Universidade Federal do Rio Grande do Sul
Programa de Pós-Graduação em Genética e Biologia Molecular

Identificação e análise de expressão de microRNAs em tecidos florais de soja (*Glycine max* (L.) Merrill)

Lorrayne Gomes Molina Gromann

Dissertação submetida ao Programa de Pós-Graduação em Genética e Biologia Molecular da UFRGS como requisito parcial para obtenção do grau de Mestre em Genética e Biologia Molecular.

Orientador: Prof. Dr. Rogério Margis

Porto Alegre, abril de 2012.

INSTITUIÇÕES E FONTES FINANCIADORAS

Esta dissertação foi desenvolvida no Laboratório de Genomas e Populações de Plantas do Centro de Biotecnologia da Universidade Federal do Rio Grande do Sul, sob orientação do Dr. Rogério Margis, através de bolsa de mestrado disponibilizada pelo CNPq.

Este trabalho foi financiado pelo consórcio GenoSoja (CNPq 5527/2007-8), GenoProt (CNPq 559636/2009-1) e projeto Estruturante de Agro-Energia (FAPERGS-FINEP).

AGRADECIMENTOS

- Aos membros da banca examinadora, Prof. Dra. Maria Helena B. Zanettini, Dra. Ana Paula Korbes e Prof. Dra. Claudia M. B. Andrade.

- Ao Prof. Dr. Rogério Margis, pela orientação e pelo fornecimento de recursos para execução deste e de outros trabalhos.

- À Dra. Andréia Turchetto Zolet, pelas contribuições como revisora desta dissertação.

- Ao PPGBM e à Universidade Federal do Rio Grande do Sul.

- Ao CNPq pela concessão da bolsa de Mestrado.

- Ao Elmo, pela sua eficiência e boa vontade em ajudar e facilitar a vida dos alunos do PPGBM.

- Aos colegas e amigos do LGPP – Laboratório de Genomas e Populações de Plantas: Fernanda (mestranda), Franceli (doutoranda), Ana Paula (mestranda), Frank (doutorando). Em especial, ao Luiz Felipe Oliveira (doutorando), pelas análises de DEGseq e construção dos heatmaps; ao Guilherme Loss e Guilherme da Fonseca (doutorandos), pelas enriquecedoras discussões e ajuda com os programas de bioinformática.

- À minha irmã Lorena e sobrinha Helena, pelos alegres finais de semana de descanso em Porto Alegre e em Passo Fundo.

- Aos meus pais, Paulo Sérgio e Rosemeire, por me apoiarem com palavras de conforto nas horas difíceis e continuarem me dando "colo" quando eu preciso.

- Ao meu marido Tiago, por todo carinho, paciência e ajuda nos momentos em que o trabalho tomou todo o meu tempo, e pelas palavras de força.

- A Deus, por me dar vida e inspiração. Porque todas as coisas são para Ele.

SUMÁRIO

LISTA DE ABREVIATURAS

AAA – Do inglês, “ATPases Associated with diverse cellular Activities”

AFB – Do inglês, “auxin signaling f-box”

AG – AGAMOUS

AGL – AGAMOUS-LIKE

AGO – Argonauta, proteína envolvida na RNAi, presente no complexo RISC, com atividade de RNAse e/ou ligação de RNAs

AKR – Aldo-ceto redutase; do inglês, “Aldo/Keto Reductase”

AP – APETALA

ARF – Fator de resposta à auxina; do inglês, “Auxine response factors”

ARF-GAP – Do inglês, “Adenosine diphosphate Ribosylation Factor-GTPase-Activating Protein”

BLAST – “Basic Local Alignment Search Tool”, ferramenta de busca ou procura de sequências nucleotídicas ou peptídicas por alinhamento básico local

blastn – Algoritmo da ferramenta BLAST que inspeciona um banco de dados de sequências codificantes, utilizando sequências peptídicas como isca.

BRD – Do inglês, “Bromodomain”

C1 – Domínio de ligação a ésteres de forbol/diacilglicerol

CAR – Biblioteca de sRNAs proveniente de amostras de carpelos

CBF-B – Subunidade B do fator de ligação a CCAAT; do inglês, “CCAAT-Binding Factor, subunit B”

CCHC – Motivo Cys-Cys-His-Cys

cDNA – DNA complementar

CDS – Sequência codante, do inglês, “coding sequence”

CNA –CORONA

CO – CONSTANS

CUC – CUP-SHAPED COTYLEDON

CYP450 – Citocromo P450; do inglês, “Cytochrome P450”

DCL – Proteína envolvida na RNAi, com atividade de RNAse e/ou ligação de dsRNAs do tipo “dicer”; do inglês, “Dicer-Like”

DHHC – Motivo Asp- His-His-Cys

DNA – Ácido desoxirribonucleico, do inglês, “desoxiribonucleic acid”

dsRNA – RNA de dupla fita, do inglês, “double strand RNA”

EF-Tu – Do inglês, “Elongation factor thermo unstable”

ELFV – Motivo Glutamato-leucina-phenylalanine-valine

EST – Biblioteca de sRNAs proveniente de amostras de estames

ESTs – Sequências expressas obtidas de cDNAs; do inglês, “Expressed Sequence Tags”

FT – FLOWERING LOCUS T

GalBL – Do inglês, “Galactose binding lectin”

GA – Giberelina

GAP – Proteína ativadora de GTPase; do inglês, “GTPase-Activating Protein”

GH – Do inglês, “Glycosyl hydrolase”

Gma – *Glycine max*

HAP 2 – Do inglês, “Heme activator protein 2”, fatores de transcrição também chamados de NF-YA

HD – Do inglês, “Homeodomain”

HD-ZIP – Fatores de transcrição; do inglês, “Homeodomain-leucine zipper proteins”

HEAT repeat – Domínio encontrado em certas proteínas citoplasmáticas, incluindo as quatro que deram origem ao acrônimo (Huntingtin, elongation factor 3 (EF3), protein phosphatase 2A (PP2A), and the yeast PI3-kinase TOR1)

HEN – ativador de HUA (ativador de agamous); do inglês, “HUA Enhancer”

HSP40 – Do inglês, “Heat Shock Protein 40 kD”

HST – HASTY

HTS: Do inglês, “High-Throughput Sequencing”

HYL1 – HYPONASTIC LEAVES1, proteína ligante de dsRNA

LAGLIDADG – Sequência consenso que deu nome à família de DNA endonuclease de "homing"

LepA/EF4 – Do inglês, “Leader peptidase A/ Elongation factor 4”

LFY – LEAFY

LRR – Repetições ricas em leucina; do inglês, “Leucine-Rich Repeat”

MAPKK – Do inglês, “Mitogen-activated protein kinase kinase”

MIP – Do inglês, “Major Intrinsic Protein”

miRNA – microRNA

mRNA – RNA mensageiro

MSF – Do inglês, “Major Facilitator Superfamily”

MuDR/Mu – Elemento autônomo da família Mutator de TEs

MULE – Do inglês, “Mutator-like elements”

NAM – Do inglês, “No apical meristem”

NB – Do inglês, “nucleotide-binding”

NER – Mecanismo de reparo do DNA por excisão de nucleotídeo; do inglês, “Nucleotide Excision Repair”

NF-YA – Do inglês, “Nuclear Factor Y Alpha-related”

nt – nucleotídeo

PCR – Reação em cadeia da polimerase, do inglês, “polymerase chain reaction”

PET – Biblioteca de sRNAs proveniente de amostras de pétalas

PH – Do inglês, “Plant pleckstrin”

phasiRNAs – siRNAs em fase; do inglês, “Phased siRNAs”

PHB – PHABULOSA

PHD – Do inglês, “Plant homeodomain”

PHV – PHAVOLUTA

PI – PISTILLATA

Pol ε –DNA Polimerase epsilon

PORR – Do inglês, “Plant Organelle RNA Recognition Domain”

PPIase – Do inglês, “Peptidylprolyl cis-trans isomerase”

PPIC – Do inglês, “Peptidylprolyl Isomerase C”

PPR – Repetição pentatricopeptídica; do inglês, “PentatricoPeptide Repeat”

Pre-miRNA – Precursor de miRNA

PTGS – Silenciamento gênico pós-transcricional, do inglês, “Post-transcriptional gene silencing”

Rab – GTPase que regula o tráfego pela membrana plasmática

RACE – Do inglês, “Rapid amplification of cDNA ends”

REV – REVOLUTA

RISC - Complexo indutor do silênciamento por RNA, do inglês, “RNA-Induced Silencing Complex”

RLK – Receptor quinase de plantas; do inglês, “Receptor-Like Kinase”

RNA – Ácido ribonucleico; do inglês, “RiboNucleic Acid”

RNA Pol – RNA polimerase

RNAi – RNA de interferência

RNAPII – RNA polimerase II

rRNA – RNA ribossomal

RT – Transcriptase reversa; do inglês, “Reverse Transcriptase”

RT-qPCR – PCR quantitativa em tempo real, do inglês, “Real-time quantitative PCR”

SAM – Do inglês, “S-adenosyl methionine”

SCRL – Do inglês, “Plant self-incompatibility response protein”

SDN1 – Nuclease de sRNAs, do inglês, “Small RNA Degrading Nuclease 1”

SE – SERRATE

SEP – SEPALLATA

siRNA – Pequeno RNA de interferência, do inglês, “short interference RNA”

SMZ – SCHLAFMÜTZE

SNZ – SCHNARCHZAPFEN

SOAP – programa de alinhamento de pequenos oligonucleotídeos, do inglês, “short oligonucleotide alignment program”

SPL – Do inglês, “SQUAMOSA Promoter-Binding Protein-Like”

sRNA – pequeno RNA, do inglês, “small RNA”

START – Do inglês,"STeroidogenic Acute Regulatory protein–related lipid Transfer"

ta-siRNAs – Do inglês, “trans-acting-siRNAs”

TE – Elemento transponível; do inglês, “Transposable Element”

TIR – Do inglês, “Toll/Interleukin-1 Receptor”

TOE – TARGET OF EAT

tRNA – RNA transportador

UFD – Do inglês, “Ubiquitin fusion degradation protein”

Zf – Dedo de zinco; do inglês, “Zinc finger”

RESUMO

A soja é uma das culturas mais importantes a nível mundial, devido à produção de óleo e a seu alto teor proteico. A fase reprodutiva é a mais importante para a produtividade da soja, visto que seu cultivo se destina principalmente à produção de grãos. Os microRNAs (miRNAs) desempenham funções essenciais em diversos aspectos do desenvolvimento reprodutivo, incluindo o florescimento, a fertilidade e o desenvolvimento da semente. A função destes pequenos RNAs (sRNAs) endógenos não codificantes é regular a expressão gênica, principalmente através de clivagem e inibição da tradução de mRNAs alvos. A identificação de miRNAs ainda não está saturada e, em soja, não há trabalhos relacionando-os aos diferentes órgãos florais, que são fundamentais na produtividade desta cultura. Neste estudo, amostras de flores, carpelos, estames e pétalas de soja foram usadas na construção de quatro bibliotecas de sRNAs sequenciadas utilizando a plataforma Solexa, gerando um total de 13.557.795 sequências. Através da análise do mapeamento das sequências de sRNAs das bibliotecas em candidatos a precursores de miRNAs de soja identificados, 276 foram considerados precursores autênticos, incluindo 143 precursores novos. Foram identificados 235 miRNAs maduros, dos quais 51 são miRNAs inéditos, pertencentes a 40 novas famílias. Os demais miRNAs identificados pertencem a 64 famílias conhecidas de miRNAs de plantas, das quais três ainda não tinham sido reportadas em soja. Todas as famílias de miRNAs que estão envolvidas na regulação do florescimento foram identificadas entre as mais frequentes nos tecidos florais de soja. Na análise de expressão pela frequência de sequências nas bibliotecas de sRNAs de carpelos, estames e pétalas, 67.2% (158) dos miRNAs identificados foram diferencialmente expressos. A maioria dos miRNAs de 22 e 24 nt diferencialmente expressos foi induzida nos carpelos, sugerindo que miRNAs destes tamanhos são importantes na regulação de processos que ocorrem especificamente nestes órgãos. Análises de expressão por PCR quantitativa (RT-qPCR) comprovaram a expressão diferencial de 19 miRNAs. O miRNA inédito denominado NF13 apresentou a maior diferença entre os tecidos, sendo fortemente induzido nos carpelos. Para os miRNAs com expressão diferencial comprovada por RT-qPCR foi feita a predição computacional dos genes alvos, para muitos dos quais já foram descritas funções relacionadas ao processo reprodutivo em plantas. O estudo da regulação destes genes pelos miRNAs em diferentes tecidos e estádios de desenvolvimento floral

contribuirá para o entendimento dos mecanismos moleculares envolvidos na reprodução da soja.

ABSTRACT

Soybean is one of the most important crops worldwide, due to the production of oil and its high protein content. The reproductive phase is considered the most important for the yield of soybean, which is mainly intended to produce the grains. MicroRNAs (miRNAs) play essential roles in various aspects of reproductive development, including flowering, fertility and seed development. The function of these endogenous small non-coding RNAs (sRNAs) is to regulate gene expression, mainly through cleavage and translation inhibition of target mRNAs. The identification of miRNAs is not yet saturated in soybeans, and there are no studies linking them to the different floral organs, which are fundamental in the productivity of this crop. In this study, samples of flowers, carpels, petals and stamens of soybeans were used in the construction of four sRNA libraries sequenced using the platform Solexa, generating a total of 13,557,795 sequences. The sRNAs sequences from four libraries were mapped in precursors candidates. Among them, 276 were considered authentic precursors, including 143 new precursors. 235 mature miRNAs were identified, of which 51 are novel miRNAs belonging to 40 new families. The other identified miRNAs belongs to 64 known plant miRNA families, of which three had not yet been reported in soybean. All miRNAs families which are involved in regulating flowering were identified among the most frequent floral tissue of soybean. Expression analysis based on the frequency of sequences in the libraries of sRNAs of carpels, stamens and petals demonstrated that 67.2% (158) corresponded to differentially expressed miRNAs. Most of 22 and 24 nt miRNAs that were differentially expressed was induced in carpels, suggesting that these miRNAs sizes are important in regulating the processes occurring specifically in these organs. Analysis of expression by quantitative PCR (RT-qPCR) confirmed the differential expression of 19 miRNAs. The novel miRNA named NF13 showed the greatest difference between the tissues and is strongly induced in the carpels. A computational prediction of targets for miRNAs with differential expression confirmed by RT-qPCR was performed. Many of the predicted targets have described functions related to the reproductive process in plants. The study of regulation of these genes by miRNAs in different tissues and stages of flower development will contribute to understanding the molecular mechanisms involved in reproduction of soybean.

1 INTRODUÇÃO

1.1 A espécie *Glycine max* e a soja cultivada

O gênero *Glycine* pertence à família das leguminosas (Fabaceae), subfamília Papilionoideae, e tribo Phaseoleae. A soja cultivada e a soja selvagem, pertencentes ao subgênero *Soja*, são plantas com ciclo de vida anual, enquanto outras 23 espécies selvagens do subgênero *Glycine* são perenes (Orf, 2010).

Evidências genéticas e moleculares sugerem que um ancestral ao gênero *Glycine* com 2*n*=2*x*=20 se originou no sudeste da Ásia. A partir deste, por auto ou alopoliploidia, surgiu uma espécie selvagem perene (2*n*=4*x*=40, desconhecida ou extinta) que migrou para a China. Houve uma subseqüente evolução para a espécie selvagem anual (2*n*=4*x*=40; *G. soja*) e finalmente para a espécie cultivada (2*n*=4*x*=40; *G. max*), as quais sofreram diploidização, pois comportam-se como diplóides na meiose (Orf, 2010).

Tanto a soja cultivada (*G. max*) quanto a selvagem (*G. soja*) exibem ampla variabilidade fenotípica. Isto inclui a morfologia e maturidade da planta, a morfologia da flor, a forma, tamanho e cor da semente, características de resistência a estresses bióticos e abióticos e características fisiológicas e bioquímicas, como o conteúdo de proteína, óleo e carboidratos da semente (Cregan, 2008; Orf, 2010).

A domesticação da soja a partir da espécie selvagem ocorreu na China central ou do sul, aproximadamente 5.000 anos atrás (Wilson, 2008). Atualmente, a soja é uma das culturas mais importantes a nível mundial. Com teor de óleo de 20%, esta oleaginosa responde por 30% do óleo vegetal produzido no mundo, além de ser nutritiva por seu alto teor protéico (40%), sendo amplamente utilizada na alimentação humana e animal (Dall´Agnol e Hirakuri, 2008; Zhang *et al*., 2008). Suas aplicações incluem ampla variedade de alimentos à base de soja, uso do óleo na produção do biodiesel e uso da proteína em substituição à carne. Além disso, a soja é uma fonte primária de produtos secundários de alto valor, como lecitina, vitaminas, nutracêuticos e anti-oxidantes (Wilson, 2008).

Os maiores produtores de soja são os Estados Unidos, Brasil e Argentina, mas a área global de produção de soja tem atingido um platô. Se esta tendência continuar ou piorar, haverá grande pressão para o ganho genético na produção para garantir suprimento suficiente de soja e de seus produtos. Neste sentido, os mais de 156.849 acessos de *G. max* das coleções de germoplasma do mundo fornecem uma base para avanços futuros em tecnologia genética necessária para prover cultivares elites de soja com adequada proteção contra pestes e doenças e maior capacidade de produção e qualidade de seus produtos (Wilson, 2008).

1.2 Crescimento e desenvolvimento das plantas de soja

O desenvolvimento da soja pode ser separado em duas fases principais: vegetativa e reprodutiva, que são subdivididas em vários estádios de crescimento, os quais podem se sobrepor, ocorrendo simultaneamente na mesma planta. A duração de cada uma é controlada pela temperatura, comprimento do dia, fatores genéticos, entre outros. Um estádio se inicia quando 50% das plantas no campo atingem este estádio específico. Os estádios vegetativos começam desde a emergência da planta no solo (VE), passando por um estádio de folhas unifolioladas (VC), a partir do qual cada estádio vegetativo é designado pela letra V seguida de um número que representa o número de nós na haste principal contendo uma folha totalmente desenvolvida (Pedersen *et al.*, 2007).

A soja é uma planta de dias curtos, pois floresce quando o comprimento da noite excede um período de tempo crítico. Variedades de soja de hábito de crescimento determinado são caracterizadas pela presença de um racemo terminal e geralmente têm entre 15 a 20 nós, enquanto algumas plantas indeterminadas têm entre 22 e 24 nós, pois continuam seu desenvolvimento vegetativo mesmo enquanto florescem e produzem vagens (Ashlock e Purcell).

As flores de soja são papilionadas, brancas ou roxas, muito pequenas (4-6 mm de comprimento), compostas por cinco pétalas que envolvem um pistilo e dez estames, dos quais nove formam um tubo ao redor do pistilo e o décimo é livre (Figura 1) (Veja, 2000).

O pólen é liberado das anteras diretamente sobre o estigma, pouco antes da abertura das flores (que ocorre pela manhã), favorecendo a autofecundação (Scaboo *et al*., 2010).

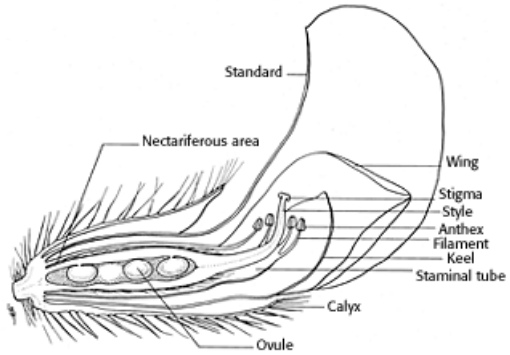


Figura 1. Seção longitudinal da flor de soja (McGregor, 1976)

Os estádios da fase reprodutiva são denominados pela letra R seguidos de números. Esta fase começa com a abertura de uma única flor em qualquer local da haste (R1). Na soja de hábito de crescimento determinado, a primeira flor geralmente aparece nos nós do ápice. O florescimento então se espalha ao longo da haste principal e dos ramos (R2) As flores formadas antes serão fertilizadas primeiro e formarão as primeiras vagens (R3). Uma vez formadas, as vagens passam por um rápido processo de expansão até atingirem seu tamanho máximo (R4). Então, ocorre o enchimento das sementes no interior das vagens (R5, R6) até que a planta atinja a maturidade fisiológica, a qual se inicia no momento em que uma vagem na haste principal atinge a coloração marrom (R7) e se completa quando a maioria das vagens e das sementes estão secas e apresentam esta coloração (R8) (Pedersen *et al*., 2007).

A fase reprodutiva é a mais importante para a produtividade da soja, visto que seu cultivo se destina basicamente à produção dos grãos. Os estádios de R1 a R6 são críticos, pois é quando o número de vagens e de sementes é determinado. O tamanho das sementes também é influenciado neste período, pois nos estádios de R3 a R6 o número de células cotiledonares é determinado (Pedersen *et al*., 2007).

1.3 Aspectos moleculares do florescimento

Flores são arranjos complexos de estruturas com funções especializadas que diferem da parte vegetativa da planta tanto na forma quanto no tipo celular. Durante o desenvolvimento vegetativo, o meristema apical produz folhas, até que inicie o desenvolvimento reprodutivo e ocorra sua transição para meristema da inflorescência, que produz meristemas florais, culminando na formação dos órgãos florais (McKim e Hay, 2010). Diversas rotas regulatórias controlam a transição da fase de crescimento vegetativo para o crescimento reprodutivo, incluindo estímulos ambientais e endógenos. Assim, o tempo para o florescimento depende da integração de respostas a esses sinais, sendo uma característica quantitativa cuja regulação gênica é muito precisa. Esta complexidade garante a ocorrência do florescimento em condições favoráveis ao sucesso reprodutivo (Terzi e Simpson, 2008).

Arabidopsis thaliana tem sido a planta modelo para a caracterização dos mecanismos moleculares que controlam o florescimento e identificação dos genes envolvidos nesse processo (revisado por Irish, 2010). Nesta espécie, o florescimento é induzido por vias de respostas geneticamente separadas, sendo as principais: fotoperíodo, vernalização, temperatura, idade da planta, concentração de giberelina e via autônoma (revisadas por Fornara *et al*., 2010).

As diversas rotas convergem em um conjunto de genes chamados de integradores florais, os quais são reprimidos até o momento do florescimento (Terzi e Simpson, 2008). Então, genes que especificam identidade de meristema floral são ativados. O fator de transcrição LEAFY (LFY) é considerado um regulador chave neste processo, pois, além de ser um integrador, também especifica identidade de meristema floral (revisado por Moyroud *et al*., 2010). APETALA1 (AP1) também tem função crucial na indução do florescimento e é parcialmente redundante com LFY no papel de especificar identidade de meristema floral. *AP1* é diretamente ativado por LFY, inibe uma série de repressores florais e controla a expressão de genes homeóticos, estando envolvido na formação de sépalas e pétalas (revisado por Kaufmann *et al*., 2010).

Em geral, flores perfeitas são compostas por quatro tipos de órgãos – sépalas, pétalas, estames e carpelos – arranjados nesta ordem em quatro anéis concêntricos, os verticilos. A identidade de órgão floral inclui sua forma e tamanho, que são determinados pelo balanço entre divisão e expansão celular. A redução na divisão celular leva ao aumento da expansão celular e *vice versa* (Dornelas *et al*., 2010). Baseado em mutantes das plantas modelo *Arabidopsis* e *Antirrhinum* apresentando transformações homeóticas, foi proposto o modelo ABC, que explica como poucos genes coordenam a formação dos quatro tipos de órgãos florais (Causier *et al*., 2010). Este modelo propõe que as proteínas que especificam identidade de órgão floral combinam-se de forma que cada verticilo é definido pela expressão de uma única classe de proteínas ou pela combinação delas. As proteínas A (AP1 e AP2) especificam sépalas no primeiro verticilo (mais externo); a co-expressão de A e B (ex. de gene B: *AP3*, *PISTILATA* - *PI*) especifica pétalas no segundo verticilo; B e C (ex. de gene C: *AGAMOUS* - *AG*) especificam estames no terceiro verticilo; C especifica carpelos no quarto verticilo (mais interno) e as proteínas D (também classificadas como proteínas de classe C mais específicas) coordenam a formação dos óvulos (Causier *et al*., 2010).

Os genes ABC, exceto *AP2*, codificam fatores de transcrição MADS-box, que ligam a motivos $CC[A/T]_6GG$ no DNA como complexos quaternários e seus possíveis alvos e funções foram revisados (Sablowski, 2010). Foi verificado que a formação destes complexos é mediada por proteínas MADS-box da classe E (SEP, anteriormente chamadas de AGAMOUS-LIKE2 - AGL2). Este modelo de multimerização prediz que os verticilos 1, 2, 3 e 4 contêm os complexos AP1/AP1/SEP/SEP, AP1/SEP/AP3/PI, AG/SEP/AP3/PI e AG/AG/SEP/SEP, respectivamente (Immink *et al*., 2010; Liu e Mara, 2010).

Outro aspecto importante no modelo ABC é o antagonismo entre as funções A e C, que define os domínios de expressão destes genes em *Arabidopsis*, mas que não tem sido demonstrado em outras espécies (Causier *et al*., 2010). Além disso, homólogos de genes da classe A em outras espécies, embora desempenhem função na regulação da identidade de meristema, não especificam identidade de sépala e pétala. Por isso, sugere-se que a função A tenha sido recentemente adquirida na evolução, representando uma nova modificação de uma função mais ancestral (Irish, 2010).

Em soja, foram identificados 28 fatores de transcrição enriquecidos nas flores, incluindo 12 MADS-box homólogos a *AP1*, *AP3*, *PI*, *AGL2/SEP1* e *AGL9/SEP3*. O ortólogo de *SEP1* foi caracterizado e mostrou ter grande importância no desenvolvimento de pétalas, como esperado. Por outro lado, diferente de *AP3* de *Arabidopsis*, que é expresso especificamente em tecidos reprodutivos, alguns genes B de soja também foram fortemente expressos em tecidos vegetativos, como raízes e folhas, sugerindo papéis adicionais para estes genes em soja (Huang *et al*., 2009).

O gene homólogo de *AGL11*, que especifica identidade de óvulo em *Arabidopsis*, também foi caracterizado em soja (*GmGAL2*), sendo mais expresso em flores e vagens, o que sugere um papel no desenvolvimento de órgão, como em *Arabidopsis*. No entanto, a função deste gene é espécie-específica e difere entre plantas de dias longos e plantas de dias curtos. Em soja, *GmGAL2* é mais expresso em dias curtos que em dias longos e sua superexpressão em *Arabidopsis* acelera o tempo de floração, sugerindo um papel adicional à formação do óvulo (Xu *et al*., 2010). O homólogo de *AGL20/SOC1* em soja (*GmGAL1*) também foi caracterizado e sua expressão foi regulada pelo ciclo circadiano, oscilando em diversos órgãos e estádios de desenvolvimento durante o ciclo de vida, sugerindo que este é um gene multifuncional no desenvolvimento da soja (Zhong *et al*., 2012).

Estes são alguns exemplos que ilustram as diferenças espécie-específicas nos mecanismos moleculares de desenvolvimento floral. Como a floração é um processo de importância vital para a agricultura por ser um dos determinantes da produtividade, é indispensável seu estudo em espécies agronomicamente importantes, como a soja.

1.4 MicroRNAs

MicroRNAs (miRNAs) são pequenos RNAs (sRNAs) endógenos não codificantes de aproximadamente 20 a 24 nucleotídeos que regulam negativamente a expressão gênica em eucariotos de maneira sequência-específica (Zhang *et al*., 2006; Voinnet, 2009; Naqvi *et al*., 2012).

O primeiro miRNA descoberto foi lin-4 em *Caenorhabditis elegans*, sendo essencial ao controle temporal normal de diversos eventos do desenvolvimento pós-embrionário (Lee *et al.*,1993). Foi verificado que tal controle se dava pela regulação dos níveis da proteína LIN-14 e que o produto do gene lin-4 não era uma proteína, mas uma sequência de RNA de 22nt. O fato de que este sRNA era complementar a sequências na região não traduzida 3' (3'UTR) do RNA mensageiro (mRNA) de lin-14 sugeriu que lin-4 regula a tradução de lin-14 via interação RNA-RNA (Lee *et al.*,1993). Anos depois, foi identificado neste mesmo organismo outro miRNA regulado temporalmente (Reinhart *et al.*, 2000): Let-7, de 21nt e complementar à 3'UTR de vários genes, controlando negativamente sua expressão em muitos animais bilaterais, incluindo seres humanos (Pasquinelli *et al*., 2000).

Os primeiros miRNAs de plantas foram descritos em *A. thaliana* (Park *et al*., 2002), tendo sido, desde então, extensivamente estudados. Os genes de miRNAs de plantas se encontram aleatoriamente distribuídos nos genomas, muitos dos quais se originaram da duplicação invertida de genes codificadores de proteínas com subsequentes mutações. A duplicação em *tandem* de miRNAs pré-existentes e subsequente dispersão pelo genoma através de rearranjos cromossômicos e a duplicação de todo o genoma foram importantes para a expansão de famílias gênicas de miRNAs. Além disso, elementos transponíveis (TEs) parecem contribuir com o surgimento de genes de miRNAs espécie-específicos (Nozawa *et al*., 2012).

A maioria dos genes de miRNAs em plantas constituem unidades transcricionais independentes, estando sob influência de seus próprios promotores, que são muito similares aos dos genes codificadores de proteínas. Assim, eles são controlados por vários fatores de transcrição e por modificações na cromatina, incluindo metilação de DNA e modificação de histonas (Meng *et al*., 2011). Portanto, estímulos externos, incluindo componentes bióticos e abióticos, levam à modulação dos níveis de expressão dos miRNAs (Naqvi *et al.*, 2012). Diferente dos animais, a maioria dos genes de miRNAs de plantas são monocistrônicos, embora algumas famílias sejam transcritas como unidades policistrônicas, como os *clusters* das famílias MIR156, MIR166 e MIR395 (Wang *et al*., 2007; Zhang *et al*., 2009).

A transcrição dos genes de miRNAs é feita por uma RNA polimerase do tipo II, originando longos miRNAs primários (pri-miRNAs). Os pri-miRNAs são comparáveis aos genes codificadores de proteínas em tamanho e também pela adição de *caps* e caudas poli(A). Uma enzima RNAse tipo III, DICER-LIKE 1 (DCL1), cliva o pri-miRNA, liberando o precursor (pré-miRNA) em forma de grampo que contém dois braços fortemente pareados (miRNA em um braço pareado com o miRNA* no outro braço). Em plantas, ainda no núcleo, o pré-miRNA é clivado novamente por DCL, em conjunto com as proteínas acessórias SERRATE (SE) e HYPONASTIC LEAVES1 (HYL1), liberando o duplex miRNA:miRNA* (Kidner e Martienssen, 2005; Zhang *et al*., 2006; Voinnet, 2009) (Figura 2). Esta clivagem pode ocorrer em duas direções: da base para a alça ("*stem-to-loop*"), como a maioria dos pré-miRNAs de plantas; ou na orientação inversa, da alça para a base ("*loop-to-base*"), como os longos precursores de MIR319 e MIR159 (Li *et al*. 2011).

O duplex miRNA:miRNA* sofre metilação em ambos os terminais 3', pela metiltransferase HEN1 e depois é exportado para o citoplasma pela proteína HASTY (HST), homóloga à exportina de animais. Uma das fitas (fita guia) é preferencialmente incorporada por uma proteína Argonauta (AGO) ao complexo RISC (Figura 2) (Kidner e Martienssen, 2005; Zhang *et al*., 2006; Voinnet, 2009). Em plantas, há muitos membros da família AGO envolvidos em diferentes vias de biogênese de sRNAs. Foi demonstrado que a incorporação do miRNA em complexos RISC contendo AGOs específicas depende do nucleotídeo da extremidade 5' do miRNA maduro (Mi *et al*., 2008) e que ocorre competição entre miRNAs e outros sRNAs para o carregamento em RISC (Meng *et al.*, 2011).

O complexo miRISC é guiado ao mRNA alvo por complementaridade de sequência com o miRNA, causando o silenciamento gênico pós-transcricional (PTGS) por clivagem ou bloqueio da tradução do mRNA (Figura 2). A clivagem é realizada pelas proteínas AGO na posição oposta aos nucleotídeos 10 e 11 da extremidade 5' do miRNA guia (Wollmann e Weigel, 2010). Uma única molécula de miRNA pode mediar o silenciamento de vários transcritos do gene alvo acoplado ao miRISC. Para um alvo ser efetivamente regulado por um miRNA, é necessária sua coexpressão temporal e espacial. Além disso, quando muitos alvos estão presentes, alguns são favorecidos em detrimento de outros,

devido às diferenças de força com que o miRNA se liga a cada um deles (Naqvi *et al*., 2012). Mais indiretamente, a clivagem guiada por alguns miRNAs sobre transcritos específicos (ex.: TAS) dá origem a pequenos RNAs de interferência que regulam a expressão de outros genes (ta-siRNAs) (Allen e Howell, 2010).

