



Universidade Federal de Goiás
Instituto de Ciências Biológicas
Programa de Pós-Graduação em Genética e Biologia Molecular

ANÁLISE DE RNAS LONGOS NÃO CODIFICANTES NO GENOMA
DE *Arabidopsis thaliana* (L.) Heynh.

VANESSA CRISTINA DA SILVA ARAÚJO

Goiânia, fevereiro de 2017

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO PARA DISPONIBILIZAR AS TESES E DISSERTAÇÕES ELETRÔNICAS NA BIBLIOTECA DIGITAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), regulamentada pela Resolução CEPEC nº 832/2007, sem ressarcimento dos direitos autorais, de acordo com a Lei nº 9610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou *download*, a título de divulgação da produção científica brasileira, a partir desta data.

1 **1. Identificação do material bibliográfico:** ☒ **Dissertação** ☐ **Tese**

1 **2. Identificação da Tese ou Dissertação**

2

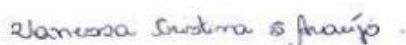
Nome completo do autor: Vanessa Cristina da Silva Araújo

Título do trabalho: Análise de RNAs longos não codificantes do genoma de *Arabidopsis thaliana*.

3. Informações de acesso ao documento:

Concorda com a liberação total do documento ☒ SIM ☐ NÃO¹

Havendo concordância com a disponibilização eletrônica, torna-se imprescindível o envio do(s) arquivo(s) em formato digital PDF da tese ou dissertação.



Assinatura do (a) autor (a)

Data: 10 / 04 / 2017

¹ Neste caso o documento será embargado por até um ano a partir da data de defesa. A extensão deste prazo suscita justificativa junto à coordenação do curso. Os dados do documento não serão disponibilizados durante o período de embargo.

Universidade Federal de Goiás
Programa de Pós-Graduação em Genética e Biologia Molecular

ANÁLISE DE RNAS LONGOS NÃO CODIFICANTES NO GENOMA
DE *Arabidopsis thaliana* (L.) Heynh.

Orientador: Prof. Dr. EVANDRO NOVAES

Dissertação apresentada à Coordenação
do Programa de Pós-Graduação em
Genética e Biologia Molecular da
Universidade Federal de Goiás

Goiânia, 2017

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Cristina da Silva Araújo, Vanessa

Análise de RNAs longos não codificantes no genoma de
Arabidopsis thaliana (L.) Heynh. [manuscrito] / Vanessa Cristina da
Silva Araújo. - 2017.

LXXX, 80 f.: il.

Orientador: Prof. Dr. Evandro Novaes.

Dissertação (Mestrado) - Universidade Federal de Goiás, Instituto
de Ciências Biológicas (ICB), Programa de Pós-Graduação em Genética
e Biologia Molecular, Goiânia, 2017.

Bibliografia.

Inclui siglas, abreviaturas, gráfico, tabelas, algoritmos, lista de
figuras, lista de tabelas.

1. *Arabidopsis*. 2. Bioinformática. 3. lncRNA. 4. RNA-Seq. I.
Novaes, Evandro, orient. II. Título.

CDU 575



SERVIÇO PÚBLICO FEDERAL
UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA E BIOLOGIA MOLECULAR

ATA DA SESSÃO PÚBLICA DE DEFESA DE DISSERTAÇÃO DE Nº 044

Aos sete dias do mês de março do ano de dois mil e dezessete (07/03/2017), às 08hs00min, no (a) Auditório do Prédio de Melhoramento de Plantas na Escola de Agronomia, reuniram-se os componentes da banca examinadora: **Prof. Dr. Evandro Novaes; Profa. Dra. Rosana Pereira Vianello e Prof. Dr. Alexandre Siqueira Guedes Coelho** para, em sessão pública presidida pelo primeiro examinador citado, procederem à avaliação da defesa de dissertação intitulada: **"Análise de RNAs longos não codificantes no genoma de *Arabidopsis thaliana*."**, em nível de mestrado, área de concentração em **Genética e Biologia Molecular**, de autoria de **Vanessa Cristina da Silva Araújo**, discente do Programa de Pós-Graduação em Genética e Biologia Molecular da Universidade Federal de Goiás. A sessão foi aberta pelo presidente, que fez a apresentação formal dos membros da banca. A palavra, a seguir, foi concedida à autora da dissertação que, em cerca de 35 minutos, procedeu à apresentação de seu trabalho. Terminada a apresentação, cada membro da banca arguiu a examinada, tendo-se adotado o sistema de diálogo sequencial. Terminada a fase de arguição, procedeu-se à avaliação da dissertação. Tendo-se em vista o que consta na Resolução nº 1294/2014 do Conselho de Ensino, Pesquisa, Extensão e Cultura (CEPEC), que regulamenta o Programa de Pós-Graduação em Genética e Biologia Molecular, a dissertação foi aprovada, considerando-se integralmente cumprido este requisito para fins de obtenção do título de Mestre em Genética e Biologia Molecular pela Universidade Federal de Goiás. A conclusão do curso dar-se-á quando da entrega da versão definitiva da dissertação na secretaria do programa, com as devidas correções sugeridas pela banca examinadora, no prazo de trinta dias a contar da data da defesa. Cumpridas as formalidades de pauta, às 11 horas e 50



SERVIÇO PÚBLICO FEDERAL
UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS

PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA E BIOLOGIA MOLECULAR

28 minutos, encerrou-se a sessão de defesa e, para constar, eu, Gleizilene Braz
29 Pereira dos Santos, Assistente em Administração do Instituto de Ciências
30 Biológicas, da Universidade Federal de Goiás, lavrei a presente ata que, após
31 lida e aprovada, será assinada pelos membros da banca examinadora em três
32 vias de igual teor.

33
34
35
36 *Evandro Novaes*
37 **Prof. Dr. Evandro Novaes**
38 **Presidente da Banca**
39 **UFG/GO**
40
41

42
43 *Rosana Pereira Vianello*
44 **Profa. Dra. Rosana Pereira Vianello**
45 **EMBRAPA ARROZ E FEIJÃO/GO**
46
47

48
49 *Alexandre Siqueira Guedes Coelho*
50 **Prof. Dr. Alexandre Siqueira Guedes Coelho**
51 **UFG/GO**

Resumo

O sequenciamento em larga escala de transcritos, via RNA-Seq, vêm mudando paradigmas ao demonstrar que a transcrição é prevalente por todo o genoma dos eucariotos. Nesses organismos, a grande maioria dos transcritos não codificam proteínas (ncRNA). Um tipo de RNA que vêm despertando grande interesse, dado sua prevalência, são os RNAs longos não codificantes (lncRNAs), que são ncRNA com mais de 200 nucleotídeos. No entanto, pouco se sabe sobre o seu papel e prevalência nos genomas de plantas, mesmo em espécies modelo como *Arabidopsis thaliana* (L.) Heynh. O objetivo desse trabalho foi identificar lncRNAs no genoma de *Arabidopsis* e caracterizar seus tamanhos, estruturas e diversidade genética. As sequências utilizadas foram obtidas de um trabalho que sequenciou RNA total de *A. thaliana*, sob diferentes regimes de luminosidade, utilizando a plataforma Illumina HiSeq 2000. Estas sequências foram mapeadas no genoma referência com o programa TopHat e montadas com o Cufflinks. Os transcritos montados foram comparados com a anotação do genoma com o Cuffcompare, afim de identificar transcritos ainda não anotados. Um total de 4.305 RNAs longos putativos foi obtido, sendo 314 (7%) senso em relação a transcritos codantes (mRNAs), 392 (9%) intergênicos, 2.216 intrônicos (52%) e 1.383 (32%) antisenso de mRNAs. Os lncRNAs obtidos foram filtrados para eliminar aqueles com potencial de codificação, bem como aqueles relacionados com a síntese rRNA, tRNA e miRNA. Após essa filtragem, foram obtidos 3.710 lncRNAs de alta confiança (HC-lncRNA), sendo que desses 58,6% ainda não foram previamente anotados. Esses HC-lncRNA representam uma baixa proporção (~1%) do genoma de *Arabidopsis thaliana*. Uma análise de enriquecimento funcional de categorias do Gene Ontology (GO) demonstrou que os genes que contém lncRNAs apresentam enriquecimento para processos ligados à localização e transporte de proteínas dentro da célula, bem como para ligação a ácidos nucléicos. Uma análise de expressão gênica identificou apenas 22 lncRNAs diferencialmente expressos entre as diferentes condições de luminosidade em que as amostras foram expostas. Utilizando os SNPs do projeto 1001 genomes, identificou-se alta diversidade nucleotídica em regiões de lncRNAs, indicando baixa conservação da estrutura primária destes transcritos. A diversidade nucleotídica em regiões de RNAs longos não codificantes é menor do que em regiões codantes, mas menor do que a diversidade observada em regiões neutras como os pseudogenes.

Palavras-chave: lncRNA, *Arabidopsis*, Bioinformática, RNA-Seq.

Abstract

Large-scale sequencing of transcripts via RNA-Seq has been changing paradigms by demonstrating that transcription is prevalent throughout the eukaryotic genome. In these organisms, the vast majority of transcripts are non-coding (ncRNA). One type of RNA that has aroused great interest, given its prevalence, is the long non-coding RNAs (lncRNAs), which are ncRNA with more than 200 nucleotides. However, little is known about the role and prevalence of these lncRNAs in plant genomes, even in model species such as *Arabidopsis thaliana* (L.) Heynh. The objective of this work was to identify lncRNAs in the *Arabidopsis* genome and to characterize their size, structure and nucleotide diversity. The sequences were obtained from a previous work that sequenced total RNA from *A. thaliana*, grown under different light regimes, using Illumina HiSeq 2000 platform. These sequences were mapped into the reference genome with TopHat and assembled with Cufflinks. The assembled transcripts were compared with the genome annotation with Cuffcompare, to identify non-annotated transcripts. A total of 4,305 long putative RNAs were obtained, with 314 (7%) being sense in relation to coding transcripts (mRNAs), 392 (9%) intergenic, 2,216 intronic (52%) and 1,383 (32%) antisense mRNAs. The lncRNAs obtained were filtered to eliminate those with coding potential, as well as those related to rRNA, tRNA and miRNA synthesis. A total of 3,710 high-confidence lncRNAs (HC-lncRNA) were obtained, of which 58.6% were not previously annotated. These HC-lncRNA encompass a low proportion (~ 1%) lncRNAs of the *Arabidopsis thaliana* genome. A functional enrichment analysis of Gene Ontology (GO) categories demonstrated that among genes containing lncRNAs there is a high proportion of categories linked to the localization and transport of proteins within the cell, as well as to nucleic acid binding. A gene expression analyses identified only 22 differentially expressed lncRNAs under the different light conditions in which samples were exposed. Using the SNP data from the 1001 genomes project, a high nucleotide diversity within lncRNAs regions was identified, indicating low conservation of the primary structure of these transcripts. The nucleotide diversity in regions of long non-coding RNAs is higher than in coding regions, but lower than in neutral regions such as pseudogenes.

Key-words: lncRNA, *Arabidopsis*, Bioinformatics, RNA-Seq

Sumário

Agradecimentos	11
Lista de Figuras	13
Lista de Tabelas	14
Lista de abreviatura	15
1 Introdução	16
2 Revisão de literatura	18
2.1 Ferramentas genômicas para a identificação de ncRNAs	19
2.1.1 Microarranjo	19
2.1.2 Sequenciadores de alto desempenho	22
2.1.3 Illumina	23
2.2 RNAs não codificantes	26
2.2.1 Micro-RNA	27
2.2.2 Pequena RNA interferência (siRNA)	30
2.3 RNAs longos não codificantes (lncRNAs)	31
2.3.1 Classificação de lncRNAs	33
2.3.2 lncRNAs em plantas	34
2.4 Bioinformática	38
3 Objetivos	40
3.1 Objetivo Geral	40
3.2 Objetivos específicos	40
4 Material e métodos	40
4.1 Fonte de Dados	40

4.2	Avaliação de qualidade, alinhamento, montagem dos transcritos ...	41
4.3	Pipeline de identificação dos RNAs longos não codificantes	42
4.4	Identificação de lncRNAs diferencialmente expressos	44
4.5	Avaliação de diversidade nucleotídica dos lncRNAs	44
4.6	Comparação dos HC-lncRNAs com os lncRNAs previamente publicados	45
5	Resultados e discussões	46
5.1	Avaliação de qualidade e alinhamento	46
5.2	Estatísticas de montagem	48
5.3	Identificação dos RNAs longos não codificantes	48
5.4	Filtragem dos lncRNAs putativos	51
5.5	Análise de enriquecimento funcional	55
5.6	Identificação de lncRNAs diferencialmente expressos	57
5.7	Diversidade nucleotídica em lncRNAs	63
5.8	Comparação dos HC-lncRNAs com lncRNAs já anotados	66
6	Considerações finais e perspectivas	67
7	Conclusões	69
8	Referências Bibliográficas	70

Agradecimentos

É sempre muito complicado agradecer a todos, que de maneira direta ou indireta ajudaram para a realização de um trabalho de tamanho teor de conhecimento. Com a certeza que muitas outras pessoas contribuíram para a realização deste sonho, vou agradecer aqui os que de forma mais intensa auxiliaram na produção deste projeto.

Agradeço primeiramente a Deus que de forma singular me conduziu até aqui. Me dotou de capacidade, saúde e de sonhos para que hoje estivesse exatamente onde estou. Chegar até aqui não foi fácil, mas me proporcionou aprendizagem para novas metas e realizações.

Ao Professor Doutor Evandro Novaes, meu orientador e exemplo de profissional, pela confiança, apoio, amizade, paciência e compreensão que sempre teve. Sonho em um dia me tornar um 1/3 do profissional que você é hoje.

Aos meus avós, por me terem dado educação, valores e por terem me ensinado a andar. A vocês que, muitas vezes, renunciaram aos seus sonhos para que eu pudesse realizar o meu, partilho a alegria deste momento.

Aos irmãos que Deus colocou em minha vida: Pamella, Raquel, Lucas, Gabriel, Herick e Letícia. Amor incondicional, sempre. Seus corações estão comigo e o meu com vocês.

A todos os meus familiares, primos (as), tios (as), pai. Não citarei nomes, para não me esquecer de ninguém. Mas há aquelas pessoas especiais que diretamente me incentivaram. Quero deixar um agradecimento em especial para quatro pessoas que Deus colocou em minha vida, que sempre me incentivaram e me impulsionaram na vida pessoal e principalmente profissional. Obrigada Fernanda, Renata, Edson e Júnior por terem acreditado em mim. Espero profundamente nunca perder a amizade que se formou.

Aos colegas e amigos do Laboratório de Bioinformática: Ivone, Mariane, Rewther, Isabela, Stela, Lucas Fagundes, Lucas Vieira, Suzy e Guilherme pelos momentos de aprendizagem, pelos conselhos e pela amizade. Quero agradecer aos meus outros amigos: Nathália, Ana Paula, Ana Gabriela, Romes,

Marcos, karinna e Guilherme pela ajuda incondicional que sempre me ofereceram, quero vocês para sempre comigo.

A Capes pelo auxílio financeiro concedido. Ao programa de Pós-Graduação em Genética e Biologia Molecular (PGBM) e UFG por ter me dado esta oportunidade de conhecimento.

Agradeço a todos vocês que de alguma forma, fizeram parte desse projeto e me ajudaram a realizar mais uma etapa da minha vida. A todos vocês, meu muito obrigada.

Lista de figuras

Figura 1. Etapas do sequenciamento pela plataforma Illumina.....	25
Figura 2. Biogênese de miRNA em plantas.....	29
Figura 3. Quatro classes de lncRNAs.....	34
Figura 4. Pipeline computacional de identificação dos HC-lncRNAs.....	44
Figura 5. Exemplo de qualidade das sequências, obtidas na plataforma Illumina	46
Figura 6. Diagrama de Venn, demonstrando a quantidade de lncRNAs de alta confiança	52
Figura 7. Principais termos Gene Ontology “GO” em que os HC-lncRNAs estão envolvidos.....	57
Figura 8. Distância entre as bibliotecas de acordo com o nível de expressão dos HC-lncRNAs identificados.....	58
Figura 9. Estrutura secundária do lncRNA TCONS_00046489.....	62
Figura 10. Diversidade nucleotídica entre diferentes posições do genoma	65
Figura 11. Diversidade nucleotídica das quatro classes de lncRNAs	66

Lista de tabelas

Tabela 1. Plataformas de sequenciamento	23
Tabela 2. Informações do material estudado	41
Tabela 3. Resultados dos softwares FastQC e TopHat	47
Tabela 4. Estatísticas de montagem dos transcritos a partir das sequências de RNA total	48
Tabela 5. Resultado programa Cuffcompare	49
Tabela 6. Resultado dos lncRNAs identificados para as quatro classes selecionadas, após a remoção dos transcritos menores que 200 nucleotídeos	50
Tabela 7. lncRNAs das quatro classes identificados em cada um dos cromossomos de <i>A. thaliana</i>	50
Tabela 8. Distribuição de comprimentos dos lncRNAs de <i>A. thaliana</i>	51
Tabela 9. Quantidade de lncRNAs por cromossomo em <i>A. thaliana</i> , após a filtragem	53
Tabela 10. Comprimentos entre as classes de HC-lncRNAs em <i>A. thaliana</i> , após a filtragem	54
Tabela 11. Resultado Blast2GO, análise de enriquecimento funcional. P= Processo biológico, F= Função molecular, C= Componente celular	55
Tabela 12. lncRNAs diferencialmente expressos, entre bibliotecas controle e biblioteca exposta somente ao escuro	59
Tabela 13. lncRNAs diferencialmente expressos, entre as bibliotecas controle e 30 minutos de luz	60
Tabela 14. lncRNAs diferencialmente expressos, entre as bibliotecas controle e três horas de luz	61
Tabela 15. lncRNAs diferencialmente expressos, entre as bibliotecas controle e 24 horas dia longo	62
Tabela 16. Quantidade de SNPs e médias por região do genoma de <i>Arabidopsis thaliana</i>	64
Tabela 17. Quantidade de SNPs e médias por classe de lncRNA	65

Lista de abreviaturas

AGO	Protein Argonaute
BAM	Binary Sequence Alignment/Map
CBC	Cap-binding complex
DCL1	Dicer-like 1
DDL	DAWDLE (An RNA binding protein)
FLC	Flowering Locus C
FPKM	Fragments Per Kilobase of Exon Per Million Fragments Mapped
GFF	General Feature Format
GO	Gene Ontology
HYL1	HYPONASTIC LEAVES 1
lincRNA	Long Intergenic Noncoding RNA
lncRNA	Long Noncoding RNA
NAT	Natural antisense transcripts
ncRNA	Noncoding RNA
NGS	Next Generation Sequencing
npcRNA	non-protein-coding RNA
ORF	Open Reading Frame
RNA	Ribonucleic Acid
SAM	Sequence Alignment/Map
SE	SERRATE (Protein)
sRNA	Small RNA
tRNA	Transfer RNA
UTR	Untranslated Region

1 Introdução

De acordo com o dogma central da biologia, a grande maioria das moléculas de RNA servem apenas como um meio de transportar a informação genética do núcleo para a síntese de proteínas no citoplasma. Com o advento das novas plataformas genômicas, como os microarranjos e as plataformas de sequenciamento de alta performance (Metzker, 2010), tem sido verificado que as moléculas de RNA ocupam muitos outros papéis. As análises de RNA total vêm apontando que uma grande proporção do genoma de mamíferos é transcrito. No genoma humano, por exemplo, mais de 70% do seu DNA é transcrito (Djebali et al. 2012), ao passo que somente 1-2% é traduzido em proteínas. Essa enorme quantidade de RNAs não codantes (ncRNA) era tida como ruído da transcrição (Hüttenhofer et al. 2005). Porém, cada vez mais evidências indicam que os ncRNA têm papel importante no funcionamento dos genomas, seja via regulação da transcrição ou via remodelagem da cromatina (Bonsasio & Shiekhhattar 2014). Estudos vêm demonstrando, por exemplo, que muitos desses ncRNAs são diferencialmente expressos em diferentes tecidos e sob estímulos bióticos e abióticos (Liu et al. 2012; Wang et al. 2014; Chekanova J. A. 2015).

