv21_pasa5	301	510	-6.02
asmbl_4469.sorghumv21_pasa9	0	253	-5.88
asmbl_6109.sorghumv21_pasa5	301	960	-5.74
asmbl_5895.sorghumv21_pasa0	90	176	-5.31
asmbl_10689.sorghumv21_pasa0	0	251	-5.27

A evolução de lincRNAs ainda é desconhecida. A determinação de lincRNAs sintênicos permitiu que estabelecêssemos o critério de conservação desses transcritos. Determinamos 661, 415 e 386 blocos sintênicos com *Z. mays*, *S. italica* e *O. sativa*. Nestes blocos, encontramos então 7, 7 e 3 lincRNAs sintênicos, respectivamente (Tabela 5). Uma vez que o mapeamento foi realizado com o software BLAT utilizando os parâmetros *default*, a identidade mínima de alinhamento foi de 90% por bloco de alinhamento. Assim, a cobertura foi utilizada como critério para determinação da conservação evolutiva de lincRNAs. O único lincRNA sintênico nas três espécies analisadas apresenta cobertura de mapeamento de 2,42

a 3,05%. Isso corresponde a sequências de ~70bp. Este dado nos indica a presença de um possível domínio conservado, em detrimento de conservação de toda a sequência do transcrito, e constituiu então o *cutoff* para determinar lincRNAs conservados, ou seja, apenas transcritos com alinhamento de pelo menos 70bp e cobertura maior do que 2,4%. O mapeamento sob esses critérios nos mostram elevada taxa de substituição de lincRNAs (Figura 14). Espécies fora do grupo de gramíneas apresentam uma pequena proporção de lincRNAs conservados (~2%). Como esperado, lincRNAs parecem apresentar um caminho evolutivo único em cada linhagem, como é evidenciado pela distribuição interespecífica de lincRNAs conservados (Figura 15). A ausência de função pode resultar em retenção diferencial desses transcritos por deriva. Ademais, novos dados que mostrem quais destas regiões homólogas resultam na transcrição de um RNA permitirão que se determine transcritos homólogos de fato.

Tabela 5. Identificação de lincRNAs sintênicos e calibração do critério de conservação

Zea mays	**Cobertura (%)**
asmbl_9502.sorghum21_pasa0	10,02
asmbl_5145.sorghum21_pasa11	2,45
asmbl_684.sorghum21_pasa1	63,67
asmbl_8935.sorghum21_pasa4	66,57
asmbl_3247.sorghum21_pasa6	52,79
asmbl_4202.sorghum21_pasa7	58,27
asmbl_415.sorghum21_pasa10	11,17
Setaria italica	
asmbl_5145.sorghum21_pasa11	3,05
asmbl_8935.sorghum21_pasa4	67,99
asmbl_3247.sorghum21_pasa6	38,43
asmbl_4202.sorghum21_pasa7	49,31
asmbl_7856.sorghum21_pasa1	13,3
asmbl_7326.sorghum21_pasa4	60,12
asmbl_4921.sorghum21_pasa8	17,79
Oryza sativa	
asmbl_1353.sorghum21_pasa0	17,63
asmbl_4142.sorghum21_pasa10	15,59
asmbl_5145.sorghumv21_pasa11	2,42

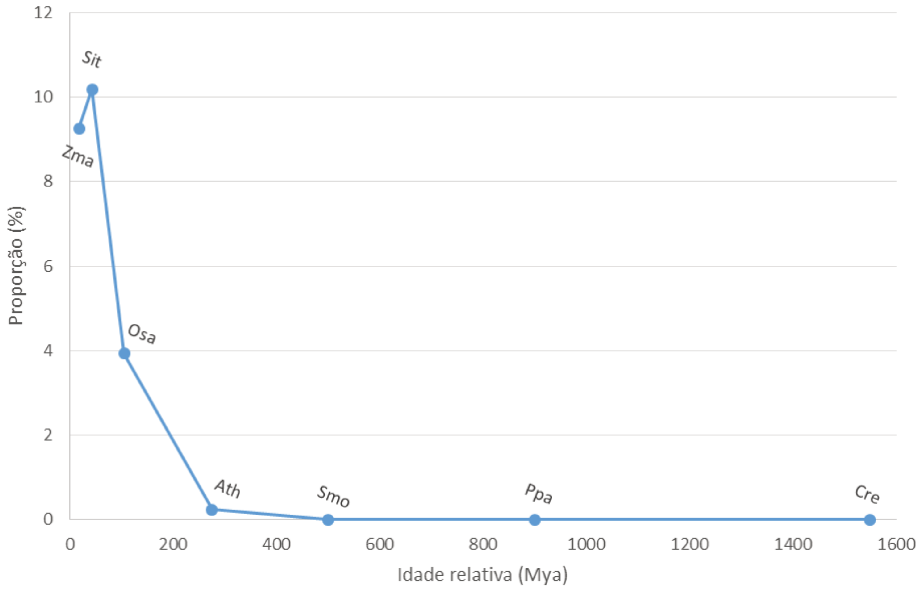


Figura 14. Conservação de lincRNAs em Viridiplantae, determinadas por mapeamento no genoma das espécies avaliadas, utilizando o *software* BLAT.

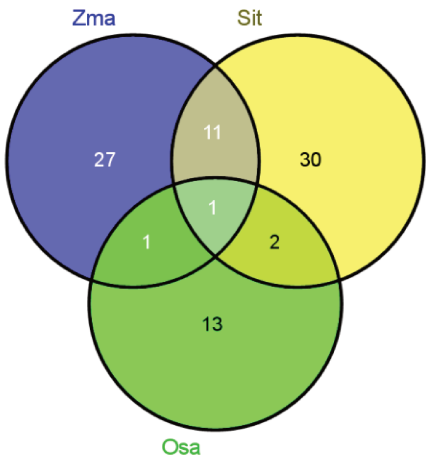


Figura 15. lincRNAs conservados em cada espécie analisada do grupo Poaceae. Conjuntos diferentes são mantidos evolutivamente quando da divergência das linhagens.

Resposta de lincRNAs a vias de sinalização

Em seguida, utilizamos o mapeamento do transcriptoma de sorgo ao genoma e quantificamos a expressão destes genes em resposta as vias de sinalização por açúcares (glicose e sacarose), pelo hormônio ABA e por stress hídrico. Apesar da variação apresentada no tamanho das bibliotecas (Figura 16), a obtenção de contagens totais que ultrapassam 4 milhões de reads por biblioteca e sobretudo a utilização de 3 réplicas por tratamento garantiram o rigor estatístico necessário para a avaliação de genes responsivos aos sinais testados.

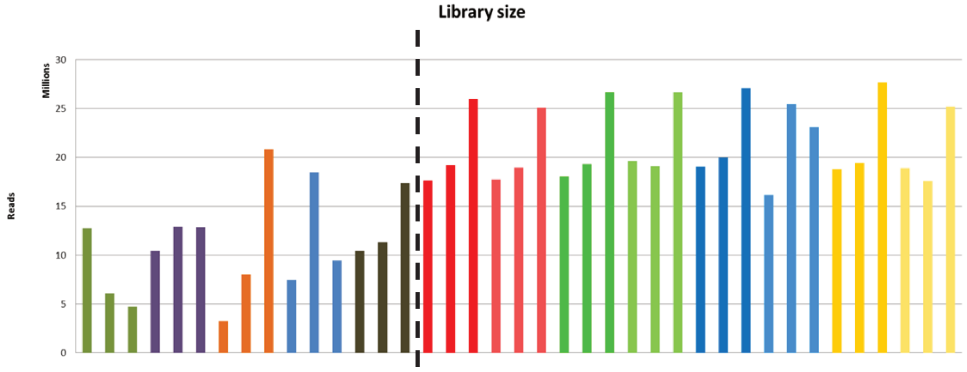
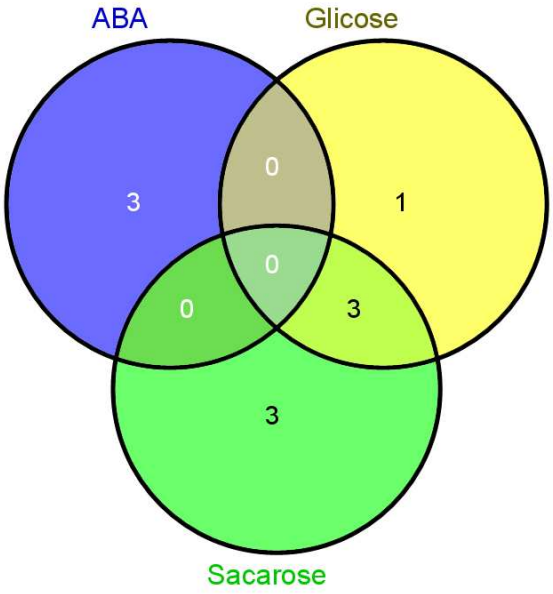


Figura 16. Tamanho das bibliotecas de sequenciamento utilizadas para análise de expressão diferencial em resposta as vias de sinalização por acúcares e ABA e em resposta a stress hídrico. As 15 blibliotecas à esquerda da linha tracejada foram geradas por Del Bem et al. As demais 24 bibliotecas foram geradas por Dugas et al., (2011).

O tratamento estatístico frequentista aplicado para determinação de expressão diferencial, a normalização *voom* e subsequente análise com o pacote *limma*, discriminam a nível de *locus* e não permitem comparações entre genes de uma mesma biblioteca, uma vez que não há normalização pelo tamanho do transcrito. Porém, a análise de variâncias entre amostras, relevante à pergunta biológica desenvolvida neste estudo, é muito robusta e

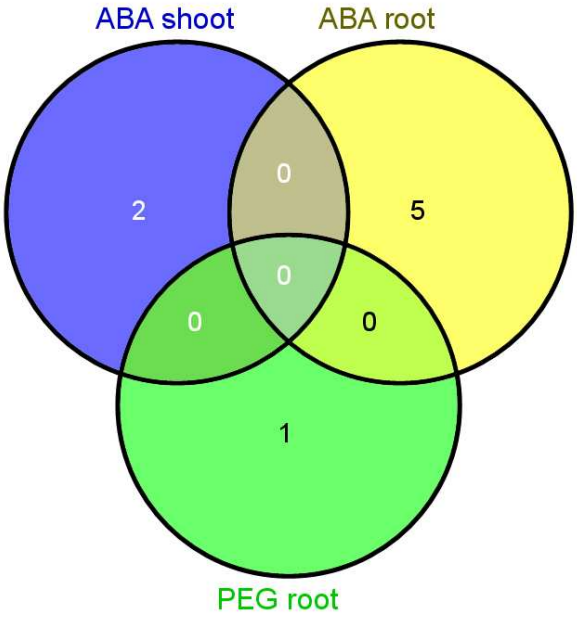
permite a identificação de genes responsivos, ao comparar verticalmente as variâncias da expressão de um mesmo gene em duas situações ou tratamentos, em relação a variância média de todos os genes. O critério restritivo de P-valor < 0,01 foi adotado para identificação de lincRNAs com alta probabilidade de responderem aos sinais testados, representando então um importante grupo alvo para estudos funcionais. Dos genes analisados, 3,7% são responsivos apenas aos sinais testados e 10,4% respondem diferencialmente devido ao tecido analisado. Estes dados sugerem que genes de lincRNAs respondem mais ao tecido do que a vias de sinalização. Isso reflete a fraca inserção destes transcritos em redes biológicas e é incompatível com a conclusão de que a fixação de lincRNAs está altamente relacionada com eventos de forte seleção positiva.



	ABA_c		$SUCR_c$		$GLUC_c$	
	UP	**DOWN**	**UP**	**DOWN**	**UP**	**DOWN**
lincRNAs	3	0	1	5	2	2

Figura 17. lincRNAs responsivos a vias de sinalição por açúcares e ABA em tratamentos de curta duração.

Como esperado, identificamos um forte *crosstalk* entre as vias de glicose e sacarose, embora seja interessante que haja tantos lincRNAs responsivos apenas a sacarose. Surpreendentemente, não detectamos lincRNAs comuns as vias de sinalização por açúcares e ABA (Figura 17).

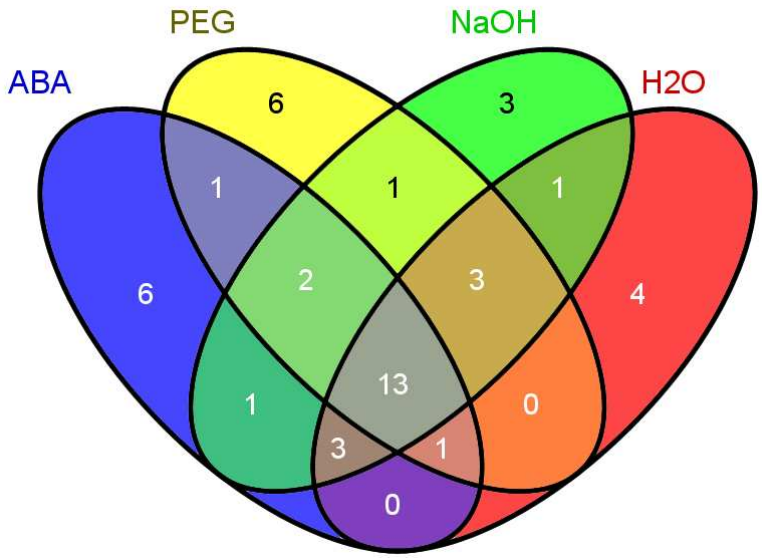


	$ABA_{l\ shoot}$		$PEG_{l\ shoot}$		$ABA_{l\ root}$		$PEG_{l\ root}$	
	UP	**DOWN**	**UP**	**DOWN**	**UP**	**DOWN**	**UP**	**DOWN**
lincRNAs	1	1	0	0	2	3	0	1

Figura 18 lincRNAs responsivos a vias de sinalição por açúcares e em resposta a stress hídrico em tratamentos de longa duração.

Os tratamentos de longa duração com ABA e PEG revelaram também independência entre os transcritos responsivos (Figura 18), sendo o tratamento com o hormônio vegetal mais enriquecido do que o tratamento de *stress* hídrico, com apenas um lincRNAs responsivo, na radícula das plantas. Encontramos mais genes responsivos ao tratamento longo com ABA do que o tratamento de curta duração, sendo que apenas o transcrito

asmbl_6865.sorghumv21_pasa5 é responsivo aos dois tratamentos (aumento da taxa transcricional no tratamento curto e no tratamento longo de raiz). Isto mostra a alta especificidade da regulação da expressão destes transcritos, especialmente no tecido radicular. Isto pode indicar um papel importante destes genes no desenvolvimento de plântulas. A resposta de lincRNAs devido ao tecido é muito mais pronunciada do que foi identificado em resposta aos sinais (Figura 19). Este dado demonstra a regulação especifica destes transcritos aparentemente de modo independente de sua função, de modo similar a variação intraespecífica identificada para lncRNAs de cana-de-açúcar (ver Capitulo 1).



	ABA		PEG		NaOH		H_2O	
	Shoot	**Root**	**Shoot**	**Root**	**Shoot**	**Root**	**Shoot**	**Root**
lincRNAs	20	7	17	10	18	9	18	7

Figura 19. lincRNAs regulados diferencialmente apenas em função do tecidos analisados. É clara a regulação mais elaborada destes transcritos na parte aérea da planta.

Foi possível determinar os lincRNAs responsivos aos tratamentos conduzidos e que também são conservados com espécies de gramíneas (Figura 20). Encontramos 5 lincRNAs responsivos e conservados com *Z. mays*, 3 com *S. italica* e apenas 1 com *O. sativa*. Estes genes podem representar um conjunto de lincRNAs alvo de seleção negativa, similar ao observado para genes codificadores. A conservação implica então em importância funcional as plantas. Segundo a teoria entrópica, a força de seleção purificadora pode sobrepujar o efeito de deriva genética sobre um gene emergente, a despeito do reduzido tamanho populacional, quando este gene implica no incremento do *fitness* do organismo, o que é corroborado pela pequena fração de lincRNAs conservados, representando poucos elementos que foram cooptados pelas redes biológicas pré-existentes.

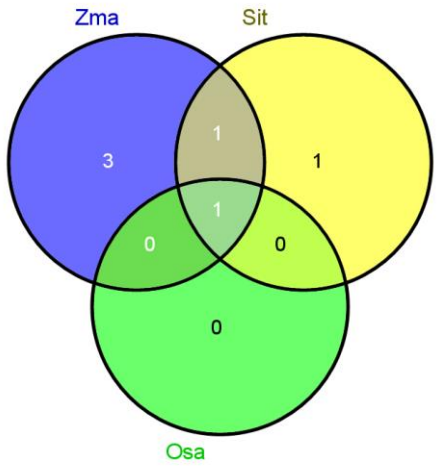


Figura 20. lincRNAs responsivos aos tratamentos e conservados com gramíneas.

Redes de coexpressão Coding-Non Coding (CNC)

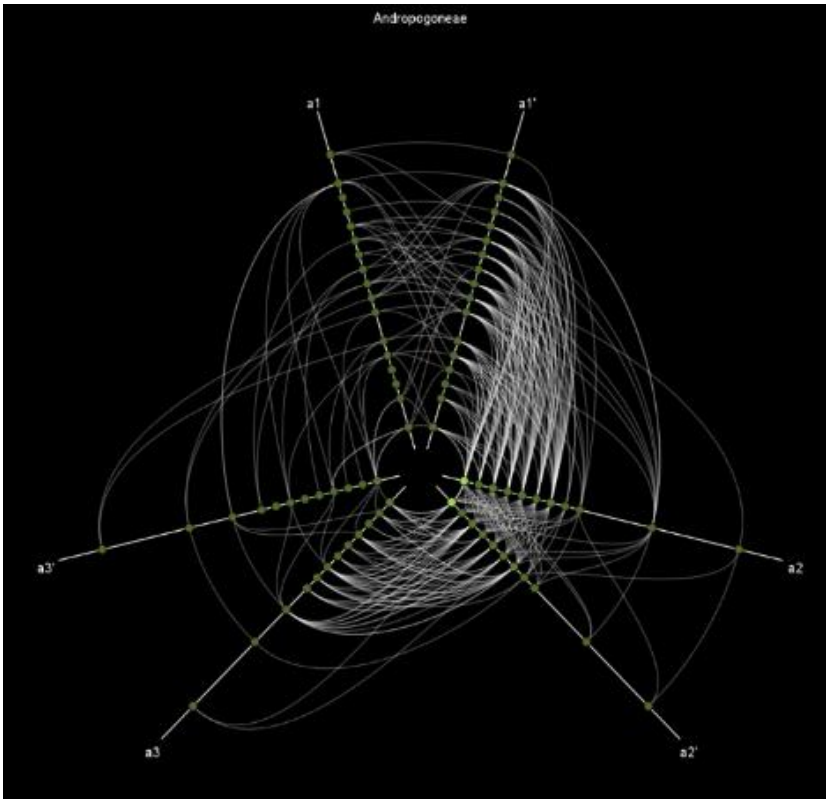
A rede de coexpressão foi construída determinando-se o coeficiente de correlação de Pearson par a par para lincRNAs e os genes codificadores de sorgo. Apesar de redes bayesianas identificarem parceiros diretos de forma mais eficiente, a natureza da análise desejada, i.e., a identificação de possíveis vias metabólicas nas quais os lincRNAs estão inseridos, requer parâmetros mais relaxados, incluindo então transcritos que interagem indiretamente. Foram determinadas 233.152 conexões entre genes codificadores, 4938 conexões entre gene codificadores e não codificadores e apenas 112 conexões entre genes não codificadores. A análise da rede revelou 241 lincRNAs com mais de 10 vizinhos que possuem um termo GO, e destes, 123 (28,47%) foram anotados devido ao enriquecimento de termos GO. A Tabela 6 resume os 16 lincRNAs responsivos aos tratamentos, dos quais 7 apresentam termos enriquecidos de GO e são portanto anotados funcionalmente. É intrigante que dos 5 genes de lincRNAs conservados em gramíneas e responsivos aos tratamentos, apenas 2 são também anotados pelo enriquecimento de termos de GO. Foram identificados termos relativos a processos basais, como componentes de membrana, até termos relativos a mecanismos de regulação gênica, como os lincRNAs TCONS_00055429 e asmbl_6865.sorghumv21_pasa5. Transcritos marcados com '*' são também sintênicos com as espécies analisadas. Os clusters foram determinados pelo método de HCCA.