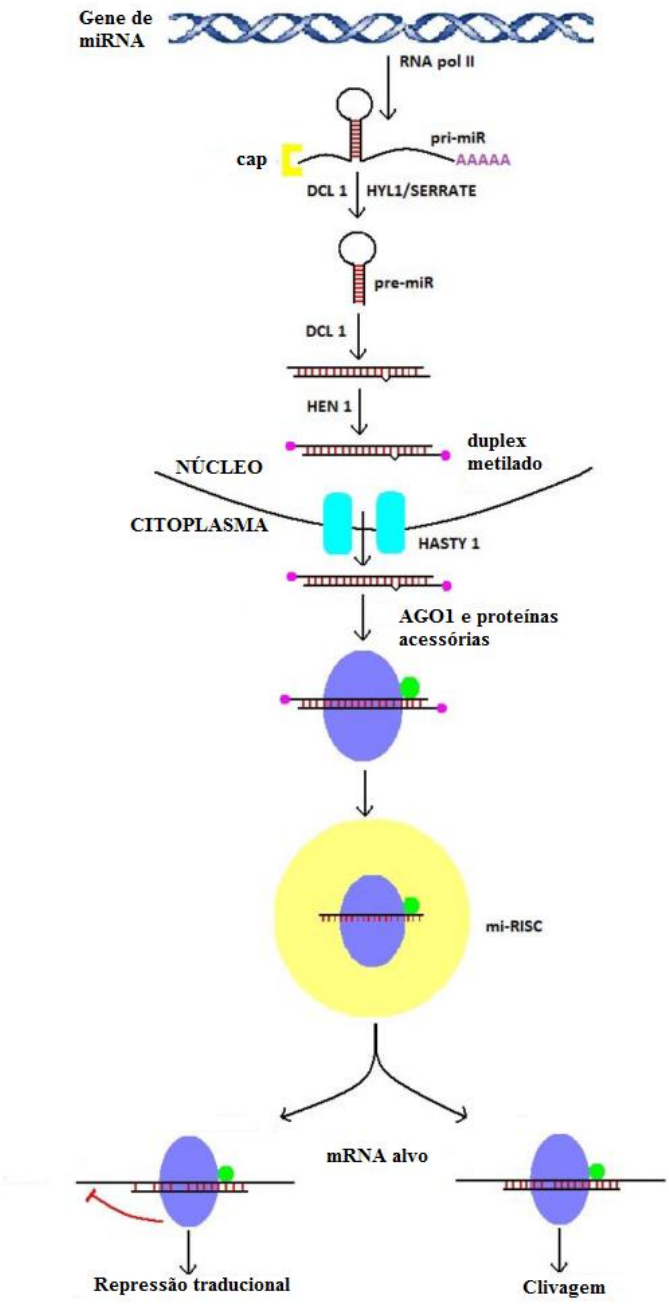


Figura 2. Resumo da biogênese e função de miRNAs em plantas (adaptado de Voinnet, 2009).

Os miRNAs estão envolvidos na regulação do metabolismo, resposta a hormônios, estresses bióticos e abióticos, diferenciação celular e desenvolvimento (revisado por Garcia, 2008 e Lima *et al*.). A função específica de um miRNA está relacionada à função de seus genes alvos que, em sua maioria, codificam para fatores de transcrição ou reguladores de crescimento (Meng *et al.*, 2011). Sua função reguladora se dá por diversos mecanismos: (i) restrição espacial dos alvos, na qual miRNA e transcritos alvos são mutuamente exclusivos; (ii) modulação dos níveis de expressão dos alvos, na qual miRNA e seus transcritos alvos compartilham o mesmo domínio espacial de expressão, pois o miRNA não suprime totalmente a expressão de seus alvos; esta modulação é muito importante nos casos em que a expressão do gene alvo deve ser finamente regulada por seu produto ser requerido em níveis específicos para um funcionamento correto; (iii) regulação temporal do acúmulo de alvos, no qual ocorrem gradientes inversos entre o acúmulo de miRNA e seus alvos com o passar do tempo, até um limite que induz a uma transição no desenvolvimento (Nag e Jack, 2010).

A meia vida dos miRNAs maduros é controlada por vários mecanismos que regulam diretamente sua estabilidade, como metilação, adenilação e uridilação, mas as consequências destas modificações parecem não ser sempre as mesmas, dependendo do contexto em que ocorrem. Por exemplo, a uridilação era considerada um mecanismo de desestabilização, mas análises *in vitro* demonstraram que miRNAs com adição de resíduos de uracil nos seus terminais 3’ não metilados foram protegidos contra a degradação pela nuclease de sRNAs de plantas SDN1 (Small RNA Degrading Nuclease 1) (revisado por Kai e Pasquinelli, 2010). Por fim, as atividades regulatórias dos miRNAs são altamente dinâmicas e afetadas por numerosos fatores desde a transcrição, o processamento do precursor, o carregamento em RISC, até o reconhecimento e regulação do mRNA alvo, e finalmente à degradação do miRNA e sua reciclagem (Meng *et al.*, 2011).

1.5 Papel dos microRNAs no florescimento e reprodução

MicroRNAs são reconhecidos como importantes reguladores do desenvolvimento e seu papel no florescimento tem sido estudado em *A. thaliana* e em algumas outras

espécies, como milho, petúnia e *Antirrhinum* (revisado por Nag e Jack, 2010). Mutações em genes envolvidos com a biogênese de miRNAs, como *DCL1*, *HYL1* e *HEN1*, resultam em diversos defeitos no florescimento (revisado por Terzi e Simpson, 2008). Na figura 3 estão ilustrados os miRNAs mais estudados no processo de florescimento de *Arabidopsis*, seus alvos e tecidos em que a regulação ocorre.

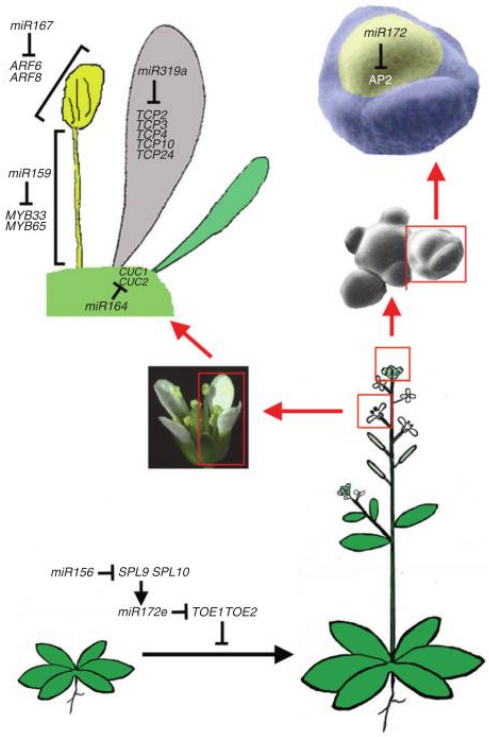


Figura 3. MicroRNAs com funções estudadas no processo de florescimento de *Arabidopsis* (imagem retirada de Nag e Jack, 2010).

O conhecimento sobre miRNAs no desenvolvimento de plantas pode ter diversas aplicações biotecnológicas. O MIR156, por exemplo, é um alvo em potencial para transformação genética visando aumentar a biomassa de culturas para a produção de bioenergia. A superexpressão leve ou moderada deste miRNA aumentou a produção de biomassa de *Panicum virgatum*, gramínea não alimentícia cuja celulose pode ser convertida em biodiesel. Este efeito foi devido ao acréscimo de perfilhos ou ao bloqueio do florescimento, que também causou aumento da produção de açúcares solubilizados e digestabilidade da forragem (Fu *et al*., 2012).

Em *Arabidopsis*, MIR172 e MIR156 regulam seus alvos temporalmente, tendo papel crítico na mudança da fase juvenil para a adulta e na indução floral (Figura 3) (revisado por Poethig, 2009). Estes miRNAs exibem perfis de expressão opostos: os níveis de MIR156 são altos durante a fase juvenil e declinam à medida em que a planta se desenvolve e se aproxima do florescimento; o contrário ocorre com o MIR172 (Wu *et al*., 2009). Na medida em que a planta se desenvolve e os níveis de MIR156 diminuem, aumentam os níveis de seus alvos SPL9 e SPL10, os quais regulam positivamente a transcrição de *MIR172*. MIR172, então, regula negativamente alguns repressores florais (Figura 3), induzindo o florescimento (revisado por Poethig, 2009). Portanto, a superexpressão de MIR156 atrasa o florescimento (Wu *et al*., 2009), enquanto sua inibição o induz por pelo menos duas vias: (i) aumento dos níveis de seus alvos SPLs, que ativam diretamente promotores florais, como *AP1* e *LFY* e (ii) regulação indireta resultando em aumento da expressão de MIR172 (Wang *et al*., 2009a).

MIR172 foi o primeiro miRNA de planta em que se verificou a ocorrência de regulação por repressão traducional (Aukerman e Sakai, 2003), embora a clivagem de alguns alvos também tenha sido comprovada (Schwab *et al*., 2005; Jung *et al*., 2007; Grant-Downton *et al*., 2009). Os alvos do MIR172 são os fatores de transcrição MADS-box *AP2*, *TOE1*, *TOE2*, *SCHNARCHZAPFEN* (*SNZ*) e *SCHLAFMÜETZE* (*SMZ*). Em *Arabidopsis*, MIR172 se acumula em dias longos (indutivos para esta espécie), diminuindo os níveis de seus alvos, que são repressores de *FLOWERING LOCUS T* (*FT*), um integrador floral. Portanto, MIR172 e seus alvos constituem uma rota única que induz o florescimento em resposta ao fotoperíodo pelo aumento da expressão de *FT* (Jung *et al*., 2007). Já na planta de dias curtos *Ipomoea nil*, o gene *InAP2-like*, ao contrário de seu homólogo *TOE1* em *Arabidopsis*, é um regulador positivo do florescimento. Assim, nesta espécie, é a diminuição dos níveis de MIR172 em dias curtos, e consequente aumento de seus alvos, que promove o florescimento (Glazińska *et al*., 2009).

MIR172, além de regular o processo de florescimento, também está envolvido na identidade de órgão floral pela regulação de *AP2* (gene da classe A), em *Arabidopsis*. Depois da indução floral, quando *MIR172* é expresso em todo o primórdio floral, seu acúmulo ocorre nos dois verticilos internos, restringindo *AP2* aos dois verticilos externos, onde o perianto é formado (Figura 3). A superexpressão de *MIR172* com um promotor

constitutivo (35S) causa a formação de carpelos no lugar do perianto, como em mutantes *ap2* (Chen, 2004), devido à ausência de AP2 nos dois verticilos externos, que permite a expansão da expressão de *AG* (gene da classe C) nestes verticilos. Assim, o módulo MIR172/AP2 é um mecanismo adicional de restrição da função A, independentemente do antagonismo com C (Zhao *et al*., 2007). Em *Antirrhinum* e *Petunia*, que não apresentam antagonismo entre as funções A e C, é MIR169 que reprime a atividade da função C no perianto ao regular a expressão de fatores de transcrição da família HAP2/NF-YA, necessários para a transcrição de genes de classe C (revisado por Nag e Jack, 2010).

MIR319a controla o tamanho e a forma dos órgãos florais em *Arabidopsis*, pela regulação dos fatores de transcrição TCP, reguladores negativos do crescimento celular e/ou promotores da diferenciação celular. Mutantes com perda de função de MIR319a apresentam pétalas e estames de tamanho reduzido (Nag *et al*., 2009). MIR166 regula a polaridade dos órgãos (inclusive dos órgãos florais) e a atividade do meristema apical caulinar e do meristema floral. Seus alvos são cinco membros da classe III dos fatores de transcrição HD-ZIP: REV (REVOLUTA), PHB (PHABULOSA), PHV (PHAVOLUTA), CNA (CORONA) e ATHB8. Plantas transgênicas de *Arabidopsis* superexpressando MIR166a apresentaram flores morfologicamente normais, mas de tamanho reduzido (Jung e Park, 2007).

Os alvos de MIR160 e MIR167 são fatores de resposta à auxina (ARFs), cuja sinalização é essencial para o desenvolvimento reprodutivo (Figura 3) (Liu *et al*., 2010). MIR159 modula a expressão de GAMYB, fatores de transcrição que regulam genes de resposta à giberelina (GA). Em *Arabidopsis*, os genes *GAMYB* estão envolvidos na indução do florescimento mediada por GA em dias curtos (pela ativação de *LFY*) e no desenvolvimento de órgãos florais (Achard *et al*., 2004). O MIR164c foi relacionado à formação dos limites entre os primórdios dos órgãos (Laufs *et al*., 2004) e à regulação do número de pétalas, pela caracterização do mutante *early extra petals1* (*eep1*), que apresenta aumento do número de pétalas diretamente correlacionado aos níveis de acúmulo de seu alvo *CUC1*, fator de transcrição da família NAC (Baker *et al*., 2005).

Além disso, miRNAs possibilitam o padrão embrionário correto por prevenir a expressão precoce de fatores de transcrição que promovem a diferenciação (incluindo SPL10, SPL11, ARF17, CNA, PHB, PHV e TCP4) (Nodine e Bartel, 2010). Mutantes

nulos *dcl1* com defeitos no desenvolvimento embrionário em *A. thaliana*, superexpressam ~50 alvos de miRNAs, demonstrando seu papel no padrão embrionário (Nodine e Bartel, 2010). Interações específicas miRNA/alvo também são requeridas para a correta formação do cotilédone durante o desenvolvimento do embrião (Laufs *et al*., 2004; Mallory *et al*., 2004).

Muitos dos módulos de regulação miRNA/mRNA alvo descritos na literatura como importantes no processo reprodutivo de plantas modelo foram detectados em degradoma de semente de soja (Song *et al*., 2011): MIR156/SBP (13 transcritos alvos), MIR159/MYB (3), MIR160/ARF (12), MIR164/NAC (10), MIR166/HD-ZIP (12), MIR167/ARF (6), MIR169/NF-Y (7), MIR172/AP2 (10), MIR319/TCP (8). Dos miRNAs com função conhecida na reprodução, todos já foram identificados em soja, no entanto nenhum foi estudado nos diferentes órgãos florais separadamente.

1.6 Metodologias de identificação de microRNAs

Para a descoberta de novos miRNAs, a aplicabilidade da genética clássica é limitada devido à redundância funcional que restringe a eficiência do silenciamento gênico. Assim, as principais estratégias empregadas na descoberta de miRNAs são: (i) bioinformática, baseada na busca por homologia com miRNAs previamente identificados em outras espécies; (ii) métodos experimentais, como clonagem e sequenciamento de bibliotecas de sRNAs. Vantagens e desvantagens de cada método e os softwares mais comumente usados para a identificação de miRNAs e predição de seu alvos foram previamente discutidos (Unver *et al*., 2009).

A construção e seqüenciamento de bibliotecas de sRNAs e a análise computacional destas seqüências tem sido um método eficiente na identificação de miRNAs. As tecnologias de sequenciamento *high throughput* (HTS), em combinação com sequências genômicas completas, permite a descoberta de novos miRNAs sem qualquer conhecimento prévio sobre o lócus de onde eles se originam (Wollmann e Weigel, 2010). Para isso, analisa-se a estrutura secundária de seqüências genômicas ou de sequências expressas

(ESTs) que contenham a seqüência do miRNA candidato. São considerados miRNAs, os candidatos que se localizem em um dos braços de uma sequência precursora que se dobre em uma estrutura secundária em forma de grampo. Pode-se também identificar miRNAs na ausência de seqüências genômicas detectando dúplex miRNA:miRNA* nas bibliotecas (Subramanian *et al*., 2008).

Em soja, foram identificados miRNAs conservados por métodos computacionais de análise de sequências genômicas e ESTs (Zhang *et al*., 2005; Zhang *et al*., 2008; Sunkar e Jagadeeswaran, 2008). Os primeiros miRNAs de soja validados experimentalmente foram extraídos de raízes recém inoculadas com *Bradyrhizobium japonicum* para investigar seu papel na simbiose com este microrganismo (Subramanian *et al*., 2008). Posteriormente, este mesmo grupo demonstrou que o aumento dos níveis de MIR482, MIR1512 e MIR1515 levam a um significante aumento no número de nódulos (Li *et al*., 2010). Foram também clonados e sequenciados miRNAs de nódulos radiculares fixadores de nitrogênio, em estádio mais avançado da simbiose legume-rizóbio (Wang *et al*. 2009b). A identificação de nove miRNAs novos em soja selvagem (*G. soja*) também foi reportada (Chen *et al*., 2009).

Foram identificados 87 miRNAs novos a partir de sequenciamento de bibliotecas de sRNAs de raízes, sementes, flores e nódulos de soja (Joshi *et al*., 2010). Em meristema apical caulinar de soja, muitos miRNAs* foram mais expressos que seus miRNAs anotados, sugerindo que eles também possuem função biológica (Wong *et al*., 2011). Recentemente, foram identificados 26 novos miRNAs em sementes de soja em desenvolvimento e seus alvos foram identificados em larga escala por análise de degradoma e alguns, por 5'RACE (Song *et al*., 2011).

Em trabalho anterior do nosso grupo, detectamos 256 miRNAs em cultivares de soja sensível e tolerante à seca e soja suscetível e resistente à ferrugem asiática. Estes miRNAs foram classificados em 28 famílias conhecidas de miRNAs e 24 famílias ainda não reportadas (Kulcheski *et al*., 2011). Posteriormente, outro trabalho de caracterização de miRNAs associados a estresse em soja identificou 50 novos miRNAs (Li *et al*., 2011). Uma estratégia similar à empregada neste estudo (identificação de miRNAs por análises *de novo*) identificou 166 miRNAs anotados e 40 miRNAs novos obtidos de 45 precursores (Zhai *et al*., 2011).

As sequências obtidas nestes estudos foram depositadas no miRBase (Griffiths-Jones *et al.*, 2008), que é um banco de dados online para sequência, nomenclatura e anotação de miRNAs. Atualmente (versão 18.0), este banco contém 18.226 sequências de pre-miRNAs, expressando 21.643 miRNAs maduros, em 168 espécies. Do Release 17.0 para o 18.0, houve acréscimo de 1.488 novos precursores e 1.929 novos miRNAs maduros. Para *Glycine max*, existem 362 precursores e 395 miRNAs maduros depositados. Como este banco continua crescendo a cada atualização, a identificação de miRNAs ainda não está saturada, portanto, seu número em soja pode ser aumentado com a descoberta de novos miRNAs em tecidos e condições distintas dos analisados previamente.

2 OBJETIVOS

2.1 Objetivo Geral

Esta dissertação visou identificar os microRNAs expressos em tecidos florais de *Glycine max* e avaliar seus padrões de expressão e possíveis funções nos diferentes órgãos florais desta espécie.

2.2 Objetivos Específicos

- Sequenciar, através de HTS, quatro bibliotecas de sRNAs de flores, carpelos, estames e pétalas de *G. max*.
- Identificar miRNAs presentes nas bibliotecas de sRNAs de tecidos florais de *G. max*, incluindo o descobrimento de novos miRNAs nesta espécie.
- Avaliar os padrões de expressão de miRNAs identificados nos diferentes órgãos florais pela frequência de sequências nas bibliotecas de sRNAs e por RT-qPCR.
- Identificar os possíveis genes alvos de miRNAs diferencialmente expressos.

3 MATERIAL E MÉTODOS

3.1 Material vegetal

Amostras de carpelos, estames, pétalas e flores inteiras (excluindo-se as sépalas) foram coletadas de plantas de soja da cultivar Urano, plantadas em novembro de 2009 no campo experimental da Universidade de Passo Fundo (UPF). Esta cultivar é de hábito determinado, precoce, com ciclo de aproximadamente 132 dias e suas flores são de cor roxa. A coleta foi realizada no dia 18 de janeiro de 2010, quando as plantas se encontravam no fim do estádio de desenvolvimento R2 (pleno florescimento), no qual as flores estavam totalmente desenvolvidas. Oito amostras de cada tipo foram utilizadas em conjunto para a construção de quatro bibliotecas de sRNAs, uma de cada órgão floral (carpelos, estames e pétalas) e uma de flores das quais apenas as sépalas foram removidas.

3.2 Extração de RNA e sequenciamento

Quando coletadas, as amostras foram imediatamente trituradas em Trizol (Invitrogen, CA, USA). RNA total foi isolado usando Trizol e seguindo as instruções do fabricante. O RNA total teve sua qualidade avaliada por eletroforese em gel de agarose 1,0% e foi quantificado usando o fluorímetro Quibit e o kit para ensaio de RNA Quant-iT, conforme as instruções do fabricante (Invitrogen, CA, USA). Aproximadamente 10 µg do RNA total de cada tipo de amostra em conjunto foi enviado à Fasteris SA (Plan-les-Ouates, Suiça) para processamento e sequenciamento usando a tecnologia Solexa, no *Illumina Genome Analyzer GAII*.

Basicamente, o processamento das amostras para a produção de bibliotecas de sRNAs por Illumina consiste nos seguintes passos: I) purificação em gel de acrilamida das bandas de RNA correspondentes à faixa de tamanho de 18 a 30 nt; II) ligação de adaptadores nos terminais 3’ e 5’ do RNA em duas etapas, cada uma seguida de purificação em gel de acrilamida; III) síntese de DNA complementar (cDNA); IV) amplificação por PCR para gerar colônias de DNA molde para o sequenciamento por Illumina. Após o sequenciamento, as sequências dos adaptadores foram removidas e sequências de tamanhos entre 18 e 26 nt foram utilizadas nas análises subsequentes.

3.3 Análise e filtragem das sequências obtidas por HTS

Foram analisadas tanto a distribuição de sequências por tamanho quanto a presença de sequências de miRNAs maduros conhecidos em cada biblioteca. Os miRNAs conhecidos foram identificados pelo ancoramento com as 330 sequências não redundantes de miRNAs maduros de soja anotados no miRBase versão 18.0, não sendo permitidos pareamentos imperfeitos (*mismatches*) ou incompletos (*overhangs*).

Para as demais análises, as sequências de baixa complexidade (menos que 3 bases diferentes, como "AGAGAGAGAGAGA"), sequências correspondentes a RNAs ribossomais (rRNAs) ou a RNAs transportadores (tRNAs) de plantas e sequências que não estão presentes no genoma da soja foram removidas, usando a versão para plantas da ferramenta de filtro do "UEA sRNA toolkit", disponível *online* (UEA sRNA tools; Moxon *et al*., 2008).

3.4 Identificação de microRNAs por análises *de novo*

A ferramenta miRCat (UEA sRNA tools; Moxon *et al*., 2008), disponível *online*, foi usada para a predição de candidatos a precursores de miRNAs. As sequências filtradas das quatro bibliotecas combinadas foram submetidas ao programa (parâmetros padrão, exceto pelo tamanho mínimo de *hairpin*, alterado para 54 nt), que primeiro ancora cada sequência no genoma da soja (DOE JGI, v1.01), requerendo complementaridade perfeita. Após, sequências estendidas tanto acima (*upstream*) quanto abaixo (*downstream*) da região de complementaridade capazes de formar estruturas secundárias em forma de grampo (*hairpin*) são consideradas candidatas a precursores de miRNAs.

Para identificar os verdadeiros precursores de miRNAs em tecidos florais de soja, as sequências de sRNAs de cada biblioteca foram mapeadas nos precursores candidatos utilizando o programa SOAP (Li *et al*., 2008), para análise de seu padrão de distribuição no mapeamento. Foram considerados precursores autênticos aqueles que apresentaram um ou mais blocos definidos de no mínimo 10 sequências de sRNAs mapeadas na mesma orientação, em pelo menos uma das bibliotecas. O número total de sequências do bloco foi

obtido pelo somatório da sequência representativa (mais abundante) e de outras sequências que se sobrepõe a ela, com no máximo 3 nt de *overhangs* em cada extremidade.

A partir dos precursores identificados como verdadeiros, a sequência mais abundante de cada bloco, repetida no mínimo 10 vezes no somatório das bibliotecas, foi eleita como miRNA maduro representativo daquele lócus. Os miRNAs maduros com sequências idênticas foram agrupados e as sequências não redundantes resultantes foram comparadas às sequências de precursores e de miRNAs maduros depositadas no miRBase (v.18.0) por blastn para serem classificados e nomeados. Foi calculada a frequência normalizada em transcritos por milhão (TPM) de cada miRNA, da seguinte maneira:
TPM = nº de sequências repetidas do miRNA/ nº total de sequências da biblioteca x 1.000.000, considerando-se como o total de sequências aquelas de 18 a 26 nt filtradas.

3.5 Análise de expressão dos microRNAs identificados pela frequência de sequências

Para a análise da expressão dos miRNAs entre as diferentes bibliotecas, foi usado o pacote do programa estatístico R denominado "DEGseq" (Wang *et al.*, 2010). Esse programa modela os dados de sequenciamento como a um processo de amostragem aleatória; portanto, o número de vezes que cada miRNA foi sequenciado segue uma distribuição binomial. Baseado neste modelo estatístico é feito um MA-plot, comumente usado nas análises de microarranjos, e é aplicado o teste exato de Fisher para identificar miRNAs diferencialmente expressos.

O número de sequências repetidas de um miRNA foi utilizado nas comparações par a par entre as bibliotecas dos órgãos florais: carpelos (CAR), estames (EST) e pétalas (PET), da seguinte maneira: EST x PET, CAR x EST e CAR x PET. Em seguida, o $\log_2$ do "*fold-change*", o valor de P e o valor de q foram computados. Os miRNAs que apresentaram expressão diferencial significativa (P<0,0001) em pelo menos uma das comparações foram representados em *heatmaps*.

3.6 Reações de PCR quantitativa (RT-qPCR)

Um subconjunto de 61 miRNAs diferencialmente expressos nas bibliotecas de sequências de sRNAs foi analisado por RT-qPCR. A amplificação de cada miRNA foi feita

utilizando três oligonucleotídeos: oligonucleotídeo *stem-loop* reverso individual para a transcrição reversa, oligonucleotídeo direto individual e oligonucleotídeo reverso universal para as reações de PCR em tempo real. O oligonucleotídeo *stem-loop* usado na síntese de cDNA foi desenhado de acordo com Chen *et al*. (2005), consistindo em 44 nt fixos que formam uma haste (dois braços pareados) e uma alça, e seis nucleotídeos no terminal 3' complementares aos seis nucleotídeos do terminal 3' do miRNA correspondente (5' GTCGTATCCAGTGCAGGGTCCGAGGTATTCGCACTGGATACGACNNNNNN 3'). O oligonucleotídeo direto específico, em geral, é idêntico à sequência completa do miRNA, e o oligonucleotídeo reverso universal (5' GTGCAGGGTCCGAGGT 3') usado em todas as reações de qPCR foi baseado na sequência fixa do oligonucleotídeo *stem-loop* (Chen *et al*., 2005).

Os oligonucleotídeo *stem-loop* de todos os miRNAs a serem testados foram agrupados em uma mesma reação, para a síntese de cDNA multiplex, que ocorreu em duas etapas: I) anelamento dos oligonucleotídeos *stem-loop* com 1 µL de RNA total (entre 500 e 100 ng) a 70^{o}C por 5 min; II) síntese da primeira fita de cDNA com a transcriptase reversa RNAse H de M-MLV, nas condições de 16^{o}C por 30 min, 42^{o}C por 30 min, 85^{o}C por 5 min e 10^{o}C por 10 min.

Foram utilizados 63 oligonucleotídeos diretos para a amplificação do conjunto de miRNAs selecionados e dos normalizadores. As reações de RT-qPCR foram realizadas em um aparelho de PCR em tempo real CFX384 Touch™ Real-Time (Bio-Rad) usando SYBR Green I (Invitrogen) para detectar a síntese do produto de amplificação. As reações foram feitas em um volume final de 10 µL contendo 5 µL de cDNA diluído (1:300), 0.1x SYBR Green I (Invitrogen), 3 mM de $MgCl_2$, 25 µM de dNTP, Tampão de PCR 1x, 0.25 U de DNA Polimerase Platinum Taq (Invitrogen) e 0.2 µM de cada oligonucleotídeo, direto e reverso.

Das mesmas amostras de RNA total utilizadas em conjunto para a construção das bibliotecas para o sequenciamento, quatro foram usadas nas reações de RT-qPCR, analisadas na forma de quadruplicatas biológicas. Em cada uma das placas de 384 poços, foram incluídas triplicatas técnicas e controles negativos. Para normalizar os dados, foram usados os Cts obtidos pela amplificação com os oligonucleotídeos diretos GmiR162, baseado na sequência do miRNA gma-MIR162 (5'- TCGATAAACCTCTGCATCCA - 3') e GmiR169, baseado na sequência do miRNA gma-MIR169a (5'-

CAGCCAAGGATGACTTGCCGG - 3'), os quais foram os mais estáveis entre estas amostras, de acordo com análises do programa NormFinder (Andersen *et al*., 2004).

As condições de amplificação foram as seguintes: uma etapa inicial de ativação da polimerase por 5 minutos a 95°C, 40 ciclos para 15 segundos a 95°C (desnaturação), 10 segundos a 60°C (anelamento) e 10 segundos a 72°C (elongação). Uma análise de curva de desnaturação foi programada para o fim da corrida de PCR com aumento da temperatura de 65-99°C. As linhas de *threshold* foram determinadas manualmente, em uma região correspondendo à faixa linear da curva de amplificação logarítmica, sendo fixadas em um ΔRn de 200 para todas as reações, usando o programa CFX Manager.

Para calcular a expressão relativa dos miRNAs, foi usado o método $2^{-\Delta\Delta Ct}$ (Livak e Schmittgen, 2001). Os resultados foram submetidos a testes de ANOVA para determinar os miRNAs que variaram significantemente entre os tecidos, além do teste-t de Student para comparar diferenças na expressão par a par. Os parâmetros estabelecidos para este teste estatístico foram: distribuição bicaudal e variâncias desiguais entre as duas amostras. As médias foram consideradas significantemente diferentes quando $P < 0.05$.

3.7 Predição de alvos

Potenciais sítios alvos de miRNAs selecionados foram buscados nas sequências de cDNA preditas dos genes de soja (DOE JGI, v1.01) obtidas no Phytozome (Phytozome v8.0: Glycine max) pelo programa *online* psRNATarget (Dai e Zhao, 2011). Foram preditos como alvos, os transcritos que tiveram quatro ou menos *mismatches* com um determinado miRNA, sendo que o pareamento G:U é contado pelo psRNATarget como 0.5 *mismatch.*

4 RESULTADOS E DISCUSSÃO

4.1 Análise das bibliotecas de pequenos RNAs

Para identificar miRNAs de soja e verificar aqueles que estão relacionados aos diferentes órgãos florais, quatro bibliotecas de sRNAs foram construídas a partir de amostras de carpelos, estames, pétalas e flores. Através do sequenciamento das bibliotecas usando a tecnologia Solexa, foram gerados 3.530.011, 3.275.049, 3.360.601 e 3.392.134 sequências, respectivamente (Tabela 1). Para as análises subsequentes, foram selecionadas somente as sequências de 18 a 26 nt, faixa de tamanho que compreende os miRNAs.

Na análise de distribuição das sequências de sRNAs por tamanho, as de 24 nt e de 21 nt foram as mais abundantes (Figura 4). Este mesmo padrão tem sido observado em dados de HTS de sRNAs em diversas outras plantas, incluindo *Vitis vinifera* (Chen *et al*., 2011), *Carthamus tinctorius* (Li *et al*., 2011), *Citrus trifoliata* (Song *et al*., 2010), *Oryza sativa* (Peng *et al*., 2011), *Populus tomentosa* (Chen *et al*., 2011) e *Medicago truncatula* (Chen *et al*., 2012a; Chen *et al*., 2012b). No entanto, em carpelos, as sequências de 22 nt foram mais abundantes que as de 21 nt (Figura 4).

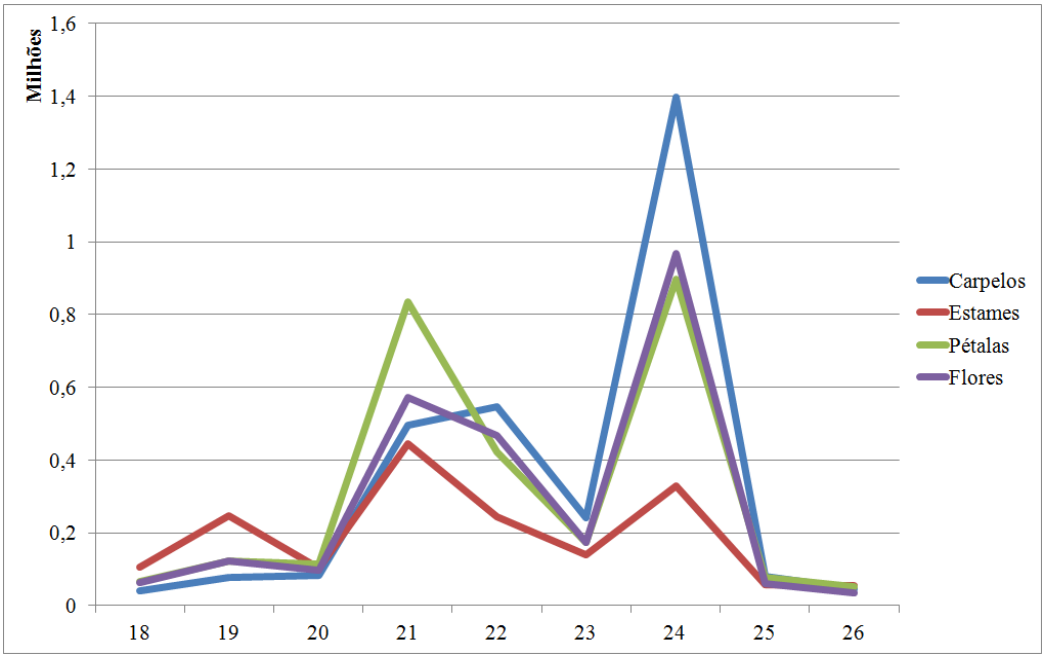


Figura 4. Distribuição por tamanho de sequências de sRNAs das quatro bibliotecas.

Para as análises subsequentes, todas as sequências correspondentes a tRNAs e rRNAs e as que não mapearam perfeitamente no genoma da soja foram removidas, restando um total de 7.678.509 sequências, contendo 3.294.577 sequências distintas (Tabela 1).

Tabela 1. Análise das sequências obtidas pelo sequenciamento das quatro bibliotecas de tecido floral de soja. Os tRNAs e rRNAs foram filtrados antes do mapeamento no genoma da soja.

	Carpelos	**Estames**	**Pétalas**	**Flores**	**Total**
	Nº total de sequências				
Total	3.530.011	3.275.049	3.360.601	3.392.134	13.557.795
18-26nt	2.976.367	1.704.711	2.728.961	2.533.416	9.943.455
t/rRNAs removidos	136.510	349.966	288.874	199.321	974.671
Mapeadas no genoma de soja*	2.466.470	1.130.917	2.076.323	2.004.799	7.678.509
	Nº sequências não redundantes				
18-26nt	1.671.684	554.189	1.235.627	1.337.710	4.020.267
t/rRNAs removidos	11.976	13.644	17.359	14.606	25.959
Mapeadas no genoma de soja	1.391.577	436.773	1.006.501	1.105.097	3.294.577

*O número de sequências de 18 a 26 nt filtradas mapeadas no genoma da soja para cada biblioteca foi usado para calcular a frequência normalizada dos miRNAs identificados.