Os RNAs não codantes geralmente incluem aqueles estruturais e/ou de função bem conhecida no metabolismo basal, tais como os RNAs transportadores, ribossomais e nucleolares. Os ncRNAs são classificados em pequenos (< 50 pb), tais como os microRNAs e RNAs de interferência, intermediários (50-200 pb) e RNAs longos não codificantes ($\geq$ 200 pb), também conhecidos como lncRNAs (Boerner & McGinnis 2012). Os lncRNAs têm despertado grande interesse, dado que ocupam grande extensão no genoma dos eucariotos e ainda são escassos os trabalhos que desvendaram suas funções moleculares. Os lncRNAs são divididos em quatro classes dependendo da sua localização: lncRNA sense, lncRNA antisense, lncRNA intergênico e lncRNA intrônico (Quan et al. 2015).

Pesquisas têm apontado lncRNAs envolvidos em múltiplas funções regulatórias (Bai et al. 2014), como na regulação da expressão gênica

transcricional, pós-transcricional e a níveis epigenéticos (Quan et al. 2015). Com isso, esses lncRNAs podem representar novas estratégias moleculares para geração de biotecnologias benéficas, seja na medicina ou na agricultura.

Os RNAs longos não codantes vêm sendo cada vez mais estudados. Porém a grande maioria dos trabalhos sobre lncRNA foi realizada com espécies animais (Mattick & Makunin 2005; Bai et al. 2014). Esses estudos com animais evidenciaram o papel de lncRNAs em muitos processos biológicos, tais como: Inativação do cromossomo X (Vallot et al., 2013; Tian et al., 2010), imprinting genômico (Keniry et al., 2012) e aplicações no tratamento de doenças como câncer e Alzheimer (Wapinski & Chang 2011).

As novas tecnologias de sequenciamento possibilitaram a descoberta de muitos candidatos a lncRNAs. A quantidade de dados gerados facilitou a identificação e caracterização destes RNAs, porém novas ferramentas computacionais foram requeridas (Iwakiri et al. 2015). Alguns estudos recentes têm descrito funções aos lncRNAs em plantas como: Sousa et al. (2001); Hamburger et al. 2002; Franco-Zorrilla et al. (2007); Ben Amor et al. (2009); Swiezewski et al. (2009); Heo J. B. & Sung S. (2011); Ding et al. (2012); Zhu et al. (2013); Yu et al. (2013); entre outros.

Muitos lncRNAs conhecidos são transcritos pela RNA Polimerase II, mas alguns lncRNAs de plantas são produzidos pela RNA Polimerase IV e V (Chekanova 2015). Alguns possuem semelhanças com mRNAs, tais como a presença de um Cap na região 5', poliadenilação na região 3' e sofrem splicing, porém não codificam proteínas (Du Toit 2013). Geralmente, os lncRNAs diferem dos mRNAs pelo baixo nível de expressão e por não conter um quadro de leitura aberta (ORF), bem como por não serem conservados em nível de sequências (Rymarquis et al. 2008). Em linhas gerais os lncRNAs podem ser considerados uma grande coleção de transcritos, poliadenilados ou não, com baixo potencial de codificar proteínas (Liu et al. 2015). A maioria das regiões intergênicas de *Arabidopsis*, arroz e milho codificam lncRNAs que são poliadenilados e transcritos pela RNA Polimerase II (Liu et al. 2012).

Esses lncRNAs podem ser peças importantes no cenário de falta de conhecimento que ainda persiste sobre a regulação da expressão gênica em

organismos eucariotos. Recentemente, esses RNAs longos não codificantes passaram a ser ativamente estudados, mas ainda restam muitas questões por serem elucidadas.

A partir do levantamento de trabalhos já publicados de lncRNAs verificamos que em plantas o conhecimento sobre quantidade, extensão, funcionalidades, síntese, regulação e biogênese desses lncRNAs ainda é bastante limitado, mesmo em espécies como *Arabidopsis thaliana*, que é organismo modelo. Alguns trabalhos com *Arabidopsis thaliana* foram descritos, como a descoberta de 6.480 lincRNAs (lncRNAs intergênicos) na análise de 200 conjuntos de dados de transcrito, com a utilização de abordagem tiling array (Liu et al. 2012) e a identificação de 37.238 lncRNAs (lncRNAs antisenses naturais), também utilizando a técnica tiling array (Wang et al. 2014). Apesar da descoberta de uma grande quantidade lncRNAs utilizando os tiling arrays, com a grande eficiência dos sequenciadores de nova geração é possível identificar novos lncRNAs através da técnica de RNA-Seq. O sequenciamento em larga escala pode ser útil para caracterização de novos transcritos e para a compreensão da função e regulação dos lncRNAs.

Este trabalho teve como objetivo analisar os RNAs longos não codificantes no genoma de *Arabidopsis thaliana*, buscando a identificação e localização destes. Além disso, pretende-se analisar a expressão desses elementos em diferentes condições de luminosidade e avaliar a diversidade genética nos RNAs longos não codificantes e em outras regiões do genoma, para verificar se esses elementos estão sob pressão de seleção. Com isso, os resultados do presente trabalho podem melhorar o conhecimento da extensão e importância dos ncRNA no genoma de plantas.

2 Revisão de Literatura

O estudo dos ncRNAs vêm se beneficiando muito do grande desenvolvimento nas plataformas de análises genômicas, tais como os sequenciadores de DNA e técnicas para análise da expressão gênica em larga escala. Um dos primeiros métodos de sequenciamento descrito foi o de Sanger

(Sanger et al 1977). Esse método passou por vários desenvolvimentos que culminaram com a criação do primeiro sequenciador automático de DNA, o ABI 370 (Applied Biosystems), no ano de 1986.

Em 1995 métodos de microarranjo e de análise serial de expressão gênica (SAGE) forneceram a oportunidade de avaliação da expressão de milhares de genes de maneira eficiente (Morozova et al. 2009). Essas plataformas causaram e vêm causando uma revolução em diversas áreas da biologia, incluindo o estudo dos ncRNAs. Atualmente com o avanço de tecnologias, novas metodologias se tornaram essenciais no cenário da genômica como o sequenciamento de nova geração, ou NGS (Next Generation Sequencing). Proporcionou um aumento exponencial na quantidade de bases sequenciadas e, alterou de forma significativa o custo benefício para atingir o conhecimento de sequências de DNA e RNA.

2.1 Ferramentas genômicas para a identificação de ncRNAs

2.1.1 Microarranjo

A técnica de microarranjo de DNA, possibilita mensurar o nível de expressão gênica, possibilitando a análise genômica (Mockler T. & Ecker J. 2005). Avanços na tecnologia de microarranjo trouxe a possibilidade de avaliar os perfis de expressão gênica em larga escala, com o objetivo de caracterizar a expressão em diferentes tipos de células sob diferentes condições e estímulos (Morozova et al. 2009). Embora o microarranjo tenha a capacidade de medir a expressão de vários genes simultaneamente, o método não abrange a detecção de novos transcritos e, por ser baseado em hibridização não permite um estudo detalhado da sequência de genes e transcritos (Morozova et al. 2009).

Métodos baseados em microarranjo também podem ser utilizados para identificar rearranjos no genoma, porém com resultados limitados. Para detecção de polimorfismos e rearranjos genômicos, as tecnologias de nova geração oferecem vantagens sobre os microarranjos (Morozova et al. 2009).

Microarranjo surgiu para revolucionar a genômica funcional, permitindo que todos os transcritos de um organismo fossem estudados em um único ensaio. Essa técnica determina os níveis de expressão de milhares de genes em paralelo. Os microarranjos primeiramente foram construídos a partir de fragmentos de cDNA roboticamente impressos e imobilizados em lâminas de vidro, similares às de microscópio (Schena et al. 1995). Embora os microarranjos de cDNA sejam úteis, esses vêm caindo em desuso pois requerem a impressão e manutenção de lâminas geralmente in house, apresentam problemas de desuniformidade entre lâminas e spots (impressões) de cDNA, e necessitam de clonagem e miniprep das bibliotecas de cDNA (Close et al. 2004).

Os microarranjos de cDNA vêm sendo substituídos pelos oligo arrays, que são construídos a partir de oligonucleotídeos sintetizados diretamente na própria lâmina (Meyers et al. 2004). Uma razão para o domínio dos oligonucleotídeos é que eles podem ser sintetizados sem placas ou diretamente em superfícies sólidas, aumentando a uniformidade entre os diferentes microarranjos e entre as diferentes sondas (spots) sintetizadas. Com isso, os microarranjos de oligonucleotídeos oferecem maior reproducibilidade nas análises quando comparados com os de cDNA. Para uma planta bem descrita como *Arabidopsis*, os clones de cDNA podem representar menos de 60% dos genes previstos (Wortmann et al. 2003).

A detecção dos níveis de expressão dos genes é baseada na hibridização de transcritos (RNA ou cDNA) nas sondas impressas (cDNA) ou sintetizadas na lâmina do microarranjo. O sinal da hibridização é obtido porque a amostra de RNA hibridizada é marcada pela incorporação de moléculas fluorescentes, sendo os mais usados “Cy3” e “Cy5”, através de uma reação de transcrição reversa (Schena 1996). Um microarranjo é uma coleção formada por sequências de DNA fixados e ordenados em áreas específicas chamadas de spots. Na lâmina, cada spot contém milhões de cópias de um único transcrito. Os segmentos são chamados de sondas, que são ligadas ao suporte onde serão realizadas reações de hibridização, por complementariedade entre a sonda e o seu mRNA correspondente transformado em cDNA, que é isolado

do tecido ou célula alvo das amostras. As sequências de cDNA a serem analisadas são marcadas com corantes fluorescentes e colocadas para reagir com sequências fixadas na superfície do microarranjo. Os corantes incorporados ao cDNA possuem espectros de emissão e excitação diferentes, permitindo a quantificação individual de até 2-3 amostras hibridizadas simultaneamente em um mesmo microarranjo.

Após o procedimento as lâminas são lavadas para a remoção dos cDNAs que não se ligaram às sondas, são expostas à ação de raios laser que irão excitar os corantes, fazendo que estes emitam luz. A lâmina é escaneada e a intensidade da fluorescência do material é medida. Cada spot de uma lâmina de microarranjo representa o valor da atividade transcricional de um determinado gene, em situações biológicas diferentes (Giachetto 2010).

Uma variação da tecnologia de microarranjo são os chamados tiling arrays, considerados poderosos instrumentos de investigação da transcrição em escala genômica (Mockler et al. 2005). Esta técnica possui os princípios básicos do experimento de microarranjo, diferindo na natureza dos fragmentos genômicos investigados, isto é as sondas.

Nos tiling arrays, as sondas geralmente cobrem o genoma inteiro (Liu 2007). A partir desta técnica é possível mensurar a expressão gênica, realizar anotação experimental de genomas, determinar a estrutura dos transcritos de genes, estudar fatores de transcrição em redes de regulação gênica, encontrar elementos ainda não descritos dos genomas, como novas sequências transcritas, elementos regulatórios, entre outros. Como todo o genoma é coberto e avaliado para verificar a existência de transcrição, esta é uma poderosa técnica para a descoberta de ncRNAs. Outra aplicação popular para o tiling array é a plataforma ChIP-chip¹⁰ que localiza no genoma regiões promotoras que interagem com diferentes fatores de transcrição. No experimento há um ensaio de imunoprecipitação da cromatina (ChIP) acoplado a hibridização em um chip de tiling array (Bulyk 2006; Liu 2007).

Alguns estudos aplicaram o método tiling array para o estudo de modificações da histona e metilação de bases citosina, combinando a análise de expressão gênica e metilação do DNA. Essas análises revelaram que a

metilação do DNA regula os níveis de expressão de muitos ncRNAs (Yazaki et al. 2007). Apesar da ampla utilização do método tiling array para identificação de novos transcritos, este geralmente subestima a complexidade do genoma, dado que sua baixa sensibilidade pode impedir a detecção de transcrições raras (Marguerat 2010).

O método tiling array tem algumas desvantagens pois requer a existência de sequência do genoma, impossibilitando seu emprego em espécies que não possuem a sequência de um genoma de referência (Quan et al. 2015). Este problema não afeta espécies modelo que têm sido amplamente estudadas com genome tiling arrays, como *Arabidopsis thaliana*. As dificuldades do tiling array podem ser superadas pela técnica de RNA-seq e outras técnicas que estão sendo criadas para identificar a complexidade do transcrito com alta precisão (Marguerat et al. 2010).

2.1.2 Sequenciadores de alto desempenho

Após a criação dos sequenciadores automatizados, esses foram aprimorados e a partir de 2005 surgiram os sequenciadores de nova geração, tais como as plataformas da Illumina, 454/Roche, Solid, HeliScope (Morozova et al. 2009). O sequenciamento de nova geração utiliza metodologias que superam as principais dificuldades do método de Sanger – isto é, necessidade de clonagem e de eletroforese para realização do sequenciamento. Com isso, essas plataformas de nova geração possuem alta eficiência e custo reduzido quando comparadas ao método de Sanger.

As tecnologias de sequenciamento de segunda geração, como 454, Illumina e Solid, incluem etapas como a preparação do DNA molde, incorporação sequencial de novos nucleotídeos, captura de imagens, sequenciamento e análise de dados. A combinação de protocolos específicos diferencia uma tecnologia de outra e determina o tipo de dado produzido em cada plataforma. Com essas diferenças, as aplicações biológicas, custo, tempo e a qualidade dos dados obtidos também diferem entre as plataformas (Metzker 2010).

Tabela 1: Plataformas de sequenciamento. Fonte: Glenn 2011.

Plataforma de sequenciamento	Roche (454)	Illumina	ABI Solid	PacBio
Química do sequenciamento	Pirosequenciamento sobre suporte sólido	Sequenciamento por síntese com terminadores reversíveis	Sequenciamento por ligação	Tempo real de molécula única
Método de amplificação	PCR em emulsão	PCR em ponte	PCR em emulsão	-
Comprimento da leitura	230-400 bp	75-250 bp	25-35 bp	1-60 kbp

O desenvolvimento das plataformas de sequenciamento de alto desempenho vem sendo acompanhado pela rápida evolução de novas ferramentas de bioinformática. A bioinformática é absolutamente necessária para as análises do grande volume de dados gerados por essas plataformas. Todo esse desenvolvimento tem causado uma profunda revolução em diversas áreas da biologia, incluindo os estudos que levaram à identificação e isolamento de muitos RNAs não codificantes de proteínas (ncRNAs).

2.1.3 Illumina

Esse método de sequenciamento desenvolvido pela empresa Solexa, foi adquirido pela Illumina que vem aprimorando o processo. O sequenciamento da Illumina é realizado por síntese nas quais os quatro nucleotídeos são adicionados sincronicamente, usando DNA polimerase e nucleotídeos terminadores marcados, sendo cada incorporação um evento único (Mardis 2008).

Os fragmentos de DNA da amostra são ligados aos adaptadores em ambas as extremidades, o que permite sua fixação ao suporte (lâmina) de sequenciamento por hibridização a um dos oligonucleotídeos fixados na lâmina. Em seguida, ocorre a clonagem in vitro dos fragmentos nesta superfície sólida de vidro através de um processo conhecido como PCR de fase sólida ou PCR em ponte (bridge PCR) (Fedurco et al. 2006; Turcatti et al. 2008). Após o passo de amplificação podem ser formados até 10 bilhões de aglomerados (clusters), onde cada aglomerado é composto por cerca de 1000 cópias clonais de uma única molécula.

As moléculas de DNA clonadas nos clusters são sequenciadas de forma paralela utilizando uma abordagem de síntese de DNA que emprega nucleotídeos especiais contendo terminadores reversíveis e molécula fluorescente com porção removível. A incorporação desses nucleotídeos impede que novas bases sejam incorporadas até que o terminador seja revertido. Os terminadores são marcados com quatro cores diferentes para distinguir entre as diferentes bases, sendo que a posição e a sequência molde de cada cluster são deduzidas pela leitura da cor em cada passo de adição do nucleotídeo (Bentley 2006). Um algoritmo atribui sequências e valores de qualidade para cada leitura, avaliando os dados do Illumina, removendo sequências com baixa qualidade, figura 1.

Atualmente, a análise do genoma por Illumina domina o mercado NGS. Tecnologias de sequenciamento de nova geração tem influenciado o estudo aprofundado de ncRNAs.