Tabela 6. Enriquecimento de termos GO para lincRNAs conservados e responsivos

lincRNA Responsivos	Cluster	GO enriquecido	Conservação
asmbl_7856.sorghumv21_pasa1	c108	membrane	Sit*
asmbl_1293.sorghumv21_pasa9	c147	-	-
asmbl_6569.sorghumv21_pasa5	c147	-	-
asmbl_5162.sorghumv21_pasa11	c150	-	-
asmbl_76.sorghumv21_pasa7	c185	serine-type endopeptidase activity, hydrolase activity, ATPase activity (coupled)	-
asmbl_3517.sorghumv21_pasa0	c19	-	-
TCONS_00061515	c242	-	-
asmbl_5145.sorghumv21_pasa11	c274	catalytic activity	Zma*, Sit*, Osa*
asmbl_4202.sorghumv21_pasa7	c276	-	Zma*, Sit*
asmbl_2900.sorghumv21_pasa9	c305	ion binding,	-
TCONS_00055429	c344	nucleic acid binding TF activity	-
asmbl_6865.sorghumv21_pasa5	c368	binding	-
TCONS_00104679	c377	-	Zma*, Sit
TCONS_00055428	c61	cell, catalytic activity	-
asmbl_3424.sorghumv21_pasa5	c65	-	Zma
asmbl_5269.sorghumv21_pasa10	no hrr10	-	-

Nós investigamos então a conectividade dos nós referente a classe de idade a que cada um pertence, como representado na Figura 21a. É possível observar um padrão claro em que genes de uma dada classe de idade tenderão a se correlacionar com genes de idades próximas (Figura 21b), embora falhamos em detectar correlação entre genes de uma mesma classe, exceto para os genes mais recentes e genes da classe Viridiplantae. A inserção de genes recém adquiridos em redes biológicas pode alterar a topologia da rede, causando um efeito deletério. Dessa forma, é esperado que genes novos interajam com genes de emergência em uma janela de tempo comum. Estas interações originam novos módulos que podem então integrar redes metabólicas *a posteriori*. Como a complexidade de redes biológicas resulta de processos de evolução neutra (Lynch, 2007), a emergência de transcritos e inserção destes em redes biológicas deve estar sujeita aos mesmos processos. Dessa forma, apenas em situações de redução do *fitness* do organismo, a seleção negativa tem poder suficiente para atuar.

a



b

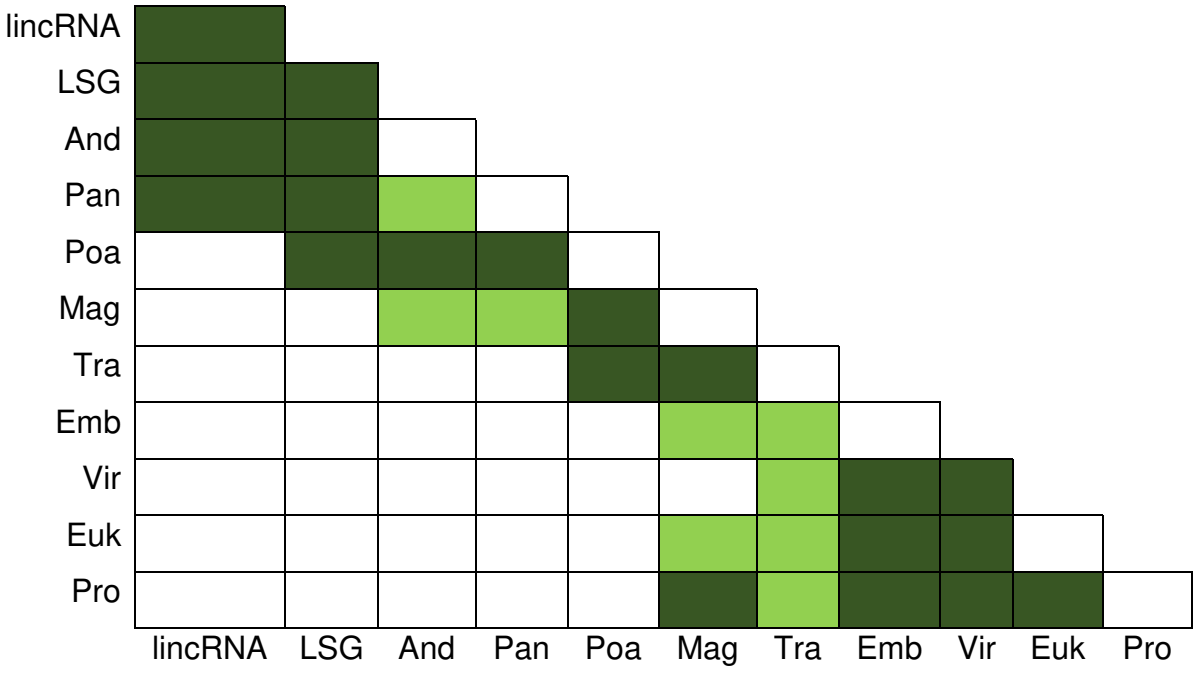


Figura 21. Correlações preferenciais significativas entre genes codificadores de diferentes classes de idade e lincRNAs. a. Conectividade entre lincRNAs (a1/a1'), LSGs (a2/a2') e genes da classe Andropogoneae (a3/a3'). b. Enriquecimento de correlações entre as classes de idade (branco – sem correlação; verde claro –

correlações significativas; verde escuro – correlações significativas e número de observações é 10% superior ao número esperado)

Nós determinamos então 10 estatísticas que descrevem a interação entre as classes de genes na rede construída (Figura 22). Apenas 3 apresentaram diferenças entre genes de classes diferentes. O parâmetro *clustering coefficient* de lincRNAs é significativamente maior do que o coeficiente de todas as classes de genes codificadores (Figura 22c). O número de conexões por nó, medido pelo *degree* (Figura 22d), é contrastante para lincRNAs e genes LSG. Ambos diferem de classes de genes mais antigos, mas lincRNAs apresentam um enriquecimento de conexões, em acordo com o elevado *clustering coefficient*, enquanto LSGs apresentam menor mediana, demonstrando que lincRNAs são mais corregulados com seus parceiros e que LSGs tem expressão mais independente. O parâmetro *eccentricity*, que é a distância máxima de cada nó até o nó mais distante dele, demonstra uma distribuição similar a genes de gramíneas, em que a excentricidade é menor do que em genes mais antigos. Estes genes emergiram há no máximo 200 milhões de anos atrás, e são relativamente novos, em comparação às demais classes. Isto reflete que estes genes ocupam posições mais centrais na rede.

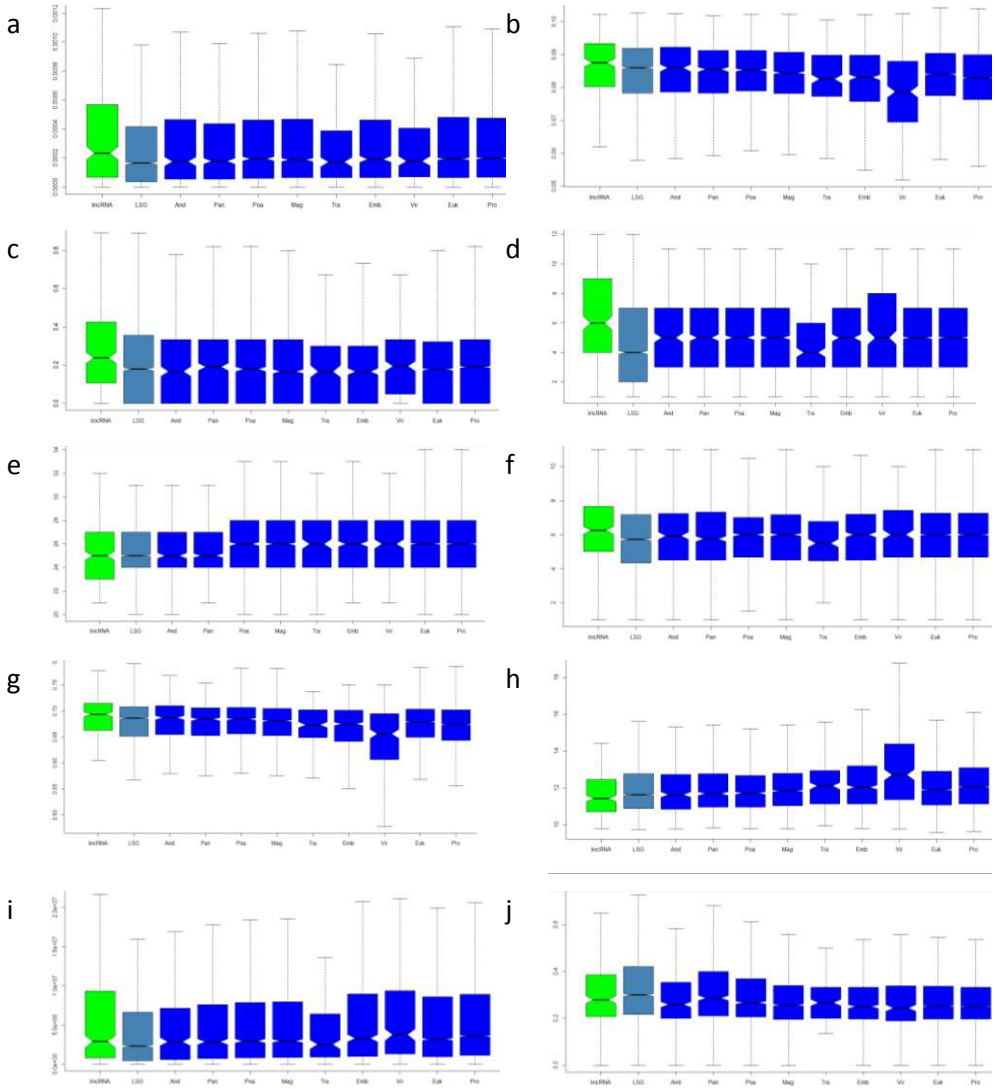


Figura 22. Estatísticas da interação de genes de diferentes classes na rede de coexpressão. a. *Betweenness centrality*; b. *Closeness centrality*; c. *Clustering coefficient*; d. *Degree*; e. *Eccentricity*; f. *Neighborhood connectivity*; g. *Radiality*; h. *Shortest path*; i. *Stress*; j. *Topological coefficient*. lincRNAs estão representados em verde, LSG, em azul claro, classes mais antigas (azul escuro).

Análise da arquitetura gênica de transcritos emergentes

O processo de emergência de genes possivelmente está correlacionado com eventos de transposição. Elementos transponíveis são muitas vezes silenciados de forma eficiente pelos genomas e esse elemento pode então ser cooptado e adquirir novas funções ou pode degenerar. No segundo caso, origina-se *de novo* um espaço de sequência de onde novos transcritos podem emergir futuramente. Dessa forma, espera-se anticorrelação entre a idade dos genes e o número de eventos antigos de inserção na proximidade dos genes. Surpreendentemente, não encontramos nenhum padrão claro quanto a idade dos genes e ilhas de *repeats* (Figura 23), em contraste ao esperado, uma vez que a emergência de lincRNAs foi correlacionada a eventos ancestrais de transposição (Kapusta et al., 2013). Porém, isto pode decorrer de viés da pequena amostragem de lincRNAs identificados.

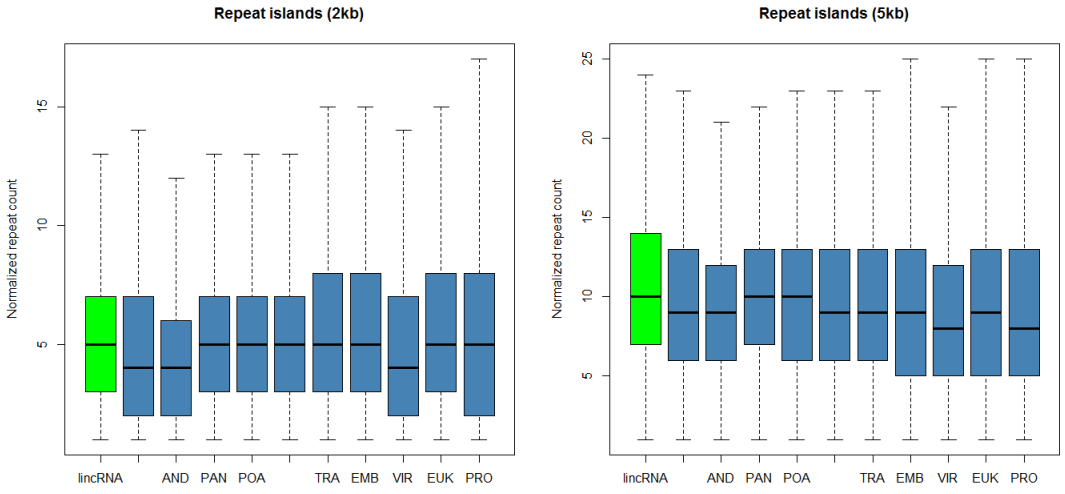


Figura 23. Distribuição de elementos repetitivos nas imediações de genes codificadores e lincRNAs

Ao considerarmos genes de emergência simultânea, esperamos que a organização da arquitetura desses genes seja semelhante. Existem evidências que respaldam o aumento da complexidade de genes codificadores e não codificadores ao longo de sua evolução (Wolf et al., 2009; Managadze et al., 2011). Assim, determinamos a arquitetura dos genes identificados de sorgo (Figura 24). Carvunis et al. (2012) demonstraram que genes codificadores que emergiram *de novo* apresentam claros padrões de arquitetura relacionados a sua idade, em que genes novos são mais curtos, menos expressos, o que revela região promotora pouco elaborada, e mais simples em termos de processamento pós-transcricional. O nível máximo de expressão de um gene reflete a complexidade de seu promotor, uma vez que transcritos com alto nível relativo de expressão apresentam elaboradas regiões promotoras. Os parâmetros tamanho do transcrito (Figura 24e), contagem de exons (Figura 24b) e nível máximo de expressão (Figura 24d) dão suporte a hipótese de padões de arquitetura comuns a genes de uma mesma idade, independentemente de ser codificador. Estes padrões possivelmente decorrem de propriedades intrínsecas do sistema genômico. A seleção natural atua então como modulador da evolução da arquitetura, embora apresente baixa eficiência. Porém, é possível observar que lincRNAs apresentam características ainda mais acentuadas que LSGs. O evento de tradução, que representa a adição de um novo nível de informação ao sistema genômico, pode aumentar levemente a eficiência de seleção, e causar assim o 'envelhecimento' de genes linhagem-específicos, o que não ocorre aos lincRNAs, que permanecem em níveis inferiores de informação, nos quais a seleção negativa apresenta eficiência muito baixa. A identificação de pressões de seleção mais acentuadas sobre LSG em detrimento de lincRNAs corrobora esta hipótese.

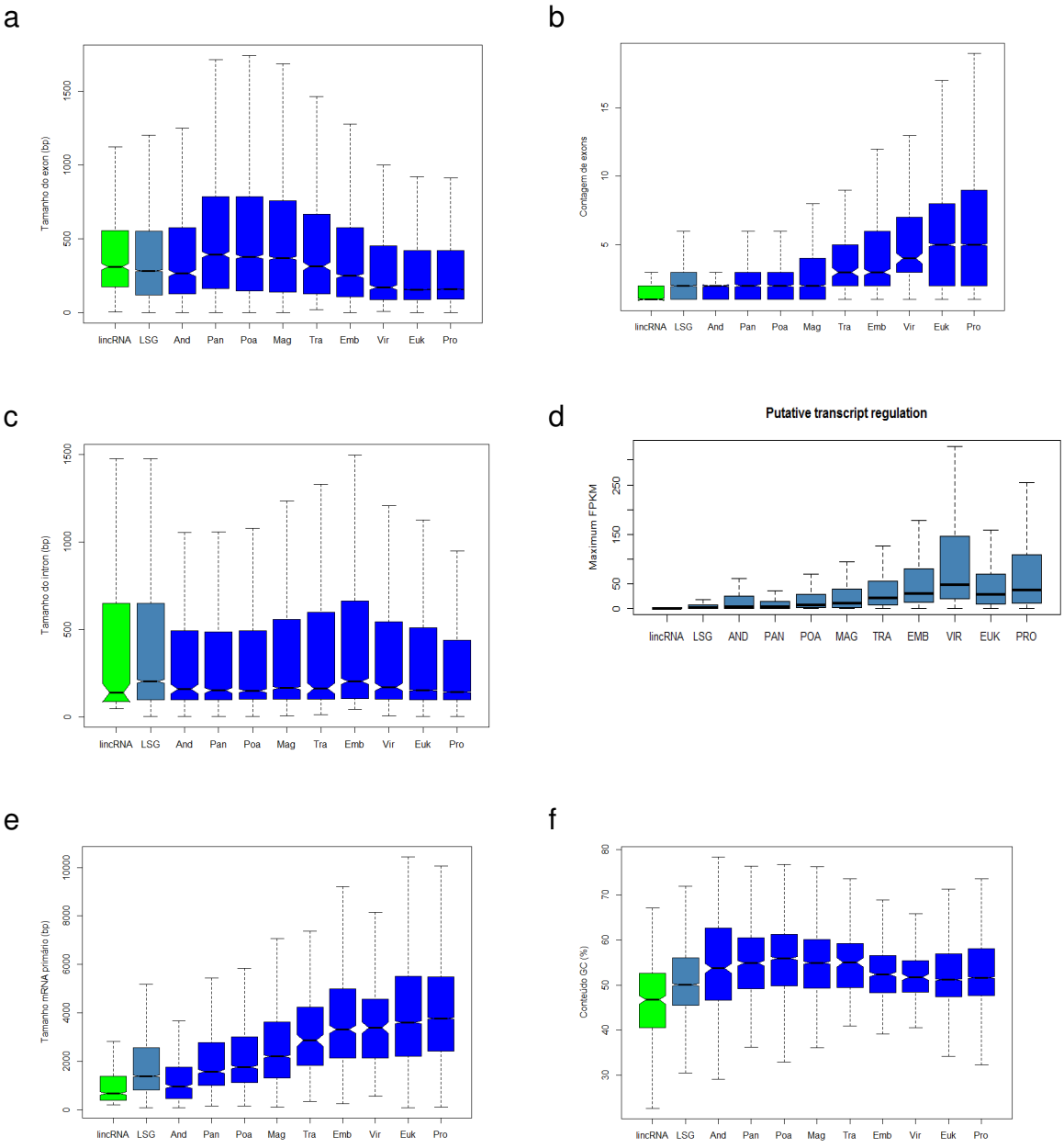


Figura 24. Análise da arquitetura gênica do transcriptoma de *S. bicolor*

Conclusões

Foi possível evidenciar que lincRNAs de sorgo, similar ao observado em modelos animais, apresentam estrutura secundária estável e são regulados por sRNAs, o que indica relevância biológica. A anotação destes genes devido ao enriquecimento de funções de seus parceiros sugere também relevância funcional.