A presença dos miRNAs conhecidos em cada biblioteca foi verificada pelo ancoramento perfeito das sequências com os miRNAs maduros de soja anotados no miRBase (v.18.0). No total, 205 miRNAs conhecidos foram detectados, representando 62,12% das 330 sequências não redundantes de miRNAs de soja depositadas neste banco de dados (Tabela 2). Pode haver diversos motivos para que miRNAs conhecidos de soja não tenham sido sequenciados neste estudo; eles podem ser fracamente expressos nos tecidos amostrados, estar representados por sequências de outros tamanhos, ou não ser miRNAs verdadeiros.

Tabela 2. Análise dos miRNAs conhecidos sequenciados, relativa às 330 sequências não redundantes de miRNAs maduros de soja anotadas no miRBase (v.18).

Biblioteca	Nº miRNAs anotados presentes	% em relação ao total de miRNAs anotados (330)
Carpelos	173	52.42
Estames	135	40.91
Pétalas	164	49.70
Flores	164	49.70
Total	205	62.12

Os 205 miRNAs conhecidos de soja detectados nas bibliotecas representaram 11% das sequências nas quatro bibliotecas (Figura 5A), correspondendo a menos de 1% da diversidade de sequências (Figura 5B). Este padrão é esperado, visto que a biogênese dos miRNAs é conhecida pela precisão da excisão de um sRNA maduro predominante e bem definido (Mohorianu *et al*., 2011), resultando em uma população de sRNAs pouco complexa, ou seja, composta por um pequeno número de sequências muito frequentes, ao contrário da população de siRNAs (Schwach *et al*., 2009). Embora o maior número de miRNAs conhecidos tenha sido detectado nos carpelos (Tabela 2), eles corresponderam a apenas 7% to total de sequências neste tecido, enquanto nas demais bibliotecas os miRNAs conhecidos representaram de 9% (flores) a 15% (pétalas) das sequências (Figura 6).

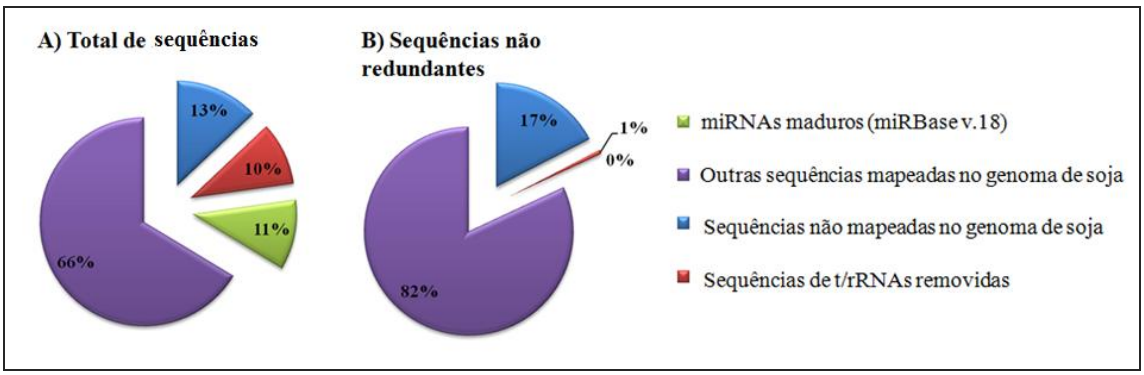


Figura 5. Análise da natureza de origem do total de sequências **(A)** e das sequências não redundantes **(B)** obtidas pelo sequenciamento das quatro bibliotecas de sRNAs de tecidos florais de soja.

Fragmentos de sequências de tRNAs e rRNAs representaram 10% do total de sequências obtidas no sequenciamento (Figura 5A) e aproximadamente 1% das sequências não redundantes (Figura 5B). A biblioteca de estames foi a que apresentou mais sequências

correspondentes a tRNAs e rRNAs, correspondentes a 21% de suas sequências, contra 5%, 11% e 8% nas bibliotecas de carpelos, pétalas e flores, respectivamente (Figura 6).

A porcentagem do total de sequências que não mapearam no genoma da soja foi de 13% (Figura 5A), não diferindo entre as bibliotecas dos tecidos florais (Figura 6). Estas sequências, juntamente com sequências provenientes de outros projetos do Genosoja, foram alvos de um estudo a parte, de metatranscriptômica, que permitiu identificar sequências correspondentes a vírus, fungos e bactérias presentes nas plantas de soja (Molina *et al*, 2012, artigo aceito para publicação no periódico Genetics and Molecular Biology – Anexo 3).

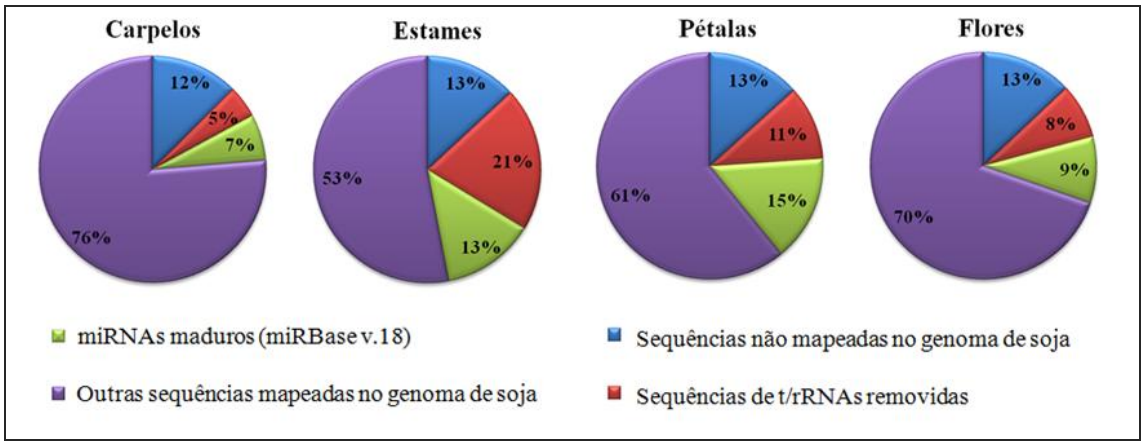


Figura 6. Análise da natureza de origem das sequências de 18 a 26 nt obtidas pelo sequenciamento de cada biblioteca de sRNAs de tecidos florais de soja.

Para esta dissertação, procurou-se identificar quais das sequências correspondentes a miRNAs anotados (Figuras 5 e 6) são de fato miRNAs representativos em tecidos florais de soja e, dentre as outras sequências mapeadas no genoma de soja (Figuras 5 e 6), quais correspondem a miRNAs ainda não anotados.

4.2 Identificação de precursores de microRNAs por análises *de novo*

Para identificar quais dentre os sRNAs presentes nas bibliotecas de tecidos florais de soja correspondem efetivamente a sequências de miRNAs, foi feita uma predição de

precursores de miRNAs por análises *de novo*. As sequências das quatro bibliotecas combinadas foram usadas na busca de sequências genômicas que potencialmente formam estruturas em grampo pela ferramenta *online* miRCat (UEA sRNA tools; Moxon *et al*., 2008) (Figura 7). Como os genomas dos eucariotos contêm milhões de sequências que potencialmente formam estruturas deste tipo (van der Burgt *et al*., 2009), os precursores candidatos foram avaliados quanto ao padrão de mapeamento das sequências de sRNAs das bibliotecas (Figura 7). O padrão de distribuição de sequências em um precursor de miRNA é tomado como uma evidência da biogênese deste miRNA. Assim, para conferir um maior grau de confiabilidade para os miRNAs anotados, o miRBase vem incorporando dados de HTS junto as suas anotações (Kozomara e Griffiths-Jones, 2011). O padrão de blocos de sequências de sRNAs no mapeamento dos precursores reflete a precisão do processamento do miRNA pela enzima DCL, que libera o miRNA maduro de um sítio específico do precursor. O fato de as sequências do bloco estarem na mesma orientação é consistente com a condição de fita simples das moléculas precursoras de miRNAs.

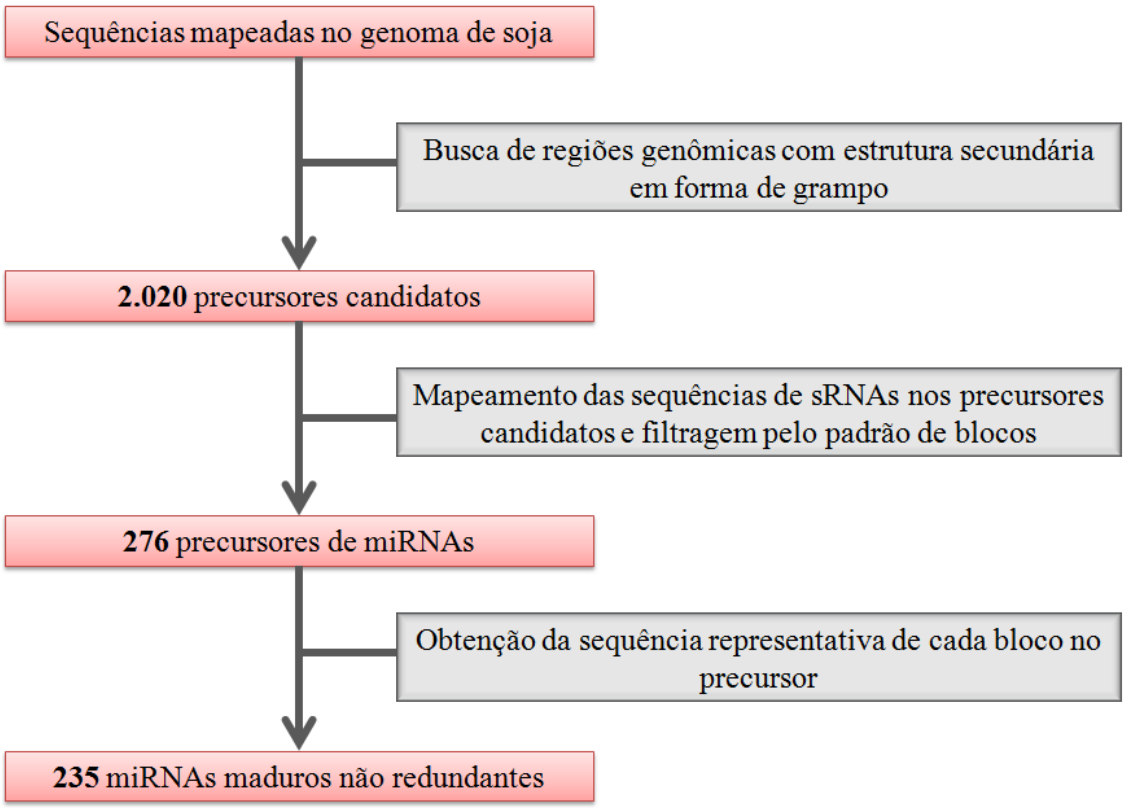


Figura 7. Fluxograma das análises empregadas na identificação consistente de miRNAs nas quatro bibliotecas de sRNAs de tecidos florais de soja.

Dos 276 *loci* de precursores de miRNAs obtidos (Figura 7), 133 correspondem a precursores já anotados no miRBase (v.18.0) (Figura 8A). Quatro deles têm sequências maiores que as anotadas, sendo denominados de precursores conhecidos estendidos (Figura 8A). Como as sequências faltantes nos precursores anotados são importantes por originar miRNAs maduros, será sugerido ao miRBase a reanotação destes precursores com as sequências descritas no Anexo 1.

Pouco mais da metade (143) dos precursores obtidos representam novos *loci* de miRNAs (Figura 8A e Anexo 1). Como esperado, a maioria está situada em regiões consideradas como intergênicas, provavelmente constituindo novos genes de miRNAs, enquanto uma pequena parte está contida em íntrons (Figura 8B). Apenas três se situam em junções de sequências codantes (CDS) com íntrons ou 3'UTR de genes (Figura 8B – CDS). Genes de miRNAs também têm sido encontrados em CDS de genes em outras plantas, mas segundo Nozawa e colaboradores (2012), muitos podem corresponder a erros de anotação, pois no genoma de *Arabidopsis*, que é o mais bem anotado, apenas dois genes de miRNA (1% do total de miRNAs conhecidos para a espécie) foram identificados em CDS.

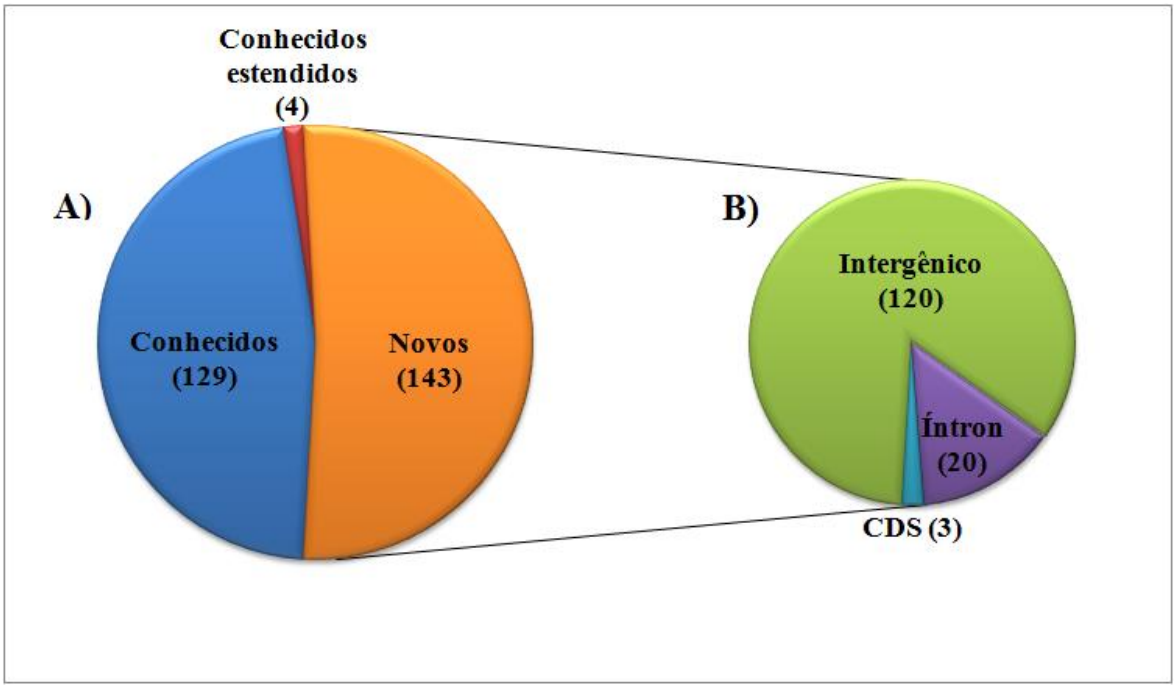


Figura 8. Classificação dos precursores de miRNAs identificados **(A)** e análise da região genômica dos novos precursores **(B)**.

4.3 MicroRNAs maduros identificados em tecidos florais de soja

Para a obtenção dos miRNAs maduros, a sequência representativa de cada bloco de cada precursor foi extraída. Sequências idênticas foram agrupadas, resultando em 235 miRNAs de sequências não redundantes (Figura 7). Este número é menor que o número de precursores, porque, apesar de um mesmo precursor poder originar mais de um miRNA maduro, um único miRNA maduro pode ser codificado em diversos *loci*.

Para a identificação destes miRNAs maduros, suas sequências foram comparadas às depositadas no miRBase por blastn. Assim, os miRNAs foram classificados em seis grupos distintos:

I) 62 miRNAs conhecidos em soja, os quais possuem sequências idênticas às depositadas no miRBase, sendo nomeados como na anotação do miRBase, combinando em um mesmo nome os membros com sequência redundante (ex.: gma-MIR156cde) (Tabelas 3 e 4);

II) 45 isoformas de miRNAs conhecidos de soja, as quais sobrepõem-se aos miRNAs maduros conhecidos quando mapeadas em seus precursores anotados, sendo nomeadas como os miRNAs conhecidos, seguidas por "_iso" (ex.: gma-MIR156bf_iso) (Tabelas 3 e 5);

III) 44 miRNAs novos em precursores conhecidos de soja, os quais se originam nos mesmos precursores, porém em braço oposto ou em fase com os miRNAs anotados; nomeados como os miRNAs conhecidos, com indicação do local em que se originam no precursor, seguido por "_new" (ex.: gma-MIR156a-3p_new) (Tabelas 3 e 6);

IV) 27 novos membros de famílias de miRNAs conhecidas em soja, nomeados pelo código da família (conforme o miRBase), seguido de números correspondentes aos *loci* dos quais eles foram obtidos, separados por vírgula (ex.: MIR156.7,8) (Tabelas 3 e 7);

V) Seis novos miRNAs de famílias já anotadas para outras espécies, mas ainda não para soja, nomeados como os novos membros de famílias de miRNAs

conhecidas em soja, antecedidos por NS- (novo em soja) (ex.: NS-MIR399) (Tabelas 3 e 7);

VI) 51 miRNAs inéditos, nomeados por NF (nova família) seguido de números sequenciais (ex.: NF10-3p) (Tabelas 3 e 8; Anexo 2).

Tabela 3. Classificação dos miRNAs maduros de sequências não redundantes obtidos dos precursores com padrão de mapeamento esperado para precursor de miRNA.

Classe do miRNA maduro	**Tamanho (nt)**								**Total**
	18	**19**	**20**	**21**	**22**	**23**	**24**	**25**	
miRNAs conhecidos em soja (miRBase v.18)	-	-	9	35	12	1	5	-	**62**
Isoformas de miRNAs conhecidos de soja	1	2	6	21	12	-	3	-	**45**
Novos miRNAs em precursores conhecidos de soja	-	-	2	33	6	1	2	-	**44**
Novos membros de famílias de miRNAs conhecidas em soja	-	-	1	22	2	1	1	-	**27**
Novos miRNAs de famílias conservadas em outras espécies, ainda não anotadas em soja	-	-	-	4	1		1	-	**6**
miRNAs inéditos	-	1	-	8	17	4	19	2	**51**
miRNAs identificados	**1**	**3**	**18**	**123**	**50**	**7**	**31**	**2**	**235**

Como esperado, mais da metade (123) dos miRNAs maduros identificados são de 21 nt. Também foram identificados miRNAs de 22 nt (50), de 24 nt (31) e de 20 nt (18). Os outros tamanhos compreenderam apenas de 1 a 7 miRNAs identificados (Tabela 3). Além disso, também como esperado (Mi *et al*. 2008), 118 (~50%) dos miRNAs identificados apresentam uma uridina no terminal 5', enquanto os demais miRNAs apresentam adenina (~20%), guanina (18%) e citosina (12%) nesta posição (Figura 9A). Os miRNAs de 24 nt tiveram tanto adenina quanto uridina na primeira posição 5' e nenhum miRNA de 20 nt foi identificado com citosina nesta posição (Figura 9A).

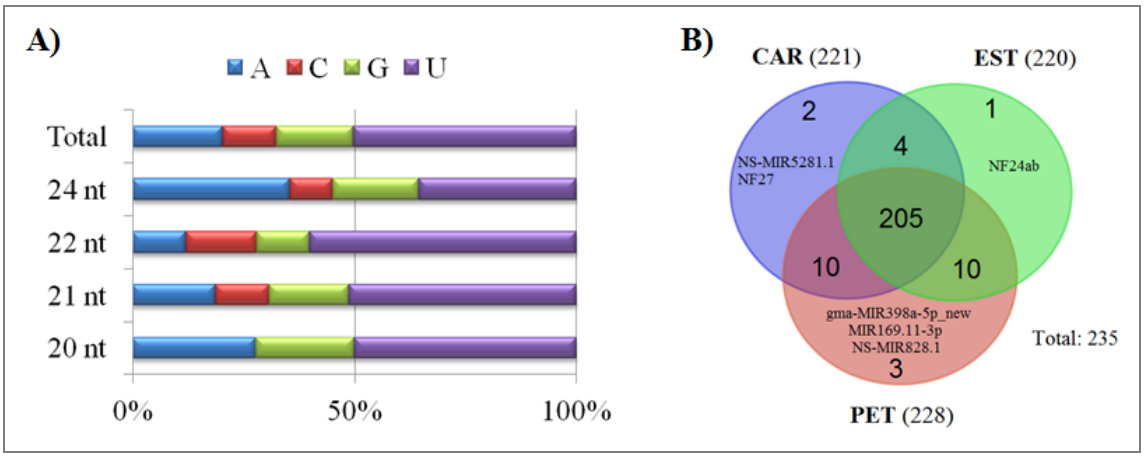


Figura 9: Análise dos miRNAs identificados nas bibliotecas de tecidos florais de soja. (A) Frequência relativa de cada nucleotídeo 5' terminal de todos os miRNAs identificados e daqueles de tamanhos mais frequentes (24 nt, 22 nt, 21 nt e 20 nt). A: adenina; C: citosina; G: guanina; U: uridina. (B) Número de miRNAs identificados nas bibliotecas de sRNAs de carpelos, estames e pétalas de soja.

A maioria (205) dos miRNAs foram detectados nos três órgãos florais, mas alguns foram detectados em bibliotecas de órgãos florais específicos (Figura 9B). Os miRNAs gma-MIR398a-5p_new (Tabela 6), MIR169.11-3p e NS-MIR828.1 (Tabela 7) estiveram presentes apenas nas pétalas (Figura 9B), possivelmente por não estarem envolvidos em processos reprodutivos nas flores de soja. Os miRNAs NS-MIR5281.1 (Tabela 7) e NF27 (Tabela 8) foram detectados em carpelos e NF24 (Tabela 8), em estames. Embora estes três miRNAs também tenham sido detectados na biblioteca de sRNAs de flores inteiras, por apresentarem uma frequência muito baixa, não poderiam ter sido identificados como miRNAs autênticos apenas com o sequenciamento desta biblioteca, mostrando a importância de se analisar bibliotecas independentes e tecidos individuais para aumentar o poder de identificação de miRNAs.

Dos 205 miRNAs maduros conhecidos e detectados nas bibliotecas (Tabela 2), apenas 62 foram confirmados na identificação *de novo* de miRNAs (Tabela 3). Os outros 143, embora tenham sido sequenciados, podem não ter apresentado o padrão de blocos esperado no mapeamento de seus precursores ou não atingiram um número de sequências suficientes para que o padrão pudesse ser detectado, por serem fracamente expressos nos tecidos florais de soja, ou podem ter sido representados por isoformas ou outros miRNAs

maduros presentes em seus precursores. Por exemplo, os miRNAs gma-MIR1512-3p_new, gma-MIR1524-3p.2_new e gma-MIR1531-3p_new foram os únicos miRNAs maduros identificados pelas análises *de novo* nos precursores de gma-MIR1512, gma-MIR1524 e gma-MIR1531, respectivamente (Tabela 6). Em um estudo anterior, gma-MIR1512 também não foi detectado em flores de soja, sendo expresso somente nas raízes (Li *et al*., 2010).

Alguns novos miRNAs originados de precursores conhecidos de soja (Tabela 6) foram mais frequentes que os próprios miRNAs anotados no miRBase identificados, como gma-MIR172cde-5p_new, gma-MIR393-3p_new, gma-MIR398b-5p_new, gma-MIR530-3p_new, gma-MIR2109-3p_new, gma-MIR4345-3p_new e gma-MIR4395-3p_new. Do mesmo modo, algumas isoformas foram mais abundantes que os miRNAs anotados correspondentes. Por exemplo, o gma-MIR167ef_iso, com 22 nt (Tabela 5), foi mais abundante nos órgãos reprodutivos (carpelos e estames) que a forma canônica conhecida de 21nt (Tabela 4) e esta mesma diferença de tamanho foi observada em pólen de *Arabidopsis* para o MIR167d (Borges *et al*., 2011).

O miRNA gma-MIR169i apresentou apenas seis sequências nas bibliotecas combinadas, não sendo identificado como miRNA maduro nas análises *de novo* cujo *cutoff* foi de 10 sequências. No entanto, seu precursor foi estendido (Anexo 1) de forma a conter dois miRNAs com maior frequência, gma-MIR169i-5p.1_new e gma-MIR169i-5p.2_new (Tabela 6). Estes miRNAs são mapeados em fase na sequência do precursor, no entanto, não se pode afirmar que eles são processados em fase, pois um deles (gma-MIR169i-5p.1_new) também pode ser transcrito a partir de um lócus adicional (Tabela 6). Um precursor que provavelmente é processado em fase é o Pre-NF12 (Anexo 1), pois apenas o lócus deste precursor contém as sequências dos miRNAs maduros (NF12-5p.1 e NF12-5p.2) no genoma da soja (Tabela 8). Foi reportada a ocorrência de miRNAs em fase em precursores de nove organismos a partir de dados públicos de sequenciamento de plantas (*Oryza sativa*, *Physcomitrella patens*, *M. truncatula* e *Populus trichocarpa*) e animais (*Homo sapiens*, *Mus musculus*, *C. elegans* e *Drosophila*). Em *Arabidopsis*, arroz e *Medicago*, precursores de MIR169 estavam entre os identificados (Zhang *et al*., 2010). Em *Arabidospsis*, foi demonstrado que a produção destes miRNAs em fase depende da mesma via de biogênese dos miRNAs, que alguns são mais expressos que os miRNAs maduros

anotados e que muitos são funcionais, como sugerido pela ocorrência, no degradoma, de produtos de clivagem nos seus sítios alvos correspondentes (Zhang *et al*., 2010).

Tabela 4. miRNAs conhecidos em soja identificados nas bibliotecas de tecidos florais de soja.

miRNA	Sequência	Tamanho (nt)	Novos *loci*	Região no precursor	Total de sequências	Frequência em TPM[1]			
						Flores	Carpelos	Estames	Pétalas
gma-MIR156a	UGACAGAAGAGAGUGAGCAC	20	Pre-MIR156.1,2,3,4,5	5p	954	170.59067	41.76	127.33	175.79
gma-MIR156cde	UUGACAGAAGAUAGAGAGCAC	21		5p	2374	342.68	54.73	438.58	508.59
gma-MIR156g	ACAGAAGAUAGAGAGCACAG	20		5p	15	2.00	1.22	2.65	2.41
gma-MIR156h	UGACAGAAGAGAGAGAGCAC	20		3p	20	4.99	0.81	2.65	2.41
gma-MIR159a-3p	UUUGGAUUGAAGGGAGCUCUA	21		3p	417844	52945.46	24523.31	55057.98	91000.77
gma-MIR159d	AGCUGCUUAGCUAUGGAUCCC	21		5p	68	7.98	10.54	7.07	8.67
gma-MIR159e-5p	GAGCUCCUUGAAGUCCAAUU	20		5p	419	37.91	49.87	19.45	95.36
gma-MIR160	UGCCUGGCUCCCUGUAUGCCA	21	Pre-MIR160.1,2,3	5p	516	76.82	85.95	84.00	26.49
gma-MIR162	UCGAUAAACCUCUGCAUCCA	20		3p	45	6.48	3.24	9.73	6.26
gma-MIR162bc	UCGAUAAACCUCUGCAUCCAG	21		3p	1689	188.05	139.88	267.04	320.28
gma-MIR164	UGGAGAAGCAGGGCACGUGCA	21	Pre-MIR164.1,2,3,4,5,6,7,8	5p	1909	200.52	227.86	210.45	340.51
gma-MIR166b/a-3p	UCGGACCAGGCUUCAUUCCCC	21	Pre-MIR166.2,3,4,5,6	3p	199142	16206.11	24858.20	52525.52	22124.69
gma-MIR166a-5p	GGAAUGUUGUCUGGCUCGAGG	21	Pre-MIR166.6,7	5p	4782	599.56	896.02	514.63	379.52
gma-MIR166h-3p	UCUCGGACCAGGCUUCAUUCC	21	Pre-MIR166.1	3p	15836	1637.57	1951.37	2505.93	2362.83
gma-MIR167abd	UGAAGCUGCCAGCAUGAUCUA	21	Pre-MIR167.1	5p	3535	421.99	247.72	157.39	915.08
gma-MIR167c	UGAAGCUGCCAGCAUGAUCUG	21		5p	1830	263.87	10.95	81.35	569.28
gma-MIR167ef	UGAAGCUGCCAGCAUGAUCUU	21		5p	7305	872.41	1609.99	1197.26	111.25
gma-MIR167g	UGAAGCUGCCAGCAUGAUCUGA	22		5p	287	31.42	21.49	38.02	61.65
gma-MIR168	UCGCUUGGUGCAGGUCGGGAA	21	Pre-MIR168.1	5p	102358	10369.12	5316.91	16283.25	24100.78
gma-MIR169a	CAGCCAAGGAUGACUUGCCGG	21		5p	48082	4426.38	1670.40	13609.31	9486.48
gma-MIR169b	CAGCCAAGGAUGACUUGCCGA	21		5p	99	8.48	10.14	38.02	6.74
gma-MIR169c	AAGCCAAGGAUGACUUGCCGA	21	Pre-MIR169.1	5p	60	6.98	2.43	9.73	13.97

gma-MIR169d	UGAGCCAAGGAUGACUUGCCGGU	23		5p	223	16.96	16.22	24.76	58.28
gma-MIR169e	AGCCAAGGAUGACUUGCCGG	20	Pre-MIR169.2,3,4,5,6	5p	4363	418.00	134.20	984.16	1002.25
gma-MIR171c	AGAUAUUGGUGCGGUUCAAUC	21	Pre-MIR171.1	5p	10	1.00	1.22	2.65	0.96
gma-MIR171efg	UGAUUGAGCCGUGCCAAUAUC	21	Pre-MIR171.3	3p	1726	154.13	86.36	371.38	377.59
gma-MIR171j	UAUUGGCCUGGUUCACUCAGA	21		5p	2879	277.33	169.47	711.81	529.78
gma-MIR172a/bh-3p	AGAAUCUUGAUGAUGCUGCAU	21		3p	286	36.41	10.95	15.92	80.91
gma-MIR172cde	GGAAUCUUGAUGAUGCUGCAG	21		3p	206	21.45	8.51	118.49	3.85
gma-MIR319d	UGGACUGAAGGGGAGCUCCUUC	22		3p	2557	273.84	426.52	273.23	311.61
gma-MIR319f	UUGGACUGAAGGGGCCUCUU	20	gma-MIR4414 estendido	3p	6971	688.35	1054.54	2273.38	201.80
gma-MIR319hjkm	UUGGACUGAAGGGAGCUCCCU	21		3p	29199	2644.65	1585.67	9142.14	4646.19
gma-MIR390a-5p	AAGCUCAGGAGGGAUAGCGCC	21	Pre-MIR390.1,2	5p	5000	626.50	268.40	1588.98	618.88
gma-MIR390b	AAGCUCAGGAGGGAUAGCACC	21	Pre-MIR390.4	5p	191	19.95	4.05	110.53	7.71
gma-MIR394c-5p	UUGGCAUUCUGUCCACCUCC	20	Pre-MIR394.1,2,3,4,5	5p	673	92.28	30.81	88.42	150.27
gma-MIR396b-3p	GCUCAAGAAAGCUGUGGGAGA	21		3p	993	149.14	67.71	59.24	221.55
gma-MIR396c/b-5p	UUCCACAGCUUUCUUGAACUU	21		5p	14704	1537.81	242.86	2024.02	4205.99
gma-MIR396d	AAGAAAGCUGUGGGAGAAUAUGGC	24		3p	14	2.99	0.41	2.65	1.93
gma-MIR396i-3p	GUUCAAUAAAGCUGUGGGAAG	21		3p	4922	743.72	45.41	580.94	1282.07
gma-MIR398c	UGUGUUCUCAGGUCGCCCCUG	21		3p	42	3.99	8.51	3.54	4.33
gma-MIR1507ab	UCUCAUUCCAUACAUCGUCUGA	22		3p	106993	13487.14	5862.63	21118.26	20040.72
gma-MIR1508c	UAGAAAGGGAAAUAGCAGUUG	21		3p	264	35.42	36.08	43.33	26.49
gma-MIR1510b-3p	UGUUGUUUUACCUAUUCCACC	21		3p	814	137.17	30.41	176.85	127.15
gma-MIR1510b-5p	AGGGAUAGGUAAAACAACUACU	22		5p	3587	465.38	348.68	487.22	598.65
gma-MIR1513	UGAGAGAAAGCCAUGACUUAC	21		5p	190	24.44	8.51	38.91	36.60
gma-MIR1515	UCAUUUUGCGUGCAAUGAUCUG	22		5p	109	15.96	3.65	9.73	27.45
gma-MIR1520kl	AAUCAGAACAUGACACGUGACAGU	24		3p	60	9.48	11.76	2.65	4.33
gma-MIR2119	UCAAAGGGAGUUGUAGGGGAA	21	Pre-MIR2119.1	3p	300	55.87	10.54	77.81	35.64
gma-MIR4355	CACUGUUGUGCUGGGUGUACCA	22		3p	15	2.49	3.65	0.88	0.00
gma-MIR4358	CAGUGCAUGACUAUAUCGCCAG	22		3p	22	1.50	5.68	1.77	1.44

gma-MIR4364a	CGCGAGAUCGCACGGAAGAAGGUU	24		3p	15	0.00	1.62	0.00	5.30
gma-MIR4370	AGUAGACUCGUCCGAUUUUGCGUA	24		5p	14	2.00	2.43	0.88	1.44
gma-MIR4382	UAUGUUAACUGAUUUCAUGGAU	22		3p	144	25.94	9.73	17.68	23.12
gma-MIR4392	UCUGCGAAAAUGUGAUUUCGGA	22	Pre-MIR4392.1	3p	285	49.38	34.06	72.51	9.63
gma-MIR4397-5p	CAUCGUUGACGCUGACUGUACG	22		5p	19	2.49	1.62	0.00	4.82
gma-MIR4407	CAGAGGAAGCAGCACUUGUACC	22		3p	41	4.99	2.03	14.15	4.82
gma-MIR4408	UAACAACAUUGGAUGAGGGUUGGA	24		3p	49	4.99	10.95	6.19	2.41
gma-MIR4412-5p	UGUUGCGGGUAUCUUUGCCUC	21		5p	205	31.92	14.60	30.06	34.20
gma-MIR4414	AGCUGCUGACUCGUUGGCUC	20	gma-MIR4414 estendido	5p	10	1.00	2.03	1.77	0.48
gma-MIR4415b/a-3p	UUGAUUCUCAUCACAACAUGG	21		3p	42	2.99	2.03	1.77	13.97
gma-MIR482bd-3p	UCUUCCCUACACCUCCCAUACC	22		3p	3648	470.87	525.04	562.38	372.29
gma-MIR530b	UGCAUUUGCACCUGCACUUUA	21	Pre-MIR530.1,2	5p	28	5.49	2.03	5.31	2.89

[1]Frequência normalizada em transcritos por milhão: TPM = nº de sequências do miRNA/ nº total de sequências na biblioteca x 1.000.000, sendo o total de sequências, as sequências de 18 a 26 nt filtradas.