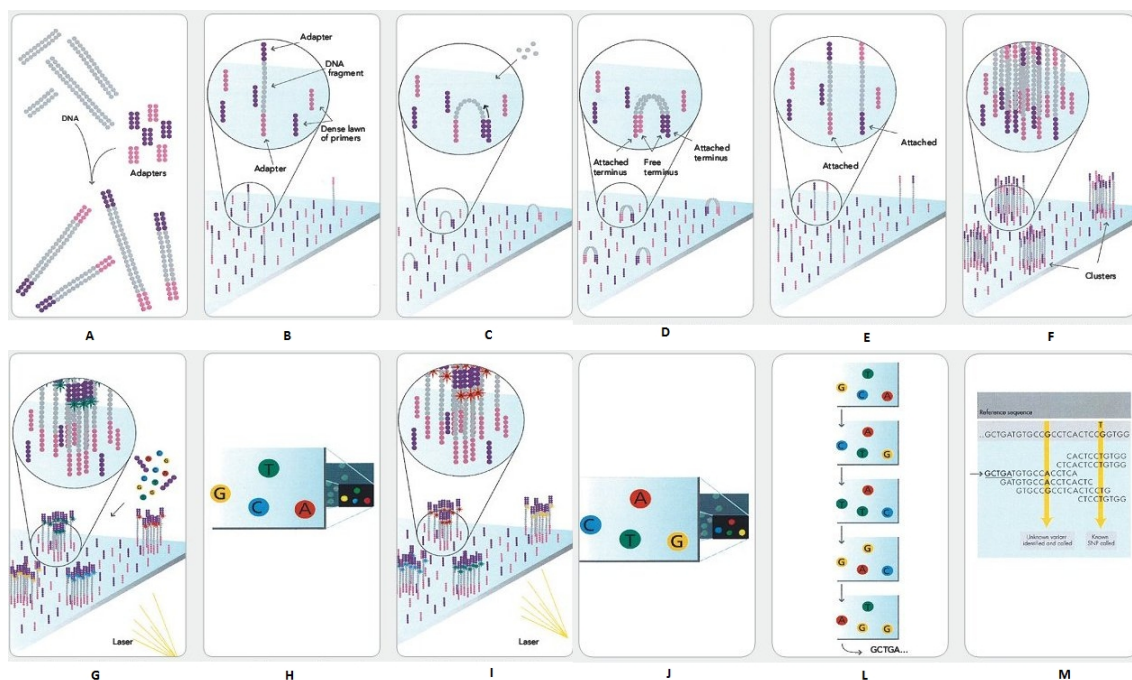


Figura 1. Etapas do sequenciamento pela plataforma Illumina. A) Os adaptadores se ligam aos fragmentos de DNA; B) Anelamento dos fragmentos fita simples aos primers; C) O adaptador da extremidade livre da molécula aderida ao suporte encontra seu oligonucleotídeo complementar no suporte, formando uma estrutura em ponte; D) A PCR é iniciada utilizando a extremidade 3' livre do oligonucleotídeo como primer; E) A ponte é desfeita pela elevação da temperatura; F) Após uma série desses ciclos, serão obtidos clusters de moléculas idênticas ligadas ao suporte; G - H) Incorporação de nucleotídeos terminadores marcados, I – J) Excitação a laser; K– L) A leitura é feita de forma sequencial, o que permite a montagem da sequência completa de cada cluster; M) Leitura da sequência do fragmento. Fonte: Carvalho & Silva (2010).

Quatro técnicas para identificação de lncRNAs envolvidos na regulação gênica, são utilizadas como: microarranjo, imunoprecipitação de RNA e cromatina, tiling array e RNA-seq (Lee C. & Kikyo N. 2012). Cada metodologia com suas vantagens e desvantagens para a compreensão destes transcritos. O RNA-seq é uma metodologia revolucionária em estudo de transcriptoma, principalmente para descoberta de novos transcritos, pela sua maior

sensibilidade e maior desempenho (Wang et al. 2009), causando uma maior acurácia em estimar a densidade de transcritos (Roberts et al. 2011).

2.2 RNAs não codificantes

A existência de ncRNAs foi proposta em 1961 por Jacob e Monod, praticamente junto com o desvendamento do código genético e consequente elucidação dos genes codificadores de proteínas. A partir desse momento, o foco foi nos genes codificadores de proteína, enquanto a área de ncRNAs evoluiu mais lentamente devido a uma crença de que esses possuem papel secundário no funcionamento celular, sendo considerados até mesmo como ruído transcricional. Mais recentemente, novas descobertas foram feitas, tais como a capacidade do RNA para realizar função catalítica, bem como a observação de que muitas sequências não codantes apresentam conservação muitas vezes similar a de regiões codantes (Gorodkin J. 2014).

Pequenos RNAs não codantes (ncRNAs) são reguladores cruciais do desenvolvimento. Estas moléculas de RNAs com comprimento de 18-30 nucleotídeos são transcritas, porém não traduzidas em proteínas. Sobre os RNAs não-codificantes, os mais conhecidos são: o RNA transportador (tRNA), responsável pelo transporte de aminoácidos e o RNA ribossomal (rRNA), que possui papel estrutural (Watson 2006). Os RNAs codificantes incluem, basicamente, os RNAs mensageiros (mRNA) que, após sintetizados, são processados e dão origem às proteínas. Os RNAs não-codificantes podem ser de vários tipos e desempenhar funções variadas dentro da célula. Dentre os principais ncRNAs estão: tRNA, rRNA, snoRNA, miRNA, siRNA, piRNA (Dias Correia 2007). Algumas classes de RNAs, tais como os tRNAs, snoRNAs ou miRNAs, se caracterizam por sequências motivos e/ou características da estrutura secundária em comum. Os tRNAs estão entre os mais bem conservados tanto em termos de sequência como de estrutura secundária (Gorodkin J. 2014; Eigen M. 1989).

Os pequenos RNAs ou sRNAs (do inglês Small RNAs) são abundantes em plantas e, geralmente, estão envolvidos no silenciamento da expressão

gênica. O fenômeno compartilha quatro passos consenso: (1) indução por fita dupla do RNA (dsRNA), (2) transformação do sRNA em 18-25 nucleotídeos, (3) metilação na região 3', (4) incorporação do sRNA no complexo efetor que é parcialmente ou totalmente complementar ao RNA alvo (Chapman & Carrington 2007). A dupla fita do RNA pode derivar de replicação de vírus, transcrição de sequências de repetição invertida ou transcrição convergente. A geração de dsRNA pode ser “geneticamente programada” em locos endógenos que produz transcrições com estruturas de hairpin loop (Voinnet 2009). Independentemente da origem, a geração de dsRNA geralmente desencadeia o mecanismo de silenciamento gênico.

Duas classes de sRNAs têm sido amplamente estudadas em muitos processos como diferenciação celular e oncogênese, são os micro-RNAs (miRNAs) e os pequenos RNAs interferência (siRNAs). Estes RNAs não codificantes servem como reguladores da expressão gênica em vários organismos. Eles se ligam a sequências complementares geralmente encontradas nas regiões UTR de genes e induz a degradação do RNA mensageiro (mRNA) alvo, regulando suas taxas de tradução (Filipowicz W. et al. 2008). Ambos, siRNAs e miRNAs, estão envolvidos no processo de silenciamento pós-transcricional e/ou interferência mediada por RNA (Kawasaki et al. 2004).

2.2.1 Micro-RNA

Os micro-RNAs (miRNAs) possuem cerca de 23 nucleotídeos de tamanho e compreendem uma grande família de ncRNAs curtos que atuam como importantes reguladores pós-transcricionais da expressão gênica em animais e plantas. Dentre algumas funções de miRNAs está o amplo fenômeno conhecido como silenciamento do RNA. Os miRNAs diferem de outras classes de RNAs pequenos em sua biogênese, pois derivam de transcritos que se dobram formando uma estrutura em grampo (hairpin) (Bartel 2009). Eles controlam a expressão de genes alvo através da repressão da tradução ou

degradação do mRNA (Edwards et al. 2009). Em plantas os miRNAs são a segunda classe mais abundante de sRNAs (Voinnet 2009).

Estudos recentes na área de genética e bioquímica buscando entender sua biogênese e modo de ação, indicam que os miRNAs de plantas compartilham muitas semelhanças com miRNAs de animais. Estes trabalhos descobriram processos reguladores que controlam o processamento e a atividade de miRNAs em plantas e mostraram que os miRNAs e siRNAs em plantas estão intimamente ligados em termos de origem e modo de ação (Voinnet 2009).

A figura 2 demonstra a via molecular do processamento e estabilidade conservada de miRNAs em plantas. Os genes que codificam os miRNA são também denominados MIR. Os precursores dos miRNAs são em sua maioria transcritos pela RNA Polimerase II (Pol II) de regiões situadas entre genes codificantes de proteínas, gerando os chamados pri-miRNAs. A proteína de ligação de RNA DAWDLE (DDL) estabiliza o pri-miRNA para conversão desses, em centros de processamento nuclear, em pre-miRNA. Esta reação envolve a ação e a interação física entre a proteína SERRATE (SE), a proteína HYPONASTIC LEAVES 1 (HYL1) que se liga a fita dupla de RNA, dicer-like1 (DCL1) e o complexo nuclear de ligação ao CAP (CBC). Posteriormente o pre-miRNA produzido pela DCL1 é então exportado para o citoplasma, possivelmente através da ação da exportina 5 ortóloga a HASTY e outros fatores desconhecidos. A fita dupla de RNA é metilada pela HEN1 para evitar sua degradação por exonucleases. Uma fita é degradada e a outra segue como fita guia onde é incorporada em proteínas ARGONAUTAS (AGO) para a reação de silenciamento.

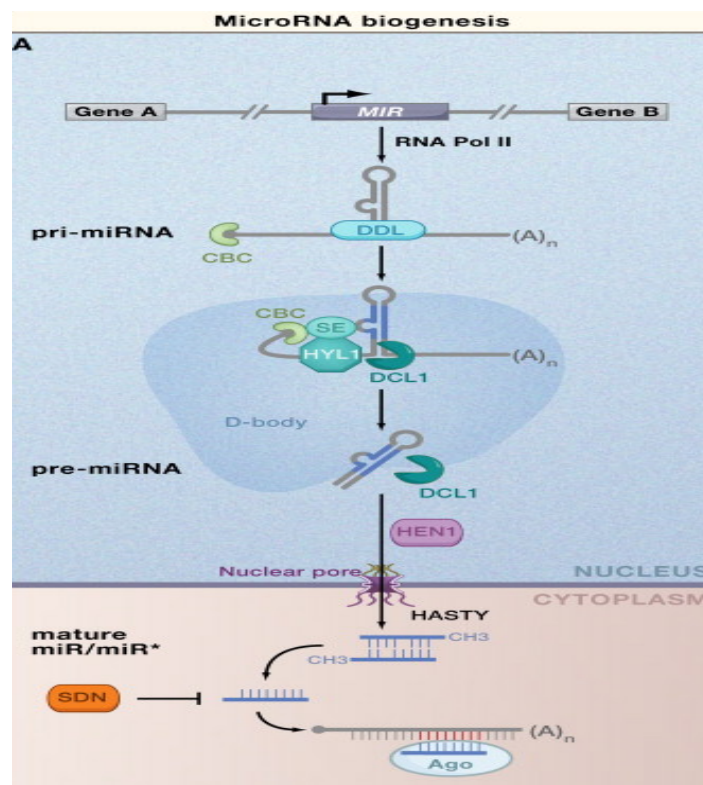


Figura 2. Biogênese de miRNA em plantas. Fonte: Voinnet 2009.

Um componente crucial na análise de miRNA é a região promotora. As sequências upstream à região de início da transcrição de genes transcritos pela RNA polimerase II modulam a expressão do gene (Megraw 2006). As sequências na região 5' upstream de genes são de primária importância na regulação gênica de *Arabidopsis* (Lee et al. 2006). MiRNAs não atuam de maneira independentemente, eles estão imersos em complexas vias que incluem interações com fatores de transcrição.

Nos promotores dos genes MIR há presença de sítios de ligação para fatores de transcrição (FTs). Dentre os FTs que se ligam ao promotor dos MIR, foram identificados fatores de transcrição do tipo AUXIN RESPONSE FACTOR (ARF), LFY que é induzido por giberelina e ativa genes homeóticos florais e BASIC HELIX-LOOP-HELIX LEU ZIPPER TRANSCRIPTION FACTOR (MYC2) que atua em vias genéticas que incrementam a capacidade de resposta ao estresse hídrico por meio do aumento da sensibilidade ao fitohormônio ácido abscísico. A descoberta dessas interações indicam que a transcrição de miRNA

está intimamente ligada à resposta das plantas aos hormônios vegetais. Além disso algumas dessas famílias de fatores de transcrição são regulados pelos próprios miRNAs, indicando um mecanismo de regulação via feedback (Megraw et al. 2006).

2.2.3 Pequenos RNAs interferência (siRNA)

Os siRNAs são os mais abundantes dentre os pequenos RNAs de plantas. Essas moléculas são responsáveis pelo mecanismo de “silenciamento gênico” que consiste em regular de forma negativa a expressão de um ou mais genes. Este mecanismo desempenha um importante papel na defesa e retrotransposons contra vírus (Vaucheret et al. 2001).

Atualmente, são reconhecidas três formas distintas de siRNAs envolvidas no controle da expressão de genes de plantas: trans-acting siRNAs (ta-siRNAs), repeat- associated siRNAs (ra-siRNAs) e natural-antisense siRNA (nat-siRNAs) (Vazquez 2006).

Os ta-siRNAs são gerados a partir de precursores TAS, que codificam transcritos não codificadores de proteínas, que são alvos de miRNAs incorporados a AGO1 e AGO7, ou seja são dependentes de miRNA (Allen et al. 2005). Os fragmentos gerados são estabilizados pela proteína SUPPRESSOR OF GENE SILENCING3 (SGS3) e convertidos em RNAs fita-dupla pela RNA polimerase dependente de RNA 6 (RDR6). A enzima DCL4, em conjunto com a enzima DRB4, cliva o RNA fita-dupla gerando o duplex imperfeito, que é metilado pela HEN1 e transportado para o citoplasma pelo fator de exportação SDE5. No citoplasma o ta-siRNA, que possui 21 nucleotídeos, é incorporado a AGO1 ou AGO7 e efetua a clivagem de seus alvos. Em contraste com outros siRNAs, que regulam o seu gene ou loci precursor, os tasiRNAs atuam em trans, regulando outros genes (Llave, 2010). Foram descritos em Arabidopsis 6 loci capazes de gerar ta-siRNAs: TAS1a, b e c, TAS2, TAS3 e TAS4 (Vazquez et al. 2004).

Os ra-siRNAs são gerados a partir de regiões genômicas com sequências repetitivas. A ação dos ra-siRNAs resulta na inibição epigenética da transcrição, principalmente de genes importantes para a transposição e

replicação de elementos genéticos móveis. Para a formação dos ra-siRNAs de 24 nucleotídeos, transcritos dos elementos repetitivos são convertidos em dsRNA pela RDR2e clivados pela DCL3. Estes servem de guia para a metilação dos loci correspondentes através de AGO4 e das metiltransferases DRM1 e DRM2. A RNA polimerase IV (Pol IV) é necessária tanto para a metilação do DNA quanto para a produção dos ra-siRNA. O complexo enzimático envolvido no silenciamento gênico transcricional tem sido denominado de RITS (RNA-induced initiation of transcriptional gene silencing) (Xie et al. 2004).

Outra forma distinta de si-RNAs são os nat-siRNAs, que são originados de dsRNAs formados por homologia de sequência de dois transcritos e presume estar envolvidos nas respostas de plantas a estresses. No trabalho de Borsani et al., em 2005 encontraram em *Arabidopsis*, dois tipos de siRNAs gerados por dois genes: P5CDH relacionado com stress e SRO5 com função desconhecida. Quando ambas as transcrições estão presentes é formando um nat-siRNA com 24 nucleotídeos. O nat-siRNA de 24 nucleotídeos serve de guia para a clivagem do transcrito de P5CDH. O produto clivado é convertido em dsRNA, sofre cortes em série pela DCL1, gerando nat-siRNAs de 21 nucleotídeos. Neste caso, apenas um dos transcritos que originaram o siRNA é clivado. Diferentemente de ta-siRNAs, essa classe atua em cis, regulando seu gene precursor (Borsani et al. 2005).

2.3 RNAs longos não codificantes (lncRNAs)

O avanço e desenvolvimento de novas plataformas genômicas vêm gerando conhecimento revolucionário nas áreas de genética e biologia molecular. Esses avanços têm alterado conceitos fundamentais da biologia, como o conceito de gene (Djebali et al. 2012) e o dogma central da biologia (Zhang et al. 2013). No centro dessa revolução estão os RNAs não codantes, cunhados como matéria negra (dark matter) transcricional devido à falta de conhecimento sobre o seu papel no funcionamento genômico, especialmente dos lncRNA (Ponting & Grant Belgard 2010). Após a descoberta da existência

de RNAs não codificantes, essa área está em constante desenvolvimento. Com isso, surge a necessidade de sistematizar os dados e resultados já descobertos. Estudos básicos envolvendo os lncRNA assumiram uma grande importância, dado a descoberta de sua preponderância nos genomas de eucariotos.

Os genomas de eucariotos produzem transcritos de vários tamanhos, de RNAs mensageiros codificantes de proteínas a transcritos curtos não codificantes. Após a descoberta de ncRNAs, boa parte dos trabalhos focaram nos pequenos ncRNAs gerando um bom conhecimento sobre seus mecanismos de regulação e suas funções em diferentes organismos. Trabalhos recentes têm revelado a importância de RNAs longos não codificantes (lncRNAs), sendo estes definidos como ncRNAs com mais de 200 nucleotídeos de comprimento. Estudos revelaram que os lncRNAs estão envolvidos na regulação da expressão gênica transcricional, pós-transcricional e em níveis epigenéticos. Porém poucos estudos até o momento exploraram as funções e mecanismos de ação dos lncRNAs, a fim de compreendê-los (Quan et al. 2015).

A grande maioria dos trabalhos sobre esse grupo de RNAs não codificantes, são estudos com animais, evidenciando o papel de lncRNAs em muitos processos biológicos, tais como: Inativação do cromossomo X (Vallot et al. 2013; Tian et al., 2010), imprinting genômico (Keniry et al. 2012) e aplicações no tratamento de doenças como câncer e Alzheimer (Wapinski O. & Chang H. Y. 2011). O trabalho de Quan et al. (2015), fornece uma revisão dos lncRNAs em animais e a partir destes estudos podem inferir perspectivas na compreensão para papéis de lncRNAs em plantas. Com essas evidências crescentes que envolvem papéis biológicos importantes aos lncRNAs em animais, torna-se necessário a análise abrangente destes em espécies vegetais.

lncRNAs que são transcritos de regiões do genoma em torno de sítios de início de transcrições, regiões enhancer, sítios de splicing e sítios de terminações de transcrições são descritos como lncRNAs instáveis (Liu et al. 2015).

A grande maioria dos lncRNAs são transcritos pela Polimerase II (Ulitsky & Bartel, 2013). Alguns lncRNAs possuem semelhanças com o mRNA, tais como cap na região 5', poliadenilação na região 3' e sofrem splicing, porém não codificam proteínas (Du Toit 2013). Geralmente, os lncRNAs diferem dos mRNAs pelo baixo nível de expressão e por não conter um quadro de leitura aberta (ORF) extenso e evolutivamente conservado (Rymarquis et al. 2008). Em linhas gerais os lncRNAs podem ser considerados uma grande coleção de transcritos, poliadenilados ou não, com baixo potencial de codificar proteínas (Liu et al. 2015). A maioria das regiões intergênicas de *Arabidopsis*, arroz e milho codificam lncRNAs que são poliadenilados e são transcritos pela Polimerase II, com alguns sendo transcritos pela Polimerase IV e/ou V (Liu et al. 2012).