Porém, a especificidade da resposta de lincRNAs aos tecidos sobreposta a resposta aos tratamentos reflete a baixa complexidade da região promotora destes genes e possivelmente o desempenho de papéis funcionais não essenciais. Esse papel não essencial, por sua vez, resulta em baixa pressão de seleção purificadora, que pode ser também evidenciada pela acentuada taxa de substituição de lincRNAs e LSGs. Não obstante, a arquitetura gênica e o comportamento de genes novos em redes biológicas sugerem padrões comuns de evolução a estes transcritos, que contrariam a visão estritamente adaptacionista e respaldam a evolução gênica constringida por fatores comuns intrínsecos ao genoma, apesar do processo evolutivo histórico modular as interações entre os genes e seu espaço adjacente.

Os resultados aqui apresentados demonstram ainda que genes de idades compatíveis, codificadores ou não codificadores, podem inicialmente formar novos módulos nas redes, tendo os lincRNAs como elementos centrais. É comum que lincRNAs sejam perdidos ao longo da evolução e, apesar de assumirem posições centrais, é provável que a redundância da rede permita que esses nós sejam extintos sem grandes prejuízos. Contudo, esse comportamento de lincRNAs dentro das sub-redes analisadas só pode ser compreendido em um contexto de evolução neutra.

Considerações Gerais

Capítulo 1

- lncRNAs de cana-de-açúcar apresentam evidencias de relevância biológicas e podem desempenhar funções importantes a planta, constituindo um grupo alvo para estudos funcionais;
- A elevada variabilidade no perfil de expressão dentro da espécie reflete a inespecificidade da região promotora, que possivelmente adquiriu módulos regulatórios distintos em cada linhagem.

Capítulo 2

- LSGs representam uma porção significativa do genoma de sorgo, sugerindo acentuada taxa de substituição;
- lincRNAs também apresentam elevada taxa de substituição, evidenciada pelo baixo nível de conservação;
- lincRNAs tem resposta transcricional preferencialmente a tecidos, embora tenhamos identificados elementos responsivos aos tratamentos na radícula, principalmente. Estes transcritos representam um conjunto de interesse para caracterização funcional;
- A conectividade e a arquitetura de genes de emergência recente apresentam padrões emergentes de evolução, o que, considerando cenário de força seletiva fracas e/ou ineficiente, sugere fatores comuns neutros que governam sua evolução;
- O efeito mais acentuado destes fatores sobre RNAs não codificadores indica que o evento de tradução, ao adicionar um novo nível informacional, pode aumentar a eficiência de seleção negativa sobre genes transcritos.

Referências Bibliográficas

Barbazuk, W. B., Emrich, S. J., Chen, H. D., Li, L., & Schnable, P. S. (2007). SNP discovery via 454 transcriptome sequencing. *The Plant Journal : For Cell and Molecular Biology*, *51*(5), 910–8.

Bardou, F., Merchan, F., Ariel, F., & Crespi, M. (2011). Dual RNAs in plants. *Biochimie*, *93*(11), 1950–4. doi:10.1016/j.biochi.2011.07.028

Bellodi, N., & Macedo, I. (1995). Quinta geração de variedades de cana-de-açúcar. COOPERATIVA DOS PRODUTORES DE CANA, AÇÚCAR E ÁLCOOL DO ESTADO DE SÃO PAULO. Technical Bulletin. Piracicaba, SP.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2013). GenBank. *Nucleic Acids Research*, *41*(Database issue), D36–42. doi:10.1093/nar/gks1195

Boerner, S., & McGinnis, K. M. (2012). Computational identification and functional predictions of long noncoding RNA in Zea mays. *PloS One*, *7*(8), e43047. doi:10.1371/journal.pone.0043047

Borevitz, J. O., & Chory, J. (2004). Genomics tools for QTL analysis and gene discovery. *Current Opinion in Plant Biology*, *7*(2), 132–6.

Bower, N. I., Casu, R. E., Maclean, D. J., Reverter, A., Chapman, S. C., & Manners, J. M. (2005). Transcriptional response of sugarcane roots to methyl jasmonate. *Plant Science*, *168*(3), 761–772.

Cai, J., Zhao, R., Jiang, H., & Wang, W. (2008). De novo origination of a new protein-coding gene in Saccharomyces cerevisiae. *Genetics*, *179*(1), 487–96. doi:10.1534/genetics.107.084491

Campalans, A., Kondorosi, A., & Crespi, M. (2004). ENOD40, a short open reading frame–containing mRNA, induces cytoplasmic localization of a nuclear RNA binding protein in Medicago truncatula. *The Plant Cell Online*, 1–14. doi:10.1105/tpc.019406.Indeed

Campbell, M. a, Zhu, W., Jiang, N., Lin, H., Ouyang, S., Childs, K. L., … Buell, C. R. (2007). Identification and characterization of

lineage-specific genes within the Poaceae. *Plant Physiology*, *145*(4), 1311–22. doi:10.1104/pp.107.104513

Capra, J. a, Pollard, K. S., & Singh, M. (2010). Novel genes exhibit distinct patterns of function acquisition and network integration. *Genome Biology*, *11*(12), R127. doi:10.1186/gb-2010-11-12-r127

Cardoso-Silva, C. B., Costa, E. A., Mancini, M. C., Balsalobre, T. W. A., Canesin, L. E. C., Pinto, L. R., … Vicentini, R. (2014). De Novo Assembly and Transcriptome Analysis of Contrasting Sugarcane Varieties. *PLoS ONE*, *9*(2), e88462. doi:10.1371/journal.pone.0088462

Carson, D., & Botha, F. . (2002). Genes expressed in sugarcane maturing internodal tissue. *Plant Cell Reports*, *20*(11), 1075–1081.

Carson, D. L., & Botha, F. C. (2000a). Preliminary Analysis of Expressed Sequence Tags for Sugarcane. *Crop Science*, *40*(6), 1769–1779.

Carson, D. L., & Botha, F. C. (2000b). Preliminary Analysis of Expressed Sequence Tags for Sugarcane. *Crop Science*, *40*, 1769–1779.

Carvunis, A.-R., Rolland, T., Wapinski, I., Calderwood, M. a, Yildirim, M. a, Simonis, N., … Vidal, M. (2012). Proto-genes and de novo gene birth. *Nature*, *487*(7407), 370–4. doi:10.1038/nature11184

Casu, R. E., Dimmock, C. M., Chapman, S. C., Grof, C. P. L., McIntyre, C. L., Bonnett, G. D., & Manners, J. M. (2004). Identification of differentially expressed transcripts from maturing stem of sugarcane by in silico analysis of stem expressed sequence tags and gene expression profiling. *Plant Molecular Biology*, *54*(4), 503–17.

Casu, R. E., Grof, C. P. L., Rae, A. L., McIntyre, C. L., Dimmock, C. M., & Manners, J. M. (2003). Identification of a novel sugar transporter homologue strongly expressed in maturing stem vascular tissues of sugarcane by expressed sequence tag and microarray analysis. *Plant Molecular Biology*, *52*(2), 371–86.

Cingolani, P., Patel, V. M., Coon, M., Nguyen, T., Land, S. J., Ruden, D. M., & Lu, X. (2012). Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. *Frontiers in Genetics*, *3*(March), 35.

Clote, P., Ferré, F., Kranakis, E., & Krizanc, D. (2005). Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency, 578–591. doi:10.1261/rna.7220505.4

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., & Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics (Oxford, England)*, *21*(18), 3674–6.

Consortium, U. (2014). Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Research*. Retrieved from http://nar.oxfordjournals.org/content/42/D1/D191.full-text-lowres.pdf

Cordeiro, G. M., Casu, R., McIntyre, C. L., Manners, J. M., & Henry, R. J. (2001). Microsatellite markers from sugarcane (Saccharum spp.) ESTs cross transferable to erianthus and sorghum. *Plant Science*, *160*(6), 1115–1123.

D'Hont, a, Grivet, L., Feldmann, P., Rao, S., Berding, N., & Glaszmann, J. C. (1996). Characterisation of the double genome structure of modern sugarcane cultivars (Saccharum spp.) by molecular cytogenetics. *Molecular & General Genetics : MGG*, *250*(4), 405–13. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8602157

Daugrois, J. H., Grivet, L., Roques, D., Hoarau, J. Y., Lombard, H., Glaszmann, J. C., & D'Hont, A. (1996). A putative major gene for rust resistance linked with a RFLP marker in sugarcane cultivar 'R570 '. *Theor. Appl. Genet*, *92*, 1059–1064.

Dekkers, J. C. M., & Hospital, F. (2002). The use of molecular genetics in the improvement of agricultural populations. *Nature Reviews. Genetics*, *3*(1), 22–32.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., … Frazer, K. A. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Research*, *22*(9), 1775–89. doi:10.1101/gr.132159.111

Dobin, A., Davis, C. a, Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., … Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, *29*(1), 15–21. doi:10.1093/bioinformatics/bts635

Dobzhansky, T. (1964). Biology, molecular and organismic. *American Zoologist*, *4*(4), 443–452. Retrieved from http://www.jstor.org/stable/3881145

Domingues, D. S., Cruz, G. M. Q., Metcalfe, C. J., Nogueira, F. T. S., Vicentini, R., Alves, C. de S., & Van Sluys, M.-A. (2012). Analysis of plant LTR-retrotransposons at the fine-scale family level reveals individual molecular patterns. *BMC Genomics*, *13*(1), 137.

Donoghue, M. T., Keshavaiah, C., Swamidatta, S. H., & Spillane, C. (2011). Evolutionary origins of Brassicaceae specific genes in Arabidopsis thaliana. *BMC Evolutionary Biology*, *11*(1), 47. doi:10.1186/1471-2148-11-47

Dugas, D. V, Monaco, M. K., Olsen, A., Klein, R. R., Kumari, S., Ware, D., & Klein, P. E. (2011). Functional annotation of the transcriptome of Sorghum bicolor in response to osmotic stress and abscisic acid. *BMC Genomics*, *12*, 514. doi:10.1186/1471-2164-12-514

Duret, L., Chureau, C., Samain, S., Weissenbach, J., & Avner, P. (2006). The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. *Science (New York, N.Y.)*, *312*(5780), 1653–5. doi:10.1126/science.1126316

Feltus, F. A., Wan, J., Schulze, S. R., Estill, J. C., Jiang, N., & Paterson, A. H. (2004). An SNP resource for rice genetics and breeding based on subspecies indica and japonica genome alignments. *Genome Research*, *14*(9), 1812–9.

Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M., & Miller, W. (1998). Genomic DNA Sequence A Computer Program for Aligning a cDNA Sequence with a Genomic DNA Sequence, *8*, 967–974.

Forde, B. G. (2002). Local and long-range signaling pathways regulating plant responses to nitrate. *Annual Review of Plant Biology*, *53*(50), 203–24. doi:10.1146/annurev.arplant.53.100301.135256

Franco-Zorrilla, J. M., Valli, A., Todesco, M., Mateos, I., Puga, M. I., Rubio-Somoza, I., … Paz-Ares, J. (2007). Target mimicry provides a new mechanism for regulation of microRNA activity. *Nature Genetics*, *39*(8), 1033–7. doi:10.1038/ng2079

Fukue, Y., Sumida, N., Tanase, J., & Ohyama, T. (2005). A highly distinctive mechanical property found in the majority of human promoters and its transcriptional relevance. *Nucleic Acids Research*, *33*(12), 3821–7. doi:10.1093/nar/gki700

Garg, R., Patel, R. K., Tyagi, A. K., & Jain, M. (2011). De novo assembly of chickpea transcriptome using short reads for gene discovery and marker identification. *DNA Research : An International Journal for Rapid Publication of Reports on Genes and Genomes*, *18*(1), 53–63.

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *Genomics (q-bio.GN); Quantitative Methods (q-bio.QM)*, 1–9.

Gibson, A. K., Smith, Z., Fuqua, C., Clay, K., & Colbourne, J. K. (2013). Why so many unknown genes? Partitioning orphans from a representative transcriptome of the lone star tick Amblyomma americanum. *BMC Genomics*, *14*, 135. doi:10.1186/1471-2164-14-135

Goff, S. a, Ricke, D., Lan, T.-H., Presting, G., Wang, R., Dunn, M., … Briggs, S. (2002). A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). *Science (New York, N.Y.)*, *296*(5565), 92–100. doi:10.1126/science.1068275

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. a, Amit, I., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–52.

Grivet, L., Hont, A. D., Dufour, P., Hamon, P., & Roquest, D. (1994). Comparative genome mapping of sugar cane with other species within the Andropogoneae tribe. *Heredity*, *73*, 500–508.

Guo, X., Gao, L., Liao, Q., Xiao, H., Ma, X., Yang, X., … Zhao, Y. (2013). Long non-coding RNAs function annotation: a global prediction method based on bi-colored networks. *Nucleic Acids Research*, *41*(2), e35. doi:10.1093/nar/gks967

Gutiérrez, R. a, Lejay, L. V, Dean, A., Chiaromonte, F., Shasha, D. E., & Coruzzi, G. M. (2007). Qualitative network models and genome-wide expression data define carbon/nitrogen-responsive molecular machines in Arabidopsis. *Genome Biology*, *8*(1), R7. doi:10.1186/gb-2007-8-1-r7

Hangauer, M. J., Vaughn, I. W., & McManus, M. T. (2013). Pervasive Transcription of the Human Genome Produces Thousands of Previously Unidentified Long Intergenic Noncoding RNAs. *PLoS Genetics*, *9*(6), e1003569. doi:10.1371/journal.pgen.1003569

Hansey, C. N., Vaillancourt, B., Sekhon, R. S., de Leon, N., Kaeppler, S. M., & Buell, C. R. (2012). Maize (Zea mays L.) genome diversity as revealed by RNA-sequencing. *PloS One*, *7*(3), e33071.

Henry, R., & Kole, C. (2010). *Genetics, Genomics and Breeding of Sugarcane*. (C. Henry, R. J.;Kole, Ed.) (1st ed.). Science Publishers.

Heo, J. B., Lee, Y.-S., & Sung, S. (2013). Epigenetic regulation by long noncoding RNAs in plants. *Chromosome Research : An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology*, *21*(6-7), 685–93. doi:10.1007/s10577-013-9392-6

Hoffmann, H. (2008). Variedades RB de cana-de-açúcar. CCA/UFSCarTechnical Bulletin1.

Hogarth, D. M., Ryan, C. C., & Taylor, P. W. J. (1993). Quantitative inheritance of rust resistance in sugarcane. *Field Crops Research*, *34*(2), 187–193.

Ideker, T., Galitski, T., & Hood, L. (2001). A new approach to decoding life: systems biology. *Annual Review of Genomics and …*. Retrieved from http://www.annualreviews.org/doi/abs/10.1146/annurev.genom.2.1.343

Ideker, T., Thorsson, V., Ranish, J. a, Christmas, R., Buhler, J., Eng, J. K., … Hood, L. (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science (New York, N.Y.)*, *292*(5518), 929–34. doi:10.1126/science.292.5518.929

Irvine, J. E. (1975). Relations of Photosynthetic Rates and Leaf and Canopy Characters to Sugarcane Yield. *Crop Science*, *15*, 671.

Iseli, C., Jongeneel, C., & Bucher, P. (1999). ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *ISMB*. Retrieved from

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.99.6106&rep=rep1&type=pdf

Iseli, C., Jongeneel, C. V., & Bucher, P. (1999). ESTScan: A Program for Detecting, Evaluating, and Reconstructing Potential Coding Regions in EST Sequences, 138–158.

Jacob, F. (1977). Evolution and tinkering. *Science*, *196*(4295), 1161–1166. Retrieved from http://adi-38.bio.ib.usp.br/ibi5023/2010/Jacob_1977.pdf

Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., … Wincker, P. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, *449*(7161), 463–7. doi:10.1038/nature06148

Jurka, J., & Kapitonov, V. (2005). Repbase Update, a database of eukaryotic repetitive elements. *… and Genome Research*. Retrieved from http://www.karger.com/Article/Pdf/84979

Kapusta, A., Kronenberg, Z., Lynch, V. J., Zhuo, X., Ramsay, L., Bourque, G., … Feschotte, C. (2013). Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs. *PLoS Genetics*, *9*(4), e1003470. doi:10.1371/journal.pgen.1003470

Kassube, S. a, Fang, J., Grob, P., Yakovchuk, P., Goodrich, J. a, & Nogales, E. (2013). Structural insights into transcriptional repression by noncoding RNAs that bind to human Pol II. *Journal of Molecular Biology*, *425*(19), 3639–48. doi:10.1016/j.jmb.2012.08.024

Khalil, A. M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., … Rinn, J. L. (2009). Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(28), 11667–72. doi:10.1073/pnas.0904715106

Kim, J., He, X., & Sinha, S. (2009). Evolution of regulatory sequences in 12 Drosophila species. *PLoS Genetics*, *5*(1), e1000330. doi:10.1371/journal.pgen.1000330

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*, 0–2. Retrieved from

http://www.genomics.arizona.edu/553/Readings/2012/Kimura1968.pdf

Kistner, C., & Matamoros, M. (2005). RNA ISOLATION USING PHASE EXTRACTION AND L I C L. In A. Márquez (Ed.), *Lotus japonicus Handbook* (Springers., pp. 123–124). Dordrecht, The Netherlands.

Knowles, D. G., & McLysaght, A. (2009). Recent de novo origin of human protein-coding genes. *Genome Research*, *19*(10), 1752–9. doi:10.1101/gr.095026.109

Koonin, E. (2004). A non-adaptationist perspective on evolution of genomic complexity or the continued dethroning of man. *CELL CYCLE-LANDES BIOSCIENCE-*, (March), 280–285. Retrieved from http://www.landesbioscience.com/journals/cc/kooninCC3-3.pdf?origin=publication_detail

Landell MGA, Campana MP, Figueiredo P, Vasconcelos ACM, Xavier MA, Bidoia MAP, Prado H, Silva MA, Miranda LLD, A. C. (2005). Variedades de cana-de-açúcar para o centro sul do Brasil.Technical Bulletin IAC 197.

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*(3), R25–R25.10.

Li, B., & Dewey, C. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. Retrieved from http://www.biomedcentral.com/content/pdf/1471-2105-12-323.pdf

Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, *12*, 323.

Li, D., Deng, Z., Qin, B., Liu, X., & Men, Z. (2012). De novo assembly and characterization of bark transcriptome using Illumina sequencing and development of EST-SSR markers in rubber tree (Hevea brasiliensis Muell. Arg.). *BMC Genomics*, *13*(1), 192.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25*(16), 2078–9.

Li, S.-W., Yang, H., Liu, Y.-F., Liao, Q.-R., Du, J., & Jin, D.-C. (2012). Transcriptome and gene expression analysis of the rice leaf folder, Cnaphalocrosis medinalis. *PloS One*, *7*(11), e47401.