Os miRNAs sublinhados foram testados quanto ao padrão de expressão nas amostras de órgãos florais de soja através de RT-qPCR.

Tabela 5. Isoformas de miRNAs conhecidos em soja identificados nas bibliotecas de tecidos florais de soja.

miRNA	Sequência	Tamanho (nt)	Novos *loci*	Região no precursor[1]	Total de sequências	Frequência em TPM[1]			
						Flores	Carpelos	Estames	Pétalas
gma-MIR156bf_iso	UUGACAGAAGAGAGAGAGCAC	21		5p	440	95.27	21.08	103.46	38.53
gma-MIR159bc_iso	AUUGGAGUGAAGGGAGCU	18		3p	18	1.00	1.22	0.88	5.78
gma-MIR166b/a-3p_iso	GGACCAGGCUUCAUUCCCC	19	Pre-MIR166.7	3p	170	7.98	26.35	50.40	15.41
gma-MIR167ef_iso	UGAAGCUGCCAGCAUGAUCUUA	22		5p	15329	1881.98	3735.30	1920.57	82.36
gma-MIR171ai_iso	UUGAGCCGUGCCAAUAUCACGA	22		3p	3252	401.04	296.37	206.03	714.73
gma-MIR171b-5p_iso	CGUGAUAUUGGUACGGCUCAUC	22		5p	254	25.94	5.27	15.92	82.36
gma-MIR172b-5p_iso	GUAGCAUCAUCAAGAUUCACA	21		5p	400	52.87	12.97	18.57	116.07
gma-MIR172cde_iso	GGAAUCUUGAUGAUGCUGCA	20	Pre-MIR172.1	3p	69	3.99	0.41	51.29	0.96
gma-MIR319egm_iso	UGUGCUUGGACUGAAGGGAGC	21		3p	63	5.99	10.14	13.26	5.30
gma-MIR319l_iso	UUUGGACUGAAGGGAGCUCCU	21	Pre-MIR319.1	3p	3486	374.60	682.76	823.23	57.79
gma-MIR390ac-3p_iso	CGCUAUCCAUCCUGAGUUUCA	21	Pre-MIR390.1	3p	913	104.25	28.79	330.71	124.74
gma-MIR390b*_iso	CGCUAUCUAUCUUGAGCUUCA	21		3p	19	0.50	0.00	12.38	1.93
gma-MIR393_iso	UCCAAAGGGAUCGCAUUGAU	20	Pre-MIR393.1,2,3,4,5	5p	58	4.49	1.62	19.45	11.08
gma-MIR394a_iso	AGCUCUGUUGGCUACACUUUG	21		3p	10	0.50	2.84	0.00	0.96
gma-MIR395abc_iso1	UGAAGUGUUUGGGGGAACUCC	21	Pre-MIR395.1,2	3p	110	9.98	2.43	2.65	6.74
gma-MIR395abc_iso2	UGAAGUGUUUGGGGGAACU	19	Pre-MIR395.7,8	3p	60	13.97	8.11	13.26	22.64
gma-MIR395abc_iso3	UGAAGUGUUUGGGGGAACUC	20	Pre-MIR395.9	3p	43	2.99	0.41	44.21	1.44
gma-MIR396ce/abi-5p_iso	UUCCACAGCUUUCUUGAACU	20		5p	2770	279.33	45.00	624.27	670.90
gma-MIR398ab_iso	UUGUGUUCUCAGGUCACCCCU	21		3p	184	20.95	16.62	52.17	20.23
gma-MIR1510a-3p_iso	UGUUGUUUUACCUAUUCCACCC	22		3p	1608	238.93	48.25	558.84	182.05
gma-MIR1510a-5p_iso	AGGGAUAGGUAAAACAAUGAC	21		5p	8727	1251.00	478.82	1332.55	1700.60
gma-MIR1511_iso	AACCAGGCUCUGAUACCAUGG	21		3p	6872	1089.88	562.75	717.12	1198.27
gma-MIR1514a_iso	UUCAUUUUUAAAAUAGGCAUUGGG	24		5p	28	4.49	2.43	2.65	4.82
gma-MIR1515_iso	UCAUUUUGCGUGCAAUGAUCU	21	Pre-MIR1515.1	5p	226	27.43	16.62	30.06	46.24
gma-MIR1516*_iso	UUGGAUACAAGUUAUAAGCUCU	22		5p	89	18.46	12.97	6.19	6.26

gma-MIR2109_iso	UGCGAGUGUCUUCGCCUCUGA	21		5p	3508	403.03	185.69	1036.33	515.33
gma-MIR1523_iso	UAUGGGAUAAAUGUGAGCUCA	21		5p	224	36.41	8.51	21.22	51.05
gma-MIR1527_iso	UAACUCAACCUUACAAAACCGG	22		5p	37	3.99	9.33	0.88	2.41
gma-MIR4345_iso	CUAAGACGGAACUUACAAAGAU	22		5p	140	20.45	19.87	14.15	16.38
gma-MIR4349_iso	GUCCCAUAUUGGCUAGAGAUAAGA	24		5p	24	4.99	1.62	5.31	1.93
gma-MIR4351_iso	UUGGGAUUCAGUUGGAGUUGG	21		5p	53	8.48	7.70	4.42	5.78
gma-MIR4367_iso	UGAACCCUAGCGAAGUAAAUCA	22		3p	47	8.48	6.08	4.42	4.82
gma-MIR4376_iso	UACGCAGGAGAGAUGACGCUG	21		5p	19	3.49	2.84	2.65	0.96
gma-MIR4395_iso	UAUGGGCUUGAGUAAGCUGCU	21		5p	10	1.00	1.62	0.88	1.44
gma-MIR4397_iso	UCCCGUCAGUGUCAAAGAUGUG	22		3p	49	8.48	5.68	6.19	5.30
gma-MIR4399_iso	ACUAACGACAGGUAGUAAAUCGAA	24	gma-MIR4399 estendido	3p	24	2.00	6.08	2.65	0.96
gma-MIR4411_iso	UAUUGUAACUAAUUUGUCGGUA	22		3p	19	4.49	2.43	0.00	1.93
gma-MIR4413_iso	UAAGAGAAUUGUAAGUCACUG	21		5p	196	36.91	7.30	15.92	41.42
gma-MIR4416_iso	UACGGGUCGCUCUCACCUAGG	21		3p	110	14.47	11.76	12.38	18.30
gma-MIR482a-3p_iso	UUCCCAAUUCCGCCCAUUCCUA	22	Pre-MIR482.1	3p	219	29.43	24.33	16.80	39.01
gma-MIR482bd_iso	UAUGGGGGGAUUGGGAAGGA	20		5p	1089	135.18	160.96	126.45	133.89
gma-MIR482c/a-5p_iso	GGAAUGGGCUGAUUGGGAAG	20		5p	262	32.42	14.19	42.44	54.90
gma-MIR482c-5p_iso	GGAAUGGGCUGAUUGGGAAGU	21	Pre-MIR482.1	5p	14864	2492.52	859.53	1311.33	3016.87
gma-MIR5036_iso	AGAGGCCCUUGGGGAGGAGUA	21		3p	11	1.00	0.81	0.00	3.37
gma-MIR5380_iso	UUAAGAAAAUGAAUGAUGAGGA	22	Pre-MIR5380.1,2	3p	64	15.46	8.51	2.65	4.33

[1]Frequência normalizada em transcritos por milhão: TPM = nº de sequências do miRNA/ nº total de sequências na biblioteca x 1.000.000, sendo o total de sequências, as sequências de 18 a 26 nt filtradas.

Os miRNAs sublinhados foram testados quanto ao padrão de expressão nas amostras de órgãos florais de soja através de RT-qPCR.

Tabela 6. Novos miRNAs presentes em sequências de pré-miRNAs já conhecidos de soja e identificados nas bibliotecas de tecidos florais de soja.

miRNA	Sequência	Tamanho (nt)	Novos *loci*	Região no precursor[1]	Total de sequências	Frequência em TPM[1]			
						Flores	Carpelos	Estames	Pétalas
gma-MIR156a-3p_new	GCUCACUUCUCUAUCUGUCAGC	22	Pre-MIR156.1	3p	802	92.78	24.33	232.55	141.11
gma-MIR156d-3p_new	GCUCUCUAUACUUCUGUCAUC	21		3p	54	5.49	0.00	13.26	13.49
gma-MIR156k-3p_new	GCUCACUUCUCUUUCUGUCAAC	22		3p	12	1.50	1.62	1.77	1.44
gma-MIR159d-3p_new	CUUCCAUAUCUGGGGAGCUUC	21		3p	39	5.49	0.00	10.61	7.71
gma-MIR160-3p_new	GCGUAUGAGGAGCCAAGCAUA	21		3p	155	25.44	15.41	24.76	18.30
gma-MIR164-3p_new	CACGUGCUCCCCUUCUCCAAC	21	Pre-MIR164.1,2,3	3p	19	2.49	3.24	1.77	1.93
gma-MIR166df-5p_new	GGAAUGGUGUCUGGUUCGAGA	21		5p	748	86.79	119.60	91.08	84.77
gma-MIR166e-5p_new	GGAAUGUUGGCUGGCUCGAGG	21	Pre-MIR166.2	5p	468	45.39	90.41	62.78	39.97
gma-MIR166g-5p_new	GGAAUGUCGUUUGGUUCGAGA	21		5p	3854	534.72	8.51	164.47	1240.17
gma-MIR167bd-3p_new	GUCAUGCUGUGCUAGCCUCACU	22		3p	41	2.99	4.05	1.77	11.08
gma-MIR167c-3p_new	UCAGGUCAUCUUGCAGCUUCA	21		3p	16	2.99	0.81	1.77	2.89
gma-MIR167g-3p_new	GAUCAUGUGGCUGCUUCACC	20		3p	106	17.46	0.41	12.38	26.97
gma-MIR168-3p_new	CCCGCCUUGCAUCAACUGAAU	21	Pre-MIR168.1	3p	754	93.77	31.22	91.96	185.42
gma-MIR169a-3p_new	GGCAAGUUGUGUUUGGCUAUG	21		3p	205	19.95	1.62	64.55	42.38
gma-MIR169c-3p_new	GGCAGGUCAUCCUCUGGCUAUA	22		3p	27	1.50	1.62	3.54	7.71
gma-MIR169i-5p.1_new	UGAAGGUAGAGAGAGUAGAUU	21	gma-MIR169i estendido, Pre-MIR169.8	5p.1	13	2.49	0.00	3.54	1.93
gma-MIR169i-5p.2_new	UGAGCCGGGAUGGCUUGCCGGCA	23	gma-MIR169i estendido	5p.2	20	0.00	1.22	0.88	7.71
gma-MIR171a-5p_new	GGAUAUUGGUCCGGUUCAAUA	21		5p	89	15.46	14.19	3.54	9.15
gma-MIR171c-3p_new	UUGAGCCGUGCCAAUAUCACA	21	Pre-MIR171.1,2	3p	818	93.28	62.03	225.48	107.40
gma-MIR171f-5p_new	AGAUAUUGGUACGGUUCAAUC	21	Pre-MIR171.2	5p	75	9.98	4.46	23.87	8.19
gma-MIR172cde-5p_new	GGAGCAUCAUCAAGAUUCACA	21		5p	555	56.86	85.55	164.47	21.19
gma-MIR1507a-5p_new	AGAGGUGUAUGGAGUGAGAGA	21		5p	262	37.91	31.22	38.02	31.79

gma-MIR1507b-5p_new	AGAGAUGUAUGGAGUGAGAGA	21		5p	337	58.86	42.57	32.72	37.08
gma-MIR1508ab-5p_new	ACUGCUAUUCCCAUUUCUAAA	21		5p	25	3.99	0.81	2.65	5.78
gma-MIR1511-5p_new	GUGGUAUCAGGUCCUGCUUCA	21		5p	510	49.88	70.95	116.72	49.61
gma-MIR1512-3p_new	UGCUUUAAGAAUUUCAGUUAU	21		3p	17	2.99	1.62	1.77	2.41
gma-MIR1524-3p.2_new	UAGGUUAUUGGAAACAAGUGG	21		3p	21	3.99	2.03	1.77	2.89
gma-MIR1531-5p_new	AUAUGGACGAAGAGAUAGGUA	21		5p	100	21.45	14.19	6.19	7.22
gma-MIR2109-3p_new	GGAGGCGUAGAUACUCACACC	21		3p	4726	526.74	417.60	850.64	808.16
gma-MIR2111-5p_new	UAAUCUGCAUCCUGAGGUUUA	21	Pre-MIR2111.1,2	5p	29	2.00	2.03	4.42	7.22
gma-MIR319ab-5p_new	AGAGCUUUCUUCAGUCCACU	20		5p	49	7.98	2.03	9.73	8.19
gma-MIR319f-5p_new	AGCUGCUGACUCGUUGGUUCG	21	gma-MIR319f estendido	5p	164	20.95	24.73	43.33	5.78
gma-MIR393-3p_new	GAUCAUGCUAUCCCUUUGGAU	21	Pre-MIR393.1,2,3,4,5	3p	172	11.47	8.51	25.64	47.68
gma-MIR396e-3p_new	AUUCAAGAUAGCUGUGGAAAA	21		3p	203	26.44	6.08	10.61	59.24
gma-MIR398a-5p_new	GGAGUGAAUCUGAGAACACAAG	22		5p	43	17.96	0.00	0.00	3.37
gma-MIR398b-5p_new	GAGUGGAUCUGAGAACACAAGG	22		5p	580	108.74	52.30	75.16	71.28
gma-MIR4345-3p_new	CAAUCUUUUUAAGUUUCGUCU	21		3p	10	1.50	0.00	3.54	1.44
gma-MIR4349-3p_new	UUAUCUUUAGCCAAUGUGGGA	21		3p	129	19.45	12.57	21.22	16.86
gma-MIR4367-5p_new	ACUAGGGUUCAGGACAAUAUUCAA	24		5p	36	3.99	6.08	1.77	5.30
gma-MIR4395-3p_new	CAGCAGCUUUCUCGGACCCAUACU	24		3p	33	4.49	7.70	3.54	0.48
gma-MIR4407-5p_new	UAAAGUGUUGCUUCGUCUAAG	21		5p	11	2.00	0.41	4.42	0.48
gma-MIR4416-5p_new	UGGGUGAGAGAAACGCGUAUC	21	Pre-MIR4416.1	5p	17	1.50	0.41	0.00	6.26
gma-MIR5036-5p_new	CUCUCCCUCAAGGGCUUCUCG	21		5p	17	0.50	1.22	1.77	5.30
gma-MIR530-3p_new	AGGUGCAGGUGCAUCUGCAGG	21	Pre-MIR530.1,2	3p	544	65.34	60.00	51.29	99.70

[1]Frequência normalizada em transcritos por milhão: TPM = nº de sequências do miRNA/ nº total de sequências na biblioteca x 1.000.000, sendo o total de sequências, as sequências de 18 a 26 nt filtradas.

Os miRNAs sublinhados foram testados quanto ao padrão de expressão nas amostras de órgãos florais de soja através de RT-qPCR.

Tabela 7. Novos membros de famílias de miRNAs conhecidas em soja ou anotadas apenas em outras espécies que foram identificados nas bibliotecas de tecidos florais de soja.

miRNA	Sequência	Tamanho (nt)	nº *hits* adicionais no genoma	Região no precursor[1]	Total de sequências	Frequência em TPM[1]			
						Flores	Carpelos	Estames	Pétalas
MIR156.6,7	UUGACAGAAGAAAGGGAGCAC	21		5p	54	6.98	7.70	12.38	3.37
MIR164.4,5-3p	CAUGUGCCCCCCUUCCCCAUC	21	1	3p*	10	2.00	0.81	1.77	0.96
MIR164.6-3p	CAUGUGCCCCUCUUCCCCAUC	21		3p*	20	4.99	0.41	0.88	3.85
MIR164.7,8-3p	CACGUGCUCUCCUUUUCCAGC	21		3p*	121	27.93	16.22	19.45	1.44
MIR166.1-5p	AAUGAGGUUUGAUCCAAGAUC	21	1	5p*	15	4.49	0.81	0.88	1.44
MIR166.3-5p	GGAAUGUCAUCUGGUUCGAGA	21		5p*	18	1.00	5.27	0.00	1.44
MIR167.2	UGAAGCUGCCAGCCUGAUCUU	21		5p	294	37.91	0.41	8.84	99.70
MIR169.10	UUGAGCCAAGGAUGACUUGCCGA	23		5p*	17	0.50	2.03	1.77	4.33
MIR169.10-3p	GGCGAGGAAUCUGGGCUCAUU	21		3p*	24	0.50	2.03	1.77	7.71
MIR169.11	CAGCCAAGGAUGACUUGCCGU	21		5p	19	2.99	0.00	4.42	3.85
MIR169.11-3p	UAACCGGCAAGUCAACUCUGGC	22	1	3p	26	2.00	0.00	0.00	10.60
MIR169.7,8	UGAGCCAGGAUGGCUUGCCGGC	22	1	5p*	79	2.00	5.27	11.50	23.60
MIR171.4	UUGAGCCGCGUCAAUAUCUCA	21	1	3p*	10398	1359.74	190.56	102.57	3412.76
MIR171.5	UUGAGCCGCGUCAAUAUCUUA	21	1	3p*	311	44.39	16.62	28.30	71.76
MIR319.2	UUUGGACCGAAGGGAGCCCCU	21		3p*	570	50.38	181.23	11.50	4.33
MIR390.3-3p	UUGGCGCUAUCUAUCCUGAGU	21	1	3p*	19	1.50	4.05	1.77	1.93
MIR393.6,7,8,9	UUCCAAAGGGAUCGCAUUGAU	21		5p*	73	7.48	1.62	5.31	23.12
MIR393.6,7-3p	UCAUGCGAUCCCUUAGGAACU	21		3p*	35	6.48	0.00	2.65	9.15
MIR395.10	CUGAAGUGUUUGGGGGAGCUU	21		3p*	24	2.00	0.00	0.88	9.15
MIR395.2-5p	AGUUCCUCUGAAUGCUUCAUA	21		5p*	199	25.94	4.87	14.15	57.31
MIR395.3	UGAAGUGUUUGGGGGAACUUU	21	4	3p*	62	3.49	4.05	38.91	0.48
MIR395.3-5p	AGUUCCUCUGAACACUUCACA	21		5p*	28	1.50	1.22	18.57	0.48
MIR395.4,5,6	UGAAGUGUUUGGGGGAACUCU	21	2 (Pre-MIR395.8,9)	3p*	51	7.48	0.81	4.42	13.97
MIR395.4,5,6-5p	AGUUCCUCUGAACGCUUCAUG	21	2	5p*	17	2.99	0.41	3.54	2.89
MIR395.9-5p	AUUUCCCUGAACACUUCAUU	20		5p*	12	0.00	0.41	1.77	4.33
MIR2111.1,2-3p	AUCCUUGGGAUGUAGAUUACC	21	1	3p*	139	9.98	3.24	8.84	48.64

MIR4406.1-3p	CGGUCCUCUCAGAAUCGAUGUAGA	24	3p	30	3.49	6.89	1.77	1.93
NS-MIR399.1,2,3	UGCCAAAGGAGAUUUGCCCAG	21	3p*	294	25.44	9.73	138.83	29.86
NS-MIR399.4,5,6,7	UGCCAAAGGAGAGUUGCCCUG	21	3p*	70	7.48	12.97	6.19	7.71
NS-MIR399.5,6-5p	GGGCUUCUCUUUAUUGGCAGG	21	5p*	35	2.49	3.65	0.88	9.63
NS-MIR399.8,9-5p	GGGCAUGUCUCUUUUGGCAGG	21	5p*	26	1.50	1.22	0.88	9.15
NS-MIR5281.1	UAUAAAUAGGAUCGGAGGUAGUAU	24	3p*	11	2.99	2.03	0.00	0.00
NS-MIR828.1	UCUUGCUCAAAUGAGUAUUCCA	22	5p	71	14.96	0.00	0.00	19.75

[1]Frequência normalizada em transcritos por milhão: TPM = nº de sequências do miRNA/ nº total de sequências na biblioteca x 1.000.000, sendo o total de sequências, as sequências de 18 a 26 nt filtradas.

(*) significa que ambas as sequências do duplex miRNA:miRNA* foram detectadas nas bibliotecas.

Tabela 8. miRNAs inéditos identificados nas bibliotecas de tecidos florais de soja.

miRNA	Sequência	Tamanho (nt)	nº *hits* adicionais no genoma	Região no precursor[1]	Total de sequências	Frequência em TPM[1]			
						Flores	Carpelos	Estames	Pétalas
NF01	CCAAAGUUGGGCUUAAGCUGUA	22		3p	339	47.89	49.87	31.83	40.46
NF02	AUAAGGCUUUGUUGUGGUAUUCCA	24		3p*	143	20.95	30.41	2.65	11.08
NF03ab-3p	GUUUGAUGAUGAUGUUACCGA	21	4	3p*	64	9.48	1.22	10.61	14.45
NF03ab-5p	UUAUCAGUAGCAUCAUCAUCA	21	1	5p*	143	18.46	3.65	14.15	39.01
NF04	UAACGACAGGUAGUAAAUUGAA	22	2	3p	115	18.95	16.22	6.19	14.45
NF05-3p	CCGGCCACCAACAAAGAAAAAAACU	25		3p*	19	2.99	0.41	10.61	0.00
NF05-5p	GUUUUUUUCUUUGUGGUGGCCGGG	24		5p*	104	10.47	1.22	67.20	1.93
NF06	GUUUGGGGCUUGUGUUUUGUGGGC	24		3p*	72	9.48	16.62	4.42	3.37
NF07-3p	ACCCAACAACAAGUGAUCCUUA	22	1	3p*	70	10.47	15.00	3.54	3.85
NF07-5p	AGGGUCACCUGUUCUUGGGUUA	22		5p*	22	3.99	5.27	0.88	0.00
NF08	AACAUGGUAUCAGGGCCUGAUAGA	24		3p	17	2.49	3.65	1.77	0.48
NF09abc	UAUUAAACGACCGAUGUAGAAAGU	24	1	3p	61	9.48	9.73	7.07	4.82
NF10-3p	UCAUAGGAGGAAUCAACUGGC	21		3p*	61	8.98	0.00	15.03	12.52
NF10-5p	CAGCUGAAUCCUCUUAUGAUC	21		5p*	16	2.00	0.00	3.54	3.85
NF11-3p	UACAUGUGCCUCUUCGUCGCUC	22		3p*	61	3.99	9.33	16.80	5.30
NF11-5p	UAGAUGAAGUUACUCUGAGCA	21		5p	24	3.49	2.03	1.77	4.82
NF12-3p	AUUAACUAGUCACAACAAUGGA	22		3p*	35	10.47	4.05	0.88	1.44
NF12-5p.1	AUUCGCACUGAAAUGGAUGUCCGU	24		5p.1*	25	2.99	4.87	2.65	1.93
NF12-5p.2	UUGUUGUGACUAGUUAAUGGGCAU	24		5p.2	52	6.98	10.14	5.31	3.37
NF13	GCCGAAGAUGAAGAGCUUUGUAU	23		5p	50	6.48	13.38	2.65	0.48
NF14	UUUGUUCUGGAUCCCUGUCGUC	22		5p*	48	4.99	10.54	6.19	2.41
NF15	UUGGCGGAAGUAAUACUAGGUA	22		3p	46	4.49	10.95	0.88	4.33
NF16	UUAGCUUCUUUCACCUUUCCC	21		3p	40	3.49	2.43	16.80	3.85
NF17	UUUUUAAAAGGUUCAGUUAGGU	22		5p	32	5.49	6.08	0.88	2.41
NF18-3p	UCAGGGAUUCAAAACAACGAAA	22		3p	30	5.99	5.27	0.88	1.93
NF18-5p	UCUGAGUCCAUGAUAUAUUAAA	22		5p*	20	4.99	3.24	0.00	0.96
NF19-3p	UUGCUGGACGUGGCGCCUUCCA	22		3p*	13	0.50	2.03	0.88	2.89
NF19-5p	AAGGCACCACUUCAGCAAUGGA	22		5p*	29	2.99	5.27	0.88	4.33
NF20	ACAGAUUGACAAUCCAUGUGAGCUA	25		5p	28	3.49	5.68	2.65	1.93

NF21	AGUUAUCAAAAGCAAAAGUUUGGA	24		3p	28	3.49	6.89	0.88	1.44
NF22	GUUGAUGUGUCACAUGGAGAUGGA	24		3p	27	2.00	2.03	0.00	8.67
NF23	GAAUGGUGAGGAUGAAAAGUAACU	24		3p*	27	2.00	5.68	0.88	3.85
NF24ab	GUUGGACUCAAGGAACCUA	19		5p	11	1.50	0.00	7.07	0.00
NF25	AAAGUGUUUGAAUCUCAAUUAGA	23	7	3p*	18	0.50	0.41	10.61	1.93
NF26	UUAAGGACUAAAACAAAACAAACA	24		3p	13	1.50	3.65	0.00	0.48
NF27	UGAGCCUUGCAGCAGUUAUGACAA	24		5p	10	1.00	3.24	0.00	0.00
NF28-3p	UUACAUGGACUAAAAAUGAGCAAA	24	4	3p*	81	14.47	14.60	2.65	6.26
NF28-5p	UGCUCAUUUUUAGUCCUGUAAGU	23		5p*	10	1.00	2.43	0.88	0.48
NF29	AAGGCAGAACGAUAUGUACGCAGA	24		3p	19	3.99	1.62	1.77	2.41
NF30	UGAUUAUGAGGUCUGACACAAA	22	13	3p	67	10.97	6.89	7.96	9.15
NF31	UCAUAGGAGAGAAAAAUAGGAAGG	24		5p	18	4.49	2.43	0.88	0.96
NF32	UAUUGUGUAAGCUUCCUAAAGAGA	24		5p	20	2.00	0.81	0.88	6.26
NF33	UAUUGGUCUUUUUGUAGUGAC	21		3p*	25	2.99	0.41	2.65	7.22
NF34	CAACCCUCCUCAGUUAGAUCUC	22		5p*	22	2.99	5.27	0.88	0.96
NF35	UAAUAGAGGGAAGAAGAUGAA	21		5p	27	4.99	2.03	1.77	4.82
NF36	UAAAAUCGAUGUAGAAAGUGCC	22		5p	26	4.99	2.43	2.65	3.37
NF37ab-3p	AUCUUGAUCGUUCAUGUUUGGU	22	1	3p*	27	5.49	2.43	0.88	4.33
NF37b-5p	CCGUAUAACCAAACAUGGAGGGU	23		5p*	22	2.99	3.24	4.42	1.44
NF38	GUAUUAGAUGCAUACUCAGUGGAU	24		5p*	22	3.49	2.43	4.42	1.93
NF39ab	UAUAGGACCGAUGUAGAAUGUGUU	24	1	3p*	94	8.48	10.95	20.34	13.00
NF40	AUAACGAUUGUGCAAGUAUAGGGA	24		3p	13	2.99	2.03	1.77	0.00

[1]Frequência normalizada em transcritos por milhão: TPM = nº de sequências do miRNA/ nº total de sequências na biblioteca x 1.000.000, sendo o total de sequências, as sequências de 18 a 26 nt filtradas.

Os miRNAs sublinhados foram testados quanto ao padrão de expressão nas amostras de órgãos florais de soja através de RT-qPCR.

(*) significa que ambas as sequências do duplex miRNA:miRNA* foram detectadas nas bibliotecas.

4.4 Famílias de microRNAs identificadas em tecidos florais de soja e seus membros 5p e 3p

Os miRNAs identificados foram agrupados em famílias. Os 51 miRNAs inéditos (Tabela 3) pertencem a 40 novas famílias (Tabela 8). Os demais miRNAs identificados pertencem a 64 famílias conhecidas de miRNAs, das quais três ainda não tinham sido reportadas em soja (MIR399, MIR5281 e MIR828) (Tabela 7). As famílias MIR399 e MIR828 já haviam sido identificadas em diversas espécies, enquanto que MIR5281, apenas em *M. truncatula* (Devers *et al*., 2011).

As principais famílias (mais frequentes) foram analisadas quanto ao número de membros (Figura 10) e quanto à frequência absoluta (número total de sequências) (Figura 11) nas quatro bibliotecas combinadas. Foi feita distinção entre miRNAs 5p e 3p, pois mesmo pertencendo à mesma família, suas sequências e, consequentemente, seus genes alvos podem ser distintos.

4.4.1 MicroRNAs e microRNAs*

Nos tecidos florais de soja, a maioria das famílias apresentou acúmulo de miRNAs de um dos braços dos precursores, especialmente a família mais abundante, MIR159, cujos membros 3p apresentaram 858 vezes mais sequências que os 5p (Figura 11). As famílias MIR2109 e MIR395 apresentaram frequências equivalentes entre miRNAs originados em ambos os braços dos precursores (Figura 11). A família MIR169 origina miRNAs maduros dos braços 5p de seus precursores para a maioria das espécies (miRBase v.18.0), como observado no presente estudo, no qual sequências correspondentes a membros 5p foram 188 vezes mais numerosos em relação aos membros 3p, os quais poderiam ser considerados como miRNAs* (Figura 11). No entanto, apesar de sua baixa frequência, estes miRNAs podem ser funcionais, visto que em *M. truncatula*, a clivagem de mRNAs alvos pelo MIR169d* foi confirmada por degradoma e seu gene alvo é relevante para a simbiose com fungos arbusculares (Devers *et al*., 2011).

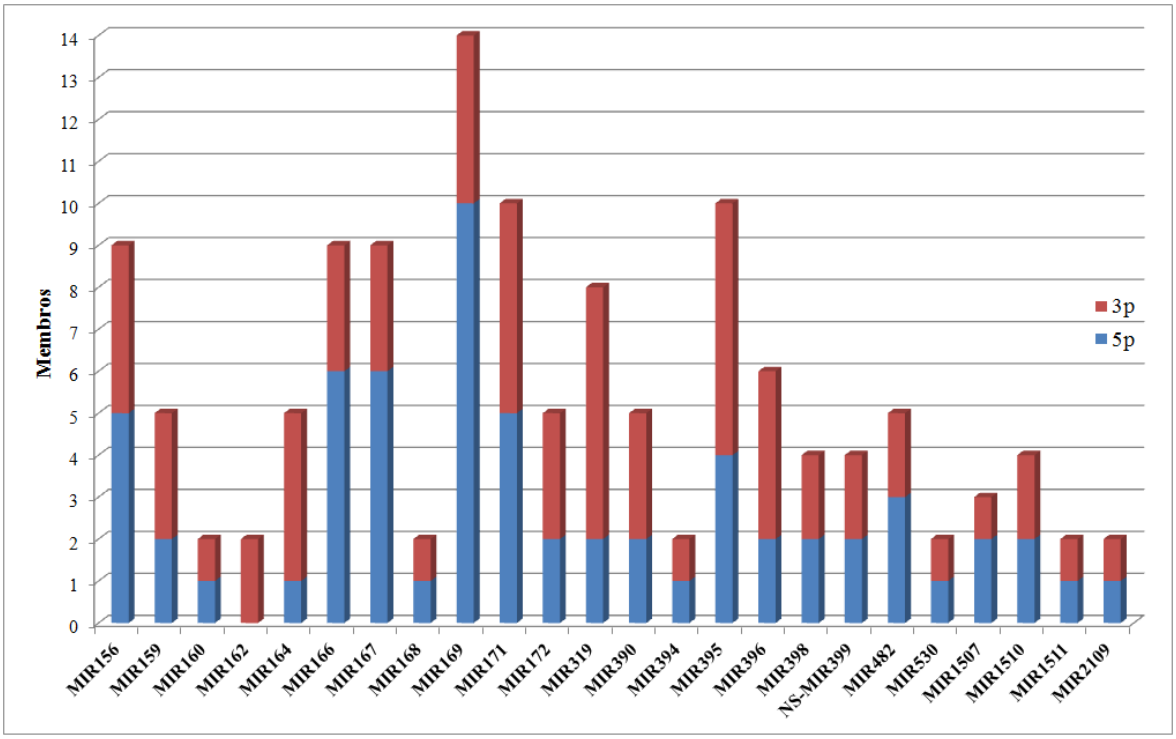


Figura 10. Número de membros de microRNAs identificados para cada uma das 24 famílias mais frequentes nas quatro bibliotecas de sRNAs de tecidos florais de soja combinadas. Foram incluídos todos os membros de cada família (5p e 3p, novos e conhecidos).

Quando apenas os miRNAs originados em um dos braços do precursor são funcionais (fitas guia), os do outro braço (fitas passageiras ou miRNAs*) ficam vulneráveis à degradação por não estarem protegidos pela associação com AGO (Kai e Pasquinelli, 2010). Assim, discrepâncias muito grandes de frequência entre MIR-5p e MIR-3p em bibliotecas de sRNAs podem indicar que apenas uma das fitas do duplex tende a ser incorporada em RISC e ser funcional em uma dada condição. No entanto, cabe lembrar que poucas moléculas de miRNA são suficientes para guiar várias rodadas de silenciamento ou ainda causar a amplificação de sinal via produção de siRNAs (Allen e Howell, 2010; Chen *et al*., 2010; Zhai *et al*., 2011). Além disso, em *Arabidopsis* foi demonstrada a funcionalidade de diversos miRNAs originados de um mesmo precursor, mesmo em baixa frequência (Zhang *et al*., 2010).