Os lncRNAs podem também apresentar um padrão de co-expressão com genes codificantes vizinhos, para assim ativar ou reprimir a expressão gênica desses vizinhos, ou atuar por exemplo em trans em alvos distantes. A interação desses elementos em cis ou trans pode ocorrer via mecanismos de regulação transcricional, por exemplo, inativação de promotor por ligação a fatores de transcrição basais, ativação de uma proteína acessória, transporte de fatores transcricionais e repressão epigenética (Ponting et al. 2009).

2.3.1 Classificação dos lncRNAs

Os lncRNAs podem ser subdivididos em quatro categorias, como demonstrado na figura 3, de acordo com sua proximidade em relação a genes codificantes de proteínas: lncRNAs sense (quando sobrepõem-se com um ou vários éxons de um gene e é transcrito a partir da mesma fita), lncRNAs antisenses (quando sobrepõem-se com um ou vários éxons de um gene e sua transcrição ocorre em fita oposta), lncRNAs intrônicos (derivam de um íntron dentro de outro gene) e lncRNAs intergênicos ocorre no intervalo entre dois genes (Quan et al. 2015).

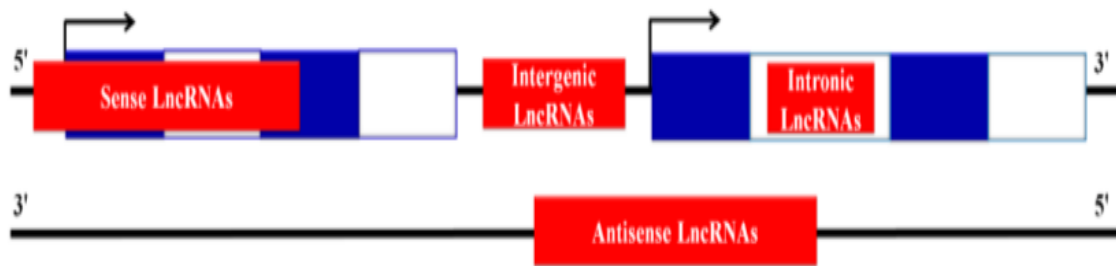


Figura 3. Quatro classes de lncRNAs. Em azul indicam éxons e em branco íntrons. As linhas em preto representam as fitas de DNA com regiões codificantes e não codificantes. Em vermelho indicam os lncRNAs, e as setas indicam a direção da transcrição. LncRNAs sense sobrepõem a genes codificantes na mesma fita; lncRNAs antisense sobrepõem com genes codificantes na fita oposta; lncRNAs intrônicos ocorrem completamente dentro de íntron e lncRNAs intergênicos ocorrem entre dois genes. Fonte: Quan et al. 2015.

Alguns autores classificam os lncRNAs em três tipos de acordo com sua origem genômica: RNAs longos intergênicos (lincRNAs), ncRNAs intrônicos (incRNAs) e os transcritos antisense natural (NATs) transcritos da cadeia complementar a de seus genes associados (Liu et al. 2015). Uma pequena quantidade de lncRNAs em animais e plantas tem sua sequência primária conservada evolutivamente englobando até mesmo alguns com funções reguladoras, sugerindo uma rápida evolução (Liu et al. 2012). Alguns lncRNAs podem se dobrar formando estruturas específicas para interagir com proteínas de ligação a RNA. Estes lncRNAs podem ter estruturas secundárias evolutivamente conservadas (Wang et al. 2014).

2.3.2 lncRNAs em plantas

Em comparação com animais, estudos de lncRNAs em plantas estão apenas começando. Vários lncRNAs foram identificados por análise do genoma inteiro usando tiling array e RNA-seq em plantas. Foram descobertos em

Arabidopsis thaliana, mais de 6.000 lncRNAs intergênicos e 37.238 transcritos pares senso e antisenso, analisados por tiling array (Liu et al. 2012).

Na espécie *Zea mays* (milho) foi realizada a caracterização de sequências de lncRNAs, utilizando dados públicos já disponíveis (Li et al. 2014). Exploraram conjuntos de dados de RNA-seq de 30 diferentes experimentos, foram identificados 20.163 lncRNAs putativos, dos quais 90% foram preditos como precursores de pequenos RNAs e 1.704 foram considerados lncRNAs de alta confiança. A maioria desses lncRNAs foram classificados como intergênicos, com apenas um éxon, e expressos em tecido específico (Li et al. 2014).

Recentemente, 867 novos lncRNAs de alta confiança foram identificados no genoma de *Zea mays*, utilizando a tecnologia de sequenciamento PacBio (Wang et al. 2016). A média de comprimento destes RNAs foi superior a outros trabalhos já publicados, com outras espécies vegetais e animais. Isso deve ter ocorrido devido ao fato do PacBio ser capaz de sequenciar longos fragmentos de DNA (> 1 kpb). Os lncRNAs foram classificados em quatro grupos de acordo com sua posição genômica, sendo 58% identificados em regiões intergênicas, 21% antisenses, 16% senses e 5% em regiões intrônicas de mRNAs. A maioria apresentou apenas um éxon, um perfil de expressão tecido específico e uma baixa expressão. O estudo observou que os lncRNAs com múltiplos exons são expressos de forma mais intensa que os lncRNAs de único éxon (Wang et al. 2016).

Um importante trabalho em *Populus tomentosa*, caracterizou lncRNAs na espécie e realizaram estudo de expressão diferencial destes RNAs em resposta ao déficit de nitrogênio (Chen et al. 2016). Foram identificados 388 candidatos a lncRNAs, destes 126 foram significativamente alterados sob essa baixa disponibilidade de nitrogênio, onde 118 foram induzidos e 8 restantes reprimidos. Baseado em sua posição genômica estes transcritos foram categorizados em quatro classes de acordo com o código da ferramenta Cuffcompare: intergênicos, antisenses, sobrepostos com exons genéricos do transcrito referência e potenciais fragmentos de novas isoformas. A maioria dos lncRNAs (87.37%) foi classificada como intergênico (Chen et al. 2016).

Em *A. thaliana*, Ben-Amor et al. (2009) encontraram 76 ncRNAs, destes 33 foram classificados como lncRNAs, por variarem de 265-1879 nucleotídeos, com uma média de comprimento de 1013 nucleotídeos. Os autores destacaram que, destes RNAs, um terço abriga um ou vários íntrons. Ainda sobre estes 76 ncRNAs, 34 locus podem dar origem a pelo menos um pequeno RNA, ou seja, abrangem locus de pequenos RNAs.

Em um estudo com 200 conjuntos de dados de transcrito de *A. thaliana*, identificaram 13.230 transcritos intergênicos, sendo que destes 6.480 foram considerados lincRNAs. Os dados vieram de experimentos utilizando o método de tiling array, com 14 mutantes de *Arabidopsis*, de 18 diferentes tratamentos ao estresse e seis diferentes órgãos/tecidos. A expressão de 2.708 lincRNAs foi detectada por experimento de sequenciamento de RNA (Liu et al. 2012).

Ainda na espécie *A. thaliana* encontraram um total de 37.238 pares de transcritos NATs, utilizando a abordagem tiling array. O comprimento variou de 200 a 12.370 nucleotídeos, com uma média de tamanho de 731 nucleotídeos. Descobriram que 2.186 pares NATs eram compostos de 2 lncNATs complementares (Wang et al. 2014).

Alguns estudos recentes tem descrito funções biológicas de lncRNAs em plantas. Ding et al. (2012) propuseram que uma determinada quantidade do lncRNA, denominado LDMAR é necessário para a fertilidade masculina sob condições de dias longos em arroz. Uma mutação espontânea de G-C levando a um polimorfismo entre 58N e 58S, altera a estrutura secundária do RNA na região. Essa alteração aumenta a metilação da região promotora de LDMAR, reduzindo o nível de transcrição do lncRNA, especificamente sob condições de dias longos. O baixo nível de LDMAR resulta morte celular programada no desenvolvimento da antera sob condições de dias longos, causando a esterilidade masculina. O lncRNA identificado fornece oportunidade adicional para revelar papéis destes em plantas.

Em *Medicago truncatula*, mostrou-se que o lncRNA Enod40 codifica dois peptídeos curtos e um ncRNA com estrutura secundária estável funcional. Uma alteração em elementos estruturais do RNA de Enod40, mas que retém

íntegros os dois peptídeos curtos, resultou na diminuição da formação de nódulos e ativação da divisão da célula cortical em *M. truncatula*. Os elementos estruturais do RNA permanecem mais conservados do que os peptídeos codificados. Enod 40 também pode alterar a localização da proteína MtRBP1. Em *M. truncatula* durante a nodulação o Enod40 interage diretamente com a proteína e muda a localização de MtRBP1 para o citoplasma em células da raiz. Nas células que não expressam Enod40, MtRBP1 é mantida no núcleo. Este evento de realocização só ocorre em plantas que expressam o Enod40 (Sousa et al. 2001; Quan et al. 2015).

Em *A. thaliana* ocorrem mecanismos que regulam a expressão do inibidor floral FLC (loco de floração C) para ajustar o tempo de floração. A regulação a nível epigenético é mediada pelo lncRNA COLDAIR, ocasionando a modificação de histona. Essa modificação na estrutura da cromatina apresenta papel crucial na regulação da expressão de FLC. Este longo RNA não codificante é classificado como lncRNA intrônico, e é do inglês COLD ASSISTED INTRONIC NONCODING RNA, sendo imprescindível para repressão do locu, possibilitando assim a floração no período de vernalização. A interação de COLDAIR com PRC2, um complexo repressivo, é necessária para o estabelecimento estável da cromatina repressiva no FLC (Heo J. B. & Sung S. 2011).

Esta interação parece ser um mecanismo evolutivamente conservado. COLDAIR é transcrito de um íntron do próprio gene alvo FLC. FLC possui um promotor para o COLDAIR e esse promotor torna-se ativo quando FLC está sendo reprimido. COLDAIR atua principalmente recrutando PRC2 para FLC durante o frio. O papel regulador do lncRNA na repressão do gene sugere que este funciona no direcionamento de maquinarias transcrpcionais para as regiões regulatórias (Heo J. B. & Sung S. 2011).

O FLC é extensivamente regulado através de modificações epigenéticas e de controle transcricional. O frio induz primeiramente COOLAIR um lncRNA antisense, transcrito da fita oposta de FLC. Atua durante a primeira fase da vernalização, acelerando o processo de desligamento de FLC, coordenando alterações de estado da cromatina. O lncRNA COOLAIR de *Arabidopsis*,

reprime a expressão de FLC ao interferir na ligação da Polimerase II (Swiezewski et al. 2009; Chekanova 2015).

Há um potencial interesse de lncRNAs na biologia vegetal, podendo estar envolvidos na regulação de importantes funções da planta, envolvidos em caracteres agronômicos. Outros estudos com plantas analisaram funções regulatórias dos RNAs longos não codificantes como: Franco-Zorrilla et al. (2007), Shin et al. (2016), Swiezewski et al. (2009), Wang et al. (2014), Zhu et al. (2013), entre outros.

2.4 Bioinformática

O objetivo principal da bioinformática é a análise computacional de sequências de moléculas como DNA, RNA e proteínas. Essas moléculas são o principal foco de estudo da Biologia Molecular. Os computadores surgiram como ferramentas importantes para os estudos nessa área no início da década de 1960 (Hagen J.B. 2000). Na década de 1990, surgiram os sequenciadores automáticos, obtendo assim uma grande quantidade de dados a serem armazenados e analisados. Portanto, computadores potentes e ferramentas computacionais rápidas foram exigidas para identificação e caracterização de resultados obtidos. Com isso, o sequenciamento foi o grande propulsor da Bioinformática.

Essa ciência envolve diversas áreas do conhecimento como: matemática, estatística, engenharia de software, ciências da computação e a biologia molecular. Consistindo em todo tipo de estudo ou de ferramenta computacional que se pode organizar ou obter informação biológica. A biologia molecular computacional compreende a área de Bioinformática onde dentro da genômica, mapeamento, sequenciamento e determinação de estrutura, são utilizadas metodologias de computação e ciências da informação (Rizk N. J. 2006).

O desenvolvimento das plataformas de sequenciamento de alto desempenho vem sendo acompanhada pela rápida evolução de novas ferramentas de bioinformática. A bioinformática é estritamente necessária para

as análises do grande volume de dados gerados por essas plataformas. Com a modernização de softwares capazes de analisar extensos volumes de dados, novos avanços têm proporcionado conhecimento elevado em Biologia Molecular. Com a descoberta de novos transcritos, sua estrutura, e predição de funções em várias espécies vegetais e animais.

3 Objetivos

3.1 Objetivo Geral

Este trabalho teve como objetivo geral identificar e caracterizar os RNAs longos não codificantes no genoma de *Arabidopsis thaliana*, analisando a diversidade nucleotídica desses elementos, bem como a sua expressão em diferentes condições de fotoperíodo.

3.2 Objetivos Específicos

- I. Estimar a densidade de RNAs não codificantes no genoma de *Arabidopsis*;
- II. Classificar os RNAs longos não codificantes de acordo com o seu tamanho e localização no genoma;
- III. Avaliar a resposta desses RNAs não codificantes a diferentes condições de fotoperíodo, através da utilização de sequências disponíveis em banco de dados genômico;
- IV. Avaliar a diversidade genética dos RNAs longos não codificantes, em comparação com outros elementos e regiões do genoma, para verificar se esses elementos estão sob pressão de seleção.

4 Material e métodos

4.1 Fonte de dados

As sequências de RNA para realização do presente trabalho foram obtidas do banco Short Read Archive (SRA) do NCBI (National Center for Biotechnology Information). As sequências são de um trabalho que sequenciou RNA total obtidos de plântulas de *Arabidopsis thaliana* (Sullivan et al. 2014) submetidas a cinco condições de luminosidade/fotoperíodo. As plântulas foram cultivadas no escuro durante sete dias, quando se iniciou o estabelecimento dos cinco tratamentos com diferentes exposições à luz: por 0min, 30min, 3h e

24h de dia longo (16h luz e 8h escuro). O RNA foi tratado com Dnase I e preparado com kit Ribo-Zero. O sequenciamento foi realizado na plataforma Illumina HiSeq 2000, usando metodologia de fita única e específica (single-end, strand-specific) com 36 ciclos para obtenção de *reads* de 36 pb. Os códigos de acesso e número das bibliotecas no SRA-NCBI estão na Tabela 2.

Tabela 2: Informações do material estudado. Nome das bibliotecas e o código de acesso no SRA (Short Read Archive) do NCBI. *LD (Long Day).

Bibliotecas	Tratamentos	Códigos de acesso	Autor
SRR1049783	Controle (dias longos: 16h claro e 8h escuro)	SRX391973	Sullivan et al. 2014
SRR1049784	7 dias escuro	SRX391974	Sullivan et al. 2014
SRR1049785	7 dias escuro + 30 minutos luz	SRX391975	Sullivan et al. 2014
SRR1049786	7 dias escuro + 3 horas luz	SRX391976	Sullivan et al. 2014
SRR1049787	7 dias no escuro + 24 horas *LD	SRX391977	Sullivan et al. 2014

4.2 Avaliação de qualidade, alinhamento e montagem dos transcritos

A avaliação da qualidade das sequências foi realizada com a ferramenta FastQC (Andrews S. 2010). A ferramenta permite também obter a porcentagem de conteúdo GC e o número de sequências de cada biblioteca. O alinhamento das sequências no genoma referência de *Arabidopsis thaliana* L. foi realizado com o programa TopHat (Trapnel et al., 2012). O genoma utilizado foi versão TAIR 10 (The *Arabidopsis* Genome Initiative 2000; Berardini et al. 2015) baixado no endereço eletrônico www.arabidopsis.org.

Os resultados do alinhamento foram analisados com o programa Cufflinks (Trapnell et al., 2012), que realizou a montagem dos transcritos alinhados no genoma referência, produzindo um arquivo GFF com localização desses transcritos. A ferramenta Cufflinks produz quatro arquivos output, dentre eles o Transcript.gtf que contém os transcritos montados de cada biblioteca, o isoforms.fpkm_tracking com o valor de número normalizado de sequências alinhadas (FPKM) para cada transcrito, o genes.fpkm_tracking que contém o valor de FPKM para os genes e o shipped.gtf um arquivo de transcritos não utilizados.

4.3 Pipeline de Identificação dos RNAs longos não codificantes.

Após o alinhamento e a montagem dos transcritos, utilizou-se a ferramenta Cuffcompare, um programa complementar do pacote Cufflinks. Esta ferramenta realizou uma comparação dos transcritos montados no programa anterior com a anotação referência (TAIR 10), criando uma união de todos os transcritos comparando suas coordenadas. Os transcritos redundantes em relação a anotação referência (marcados com o símbolo “=”), não foram considerados nas etapas subsequentes.

Os transcritos não redundantes classificados como senses (“e”), antisenses (“x”), intergênicos (“u”) e intrônicos (“i”) foram escolhidos para as análises posteriores. Dentre esses transcritos, foram selecionados aqueles maiores que 200 nucleotídeos utilizando um script em Perl. Esses transcritos foram considerados lncRNAs putativos. Posterior a esta etapa, as sequências foram filtradas, utilizando cinco ferramentas para remoção de transcritos com potencial de codificação.

Para predição de ORF, foi utilizada a ferramenta ORF-Predictor (Min et al., 2005), onde comprimentos de ORFs maiores que 100 aminoácidos, foram retirados da análise. Foi calculado o potencial de codificação dos transcritos, utilizando a ferramenta CPC (Coding Potential Calculator), explorando os parâmetros default (Kong et al. 2007) e foram selecionados os transcritos denominados não codificantes.

Para a identificação do conjunto de transcritos precursores de miRNA, utilizou-se o Blastn e a ferramenta IntersectBed do Bedtools (Quinlan & Hall, 2010). As sequências foram alinhadas com o arquivo de miRNAs baixado no Banco de Dados do miRBase (Kozomara A. & Griffiths-Jones S. 2014) que contém um amplo repertório de miRNAs já descritos. As ferramentas possibilitaram a verificação de sobreposição entre os arquivos, onde precursores de miRNAs foram removidos da análise.