Liao, Q., Liu, C., Yuan, X., Kang, S., Miao, R., Xiao, H., … Zhao, Y. (2011). Large-scale prediction of long non-coding RNA functions in a coding-non-coding gene co-expression network. *Nucleic Acids Research*, *39*(9), 3864–78. doi:10.1093/nar/gkq1348

Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, a H., & Ecker, J. R. (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, *133*(3), 523–36.

Liu, J., Jung, C., Xu, J., Wang, H., Deng, S., Bernad, L., … Chua, N.-H. (2012). Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in Arabidopsis. *The Plant Cell*, *24*(11), 4333–45. doi:10.1105/tpc.112.102855

Liu, M., Qiao, G., Jiang, J., Yang, H., Xie, L., Xie, J., & Zhuo, R. (2012). Transcriptome sequencing and de novo analysis for Ma bamboo (Dendrocalamus latiflorus Munro) using the Illumina platform. *PloS One*, *7*(10), e46766.

Liu, S., Li, W., Wu, Y., Chen, C., & Lei, J. (2013). De Novo Transcriptome Assembly in Chili Pepper (Capsicum frutescens) to Identify Genes Involved in the Biosynthesis of Capsaicinoids. *PloS One*, *8*(1), e48156. doi:10.1371/journal.pone.0048156

Lorenz, R., Bernhart, S. H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., & Hofacker, I. L. (2011). ViennaRNA Package 2.0. *Algorithms for Molecular Biology : AMB*, *6*(1), 26.

Lu, T., Lu, G., Fan, D., Zhu, C., Li, W., Zhao, Q., … Han, B. (2010). Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. *Genome Research*, *20*(9), 1238–49.

Luo, F., Yang, Y., Zhong, J., Gao, H., Khan, L., Thompson, D. K., & Zhou, J. (2007). Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory. *BMC Bioinformatics*, *8*, 299. doi:10.1186/1471-2105-8-299

Lynch, M. (2007). The evolution of genetic networks by non-adaptive processes. *Nature Reviews. Genetics*, *8*(10), 803–13. doi:10.1038/nrg2192

Lynch, M., & Conery, J. S. (2003). The origins of genome complexity. *Science (New York, N.Y.)*, *302*(5649), 1401–4. doi:10.1126/science.1089370

Ma, H.-M., Schulze, S., Lee, S., Yang, M., Mirkov, E., Irvine, J., … Paterson, A. (2004). An EST survey of the sugarcane transcriptome. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, *108*(5), 851–63.

Managadze, D., Lobkovsky, A. E., Wolf, Y. I., Shabalina, S. a, Rogozin, I. B., & Koonin, E. V. (2013). The vast, conserved mammalian lincRNome. *PLoS Computational Biology*, *9*(2), e1002917. doi:10.1371/journal.pcbi.1002917

Managadze, D., Rogozin, I. B., Chernikova, D., Shabalina, S. a, & Koonin, E. V. (2011). Negative correlation between expression level and evolutionary rate of long intergenic noncoding RNAs. *Genome Biology and Evolution*, *3*, 1390–404. doi:10.1093/gbe/evr116

Mancini, M. C., Leite, D. C., Perecin, D., Bidóia, M. a. P., Xavier, M. a., Landell, M. G. a., & Pinto, L. R. (2012). Characterization of the Genetic Variability of a Sugarcane Commercial Cross Through Yield Components and Quality Parameters. *Sugar Tech*, *14*(2), 119–125.

Marconi, T. G., Costa, E. A., Miranda, H. R., Mancini, M. C., Cardoso-Silva, C. B., Oliveira, K. M., … Souza, A. P. (2011). Functional markers for gene mapping and genetic diversity studies in sugarcane. *BMC Research Notes*, *4*(1), 264.

Marguerat, S., & Bähler, J. (2010). RNA-seq: from technology to biology. *Cellular and Molecular Life Sciences : CMLS*, *67*(4), 569–79.

McCormick, A. J., Cramer, M. D., & Watt, D. A. (2006). Sink strength regulates photosynthesis in sugarcane. *The New Phytologist*, *171*(4), 759–70.

Metzgar, D., Bytof, J., & Wills, C. (2000). Selection against frameshift mutations limits microsatellite expansion in coding DNA. *Genome Research*, *10*(1), 72–80.

Ming, R., Liu, S. C., Lin, Y. R., da Silva, J., Wilson, W., Braga, D., … Paterson, a H. (1998). Detailed alignment of saccharum and sorghum chromosomes: comparative organization of closely related diploid and polyploid genomes. *Genetics*, *150*(4), 1663–82.

Ministério da Agricultura, P. e A. (2013). Acompanhamento de safra brasileira : cana-de-açúcar Safra 2012/2013 Terceiro levantamento. *Companhia Nacional de Abastecimento*.

Moore, P. H., Botha, F. ., Furbank, R. ., & Grof, C. P. . (1996). *Intensive sugarcane production: Meeting the challenge beyond 2000*. (Keating BA and Wilson JR, Ed.) (p. 141). Oxon, UK: CAB International.

Morozova, O., & Marra, M. a. (2008). Applications of next-generation sequencing technologies in functional genomics. *Genomics*, *92*(5), 255–64.

Mutwil, M., Usadel, B., Schütte, M., Loraine, A., Ebenhöh, O., & Persson, S. (2010). Assembly of an interactive correlation network for the Arabidopsis genome using a novel heuristic clustering algorithm. *Plant Physiology*, *152*(1), 29–43. doi:10.1104/pp.109.145318

Novaes, E., Drost, D. R., Farmerie, W. G., Pappas, G. J., Grattapaglia, D., Sederoff, R. R., & Kirst, M. (2008). High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome. *BMC Genomics*, *9*, 312.

Ogasawara, J., & Morishita, S. (2003). A fast and sensitive algorithm for aligning ESTs to the human genome. *Journal of Bioinformatics and …*, *1*(2), 363–386. Retrieved from http://www.worldscientific.com/doi/abs/10.1142/S0219720003000058

Ouyang, S., & Buell, C. R. (2004). The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants. *Nucleic Acids Research*, *32*(Database issue), D360–3. doi:10.1093/nar/gkh099

Papini-Terzi, F. S., Rocha, F. R., Vêncio, R. Z. N., Felix, J. M., Branco, D. S., Waclawovsky, A. J., … Souza, G. M. (2009). Sugarcane genes associated with sucrose content. *BMC Genomics*, *10*, 120. doi:10.1186/1471-2164-10-120

Patel, R. K., & Jain, M. (2012). NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PloS One*, *7*(2), e30619.

Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., … Rokhsar, D. S. (2009). The Sorghum bicolor genome and the diversification of grasses. *Nature*, *457*(7229), 551–6. doi:10.1038/nature07723

Pinto, L. R., Oliveira, K. M., Ulian, E. C., Garcia, A. A. F., & de Souza, A. P. (2004). Survey in the sugarcane expressed sequence tag database (SUCEST) for simple sequence repeats. *Genome / National Research Council Canada = Génome / Conseil National de Recherches Canada*, *47*(5), 795–804.

Quint, M., Drost, H.-G., Gabel, A., Ullrich, K. K., Bönn, M., & Grosse, I. (2012). A transcriptomic hourglass in plant embryogenesis. *Nature*, *490*(7418), 98–101. doi:10.1038/nature11394

Ramu, P., Kassahun, B., Senthilvel, S., Ashok Kumar, C., Jayashree, B., Folkertsma, R. T., … Hash, C. T. (2009). Exploiting rice-sorghum synteny for targeted development of EST-SSRs to enrich the sorghum genetic linkage map. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, *119*(7), 1193–204.

Rolland, F., Baena-Gonzalez, E., & Sheen, J. (2006a). Sugar sensing and signaling in plants: conserved and novel mechanisms. *Annual Review of Plant Biology*, *57*, 675–709. doi:10.1146/annurev.arplant.57.032905.105441

Rolland, F., Baena-Gonzalez, E., & Sheen, J. (2006b). Sugar sensing and signaling in plants: conserved and novel mechanisms. *Annual Review of Plant Biology*, *57*, 675–709. doi:10.1146/annurev.arplant.57.032905.105441

Rook, F., Hadingham, S. a., Li, Y., & Bevan, M. W. (2006). Sugar and ABA response pathways and the control of gene expression. *Plant, Cell and Environment*, *29*(3), 426–434. doi:10.1111/j.1365-3040.2005.01477.x

Rutter, M. T., Cross, K. V, & Van Woert, P. a. (2012). Birth, death and subfunctionalization in the Arabidopsis genome. *Trends in Plant Science*, *17*(4), 204–12. doi:10.1016/j.tplants.2012.01.006

Sabino, J. (1997). Sexta geração de variedades de cana-de-açúcar. COOPERATIVA DE PRODUTORES DE CANA, AÇÚCAR E ÁLCOOL DO ESTADO DE SÃO PAULO LTDA. Technical Bulletin. Piracicaba, SP.

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., … Jackson, S. a. (2010). Genome sequence of the palaeopolyploid soybean. *Nature*, *463*(7278), 178–83. doi:10.1038/nature08670

Seffens, W., & Digby, D. (1999). mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. *Nucleic Acids Research*, *27*(7), 1578–84. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=148359&tool=pmcentrez&rendertype=abstract

Silveira, A. B., Trontin, C., Cortijo, S., Barau, J., Del Bem, L. E. V., Loudet, O., … Vincentz, M. (2013). Extensive natural epigenetic variation at a de novo originated gene. *PLoS Genetics*, *9*(4), e1003437. doi:10.1371/journal.pgen.1003437

Smale, S. T., & Kadonaga, J. T. (2003). The RNA polymerase II core promoter. *Annual Review of Biochemistry*, *72*, 449–79. doi:10.1146/annurev.biochem.72.121801.161520

Sun, J., Zhou, M., Mao, Z.-T., Hao, D.-P., Wang, Z.-Z., & Li, C.-X. (2013). Systematic analysis of genomic organization and structure of long non-coding RNAs in the human genome. *FEBS Letters*, *587*(7), 976–82. doi:10.1016/j.febslet.2013.02.036

Tai, P. Y. P., Miller, J. D., & Dean, J. L. (1981). INHERITANCE OF RESISTANCE TO RUST IN SUGARCANE. *Field Crops Research*, *4*, 261–268.

Tautz, D., & Domazet-Lošo, T. (2011). The evolutionary origin of orphan genes. *Nature Reviews. Genetics*, *12*(10), 692–702. doi:10.1038/nrg3053

Tomato, T., & Consortium, G. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, *485*(7400), 635–41. doi:10.1038/nature11119

Trick, M., Long, Y., Meng, J., & Bancroft, I. (2009). Single nucleotide polymorphism (SNP) discovery in the polyploid Brassica napus

using Solexa transcriptome sequencing. *Plant Biotechnology Journal*, *7*(4), 334–46.

Ulitsky, I., & Bartel, D. P. (2013). lincRNAs: Genomics, Evolution, and Mechanisms. *Cell*, *154*(1), 26–46. doi:10.1016/j.cell.2013.06.020

Varshney, R. K., Nayak, S. N., May, G. D., & Jackson, S. a. (2009). Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends in Biotechnology*, *27*(9), 522–30.

Vettore, A. L., da Silva, F. R., Kemper, E. L., Souza, G. M., da Silva, A. M., Ferro, M. I. T., … Arruda, P. (2003). Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane. *Genome Research*, *13*(12), 2725–35.

Vettore, A. L., Silva, F. R., Kemper, E. L., & Arruda, P. (2001). The libraries that made SUCEST, *24*, 1–7.

Vicentini, R., Bem, L. E. V., Sluys, M. a., Nogueira, F. T. S., & Vincentz, M. (2012). Gene Content Analysis of Sugarcane Public ESTs Reveals Thousands of Missing Coding-Genes and an Unexpected Pool of Grasses Conserved ncRNAs. *Tropical Plant Biology*, *5*(2), 199–205. doi:10.1007/s12042-012-9103-z

Vincentz, M., Cara, F. A. A., Okura, V. K., da Silva, F. R., Pedrosa, G. L., Hemerly, A. S., … Menck, C. F. M. (2004). Evaluation of monocot and eudicot divergence using the sugarcane transcriptome. *Plant Physiology*, *134*(3), 951–9.

Wakeley, J. (1996). The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance. *Tree*, *11*(4), 158– 162.

Wolf, Y. I., Novichkov, P. S., Karev, G. P., Koonin, E. V, & Lipman, D. J. (2009). The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(18), 7273–80. doi:10.1073/pnas.0901808106

Yu, J., Hu, S., Wang, J., Wong, G. K.-S., Li, S., Liu, B., … Yang, H. (2002). A draft sequence of the rice genome (Oryza sativa L. ssp. indica). *Science (New York, N.Y.)*, *296*(5565), 79–92. doi:10.1126/science.1068037

Zhai, J., Jeong, D., Paoli, E. De, Park, S., Rosen, B. D., Yan, Z., … Meyers, B. C. (2011). MicroRNAs as master regulators of the plant NB-LRR defense gene family via the production of phased , trans -acting siRNAs, 2540–2553. doi:10.1101/gad.177527.111.infection

Zhang, B., & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, *4*, Article17. doi:10.2202/1544-6115.1128

Zhao, L., Saelao, P., Jones, C. D., & Begun, D. J. (2014). Origin and spread of de novo genes in Drosophila melanogaster populations. *Science (New York, N.Y.)*, *343*(6172), 769–72. doi:10.1126/science.1248286

Zhou, Q., Zhang, G., Zhang, Y., Xu, S., Zhao, R., Zhan, Z., … Wang, W. (2008). On the origin of new genes in Drosophila. *Genome Research*, *18*(9), 1446–55. doi:10.1101/gr.076588.108

Zhu, Q.-H., & Wang, M.-B. (2012). Molecular Functions of Long Non-Coding RNAs in Plants. *Genes*, *3*(4), 176–190. doi:10.3390/genes3010176

Anexos

Anexo 1

Artigo publicado na revista PLOS ONE
(DOI:10.1371/journal.pone.0088462)

De novo assembly and transcriptome analysis of contrasting sugarcane varieties

Claudio Benicio Cardoso-Silva[1☼], Estela Araujo Costa[1☼], Melina Cristina Mancini[1], Thiago Willian Almeida Balsalobre[1], Lucas Eduardo Costa Canesin[1], Luciana Rossini Pinto[2], Monalisa Sampaio Carneiro[3], Antonio Augusto Franco Garcia[4], Anete Pereira de Souza[1,5], Renato Vicentini[1§]

[1]Center for Molecular Biology and Genetic Engineering (CBMEG), University of Campinas (UNICAMP), Campinas, SP, Brazil

[2]Centro Avançado da Pesquisa Tecnológica do Agronegócio de Cana (IAC/Apta), Ribeirão Preto, SP, Brazil

[3]Departamento de Biotecnologia e Produção Vegetal e Animal, Centro de Ciências Agrárias, Universidade Federal de São Carlos, Araras, SP, Brazil

[4]Departamento de Genética, Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, Piracicaba, SP, Brazil

[5]Departamento de Biologia Vegetal, Instituto de Biologia, Universidade Estadual de Campinas (UNICAMP), Campinas, SP, Brazil

[§]Corresponding author

Email Address: shinapes@unicamp.br

[☼]These authors contributed equally to this work.

Abstract

Sugarcane is an important crop and a major source of sugar and alcohol. In this study, we performed *de novo* assembly and transcriptome annotation for six sugarcane genotypes involved in bi-parental crosses. The *de novo* assembly of the sugarcane transcriptome was performed using short reads generated using the Illumina RNA-Seq platform. We produced more than 400 million reads, which were assembled into 72,269 unigenes. Based on a similarity search, the unigenes showed significant similarity to more than 28,788 sorghum proteins, including a set of 5,272 unigenes that are not present in the public sugarcane EST databases; many of these unigenes are likely putative undescribed sugarcane genes. From this collection of unigenes, a large number of molecular markers were identified, including 5,106 simple sequence repeats (SSRs) and 708,125 single-nucleotide polymorphisms (SNPs). This new dataset will be a useful resource for future genetic and genomic studies in this species.

Keywords

Sugarcane; RNA-Seq; *de novo* assembly; transcriptome; SNP; lncRNAs

Background

Sugarcane belongs to the grass family (Poaceae), which is an economically important seed plant family that includes maize, wheat, rice, sorghum and many types of grasses. The sugarcane crop is the main source of both sugar and alcohol, accounting for two-thirds of the world's sugar production (D. L. Carson & Botha, 2000a). It is estimated that approximately 653.81 million tons of sugarcane will be produced during the 2013/2014 harvest in Brazil, surpassing the production of the last harvest (Ministério da Agricultura, 2013).

Modern sugarcane varieties are derived from interspecific hybridization between *Saccharum officinarum* and *Saccharum spontaneum*, resulting in highly polyploid and aneuploid plants. Indeed, the chromosome number of these varieties ranges from 80 to 140. Modern varieties of sugarcane typically exhibit more than eight homologous copies of each basic chromosome from *S. officinarum* and several

copies of the homologous chromosomes from *S. spontaneum* (Ming et al., 1998). Therefore, sugarcane cultivars are highly heterozygous, presenting several different alleles at each locus, and this high level of genetic complexity creates challenges during conventional and molecular breeding programs.

Recent technological developments have the potential to greatly increase our understanding of sugarcane plants through the application of emerging genomic technologies, and the use of next-generation sequencing (NGS) technologies could have significant implications for crop genetics and breeding. Although the sequencing of large genomes remains expensive, even using NGS technologies (S.-W. Li et al., 2012), transcriptome sequencing can provide information regarding the gene content of a species and can complement genome sequencing approaches.

RNA sequencing (RNA-Seq) has been applied as a tool for transcriptome analysis in many species, such as *Arabidopsis thaliana* (Lister et al., 2008), *Brassica* spp. (Trick, Long, Meng, & Bancroft, 2009), rice (Lu et al., 2010) and maize (Hansey et al., 2012). RNA-Seq has several advantages, including (i) allowing more precise measurement of the levels of transcripts and their isoforms than other methods, (ii) presenting the potential for the development of SNPs that can be used to detect allele-specific expression because the same base is sequenced multiple times, (iii) the ability to identify reads containing post-transcriptional modifications or rearranged sequences that cannot be mapped directly to the genome (Marguerat & Bähler, 2010) and (iv) allowing the identification of species-specific genes (Morozova & Marra, 2008). Moreover, the availability of a large number of genetic markers developed using NGS technologies is facilitating trait mapping and marker-assisted breeding (Varshney, Nayak, May, & Jackson, 2009).

In plant breeding programs, genotypes of interest to breeders, such as the parental genotypes of mapping populations, can be sequenced using NGS technologies. More than one genotype can be employed to generate sequence data with these technologies, and these data can be aligned using genome or transcriptome sequencing data for model or major crop species that are closely related to the species of interest (Varshney et al., 2009). This approach has also been applied for marker discovery in some crop species, such as eucalyptus

(Novaes et al., 2008), maize (Barbazuk, Emrich, Chen, Li, & Schnable, 2007) and chickpea, and has been used to identify SNPs between the parental genotypes of mapping populations. These SNPs can then be employed to develop markers for marker-deficient crops to allow trait mapping through marker-assisted selection (MAS).