Refletindo os múltiplos eventos de duplicação que ocorreram no genoma da soja (Cannon, 2008), muitos miRNAs são codificados em diversos *loci*, porém é difícil

determinar quais genes são ativos em uma determinada situação. Neste sentido, a detecção de uma sequência do braço oposto do precursor pode ser uma evidência de que o gene que a codifica está de fato sendo transcrito quando tal sequência só existe naquele lócus. Por exemplo, foram encontrados oito *loci* para gma-MIR164, adicionais ao anotado. Este miRNA, apesar de ser o único membro 5p da família (Figura 10), é 11 vezes mais frequente que os quatro membros 3p somados (Figura 11). No entanto, estes membros 3p fornecem evidência de que pelo menos quatro destes *loci* adicionais descobertos neste estudo são ativos nos tecidos florais de soja.

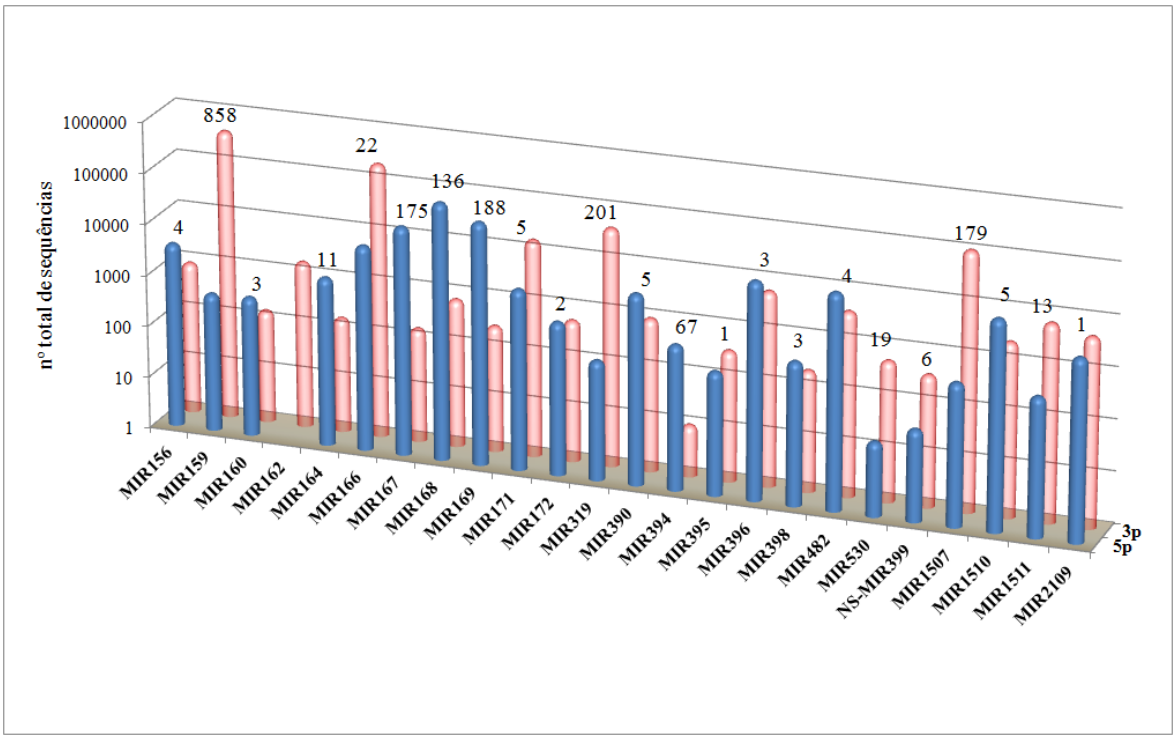


Figura 11. Frequência absoluta das 24 famílias de miRNAs mais frequentes nas quatro bibliotecas combinadas de sRNAs de tecido floral de soja. Foram contadas as sequências correspondentes a todos os membros de cada família (5p e 3p, novos e conhecidos). Os números acima das barras representam o "*fold change*" entre as frequências de membros 5p e 3p, dado pela razão entre eles.

4.4.2 Famílias de microRNAs frequentes em tecidos florais de soja

Os miRNAs filogenetimente conservados geralmente são altamente abundantes e os espécie-específicos, em geral são raros, pois evoluíram recentemente e participam da regulação de processos mais específicos (Voinnet, 2009). Como esperado, as 40 novas famílias de miRNAs espécie-específicos identificados nos tecidos florais de soja apresentaram frequências relativamente baixas, variando entre 10 e 339 sequências (Tabela 8). Portanto, estas famílias não estiveram entre as 24 mais frequentes nas bibliotecas combinadas.

As famílias MIR167, MIR172 e MIR390, que estiveram entre as 24 famílias mais frequentes no presente estudo, também foram fortemente expressas durante os estádios florais de tomate comparados a oito estádios de desenvolvimento do fruto (Mohorianu *et al*., 2011). As famílias MIR1507, MIR1510, e MIR2109 também foram altamente frequentes nas bibliotecas de tecidos florais de soja (Figura 11), como já observado em diversas bibliotecas de sRNAs de leguminosas (Zhai *et al*., 2011). Porém, em análise de expressão por *Northern blot*, MIR1507 foi expresso em raiz, haste, folha e semente madura, mas ausente nas flores de soja (Li *et al*., 2010). Neste mesmo trabalho, gma-MIR1511 foi expresso em todos os tecidos, corroborando com sua alta frequência nas bibliotecas de tecidos florais de soja. Em feijão, no entanto, MIR1511 foi detectado somente em condições de estresse (Valdés-López *et al*., 2010).

A família MIR482 esteve entre as 24 mais frequentes neste estudo (Figura 11). Foi confirmado por 5'RACE a clivagem de um sítio alvo de gma-miR482a-3p (Li *et al*., 2010) presente em dois genes de soja (Glyma16g00400 e Glyma12g28730), ambos codificando para uma proteína similar (87%) a uma quinase 41 *shaggy-like* de *Arabidopsis*. Esta proteína tem função predita na sinalização extracelular que regula a transcrição em células em diferenciação e é expressa exclusivamente em inflorescências (Yamada *et al*., 2003).

MIR168 foi representada essencialmente por gma-MIR168 (5p), o qual foi altamente frequente nas bibliotecas (Figura 11). Evidências apontam para a regulação entre AGO1 e MIR168 e para a importância deste mecanismo no desenvolvimento de *Arabidopsis* (Vaucheret *et al*., 2004). Em soja, a clivagem de um sítio alvo presente em

dois mRNAs codificando para proteínas AGO foi verificada no degradoma de sementes (Song *et al*., 2011). Outro gene alvo (Glyma11g04200) deste miRNA foi validado em soja por 5'RACE (Subramanian *et al*., 2008) e codifica um receptor quinase (RLK).

Para MIR160, somente um lócus foi identificado, o qual origina miRNAs maduros em ambos os braços do precursor (Figura 11). Doze genes alvos de gma-MIR160 foram identificados no degradoma de semente de soja, todos eles codificando ARFs (Song *et al*., 2011). Mutantes de *Arabidopsis* com perda de função deste miRNA apresentaram flores irregulares, órgãos florais dentro dos carpelos, fertilidade reduzida, sementes aberrantes, número anormal de órgãos florais, sépalas e pétalas estreitas (Liu *et al*., 2010). Portanto, este miRNA é importante em diversos aspectos do desenvolvimento floral, o que justifica sua abundância nos tecidos florais de soja.

Além disso, todas as outras famílias de miRNAs que têm sido relacionadas ao florescimento e desenvolvimento reprodutivo de plantas foram detectadas e estão presentes entre as 24 famílias mais frequentes nas bibliotecas de sRNAs de tecidos florais de soja (Figura 11).

4.5 Análise de expressão dos microRNAs identificados entre as bibliotecas de sRNAs de órgãos florais de soja por DEGseq

Assumindo que a frequência de sequências correspondentes a miRNAs em bibliotecas de sRNAs pode ser utilizada como um índice para a estimativa da abundância relativa destes miRNAs, a contagem de sequências de cada miRNA identificado foi utilizada em comparações par a par entre as bibliotecas de carpelos, estames e pétalas, usando o pacote do programa estatístico R "DEGseq" (Wang *et al*., 2010).

O maior número de miRNAs diferencialmente expressos (129), tanto reprimidos quanto induzidos, ocorreu na comparação entre carpelos e pétalas, enquanto a comparação entre pétalas e estames apresentou o maior número de miRNAs (149) igualmente expressos (Figura 12A). Estas diferenças podem refletir o fato de que estames e pétalas compartilham o domínio de expressão dos genes de classe B, sendo menos contrastantes em termos de

expressão gênica que carpelos e pétalas, que não compartilham domínios de expressão dos genes ABC.

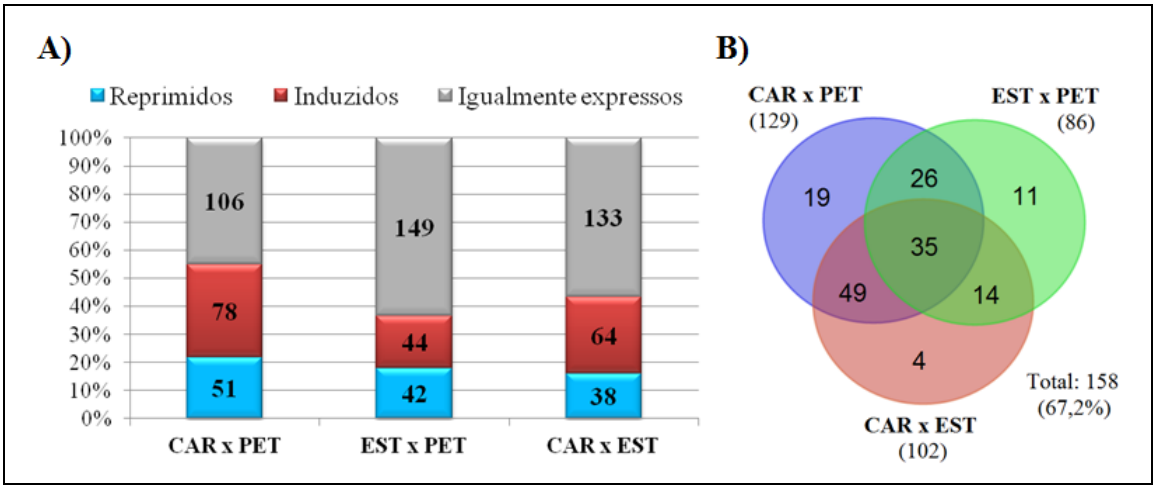


Figura 12. Análise de expressão dos miRNAs maduros identificados entre as bibliotecas de sequências de sRNAs de órgãos florais de soja. **(A)** Análise da expressão de todos os 235 miRNAs identificados; **(B)** Distribuição dos miRNAs diferencialmente expressos entre as comparações par a par. A indução/repressão é relativa à primeira biblioteca citada na comparação. PET: pétalas; CAR: carpelos; EST: estames.

No total, 158 miRNAs, representado 67.2% dos miRNAs identificados, foram diferencialmente expressos, sendo 34 miRNAs diferencialmente expressos em apenas uma das comparações e 35 miRNAs, em todas as comparações (Figura 12B). Outros 26 miRNAs apresentaram diferenças de expressão significativas entre as pétalas e os tecidos reprodutivos, sem diferença entre carpelos e estames; 14 foram diferencialmente expressos nos estames em relação a pétalas e carpelos e 49 foram diferencialmente expressos em carpelos em relação aos demais tecidos, que não diferiram entre si (Figura 12B). Nos carpelos, mais miRNAs foram induzidos do que reprimidos: na comparação com pétalas, 78 miRNAs foram induzidos e 51 reprimidos; e na comparação com estames, 64 foram induzidos e 38 reprimidos (Figura 12A).

4.5.1 Tamanho de sequência dos microRNAs e seu padrão de expressão

Ao analisar em separado o padrão de expressão dos miRNAs das principais classes de tamanho, observa-se que aproximadamente a metade dos miRNAs de 21 nt diferencialmente expressos são reprimidos e metade são induzidos, em todas as comparações (Figura 13). Por outro lado, a grande maioria dos miRNAs de 22 e 24 nt diferencialmente expressos é induzida nos carpelos, comparados tanto a pétalas como a estames. Já nas comparações entre estames e pétalas, este viés para a expressão diferencial de miRNAs de um determinado tamanho não é observado (Figura 13).

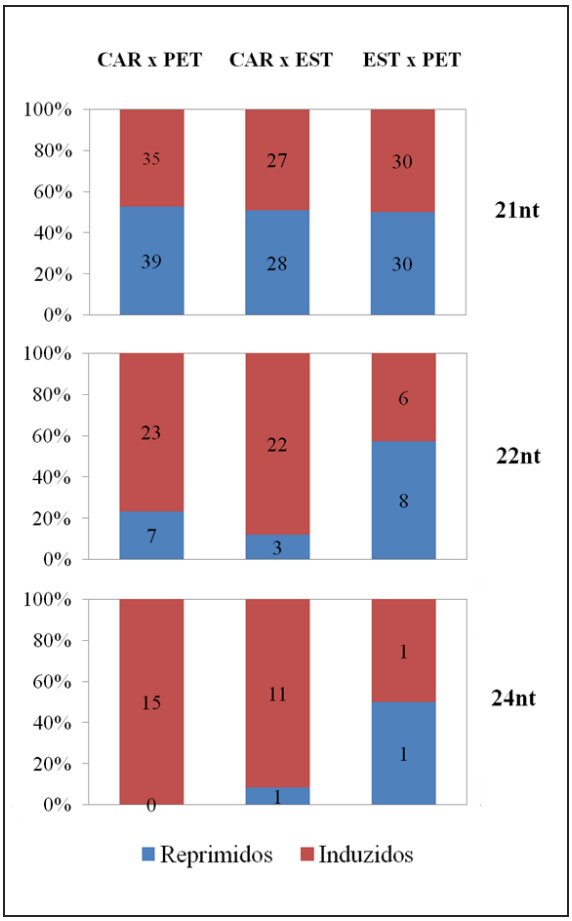


Figura 13. Análise do padrão de expressão dos miRNAs identificados dos principais tamanhos (21, 22 e 24nt) nas comparações par a par entre três bibliotecas de sRNAs de órgãos florais de soja. Os números de miRNAs estão indicados dentro das barras. PET: pétalas; CAR: carpelos; EST: estames.

Cabe ainda lembrar que a biblioteca de carpelos apresentou um pico maior de frequência de sequências de 24 nt que as demais bibliotecas e foi a única cujos sRNAs de 22 nt foram mais abundantes que os de 21 nt (Figura 4). O mesmo foi reportado em órgãos que também estão relacionados ao desenvolvimento do embrião e do fruto em outras plantas, como óvulos em desenvolvimento (de três dias anteriores a três dias posteriores à antese) de plantas de algodão (Wang *et al*., 2011a) e sementes de milho nas fases iniciais da germinação (Wang *et al*., 2011b). Em tomate, foram sequenciadas bibliotecas de sRNAs em 10 pontos do desenvolvimento da flor e do fruto, compreendendo dois estádios florais: botão e flor madura. Os sRNAs obtidos foram agrupados pelo padrão de expressão nas bibliotecas, independente de similaridade de sequência e um grupo altamente expresso em botões florais e flores apresentou um forte viés para sRNAs de 22 nt. Além disso, um viés para a produção de sRNAs de 24 nt foi particularmente forte no botão floral, decrescendo gradualmente nos estádios posteriores (Mohorianu *et al*., 2011).

Este viés de tamanho de sRNAs (incluindo os miRNAs), como observado nos carpelos, pode ocorrer devido a mecanismos que favoreçam a produção de tamanhos específicos de sRNAs, como a clivagem diferencial de moléculas precursoras (Ebhardt *et al*., 2010) ou que favoreçam sua estabilidade, como a maior abundância de um determinado tipo de AGO com mais afinidade de associação a sRNAs de tamanhos específicos (Mi *et al*., 2008). Foi observado, por imunoprecipitação, que a maioria dos sRNAs associados com AGO1 e AGO2 são de 21 nt, os associados com AGO4 são de 24 nt, os associados a AGO5 são de 21, 22 e 24 nt (Mi *et al*., 2008); e os associados a AGO9 são principalmente de 24 nt (Olmedo-Monfil *et al*., 2010). As proteínas AGO5 e AGO9 já foram relacionadas ao desenvolvimento dos tecidos reprodutivos femininos de *Arabidopsis*: AGO5 é expressa nos tecidos somáticos dos óvulos e é fundamental para promover a iniciação da megagametogênese (Tucker *et al*., 2012) e AGO9, associada a siRNAs de 24 nt, promove o silenciamento de TEs nos gametas femininos e suas células acessórias (Olmedo-Monfil *et al*., 2010).

Quanto aos miRNAs especificamente, tem sido observado que o tamanho está relacionado com as suas funções. Em arroz, por exemplo, foram identificados miRNAs de 24 nt associados a AGO4 que guiam a metilação de DNA a aproximadamente 80 nt de distância de seus sítios alvos (Wu *et al*., 2010). Estes miRNAs foram denominados

lmiRNAs (*long* miRNAs) e requerem DCL3, em vez de DCL1, para sua biogênese (Wu *et al*., 2010). Muitos miRNAs de 22 nt estão envolvidos nas clivagens iniciais que levam à produção de siRNAs secundários. A expressão transiente em tabaco confirmou que os miRNAs identificados como iniciadores da produção de siRNAs só realizam esta função quando expressos na forma de 22 nt. Por outro lado, ao aumentar o tamanho de MIR319, que não exerce essa função, de 21 para 22 nt, ele se torna capaz de guiar a produção de siRNAs (Chen *et al*., 2010). Além disso, a análise de dados de sequenciamento de diversas leguminosas mostrou que *M. truncatula* e soja produzem miRNAs de 22 nt em abundância, muitos deles envolvidos na produção de siRNAs em fase (*phasi*RNAs), especialmente pela clivagem de domínios conservados em genes de resistência com repetições ricas em leucina (NB-LRR) (Zhai *et al.*, 2011).

4.5.2 Agrupamento dos microRNAs de acordo com seu padrão de expressão

Os padrões de expressão dos miRNAs identificados que foram diferencialmente expressos em pelo menos uma das comparações feitas com DEGseq foram representadas em *heatmaps* (Figuras 14, 15, 16, 17 e 18).

O miRNA gma-MIR164 foi reprimido nos estames (Figura 14). Em *Arabidopsis*, este miRNA apresentou baixos níveis (Chambers e Shuai, 2009) ou não foi detectado (Grant-Downton *et al*., 2009) em pólen maduro. MIR164 regula genes da família de fatores de transcrição NAC, requeridos na formação dos limites entre os primórdios dos órgãos por suprimir o crescimento celular entre eles (Laufs *et al*., 2004). Em soja, 10 mRNAs alvos codificando fatores de transcrição NAC já foram confirmados pela análise de degradoma (Song *et al*., 2011).

A família MIR166 foi a segunda maior família em número de sequências (Figura 11). No geral, esta família foi induzida nos carpelos (Figuras 14, 15, 16 e 17). Interessantemente, gma-MIR166g-5p_new apresentou um padrão de expressão inverso aos demais miRNAs desta família (Figura 16), com maiores níveis em pétalas, seguido por estames e por último, carpelos. Estes dados corroboram com um estudo que mostrou, com

hibridizações *in situ*, que gma-MIR166ab e gma-MIR166ab* apresentam localizações distintas na folha de soja em desenvolvimento (Wong *et al*., 2011). Isto sugere que em um tecido o miRNA produzido em um braço do precursor é usado como fita guia, enquanto que em outro tecido, o contrário ocorre. Portanto, ambos os miRNAs são funcionais e suas funções provavelmente divergem.

Os miRNAs maduros originados do precursor gma-MIR172cde (miRNA anotado - Figura 14, seu iso - Figura 15; e miRNA 5p - Figuras 16), foram induzidos nos tecidos reprodutivos (carpelos e estames), enquanto que gma-MIR172a/bh-3p e gma-MIR172b-5p_iso, foram reprimidos (Figuras 14 e 15, respectivamente). Diferentes membros desta família podem desempenhar funções distintas, visto que linhagens transgênicas superexpressando MIR172a e MIR172b, mas não a linhagem superexpressando MIR172c, apresentaram transformações homeóticas (carpelos em lugar do perianto) em *Arabidopsis* (Chen, 2004). Portanto, pode-se esperar que membros desta família não apresentem o mesmo padrão de expressão entre os tecidos.

Membros da família MIR393, foram reprimidos nos carpelos (ex.: MIR393.6,7,8,9 – Figura 17). Um alvo de gma-MIR393 (Glyma02g07240) validado por 5'RACE (Subramanian *et al*., 2008) codifica uma proteína similar (73%) a uma proteína F-BOX envolvida na sinalização da auxina (AFB3) de *Arabidopsis*. Esta proteína apresenta expressão ubíqua, com maiores níveis em flores de *Arabidopsis*, e sua função inclui a regulação da embriogênese pela auxina. Portanto, os baixos níveis de MIR393 nos carpelos pode ser devido à necessidade da presença deste transcrito nos carpelos, onde a embriogênese ocorre.

Os membros da família MIR396 foram reprimidos nos tecidos reprodutivos em relação às pétalas (ex.: gma-MIR396c/b-5p – Figura 14). A clivagem de um sítio alvo presente em dois genes de soja foi confirmada por 5'RACE (Subramanian *et al*., 2008). Ambos codificam proteases de cisteína (Cathepsin L), as quais já foram caracterizadas em soja, estando envolvidas na mobilização de proteínas de reserva nos cotilédones durante as primeiras fases do desenvolvimento da plântula (Asano *et al*., 1999; Kodera *et al*., 2005).

O miRNA NS-MIR399.1,2,3 (5p), membro mais frequente da nova família de soja MIR399, foi induzido em estames em relação aos demais órgãos, enquanto dois membros

3p apresentaram um padrão oposto, sendo reprimido nos estames comparados com as pétalas (Figura 17). Em pólen de *Arabidopsis*, MIR399 regula o metabolismo do fosfato, necessário para o acúmulo de reservas e possivelmente para o crescimento heterotrófico do tubo polínico (Grant-Downton *et al*., 2009).

Quanto aos miRNAs inéditos, o padrão de expressão dominante pelas análise de DEGseq foi a indução nos carpelos, em ambas as comparações, ou somente na comparação com pétalas (Figura 18).

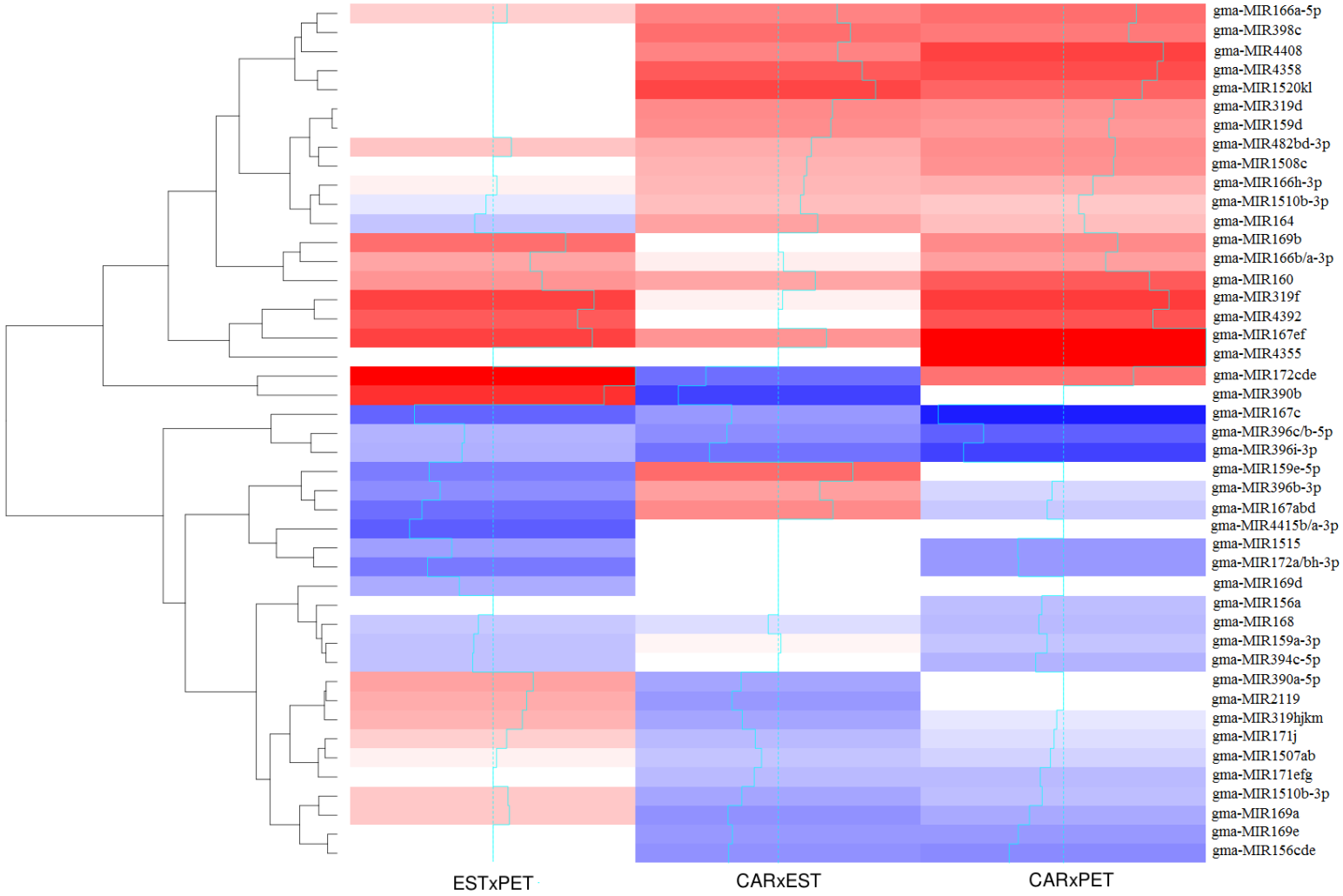


Figura 14: *Heatmap* dos miRNAs conhecidos diferencialmente expressos em tecidos florais de soja. Esta figura foi gerada a partir das comparações par a par entre as bibliotecas de carpelos (CAR), estames (EST) e pétalas (PET), feitas por DEGseq. As cores na escala do vermelho representam indução e as cores na escala do azul representam repressão do miRNA na primeira biblioteca citada. Em branco estão as comparações não significativas. Os histogramas representam a proporção de sequências do respectivo miRNA em cada uma das bibliotecas da comparação.

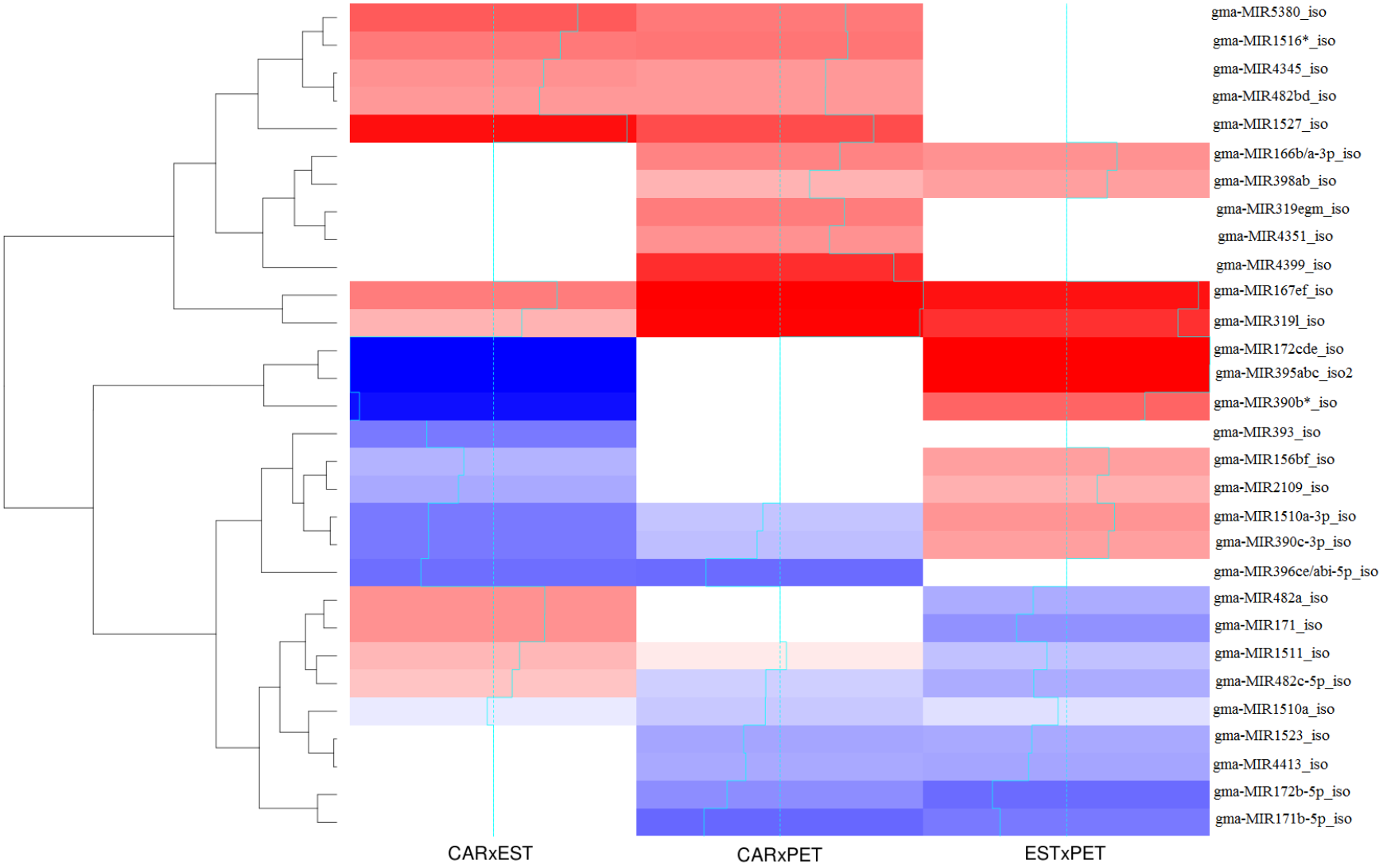


Figura 15: *Heatmap* das isoformas de miRNAs conhecidos diferencialmente expressos em tecidos florais de soja. Esta figura foi gerada a partir das comparações par a par entre as bibliotecas de carpelos (CAR), estames (EST) e pétalas (PET), feitas por DEGseq. As cores na escala do vermelho representam indução e as cores na escala do azul representam repressão do miRNA na primeira biblioteca citada. Em branco estão as comparações não significativas. Os histogramas representam a proporção de sequências do respectivo miRNA em cada uma das bibliotecas da comparação.

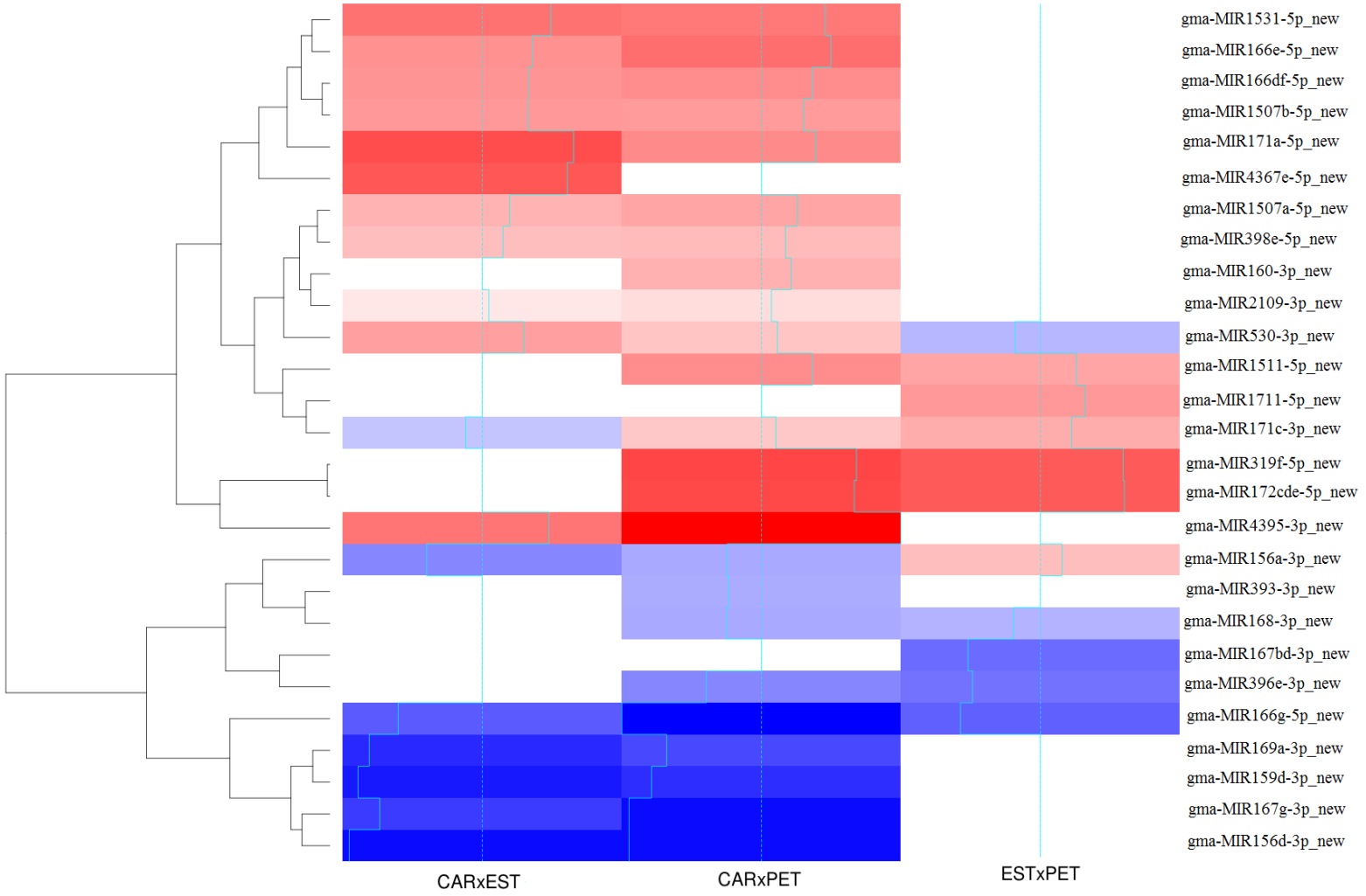


Figura 16: *Heatmap* dos novos miRNAs maduros de precursores conhecidos diferencialmente expressos em tecidos florais de soja. Esta figura foi gerada a partir das comparações par a par entre as bibliotecas de carpelos (CAR), estames (EST) e pétalas (PET), feitas por DEGseq. As cores na escala do vermelho representam indução e as cores na escala do azul representam repressão do miRNA na primeira biblioteca citada. Em branco estão as comparações não significativas. Os histogramas representam a proporção de sequências do respectivo miRNA em cada uma das bibliotecas da comparação.