O potencial de codificação dos lncRNAs putativos foi ainda investigado através de busca por homologies contra o banco de peptídeos *nr* de Embryophytas, utilizando a ferramenta Blastx com *E*-value de 10^{-5} . O programa ncRNA-Agent foi também utilizado na filtragem dos lncRNAs putativos, para investigar outras classes de RNAs, como RNA transportador, RNA nucleolar, entre outros. Após as cinco etapas de filtragem (ORF-Predictor, CPC, miRBase, *nr* de Embryophytas e ncRNA-Agent), os transcritos remanescentes foram denominados lncRNAs de alta confiança (HC-lncRNAs).

Com um script em Perl foi criado um arquivo com o nome do lncRNA, cromossomo, sua localização genômica, tamanho (em pb), sua classificação e fita para cada HC-lncRNA. Esse arquivo foi utilizado para a caracterização desses elementos na plataforma R (R Development Core Team 2011). O pipeline computacional utilizado para a identificação dos lncRNAs, segue na figura 4.

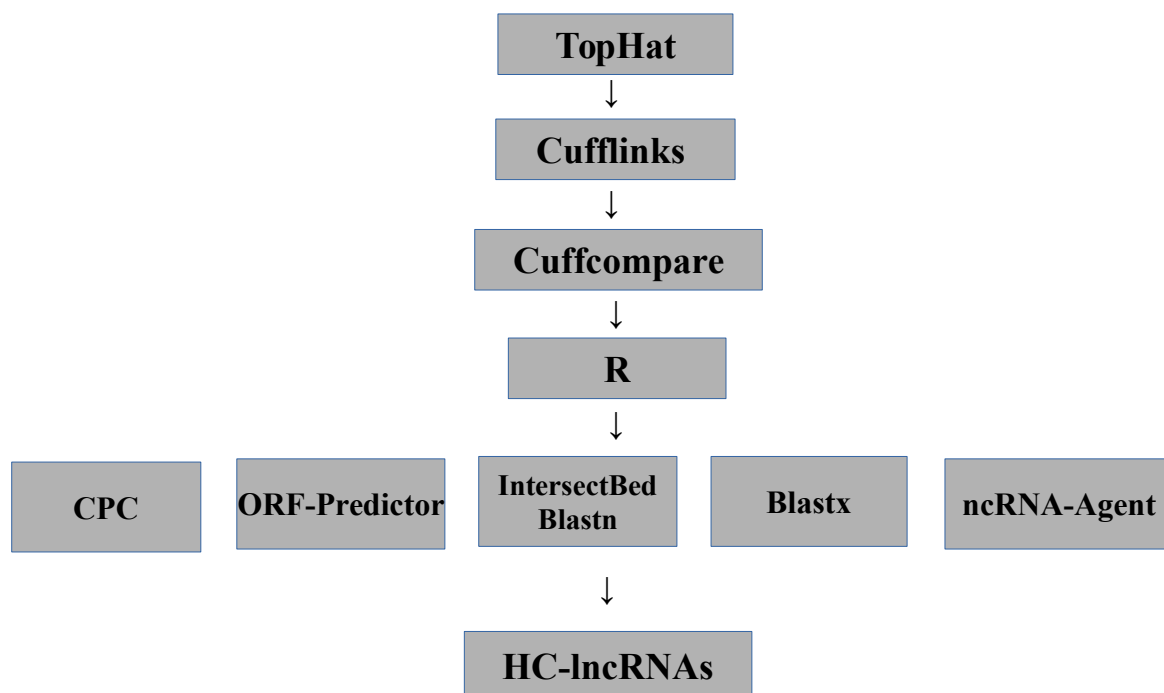


Figura 4. Pipeline computacional utilizado para a identificação dos HC-lncRNAs.

4.4 Identificação de lncRNAs diferencialmente expressos

Para a análise de expressão gênica diferencial, as reads de cada biblioteca alinhadas no genoma com o programa TopHat, foram canalizadas com o programa HTSeq. Este programa realiza a contagem do número de sequências de cada biblioteca alinhadas em cada HC-lncRNA (Anders S. et al. 2014). Os arquivos gerados foram analisados com o programa EdgeR do pacote Bioconductor (Storey et al. 2015) para verificação de lncRNAs diferencialmente expressos.

4.5 Avaliação da diversidade nucleotídica dos lncRNAs contra diferentes regiões do genoma de *Arabidopsis thaliana*

Para comparação da diversidade nucleotídica dos HC-lncRNAs contra diferentes regiões do genoma de *A. thaliana*, utilizou-se os dados do projeto 1001 genomas (Alonso-Blanco et al. 2016). Utilizando-se o pacote VCFTools (Danecek et al. 2011) obteve-se a diversidade nucleotídica π (Nei M. 1987) para cada um dos 10.707.430 SNPs (Polimorfismos de Nucleotídeos Únicos) do projeto 1001. A ferramenta IntersectBED do BEDTools (Quilan & Hall 2010) foi então utilizada para identificar as estimativas de π posicionadas na sequência de cada HC-lncRNA, bem como dos diferentes elementos anotados no genoma (TAIR 10). A tabela BED gerada foi manipulada com um script em Perl e analisada na plataforma R para estimativa da diversidade nucleotídica de cada elemento (somatório dos π / tamanho do elemento). Também no R foram feitas análises e gráficos com o objetivo de comparar a diversidade nucleotídica dos lncRNAs com os outros elementos anotados no genoma, especialmente as regiões codantes (éxons).

Para a predição de estrutura secundária do HC-lncRNA utilizamos a ferramenta RNAfold (Lorenz et al. 2011), que calcula a energia livre mínima da estrutura dos RNAs. No estudo para demonstração de sua estrutura, aplicamos o programa no HC-lncRNA TCONS_00046489, que foi diferencialmente expresso entre as bibliotecas controle e sete dias no escuro.

4.6 Comparação dos HC-lncRNAs com os lncRNAs previamente publicados

Para identificar, dentre os nossos HC-lncRNAs, novos lncRNAs ainda não anotados no genoma de *A. thaliana*, utilizou-se o banco de dados PlncDB: Plant Long noncoding RNA Database (Jin et al. 2013). Esse banco contém um total de 16.227 lncRNAs de *A. thaliana* previamente publicados. As posições dos nossos HC-lncRNAs foram comparados contra as posições dos elementos do PlncDB com o IntersectBed (Quinlan & Hall, 2010) para identificar os lncRNAs em comum entre os dados, distinguindo os transcritos ainda não anotados.

5 Resultados

5.1 Avaliação de qualidade e alinhamento

A análise da qualidade das sequências, com o programa fastQC, verificou altos valores de qualidade, com valor Phred médio sempre acima de 30 para todas as 36 pb (Figura 5). Os resultados do controle de qualidade também indicam que não houve uma queda na qualidade comumente observada no final das sequências Illumina. Com isso e com o fato de que as sequências seriam posteriormente alinhadas no genoma de referência, optou-se pela não retirada das bases com baixa qualidade das cinco bibliotecas analisadas figura 5.

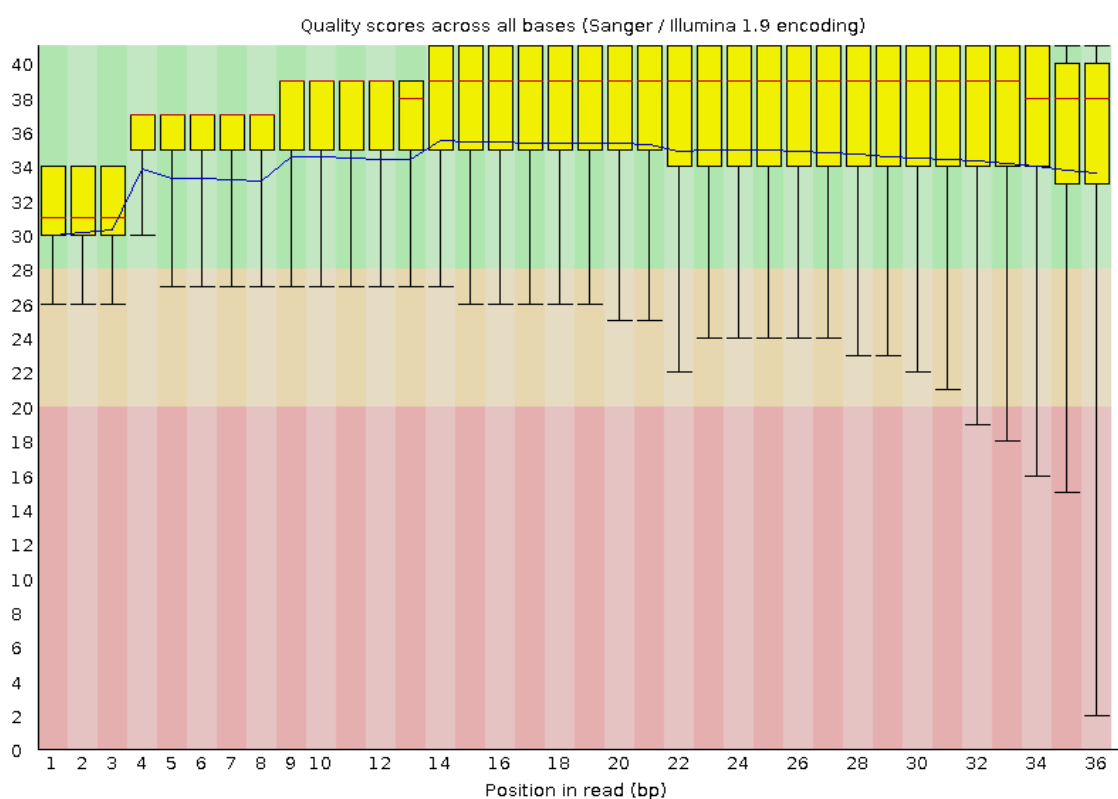


Figura 5: Exemplo de qualidade das sequências, obtidas na plataforma Illumina. Verde: Boa qualidade (phred > 28); laranja: qualidade razoável (20 < phred < 28); vermelho: qualidade ruim (phred < 20). No eixo “x” estão as posições das bases da read e no eixo “y” as notas de qualidade phred.

Em seguida, as sequências de cada biblioteca foram alinhadas no genoma referência de *A. thaliana*, versão TAIR10. A Tabela 3 demonstra os resultados do alinhamento. Com este resultado podemos observar a quantidade de sequências por biblioteca sequenciada, e porcentagem de alinhamento de cada biblioteca ao genoma referência. Como esperado, a taxa de alinhamento no genoma de referência foi alta, com média de 90,32%, variando de 88,2 %, para a biblioteca de 3 horas de luz, a 93,7 % para a biblioteca controle. O conteúdo GC foi de 44 % para todas as bibliotecas analisadas.

Tabela 3: Resultados dos softwares FastQC e TopHat. Com total de sequências por biblioteca amostrada, comprimento das sequências, conteúdo GC, plataforma utilizada para o sequenciamento e a porcentagem de sequências que foram alinhadas no genoma referência.

Biblioteca	Controle	Escuro	Escuro + 30 min luz	Escuro + 3 horas luz	Escuro + 24 horas LD
Total de sequências	47628234	62154258	84239683	61981925	75467965
Comprimento de sequências	36	36	36	36	36
% GC	44	44	44	44	44
Sequenciamento	Illumina	Illumina	Illumina	Illumina	Illumina
Alinhamento	93.7%	91.1%	88.7%	88.2%	89.9%
Alinhamentos múltiplos	10.9%	6.7%	6.6%	6.5%	6.4%

5.2 Estatísticas de montagem

O processo de montagem utilizando as sequências gerou em média 53.456 transcritos por biblioteca, com mais de 230 mil éxons no total. O comprimento médio dos transcritos foi de 1.882, variando de microRNAs (21 pb) a um transcrito de 58.520 pb. As estatísticas da montagem seguem na Tabela 4.

Tabela 4: Estatísticas de montagem dos transcritos a partir das sequências de RNA total.

Biblioteca	Controle	Escuro	Escuro + 30 min de luz	Escuro + 3 horas de luz	Escuro + 24 horas LD
Número de transcritos	48610	53460	54949	55369	54892
Média dos transcritos	2053	1884	1834	1817	1822
Comprimento mínimo	21	21	21	21	21
Comprimento máximo	58520	58520	58520	58520	58520
Éxons	232655	237478	238620	239658	238230

5.3 Identificação dos RNAs longos não codificantes

Os transcritos montados foram comparados com a anotação do genoma utilizando a ferramenta Cuffcompare, afim de identificar os transcritos não anotados no genoma de *Arabidopsis*. O Cuffcompare classifica com um código (class_code) os transcritos de acordo com sua posição genômica em relação aos transcritos previamente anotados. Como demonstrado na Tabela 5, para as análises subsequentes, foram selecionadas quatro classes como potenciais

ncRNAs. Essas classes foram “i”, “u”, “e” e “x”, representando transcritos intrônicos, intergênicos, senses e antisenses, respectivamente, em relação aos transcritos anotados na versão TAIR10 do genoma. No total foram identificados 31.900 transcritos nessas quatro classes (Tabela 5).

As sequências menores que 200 nucleotídeos foram removidas através de um script Perl, ficando assim com 4.305 lncRNAs putativos. Desses, 314 (7%) foram senso em relação a transcritos codantes (mRNAs), 392 (9%) intergênicos, 2.216 intrônicos (52%) e 1.383 (32%) antisenso de mRNAs anotados no TAIR 10, como mostra a Tabela 6.

Tabela 5: Resultado do programa Cuffcompare, com a comparação dos transcritos com um arquivo de anotação e os respectivos class_codes identificados.

Classes	Transcritos	Porcentagem	Descrição
.	2348	2.71%	Indica múltiplas classificações
=	42056	48.58%	Match completo com transcrito previamente anotado
c	22	0.02%	Contido
e	374	0.43%	Transfrag de único éxon sobreposto a um exon referência
i	17908	20.68%	Transfrag dentro de um íntron
j	3880	4.48%	Potencialmente uma nova isoforma (fragmento)
o	119	0.13%	Éxon genérico sobreposto a um transcrito referência
u	9819	11.34%	Transcrito intergênico

Classes	Transcritos	Porcentagem	Descrição
x	3799	4.38%	Éxon sobreposto com referência na fita oposta
s	1	0.00%	Íntron do transfrag sobreposto a um íntron referência na fita oposta
Total	86559	100.00%	-

Tabela 6: Resultado dos lncRNAs identificados para as quatro classes selecionadas, após a remoção dos transcritos menores que 200 nucleotídeos.

Classes	Senses (e)	Intrônicos (i)	Intergênicos (u)	Antisenses (x)
LncRNAs	314	2216	392	1383

A Tabela 7 demonstra a quantidade de RNAs longos não codificantes, pelos cromossomos de *A. thaliana*, dentre eles o cromossomo mitocondrial, que contém o menor número de lncRNAs. Os resultados demonstram uma boa distribuição dos lncRNAs em todos os cromossomos. Curiosamente, em mitocôndria, houve um enriquecimento significativo de lncRNAs intergênicos quando comparado ao esperado.

Tabela 7: lncRNAs das quatro classes identificados em cada um dos cromossomos de *A. thaliana*.

	Senses	Intrônicos	Intergênicos	Antisenses
Cromossomo 1	93	590	86	339
Cromossomo 2	47	334	52	208
Cromossomo 3	63	428	58	282
Cromossomo 4	43	342	66	209

	Senses	Intrônicos	Intergênicos	Antisenses
Cromossomo 5	68	521	96	333
Cromossomo M	0	1	34	12
			Total	4.305

Os lncRNAs possuem em média 463 bp de comprimento e contêm poucos éxons, com 93% contendo apenas um éxon. Esse tamanho e número de éxons é bem menor do que os valores encontrados para os genes codantes de *Arabidopsis*. O tamanho dos lncRNAs variou entre as classes, contendo um comprimento maior nos transcritos sense (média de 925bp) e menor nos transcritos de região intrônica (299 bp), como podemos visualizar na Tabela 8.

Tabela 8: Distribuição de comprimentos dos lncRNAs de *A. thaliana*.

	Senses	Intrônicos	Intergênicos	Antisenses
Comprimento máximo	7003	2642	2633	6485
Comprimento mínimo	200	200	200	200
Comprimento médio	925	299	318	662

5.4 Filtragem de lncRNAs putativos

Os lncRNAs putativos foram filtrados para eliminar aqueles com potencial de codificação protéica, bem como aqueles relacionados com a síntese de rRNA, tRNA e miRNA. Após essa filtragem, foram obtidos 3.710 lncRNAs de alta confiança, denominados HC-lncRNAs (figura 6).

A ferramenta CPC (Calculator Potential Coding) filtrou 295 (6.8%) lncRNAs. Com o software ORF Predictor, usado para predição de ORFs

maiores que 100 aminoácidos, foram filtrados um total de 329 (7,6%) lncRNAs.

O Blastx foi utilizado para eliminar lncRNAs com homologia contra as sequências de Embryophyta do banco de dados nr (NCBI). Utilizando um *E*-value de 10^{-5} , foram filtrados um total de 375 (8,7%) lncRNAs com potencial de codificação. Para a filtragem de RNAs transportadores, RNAs ribossomais, RNAs nucleolares, RNAs nucleares, RNAs interferências, entre outros utilizou-se a ferramenta ncRNA-Agents. Nesta etapa foram eliminados 88 (2%) lncRNAs putativos.

Com o algoritmo Blastn, foram identificadas homologias entre os lncRNAs e os miRNAs previamente anotados em *A. thaliana*. O IntersectBed também foi utilizado para identificação de possíveis miRNAs já anotados no TAIR10. Após esta verificação obtivemos 21 lncRNAs precursores de miRNAs em comum entre os dois softwares. Com a remoção destes, totalizou 4.284 lncRNAs putativos, a partir deste método de análise.

Após a filtragem de todos os transcritos com potencial de codificação de proteínas e outros ncRNAs (tRNA, rRNA, microRNA, etc.), foram obtidos um total de 3.710 lncRNAs de alta confiança, denominado HC-lncRNA.

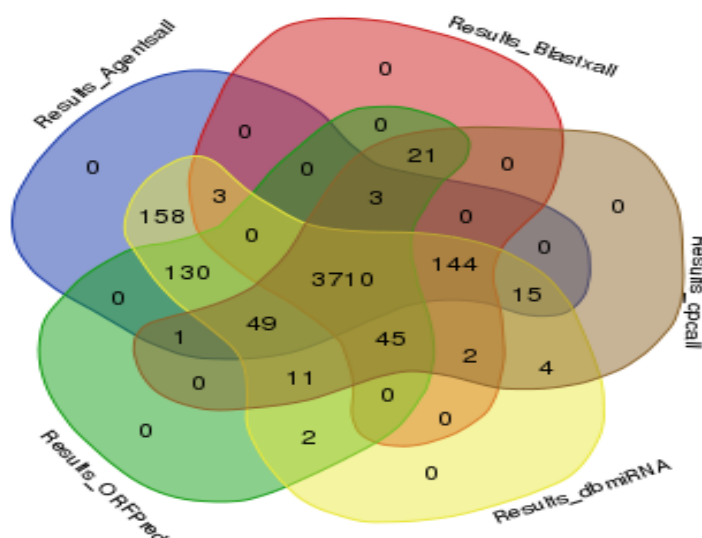


Figura 6: Diagrama de Venn, demonstrando a quantidade de lncRNAs de alta confiança, que passaram pelo processo de filtragem dos cinco programas utilizados, sendo denominados HC-lncRNAs.