Despite its economic importance, no published genome sequence is currently available for sugarcane. Instead, the basic resource used for the study of sugarcane gene sequences is the substantial expressed sequence tag (EST) information available in public databases. Transcriptome studies in sugarcane began in South Africa (D. Carson & Botha, 2002; D. L. Carson & Botha, 2000b), and the largest EST collection (~238,000 ESTs) was developed through the Brazilian SUCEST project (Vettore et al., 2003; Vettore, Silva, Kemper, & Arruda, 2001). Researchers in Australia (Bower et al., 2005; Casu et al., 2003, 2004) and the USA (Ma et al., 2004) have generated three additional libraries containing 10,000 ESTs each. Currently, all of the reported ESTs are collected in the Sugarcane Gene Index, version 3.0, which contains 282,683 ESTs and 499 complete cDNA sequences, resulting in 121,342 unique assembled sequences, or unigenes. There are still more than 10,000 sugarcane coding genes that have yet to be identified (Vicentini, Bem, Sluys, Nogueira, & Vincentz, 2012), highlighting the need for new sequencing efforts in the sugarcane transcriptome. This information would increase the panel of potential molecular markers and sequence information available for sugarcane breeding programs, resulting in biotechnological improvements. In the present study, using the Illumina GA IIx sequencing platform, we performed *de novo* transcriptome sequencing in six sugarcane genotypes that are employed as parents in Brazilian Sugarcane Breeding Programs. We identified conserved genes that have not previously been described in sugarcane, and these data will be useful for future genome assembly and marker identification.

Materials and Methods

Ethics Statement

We confirm that no specific permits were required for the described field studies. This work was a collaborative research project developed by researchers from UNICAMP, ESALQ/USP, IAC/Apta (Instituto Agronômico de Campinas) and UFSCar-RIDESA (Universidade Federal de São Carlos-Rede Interinstitucional de Desenvolvimento do Setor Sucroalcooleiro) (all from Brazil). We also confirm that the field studies did not involve endangered or protected species.

Plant Materials and RNA Extraction

Six genotypes were included in this study. IACSP96-3046 and IACSP95-3018 are the parents of a mapping population from the Sugarcane Breeding Program at IAC/Apta. IACSP95-3018 is a promising clone that is also used as a parent in the breeding program. IACSP93-3046 is a variety that exhibits good tillering, an erect stool habit and resistance to rust.

SP81-3250 x RB925345 and SP80-3280 x RB835486 are the parents of two different mapping populations from the Sugarcane Breeding Program at UFSCar, which is part of RIDESA. These parents exhibit contrasting properties: SP81-3250 and SP80-3280 are resistant to rust (Bellodi & Macedo, 1995; Sabino, 1997), whereas RB925345 and RB835486 are susceptible (Hoffmann, 2008). All of the examined genotypes display high levels of sucrose.

Leaves at the third position (McCormick et al., 2006) were collected from one plant per genotype and immediately frozen, and total RNA was extracted using a modified protocol (Kistner & Matamoros, 2005). The integrity and quantity of the isolated RNA were assessed using a 2100 Bioanalyzer (Agilent). Equal quantities of high-quality RNA from each genotype were pooled for cDNA synthesis.

mRNA-Seq Library Construction for Illumina Sequencing

Paired-end Illumina mRNA libraries were generated from 4 μg of total RNA in accordance with the manufacturer's instructions for mRNA-Seq Sample Preparation

(Illumina Inc., San Diego, CA, USA). The quality of the library was assessed using a 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, USA).

Cluster amplification was performed using the TruSeq PE Cluster Kit and a cBot (Illumina), and each sample was sequenced in a separate GAIIx lane using the TruSeq SBS 36 Cycle Kit (Illumina). The read length was 72 bp.

Sequence Data Analysis and Assembly

The raw data generated by Illumina sequencing were converted from the BCL format to qSeq using Off-line Basecaller, v.1.9.4 (OLB) software. The qSeq files were transformed in FastQ files, which contain sequences that are 72 bp in length, using a custom script. Low-quality sequences were removed; these sequences included reads with ambiguous bases, reads with less than 70 bases, and reads with a Phred quality score Q ≤ 20 using the NGS QC toolkit (Patel & Jain, 2012). All reads were deposited in the National Center for Biotechnology Information (NCBI) database and can be found under accession number SRA073690.

All datasets were combined, and the sequenced reads were assembled using Trinity (http://trinityrnaseq.sourceforge.net/), which is a program developed specifically for *de novo* transcriptome assembly from short-read RNA-Seq data that recovers transcript isoforms efficiently and sensitively using the de Bruijn graph algorithm (Grabherr et al., 2011). The optimal assembly results were chosen according to an evaluation of the assembly encompassing the total number of contigs, the distribution of contig lengths, the N50 statistic and the average coverage. The assembled transcripts were based on the main isoform of each transcript, and only contigs with lengths of greater than 300 bp were included in the downstream analysis.

To identify the genotypic contribution to each transcript, reads from each library were mapped against the assembly generated from all libraries using the bowtie aligner (Langmead et al., 2009). The BAM files generated by bowtie were then used to estimate the transcript-level abundance for each library using the RSEM (RNA-Seq by Expectation Maximization) software (Bo Li & Dewey, 2011).

Functional Annotation of Sugarcane Transcripts

The assembled sequences were compared against the NCBI non-redundant protein database (NR) using BLASTX with a cut-off E-value of 10^{-6}. To annotate the assembled sequences according to Gene Ontology (GO) terms (The Gene Ontology Consortium, 2000), the above BLAST results were analyzed using Blast2GO (Conesa et al., 2005) to determine and compare gene functions. The GO terms were assigned to the representative transcripts for each sample through an enrichment analysis using Fisher's exact test (p-value < 0.01), with a false discovery rate (FDR) correction in terms of biological processes and molecular functions. The transcript sequences were also aligned against the *Viridiplantae*, grass and sorghum protein databases (http://www.phytozome.org/) using BLASTX and against the Sugarcane Gene Index (http://compbio.dfci.harvard.edu/tgi/) using BLASTN; in both alignments, a cut-off E-value of 10^{-6} was applied. The BLAST search was limited to the first ten significant query hits, and the gene names were assigned to each query based on the highest score. Transcripts that showed similarity to *Viridiplantae* proteins were aligned against the sorghum genome using sim4 software (Florea, Hartzell, Zhang, Rubin, & Miller, 1998). Open reading frames (ORFs) were predicted using a script available in the TransDecoder package (http://transdecoder.sourceforge.net/), with 300 bp as the minimum ORF length. Those transcripts showing predicted ORFs were aligned against grass proteins using the STRING database, v.9.05 (http://string-db.org), to predict Clusters of Orthologous Groups (COG).

To further characterize the subset of unigenes that did not show similarity to any known plant proteins, we applied a computational strategy to mine putative long non-coding RNA (lncRNA) data. We first aligned all 121,342 EST unigenes to *Viridiplantae* proteins and to the GenBank NR database using BLASTX. Those EST unigenes that did not align with any proteins were then mapped to the *Sorghum bicolor* genome, obtaining at least 70% coverage and a maximum intron size of 15 kb. The coding probability of the positively mapped unigenes was then evaluated by removing sequences with potential ORFs longer than 100 aa using ESTScan (Christian Iseli, Jongeneel, & Bucher, 1999). We further investigated the functional role of the remaining unigenes and putative lncRNAs by searching for three indirect

indications of functionality: we examined the stability of the secondary structure using the Vienna package (Lorenz et al., 2011), normalized to the Z-score index (Clote et al., 2005); we mapped the small RNAs (sRNAs) (Domingues et al., 2012) against sugarcane unigenes; and we analyzed the sequence similarities between the unigenes and *S. bicolor* ESTs (BLASTN, E-value ≤ $1e^{-5}$). Only EST unigenes with at least one indirect piece of functional evidence were analyzed further. The putative lncRNAs were then aligned to the 18,910 assembled transcripts that showed no similarity to any plant protein but were successfully mapped to *S. bicolor* (Text S4). Only hits with an E-value below $1e^{-5}$ and coverage higher than 40% were considered positive.

Putative Molecular Markers

We utilized the MISA program (http://pgrc.ipk-gatersleben.de/misa/) to search for simple sequence repeat (SSR) motifs in the unigenes; the MISA script can identify both perfect and compound (interrupted by a certain number of bases) motifs. To identify the presence of SSRs, only motifs of two to six nucleotides were considered, and the minimum repeat unit was defined as six for dinucleotide motifs and five for tri-, tetra-, penta- and hexanucleotide motifs. A compound motif was defined as two or more SSR motifs interrupted by sequences of up to 100 bp.

To identify putative single-nucleotide polymorphisms (SNPs) in the sugarcane transcript assembly, we first separately mapped all of the short reads from each library to the assembly using the Burrows-Wheeler Aligner (BWA). Next, FreeBayes (Garrison & Marth, 2012) and SAMtools (H. Li et al., 2009) were used to detect the variable positions of SNPs from the consensus sugarcane assembly. The FreeBayes tool allowed us to identify genetic variants in the polyploid organisms. The putative SNPs were then filtered using the varFilter command, where variants were called only for positions with a minimal mapping quality (-Q) and coverage (-d) of 25. To compare the composition of the SNP variation in the parental genotype, unique and shared SNPs were extracted using an in-house script. The transition and transversion ratios were calculated using the tstv tool developed by SnpSift software (Cingolani et al., 2012).

Results and Discussion

De novo assembly of the sugarcane transcriptome

The libraries sequenced using the Illumina platform produced a total of 610,232,490 paired-end (PE) sequence reads, each of which was 72 bp in length. We filtered the sequence data for low-quality reads, resulting in 445,374,504 high-quality PE trimmed reads (97.67%), which were used to obtain the *de novo* assembly. An overview of the sequencing procedure is presented in Table 1. The *de novo* assembly generated 119,768 transcripts when all isoforms were considered. These transcripts represent a total of 72,269 unigenes that were considered for downstream analysis (Text S1). The length of the unigenes ranged from 300 bp to ~7 kb, with a mean length of 921 bp, an N50 equal to 1,367 bp and 46.39% GC content. The average length of the assembled unigenes was greater than those obtained from chickpea (523 bp) (Garg et al., 2011), rubber trees (485 bp) (D. Li, Deng, Qin, Liu, & Men, 2012) and bamboo (736 bp) (M. Liu et al., 2012) using similar sequencing technologies. Considering the N50 values, the values for the sugarcane unigenes were greater than those for rubber trees (592 bp), bamboo (1,132 bp) and chili pepper (1,076 bp) (S. Liu, Li, Wu, Chen, & Lei, 2013), which were also assembled using short reads generated by the Illumina platform. In total, we obtained 18,624 (27.21%) unigenes longer than 1 kb and 7,657 (10.6%) unigenes longer than 2 kb. The length distributions of the unigenes are shown in Table 2, revealing that more than 40,000 unigenes (55.76%) were longer than 500 bp. These unigenes were submitted to an ORF predictor using TransDecoder, and we detected 33,673 (46.59%) unigenes with ORFs, with 9,350 (12.94%) presenting complete ORFs.

Unigene annotation

The 72,269 sugarcane unigenes were analyzed for sequence similarity against the *Viridiplantae* (comprising all green plants) and grass (*S. bicolor, Oryza sativa, Zea mays, Panicum virgatum, Setaria italica* and *Brachypodium virgatum*) datasets

through BLASTX searches. The unigenes were also compared against the sugarcane EST database via a BLASTN search (Table 3). A total of 35,456 (49.06%) unigenes showed significant similarity to *Viridiplantae*. The high percentage of sugarcane unigenes obtained in this study that did not match the *Viridiplantae* protein database (50.84%) indicates that there is potential for the discovery of as-yet-undescribed and novel genes in sugarcane, although most of these unigenes may encode non-coding RNAs. In fact, more than 26% of the unigenes in this set exhibited high similarity to intergenic regions of the sorghum genome (Figure 1). Additionally, the significance of a BLAST search depends on the length of the query sequence; therefore, short sequences are rarely matched to known genes (Novaes et al., 2008), or these sequences may represent rapidly evolving sequences that have diverged substantially from their homologs (Vincentz et al., 2004).

In turn, alignment of the unigenes against the grass protein database returned 34,814 significant hits. When considering the hits by species, 28,788 unigenes showed significant similarity to sorghum, corresponding to 98% of sorghum proteins (Figure 1). These results were expected, as comparative genomic studies (Grivet, Hont, Dufour, Hamon, & Roquest, 1994) have revealed conservation and synteny among the sugarcane and sorghum genomes. The sugarcane transcriptome also significantly matched that of rice, with approximately 29,285 unigenes (corresponding to 28,732 unique protein accessions) showing significant similarity to rice proteins.

To investigate previously unidentified potential genes in sugarcane, we compared the unigenes against the sugarcane transcripts deposited in public databases and performed BLAST searches to detect possible similarities with the SoGI database (*S. officinarum*). Furthermore, the unigenes that did not show similarity to sugarcane ESTs were compared against sorghum proteins. Approximately 22,171 unigenes exhibited significant similarity to sorghum proteins and sugarcane transcripts (Figure 1). The remaining 5,272 unigenes (Text S3) showed significant similarity to sorghum and rice proteins but not to the sugarcane transcripts that were considered to be putative new sugarcane genes (Figure 1). By examining the presence of candidate coding regions in these unigenes, we identified

4,895 sequences that contained ORFs, with 732 unigenes containing complete ORFs. These unigenes represent genes that have not yet been described for sugarcane.

Clusters of Orthologous Groups (COG) classification

COG classification was performed for the transcriptome data, and a total of 7,519 unigenes were identified (Figure 2). These unigenes were classified into 23 COG categories, with the largest number of unigenes being grouped in the 'replication, recombination and repair' cluster (20.49%), followed by the 'general function prediction only' cluster (17.05%) and the 'posttranslational modification, protein turnover and chaperones' cluster (7.39%). These three categories are the same categories that are highly represented in sorghum (Figure 2).

A total of 19 of the 23 COG categories were present in the transcriptome data, and at least 60% of the sugarcane unigenes were annotated when compared with the annotation of sorghum genes in the COG categories.

The categories 'energy production and conversion' (3.72%), 'carbohydrate transport and metabolism' (5%) and 'defense mechanisms' (2%) exhibited at least 56% of the expected genes compared with the sorghum genes. These categories should be considered to represent gene sequences showing a high potential for the development of molecular markers in sugarcane breeding programs. Therefore, the likelihood of these markers being associated with agronomic traits of interest in QTL mapping and marker-assisted selection (MAS) (Dekkers & Hospital, 2002) is increased.

Gene Ontology enrichment analyses

The identification of functional classes that differ statistically between two lists of terms is a typical data-mining approach applied in functional genomics research (Conesa et al., 2005). In this work, we were interested in identifying which functions were distinctly represented among the different sugarcane genotypes. A total of 14,983 unigenes (Text S2) were annotated based on BLAST matches to known proteins in the NR database and were assigned to GO classes representing 39

terms, including some (10) that contain important information related to the enriched genotype (Figure 3).

Genes responsible for disease resistance, corresponding to the categories ‘signaling,’ ‘response to stimulus,’ ‘cellular response to stimulus,’ ‘response to chemical stimulus’ and ‘response to auxin stimulus’, were enriched in the SP81-3250, SP80-3280 and IACSP93-3046 genotypes, with IACSP93-3046 being represented in all of these categories (Figure 3). These three genotypes exhibit resistance to rust (Bellodi & Macedo, 1995; Landell MGA, Campana MP, Figueiredo P, Vasconcelos ACM, Xavier MA, Bidoia MAP, Prado H, Silva MA, Miranda LLD, 2005; Sabino, 1997), whereas the other genotypes, RB925345, RB835486 and IACSP95-3018, are susceptible to rust (Hoffmann, 2008; Mancini et al., 2012). Common sugarcane rust, caused by the fungus *Puccinia melanocephala*, is a disease that occurs worldwide and can result in large losses of sugar tonnage in susceptible varieties (Daugrois et al., 1996). Rust resistance is generally considered to be a quantitatively inherited trait showing a high degree of heritability and a strong additive genetic variance component (Hogarth, Ryan, & Taylor, 1993; Tai, Miller, & Dean, 1981).

The obtained enriched terms suggest that these three genotypes harbor transcripts that are involved in stimulus response pathways and probable disease responses. These results are correlated with the characteristics of resistance and susceptibility in these varieties.

Another important characteristic of sugarcane crops is their accumulation of sucrose. Wild sugarcane species produce less than 4% fresh weight of sucrose, whereas high-yield varieties can produce sucrose contents of up to 20% of their fresh weight (Irvine, 1975). The major differences between these varieties is based on sugar transport and metabolism in storage tissues (Moore, Botha, Furbank, & Grof, 1996). The entire network involving sucrose synthesis, accumulation, storage and retention is a complex system in which several metabolic pathways interact with each other (Henry & Kole, 2010). The most important aspect of this network is transport, which chiefly involves specific carrier molecules, ion transport and active transport and depends on the amount of available ATP. Within this context, we

observed some genotypes that were enriched in categories related to this network, particularly the transport process. These categories included 'organic substance transport' (SP81-3250, RB925345, SP80-3280, IACSP96-3046 and IACSP95-3018), 'substrate-specific transporter activity', 'substrate-specific transmembrane transporter activity' (SP81-3250 and SP80-3280), 'ion transmembrane transport' (SP81-3250 and IACSP93-3046) and 'transporter activity' (SP81-3250, SP80-3280, and IACSP93-3046).

Important categories involved in sugar transport and metabolism in storage tissues include the 'monosaccharide metabolic process,' 'glucose metabolic process,' 'small molecule biosynthetic process' and 'small molecule metabolic process' categories. The terms in the first and second categories were only enriched in the SP81-3250 genotype, whereas the terms in the third category were enriched in both the IACSP93-3046 and IACSP95-3018 genotypes. All genotypes showed enrichment in the last category, although SP80-3280 was the least represented.

All of the genotypes were enriched for transcripts involved in this complex network of sucrose synthesis, accumulation, storage and retention, and these results were corroborated by the agronomic characteristics of the plants. All of these genotypes produce high levels of sucrose, in accordance with the agronomic description of the genotypes SP81-3250 (Bellodi & Macedo, 1995), RB925345, RB835486, SP80-3280, IACSP93-3046 (Landell MGA, Campana MP, Figueiredo P, Vasconcelos ACM, Xavier MA, Bidoia MAP, Prado H, Silva MA, Miranda LLD, 2005) and IACSP95-3018.

Putative lncRNAs

Among the initial set of 121,342 EST retrieved unigenes, 23,529 showed no similarity to any known plant protein. These unigenes were mapped to the *S. bicolor* genome, resulting in 4,476 positive hits, with only 1,884 not exhibiting an ORF or presenting an ORF shorter than 100 aa. This subset comprised the putative sugarcane lncRNAs that are publicly available. We found that for ~4% of these sequence, there were small RNAs (sRNAs) that mapped to their sequence, with ~59% showing similarity to *S. bicolor* and ~39% showing a highly stable secondary

structure. In total, 1,446 non-redundant putative lncRNAs were identified that showed indirect evidence of functionality (Figure S1). We then compared this inclusive set (1,884 sequences) with the 18,910 assembled transcripts that lacked similarity to plant proteins. We observed 358 putative lncRNAs represented among the assembled transcripts, with ~42% of these sequences showing a highly stable secondary structure and ~40% showing evidence of transcription in the *S. bicolor* EST dataset. None of the unigenes to which sRNAs were mapped were similar to any assembled transcript. Finally, we compared the expression profiles of the putative lncRNAs between the different genotypes, which suggested that these transcripts may display genotype-specific expression patterns, as shown in Figure 4. A hierarchical clustering analysis revealed a pattern of separation between the genotypes from the different breeding programs, a result that is in accordance with the observation that the varieties from the same breeding program have the same genetic basis. We observed that the plant lncRNAs may display elevated intraspecific variation in expression, and several recent works have demonstrated that these transcripts exhibit tissue- and cell-specific expression patterns (Derrien et al., 2012; Guo et al., 2013; Hangauer et al., 2013; J. Liu et al., 2012). This study adds information regarding the dynamic involvement of these transcripts and reveals putative targets for further investigation (Kapusta et al., 2013; Sun et al., 2013).