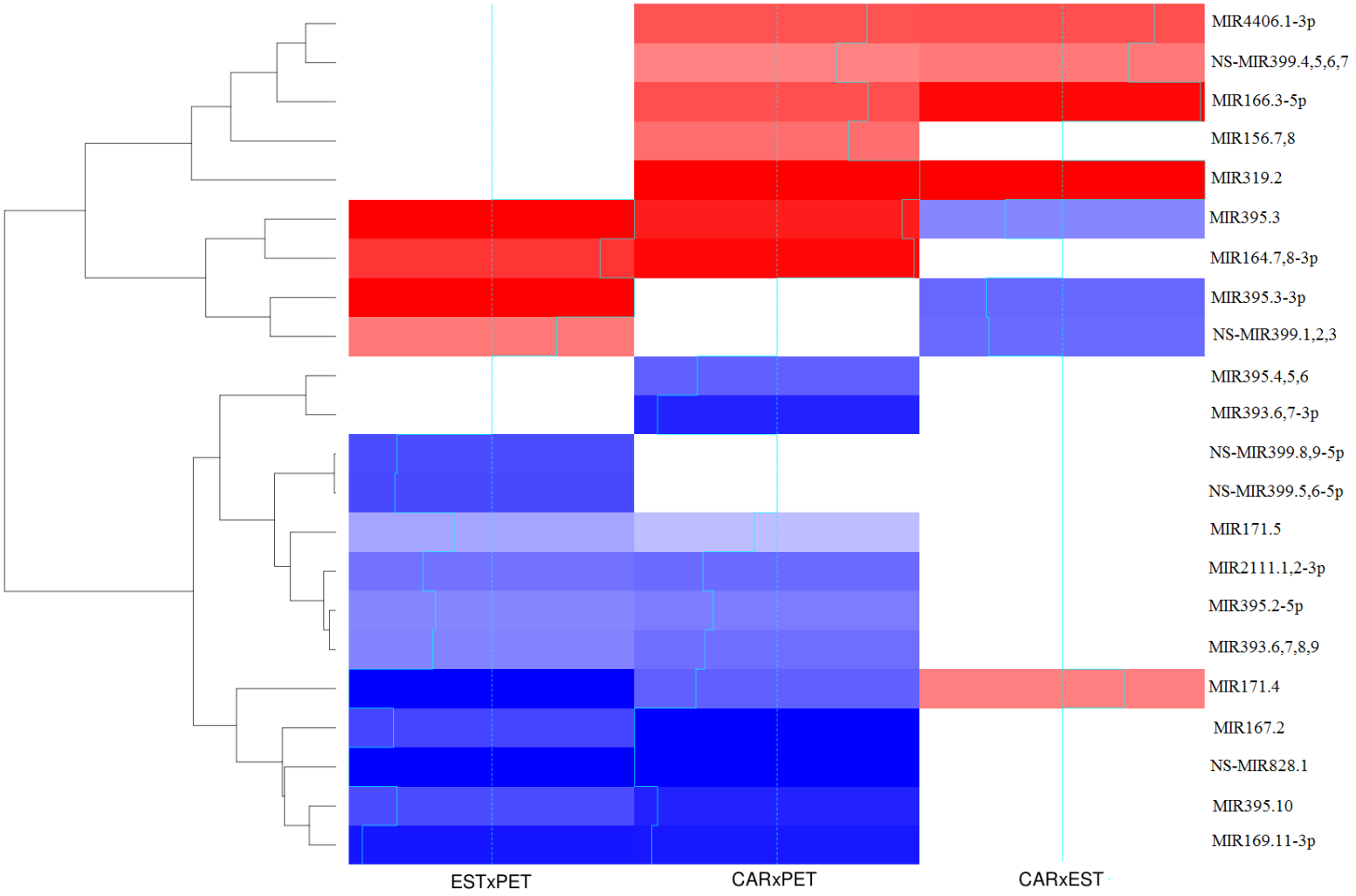


Figura 17: *Heatmap* dos novos membros de famílias de miRNAs conhecidas (já anotadas em soja ou apenas em outras espécies) diferencialmente expressos em tecidos florais de soja. Esta figura foi gerada a partir das comparações par a par entre as bibliotecas de carpelos (CAR), estames (EST) e pétalas (PET), feitas por DEGseq. As cores na escala do vermelho representam indução e as cores na escala do azul representam repressão do miRNA na primeira biblioteca citada. Em branco estão as comparações não significativas. Os histogramas representam a proporção de sequências do respectivo miRNA em cada uma das bibliotecas da comparação.

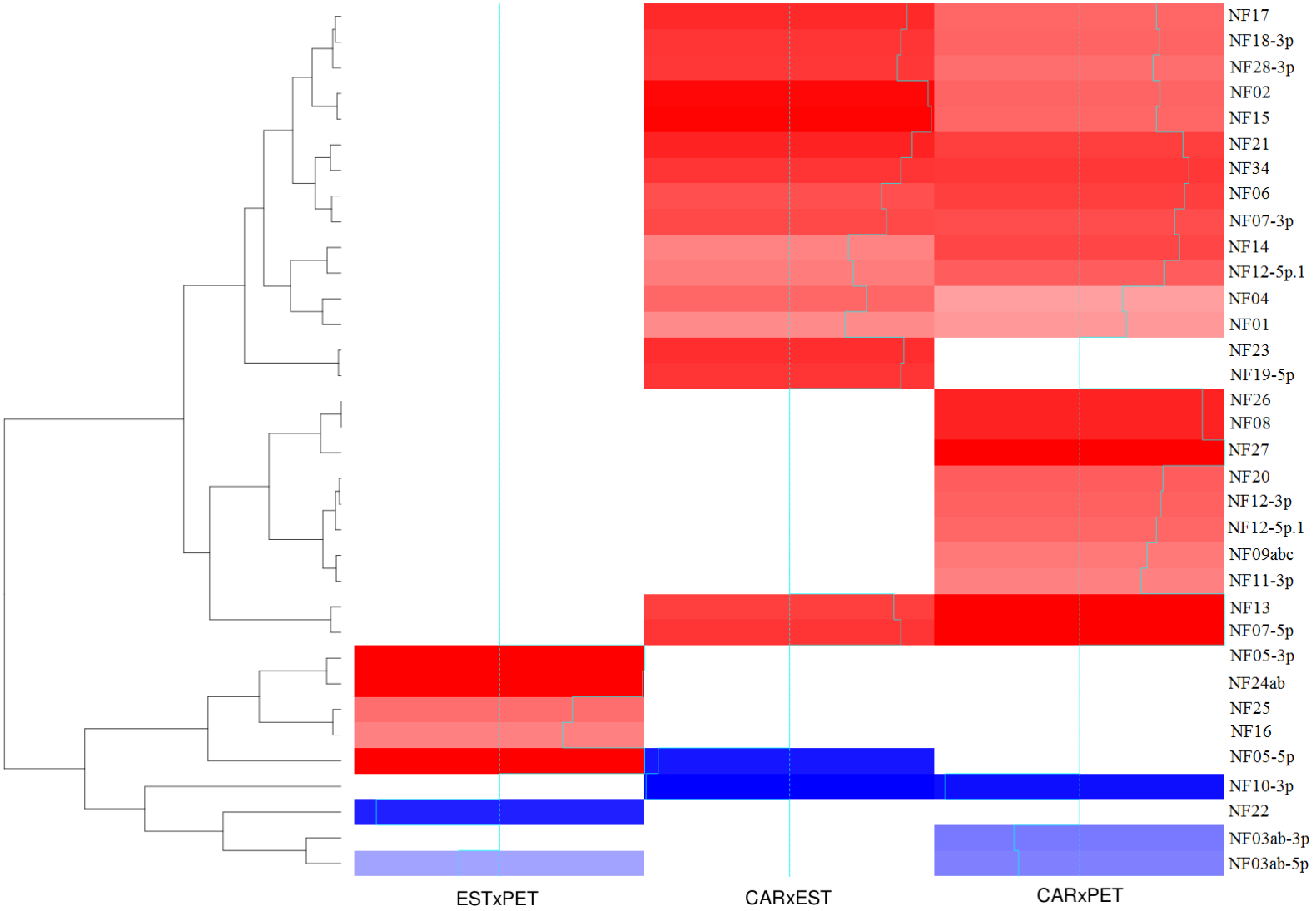


Figura 18: *Heatmap* dos miRNAs inéditos diferencialmente expressos em tecidos florais de soja. Esta figura foi gerada a partir das comparações par a par entre as bibliotecas de carpelos (CAR), estames (EST) e pétalas (PET), feitas por DEGseq. As cores na escala do vermelho representam indução e as cores na escala do azul representam repressão do miRNA na primeira biblioteca citada. Em branco estão as comparações não significativas. Os histogramas representam a proporção de sequências do respectivo miRNA em cada uma das bibliotecas da comparação.

4.6 Análise do padrão de expressão de microRNAs por RT-qPCR em amostras de órgãos florais de soja

Para confirmar os padrões de expressão diferencial dos miRNAs obtidos pela sua frequência nas bibliotecas de sRNAs de órgãos florais de soja, um subconjunto de 61 miRNAs (Tabelas 4, 5, 6 e 8) diferencialmente expressos nas bibliotecas foi analisado por RT-qPCR, incluindo 32 dos 34 miRNAs inéditos diferencialmente expressos. Os Cts das reações utilizando os oligonucleotídeos com sequências idênticas às de gma-MIR162 e gma-MIR169a foram usados como normalizadores.

Dos 32 miRNAs inéditos testados, quatro (NF05-3p, NF08, NF24ab e NF34) não puderam ser amplificados, apresentando amplificações e temperaturas de fusão comparáveis às dos controles negativos. Suas baixas frequências nas bibliotecas (Figura 9 – NF24ab só foi detectado em estames; e Tabela 8) indicam que provavelmente a amplificação não ocorreu por eles serem muito pouco expressos e as sequências de seus oligonucleotídeos propiciarem a formação de dímeros e amplificação inespecífica.

4.6.1 Diferenças entre as análises de expressão por DEGseq e RT-qPCR

Dos 61 miRNAs testados com RT-qPCR, 24 não apresentaram diferenças de expressão significativas entre os órgãos florais e 14 tiveram diferenças significativas de expressão, porém entre comparações diferentes ou mostrando relações opostas às obtidas por DEGseq. A existência de discrepâncias entre análises de expressão realizadas através de métodos diferentes são comuns, principalmente para genes que são expressos em baixos níveis (Nagalakshmi *et al*., 2008; Bloom *et al*., 2009), como ocorreu com a maioria dos miRNAs inéditos testados, os quais são pouco expressos.

Como muitos miRNAs de uma mesma família têm sequências maduras muito similares, mesmo usando oligonucleotídeos *stem-loop*, que conferem uma maior especificidade na amplificação nas reações de RT-qPCR, alguns membros e/ou isoformas

dificilmente podem ser distinguidos (Chen *et al*., 2012b), especialmente quando diferem apenas na extremidade 5'. Como membros de uma família podem se expressar diferentemente, a presença de um membro pode compensar a falta de outro em um tecido, resultando em níveis de expressão estáveis entre os tecidos. Isto explica a ocorrência de um menor número de diferenças de expressão significativas em RT-qPCR que nas análises por DEGseq, nas quais cada sequência foi analisada individualmente, e sequências com apenas um nucleotídeo de diferença foram distinguidas.

Por um lado, não se sabe até que ponto se deve analisar separadamente miRNAs que podem ser coexpressos e agir sobre os mesmos sítios alvo devido à similaridade de suas sequências. Por outro lado, em famílias de miRNAs com múltiplas cópias de um precursor, é difícil determinar quais miRNAs são de fato isoformas originadas de um mesmo precursor (Guo e Lu, 2010). E, ainda que sejam realmente coexpressos, diferenças de apenas 1 nt podem de fato fazer miRNAs atuarem de maneiras diferentes (Ebhardt *et al*., 2010), o que irá refletir em seus níveis de expressão, tornando desaconselhável agrupá-los. Diante disso, uma boa estratégia de análise de expressão de miRNAs é usar a frequência de sequências da forma mais abundante do miRNA (Guo e Lu, 2010). No entanto, comparado a microarranjo e análise por dados de sequenciamento, RT-qPCR foi considerado o método mais confiável para medir a expressão diferencial de miRNAs (Git *et al*., 2010).

4.6.2 Validação da expressão diferencial de microRNAs por RT-qPCR

Dos miRNAs testados por RT-qPCR, 19 apresentaram diferenças de expressão significativas e comparáveis às diferenças apontadas por DEGseq. Entre eles, 12 são miRNAs conhecidos ou suas isoformas, um é novo miRNA presente em precursor conhecido de soja e seis são miRNAs inéditos. Para estes miRNAs as expressões relativas calculadas a partir dos dados de RT-qPCR estão representadas na Figura 19.

O único miRNA que apresentou um "*fold change*" maior que dois nestas análises foi o miRNA inédito NF13, cujos níveis foram aumentados nos carpelos em relação aos

demais órgãos testados. Embora as diferenças de expressão dos demais miRNAs entre os tecidos tenham sido pequenas ("*fold changes*" menores que dois), sua demonstração através tanto de DEGseq quanto de RT-qPCR confere confiabilidade a estes resultados.

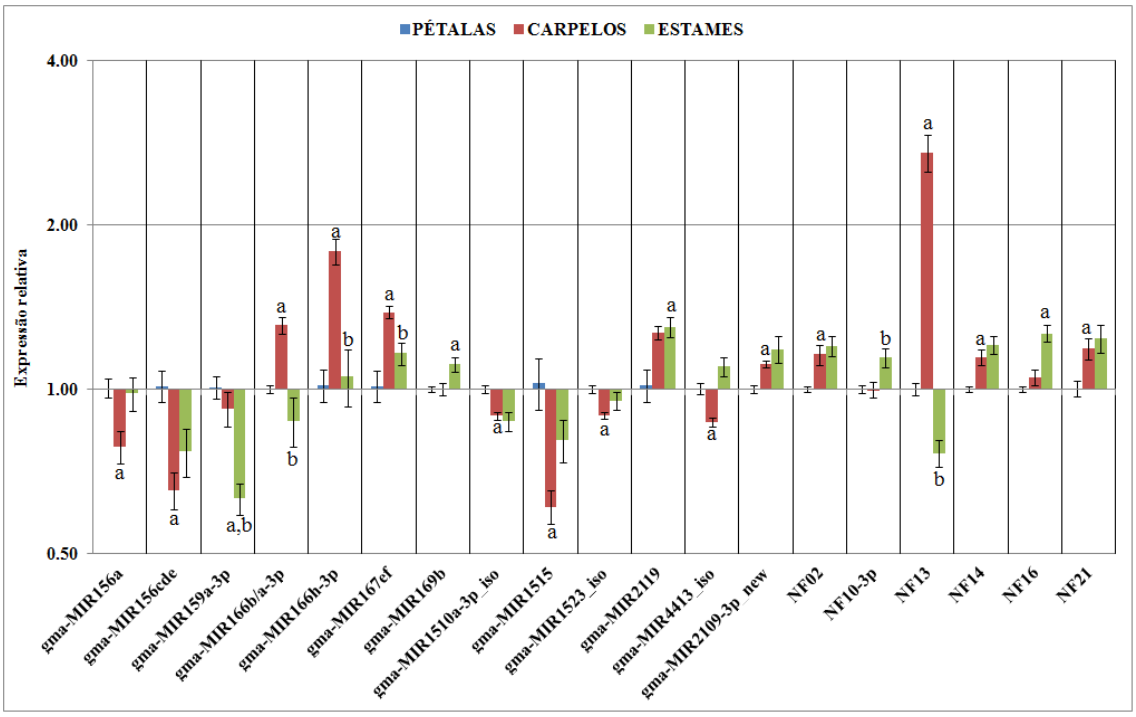


Figura 19. Expressão relativa de miRNAs em três tecidos florais de soja avaliada por RT-qPCR. Amostras que diferem significativamente tanto por RT-qPCR (ANOVA e teste t de Student; P<0,005) e por DEGseq (P<0,001) estão marcadas com: "a" para diferenças entre pétalas e carpelos ou estames; e "b" para diferenças entre carpelos e estames.

4.7 Potenciais genes alvos dos microRNAs com expressão diferencial duplamente comprovada por DEGseq e RT-qPCR

Para os 19 miRNAs com diferenças de expressão demonstradas tanto por DEGseq quanto por RT-qPCR (Figura 19 e Tabela 9), foi feita a predição de alvos computacionalmente. Alguns dos alvos preditos são alvos já descritos para os miRNAs envolvidos no florescimento, muitos dos quais já foram comprovados em soja (Tabela 10).

Para gma-MIR2109-3p_new, que foi induzido em carpelos em comparação com pétalas (Figura 19 e Tabela 9), os alvos preditos foram dois fatores de elongação da tradução. Outros miRNAs que tiveram sua expressão levemente aumentada nos carpelos em relação às pétalas foram NF02, NF14 e NF21 (Figura 19 e Tabela 9). O gene alvo em potencial de NF02 codifica uma O-metiltransferase (Tabela 10), membro de uma família de enzimas envolvida na biossíntese de diversos compostos secundários (Kim *et al*., 2006). NF14 teve um alvo predito sem função anotada (Tabela 10).

Tabela 9. miRNAs diferencialmente expressos nas comparações entre os três órgãos florais de soja. Foram considerados somente os miRNAs que diferiram significativamente entre os tecidos tanto por RT-qPCR (ANOVA e teste t de Student; P<0,05) e por DEGseq (P<0,001). PET: pétalas; CAR: carpelos; EST: estames.

CAR x PET		**EST x PET**		**CAR x EST**	
Induzidos	**Reprimidos**	**Induzidos**	**Reprimidos**	**Induzidos**	**Reprimidos**
gma-MIR166b/a-3p	gma-MIR156a	gma-MIR169b	gma-MIR159a-3p	gma-MIR159a-3p	NF10-3p
gma-MIR166h-3p	gma-MIR156cde	gma-MIR2119		gma-MIR166b/a-3p	
gma-MIR167ef	gma-MIR1510a-3p	NF16		gma-MIR166h-3p	
gma-MIR2109-3p_new	gma-MIR1515			gma-MIR167ef	
NF02	gma-MIR1523			NF13	
NF13	gma-MIR4413				
NF14					
NF21					

Para NF21 foram preditos como alvos dois transcritos codificando proteínas contendo dedos de zinco do tipo DHHC (Tabela 10). Em *Arabidopsis*, o gene *TIP1* (*Tip Growth Defect 1*), contendo estes domínios, foi clonado e caracterizado codificando uma S-acil transferase envolvida na palmitoilação de proteínas que regula o crescimento celular (Hemsley *et al*., 2005). *TIP1* é pouco expresso em flores de *Arabidopsis* (Hemsley *et al*., 2005), mas At5g04270, outro gene que também codifica uma proteína dedo de zinco do tipo DHHC, é altamente expresso em flores (Xiang *et al*., 2010).

Tabela 10. Potenciais alvos dos miRNAs com a expressão diferencial demonstrada por DEGseq e RT-qPCR. Os genes escritos em negrito foram previamente identificados como alvos por métodos experimentais (detalhes no texto).

miRNA	Domínios/Proteína	Função	Genes alvo	Região
gma-MIR156a	HSP40	Chaperona	Glyma13g28560; Glyma15g10560	CDS
	MAPKK	Transdução de sinal	Glyma03g40620	CDS
gma-MIR156cde	SPL	Fator de transcrição	Glyma07g31880; Glyma13g24590	3'UTR
		Não anotada	Glyma06g47590	CDS
gma-MIR156cde e gma-MIR156a	SPL	Fator de transcrição	**Glyma04g37390**; Glyma12g27330; Glyma13g35000	3'UTR
	SPL	Fator de transcrição	**Glyma02g13370;Glyma02g30670; Glyma03g27200;** Glyma03g29900; Glyma05g00200; **Glyma05g38180; Glyma06g17700;Glyma08g01450; Glyma11g36980;Glyma17g08840; Glyma18g36960;Glyma19g26390;** Glyma19g32800	CDS
		Não anotada	Glyma16g05900; Glyma18g00890	CDS
gma-MIR159a-3p	GalBL-GH35	Metabolismo de carboidratos	Glyma12g03650	CDS
	MuDR/Mu transposase; transcriptase reversa	Elemento transponível	Glyma19g25310	CDS
	MYB	Fator de transcrição	Glyma03g26830; Glyma04g15150; Glyma06g47000; Glyma07g14480; Glyma13g04030; **Glyma13g25720; Glyma15g35860; Glyma20g11040**	CDS
	PORR	Fator de *splicing*	Glyma04g10930	3'UTR
	SCRL	Auto-incompatibilidade	Glyma18g51290	3'UTR
		Não anotada	Glyma13g34710; Glyma15g08630; Glyma20g35390	5'UTR
gma-MIR166b/a-3p	HD-START-MEKHLA	Fator de transcrição	**Glyma06g09100**	CDS
gma-MIR166b/a-3p e gma-MIR166h-3p	HD-START-MEKHLA	Fator de transcrição	**Glyma07g01940; Glyma07g01950**	CDS
	HD-ZIP III	Fator de transcrição	**Glyma05g30000;Glyma08g13110; Glyma08g21620;Glyma09g02750; Glyma11g20520;Glyma12g08080; Glyma15g13640**	CDS
gma-MIR166h-3p	Amônia permease	Transporte de amônio	Glyma18g43540	3'UTR
	C1-PHD-finger	Fator de transcrição	Glyma06g11370	CDS
	Esqualeno sintase	Biossíntese de esterol	Glyma08g41890	CDS
gma-MIR167ef	BRD	Fator de transcrição	Glyma06g18070	CDS
	DEAD box-RNA helicase	Helicase	Glyma01g43960; Glyma11g01430	CDS
		Não anotada	Glyma04g10970	3'UTR

gma-MIR169b	ARF-GAP	Transporte de vesículas	Glyma19g35620	CDS
	CBF-B/NF-YA	Fator de transcrição	Glyma02g35190; Glyma02g47380; Glyma07g04050; **Glyma08g45030; Glyma09g07960;** Glyma10g10240; **Glyma13g16770;** Glyma14g01360; **Glyma15g18970;Glyma17g05920; Glyma19g38800 (2)**	3'UTR
gma-MIR1510a-3p_iso	LRR	Proteína de resistência	Glyma06g43850; Glyma19g07660; Glyma12g15850; Glyma12g34020; Glyma13g03770; Glyma13g25440; Glyma13g25950; Glyma13g26250; Glyma13g26310; Glyma16g26270; Glyma16g24940; Glyma18g12030; Glyma18g14810; Glyma19g02670; Glyma19g07680; Glyma19g07650; Glyma19g07700	CDS
	TIR	Proteína de resistência	Glyma04g39740; Glyma12g27800; Glyma16g22620	CDS
		Não anotada	Glyma06g36310	CDS
gma-MIR1515	Aquaporina (MIP)	Transporte de água	Glyma02g08110	3'UTR
	MYB	Fator de transcrição	Glyma08g27660	CDS
	NAM	Fator de transcrição	Glyma07g31220	CDS
	RNAse III	Nuclease de dsRNA	**Glyma09g02920**; Glyma09g02930; Glyma19g44390	CDS
gma-MIR1523_iso	PH-like	Não anotada	Glyma05g04760	5'UTR
	Receptor LRR-RLK	Transdução de sinal	Glyma16g23450; Glyma16g23560; Glyma16g28570	CDS
	UFD1	Ubiquitina	Glyma15g04140	CDS
gma-MIR2109-3p_new	GTPase EF-LepA	Fator de elongação	Glyma18g46900; Glyma09g39400	CDS
gma-MIR2119	Álcool desidrogenase (ADH)	Metabolismo	Glyma13g09530; Glyma14g24860	CDS
	AP2-*like*	Fator de transcrição	Glyma13g37450	CDS
	CYP450	Carreador de elétrons	Glyma20g00740; Glyma20g00750	CDS
	MSF	Transporte transmembrana	Glyma16g34220	CDS
	NER-RNAPII-HEAT-RT	Reparo de DNA/ trancrição reversa	Glyma13g43100	CDS
	Subtilase	Protease	Glyma05g30460; Glyma08g13590; Glyma15g21920	CDS
	Zf-CCHC	Ligante de RNA/DNA	Glyma13g15790	3'UTR
gma-MIR4413_iso NF16	PPRP	Ligante de RNA	Glyma16g28020; Glyma16g25410; Glyma16g27790; Glyma09g07250; Glyma07g11410; Glyma14g38270; Glyma09g30720; Glyma09g30580; Glyma16g32210; Glyma09g07290; Glyma09g39260; Glyma09g30940; Glyma07g11290; Glyma16g31950	CDS
gma-MIR4413_iso	PPRP	Ligante de RNA	Glyma09g07300; Glyma16g27600; Glyma01g44420; Glyma11g01110; Glyma09g30500; Glyma15g24040	CDS
	Quinesina	Movimento ao longo do microtúbulo	Glyma20g37780	CDS

gma-MIR4413_iso	Rhodanese-like PPIC-type PPIase	Chaperona	Glyma14g04000	3'UTR
NF02	O-methyltransferase	Metabolismo secundário	Glyma14g00800	CDS
NF10-3p	ATPase AAA	Diversas funções	Glyma03g27900; Glyma19g30710	CDS
	GH47	Metabolismo de carboidratos	Glyma05g01830	CDS
	Rab-GAP	Transporte de vesículas	Glyma12g34110; Glyma13g36430	CDS
	Receptor LRR-RLK	Transdução de sinal	Glyma02g43150	CDS
	RT-integrase-GAG	Elemento transponível	Glyma04g27590	CDS
		Não anotada	Glyma01g19400	CDS
NF13	AKR	Metabolismo de aldeídos	Glyma08g06840	3'UTR
	M18 aspartil aminopeptidase	Protease	Glyma10g34800; Glyma20g32790	CDS
NF14		Não anotada	Glyma02g47930	CDS
NF16	Carboxil metiltransferase dependent de SAM	Metiltransferase	Glyma08g43460	CDS
	ELFV desidrogenase	Metabolismo de aminoácidos	Glyma19g28770	5'UTR
	Pol ε, subunidade catalítica A	DNA polimerase	Glyma11g36140	CDS
	PPR-LAGLIDADG	Endonuclease de "homing"	Glyma20g01250	CDS
	PPRP	Ligante de RNA	Glyma13g41620; Glyma02g31470; Glyma09g30160; Glyma09g30740; Glyma07g11500; Glyma01g41010; Glyma20g22740; Glyma20g22770; Glyma16g06320	CDS
NF21	Zf-DHHC	S-acil transferase	Glyma03g27410; Glyma19g30360	CDS

Os miRNAs gma-MIR166b/a-3p e gma-MIR166h-3p, membros mais abundantes da família MIR166, foram induzidos nos carpelos em relação aos demais órgãos (Tabela 9; Figura 19). Os mRNAs alvos destes miRNAs codificam fatores de transcrição HD-ZIP de classe III (Tabela 10), já validados em soja por degradoma e/ou 5'RACE (Song *et al*., 2011). Foi reportado que a regulação de dois destes fatores de transcrição (PHB e PHV) mediada por MIR166 é importante para o padrão embrionário em *Arabidospsis* (Grigg *et al*., 2009). Possivelmente, o mesmo mecanismo ocorre nos carpelos de soja, onde os embriões se desenvolvem.

O miRNA gma-MIR167ef também apresentou maior expressão nos carpelos que em pétalas e estames (Tabela 9; Figura 19). Consistente com este resultado, hibridizações *in situ* em *Nicotiana benthamiana* e *Arabidopsis* mostraram que os níveis de MIR167 aumentam após a diferenciação do óvulo, sendo continuamente expresso em sementes prematuras (Válóczi *et al*., 2006). Para gma-MIR167g, um mRNA alvo (Glyma18g05330) foi identificado por 5'RACE (Joshi *et al*., 2010), codificando um fator de resposta à auxina (ARF), similar (78,3%) a ARF8 de Arabidopsis. Os fatores de transcrição ARF6 e ARF8 regulam a expressão de genes de resposta à auxina e são requeridos para a elongação celular em vários órgãos florais e seu domínio espacial de expressão é definido por MIR167 nos estames e óvulos, onde desempenham importantes funções no seu desenvolvimento (Wu *et al*., 2006) e na germinação do pólen (Ru *et al*., 2006). No entanto, para o membro gma-MIR167ef, os genes alvos preditos foram duas helicases de RNA, um fator de transcrição com domínio BRD e um sem função anotada (Tabela 10). Uma helicase de RNA foi previamente predita como alvo para membros desta família de miRNAs em soja (Zeng *et al*., 2010).

Para NF10-3p, que foi fracamente induzido em estames em relação aos carpelos, (Tabela 9; Figura 19), foram preditos oito mRNAs, dois codificando para ATPases da família AAA, dois para ativadores de GTPase envolvidos no transporte de vesículas, um para um receptor quinase, um para uma glicosil hidrolase e outra codificando alguns domínios característicos de retrovírus e TEs.

Gma-MIR2119 foi induzido nos estames em relação a pétalas (Tabela 9; Figura 19) e possui diversos alvos em potencial diferentes (Tabela 10). Além de alvos já descritos,

como álcool desidrogenases (ADH), em *M.truncatula* (Devers *et al*., 2011), foram preditos dois transcritos alvos que codificam o citocromo P450 (CYP450). As plantas possuem grande número de diferentes enzimas CYP450 para a síntese de metabólitos secundários, muitos dos quais são importantes pigmentos florais (Holton *et al*., 1993). Em arroz, foi identificado um membro da família de CYP450 (CYP704B2) que é requerido para a biossíntese de cutina na antera e para a formação da exina no pólen durante o desenvolvimento reprodutivo vegetal masculino. Uma mutação no gene que codifica este citocromo causa esterilidade masculina (Li *et al*., 2010).

O miRNA gma-MIR169b também foi levemente induzido em estames em relação às pétalas (Tabela 9; Figura 19). Este miRNA possui sítios de clivagem na região 3'UTR de fatores de transcrição NF-YA (Tabela 10), os quais já foram validados em soja por análise de degradoma (Song *et al*., 2011). Em *Arabidopsis*, NF-Y promove a transcrição de genes da classe C (ex.: *AG*), necessários ao desenvolvimento dos estames e carpelos (Hong *et al*., 2003). Portanto, este miRNA possivelmente modula os níveis de NF-Y, e consequentemente, de proteínas da classe C nos estames de soja. Além destes alvos clássicos da família MIR169, foi predito como alvo para gma-MIR169b, um mRNA que codifica uma proteína ativadora de GTPase (ARF-GAP), envolvida no transporte de vesículas.

Dois membros da família MIR156, gma-MIR156a e gma-MIR156cde, foram reprimidos nos carpelos (Tabela 9; Figura 19). Os genes alvos de MIR156 codificam fatores de transcrição SPL que agem na transição da fase vegetativa para a reprodutiva (Gandikota *et al*., 2007; Yamaguchi *et al*., 2009). Foram preditos como alvos destes miRNAs, 18 mRNAs codificando SPLs (Tabela 10), dos quais 11 já haviam sido identificados no degradoma de soja (Song *et al*., 2011). Em *Arabidopsis*, *SPOROCYTELESS*, um gene de SPL, é um alvo direto de *AG*, gene de função C. Portanto, este gene deve ser expresso nos carpelos e estames (Ito *et al*., 2004). Além dos alvos clássicos de MIR156, nesta análise, também foram preditos como alvos, dois genes codificando proteínas *heat shock* e um codificando uma proteína quinase 2 ativada por mitógeno (MAPKK) (Tabela 10).

O miRNA gma-MIR159a-3p foi reprimido nos estames em comparação com pétalas e carpelos (Tabela 9; Figura19) e seus alvos são fatores de transcrição MYB

(Tabela 10). Este padrão de expressão é esperado, visto que este miRNA, com exatamente a mesma sequência nucleotídica, é expresso em todos os órgãos florais de *Arabidopsis*, exceto nas anteras, pois sua ausência permite a expressão de MYB33 e MYB65, que são necessários ao desenvolvimento destes órgãos (Allen *et al*., 2007). Por isso, a superexpressão de MIR159 em plantas transgênicas de *Arabidopsis* comprometeu a fertilidade masculina devido à formação de anteras indeiscentes (Achard *et al*., 2004; Schwab *et al*., 2005). Outros alvos também foram encontrados para gma-MIR159a-3p, incluindo um mRNA que codifica uma proteína de resposta à auto-incompatibiliadade.

Gma-MIR1515 foi reprimido em carpelos em relação a pétalas (Tabela 9; Figura 19) e seus potenciais alvos incluem fatores de transcrição MYB e um NAM (No Apical Meristem), uma aquaporina e três mRNAs codificando RNAses do tipo III (Tabela 10). Para um dos mRNAs de RNAse III, a clivagem foi previamente comprovada em soja por 5'RACE (Li *et al*., 2010). Gma-MIR1523 também foi reprimido nos carpelos na comparação com pétalas (Tabela 9; Figura 19) e seus alvos codificam três receptores quinase, uma ubiquitina e uma plecstrina.