A Tabela 9 mostra que os HC-lncRNAs foram distribuídos entre os cinco cromossomos de *A. thaliana*, sendo mais frequentes no cromossomo 1. Após a filtragem, o número de transcritos senses, teve uma diminuição significativa, com 91% desses sendo eliminados por conter potencial de codificação. O cromossomo 1, foi o que mais obteve RNAs longos, sendo o cromossomo com comprimento máximo, entre cromossomos nucleares e organelares. A densidade entre os cromossomos nucleares não diferiu entre os cromossomos nucleares, no entanto houve um aumento considerável de RNAs longos no cromossomo mitocondrial (tabela 9).

A média de comprimento dos transcritos das quatro classes analisadas reduziu, atingindo uma média de 374 nucleotídeos, entre as classes apresentadas (Tabela 10).

Tabela 9: Quantidade de lncRNAs por cromossomo em *A. thaliana*, após a filtragem.

Cromossomos	Classes				Densidade
	e	i	u	x	
Chr 1	5	586	83	284	3,0/100kb
Chr 2	4	333	51	162	3,0/100kb
Chr 3	7	419	53	231	3,0/100kb
Chr 4	4	340	59	174	3,0/100kb
Chr 5	7	514	86	276	3,0/100kb
Chr Mitocondrial	0	0	29	3	9,0/100kb
Total	27	2192	361	1130	-

Tabela 10: Comprimentos entre as classes de HC-lncRNAs em *A. thaliana*, após a filtragem.

	Senses	Intrônicos	Intergênicos	Antisenses
Comprimento máximo	1134	2642	2230	6896
Comprimento mínimo	206	200	200	200
Comprimento médio	397	298	305	545

Neste trabalho 3.710 lncRNAs de alta confiança foram encontrados. Desses, 58,6% são aparentemente inéditos, na medida em que não estão contidos no PlncDB. lncRNAs tem importantes papéis em vários processos biológicos de plantas e animais (Bai et al. 2014). Estes RNAs estão sendo alvo de muitas pesquisas recentes em biologia molecular (Derrien et al. 2012). A classe de ncRNAs mais identificada em vários trabalhos, são os lncRNAs intergênicos, como por exemplo no trabalho Liu et al. (2012). Em nosso trabalho, por outro lado, a classe mais representada foi a de lncRNA intrônicos, ou seja, aqueles contidos em posições de íntrons de regiões codificantes de proteínas. Os íntrons em humanos, são ricos em RNAs não codificantes, podendo conter diferentes ncRNAs, como por exemplo, os lncRNAs foco de nosso trabalho (Rearick et al. 2010). Estes ncRNAs podem regular a expressão gênica de diferentes maneiras, entre elas o silenciamento gênico transcricional, tipo bastante aplicado em plantas (Rearick et al. 2010).

Os lncRNAs apresentam poucos éxons em sua sequência, com a maioria (93%) dos HC-lncRNAs identificados neste trabalho contendo apenas um éxon. Esse resultado é corroborado por outros trabalhos (Li et al., 2014; Chen et al., 2016; Shin et al., 2016).

5.5 Análise de enriquecimento funcional

Para uma análise de enriquecimento funcional de categorias do Gene Ontology (GO), de genes com co-localização aos HC-lncRNAs utilizamos a ferramenta Blast2go. A co-localização com genes codantes anotados no TAIR10 ocorreu para os lncRNAs de classes “e”, “i” e “x”. O resultado como indicado na Tabela 11, demonstrou que dentre os genes que contém lncRNAs, existe enriquecimento para genes ligados a processos biológicos de localização e transporte de proteínas dentro da célula. Na classificação quanto a componente celular temos enriquecimento significativo para localização no núcleo e membrana (figura 7). Já em relação às possíveis funções moleculares temos enriquecimento para ligação a ácido nucléico (figura 7). Essa ligação em ácidos nucléicos e a localização de muitos genes com lncRNA no núcleo, indicam que lncRNAs podem estar atuando em fatores de transcrição em cis, corroborando estudos prévios que indicam o papel dos lncRNAs na regulação gênica (Bai et al. 2014). De maneira geral, esses enriquecimentos funcionais indicam que a presença de lncRNAs em genes codantes não é aleatória, mas parece ocorrer mais frequentemente em genes relacionados ao endereçamento de outras proteínas e em fatores de transcrição.

Tabela 11: Resultado da análise de enriquecimento funcional utilizando o programa Blast2GO. P= Processo biológico, F= Função molecular, C= Componente celular.

GO-ID	Term	Categoria	FDR	P-Value
GO:0033036	Macromolecule localization	P	7.021825E-4	3.001089E-7
GO:0051641	Cellular localization	P	7.021825E-4	4.698868E-7
GO:0051649	Establishment of localization in cell	P	7.021825E-4	8.771799E-7
GO:0046907	Intracellular transport	P	7.021825E-4	8.771799E-7

GO-ID	Term	Categoria	FDR	P-Value
GO:0045184	Establishment of protein localization	P	3.541634E-3	8.40368E-6
GO:0005634	Nucleos	C	3.541634E-3	8.514655E-6
GO:0008104	Protein localization	P	3.541634E-3	8.848554E-6
GO:0051179	Localization	P	5.016212E-3	1.508249E-5
GO:0051234	Establishment of localization	P	5.016212E-3	1.614647E-5
GO:00115031	Protein localization	P	5.016212E-3	1.723246E-5
GO:0003676	Nucleic acid binding	F	5.820973E-3	2.181502E-5
GO:0006810	Transport	P	6.071182E-3	2.464877E-5
GO:0043231	Intracellular membrane-bounded organelle	C	6.317825E-3	3.096465E-5
GO:0043227	Membrane-bounded organelle	C	6.317825E-3	3.096465E-5
GO:0070727	Cellular macromolecule localization	P	6.317825E-3	3.354248E-5
GO:0064613	Cellular protein localization	P	6.317825E-3	3.354248E-5
GO:0006886	Intracellular protein transport	P	1.160817E-2	6.525515E-5

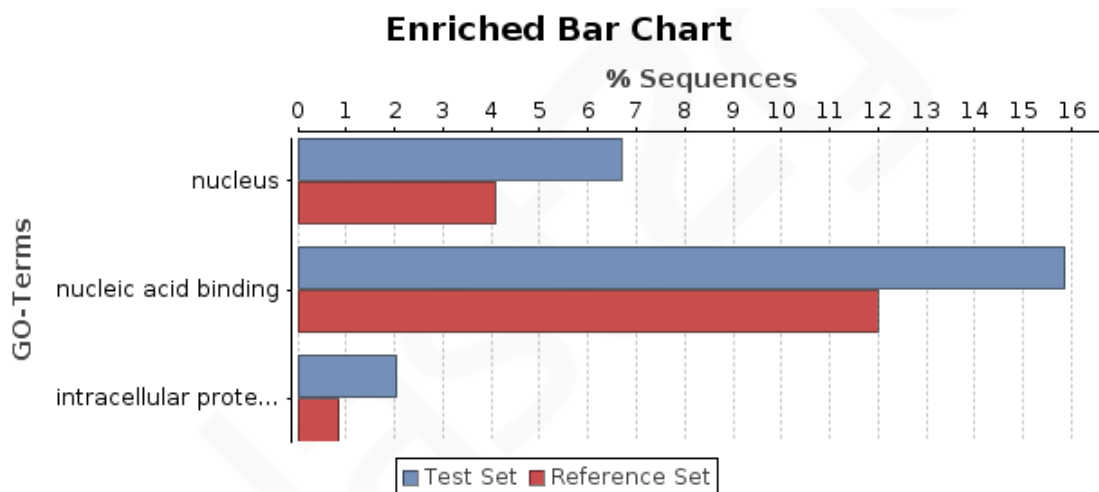


Figura 7: Principais termos Gene Ontology “GO” em que os genes que contém HC-lncRNAs estão envolvidos.

Uma das principais funções dos lncRNAs parece ser agir em processo epigenético, podendo atuar em cis ou trans, em genes próximos que tem potencial codificante (Qureshi et al. 2010). A análise de enriquecimento funcional mostrou que os genes relacionados com os lncRNAs identificados podem estar envolvidos em processos de regulação gênica, como por exemplo através da ligação a ácidos nucleicos. A partir desta evidência, é possível que estes lncRNAs possam estar regulando estes genes adjacentes.

5.6 Identificação de lncRNAs diferencialmente expressos em resposta a luz

As sequências deste trabalho foram obtidas de trabalho prévio que estudou o efeito da luz na regulação do transcrito, via remodelação da cromatina (Sullivan et al. 2014). Sendo a luz um importante fator ambiental para a regulação do crescimento e desenvolvimento das plantas, buscou-se no presente trabalho avaliar se os HC-lncRNAs identificados também podem atuar na fotomorfogênese vegetal.

Um dendrograma construído com base nas correlações entre os níveis de expressão dos lncRNAs nas diferentes amostras, indica consistência entre as distâncias e a intensidade de luz recebida pelas plantas. Assim, existe indicativo de que a luz, fator essencial para o desenvolvimento vegetal, afeta a expressão dos lncRNAs.

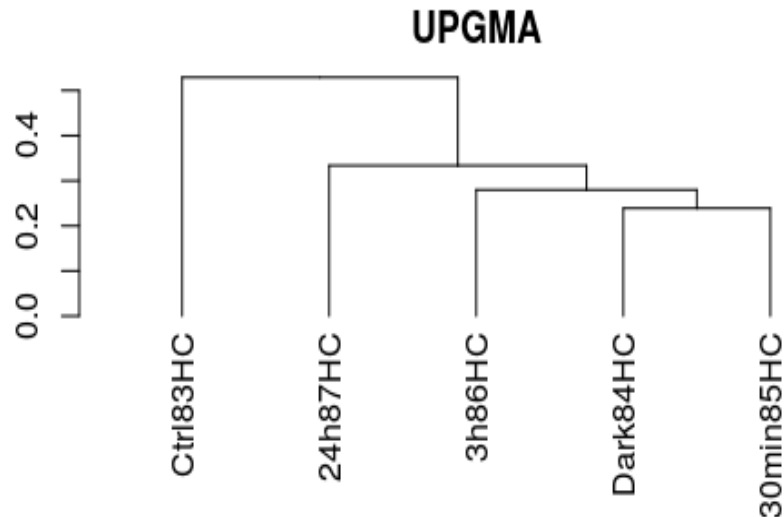


Figura 8: Distância entre as bibliotecas de acordo com o nível de expressão dos HC-lncRNAs identificados. O dendrograma foi construído com base no agrupamento, via UPGMA, das distâncias obtidas a partir do complemento da correlação de Spearman entre o nível de expressão (contagem de reads mapeados) de cada lncRNA.

Para testar a hipótese de que os lncRNAs são afetados pela luz, realizou-se uma análise de expressão diferencial via RNA-Seq, isto é através da análise da contagem das sequências de cada biblioteca (amostra) mapeadas em cada lncRNA. Um total de 14 lncRNAs diferencialmente expressos foram identificados entre a biblioteca controle (16h claro e 8h escuro) e a biblioteca exposta somente ao escuro (Tabela 12). Um total de 13 e 16 lncRNAs diferencialmente expressos foram identificados entre as amostras controle e 30 minutos e controle e 3 horas expostas a luz respectivamente (tabelas 13 e 14). Em adição, a biblioteca controle comparada à biblioteca de 24 horas LD gerou um total de sete lncRNAs diferencialmente expressos (Tabela 15). Um total de 22 lncRNAs diferencialmente expressos foram identificados significativamente, utilizando FDR < 0.05 e fold change > 2 ou < -2.

Os locus XLOC_008359, XLOC_010820, XLOC_060027, identificados como diferencialmente expressos, estão envolvidos em núcleo e proteína de ligação a membrana como componente celular e localização como processo biológico. A análise de expressão demonstrou uma diferença significativa entre

a expressão de lncRNAs entre as bibliotecas estudadas, como demonstrado na figura 8.

Tabela 12: lncRNAs diferencialmente expressos, entre bibliotecas controle e biblioteca exposta somente ao escuro.

LncRNA	Classe	Cromossomo	FDR	LogFC
TCONS_00046489	antisense	Chr3	0.000826	-7.58
TCONS_00044801	intrônico	Chr3	0.004016	10.77
TCONS_00019188	intergênico	Chr1	0.004016	10.00
TCONS_00080954	antisense	Chr5	0.020024	9.12
TCONS_00002035	antisense	Chr1	0.020440	-5.28
TCONS_00018015	antisense	Chr1	0.020440	5.93
TCONS_00021340	antisense	Chr1	0.020440	5.93
TCONS_00084987	intrônico	Chr5	0.020440	8.80
TCONS_00049097	sense	Chr3	0.020440	8.76
TCONS_00064876	antisense	Chr4	0.032055	-4.58
TCONS_00080348	sense	Chr5	0.032252	8.44
TCONS_00084016	antisense	Chr5	0.037017	-4.60
TCONS_00083294	antisense	Chr5	0.039113	4.37
TCONS_00033946	antisense	Chr2	0.043798	-4.20

Tabela 13: LncRNAs diferencialmente expressos, entre as bibliotecas controle e 30 minutos de luz.

LncRNA	Classe	Cromossomo	FDR	LogFC
TCONS_00064876	antisense	Chr4	0.000721	-8.46
TCONS_00046489	antisense	Chr3	0.000783	-7.20
TCONS_00002035	antisense	Chr1	0.000877	-10.77
TCONS_00033946	antisense	Chr2	0.007431	-5.72
TCONS_00044801	intrônico	Chr3	0.010727	10.02
TCONS_00048963	antisense	Chr3	0.022365	-4.87
TCONS_00083294	antisense	Chr5	0.022365	4.82
TCONS_00019188	intergênico	Chr1	0.022365	9.45
TCONS_00080954	antisense	Chr5	0.027950	9.29
TCONS_00009028	antisense	Chr1	0.028761	5.14
TCONS_00080231	antisense	Chr2	0.033448	9.09
TCONS_00032291	intergênico	Chr2	0.037752	9.00
TCONS_00049097	sense	Chr3	0.041108	8.91

Tabela 14: LncRNAs diferencialmente expressos, entre as bibliotecas controle e três horas de luz.

LncRNA	Classe	Cromossomo	FDR	LogFC
TCONS_00064876	antisense	Chr4	0.000194	-9.23
TCONS_00046489	antisense	Chr3	0.009871	-5.80
TCONS_00019188	intergênico	Chr1	0.023962	9.79
TCONS_00080954	antisense	Chr5	0.023962	9.72
TCONS_00049097	sense	Chr3	0.025646	8.86
TCONS_00082591	antisense	Chr5	0.025646	8.85
TCONS_00080231	antisense	Chr5	0.025646	8.85
TCONS_00044801	intrônico	Chr3	0.031229	8.61
TCONS_00024804	antisense	Chr2	0.031229	-4.55
TCONS_00019369	antisense	Chr1	0.031229	5.21
TCONS_00009028	antisense	Chr1	0.043711	4.99
TCONS_00080348	sense	Chr5	0.043783	8.25
TCONS_00037365	antisense	Chr3	0.043972	5.19
TCONS_00079383	antisense	Chr5	0.043972	5.19
TCONS_00083294	antisense	Chr5	0.043972	4.22
TCONS_00065464	antisense	Chr4	0.043972	-4.46

Tabela 15: LncRNAs diferencialmente expressos, entre as bibliotecas controle e 24 horas dia longo.

LncRNA	Classe	Cromossomo	FDR	LogFC
TCONS_00064876	antisense	Chr4	0.000244	-11.75
TCONS_00044801	intrônico	Chr3	0.014267	9.52
TCONS_00019188	intergênico	Chr1	0.014267	9.39
TCONS_00049097	sense	Chr3	0.016532	9.20
TCONS_00024804	antisense	Chr2	0.032035	-4.80
TCONS_00009028	antisense	Chr1	0.042789	5.35
TCONS_00014870	antisense	Chr1	0.049730	-5.72

Com a utilização da ferramenta RNAfold obtivemos a estrutura secundária do lncRNA diferencialmente expresso com o menor FDR entre as bibliotecas controle e 7 dias no escuro, como demonstrado na figura 9.

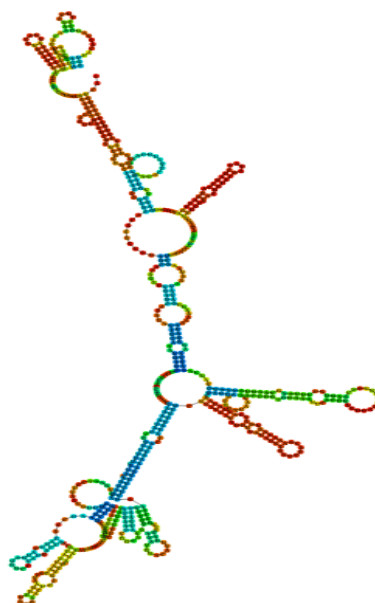


Figura 9: Estrutura secundária do lncRNA TCONS_00046489.

Mesmo com alguns estudos demonstrando os importantes papéis dos lncRNAs em plantas, em diversas fases de desenvolvimento ou processos biológicos, o estudo de Lv et al. 2016 mostrou que poucos lncRNAs parecem estar envolvidos na resposta de *Zea mays* a estresses abióticos. Naquele trabalho foram identificados poucos lncRNAs intergênicos e intrônicos diferencialmente expressos em resposta ao déficit e suficiência de nitrogênio (Lv et al. 2016). Nossos resultados corroboram isso na medida em que somente 22 HC-lncRNAs foram diferencialmente expressos em resposta às diferentes condições de luz. Apesar desses estudos iniciais, pode-se dizer que as funções dos lncRNAs permanecem ainda imensamente desconhecidas.