Marker discovery

SSR discovery

Expressed sequence tag/simple sequence repeat (EST-SSR) markers are well established as important tools for researchers assessing genetic diversity and are useful in the development of genetic maps, comparative genomics and MAS breeding. Thus, the unigene sequences were searched for repeat motifs to explore the SSR profiles in the sugarcane transcriptome. A total of 5,106 SSRs were obtained from 4,616 unigene sequences (7.96%), and 576 of the unigenes contained more than one SSR (Text S7). Of these unigenes, 189 exhibited compound SSR formation. Trinucleotide repeat motifs were the most abundant,

accounting for 2,585 SSRs (50.63%) in 2,318 unigene sequences; dinucleotide repeat motifs accounted for 1,927 SSRs (37.74%) in 1,732 unigenes; and other motifs accounted for 594 SSRs (11.63%) in 1,708 unigenes (Table 4). The relative percentage of the sequences containing SSRs was higher than that obtained in the SUCEST (Sugarcane Expressed Sequence Tag database) study, in which 2,005 clusters containing SSRs were found among 43,141 clusters (4.64%) (Pinto, Oliveira, Ulian, Garcia, & de Souza, 2004).

The most abundant motifs included the dinucleotide AG motif (49.9%) and the trinucleotide CCG (17%) and ACC (4.7%) motifs. These results are similar to those of the SSR motif analysis performed in sorghum (Ramu et al., 2009). Additionally, CCG and ACC were the most commonly found motifs in the SUCEST study (Pinto et al., 2004), and CCG was the motif that was identified most often by Cordeiro *et al*. (Cordeiro, Casu, McIntyre, Manners, & Henry, 2001). The most frequent tetranucleotide motif found in the present study was AAAG. The overall frequency of SSRs was observed to be 1/1.6 kb.

The prevalence of trimeric motifs over other SSR repeats may be explained based on the risk of frameshift mutations that may occur when microsatellites alternate in size (Metzgar, Bytof, & Wills, 2000). Furthermore, a large number of trinucleotide coding repeats appear to be controlled primarily by mutation pressure.

The development of SSR markers associated with important agronomic traits can be used to assist in the selection of varieties during the early stages of MAS breeding programs and can be helpful in the selection of the best parents for crossing (Marconi et al., 2011). Consequently, the application of such markers supports breeding programs by significantly reducing the time and cost involved in developing new varieties and can help bypass barriers in sugarcane breeding programs.

SNP discovery

A total of 708,125 putative SNP positions were identified (Text S5), with a density of 1 SNP per 86 bp. The frequency of SNPs found in the sugarcane genes was higher than has been observed in other grasses, such as rice and sorghum, which exhibit a frequency of ≥1 SNP per 300 bp (Feltus et al., 2004). The observed number

of transitions was 456,666, and 254,658 transversions were detected, with the number of the former being 1.79 times that of the latter. Transitions were most likely more frequent because they are more tolerated by natural selection as the tendency to generate synonymous mutations in coding sequences is related to the number of transversions (Wakeley, 1996).

We identified SNPs in 58,903 different unigenes, which represent 81.50% of the total unigenes. Considering the number of unigenes without SNPs, we verified that 10,516 (79%) are unigenes with a length of less than 500 bp. Considering only those unigenes with predicted ORFs (33,673 unigenes), we found a total of 289,969 SNPs (37.5% of the total detected SNPs).

To detect different heterozygous SNPs between the parents from each mapping population, the reads from each genotype were mapped against all the unigenes (Text S6). Figure 5 shows the heterozygous SNPs that were detected, and the unique and shared SNPs in each parent from the mapping populations were evaluated. The percentages of SNPs that were common in the three mapping populations, IACSP95-3018 x IACSP93-3046 (32.86%), SP81-3250 x RB925345 (32.42%) and SP80-3280 x RB835486 (34.06%), were similar, and these SNPs may thus be polymorphic between the parents. As sugarcane is a polyploid species, polymorphisms can be generated from a different number of allelic copies present in each genotype. However, such polymorphisms are difficult to validate (Garcia *et al* 2013, *submitted*).

The SNPs that were unique to each genotype (Figure 5) exhibited a higher probability of association with the contrasting agronomic traits of interest. Because polymorphism markers between parents are important for generating saturated genetic mapping in mapping populations, these SNPs are a source of data for generating markers associated with quantitative trait loci (QTLs). Such functional molecular markers have been broadly applied for the genetic improvement of several crops (Borevitz & Chory, 2004).

According to the Gene Ontology annotation, we identified SNPs in 6,712 unigenes with annotation information, representing 44.80% of the unigenes included in the enrichment analyses. Some categories exhibited important results related to the

genotype (Figure 3), particularly those associated with disease resistance. In the 'signaling' category, we identified 161 unigene sequences with SNPs, whereas we identified 477 unigenes with SNPs in the 'response to stimulus' category. These unigenes likely represent source data for the development of functional markers related to disease resistance.

When we analyzed the categories related to sucrose synthesis, accumulation, storage and retention, we also observed unigenes with SNPs in the 'organic substance transport' (226), 'substrate-specific transporter activity' (196) and 'ion transmembrane transport' (53) clusters. Equally important categories involving sugar transport and metabolism in storage tissues, such as the 'glucose metabolic process' (43), 'small molecule biosynthetic process' (133) and 'small molecule metabolic process' (414) categories, also containing unigene sequences with SNPs.

All of these unigene sequences with SNPs represent an important source of data. These sequences could be priority candidates for the development of specific functional markers and could be very useful in further genetic or genomic studies in sugarcane.

Conclusion

This is the first publicly available sugarcane transcriptome sequencing study performed using NGS technology to investigate the entire sugarcane transcriptome, and our data provide the most comprehensive transcriptome resource currently available for sugarcane. In addition, polymorphisms associated with candidate genes potentially involved in the stimulus response, energy production and growth were identified among the contrasting varieties and deserve future investigation. Based on the enrichment analysis, we identified putative genes related to disease and the accumulation of sucrose. Additionally, a large number of SNPs and SSRs were identified, and marker development would be a useful resource for future genetic or genomic studies of this species. Finally, this work contributed information on 5,000 undescribed genes, which is more than half of the expected sugarcane genes that are missing from sugarcane databases.

References

Barbazuk, W. B., Emrich, S. J., Chen, H. D., Li, L., & Schnable, P. S. (2007). SNP discovery via 454 transcriptome sequencing. *The Plant Journal : For Cell and Molecular Biology*, *51*(5), 910–8.

Bardou, F., Merchan, F., Ariel, F., & Crespi, M. (2011). Dual RNAs in plants. *Biochimie*, *93*(11), 1950–4. doi:10.1016/j.biochi.2011.07.028

Bellodi, N., & Macedo, I. (1995). Quinta geração de variedades de cana-de-açúcar. COOPERATIVA DOS PRODUTORES DE CANA, AÇÚCAR E ÁLCOOL DO ESTADO DE SÃO PAULO. Technical Bulletin. Piracicaba, SP.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2013). GenBank. *Nucleic Acids Research*, *41*(Database issue), D36–42. doi:10.1093/nar/gks1195

Boerner, S., & McGinnis, K. M. (2012). Computational identification and functional predictions of long noncoding RNA in Zea mays. *PloS One*, *7*(8), e43047. doi:10.1371/journal.pone.0043047

Borevitz, J. O., & Chory, J. (2004). Genomics tools for QTL analysis and gene discovery. *Current Opinion in Plant Biology*, *7*(2), 132–6.

Bower, N. I., Casu, R. E., Maclean, D. J., Reverter, A., Chapman, S. C., & Manners, J. M. (2005). Transcriptional response of sugarcane roots to methyl jasmonate. *Plant Science*, *168*(3), 761–772.

Cai, J., Zhao, R., Jiang, H., & Wang, W. (2008). De novo origination of a new protein-coding gene in Saccharomyces cerevisiae. *Genetics*, *179*(1), 487–96. doi:10.1534/genetics.107.084491

Campalans, A., Kondorosi, A., & Crespi, M. (2004). ENOD40, a short open reading frame–containing mRNA, induces cytoplasmic localization of a nuclear RNA binding protein in Medicago truncatula. *The Plant Cell Online*, 1–14. doi:10.1105/tpc.019406.Indeed

Campbell, M. a, Zhu, W., Jiang, N., Lin, H., Ouyang, S., Childs, K. L., … Buell, C. R. (2007). Identification and characterization of lineage-specific genes within the Poaceae. *Plant Physiology*, *145*(4), 1311–22. doi:10.1104/pp.107.104513

Capra, J. a, Pollard, K. S., & Singh, M. (2010). Novel genes exhibit distinct patterns of function acquisition and network integration. *Genome Biology*, *11*(12), R127. doi:10.1186/gb-2010-11-12-r127

Cardoso-Silva, C. B., Costa, E. A., Mancini, M. C., Balsalobre, T. W. A., Canesin, L. E. C., Pinto, L. R., … Vicentini, R. (2014). De Novo Assembly and Transcriptome Analysis of Contrasting Sugarcane Varieties. *PLoS ONE*, *9*(2), e88462. doi:10.1371/journal.pone.0088462

Carson, D., & Botha, F. . (2002). Genes expressed in sugarcane maturing internodal tissue. *Plant Cell Reports*, *20*(11), 1075–1081.

Carson, D. L., & Botha, F. C. (2000a). Preliminary Analysis of Expressed Sequence Tags for Sugarcane. *Crop Science*, *40*(6), 1769–1779.

Carson, D. L., & Botha, F. C. (2000b). Preliminary Analysis of Expressed Sequence Tags for Sugarcane. *Crop Science*, *40*, 1769–1779.

Carvunis, A.-R., Rolland, T., Wapinski, I., Calderwood, M. a, Yildirim, M. a, Simonis, N., … Vidal, M. (2012). Proto-genes and de novo gene birth. *Nature*, *487*(7407), 370–4. doi:10.1038/nature11184

Casu, R. E., Dimmock, C. M., Chapman, S. C., Grof, C. P. L., McIntyre, C. L., Bonnett, G. D., & Manners, J. M. (2004). Identification of differentially expressed transcripts from maturing stem of sugarcane by in silico analysis of stem expressed sequence tags and gene expression profiling. *Plant Molecular Biology*, *54*(4), 503–17.

Casu, R. E., Grof, C. P. L., Rae, A. L., McIntyre, C. L., Dimmock, C. M., & Manners, J. M. (2003). Identification of a novel sugar transporter homologue strongly expressed in maturing stem vascular tissues of sugarcane by expressed sequence tag and microarray analysis. *Plant Molecular Biology*, *52*(2), 371–86.

Cingolani, P., Patel, V. M., Coon, M., Nguyen, T., Land, S. J., Ruden, D. M., & Lu, X. (2012). Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift. *Frontiers in Genetics*, *3*(March), 35.

Clote, P., Ferré, F., Kranakis, E., & Krizanc, D. (2005). Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency, 578–591. doi:10.1261/rna.7220505.4

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., & Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics (Oxford, England)*, *21*(18), 3674–6.

Consortium, U. (2014). Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Research*. Retrieved from http://nar.oxfordjournals.org/content/42/D1/D191.full-text-lowres.pdf

Cordeiro, G. M., Casu, R., McIntyre, C. L., Manners, J. M., & Henry, R. J. (2001). Microsatellite markers from sugarcane (Saccharum spp.) ESTs cross transferable to erianthus and sorghum. *Plant Science*, *160*(6), 1115–1123.

D'Hont, a, Grivet, L., Feldmann, P., Rao, S., Berding, N., & Glaszmann, J. C. (1996). Characterisation of the double genome structure of modern sugarcane cultivars (Saccharum spp.) by molecular cytogenetics. *Molecular & General Genetics : MGG*, *250*(4), 405–13. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8602157

Daugrois, J. H., Grivet, L., Roques, D., Hoarau, J. Y., Lombard, H., Glaszmann, J. C., & D'Hont, A. (1996). A putative major gene for rust resistance linked with a RFLP marker in sugarcane cultivar 'R570 '. *Theor. Appl. Genet*, *92*, 1059–1064.

Dekkers, J. C. M., & Hospital, F. (2002). The use of molecular genetics in the improvement of agricultural populations. *Nature Reviews. Genetics*, *3*(1), 22–32.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., … Frazer, K. A. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Research*, *22*(9), 1775–89. doi:10.1101/gr.132159.111

Dobin, A., Davis, C. a, Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., … Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, *29*(1), 15–21. doi:10.1093/bioinformatics/bts635

Dobzhansky, T. (1964). Biology, molecular and organismic. *American Zoologist*, *4*(4), 443–452. Retrieved from http://www.jstor.org/stable/3881145

Domingues, D. S., Cruz, G. M. Q., Metcalfe, C. J., Nogueira, F. T. S., Vicentini, R., Alves, C. de S., & Van Sluys, M.-A. (2012). Analysis of plant LTR-retrotransposons at the fine-scale family level reveals individual molecular patterns. *BMC Genomics*, *13*(1), 137.

Donoghue, M. T., Keshavaiah, C., Swamidatta, S. H., & Spillane, C. (2011). Evolutionary origins of Brassicaceae specific genes in Arabidopsis thaliana. *BMC Evolutionary Biology*, *11*(1), 47. doi:10.1186/1471-2148-11-47

Dugas, D. V, Monaco, M. K., Olsen, A., Klein, R. R., Kumari, S., Ware, D., & Klein, P. E. (2011). Functional annotation of the transcriptome of Sorghum bicolor in response to osmotic stress and abscisic acid. *BMC Genomics*, *12*, 514. doi:10.1186/1471-2164-12-514

Duret, L., Chureau, C., Samain, S., Weissenbach, J., & Avner, P. (2006). The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. *Science (New York, N.Y.)*, *312*(5780), 1653–5. doi:10.1126/science.1126316

Feltus, F. A., Wan, J., Schulze, S. R., Estill, J. C., Jiang, N., & Paterson, A. H. (2004). An SNP resource for rice genetics and breeding based on subspecies indica and japonica genome alignments. *Genome Research*, *14*(9), 1812–9.

Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M., & Miller, W. (1998). Genomic DNA Sequence A Computer Program for Aligning a cDNA Sequence with a Genomic DNA Sequence, *8*, 967–974.

Forde, B. G. (2002). Local and long-range signaling pathways regulating plant responses to nitrate. *Annual Review of Plant Biology*, *53*(50), 203–24. doi:10.1146/annurev.arplant.53.100301.135256

Franco-Zorrilla, J. M., Valli, A., Todesco, M., Mateos, I., Puga, M. I., Rubio-Somoza, I., … Paz-Ares, J. (2007). Target mimicry provides a new mechanism for regulation of microRNA activity. *Nature Genetics*, *39*(8), 1033–7. doi:10.1038/ng2079

Fukue, Y., Sumida, N., Tanase, J., & Ohyama, T. (2005). A highly distinctive mechanical property found in the majority of human promoters and its transcriptional relevance. *Nucleic Acids Research*, *33*(12), 3821–7. doi:10.1093/nar/gki700

Garg, R., Patel, R. K., Tyagi, A. K., & Jain, M. (2011). De novo assembly of chickpea transcriptome using short reads for gene discovery and marker identification. *DNA Research : An International Journal for Rapid Publication of Reports on Genes and Genomes*, *18*(1), 53–63.

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *Genomics (q-bio.GN); Quantitative Methods (q-bio.QM)*, 1–9.

Gibson, A. K., Smith, Z., Fuqua, C., Clay, K., & Colbourne, J. K. (2013). Why so many unknown genes? Partitioning orphans from a representative transcriptome of the lone star tick Amblyomma americanum. *BMC Genomics*, *14*, 135. doi:10.1186/1471-2164-14-135

Goff, S. a, Ricke, D., Lan, T.-H., Presting, G., Wang, R., Dunn, M., … Briggs, S. (2002). A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). *Science (New York, N.Y.)*, *296*(5565), 92–100. doi:10.1126/science.1068275

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. a, Amit, I., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–52.

Grivet, L., Hont, A. D., Dufour, P., Hamon, P., & Roquest, D. (1994). Comparative genome mapping of sugar cane with other species within the Andropogoneae tribe. *Heredity*, *73*, 500–508.

Guo, X., Gao, L., Liao, Q., Xiao, H., Ma, X., Yang, X., … Zhao, Y. (2013). Long non-coding RNAs function annotation: a global prediction method based on bi-colored networks. *Nucleic Acids Research*, *41*(2), e35. doi:10.1093/nar/gks967

Gutiérrez, R. a, Lejay, L. V, Dean, A., Chiaromonte, F., Shasha, D. E., & Coruzzi, G. M. (2007). Qualitative network models and genome-wide expression data define carbon/nitrogen-responsive molecular machines in Arabidopsis. *Genome Biology*, *8*(1), R7. doi:10.1186/gb-2007-8-1-r7

Hangauer, M. J., Vaughn, I. W., & McManus, M. T. (2013). Pervasive Transcription of the Human Genome Produces Thousands of Previously Unidentified Long Intergenic Noncoding RNAs. *PLoS Genetics*, *9*(6), e1003569. doi:10.1371/journal.pgen.1003569

Hansey, C. N., Vaillancourt, B., Sekhon, R. S., de Leon, N., Kaeppler, S. M., & Buell, C. R. (2012). Maize (Zea mays L.) genome diversity as revealed by RNA-sequencing. *PloS One*, *7*(3), e33071.

Henry, R., & Kole, C. (2010). *Genetics, Genomics and Breeding of Sugarcane*. (C. Henry, R. J.;Kole, Ed.) (1st ed.). Science Publishers.

Heo, J. B., Lee, Y.-S., & Sung, S. (2013). Epigenetic regulation by long noncoding RNAs in plants. *Chromosome Research : An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology*, *21*(6-7), 685–93. doi:10.1007/s10577-013-9392-6

Hoffmann, H. (2008). Variedades RB de cana-de-açúcar. CCA/UFSCarTechnical Bulletin1.

Hogarth, D. M., Ryan, C. C., & Taylor, P. W. J. (1993). Quantitative inheritance of rust resistance in sugarcane. *Field Crops Research*, *34*(2), 187–193.

Ideker, T., Galitski, T., & Hood, L. (2001). A new approach to decoding life: systems biology. *Annual Review of Genomics and …*. Retrieved from http://www.annualreviews.org/doi/abs/10.1146/annurev.genom.2.1.343

Ideker, T., Thorsson, V., Ranish, J. a, Christmas, R., Buhler, J., Eng, J. K., … Hood, L. (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science (New York, N.Y.)*, *292*(5518), 929–34. doi:10.1126/science.292.5518.929

Irvine, J. E. (1975). Relations of Photosynthetic Rates and Leaf and Canopy Characters to Sugarcane Yield. *Crop Science*, *15*, 671.