Os miRNAs gma-MIR4413_iso e NF16 apresentaram padrões de expressão distintos, o primeiro sendo reprimido nos carpelos e o segundo, induzido nos estames (Tabela 9; Figura 19). Para ambos os miRNAs foram preditos mRNAs alvos em comum (Tabela 10), que codificam proteínas com repetições pentatricopeptídicas (PPR). No entanto, estes miRNAs se ligam a sítios alvo em regiões diferentes dos mesmos mRNAs, com os sítios alvo de gma-MIR4413_iso localizados abaixo dos sítios alvo de NF16. A família de proteínas PPR é muito numerosa, sendo regulada por mais de um miRNA e também por siRNAs, provavelmente para minimizar o número de cópias ativas de PPR e suprimir a dosagem gênica (Schmitz-Linneweber e Small, 2008). Dentre as diversas funções que estas proteínas desempenham em plantas, algumas estão relacionadas ao desenvolvimento reprodutivo. Um exemplo é a restauração da fertilidade masculina pela supressão da expressão de genes mitocondriais associados com a esterilidade citoplasmática (Bentolila *et al*., 2002; Brown *et al*., 2003; Kazama e Toriyama, 2003). Além disso, o estudo de mutantes *knockout* de genes *pprp* revelou o papel essencial de pelo menos 7 PPRPs na embriogênese de *Arabidopsis* (Cushing *et al*., 2005).

A grande família de proteínas de resistência a doenças NB-LRR também é regulada por mais de um miRNA. Um estudo recente mostrou como um pequeno número de miRNAs são capazes de regular esta família gênica tão numerosa nos genomas de plantas (Zhai *et al*., 2011). Entre os miRNAs reguladores desta família, está MIR1510, como predito em *Phaseolus vulgaris* (Valdés-López *et al*., 2010) e soja (Wong *et al*., 2011; Kulcheski *et al*., 2011). No presente estudo, vinte membros desta família de proteínas de resistência foram preditos como mRNAs alvos de gma-MIR1510a-3p_iso (Tabela 10), que foi levemente reprimido nos carpelos em relação às pétalas (Tabela 9; Figura 19). Muitos dos genes alvos preditos foram previamente identificados como genes *PHAS*, cujos transcritos produzem phasiRNAs (Zhai *et al*., 2011).

O miRNA inédito NF13 foi o que apresentou maior diferença de expressão entre os órgãos florais, sendo fortemente induzido nos carpelos em todas as comparações (Tabela 9; Figura 19). Os alvos preditos para este miRNA são dois transcritos que codificam duas enzimas aspartil aminopeptidases e uma aldo-ceto redutase (AKR) (Tabela 10). AKRs têm sido relacionadas à detoxificação de aldeídos reativos em diversas plantas (Simpson *et al*., 2009; Yamauchi *et al*., 2011; Turóczy *et al*., 2011; Narawongsanont *et al*., 2012). O gene de ARK caracterizado em soja, *GmARK1*, apresenta expressão raiz-específica (Hur *et al*., 2009). Quanto às aspartil aminopeptidases, uma enzima desta família já foi isolada e caracterizada em cotilédones de soja e parece estar envolvida na germinação (Asano *et al*., 2010), como em *Pinus banksiana*, no qual apresentou alta atividade nas sementes secas (Bourgeois e Malek, 1991). Como os níveis de NF13 foram fortemente aumentados nos carpelos, este miRNA pode estar regulando negativamente esta enzima nas sementes em formação, prevenindo sua expressão até o momento da germinação. Este modelo estaria de acordo com a proposta de Nodine e Bartel (2010), de que, em plantas, os miRNAs regulam as transições no desenvolvimento pela repressão de mRNAs que devem agir posteriormente. Nos embriões de *Arabidopsis* no estádio de oito células, por exemplo, foi verificado que miRNAs reprimem genes que devem ser expressos nas células filhas, em um momento posterior do desenvolvimento do embrião. Portanto, em vez de atenuar funções pré-existentes, como ocorre em animais, os miRNAs previnem a ação precoce de genes em plantas (Nodine e Bartel, 2010).

Embora muitos dos genes preditos como alvos já tenham tido a clivagem comprovada experimentalmente, por análise de degradoma ou 5'RACE, muitos ainda precisam ser validados, especialmente os alvos preditos para os miRNAs inéditos.

5 CONCLUSÕES E PERSPECTIVAS

Os baixos níveis de expressão dos miRNAs espécie e tecido-específicos dificultam sua detecção por métodos convencionais. Embora centenas de miRNAs já tenham sido descritos em soja, inclusive em flor (Joshi *et al*., 2010), miRNAs frequentes em órgãos pequenos como carpelos e estames acabam não sendo detectados por um efeito de diluição no RNA total quando flores inteiras são amostradas. No presente estudo, o sequenciamento de bibliotecas de tecidos florais de soja por HTS possibilitou a descoberta de 51 miRNAs inéditos pertencentes a 40 novas famílias de miRNAs específicos de soja. Também foi possível detectar 45 isoformas de miRNAs conhecidos e 33 novos membros de famílias conhecidas de miRNAs de plantas, incluindo seis pertencentes a três famílias que foram descritas pela primeira vez em soja. Além disso, este trabalho forneceu validação adicional da precisão das sequências de 62 miRNAs maduros anotados no miRBase para soja.

As análises de expressão nas bibliotecas de sRNAs por DEGseq mostraram um acúmulo de miRNAs de 22 e 24 nt induzidos nos carpelos. Este resultado sugere que possivelmente há algum mecanismo que favoreça a produção e/ou a estabilidade de sRNAs destes tamanhos, os quais devem desempenhar papéis específicos no desenvolvimento reprodutivo nestes órgãos.

A compreensão dos mecanismos fisiológicos, bioquímicos e genéticos dos processos reprodutivos deve tornar possível o desenvolvimento de plantas de soja com maior produtividade de grãos. Os resultados desta dissertação demonstraram que diversos miRNAs são diferencialmente expressos entre os órgãos florais desta espécie. Para um melhor entendimento sobre o papel destes miRNAs nestes órgãos, podem ser realizadas análises de expressão de seus genes alvos e confirmação de seu silenciamento por metodologias experimentais, como a detecção dos produtos de clivagem por 5'RACE e sequenciamento do degradoma de RNAs. Uma perspectiva para estudos posteriores é a investigação do papel do miRNA inédito denominado NF13, visto que apresentou a maior diferença de expressão entre os miRNAs testados por RT-qPCR, sendo fortemente induzido nos carpelos, onde são formados os grãos.

Os níveis de miRNAs maduros detectados neste estudo são o resultado final, no momento da extração de RNA, do efeito combinatório de sua transcrição, processamento, carregamento em RISC, reconhecimento e regulação do mRNA alvo, sua degradação e reciclagem (Meng *et al*., 2011). Além disso, o silenciamento gênico guiado pelos miRNAs

envolve eliminar transcritos de processos de desenvolvimento anteriores, modular a expressão de genes ativos nos processos presentes (Borges *et al*., 2011) e prevenir a expressão de genes que devem ser expressos posteriormente (Nodine e Bartel, 2010).

Embora não tenham sido amostrados diferentes estádios de desenvolvimento dos tecidos florais analisados, inferências sobre a função de miRNAs na regulação de processos florais puderam ser feitas a partir destes dados, visto que a expressão de muitos genes envolvidos no desenvolvimento floral persiste na flor madura. Por exemplo, *AG*, além de ser expresso no primórdio floral para especificar a formação dos órgãos reprodutivos nos verticilos internos, também se expressa em estádios mais avançados, em regiões distintas destes órgãos, sugerindo que *AG* também está envolvido na maturação de estames e carpelos (Ito *et al*., 2004). O gene *GmAP1* de soja (Chi *et al*., 2011), que codifica uma proteína semelhante a AP1, que regula o tempo de florescimento e especifica a formação dos órgãos florais vegetativos em *Arabidopsis*, foi expresso principalmente em sépalas e pétalas, mesmo obtidas de flores de soja já maduras. Em estudo de transcriptoma de sépalas, pétalas, estames e carpelos de papoula (*Eschscholzia californica*), apesar de os órgãos florais já estarem desenvolvidos, a expressão de fatores de transcrição MADS-box esteve de acordo com o modelo ABC proposto para o desenvolvimento destes órgãos (Yu *et al*., 2011). Por exemplo, homólogos de *PI* e *AP3* (genes da classe B) foram mais expressos em pétalas e estames e o homólogo de *AG* (gene da classe C) foi mais expresso em estames e carpelos.

Por isso, seria interessante investigar a expressão dos miRNAs diferencialmente expressos e seus genes alvos em diferentes pontos do tempo, ao longo do desenvolvimento da flor e da semente, em particular através da utilização da metodologia de hibridação *in situ*. Estes estudos poderão ampliar o entendimento sobre a dinâmica da regulação dos níveis destes miRNAs e dos seus genes alvos e sua importância para os processos reprodutivos da soja.

6 REFERÊNCIAS BIBLIOGRÁFICAS

Achard P, Herr A, Baulcombe DC, Harberd NP (2004) Modulation of floral development by a gibberellin-regulated microRNA. Development 131: 3357–3365.

Allen E. e Howell MD. (2010) miRNAs in the biogenesis of trans-acting siRNAs in higher plants. Semin Cell Dev Biol 21: 798-804.

Allen RS, Li J, Stahle MI, Dubroué A, Gubler F, Millar AA (2007) Genetic analysis reveals functional redundancy and the major target genes of the Arabidopsis miR159 family. Proc Natl Acad Sci U S A 104: 16371-16376.

Andersen CL, Jensen JL, Orntoft TF (2004) Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. Cancer research 64: 5245–5250.

Asano M, Suzuki S, Kawai M, Miwa T, Shibai H (1999) Characterization of Novel Cysteine Proteases from Germinating Cotyledons of Soybean [Glycine max (L.) Merrill]. J. Biochem. 126: 296-301.

Asano M, Nakamura N, Kawai M, Miwa T, Nio N (2010) Purification and Characterization of an N-Terminal Acidic Amino Acid-Specific Aminopeptidase from Soybean Cotyledons (*Glycine max*). Biosci Biotechnol Biochem 74: 113–118.

Ashlock L, Purcell L. Growth and Development. In: Arkansas Soybean handbook. pp. 7–12.

Aukerman M, Sakai H (2003) Regulation of flowering time and floral organ identity by a microRNA and its APETALA2-1ike target genes. Plant Cell 15: 2730–2741.

Baker CC, Sieber P, Wellmer F, Meyerowitz EM (2005) The *early extra petals1* mutant uncovers a role for microRNA miR164c in regulating petal number in Arabidopsis. Current Biology 15: 303–315.

Bentolila S, Alfonso AA, Hanson MR (2002) A pentatricopeptide repeat-containing gene restores fertility to cytoplasmic male-sterile plants. Proc Natl Acad Sci U S A 99: 10887–10892.

Bloom JS, Khan Z, Kruglyak L, Singh M, Caudy AA (2009) Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. BMC genomics 10: 221.

Borges F, Pereira PA, Slotkin RK, Martienssen RA, Becker JD (2011) MicroRNA activity in the Arabidopsis male germline. J Exp Bot 62: 1611–1620.

Bourgeois J, Malek L (1991) Purification and characterization of an aspartyl proteinase from dry jack pine seeds. Seed Sci Res 1: 139–147.

Brown GG, Formanová N, Jin H, Wargachuk R, Dendy C, Patil P, Laforest M, Zhang J, Cheung WY, Landry BS (2003) The radish Rfo restorer gene of Ogura cytoplasmic male

sterility encodes a protein with multiple pentatricopeptide repeats. The Plant journal 35: 262–272.

Cannon S (2008) Legume Comparative Genomics. In: Stacey G, editor. Genetics and Genomics of Soybean. Springer Verlag.

Causier B, Schwarz-Sommer Z, Davies B (2010) Floral organ identity: 20 years of ABCs. Semin Cell Dev Biol 21: 73–79.

Chambers C, Shuai B (2009) Profiling microRNA expression in Arabidopsis pollen using microRNA array and real-time PCR. BMC Plant Biol 9: 87.

Chen X (2004) A microRNA as a translational repressor of APETALA2 in Arabidopsis flower development. Science 303: 2022–2025.

Chen C, Ridzon DA, Broomer AJ, Zhou Z, Lee DH, Nguyen JT, Barbisin M, Xu NL, Mahuvakar VR, Andersen MR *et al*. (2005) Real-time quantification of microRNAs by stem-loop RT-PCR. Nucleic Acids Res 33: e179.

Chen H-M, Chen L-T, Patel K, Li Y-H, Baulcombe DC, Wu S-H (2010) 22-Nucleotide RNAs trigger secondary siRNA biogenesis in plants. Proc Natl Acad Sci U S A 107: 15269–15274.

Chen L, Wang T, Zhao M, Tian Q, Zhang W-H (2012a) Identification of aluminum-responsive microRNAs in *Medicago truncatula* by genome-wide high-throughput sequencing. Planta 235: 375–386.

Chen L, Wang T, Zhao M, Zhang W (2012b) Ethylene-responsive miRNAs in roots of *Medicago truncatula* identified by high-throughput sequencing at whole genome level. Plant Science 184: 14–19.

Chen L, Zhang Y, Ren Y, Xu J, Zhang Z, Wang Y (2011) Genome-wide identification of cold-responsive and new microRNAs in *Populus tomentosa* by high-throughput sequencing. Biochem Biophys Res Commun 417: 892–896.

Chen R, Hu Z, Zhang H (2009) Identification of microRNAs in wild soybean (*Glycine soja*). J Integr Plant Biol 51: 1071–1079.

Chi Y, Huang F, Liu H, Yang S, Yu D (2011) An APETALA1-like gene of soybean regulates flowering time and specifies floral organs. Journal of plant physiology 168: 2251–2259.

Cregan PB (2008) Soybean molecular genetic diversity. In: Stacey G, editor. Genetics and Genomics of Soybean. Springer Verlag. pp. 17–34.

Cushing DA, Forsthoefel NR, Gestaut DR, Vernon DM (2005) Arabidopsis emb175 and other ppr knockout mutants reveal essential roles for pentatricopeptide repeat (PPR) proteins in plant embryogenesis. Planta 221: 424–436.

Dai X, Zhao PX (2011) psRNATarget: a plant small RNA target analysis server. Nucleic Acids Res 39: W155–W159.

Dall´Agnol A, Hirakuri MH (2008) Realidade e perspectivas do Brasil na produção de alimentos e agroenergia, com ênfase na soja. Embrapa soja - Circular técnica 59: 1-8

Devers E a, Branscheid A, May P, Krajinski F (2011) Stars and symbiosis: microRNA- and microRNA*-mediated transcript cleavage involved in arbuscular mycorrhizal symbiosis. Plant physiology 156: 1990–2010.

Dornelas MC, Patreze CM, Angenent GC, Immink RGH (2010) MADS: the missing link between identity and growth? Trends Plant Sci. 16: 89-97.

Ebhardt HA, Fedynak A, Fahlman RP (2010) Naturally occurring variations in sequence length creates microRNA isoforms that differ in argonaute effector complex specificity. Silence 1: 1-6.

Fornara F, de Montaigu A, Coupland G (2010) SnapShot: Control of Flowering in Arabidopsis. Cell 141: 550–550.e2.

Fu C, Sunkar R, Zhou C, Shen H, Zhang J-Y, Matts J, Wolf J, Mann DGJ, Stewart CN, Tang Y, Wang Z-Y. (2012) Overexpression of miR156 in switchgrass (*Panicum virgatum* L.) results in various morphological alterations and leads to improved biomass production. Plant biotechnol J 2010:1-10

Gandikota M, Birkenbihl RP, Höhmann S, Cardon GH, Saedler H, Huijser P (2007) The miRNA156/157 recognition element in the 3' UTR of the Arabidopsis SBP box gene *SPL3* prevents early flowering by translational inhibition in seedlings. Plant J 49: 683–693.

Garcia D (2008) A miRacle in plant development: role of microRNAs in cell differentiation and patterning. Semin Cell Dev Biol 19: 586–595.

Git A, Dvinge H, Osborne M, Kutter C, Hadfield J, Bertone P, Caldas C (2010) Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression. Cancer Research 16:991–1006.

Glazińska P, Zienkiewicz A, Wojciechowski W, Kopcewicz J (2009) The putative miR172 target gene In *APETALA2-like* is involved in the photoperiodic flower induction of *Ipomoea nil*. J Plant Physiol 166: 1801–1813.

Grant-Downton R, Hafidh S, Twell D, Dickinson HG (2009) Small RNA pathways are present and functional in the angiosperm male gametophyte. Mol Plant 2: 500–512.

Grant-Downton R, Le Trionnaire G, Schmid R, Rodriguez-Enriquez J, Hafidh S, Mehdi S, Twell D, Dickinson H (2009) MicroRNA and tasiRNA diversity in mature pollen of Arabidopsis thaliana. BMC Genomics 10: 643.

Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ (2008) miRBase: tools for microRNA genomics. Nucleic Acids Res 36: 154–158.

Grigg SP, Galinha C, Kornet N, Canales C, Scheres B, Tsiantis M (2009) Repression of apical homeobox genes is required for embryonic root development in Arabidopsis. Current biology 19: 1485–1490.

Guo L, Lu Z (2010) Global expression analysis of miRNA gene cluster and family based on isomiRs from deep sequencing data. Comput Biol Chem 34: 165-71.

Hemsley PA, Kemp AC, Grierson CS (2005) The TIP GROWTH DEFECTIVE1 S-acyl transferase regulates plant cell growth in Arabidopsis. Plant cell 17: 2554–2563.

Holton TA, Brugliera F, Lester DR, Tanaka Y, Hyland CD, Menting JG, Lu CY, Farcy E, Stevenson TW, Cornish EC (1993) Cloning and expression of cytochrome P450 genes controlling flower colour. Nature, 366, 276-9.

Hong RL, Hamaguchi L, Busch MA, Weigel D (2003) Regulatory elements of the floral homeotic gene AGAMOUS identified by phylogenetic footprinting and shadowing. Plant Cell 15: 1296-1309.

Huang F, Chi Y, Gai J, Yu D (2009) Identification of transcription factors predominantly expressed in soybean flowers and characterization of *GmSEP1* encoding a SEPALLATA1-like protein. Gene 438: 40–48.

Hur Y-S, Shin K-H, Kim S, Nam KH, Lee M-S, Chun J-Y, Cheon C-I (2009) Overexpression of *GmAKR1*, a stress-induced aldo/keto reductase from soybean, retards nodule development. Mol Cells 27: 217–223.

Immink RGH, Kaufmann K, Angenent GC (2010) The "ABC" of MADS domain protein behaviour and interactions. Semin Cell Dev Biol 21: 87–93.

Irish VF (2010) The flowering of Arabidopsis flower development. Plant J. 61: 1014–1028.

Ito T, Wellmer F, Yu H, Das P, Ito N, Alves-Ferreira M, Riechmann JL, Meyerowitz E (2004) The homeotic protein AGAMOUS controls microsporogenesis by regulation of SPOROCYTELESS. Nature 430: 356–360.

Joshi T, Yan Z, Libault M, Jeong D-H, Park S, Green PJ, Sherrier DJ, Farmer A, May G, Meyers BC (2010) Prediction of novel miRNAs and associated target genes in *Glycine max*. BMC Bbioinformatics 11 (Suppl 1): S14.

Jung J-H, Park C-M (2007) MIR166/165 genes exhibit dynamic expression patterns in regulating shoot apical meristem and floral development in Arabidopsis. Planta 225: 1327–1338.

Jung J-H, Seo Y-H, Seo PJ, Reyes JL, Yun J, Chuab N-H, Park C-M. (2007) The GIGANTEA-regulated microRNA172 mediates photoperiodic flowering independent of CONSTANS in Arabidopsis. Plant Cell 19: 2736–2748.

Kai ZS, Pasquinelli AE (2010) MicroRNA assassins: factors that regulate the disappearance of miRNAs. Nat Struct Mol Biol 17: 5–10.

Kaufmann K, Wellmer F, Muiño JM, Ferrier T, Wuest SE, et al. (2010) Orchestration of floral initiation by APETALA1. Science 328: 85–89.

Kazama T, Toriyama K (2003) A pentatricopeptide repeat-containing gene that promotes the processing of aberrant atp6 RNA of cytoplasmic male-sterile rice. FEBS letters 544: 99–102.

Kidner C A, Martienssen R A (2005) The developmental role of microRNA in plants. Curr Opin Plant Biol 8: 38–44.

Kim BG, Lee HJ, Park Y, Lim Y, Ahn J-H (2006) Characterization of an O-methyltransferase from soybean. Plant physiology and biochemistry 44: 236–241.

Kodera T, Asano M, Kawai M, Miwa T, Nio N (2005) The Effective Methods in Refolding and Activation of Cathepsin L-like Soybean Protease D3. Journal of Food Science 70: 495-502.

Kozomara A, Griffiths-Jones S (2011) miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res 39: 152–157.

Kulcheski F, de Oliveira L, Molina L, Almerao M, Rodrigues F, Marcolino J, Barbosa JF, Stolf-Moreira F, Nepomuceno AL, Marcelino-Guimaraes FC. (2011) Identification of novel soybean microRNAs involved in abiotic and biotic stresses. BMC Genomics 12: 307.

Laufs P, Peaucelle A, Morin H, Traas J (2004) MicroRNA regulation of the *CUC* genes is required for boundary size control in Arabidopsis meristems. Development (Cambridge, England) 131: 4311–4322.

Lee RC, Feinbaum RL, Ambros V (1993) The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 75: 843–854.

Li H, Deng Y, Wu T, Subramanian S, Yu O (2010) Mis-expression of miR482, miR1512, and miR1515 increases soybean nodulation. Plant Physiology 153: 1759-70.

Li H, Dong Y, Sun Y, Zhu E, Yang J, Liu X, Xue P, Xiao Y, Yang S, Wu J, *et al*. (2011) Investigation of the microRNAs in safflower seed, leaf, and petal by high-throughput sequencing. Planta 233: 611–619.

Li H, Dong Y, Yin H, Wang N, Yang J, Liu X, Wang Y, Wu J, Li X (2011) Characterization of the stress associated microRNAs in *Glycine max* by deep sequencing. BMC Plant Biol 11: 170.

Li R, Li Y, Kristiansen K, Wang J (2008) SOAP: short oligonucleotide alignment program. Bioinformatics (Oxford, England) 24: 713–714.

Lima JC, Morais GL, Margis R. Plant microRNAs as central regulators during development and abiotic stress responses. Artigo submetido Plant Biol (2011).

Liu C, Xi W, Shen L, Tan C, Yu H (2009) Regulation of floral patterning by flowering time genes. Dev Cell 16: 711–722.

Liu Q, Chen Y-Q (2009) Insights into the mechanism of plant development: interactions of miRNAs pathway with phytohormone response. Biochem Biophys Res Commun 384: 1–5.

Liu X, Huang J, Wang Y, Khanna K, Xie Z, Owen HA, Zhao D (2010) The role of floral organs in carpels, an Arabidopsis loss-of-function mutation in microRNA160a, in organogenesis and the mechanism regulating its expression. Plant J 62: 416–428.

Liu Z, Mara C (2010) Regulatory mechanisms for floral homeotic gene expression. Semin Cell Dev Biol 21: 80–86.

Livak KJ, Schmittgen TD (2001) Analysis of relative gene expression data using Real-Time Quantitative PCR and the $2^{-\Delta\Delta CT}$ method. Gene Expr 25: 402–408.

Mallory A, Dugas D, Bartel D, Bartel B (2004) MicroRNA regulation of NAC-domain targets is required for proper formation and separation of adjacent embryonic, vegetative, and floral organs. Current Biology 14: 1035–1046.

McGregor S (1976) Legumes and some relatives. In: Insect Pollination Of Cultivated Crop Plants (USDA).

McKim S, Hay A (2010) Patterning and evolution of floral structures - marking time. Curr Opin Genet Dev 20: 448–453.

Meng Y, Shao C, Wang H, Chen M (2011) The regulatory activities of plant microRNAs: a more dynamic perspective. Plant Physiology 157: 1583-1595.

Mi S, Cai T, Hu Y, Chen Y, Hodges E, Ni F, Wu L, Li S, Zhou H, Long C. *et al*. (2008) Sorting of small RNAs into Arabidopsis Argonaute complexes is directed by the 5' terminal nucleotide. Cell 133: 116–127.

Mohorianu I, Schwach F, Jing R, Lopez-Gomollon S, Moxon S, Szittya G, Sorefan K, Moulton V, Dalmay T (2011) Profiling of short RNAs during fleshy fruit development reveals stage-specific sRNAome expression patterns. Plant J 67: 232–246.

Moxon S, Schwach F, Dalmay T, Maclean D, Studholme DJ, Moulton V (2008) A toolkit for analysing large-scale plant small RNA datasets. Bioinformatics (Oxford, England) 24: 2252–2253.

Moyroud E, Kusters E, Monniaux M, Koes R, Parcy F (2010) LEAFY blossoms. Trends Plant Sci 15: 346–352.

Nag A, Jack T (2010) Sculpting the flower; the role of microRNAs in flower development. Plant Development 91: 349–377.

Nag A, King S, Jack T (2009) miR319a targeting of *TCP4* is critical for petal growth and development in Arabidopsis. Proc Natl Acad Sci U S A 106: 22534–22539.

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320: 1344–1349.

Naqvi AR, Sarwat M, Hasan S, Choudhury NR (2012) Biogenesis, functions and fate of plant microRNAs. Journal of Cellular Physiology. Jan: 1–23.

Narawongsanont R, Kabinpong S, Auiyawong B, Tantitadapitak C (2012) Cloning and characterization of AKR4C14, a rice aldo-keto reductase, from Thai Jasmine rice. Protein J 31: 35–42.

Nodine MD, Bartel DP (2010) MicroRNAs prevent precocious gene expression and enable pattern formation during plant embryogenesis. Genes Dev 24: 2678–2692.

Nozawa M, Miura S, Nei M (2012) Origins and evolution of microRNA genes in plant species. Genome Biol Evol 81: 1–35.

Olmedo-Monfil V, Duran-Figueroa N, Arteaga-Vázquez M, Demesa-Arévalo E, Autran D, et al. (2010) Control of female gamete formation by a small RNA pathway in Arabidopsis. Nature 464: 628–634.

Orf J (2010) Introduction. In: Bilyeu K, Ratnaparkhe MB, Kole C, editors. Genetics, genomics, and breeding of soybeans. Science Publishers, Inc. pp. 1–18.

Park W, Li J, Song R, Messing J, Chen X (2002) CARPEL FACTORY, a Dicer homolog, and HEN1, a novel protein, act in microRNA metabolism in Arabidopsis thaliana. Current Biology. 12: 1484–1495.

Pasquinelli AE, Reinhart BJ, Slack F, Martindale MQ, Kuroda MI, et al. (2000) Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. Nature 408: 86–89.

Pedersen P, Kumudini S, Board J, Conley S (2007) Soybean Growth and Development. In: Dorrance AE, Draper MA, Hershman DE, editors. Using foliar fungicides to manage soybean rust. Columbus, OH. pp. 41–47.

Peng T, Lv Q, Zhang J, Li J, Du Y, Zhao Q (2011) Differential expression of the microRNAs in superior and inferior spikelets in rice (*Oryza sativa*). J Exp Bot 62: 4943–4954.

Poethig RS (2009) Small RNAs and developmental timing in plants. Curr Opin Genet Dev 19: 374–378.

Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, et al. (2000) The 21-nucleotide let-7 RNA regulates developmental timing in *Caenorhabditis elegans*. Nature 403: 901–906.

Ru P, Xu L, Ma H, Huang H (2006) Plant fertility defects induced by the enhanced expression of microRNA167. Cell research 16: 457–465.

Sablowski R (2010) Genes and functions controlled by floral organ identity genes. Semin Cell Dev Biol 21: 94–99.

Scaboo AM, Chen P, Sleper DA, Clark KM (2010) Classical Breeding and Genetics of Soybean. In: Bilyeu K, Ratnaparkhe MB, Kole C, editors. Genetics, genomics, and breeding of soybeans. Science Publishers, Inc. pp. 19–54.

Schiefelbein J, Galway M, Masucci J, Ford S (1993) Pollen tube and root-hair tip growth is disrupted in a mutant of *Arabidopsis thaliana*. Plant Physiology 103: 979–985.

Schmitz-Linneweber C, Small I (2008) Pentatricopeptide repeat proteins: a socket set for organelle gene expression. Trends Plant Sci 13: 663–670.

Schwab R, Palatnik JF, Riester M, Schommer C, Schmid M, Weigel D. (2005) Specific effects of microRNAs on the plant transcriptome. Dev Cell 8: 517–527.

Schwach F, Moxon S, Moulton V, Dalmay T (2009) Deciphering the diversity of small RNAs in plants: the long and short of it. Brief Funct Genomic Proteomic 8: 472–481.

Simpson PJ, Tantitadapitak C, Reed AM, Mather OC, Bunce CM, White SA, Ride JP (2009) Characterization of two novel aldo-keto reductases from Arabidopsis: expression patterns, broad substrate specificity, and an open active-site structure suggest a role in toxicant metabolism following stress. Journal of Molecular Biology 392: 465–480.

Song C, Wang C, Zhang C, Korir NKK, Yu H, Ma Z, Fan J (2010) Deep sequencing discovery of novel and conserved microRNAs in trifoliate orange (*Citrus trifoliata*). BMC Genomics 11: 431.

Song Q-X, Liu Y-F, Hu X-Y, Zhang W-K, Ma B, Chen S-Y, Zhang JS (2011) Identification of miRNAs and their target genes in developing soybean seeds by deep sequencing. BMC Plant Biol 11: 1-16.

Subramanian S, Fu Y, Sunkar R, Barbazuk WB, Zhu J-K, Yu O (2008) Novel and nodulation-regulated microRNAs in soybean roots. BMC Genomics 9: 1-14.

Sunkar R, Jagadeeswaran G (2008) In silico identification of conserved microRNAs in large number of diverse plant species. BMC Plant Biol 8: 1-13.

Terzi LC, Simpson GG (2008) Regulation of flowering time by RNA processing. Curr Top Microbiol Immunol 326: 201–218.

Tucker MR, Okada T, Hu Y, Scholefield A, Taylor JM, Koltunow AMG (2012) Somatic small RNA pathways promote the mitotic events of megagametogenesis during female reproductive development in Arabidopsis. Development (Cambridge, England) 139: 1399–1404.

Turóczy Z, Kis P, Török K, Cserháti M, Lendvai A, Dudits D, Horváth GV (2011) Overproduction of a rice aldo-keto reductase increases oxidative and heat stress tolerance by malondialdehyde and methylglyoxal detoxification. Plant Mol Biol 75: 399–412.

Unver T, Namuth-Covert DM, Budak H (2009) Review of current methodological approaches for characterizing microRNAs in plants. Int J Plant Genomics 2009:1-11.

Valdés-López O, Yang SS, Aparicio-Fabre R, Graham PH, Reyes JL, et al. (2010) MicroRNA expression profile in common bean (*Phaseolus vulgaris*) under nutrient deficiency stresses and manganese toxicity. New Phytol 187: 805–818.

Válóczi A, Várallyay E, Kauppinen S, Burgyán J, Havelda Z (2006) Spatio-temporal accumulation of microRNAs is highly coordinated in developing plant tissues. Plant J 47: 140–151.

van der Burgt A, Fiers MWJE, Nap J-P, van Ham RCHJ (2009) In silico miRNA prediction in metazoan genomes: balancing between sensitivity and specificity. BMC Genomics 10: 204.

Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, de Paepe A, Speleman F (2002) Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biol 3: 1-12.

Vaucheret H, Vazquez F, Crété P, Bartel DP (2004) The action of ARGONAUTE1 in the miRNA pathway and its regulation by the miRNA pathway are crucial for plant development. Gen Dev 18: 1187–1197.

Vega C (2000) Reproductive Allometry in Soybean, Maize and Sunflower. Ann Bot 85: 461–468.

Voinnet O (2009) Origin, biogenesis, and activity of plant microRNAs. Cell 136: 669–687.

Wang S, Zhu Q-H, Guo X, Gui Y, Bao J, Helliwell C, Fan L (2007) Molecular evolution and selection of a gene encoding two tandem microRNAs in rice. FEBS Lett 581: 4789–4793.

Wang J-W, Czech B, Weigel D (2009a) miR156-regulated SPL transcription factors define an endogenous flowering pathway in *Arabidopsis thaliana*. Cell 138: 738–749.

Wang Y, Li P, Cao X, Wang X, Zhang A, Li X (2009b) Identification and expression analysis of miRNAs from nitrogen-fixing soybean nodules. Biochem Biophys Res Commun 378: 799–803.

Wang L, Feng Z, Wang X, Zhang X (2010) DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics (Oxford, England) 26: 136–138.

Wang Z-M, Xue W, Dong C-J, Jin L-G, Bian S-M, Wang C, Wu X-Y, Liu J-Y (2011a) A comparative miRNAome analysis reveals seven fiber initiation-related and 36 novel miRNAs in developing cotton ovules. Mol Plant 2011: 1-12.

Wang L, Liu H, Li D, Chen H (2011b) Identification and characterization of maize microRNAs involved in the very early stage of seed germination. BMC Genomics 12: 154.

Wilson RF (2008) Soybean:Market Driven Research Needs. In: Stacey G, editor. Genetics and Genomics of Soybean. Springer Verlag. pp. 3–16.

Wollmann H, Weigel D (2010) Small RNAs in flower development. Eur J Cell Biol 89: 250–257.

Wong CE, Zhao Y-T, Wang X-J, Croft L, Wang Z-H, Haerizadeh F, Mattick JS, Singh MB, Carroll BJ, Bhalla PL (2011) MicroRNAs in the shoot apical meristem of soybean. J Exp Bot 62: 2495-2506.

Wu G, Park MY, Conway SR, Wang J-W, Weigel D, Poethig RS (2009) The sequential action of miR156 and miR172 regulates developmental timing in Arabidopsis. Cell 138: 750–759.

Wu L, Zhou H, Zhang Q, Zhang J, Ni F, Liu C, Qi Y (2010) DNA methylation mediated by a microRNA pathway. Molecular Cell 38: 465–475.