5.7 Diversidade nucleotídica em lncRNAs e outras regiões do genoma

Uma pequena quantidade de lncRNAs em animais e plantas tem sua sequência primária conservada evolutivamente. Essa baixa conservação de sequências ocorre mesmo em lncRNAs com funções reguladoras, sugerindo uma rápida evolução (Liu et al., 2012). Para avaliar esta hipótese em *Arabidopsis* verificamos a diversidade nucleotídica presente nas regiões do genoma onde foram anotados os HC-lncRNAs. Com os SNPs já identificados no trabalho de 1001 genomes (The 1001 Genomes Consortium), verificamos se em regiões de lncRNAs a densidade de SNPs apresentam-se maior do que em outras regiões do genoma (Tabela 16). Nossa hipótese é de que, conforme observado em animais, os lncRNAs de *Arabidopsis* também não apresentarão conservação em sua estrutura primária, ocorrendo uma maior diversidade nucleotídica quando comparado a sequências codantes.

Os HC-lncRNAs identificados nesse trabalho possuem diversidade genética média variando de 0,0056 para lncRNA em regiões intergênicas (u) a 0,0045 para lncRNAs que englobam pelo menos parte da região codante (e), como demonstrado na Tabela 17. Esses valores são similares a média de diversidade genética observada em regiões UTR (0,0055 para 5'UTR e 0,0053 para 3' UTR), superiores a diversidade obtida em regiões codantes (0,0036 para éxons) e inferiores aos valores observados em pseudogenes (0,0077). A partir desta interpretação podemos inferir que estas regiões de lncRNAs

diversidade nucleotídica intermediária. Apesar da baixa conservação destes transcritos em estrutura primária, estes apresentam-se menor diversidade quando comparado a regiões neutras como os pseudo-genes.

Tabela 16: Quantidade de SNPs e médias por região do genoma de *Arabidopsis thaliana*.

feature	n	pi médio	pi mediano	Desvio padrão de pi
exon	201009	0.00364507	0.00176883	0.0051006
five_prime_UTR	32134	0.00548838	0.00270385	0.00946189
miRNA	179	0.00481873	0.00303576	0.00520925
mRNA	34966	0.00482885	0.00321353	0.00438207
ncRNA	475	0.00551214	0.00396486	0.00485126
pseudogenic_exon	1223	0.00725795	0.0059727	0.00616658
pseudogenic_transcript	895	0.0077263	0.00660739	0.00581028
snoRNA	71	0.00344629	0.00197112	0.00456873
snRNA	13	0.004982	0.00363355	0.00385484
three_prime_UTR	29917	0.00528161	0.00319923	0.00688072
tRNA	522	0.00229686	0.00077484	0.00438588

Tabela 17: Quantidade de SNPs e médias por classe de lncRNA.

feature	n	pi médio	pi mediano	Desvio padrão de pi
e	311	0.00456967	0.00329556	0.00413202
i	2215	0.00487142	0.00296517	0.00547828
u	339	0.00568767	0.00389467	0.00535648
x	1361	0.00518326	0.00344377	0.00520055

Os dados mostram uma diversidade maior em regiões de RNAs não codificantes e transcritos pseudogênicos, em relação a regiões de RNAs codificantes de proteínas, como demonstrado na figura 10. Já os HC-lncRNAs deste trabalho apresentam diversidade nucleotídica intermediária em relação aos exons e pseudogenes (figura 11). O perfil de diversidade observado nos HC-lncRNAs é similar a distribuição obtida para os ncRNAs previamente anotados no genoma de *Arabidopsis* (Tabela 10).

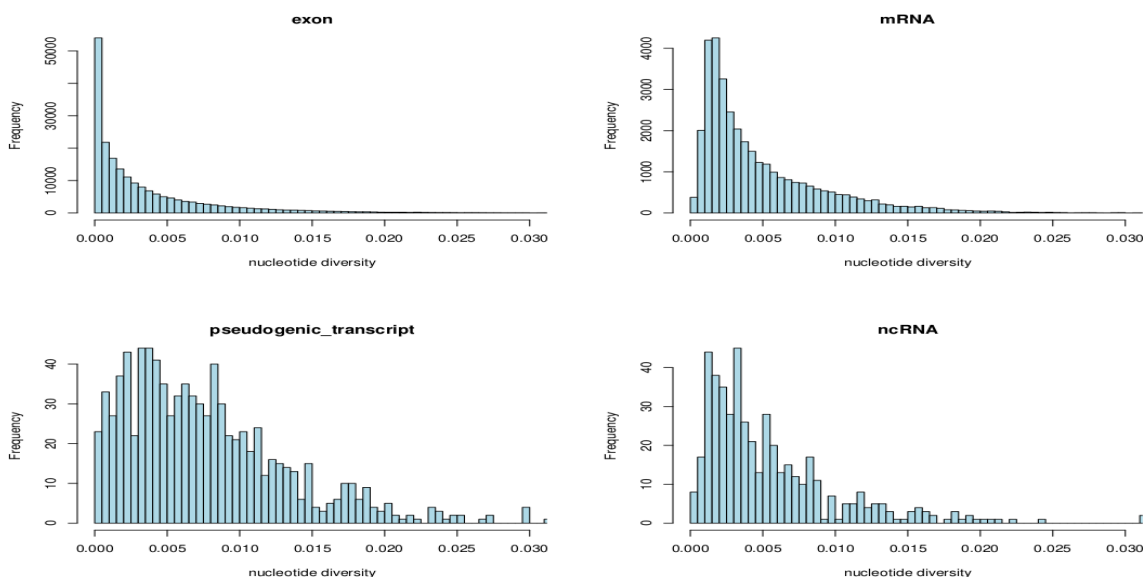


Figura 10: Diversidade nucleotídica entre diferentes posições do genoma: éxon, RNA mensageiro (mRNA), transcritos pseudogênicos e RNAs não codificantes.

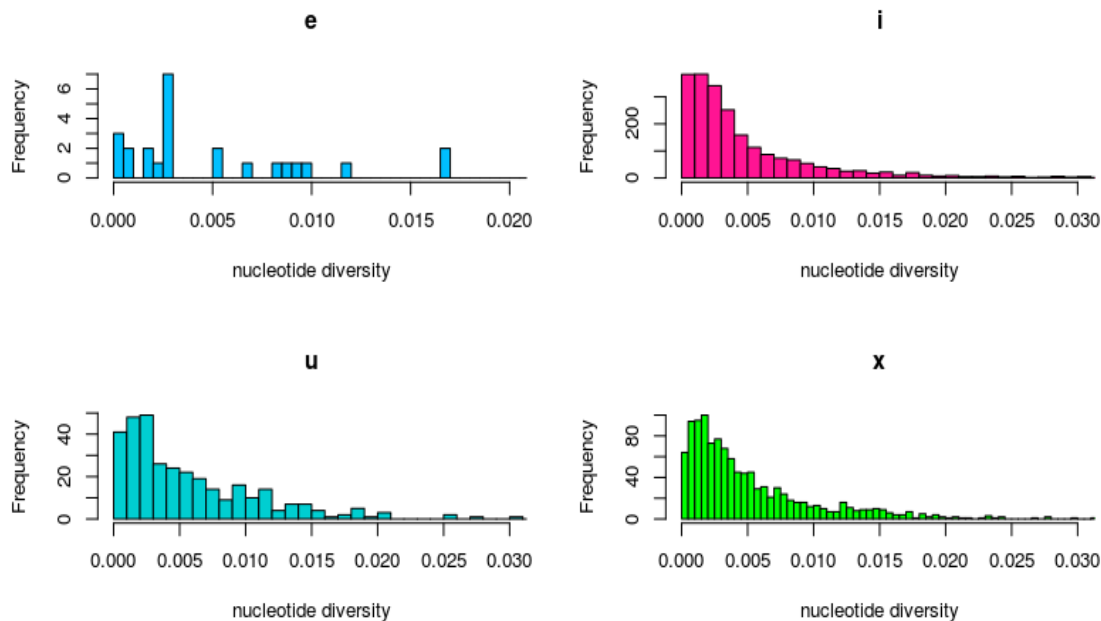


Figura 11: Diversidade nucleotídica das quatro classes de lncRNAs: senses (e), intrônicos (i), intergênicos (u) e antisenses (x).

Este trabalho confirma a observação de que a maior parte dos lncRNAs são pouco conservados em estrutura primária. Apesar disso, existe também lncRNAs identificados com uma baixa diversidade nucleotídica. Em mamíferos também existem alguns lncRNAs mais bem conservados dependendo de sua posição genômica (Guttman et al. 2009).

5.8 Comparação dos HC-lncRNAs com lncRNAs já anotados.

Utilizando a ferramenta IntersectBed comparou-se as posições genômicas dos lncRNAs identificados neste trabalho com aqueles previamente anotados, por outros trabalhos, no genoma de *Arabidopsis*. Para isso utilizou-se o banco de dados PLncDB de lncRNAs de *Arabidopsis* (Jin et al. 2013). A análise demonstra que, dentre os HC-lncRNAs deste trabalho, 2.173 (58,6%) podem ser considerados novos, sendo 15 senses, 1.371 intrônicos, 179 intergênicos e 608 anisenses.

As anotações de lncRNAs em muitas espécies de plantas, até o presente momento, foram realizadas por metodologias como sequenciamento de EST, microarranjo e tiling array. Atualmente, o baixo custo do sequenciamento e os novos avanços na área de bionformática, com criação de novos softwares,

possibilitaram amplos estudos com RNA-seq. Até o momento, esses trabalhos têm focado principalmente em animais, para descoberta de novos transcritos ainda não anotados.

Este estudo foi proposto para ampliar o conhecimento básico sobre os lncRNAs em plantas, gerando a identificação de novos transcritos e a avaliação de quais destes elementos respondem às condições de déficit de luz. Estudos têm demonstrado que os lncRNAs são expressos em tecidos específicos e espécies específicas (Wang et al. 2016), relacionado também com condições variadas em que as plantas são expostas. Uma análise de transcrito de células de tecidos específicos provavelmente geraria a descoberta de ainda mais lncRNAs em plantas (Yamada M. 2016). Esses estudos são importantes, uma vez que as funções destes RNAs ainda não são bem caracterizadas, principalmente em plantas.

A identificação de lncRNAs em diferentes conjuntos de dados em *Arabidopsis* (Liu et al. 2012), demonstrou que: a) estes RNAs são expressos em baixa quantidade quando comparado a expressão de RNAs mensageiros; b) são tecido específico, expressando sob alguma condição em particular. O fato dos lncRNAs serem tecido-específico, indica que a presença destes transcritos é diluído quando se sequencia RNA extraído de toda a planta (Li et al. 2016), como no caso do presente estudo. Este fato talvez explique a descoberta de poucos lncRNAs neste trabalho.

Neste trabalho foram apresentados resultados sobre a identificação e caracterização inicial de novos lncRNAs no genoma de *Arabidopsis*. A análise da expressão desses lncRNAs em resposta a diferentes condições de luminosidade, indicam que poucos desses elementos participam do controle da fotomorfogênese em *A. thaliana*.

6 Considerações finais e perspectivas

O entendimento acerca de funções biológicas dos lncRNAs ainda é bem limitado. O nosso estudo representa um passo inicial em busca da

compreensão das reais funções moleculares destes transcritos. Os HC-lncRNAs aqui identificados podem ser alvo de estudos futuros a fim de melhor compreender esta classe de RNAs. Estas descobertas enriquecem nosso conhecimento sobre os lncRNAs em *A. thaliana*, podendo assim expandir os métodos de identificação destes transcritos.

Como trabalho futuro é possível submeter *A. thaliana* ou outra espécie de planta a novos experimentos, utilizando diferentes condições ambientais que alterem o desenvolvimento vegetal, afim de descobrir novos RNAs longos não codificantes e os seus possíveis papéis no desenvolvimento vegetal. Para esses novos experimentos podem ser feitas análises em tecidos específicos. Esses novos testes avaliariam o papel central destes transcritos na planta, validando a eficácia desta abordagem proposta em diferentes condições ambientais e fase de desenvolvimento. Vale ressaltar que as análises realizadas nesse estudo, podem servir de base para novos trabalhos na mesma ou em outras espécies vegetais.

7 Conclusões

Com os resultados obtidos neste trabalho podemos concluir que:

- Devido aos seus pequenos tamanhos, somente ~1% do genoma de *Arabidopsis thaliana* é expresso em lncRNAs.
- Nas condições do presente estudo, utilizando todos os tecidos de plântulas, 22 lncRNAs tiveram expressão gênica diferencial em resposta a diferentes regimes de luz.
- A diversidade nucleotídica em regiões de RNAs longos não codificantes é maior do que em regiões codantes, mas menor do que a diversidade observada em regiões neutras como os pseudogenes.

8 Referências bibliográficas

- ALONSO-BLANCO C. et al. 2016. 1.135 Genomes Reveal the Global Pattern of Polymorphism in *Arabidopsis thaliana*. **Cell**. v. 166, n° 2, p. 481-491.
- ALLEN E., XIE Z., GUSTAFSON A. M., CARRINGTON J. C., 2005. MicroRNA-directed phasing during trans-acting siRNA biogenesis in plants. **Cell**. v. 121, n° 2, p. 207-221.
- ANDERS S., PYL T. P., HUBER W., 2014. HTSeq — A Python framework to work with high-throughput sequencing data. **Bioinformatics**. v. 31, n° 2, p. 166-169.
- ANDREWS S., 2010. FastQC: a quality control tool for high throughput sequence data. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>
- BAI, Y., DAÍ X., HARRISON A. P., CHEN M., 2014. RNA regulatory networks in animals and plants: a long noncoding RNA perspective. **Briefings in functional genomics**. v. 14, n° 2, p. 1–2.
- Bartel, D. P., 2009. "MicroRNAs: target recognition and regulatory functions". **Cell**. v. 136, n° 2, p. 215-233.
- BEN AMOR B., WIRTH S., MERCHAN F., LAPORTE P., D'AUBENTON-CARAFAY., HIRSCH J., MAIZEL A., MALLORY A., LUCAS A., DERAGON J. M., VAUCHERET H., THERMES C., CRESPI M. 2009. Novel long non-protein coding RNAs involved in *Arabidopsis* differentiation and stress responses. **Genome Res**. v. 19, n° 1, p. 57-69.
- BENTLEY D. R., 2006. Whole-genome re-sequencing. **Curr. Opin. Genet. Dev**. v. 16, n° 6, p. 545–552.
- BERARDINI T. Z., REISER L., LI D., MEZHERITSKY Y., MULLER R., STRAIT E., HUALA E., 2015. The arabidopsis information resource: Making and mining the “gold standard” annotated reference plant genome. **Genesis**. v. 53, n° 8, p. 474-485.
- BOERNER, S. & MCGINNIS, K.M., 2012. Computational identification and functional predictions of long noncoding RNA in *Zea mays*. **PLoS ONE**. v. 7.

- BONSASIO, R. & SHIEKHATTAR, R., 2014. Regulation of Transcription by Long Noncoding RNAs. **Annu. Rev. Genet.** v. 48, p.433–455.
- BORSANI O., ZHU J., VERSLUES P., SUNKAR R., ZHU J. K., 2005. Endogenous siRNAs Derived from a Pair of Natural cis-Antisense Transcripts Regulate Salt Tolerance in *Arabidopsis*. **Cell.** v. 123, n° 7, p. 1279–1291.
- BULYK M. L., 2006. Dna microarray Technologies for measuring protein-DNA interactions. **Current opinion in biotechnology.** v. 17, n° 4, p. 422-430.
- CARVALHO M. C. C. G., SILVA D. C. G. 2010. Sequenciamento de DNA de nova geração e suas aplicações na genômica de plantas. **Ciência Rural, Santa Maria.** v. 40, n. 3, p. 735 –744.
- CHAPMAN E. J., CARRINGTON J. C., 2007. Specialization and evolution of endogenous small RNA pathways. **Nat. Rev. Genet.** v. 8, n° 11, p. 884-896.
- CHEKANOVA J. A., 2015. Long non-coding RNAs and their functions in plants. **Current Opinion in Plant Biology.** v. 27, p.207-216.
- CHEN M., WANG C., BAO H., CHEN H., WANG Y., 2016. Genome-wide identification and characterization of novel lncRNAs in *Populus* under nitrogen deficiency. **Molecular Genetics Genomics.** v. 291, n° 4.
- CLOSE, T. J.; WANAMAKER, S.; CALDO, R. A.; TURNER, S. M.; ASHLOCK, D. A.; DICKERSON, J. A.; WING, R. A.; MUEHLBAUER, G. J.; KLEINHOF, A.; WISE, R. P., 2004. A new resource for cereal genomics: 22K barley Gene Chip comes of age. **Plant Physiology.** v. 134, n° 3, p. 960–968.
- DANECEK P., AUTON A., ABECASIS G., ALBERS C. A., BANKS E., DEPRISTO M. A., HANDSAKER R., LUNTER G., MARTH G., SHERRY S. T., MCVEAN G., DURBIN R. and 1000 Genomes Project Analysis Group, 2011. The Variant Call Format and VCFtools. **Bioinformatics.** v. 27, n° 15, p. 156-158.
- DERRIEN T., JOHNSON R., BUSSOTTI G., TANZER A., DJEBALI S., TILGNER H., GUERNEC G., MARTIN D., MERKEL A., KNOWLES D. G., LAGARDE J., VEERAVALLI L., RUAN X., RUAN Y., LASSMANN T., CARNINCI P., BROWN J. B., LIPOVICH L., GONZALES J. M., THOMAS M., DAVIS C. A., SHIEKHATTAR R., GINGERAS T. R., HUBBARD T. J., NOTREDAME C.,

GUIGÓ R., 2012. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. **Genome Res.** v. 22, n° 9, p. 1775-1789.

DIAS CORREIA J. H. R., 2007. Funcionalidades dos RNA não codificantes (ncRNA) e pequenos RNA reguladores, nos mamíferos. **Revista electrónica de Veterinaria.** v. 8, n° 10.

DING J., LU Q., OUYANG Y., MAO H., ZHANG P., YAO J., XU C., LI X., XIAO J., ZHANG Q., 2012. A long noncoding RNA regulates photoperiod-sensitive male sterility, an essential component of hybrid rice. **PNAS.** v. 109, n° 7, p. 2654-2659.