Iseli, C., Jongeneel, C., & Bucher, P. (1999). ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *ISMB*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.99.6106&rep=rep1&type=pdf

Iseli, C., Jongeneel, C. V., & Bucher, P. (1999). ESTScan: A Program for Detecting, Evaluating, and Reconstructing Potential Coding Regions in EST Sequences, 138–158.

Jacob, F. (1977). Evolution and tinkering. *Science*, *196*(4295), 1161–1166. Retrieved from http://adi-38.bio.ib.usp.br/ibi5023/2010/Jacob_1977.pdf

Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., … Wincker, P. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, *449*(7161), 463–7. doi:10.1038/nature06148

Jurka, J., & Kapitonov, V. (2005). Repbase Update, a database of eukaryotic repetitive elements. *… and Genome Research*. Retrieved from http://www.karger.com/Article/Pdf/84979

Kapusta, A., Kronenberg, Z., Lynch, V. J., Zhuo, X., Ramsay, L., Bourque, G., … Feschotte, C. (2013). Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs. *PLoS Genetics*, *9*(4), e1003470. doi:10.1371/journal.pgen.1003470

Kassube, S. a, Fang, J., Grob, P., Yakovchuk, P., Goodrich, J. a, & Nogales, E. (2013). Structural insights into transcriptional repression by noncoding RNAs that bind to human Pol II. *Journal of Molecular Biology*, *425*(19), 3639–48. doi:10.1016/j.jmb.2012.08.024

Khalil, A. M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., … Rinn, J. L. (2009). Many human large intergenic noncoding RNAs associate

with chromatin-modifying complexes and affect gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(28), 11667–72. doi:10.1073/pnas.0904715106

Kim, J., He, X., & Sinha, S. (2009). Evolution of regulatory sequences in 12 Drosophila species. *PLoS Genetics*, *5*(1), e1000330. doi:10.1371/journal.pgen.1000330

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*, 0–2. Retrieved from http://www.genomics.arizona.edu/553/Readings/2012/Kimura1968.pdf

Kistner, C., & Matamoros, M. (2005). RNA ISOLATION USING PHASE EXTRACTION AND L I C L. In A. Márquez (Ed.), *Lotus japonicus Handbook* (Springers., pp. 123–124). Dordrecht, The Netherlands.

Knowles, D. G., & McLysaght, A. (2009). Recent de novo origin of human protein-coding genes. *Genome Research*, *19*(10), 1752–9. doi:10.1101/gr.095026.109

Koonin, E. (2004). A non-adaptationist perspective on evolution of genomic complexity or the continued dethroning of man. *CELL CYCLE-LANDES BIOSCIENCE-*, (March), 280–285. Retrieved from http://www.landesbioscience.com/journals/cc/kooninCC3-3.pdf?origin=publication_detail

Landell MGA, Campana MP, Figueiredo P, Vasconcelos ACM, Xavier MA, Bidoia MAP, Prado H, Silva MA, Miranda LLD, A. C. (2005). Variedades de cana-de-açúcar para o centro sul do Brasil.Technical Bulletin IAC 197.

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*(3), R25–R25.10.

Li, B., & Dewey, C. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. Retrieved from http://www.biomedcentral.com/content/pdf/1471-2105-12-323.pdf

Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, *12*, 323.

Li, D., Deng, Z., Qin, B., Liu, X., & Men, Z. (2012). De novo assembly and characterization of bark transcriptome using Illumina sequencing and development of EST-SSR markers in rubber tree (Hevea brasiliensis Muell. Arg.). *BMC Genomics*, *13*(1), 192.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25*(16), 2078–9.

Li, S.-W., Yang, H., Liu, Y.-F., Liao, Q.-R., Du, J., & Jin, D.-C. (2012). Transcriptome and gene expression analysis of the rice leaf folder, Cnaphalocrosis medinalis. *PloS One*, *7*(11), e47401.

Liao, Q., Liu, C., Yuan, X., Kang, S., Miao, R., Xiao, H., … Zhao, Y. (2011). Large-scale prediction of long non-coding RNA functions in a coding-non-coding gene co-expression network. *Nucleic Acids Research*, *39*(9), 3864–78. doi:10.1093/nar/gkq1348

Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, a H., & Ecker, J. R. (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, *133*(3), 523–36.

Liu, J., Jung, C., Xu, J., Wang, H., Deng, S., Bernad, L., … Chua, N.-H. (2012). Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in Arabidopsis. *The Plant Cell*, *24*(11), 4333–45. doi:10.1105/tpc.112.102855

Liu, M., Qiao, G., Jiang, J., Yang, H., Xie, L., Xie, J., & Zhuo, R. (2012). Transcriptome sequencing and de novo analysis for Ma bamboo (Dendrocalamus latiflorus Munro) using the Illumina platform. *PloS One*, *7*(10), e46766.

Liu, S., Li, W., Wu, Y., Chen, C., & Lei, J. (2013). De Novo Transcriptome Assembly in Chili Pepper (Capsicum frutescens) to Identify Genes Involved in the Biosynthesis of Capsaicinoids. *PloS One*, *8*(1), e48156. doi:10.1371/journal.pone.0048156

Lorenz, R., Bernhart, S. H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., & Hofacker, I. L. (2011). ViennaRNA Package 2.0. *Algorithms for Molecular Biology : AMB*, *6*(1), 26.

Lu, T., Lu, G., Fan, D., Zhu, C., Li, W., Zhao, Q., … Han, B. (2010). Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. *Genome Research*, *20*(9), 1238–49.

Luo, F., Yang, Y., Zhong, J., Gao, H., Khan, L., Thompson, D. K., & Zhou, J. (2007). Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory. *BMC Bioinformatics*, *8*, 299. doi:10.1186/1471-2105-8-299

Lynch, M. (2007). The evolution of genetic networks by non-adaptive processes. *Nature Reviews. Genetics*, *8*(10), 803–13. doi:10.1038/nrg2192

Lynch, M., & Conery, J. S. (2003). The origins of genome complexity. *Science (New York, N.Y.)*, *302*(5649), 1401–4. doi:10.1126/science.1089370

Ma, H.-M., Schulze, S., Lee, S., Yang, M., Mirkov, E., Irvine, J., … Paterson, A. (2004). An EST survey of the sugarcane transcriptome. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, *108*(5), 851–63.

Managadze, D., Lobkovsky, A. E., Wolf, Y. I., Shabalina, S. a, Rogozin, I. B., & Koonin, E. V. (2013). The vast, conserved mammalian lincRNome. *PLoS Computational Biology*, *9*(2), e1002917. doi:10.1371/journal.pcbi.1002917

Managadze, D., Rogozin, I. B., Chernikova, D., Shabalina, S. a, & Koonin, E. V. (2011). Negative correlation between expression level and evolutionary rate of long intergenic noncoding RNAs. *Genome Biology and Evolution*, *3*, 1390–404. doi:10.1093/gbe/evr116

Mancini, M. C., Leite, D. C., Perecin, D., Bidóia, M. a. P., Xavier, M. a., Landell, M. G. a., & Pinto, L. R. (2012). Characterization of the Genetic Variability of a Sugarcane Commercial Cross Through Yield Components and Quality Parameters. *Sugar Tech*, *14*(2), 119–125.

Marconi, T. G., Costa, E. A., Miranda, H. R., Mancini, M. C., Cardoso-Silva, C. B., Oliveira, K. M., … Souza, A. P. (2011). Functional markers for gene mapping and genetic diversity studies in sugarcane. *BMC Research Notes*, *4*(1), 264.

Marguerat, S., & Bähler, J. (2010). RNA-seq: from technology to biology. *Cellular and Molecular Life Sciences : CMLS*, *67*(4), 569–79.

McCormick, A. J., Cramer, M. D., & Watt, D. A. (2006). Sink strength regulates photosynthesis in sugarcane. *The New Phytologist*, *171*(4), 759–70.

Metzgar, D., Bytof, J., & Wills, C. (2000). Selection against frameshift mutations limits microsatellite expansion in coding DNA. *Genome Research*, *10*(1), 72–80.

Ming, R., Liu, S. C., Lin, Y. R., da Silva, J., Wilson, W., Braga, D., … Paterson, a H. (1998). Detailed alignment of saccharum and sorghum chromosomes: comparative organization of closely related diploid and polyploid genomes. *Genetics*, *150*(4), 1663–82.

Ministério da Agricultura, P. e A. (2013). Acompanhamento de safra brasileira : cana-de-açúcar Safra 2012/2013 Terceiro levantamento. *Companhia Nacional de Abastecimento*.

Moore, P. H., Botha, F. ., Furbank, R. ., & Grof, C. P. . (1996). *Intensive sugarcane production: Meeting the challenge beyond 2000*. (Keating BA and Wilson JR, Ed.) (p. 141). Oxon, UK: CAB International.

Morozova, O., & Marra, M. a. (2008). Applications of next-generation sequencing technologies in functional genomics. *Genomics*, *92*(5), 255–64.

Mutwil, M., Usadel, B., Schütte, M., Loraine, A., Ebenhöh, O., & Persson, S. (2010). Assembly of an interactive correlation network for the Arabidopsis genome using a novel heuristic clustering algorithm. *Plant Physiology*, *152*(1), 29–43. doi:10.1104/pp.109.145318

Novaes, E., Drost, D. R., Farmerie, W. G., Pappas, G. J., Grattapaglia, D., Sederoff, R. R., & Kirst, M. (2008). High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome. *BMC Genomics*, *9*, 312.

Ogasawara, J., & Morishita, S. (2003). A fast and sensitive algorithm for aligning ESTs to the human genome. *Journal of Bioinformatics and …*, *1*(2), 363–386. Retrieved from http://www.worldscientific.com/doi/abs/10.1142/S0219720003000058

Ouyang, S., & Buell, C. R. (2004). The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants. *Nucleic Acids Research*, *32*(Database issue), D360–3. doi:10.1093/nar/gkh099

Papini-Terzi, F. S., Rocha, F. R., Vêncio, R. Z. N., Felix, J. M., Branco, D. S., Waclawovsky, A. J., … Souza, G. M. (2009). Sugarcane genes associated with sucrose content. *BMC Genomics*, *10*, 120. doi:10.1186/1471-2164-10-120

Patel, R. K., & Jain, M. (2012). NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PloS One*, *7*(2), e30619.

Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., … Rokhsar, D. S. (2009). The Sorghum bicolor genome and the diversification of grasses. *Nature*, *457*(7229), 551–6. doi:10.1038/nature07723

Pinto, L. R., Oliveira, K. M., Ulian, E. C., Garcia, A. A. F., & de Souza, A. P. (2004). Survey in the sugarcane expressed sequence tag database (SUCEST) for simple sequence repeats. *Genome / National Research Council Canada = Génome / Conseil National de Recherches Canada*, *47*(5), 795–804.

Quint, M., Drost, H.-G., Gabel, A., Ullrich, K. K., Bönn, M., & Grosse, I. (2012). A transcriptomic hourglass in plant embryogenesis. *Nature*, *490*(7418), 98–101. doi:10.1038/nature11394

Ramu, P., Kassahun, B., Senthilvel, S., Ashok Kumar, C., Jayashree, B., Folkertsma, R. T., … Hash, C. T. (2009). Exploiting rice-sorghum synteny for targeted development of EST-SSRs to enrich the sorghum genetic linkage map. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, *119*(7), 1193–204.

Rolland, F., Baena-Gonzalez, E., & Sheen, J. (2006a). Sugar sensing and signaling in plants: conserved and novel mechanisms. *Annual Review of Plant Biology*, *57*, 675–709. doi:10.1146/annurev.arplant.57.032905.105441

Rolland, F., Baena-Gonzalez, E., & Sheen, J. (2006b). Sugar sensing and signaling in plants: conserved and novel mechanisms. *Annual Review of Plant Biology*, *57*, 675–709. doi:10.1146/annurev.arplant.57.032905.105441

Rook, F., Hadingham, S. a., Li, Y., & Bevan, M. W. (2006). Sugar and ABA response pathways and the control of gene expression. *Plant, Cell and Environment*, *29*(3), 426–434. doi:10.1111/j.1365-3040.2005.01477.x

Rutter, M. T., Cross, K. V, & Van Woert, P. a. (2012). Birth, death and subfunctionalization in the Arabidopsis genome. *Trends in Plant Science*, *17*(4), 204–12. doi:10.1016/j.tplants.2012.01.006

Sabino, J. (1997). Sexta geração de variedades de cana-de-açúcar. COOPERATIVA DE PRODUTORES DE CANA, AÇÚCAR E ÁLCOOL DO ESTADO DE SÃO PAULO LTDA. Technical Bulletin. Piracicaba, SP.

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., … Jackson, S. a. (2010). Genome sequence of the palaeopolyploid soybean. *Nature*, *463*(7278), 178–83. doi:10.1038/nature08670

Seffens, W., & Digby, D. (1999). mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. *Nucleic Acids Research*, *27*(7), 1578–84. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=148359&tool=pmcentrez&rendertype=abstract

Silveira, A. B., Trontin, C., Cortijo, S., Barau, J., Del Bem, L. E. V., Loudet, O., … Vincentz, M. (2013). Extensive natural epigenetic variation at a de novo originated gene. *PLoS Genetics*, *9*(4), e1003437. doi:10.1371/journal.pgen.1003437

Smale, S. T., & Kadonaga, J. T. (2003). The RNA polymerase II core promoter. *Annual Review of Biochemistry*, *72*, 449–79. doi:10.1146/annurev.biochem.72.121801.161520

Sun, J., Zhou, M., Mao, Z.-T., Hao, D.-P., Wang, Z.-Z., & Li, C.-X. (2013). Systematic analysis of genomic organization and structure of long non-coding RNAs in the human genome. *FEBS Letters*, *587*(7), 976–82. doi:10.1016/j.febslet.2013.02.036

Tai, P. Y. P., Miller, J. D., & Dean, J. L. (1981). INHERITANCE OF RESISTANCE TO RUST IN SUGARCANE. *Field Crops Research*, *4*, 261–268.

Tautz, D., & Domazet-Lošo, T. (2011). The evolutionary origin of orphan genes. *Nature Reviews. Genetics*, *12*(10), 692–702. doi:10.1038/nrg3053

Tomato, T., & Consortium, G. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, *485*(7400), 635–41. doi:10.1038/nature11119

Trick, M., Long, Y., Meng, J., & Bancroft, I. (2009). Single nucleotide polymorphism (SNP) discovery in the polyploid Brassica napus using Solexa transcriptome sequencing. *Plant Biotechnology Journal*, *7*(4), 334–46.

Ulitsky, I., & Bartel, D. P. (2013). lincRNAs: Genomics, Evolution, and Mechanisms. *Cell*, *154*(1), 26–46. doi:10.1016/j.cell.2013.06.020

Varshney, R. K., Nayak, S. N., May, G. D., & Jackson, S. a. (2009). Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends in Biotechnology*, *27*(9), 522–30.

Vettore, A. L., da Silva, F. R., Kemper, E. L., Souza, G. M., da Silva, A. M., Ferro, M. I. T., … Arruda, P. (2003). Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane. *Genome Research*, *13*(12), 2725–35.

Vettore, A. L., Silva, F. R., Kemper, E. L., & Arruda, P. (2001). The libraries that made SUCEST, *24*, 1–7.

Vicentini, R., Bem, L. E. V., Sluys, M. a., Nogueira, F. T. S., & Vincentz, M. (2012). Gene Content Analysis of Sugarcane Public ESTs Reveals Thousands of Missing Coding-Genes and an Unexpected Pool of Grasses Conserved ncRNAs. *Tropical Plant Biology*, *5*(2), 199–205. doi:10.1007/s12042-012-9103-z

Vincentz, M., Cara, F. A. A., Okura, V. K., da Silva, F. R., Pedrosa, G. L., Hemerly, A. S., … Menck, C. F. M. (2004). Evaluation of monocot and eudicot divergence using the sugarcane transcriptome. *Plant Physiology*, *134*(3), 951–9.

Wakeley, J. (1996). The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance. *Tree*, *11*(4), 158– 162.

Wolf, Y. I., Novichkov, P. S., Karev, G. P., Koonin, E. V, & Lipman, D. J. (2009). The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(18), 7273–80. doi:10.1073/pnas.0901808106

Yu, J., Hu, S., Wang, J., Wong, G. K.-S., Li, S., Liu, B., … Yang, H. (2002). A draft sequence of the rice genome (Oryza sativa L. ssp. indica). *Science (New York, N.Y.)*, *296*(5565), 79–92. doi:10.1126/science.1068037

Zhai, J., Jeong, D., Paoli, E. De, Park, S., Rosen, B. D., Yan, Z., … Meyers, B. C. (2011). MicroRNAs as master regulators of the plant NB-LRR defense gene family via the production of phased , trans -acting siRNAs, 2540–2553. doi:10.1101/gad.177527.111.infection

Zhang, B., & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, *4*, Article17. doi:10.2202/1544-6115.1128

Zhao, L., Saelao, P., Jones, C. D., & Begun, D. J. (2014). Origin and spread of de novo genes in Drosophila melanogaster populations. *Science (New York, N.Y.)*, *343*(6172), 769–72. doi:10.1126/science.1248286

Zhou, Q., Zhang, G., Zhang, Y., Xu, S., Zhao, R., Zhan, Z., … Wang, W. (2008). On the origin of new genes in Drosophila. *Genome Research*, *18*(9), 1446–55. doi:10.1101/gr.076588.108

Zhu, Q.-H., & Wang, M.-B. (2012). Molecular Functions of Long Non-Coding RNAs in Plants. *Genes*, *3*(4), 176–190. doi:10.3390/genes3010176

Tables

Table 1. Summary of Illumina transcriptome sequencing data for the sugarcane varieties included in this study.

Sample	Read length (bp)	Raw data	Trimmed data	GC (%)	Q20 (%)
SP95-3018	72+72	84,105,462	64,906,391	49.04	98.09
SP81-3250	72+72	103,971,718	71,002,186	47.52	97.32
RB925345	72+72	112,124,334	77,476,268	46.91	97.11
SP80-3280	72+72	101,983,186	73,160,814	47.59	97.56
RB835486	72+72	119,280,444	87,873,521	46.62	97.66
SP93-3046	72+72	88,767,346	70,955,324	48.07	98.25

Table 2. Summary of the *de novo* assembly results for the sugarcane transcriptome.

Unigene length (bp)	Total unigenes	Percentage
300-500	31,971	44.24%
500-1000	20,634	28.55%
1000-2000	12,007	16.61%
2000-3000	4,827	6.68%
3000-4000	1,790	2.47%
4000-5000	636	0.88%
>5000	404	0.56%
Total length (bp)	66,572,642	-
Unigenes	72,269	-
N50 length	1,367	-
GC (%)	46.39	-

Table 3. Summary of the annotation of each database.

Database	Number of unigenes	Number of proteins matched	Percentage of unigenes[a]
Viridiplantae proteins	35,456	34,969	49.06%
Grass proteins	34,814	34,304	48.17%
Sorghum proteins	28,788	28,030	39.83%
Hits against sorghum proteins and sugarcane ESTs	22,171	20,969	30.68%
Total of no-hit unigenes	36,813	-	50.94%
No-hit unigenes with high similarity to the sorghum genome	18,910	-	26,16

[a]Percentage relative to the total number of sugarcane unigenes.

Table 4. Summary of the simple sequence repeat (SSR) types in the sugarcane transcriptome.