DJEBALI S., DAVIS C. A., MERKEL A., DOBIN A., LASSMANN T., MORTAZAVI A., TANZER A., LAGARDE J., LIN W., SCHLESINGER F., XUE C., MARINOV G. K., KHATUN J., WILLIAMS B. A., ZALESKI C., ROZOWSKY J., RODER M., KOKOCINSKI F., ABDELHAMID R. F., ANTOSHECHKIN I., BAER M. T., BAR N. S., BATUT P., BELL K., BELL I., CHAKRABORTTY S., CHEN X., CHRAST J., CURADO J., DERRIEN T., DRENKOW J., DUMAIS E., DUMAIS J., DUTTAGUPTA R., FALCONNET E., FASTUCA M., FEJES-TOTH K., FERREIRA P., FOISSAC S., FULLWOOD M. J., GAO H., GONZALEZ D., GORDON A., GUNAWARDENA H., HOWALD C., JHA S., JOHNSON R., KAPRANOV P., KING B., KINGSWOOD C., LUO O. J., PARK E., PERSAUD K., PREALL J. B., RIBECA P., RISK B., ROBYR D., SAMMETH M., SCHAFFER L., SEE L. H., SHAHAB A., SKANCKE J., SUZUKI A. M., TAKAHASHI H., TILGNER H., TROUT D., WALTERS N., WANG H., WROBEL J., YU Y., RUAN X., HAYASHIZAKI Y., HARROW J., GERSTEIN M., HUBBARD T., REYMOND A., ANTONARAKIS S. E., HANNON G., GIDDINGS M. C., RUAN Y., WOLD B., CARNINCI P., GUIGÓ R., GINGERAS T. R., 2012. Landscape of transcription in human cells. **Nature.** v. 489, n° 7414, p. 101-108.

DU TOIT A., 2013. Non-coding RNA: RNA stability control by Pol II. **Nat. Rev. Mol. Cell Biol.** v. 14, n° 3, p. 128-129.

EDWARDS D., STAJICH J., HANSEN D., 2009. Bioinformatics: Tools and Applications. **Springer.** cap. 4, p. 65-96.

EIGEN M., LINDEMANN B. F., TIETZE M., WINKLER- OSWATITSCH R, DRESS A. W. M., VON HAESELER A., 1989. How old is the genetic code? Statistical geometry of tRNA provides an answer. **Science**. v. 244, n° 4905, p. 673–679.

FEDURCO M., ROMIEU A., WILLIAMS S., LAWRENCE I., TURCATTI, G. 2006. BTA a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. **Nucleic Acids Res**. v. 34, n. 3.

FILIPOWICZ W., BHATTACHARYYA S. N., SONENBERG N. 2008. Mechanisms of post-transcriptional regulation by microRNAs: Are the answers in sight? **Nat. Rev. Genet**. v. 9, n° 2, p. 102–104.

FRANCO-ZORRILLA J. M., VALLI A., TODESCO M., MATEOS I., PUGA M. I., RUBIO-SOMOZA I., LEYVA A., WEIGEL D., GARCIA J. A., PAZ-ARES J. 2007. Target mimicry provides a new mechanism for regulation of microRNA activity. **Nat. Genet**. v. 39, n° 8, p. 1033-1037.

GIACHETTO P. F. 2010. A tecnologia de microarranjos na identificação de genes de interesse na bovinocultura. Embrapa Informática Agropecuária. 21. ed.

GLENN, T. C., 2011. Field guide to next-generation DNA sequencers. **Molecular Ecology Resources**. v.11, n° 5, p. 759-769.

GORODKIN J., RUZZO W. L., 2014. Rna sequence, structure, and function: Computational and Bioinformatic methods. **Humana Press**. cap 10, p. 199-214.

GUTTMAN M., AMIT I., GARBER M., FRENCH C., LIN M. F., FELDSEDER D., HUARTE M., ZUK O., CAREY B. W., CASSADY J. P., CABILI M., JAENISCH R., MIKKELSEN T. S., JACKS T., HACHOEN N., BERNSTEIN E., KELLIS M., REGEV A., RINN J. L., LANDER E. S., 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. **Nature**. v. 458, n° 7235, p. 223–227.

HAGEN J. B. 2000. The origins of bioinformatics. **Nature Reviews**. Genetics. v. 1, p. 231-236.

HAMBURGER D., REZZONICO E., MAC-DONALD-COMBER PETÉTOT J.,

- SOMERVILLE C., POIRIER Y., 2002. Identification and characterization of the *Arabidopsis* PHO1 gene involved in phosphate loading to the xylem. **Plant Cell**. v. 14, n° 4, p. 889-902.
- HEO J. B. & SUNG S., 2011. Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. **Science**. v. 331, n° 6013, p. 76-79.
- HUTTENHOFER, A., SCHATTNER, P. & POLACEK, N., 2005. Non-coding RNAs: hope or hype? **Trends in genetics** : TIG, v. 21, n° 5, p. 289–97.
- IWAKIRI J.; MICHIAKI H.; ASAI K., 2015. Bioinformatics tools for lncRNA research. **Biochimica et Biophysica Acta**. v. 1859, n° 1, p. 23-30.
- JIN J., LIU J., WANG H., WONG L., CHUA N. H., 2013. PLncDB: plant long non-coding RNA database. **Bioinformatics**. v. 29, n° 8, p. 1068-1071.
- KAWASAKI H., TAIRA K., WADHWA R., 2004. World of small RNAs: from ribozymes to siRNA and miRNA. **Differentiation**. v. 72, n° 2, p. 58-64.
- KENIRY A., OXLEY D., MONNIER P., KYBA M., DANDOLO L., SMITS G., REIK W., 2012. The H19 lincRNA is a developmental reservoir of miR-675 that suppresses growth and Igf1r. **Nat. Cell Biol**. v. 14, n° 7, p. 659–665.
- KENIRY A., OXLEY D., MONNIER P., KYBA M., DANDOLO L., SMITS G., REIK W., 2012. The H19 lincRNA is a developmental reservoir of miR-675 that suppresses growth and Igf1r. **Nat. Cell Biol**. v. 14, n° 7, p. 659–665.
- KONG L., ZHANG Y., YE Z. Q., LIU X. Q., ZHAO S. Q., WEI L., GAO G., 2007. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. **Nucleic Acids Res**. v. 35, p. W345–W349.
- KOZOMARA A., GRIFFITHS-JONES S., 2014. miRBase: annotating high confidence microRNAs using deep sequencing data. **Nucleic Acids Res**. v. 42, n° D68-73.
- LEE, J.Y., COLINAS, J., WANG, J.Y., MACE, D., OHLER, U., AND BENFEY, P.N., 2006. Transcriptional and posttranscriptional regulation of transcription factor expression in *Arabidopsis* roots. **Proc. Natl. Acad. Sci**. v. 103, n° 15, p. 6055–6060.
- LEE C & KIKYO N. 2012. Strategies to identify long noncoding RNAs involved

in gene regulation. **Cell & Bioscience**. v. 2, n° 37.

LLAVE , C., 2010. Virus-derived small-interfering RNAs at the core of plant-virus interactions. Trends in Plant Science. **Cell**. v. 15, n° 12, p. 701-707.

LI L., EICHTEN S. R., SHIMIZU R., PETSCH K., YEH C. T., WU W., CHETTOOR A. M., GIVAN S., COLE R. A., FOWLER J. E., EVANS M. S. M., SCANLON M. J., YU J., SCHNABLE P. S., TIMMERMANS M. C. P., SPRINGER N. M., MUEHLBAUER G. J., 2014. Genome-Wide discovery and characterization of maize long non-coding RNAs. **Genome Biology**. v. 15, n° 2.

LI S., YAMADA M., HAN X., OHLER U., BENFEY P. N., 2016. High-Resolution Expression Map of the *Arabidopsis* Root Reveals Alternative Splicing and lincRNA Regulation. **Developmental Cell**. v. 39, n° 4, p. 508-522.

LIU X. S., 2007. Getting started in tiling microarray analysis. **PLOS Computational Biology**. v. 3, n° 10, p. 1842-1844.

LIU J., JUNG C., XU J., WANG H., DENG S., BERNAD L., ARENAS-HUERTERO C., CHUA N. H., 2012. Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in *Arabidopsis*. **The Plant cell**. v. 24, n° 11, p. 4333–4345.

LIU J., WANG H., CHUA N. H., 2015. Long noncoding RNA transcriptome of plants. **Plant Biotechnology Journal**. v. 13, n° 3, p. 319-328.

LORENZ R., BERNHART S. H., SIEDERDISSEN C. H. Z., TAHER H., FLAMM C., STADLER P. F., HOFACKER I. L., 2011. ViennaRNA Package 2.0. **Open Access**. v. 6, n°26.

LV Y., LIANG Z., GE M., QI W., ZHANG T., LIN F., PENG Z., ZHAO H., 2016. Genome-wide identification and functional prediction of nitrogen-responsive intergenic and intronic long non-coding RNAs in maize (*Zea mays* L.). **BMC Genomics**. v. 17, n° 350.

MARDIS, E.R., 2008. Next-Generation DNA sequencing methods. **Annu. Rev. Genomics Hum. Genet.** v. 9, p. 387–402.

MARGUERAT S., BAHLER J., 2010. RNA-seq: from technology to biology.

Cellular and Molecular Life Sciences. v. 67, n° 4, p. 569-579.

MATTICK, J.S. & MAKUNIN, I. V., 2005. Small regulatory RNAs in mammals. **Human Molecular Genetics.** v. 14, n°1, p. R121-132.

MEGRAW M., BAEV V., RUSINOV V., JENSEN S. T., KALANTIDIS K., HATZIGEORGIOU A. G., 2006. MicroRNA promoter element discovery in *Arabidopsis*. **Cold Spring Harbor Laboratory Press.** v. 12, n° 9, p. 1612-1619.

MEYERS, B. C.; GALBRAITH, D. W.; NELSON, T.; AGRAWAL, V., 2004. Methods for transcriptional profiling in plants. Be fruitful and replicate. **Plant Physiology.** v. 135, n° 2, p. 637-652.

METZKER, M.L., 2010. Sequencing technologies - the next generation. Nature reviews. **Genetics.** v. 11, n° 1, p.31–46.

MIN X. J, BUTLER G., STORMS R., TSANG A., 2005. OrfPredictor: predicting protein-coding regions in EST-derived sequences. **Nucleic Acids Res.** v. 33, p. W677–W680.

MOCKLER T. C. & ECKER J. R., 2005. Applications of DNA tiling arrays for whole-genome analysis. **Genomics.** v. 85, n° 1, p. 1-15.

MOROZOVA O., HIRST M., MARRA M. A., 2009. Applications of New Sequencing Technologies for Transcriptome Analysis. **Annu. Rev. Genomics Hum. Genet.** v. 10, p. 135-151.

NEI M., 1987. Molecular Evolutionary Genetics. Columbia University Press, New York, NY.

PAUL P., DHANDAPANI V., CHOI S. R., LIM Y. P., 2016. Genome wide identification and functional prediction of long non-coding RNAs in *Brassica rapa*. **Genes Genomics.** v. 38, n° 6, p. 547-555.

PONTING C. P., OLIVER P. L., REIK W., 2009. Evolution and functions of long noncoding RNA. **Cell.** v. 136, n° 4, p. 629-641.

QUAN, M., CHEN, J. & ZHANG, D., 2015. Exploring the Secrets of Long Noncoding RNAs. **International Journal of Molecular Sciences.** v. 16, n° 3, p. 5467–5496.

- QURESHI I. A., MATTICK J. S., MEHLER M. F., 2010. Long non-coding RNAs in nervous system function and disease. **Brain Res.**1338, p. 20–35.
- QUINLAN, A. R., & HALL, I. M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. **Bioinformatics** v. 26, n° 6, p. 841–842.
- R Development Core Team, 2011. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL: <http://www.R-project.org/>.
- REARICK D., PRAKASH A., MCSWEENEY A., SHEPARD S. S., FEDOROVA L., FEDOROV A., 2010. Critical association of ncRNA with introns. **Nucleic Acids Research.** v. 39, n° 6, p. 2357–2366.
- RYMARQUIS L. A., KASTENMAYER J. P., HUTTENHOFER A. G., GREEN P. J., 2008. Diamonds in the rough: mRNA-like non-coding RNAs. **Trends Plant Sci.** v. 13, n° 7, p. 329-334.
- RIZK N. J. 2006. Parallel Computing for Bioinformatics and Computational Biology. Models, Enabling Technologies, And Case Studies. Edited by Albert Y. Zomaya. Wiley-Interscience.
- ROBERTS A., PIMENTEL H., TRAPNELL C., PACHTER L., 2011. Identification of novel transcripts in annotated genomes using RNA-Seq. **Bioinformatics.** v. 27, n° 17, p. 2325–2329.
- SANGER F., NICKLEN S., COULSON A. R., 1977. DNA sequencing with chain-terminating inhibitors. **Proceedings of the National Academy of Sciences.** v. 74, n° 12, p. 5463-5467.
- SCHENA M., SHALON, D., DAVIS, R. W., BROWN P. O., 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. **Science**, v. 270, n° 5235 p. 467-470.
- SCHENA M., 1996. Genome analysis with gene expression microarrays. **Bioessays.** v. 18, n° 5, p. 427-431.
- SHIN S. I., AHN H., JUNG I., RHEE S., KIM J., KWON . B., 2016. Novel drought-responsive regulatory coding and non-coding transcripts from *Oryza Sativa* L. **Genes Genomics.** v. 38, n° 10, p. 949-960.

- SOUSA C., JOHANSSON C., CHARON C., MANYANI H., SAUTTER C., KONDOROSI A., CRESPI M., 2001. Translation and structural requirements of the early nodulin gene Enod40, a short-open reading frame-containing RNA, for elicitation of a cell-specific growth response in the alfalfa root cortex. **Molecular and Cellular Biology**. v. 21, n° 1, p. 354-366.
- STOREY J. D., LEEK J. T., BASS A. J., 2015. edge: Extraction of Differential Gene Expression. R package version 2.6.0.
- SULIVAN A. M., ARSOV-SKY A. A., QUEITSCH C., STAMATOYANNOPOULOS J. A., 2014. Mapping and Dynamics of Regulatory DNA and Transcription Factor Networks in *A. thaliana*. **Cell Reports**. v. 8, n° 6, p. 2015-2030.
- STOREY J. D., LEEK J. T., BASS A. J., 2015. *edge: Extraction of Differential Gene Expression*. R package version 2.6.0
- SWIEZEWSKI S., LIU F., MAGUSIN A., DEAN C. 2009. Cold-induced silencing by long antisense transcripts of an *Arabidopsis* polycomb target. **Nature**. v. 462, p. 799–802.
- TIAN D., SUN S., LEE J.T., 2010. The long noncoding RNA, Jpx, is a molecular switch for X chromosome inactivation. **Cell**. v. 143, n° 3, p. 390–403.
- The 1001 Genomes Consortium, 2016. 1135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. **Cell**, v. 166, n° 2, pag. 481-491.
- The *Arabidopsis* Genome Initiative, 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. **Nature**. v. 408, p. 796-815.
- TRAPNELL C., ROBERTS A., GOFF L., PERTEA G., KIM D., KELLEY D. R., PIMENTEL H., SALZBERG S. L., RINN J. L., PACHTER L., 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. **Nature Protocols**. v.7, p. 562-578.
- TURCATTI G., ROMIEU A., FEDURCO M., TAIRI A. P. A., 2008. New class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. **Nucleic Acids Res**. v. 36, n° 4.
- ULITSKY I. & BARTEL, 2013. lincRNAs: Genomics, Evolution and Mechanisms.

Cell. v. 154, n° 1, p. 26-46.

VALLOT C., HURET C., LESECQUE Y., RESCH A., OUDRHIRI N., BENNACEUR-GRISCELLI A., DURET L., ROUGEULLE C., 2013. XACT, a long noncoding transcript coating the active X chromosome in human pluripotent cells. **Nat. Genet.** v. 45, p. 239–241.

VAZQUEZ F., VAUCHERET H., RAJAGOPALAN R., LEPERS C., GASCIOLLI V., 2004. Endogenous trans-acting siRNAs regulate the accumulation of *Arabidopsis* mRNAs. **Molecular Cell**, v.16, n° 1, p. 69–79.

VAZQUEZ F., 2006. *Arabidopsis* endogenous small RNAs: highways and byways. **TRENDS in Plant Science.** v. 11, n° 9, p. 460-468.

VAUCHERET H. & FAGARD M., 2001. Transcriptional gene silencing in plants: targets, inducers and regulators. **TRENDS in Genetics.** v. 17, n. 1, p. 29-35.

VOINNET O., 2009. Origin, Biogenesis, and Activity of Plant MicroRNAs. **Cell Press.** v.136, n° 4, p. 669-687.

WANG Z., GERSTEIN M., SNYDER M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. **Nat Rev Genet.** v. 10, n° 1, p. 57–63.

WANG Y., FAN X., LIN F., HE G., TERZAGHI W., ZHU D., DENG X. W. 2014. *Arabidopsis* noncoding RNA mediates control of photomorphogenesis by red light. **Proc. Natl. Acad. Sci.** v. 111, n° 28, p.10359 –10364.

WANG H., CHUNG P. J., LIU J., JANG I. C., KEAN M., XU J., CHUA N. H., 2014. Genome-wide identification of long noncoding natural antisense transcripts and their responses to light in *Arabidopsis*. **Genome Research.** v. 24, n° 3, p. 444–453.

WANG B., TSENG E., REGULSKI M., CLARCK T. A., HON T., JIAO Y., LU Z., OLSON A., STEIN J. C., WARE D., 2016. Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. **Nature Communications.** v. 24, n° 7.

WAPINSKI O. & CHANGH. Y., 2011. Long noncoding RNAs and human disease. **Cell Biology.** v. 21, n° 6, p. 354-361.

WATSON, J. D., BAKER, T .A., BELL S.P., GANN A., LEVINE M., LOSICK R., 2006. *Biologia Molecular do Gene*. 5.a ed. Porto Alegre, **Artmed**, p. 728.

WORTMAN J. R., HAAS B. J., HANNICK L. I., SMITH R. K. JR, MAITI R, RONNING CM, CHAN AP, YU C, AYELE M., WHITELOW CA, 2003. Annotation of the *Arabidopsis* genome. **Plant Physiol.** v. 132, n° 2, p. 461–468.

XIE Z., JOHANSEN L.K., GUSTAFSON A. M., KASSCHAU K.D., LELLIS A.D., 2004. Genetic and functional diversification of small RNA pathways in plants. **PLoS Biology.** v. 2, n° 5.

YAMADA M., 2016. Functions of long intergenic non-coding (linc) RNAs in plants. **J Plant Res.** Springer.

YAZAKI J., GREGORY B. D., ECKER J. R., 2007. Mapping the genome landscape using tiling array technology. **Current Opinion in Plant Biology.** v. 10, n° 5, p. 1-9.

ZHU Q. H., STEPHEN S., TAYLOR J., HELLIWELL C. A., WANG M. B., 2013. Long noncoding RNAs responsive to *Fusarium oxysporum* infection in *Arabidopsis thaliana*. **New Phytol.** v. 201, n° 2, p. 574-584.

YU X., YANG J., LI X., LIU X., SUN C., WU F., HE Y., 2013. Global analysis of cis-natural antisense transcripts and their heat-responsive nat-siRNAs in *Brassica rapa*. **BMC Plant Biol.** v.13, n° 208.