Repeat motif	Number[a]	Unigenes[b]	Percentage (%)[c]
Di-nucleotide			
AC/GT	551		
AG/CT	962		
AT/TA	336		
CG/GC	78		
Total	**1,927**	**1,732**	**37.74**
Tri-nucleotide			
AAC/GTT	141		
AAG/CTT	152		
AAT/ATT	60		
AGC/GCT	219		
ACG/CGT	197		
AGT/ACT	62		
ACC/GGT	122		
AGG/CCT	252		
ACA/TGT	97		
AGA/TCT	46		
ATA/TAT	24		
ATC/GAT	42		
ATG/CAT	43		
CAC/GTG	69		
CAG/CTG	228		
CCG/CGG	442		
CGC/GCG	241		
CTC/GAG	148		
Total	**2,585**	**2,318**	**50.63**
Other motifs[d]	**594**	**1,708**	**11.63%**
Total	**5,106**	**5,758**	**-**

[a]Number of the total SSRs (di-, tri- and other motifs).

[b]Number of unigene sequences containing SSRs.

[c]The relative percentage of SSRs with different repeat motifs among the total SSRs.

[d]The total number of SSRs of other sizes.

Figure legends

Figure 1. Proportions of sugarcane transcripts showing homology to sugarcane unigenes and sorghum and rice proteins. For annotation, the best BLASTX/N hit against the protein or nucleotide sequences of the reference organisms was employed, with an E-value cut-off of ≤ 10^{-6}. The number between the parentheses indicates the number of different proteins/unigenes in each species (sugarcane[a], sorghum[b] and rice[c]). The number outside of the Venn diagram indicates no-hit transcripts and the number of transcripts[d] that mapped to the sorghum genome.

Figure 2. Histogram of the Clusters of Orthologous Groups (COG) classifications of the sugarcane transcripts and sorghum proteins.

Figure 3. Enrichment of Gene Ontology terms for each sugarcane variety.

Figure 4. Hierarchical clustering of the 358 putative sugarcane lncRNAs. The expression patterns allowed the identification of the genotypes based on their ability to store sucrose and according to the bi-parental crosses involved in the different mapping populations.

Figure 5. Unique and shared heterozygous putative SNPs in the parental genotypes of the three sugarcane mapping populations.

Figures

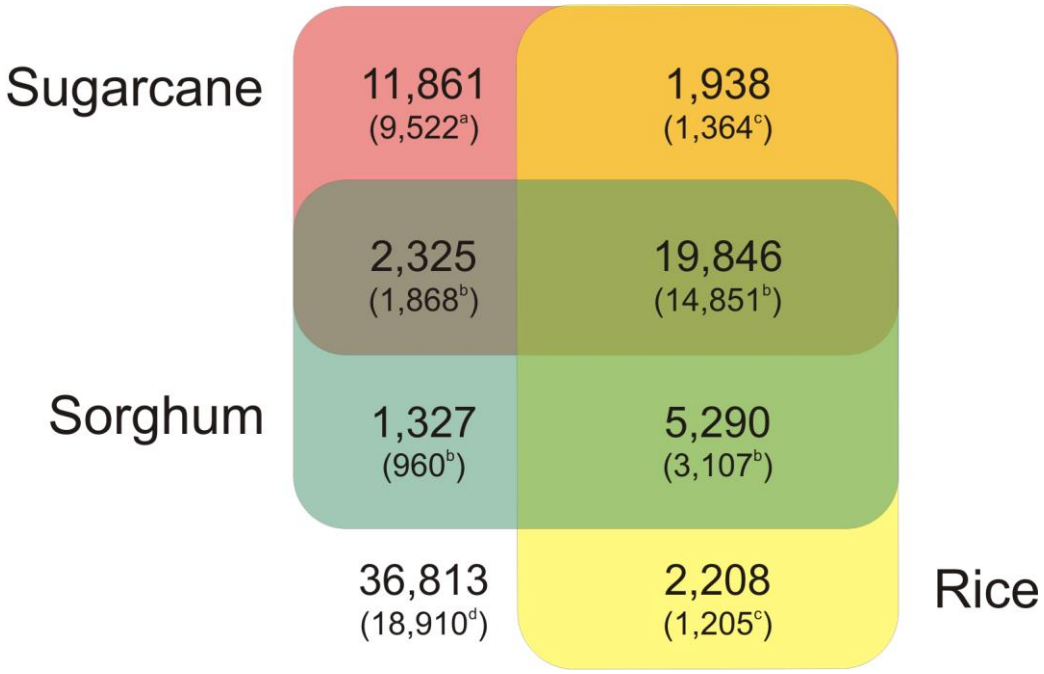


Figure 1.

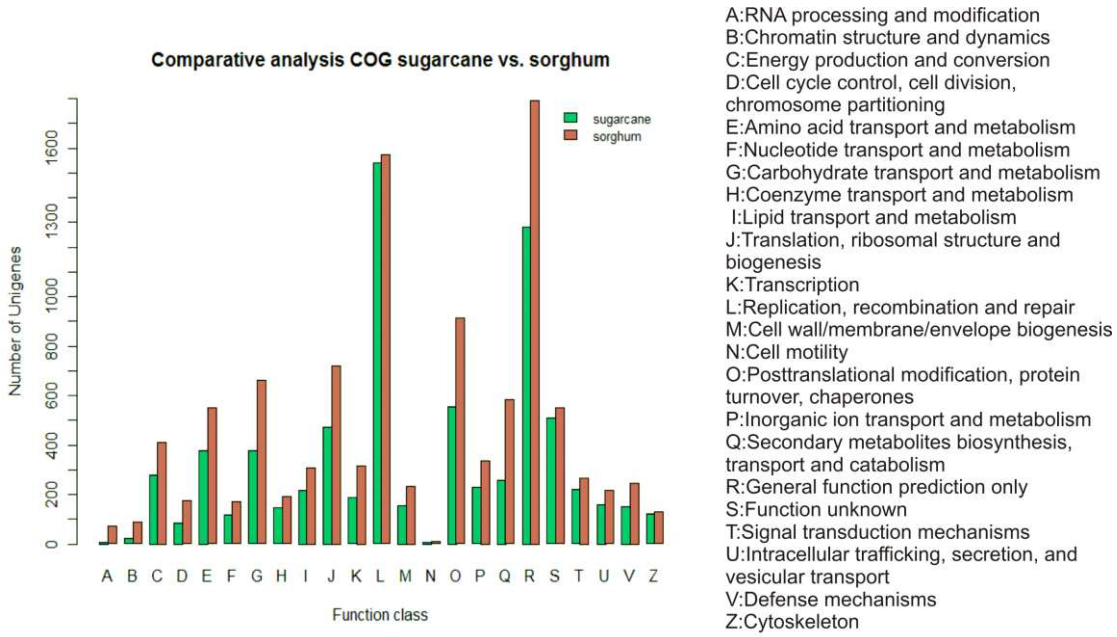


Figure 2.

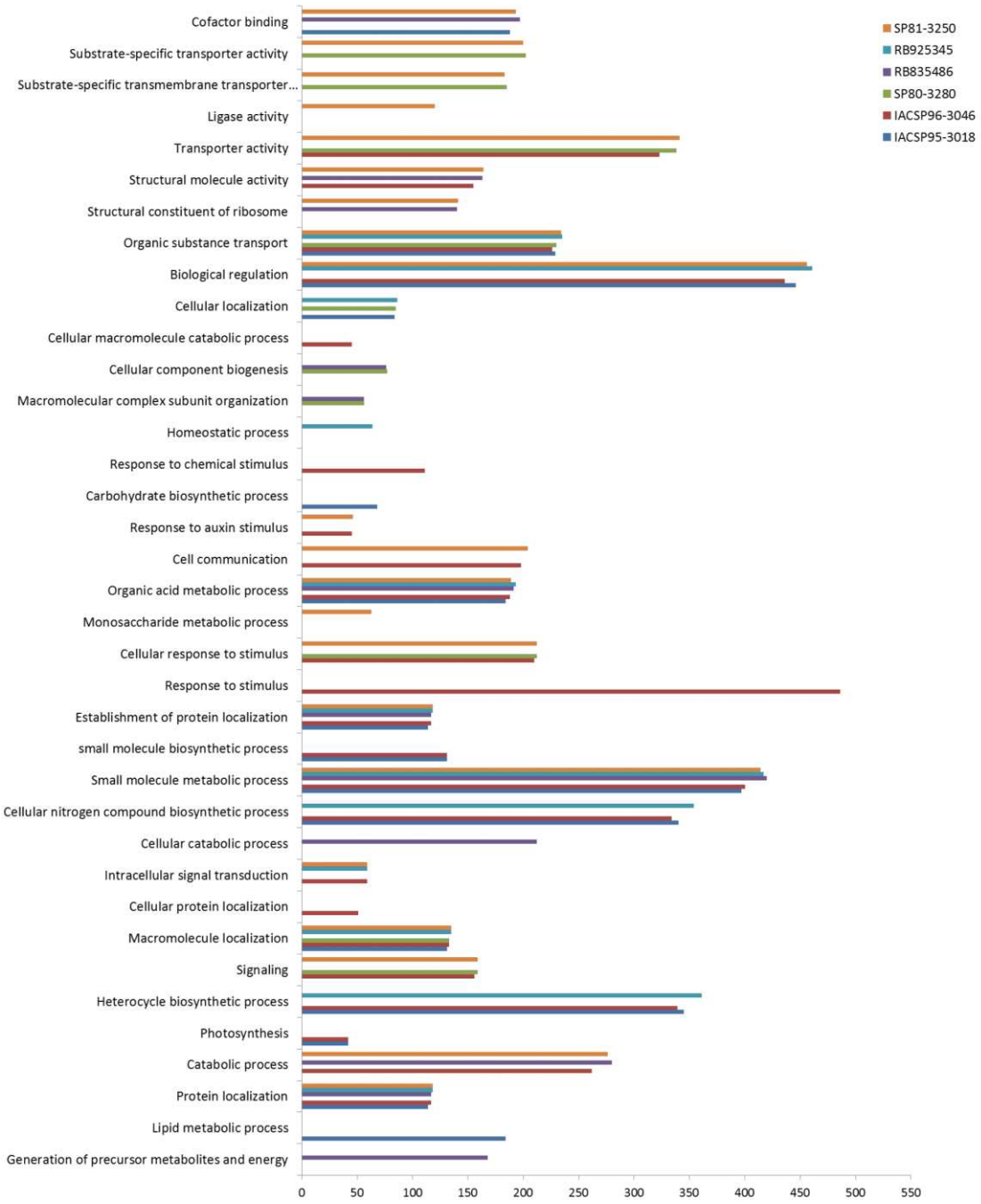

Cofactor binding
Substrate-specific transporter activity
Substrate-specific transmembrane transporter...
Ligase activity
Transporter activity
Structural molecule activity
Structural constituent of ribosome
Organic substance transport
Biological regulation
Cellular localization
Cellular macromolecule catabolic process
Cellular component biogenesis
Macromolecular complex subunit organization
Homeostatic process
Response to chemical stimulus
Carbohydrate biosynthetic process
Response to auxin stimulus
Cell communication
Organic acid metabolic process
Monosaccharide metabolic process
Cellular response to stimulus
Response to stimulus
Establishment of protein localization
small molecule biosynthetic process
Small molecule metabolic process
Cellular nitrogen compound biosynthetic process
Cellular catabolic process
Intracellular signal transduction
Cellular protein localization
Macromolecule localization
Signaling
Heterocycle biosynthetic process
Photosynthesis
Catabolic process
Protein localization
Lipid metabolic process
Generation of precursor metabolites and energy
0
50
100
150
200
250
300
350
400
450
500
550
SP81-3250
RB925345
RB835486
SP80-3280
IACSP96-3046
IACSP95-3018

Figure 3.

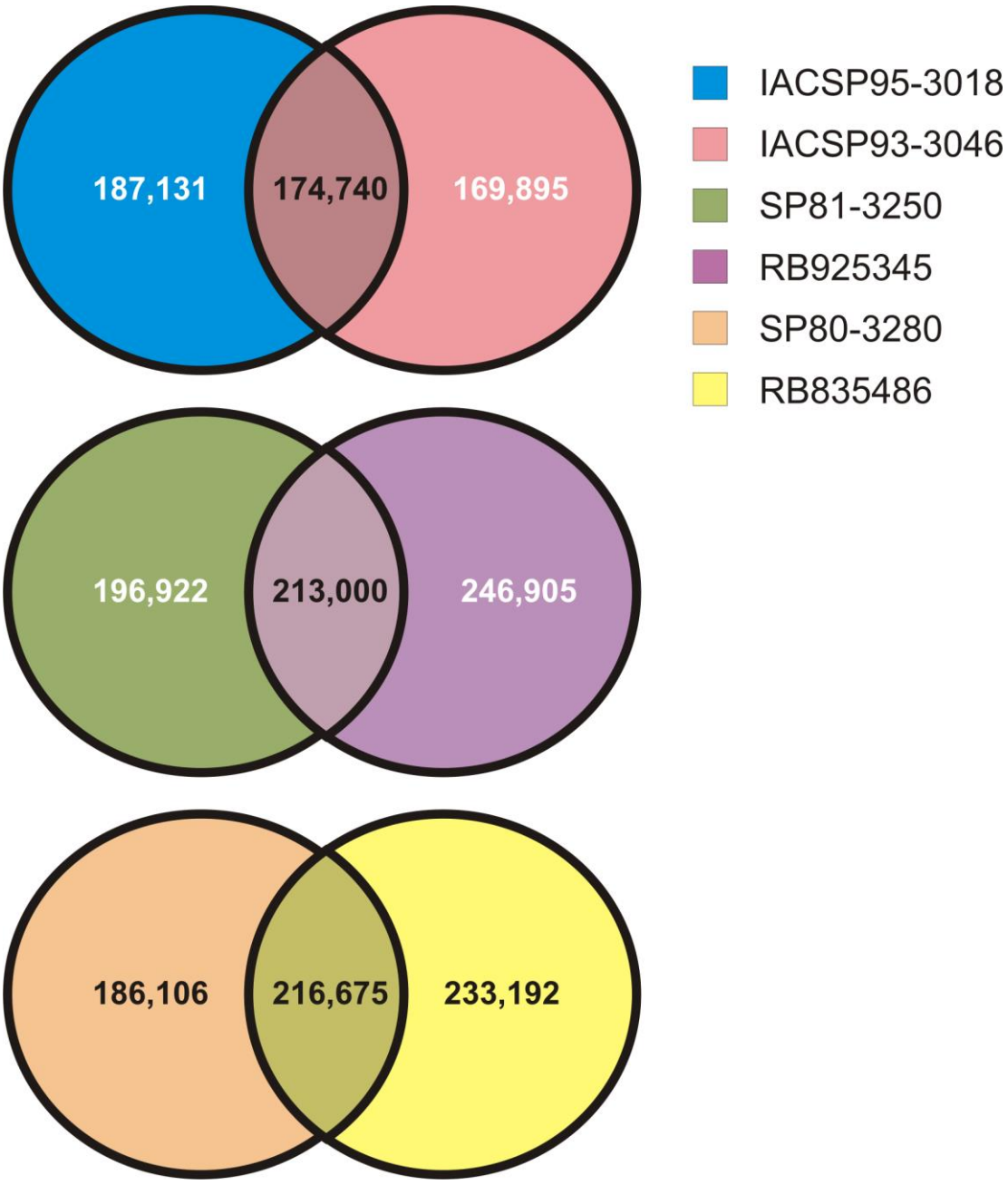

187,131
174,740
169,895
196,922
213,000
246,905
186,106
216,675
233,192
IACSP95-3018
IACSP93-3046
SP81-3250
RB925345
SP80-3280
RB835486

Figure 4.

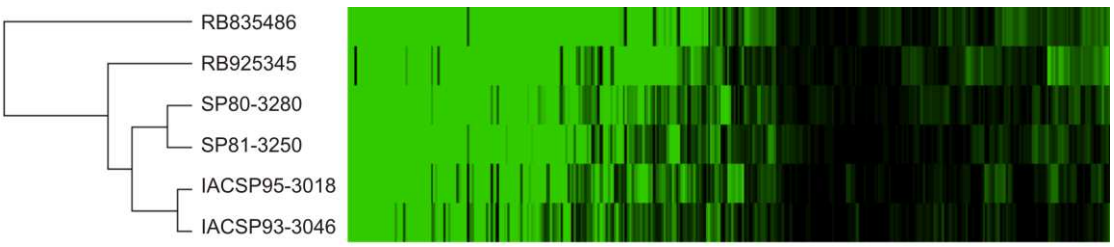

RB835486
RB925345
SP80-3280
SP81-3250
IACSP95-3018
IACSP93-3046

Figure 5.

Supporting information

Text S1. Unigene sequences in FASTA format.

Text S2. Gene ontology enrichment annotation for the transcripts of each genotype.

Text S3. Putative previously unknown sugarcane transcripts showing the best matches to sorghum proteins.

Text S4. List of 18,910 putative sugarcane ncRNAs with high coverage in the sorghum genome.

Text S5. List of 708,125 putative SNP positions identified in this study.

Text S6. List of putative SNPs identified in each genotype.

Text S7. List of 5,106 putative SSR positions identified in this study.

Figure S1. Venn diagram showing the classification of the identified putative sugarcane lncRNAs in the EST data (A) and RNA-Seq data (B).

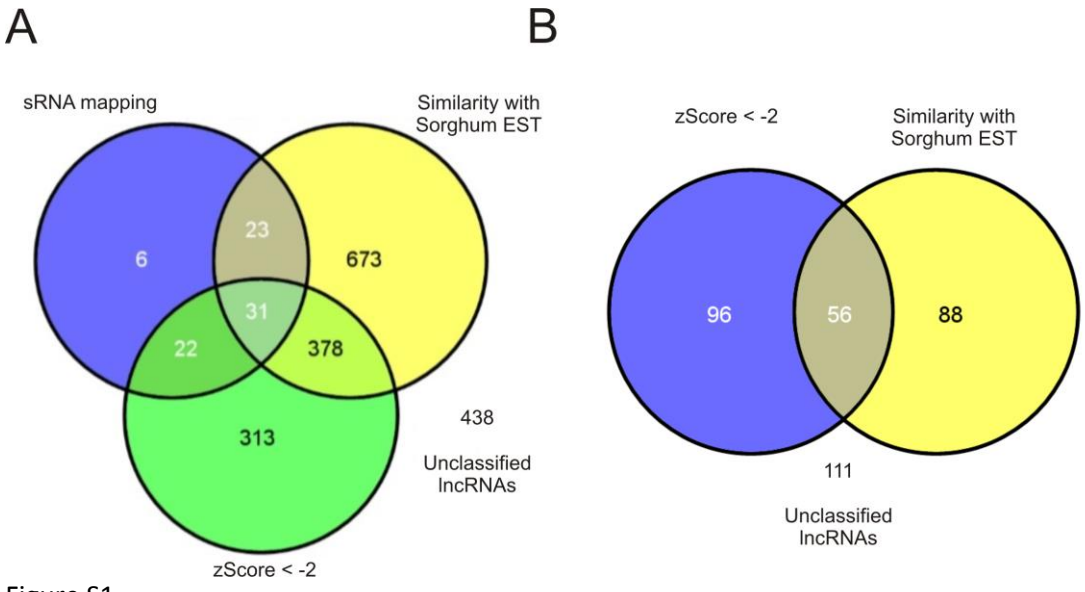


Figure S1.