Wu M-F, Tian Q, Reed JW (2006) Arabidopsis microRNA167 controls patterns of *ARF6* and *ARF8* expression, and regulates both female and male reproduction. Development (Cambridge, England) 133: 4211–4218.

Xiang J, Lin J, Tang D, Zhou B, Guo M, He R, Huang X, Zhao X, Liu X (2010) A DHHC-type zinc finger protein gene regulates shoot branching in Arabidopsis. Journal of Biotechnology 9: 7759–7766.

Xu J, Zhong X, Zhang Q, Li H (2010) Overexpression of the *GmGAL2* gene accelerates flowering in Arabidopsis. Plant Mol Biol Rep 28: 704–711.

Yamada K, Lim J, Dale JM, Chen H, Shinn P, Palm CJ, Southwick AM, Wu HC, Kim C, Nguyen M *et al*. (2003) Empirical analysis of transcriptional activity in the Arabidopsis genome. Science 302: 842-846.

Yamaguchi A, Wu M-F, Yang L, Wu G, Poethig RS, Wagner D (2009) The microRNA-regulated SBP-Box transcription factor SPL3 is a direct upstream activator of *LEAFY*, *FRUITFULL*, and *APETALA1*. Dev Cell 17: 268–278.

Yamauchi Y, Hasegawa A, Taninaka A, Mizutani M, Sugimoto Y (2011) NADPH-dependent reductases involved in the detoxification of reactive carbonyls in plants. The Journal of Biological Chemistry 286: 6999–7009.

Yu X, Wang H, Lu Y, de Ruiter M, Cariaso M, Prins M, van Tunen A, He Y (2011) Identification of conserved and novel microRNAs that are responsive to heat stress in *Brassica rapa*. J Exp Bot 63: 1025-38.

Zahn LM, Ma X, Altman NS, Zhang Q, Wall PK, Tian D, Gibas CJ, Gharaibeh R, Leebens-Mack JH, de Pamphilis CW, *et al*. (2010) Comparative transcriptomics among floral organs of the basal eudicot *Eschscholzia californica* as reference for floral evolutionary developmental studies. Genome Biol 11: 1-21.

Zeng HQ, Zhu YY, Huang SQ, Yang ZM (2010) Analysis of phosphorus-deficient responsive miRNAs and cis-elements from soybean (*Glycine max* L.). J Plant Physiol 167: 1289-1297.

Zhai J, Jeong D-H, De Paoli E, Park S, Rosen BD, Li Y, González AJ, Yan Z, Kitto SL, Grusak MA, et al. (2011) MicroRNAs as master regulators of the plant NB-LRR defense gene family via the production of phased, trans-acting siRNAs. Genes Dev 25: 2540–2553.

Zhang B, Pan X, Cobb GP, Anderson T a (2006) Plant microRNA: a small regulatory molecule with big impact. Developmental Biology 289: 3–16.

Zhang B, Pan X, Stellwag EJ (2008) Identification of soybean microRNAs and their targets. Planta 229: 161–182.

Zhang BH, Pan XP, Wang QL, Cobb GP, Anderson TA (2005) Identification and characterization of new plant microRNAs using EST analysis. Cell Research 15: 336–360.

Zhang L, Chia J-M, Kumari S, Stein JC, Liu Z, Narechania A, Maher CA, Katherine Guil K, McMullen MD, Ware D (2009) A genome-wide characterization of microRNA genes in maize. PLoS Genetics 5: 1-16.

Zhang W, Gao S, Zhou X, Xia J, Chellappan P, Zhou X, Zhang X, Jin H (2010) Multiple distinct small RNAs originate from the same microRNA precursors. Genome Biol 11: 1-8.

Zhao L, Kim Y, Dinh TT, Chen X (2007) miR172 regulates stem cell fate and defines the inner boundary of *APETALA3* and *PISTILLATA* expression domain in Arabidopsis floral meristems. Plant J 51: 840–849.

Zhong X, Dai X, Xv J, Wu H, Liu B, et al. (2012) Cloning and expression analysis of *GmGAL1*, *SOC1* homolog gene in soybean. Mol Biol Rep 1-8.

Fontes eletrônicas de consulta e análise

miRBase: the microRNA database. Disponível em: http://www.mirbase.org/

Phytozome v8.0: *Glycine max*. Disponível em: http://www.phytozome.net/soybean.

psRNATarget: A Plant Small RNA Target Analysis Server. Disponível em: http://plantgrn.noble.org/psRNATarget/

UEA sRNA tools. Disponível em: http://srna-tools.cmp.uea.ac.uk/plant/cgi-bin/srna-tools.cgi.

ANEXOS

Anexo 1. Tabela dos novos de novos precursores de microRNAs e precursores conhecidos estendidos identificados no estudo de miRNAs em tecidos florais de soja.

Precursor	**Braço de origem do miRNA maduro**	***Loci***	**Região genômica**
gma-MIR169i estendido	5p (2)	Gm13:355,457..355,599 [-]	intergênico
gma-MIR319f estendido	5p/3p	Gm10:48,317,372..48,317,605 [-]	intergênico
gma-MIR4399 estendido	3p	Gm03:25,185,780..25185979 [+]	intron-cds-inter Glyma03g19830
gma-MIR4414 estendido	3p	Gm20:35,662,941..35,663,178 [+]	intergênico
Pre-MIR1515.1	5p	Gm11:6,646,776..6,646,956 [-]	intergênico
Pre-MIR156.1	5p/3p	Gm02:39,172,646..39,172,790 [+]	intergênico
Pre-MIR156.2	5p	Gm04:4,257,054..4,257,165 [+]	intergênico
Pre-MIR156.3	5p	Gm13:20,521,451..20,521,575 [+]	intergênico
Pre-MIR156.4	5p	Gm17:38,431,844..38,431,974 [-]	intergênico
Pre-MIR156.5	5p	Gm17:4,291,666..4,291,757 [-]	intergênico
Pre-MIR156.6	5p	Gm01:55,282,657..55,282,783 [+]	intergênico
Pre-MIR156.7	5p	Gm11:453,207..453,318 [-]	intergênico
Pre-MIR160.1	5p	Gm03:41,268,387..41,268,531 [-]	intron Glyma03g33780
Pre-MIR160.2	5p	Gm10:4,733,466..4,733,623 [-]	intergênico
Pre-MIR160.3	5p	Gm19:43,795,925..43,796,070 [-]	intergênico
Pre-MIR164.1	5p/3p	Gm20:41,936,354..41,936,504 [-]	intergênico
Pre-MIR164.2	5p/3p	Gm02:1,511,582..1,511,694 [+]	intergênico
Pre-MIR164.3	5p/3p	Gm03:46,896,200..46,896,353 [+]	intergênico
Pre-MIR164.4	5p/3p	Gm09:46,253,825..46,254,026 [-]	intergênico
Pre-MIR164.5	5p/3p	Gm01:43,367,838..43,368,006 [+]	intergênico
Pre-MIR164.6	5p/3p	Gm18:53,778,985..53,779,158 [+]	intergênico
Pre-MIR164.7	5p/3p	Gm19:48,157,195..48,157,299 [+]	intergênico
Pre-MIR164.8	5p/3p	Gm03:45,537,760..45,537,879 [+]	intergênico
Pre-MIR166.1	5p/3p	Gm19:36,649,682..36,649,868 [-]	intergênico
Pre-MIR166.2	5p/3p	Gm05:37,747,278..37,747,371 [-]	intergênico
Pre-MIR166.3	5p/3p	Gm03:39,519,813..39,519,973 [-]	intergênico
Pre-MIR166.4	3p	Gm07:10,198,785..10,198,984 [+]	intergênico
Pre-MIR166.5	3p	Gm09:37,125,211..37,125,372 [-]	intergênico
Pre-MIR166.6	5p/3p	Gm04:25,372,866..25,373,022 [+]	intergênico
Pre-MIR166.7	5p/3p	Gm09:33,908,882..33,909,036 [+]	intergênico
Pre-MIR167.1	5p	Gm03:39,319,041..39,319,195 [+]	cds-3UTR Glyma03g31410
Pre-MIR167.2	5p	Gm02:307,506..307,629 [-]	intergênico
Pre-MIR168.1	5p/3p	Gm01:48,070,305..48,070,428 [-]	intergênico
Pre-MIR169.1	5p	Gm15:14,202,424..14,202,597 [+]	intergênico

Pre-MIR169.10	5p/3p	Gm15:14,176,133..14,176,271 [+]	intergênico
Pre-MIR169.11	5p/3p	Gm09:42,642,439..42,642,660 [-]	intergênico
Pre-MIR169.2	5p	Gm09:5,299,542..5,299,763 [+]	intergênico
Pre-MIR169.3	5p	Gm13:371,080..371,201 [-]	intergênico
Pre-MIR169.4	5p	Gm15:14,188,469..14,188,641 [+]	intergênico
Pre-MIR169.5	5p	Gm15:14,191,146..14,191,325 [+]	intergênico
Pre-MIR169.6	5p	Gm17:4,864,158..4,864,292 [-]	intergênico
Pre-MIR169.7	5p	Gm17:4,857,065..4,857,207 [-]	intergênico
Pre-MIR169.8	5p (2)	Gm13:363,824..363,962 [-]	intergênico
Pre-MIR171.1	5p/3p	Gm17:1,007,441..1,007,605 [+]	intergênico
Pre-MIR171.2	5p/3p	Gm13:30,650,796..30,650,885 [+]	intergênico
Pre-MIR171.3	3p	Gm08:921,797..921,896 [-]	intergênico
Pre-MIR171.4	3p	Gm06:46,773,062..46,773,167 [-]	intergênico
Pre-MIR171.5	3p	Gm12:35,489,092..35,489,187 [+]	intergênico
Pre-MIR172.1	3p	Gm13:41,635,473..41,635,640 [-]	intergênico
Pre-MIR2111.1	5p/3p	Gm07:16,377,773..16,377,895 [-]	intergênico
Pre-MIR2111.2	5p/3p	Gm18:44,683,049..44,683,157 [-]	intron-cds Glyma18g37410
Pre-MIR2119.1	3p	Gm02:10,988,187..10,988,282 [-]	intron Glyma02g12710
Pre-MIR319.1	3p	Gm06:22,581,687..22,581,785 [+]	intergênico
Pre-MIR319.2	3p	Gm13:27,068,574..27,068,690 [-]	intergênico
Pre-MIR390.1	5p/3p	Gm01:42,335,608..42,335,718 [+]	intergênico
Pre-MIR390.2	5p	Gm18:53,278,028..53,278,166 [+]	intergênico
Pre-MIR390.3	3p	Gm18:5,047,763..5,047,873 [-]	intergênico
Pre-MIR390.4	5p/3p	Gm02:44,954,729..44,954,915 [+]	intergênico
Pre-MIR393.1	5p/3p	Gm02:47,136,182..47,136,341 [-]	intergênico
Pre-MIR393.2	5p/3p	Gm01:41,428,795..41,428,984 [-]	intergênico
Pre-MIR393.3	5p/3p	Gm14:5,053,414..5,053,567 [+]	intergênico
Pre-MIR393.4	5p/3p	Gm18:2,409,721..2,409,880 [-]	intergênico
Pre-MIR393.5	5p/3p	Gm20:24,956,244..24,956,363 [-]	intergênico
Pre-MIR393.6	5p/3p	Gm09:31,624,534..31,624,673 [+]	intergênico
Pre-MIR393.7	5p/3p	Gm16:33,891,074..33,891,220 [+]	intergênico
Pre-MIR393.8	5p	Gm10:45,859,041..45,859,169 [+]	intergênico
Pre-MIR393.9	5p	Gm20:38,670,383..38,670,517 [-]	intergênico
Pre-MIR394.1	5p	Gm08:9,880,035..9,880,187 [+]	intergênico
Pre-MIR394.2	5p	Gm14:46,995,379..46,995,532 [-]	intergênico
Pre-MIR394.3	5p	Gm15:3,767,148..3,767,290 [+]	intergênico
Pre-MIR394.4	5p	Gm05:35,743,160..35,743,313 [+]	intergênico
Pre-MIR394.5	5p	Gm06:1,502,431..1,502,568 [-]	intergênico
Pre-MIR395.1	3p	Gm01:4,813,248..4,813,391 [-]	intergênico
Pre-MIR395.10	3p	Gm18:16,305,069..16,305,195 [-]	intergênico
Pre-MIR395.2	5p/3p	Gm02:1,730,678..1,730,791 [+]	intergênico
Pre-MIR395.3	5p/3p	Gm08:40,810,567..40,810,703 [-]	intergênico
Pre-MIR395.4	5p/3p	Gm01:4,797,901..4,798,016 [-]	intergênico
Pre-MIR395.5	5p/3p	Gm01:4,810,809..4810946 [-]	intergênico
Pre-MIR395.6	5p/3p	Gm02:1,736,324..1,736,468 [+]	intergênico
Pre-MIR395.7	3p	Gm01:4,788,545..4,792,485 [+]	intergênico
Pre-MIR395.8	3p	Gm08:40,838,512..40,838,659 [-]	intergênico

Pre-MIR395.9	5p/3p	Gm02:1,772,395..1,772,531 [+]	intergênico
Pre-MIR4392.1	3p	Gm20:156,620..156,765 [-]	intergênico
Pre-MIR4406.1	3p	Gm14:21,903,406..21,903,531 [+]	intergênico
Pre-MIR4416.1	5p	Gm03:38,094,623..38,094,776 [-]	intergênico
Pre-MIR482.1	5p/3p	Gm18:61,452,879..61,453,028 [-]	intergênico
Pre-MIR530.1	5p/3p	Gm11:16,833,250..16,833,441 [-]	intergênico
Pre-MIR530.2	5p/3p	Gm13:40,319,183..40,319,372 [-]	intergênico
Pre-MIR5380.1	3p	Gm15:18,047,158..18,047,244 [+]	intergênico
Pre-MIR5380.2	3p	Gm16:11,320,145..11,320,309 [-]	intergênico
Pre-NF01	3p	Gm02:8,454,911..8,455,158 [-]	intron Glyma02g10570
Pre-NF02	3p	Gm14:558,487..558,711 [-]	intron Glyma14g01040
Pre-NF03a	5p/3p	Gm07:10,001,901..10,002,013 [+]	intergênico
Pre-NF03b	5p/3p	Gm07:10,004,513..10,004,625 [+]	intergênico
Pre-NF04	3p	Gm15:17,904,646..17,904,916 [-]	intergênico
Pre-NF05	5p/3p	Gm08:40,662,138..40,662,278 [+]	intergênico
Pre-NF06	3p	Gm13:8,342,024..8,342,234 [+]	intron Glyma13g08090
Pre-NF07	5p/3p	Gm15:39,190,763..39,191,053 [-]	intergênico
Pre-NF08	3p	Gm16:19,925,619..19925903 [+]	intergênico
Pre-NF09a	3p	Gm02:39,925,519..39,925,662 [-]	intergênico
Pre-NF09b	3p	Gm15:31,393,116..31,393,303 [+]	intergênico
Pre-NF09c	3p	Gm18:42,937,102..42,937,265 [+]	intergênico
Pre-NF10	5p/3p	Gm12:648,893..649,030 [+]	intergênico
Pre-NF11	5p/3p	Gm14:6,304,120..6,304,369 [-]	intergênico
Pre-NF12	5p/3p	Gm20:42,606,806..42,607,020 [+]	intergênico
Pre-NF13	5p	Gm16:7,236,797..7,236,894 [-]	intergênico
Pre-NF14	5p	Gm17:27,739,062..27,739,312 [+]	intron Glyma17g26570
Pre-NF15	3p	Gm03:20,886,972..20,887,185 [-]	intron Glyma03g16440
Pre-NF16	3p	Gm17:14,170,467..14,170,651 [-]	intron Glyma17g17360
Pre-NF17	5p	Gm03:46,862,468..46,862,679 [+]	intron Glyma03g41350
Pre-NF18	5p/3p	Gm13:34,948,270..34,948,416 [-]	intron Glyma13g32940
Pre-NF19	5p/3p	Gm14:4,584,761..4,585,058 [-]	intron Glyma14g06340
Pre-NF20	5p	Gm20:44,450,038..44,450,191 [-]	intergênico
Pre-NF21	3p	Gm20:13,840,325..13,840,460 [+]	intron Glyma20g10030
Pre-NF22	3p	Gm16: 6,256,970.. 6,257,129 [+]	intron Glyma16g06930
Pre-NF23	3p	Gm10:37,397,199..37,397,385 [-]	intergênico
Pre-NF24a	5p	Gm07:44,231,008..44,231,088 [+]	intergênico
Pre-NF24b	5p	Gm17:493,663..493,743 [-]	intergênico
Pre-NF25	3p	Gm03:2,812,304..2,812,468 [+]	intergênico

Pre-NF26	3p	Gm04:5,821,402..5,821,618 [+]	intron Glyma04g07460
Pre-NF27	5p	Gm10:47,656,000..47,656,302 [+]	intergênico
Pre-NF28	5p/3p	Gm08:11,432,998..11,433,200 [-]	intergênico
Pre-NF29	3p	Gm13:41,358,321..41,358,477 [+]	intron Glyma13g40930
Pre-NF30	3p	Gm13:11,972,958..11,973,052 [-]	intergênico
Pre-NF31	5p	Gm12:37,542,998..37,543,242 [-]	intron Glyma12g34410
Pre-NF32	5p	Gm01:8,573,823..8,573,906 [+]	intron Glyma01g07770
Pre-NF33	3p	Gm02:6,233,520..6,233,676 [+]	intergênico
Pre-NF34	5p	Gm02:48,299,297..48,299,447 [+]	intron Glyma02g43540
Pre-NF35	5p	Gm10:43,308,727..43,308,855 [-]	intergênico
Pre-NF36	5p	Gm07:12,248,009..12,248,194 [+]	intergênico
Pre-NF37a	3p	Gm14:25,154,100..25,154,318 [+]	intergênico
Pre-NF37b	5p/3p	Gm19:40,112,991..40,113,140 [+]	intergênico
Pre-NF38	5p	Gm15:12,954,693..12,954,874 [+]	cds-intron-cds Glyma15g16650
Pre-NF39a	3p	Gm18:42,254,378..42,254,587 [+]	intergênico
Pre-NF39b	3p	Gm19:11,095,283..11,095,442 [-]	intergênico
Pre-NF40	3p	Gm16:1,163,892..1,164,088 [+]	intron Glyma16g01590
Pre-NS-MIR399.1	3p	Gm05:34,967,635..34,967,795 [-]	intergênico
Pre-NS-MIR399.2	3p	Gm08:9,118,494..9,118,632 [-]	intergênico
Pre-NS-MIR399.3	3p	Gm08:9,126,507..9,126,650 [-]	intergênico
Pre-NS-MIR399.4	3p	Gm09:34,181,494..34,181,659 [+]	intergênico
Pre-NS-MIR399.5	5p/3p	Gm10:46,275,285..46275438 [+]	intergênico
Pre-NS-MIR399.6	5p/3p	Gm20:38,251,181..37,397,385 [-]	intergênico
Pre-NS-MIR399.7	3p	Gm16:35,612,482..35,612,634 [+]	intergênico
Pre-NS-MIR399.8	5p	Gm09:34,174,598..34,174,761 [+]	intergênico
Pre-NS-MIR399.9	5p	Gm16:35,606,632..35,606,797 [+]	intergênico
Pre-NS-MIR5281.1	3p	Gm03:27,274,605..27,274,751 [-]	intron Glyma03g21620
Pre-NS-MIR828.1	5p	Gm12:3,213,744..3,213,893 [+]	intergênico

Anexo 2: Estruturas secundárias e mapeamento dos precursores dos miRNAs inéditos. A identificação, o lócus, a orientação e a região genômica de cada precursor estão indicados. Os nucleotídos pareados e não pareados estão representados por parênteses e pontos, respectivamente, abaixo das sequências dos precursores. Ao final da representação da estrutura secundária está indicada a energia livre mínima (MFE). As sequências dos miRNAs maduros 5p e 3p representativos de cada precursor estão sublinhadas ou marcadas em cinza, respectivamente.

```
>Pre-NF01   Gm02:8,454,911..8,455,158 [-] intron_Glyma02g10570
GAGATAAATTCTGCCAAGGCTTAATCCTCCACTTTGGTACGGCTTAAGTTCAACTTTGGAGGAAAAACATGGATGTTATGGTGCTTTAAGCCCTGCAACTTTAGTGCGGCTTAAGC
(((((.....(((((((((.((((......(((((((((((((((((((((((((((((((((.((((((((((.((((((.(((((.((.((.(((.((((((((((..
ACGACCCTTAAGCATGCCAGTTTTTGATTAAGCTGCATCAAACTTTCAGGCTTAAGCACAGTAGCATCCATATTTTTCCTCCAAAGTTGGGCTTAAGCTGTACCAAAGTTTAATCC
.(((.....((((......)))))))..)))))))))).)))..)).)))))))))))))).)))))))))).)))))))))))))))))))))))))))))))))))))).))
TTGGCAGTAAAATCTC
)))))))....)))))  (-131.30)]
```

Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)
ACTTTGGTACGGCTTAAGTTCAAC (3)	31	24
CTTTGGTACGGCTTAAGTTCAACT (1)	32	24
TTTGGTACGGCTTAAGTTCAAC (2)	33	22
TTGGTACGGCTTAAGTTCAAC (1)	34	21
GGAAAAACATGGATGTTATGGTGC (1)	61	24
GAAAAACATGGATGTTATGGT (2)	62	21
GAAAAACATGGATGTTATGGTG (2)	62	22
GAAAAACATGGATGTTATGGTGCT (9)	62	24
CTGCATCAAACTTTCAGGCTTAA (2)	149	23
TTAAGCACAGTAGCATCCAT (1)	168	20
TTAAGCACAGTAGCATCCATATT (1)	168	23
CACAGTAGCATCCATATTTTTC (1)	173	22
GTAGCATCCATATTTTTCCTC (1)	177	21
GCATCCATATTTTTCCTCCAAAGT (8)	180	24
TTTCCTCCAAAGTTGGGCTTAAGC (1)	191	24

Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)
TTCCTCCAAAGTTGGGCTTAAG (4)	192	22
TTCCTCCAAAGTTGGGCTTAAGC (1)	192	23
CCTCCAAAGTTGGGCTTAAGCT (1)	194	22
CTCCAAAGTTGGGCTTAAGCTGT (1)	195	23
CTCCAAAGTTGGGCTTAAGCTGTA (21)	195	24
TCCAAAGTTGGGCTTAAGCTGT (1)	196	22
TCCAAAGTTGGGCTTAAGCTGTA (7)	196	23
TCCAAAGTTGGGCTTAAGCTGTAC (7)	196	24
CCAAAGTTGGGCTTAAGCT (2)	197	19
CCAAAGTTGGGCTTAAGCTGTA (339)	197	22
CCAAAGTTGGGCTTAAGCTGTAC (4)	197	23
CCAAAGTTGGGCTTAAGCTGTACC (62)	197	24
CAAAGTTGGGCTTAAGCTGTA (15)	198	21
CAAAGTTGGGCTTAAGCTGTAC (1)	198	22
CAAAGTTGGGCTTAAGCTGTACC (2)	198	23
CAAAGTTGGGCTTAAGCTGTACCA (8)	198	24

>Pre-NF02 Gm14:558,487..558,711 [-] intron_Glyma14g01040
AGAATATTTCATGTGTACAAGTCATATACTATGTAAAATGATTTCAGAATATAAATTATCTTTATCACCTGCTTTATGTGGAATACCACAACAAAGTCTTATCCCACTAGGTCACC
((((((((((((((((((((....((((..((((((((((((((((((((((((((..(((((((((.(((((((((((((((((((((((((((((((((((...))
TAGTGGG**ATAAGGCTTTGTTGTGGTATTCCA**CATAAAATAGGTGATACTGATAATTTATATTCTGGAATCATTTTACTCAGTAAACAACTTGTACACATGAAATATTTT
))))))))))))))))))))))))))))))))))).)))))))..))))))))))))))))))))))))..))))....))))))))))))))))))
(-150.10)

Sequência (nº de *reads*)	Posição inicial	Tamanho (nt)
TGTAAAATGATTTCAGAATATAAA (1)	32	24
TCTTTATCACCTGCTTTATGTG (2)	59	22
(nt)TTTATCACCTGCTTTATGTGGA (8)	61	22
TTATCACCTGCTTTATGTGGA (1)	62	21
GAATACCACAACAAAGTCTTATCC (1)	81	24
ATACCACAACAAAGTCTTATCCCA (7)	83	24
CCACAACAAAGTCTTATCCCAC (1)	86	22
CCACAACAAAGTCTTATCCCACT (3)	86	23
CCACAACAAAGTCTTATCCCACTA (1)	86	24
CACCTAGTGGGATAAGGC (1)	113	18
CACCTAGTGGGATAAGGCTTTGT (1)	113	23
CACCTAGTGGGATAAGGCTTTGTT (1)	113	24

Sequência (nº de *reads*)	Posição inicial	Tamanho
GTGGGATAAGGCTTTGTTGTGG (1)	119	22
GTGGGATAAGGCTTTGTTGTGGT (25)	119	23
GTGGGATAAGGCTTTGTTGTGGTA (49)	119	24
TGGGATAAGGCTTTGTTGTGGT (2)	120	22
TGGGATAAGGCTTTGTTGTGGTA (22)	120	23
TGGGATAAGGCTTTGTTGTGGTAT (9)	120	24
TGGGATAAGGCTTTGTTGTGGTATT (2)	120	25
ATAAGGCTTTGTTGTGGTATT (1)	124	21
ATAAGGCTTTGTTGTGGTATTCC (1)	124	23
ATAAGGCTTTGTTGTGGTATTCCA (143)	124	24
TAAGGCTTTGTTGTGGTATTC (3)	125	21
TGGTATTCCACATAAAATAGGT (1)	138	22
TATATTCTGGAATCATTTTACTC (1)	173	23

>Pre-NF03a Gm07:10,001,901.. 10,002,013 [+] intergênico (duplex)
GAACTAAACAGATCTCGTAGTTAT**TTATCAGTAGCATCATCATCA**AAGTTCTTTACATGAAA**GTTTGATGATGATGTTACCGA**TAATGAGCTACGAGAGTGTTTTTAGCCGTC
.....(((((..((((((((((..((((((.((((((((((((((((.(((.......))).)).)))))))))))))))).)))))..))))))))))).)))))........
(-48.90)

Sequência (nº de *reads*)	Posição inicial	Tamanho (nt)
TCTCGTAGTTATTTATCAGTA (1)	13	21
TTTATCAGTAGCATCATCATC (20)	24	21
TTTATCAGTAGCATCATCATCA (1)	24	22
TTTATCAGTAGCATCATCATCAAA (1)	24	24
TTATCAGTAGCATCATCAT (1)	25	19
TTATCAGTAGCATCATCATC (4)	25	20
TTATCAGTAGCATCATCATCA (143)	25	21
TTATCAGTAGCATCATCATCAA (1)	25	22
TTATCAGTAGCATCATCATCAAAGT (1)	25	25

Sequência (nº de *reads*)	Posição inicial	Tamanho (nt)
TTATCAGTAGCATCATCATCAAAGTT (1)	25	26
ATCAGTAGCATCATCATCAAA (9)	27	21
ATCAGTAGCATCATCATCAAAGTT (3)	27	24
AGTAGCATCATCATCAAAGTT (2)	30	21
GTAGCATCATCATCAAAG (1)	31	18
AAAGTTCTTTACATGAAAGTTTGA (1)	45	24
GTTTGATGATGATGTTACCG (2)	63	20
GTTTGATGATGATGTTACCGA (64)	63	21
TGATGATGATGTTACCGATAA (1)	66	21

>Pre-NF03b Gm07:10,004,513..10,004,625 [+] intergênico
GAACTAAACAGATCTCGTAGTTAT**TTATCAGTAGCATCATCATCA**AAGTTCTTAACATAAAA**GTTTGATGATGATGTTACCGA**TAAAGAGCTACGAGAGTGTTTTTAGCCGTC
.....(((((..((((((((((.((((((.((((((((((((((((((.((..........)).)))))))))))))))))).)))))).)))))))))).)))))........
(-48.40)

Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)	Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)
TCTCGTAGTTATTTATCAGTA (1)	13	21	ATCAGTAGCATCATCATCAAAGTT (3)	27	24
TTTATCAGTAGCATCATCATC (20)	24	21	AGTAGCATCATCATCAAAGTT (2)	30	21
TTTATCAGTAGCATCATCATCA (1)	24	22	GTAGCATCATCATCAAAG (1)	31	18
TTTATCAGTAGCATCATCATCAAA (1)	24	24	GTTTGATGATGATGTTACCG (2)	63	20
TTATCAGTAGCATCATCAT (1)	25	19	**GTTTGATGATGATGTTACCGA** (64)	63	21
TTATCAGTAGCATCATCATC (4)	25	20	TGATGATGATGTTACCGATAA (1)	66	21
TTATCAGTAGCATCATCATCA (143)	25	21	TTTATCGGTAACATCATCATC (1)	67	21
TTATCAGTAGCATCATCATCAA (1)	25	22	GATGATGATGTTACCGATAAAGAG (1)	67	24
TTATCAGTAGCATCATCATCAAAGT (1)	25	25	ATGATGATGTTACCGATAAAGA (2)	68	22
TTATCAGTAGCATCATCATCAAAGTT (1)	25	26	TGATGATGTTACCGATAAAGA (2)	69	21
ATCAGTAGCATCATCATCAAA (9)	27	21	TTACCGATAAAGAGCTACGAG (3)	77	21
			TACCGATAAAGAGCTACGAGA (9)	78	21

Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)

>Pre-NF04 Gm15:17,904,646..17,904,916 [-] intergênico
TCAACTAAGTATCCAAAAAAATTTATATTTTAATTTACTATCTGATATTAGTCTCCTTCTATTAAGTAATGACGTAGCGGACGGAGTGTCACATCATCAT
((((.(((.((((..((((((.....((((((((((((((((...((((((((((..((((((.((((.((((((((((((((((..((.(((.
GTCTTATCACTAACACACATTGTCATGTCATCACCTTTAATGGCGTAACAATGATGTGTCAGTGATTAGGCTTGACGGCATGACACTTCGTCTGCTACAT
((((.((((((.((((((((((.(((((((........)))))))).)))))).)))))).)))))).)))).)))..)))))))))))))))))).)
TATCACTTAACTGAAGGGGAC**TAACGACAGGTAGTAAATTGAA**ACAATTTTTTTTTCAGATAGTTATTTGA
)))..))))))..)))))))))))))...)))))))))))))))))).....))))))..))))..)))..)))) (-125.40)

Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)	Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)
AAGGAGACTAATATCAGATAGTAA (1)	35	24	GTAGCGGACGGAGTGTCACATCAT (1)	74	24
ATTAAGTAATGACGTAGCGGACGG (1)	61	24	ACATGATGATGTGACACTCCGTCC (2)	79	24
TAAGTAATGACGTAGCGGACGGAG (2)	63	24	GATAAGACATGATGATGTGACACT (1)	85	24
AAGTAATGACGTAGCGGACGGAGT (8)	64	24	TGATAAGACATGATGATGTGACACT (1)	85	25
AGTAATGACGTAGCGGACGGAGTG (2)	65	24	GTGATAAGACATGATGATGTGA (1)	89	22
GTAATGACGTAGCGGACGGAGTGT (1)	66	24	AACACACATTGTCATGTCATCACC (1)	112	24
TAATGACGTAGCGGACGGAGTGTC (1)	67	24	CATGTCATCACCTTTAATGGCGTA (1)	124	24
ATGACGTAGCGGACGGAGTGTCACA (1)	69	25	TAATGGCGTAACAATGATGTGT (1)	138	22

CGTAACAATGATGTGTCAGTGA (1)	144	22
CAATGATGTGTCAGTGATTAGGCT (2)	149	24
AATGATGTGTCAGTGATTAGGCT (1)	150	23
AATGATGTGTCAGTGATTAGGCTT (2)	150	24
GTCATGCCGTCAAGCCTAATCACT (1)	161	24
GATTAGGCTTGACGGCATGACACT (1)	164	24
AGTGATAATGTAGCAGACGAAGTG (2)	184	24
TAAGTGATAATGTAGCAGACGAAG (3)	186	24
ACATTATCACTTAACTGAAGGGGA (2)	197	24
ATTATCACTTAACTGAAGGGGACT (1)	199	24
AAGGGGACTAACGACAGGTAGTAA (1)	214	24
ATTTACTACCTGTCGTTAGTCCCC (2)	216	24
GGGACTAACGACAGGTAGTAAA (1)	217	22
GGACTAACGACAGGTAGTAAA (1)	218	21
GGACTAACGACAGGTAGTAAATT (1)	218	23
ACTAACGACAGGTAGTAAAT (1)	220	20
ACTAACGACAGGTAGTAAATTGAA (1)	220	24
CTAACGACAGGTAGTAAATTG (1)	221	21
CTAACGACAGGTAGTAAATTGA (58)	221	22
CTAACGACAGGTAGTAAATTGAA (5)	221	23
TAACGACAGGTAGTAAATTGA (3)	222	21
TAACGACAGGTAGTAAATTGAA (115)	222	22
TAACGACAGGTAGTAAATTGAAA (2)	222	23
AACGACAGGTAGTAAATTGAA (1)	223	21
AACGACAGGTAGTAAATTGAAA (2)	223	22
AACGACAGGTAGTAAATTGAAACA (8)	223	24
TGTTTCAATTTACTACCTGTCGTT (1)	223	24
AACGACAGGTAGTAAATTGAAACAA (1)	223	25
ACGACAGGTAGTAAATTGAAACA (2)	224	23
ACGACAGGTAGTAAATTGAAACAA (1)	224	24
CGACAGGTAGTAAATTGAAACA (2)	225	22

>Pre-NF05 Gm08:40,662,138..40,662,278 [+] intergênico (duplex)

GGTTCTTCAATTTTGTA**GTTTTTTTCTTTGTGGTGGCCGGG**TTAAACTTCAATTAATAATTATTTGGAGTCCAAACCTTGGCAGC

(((((.(((...(((((((((((((((((((((((((.....(((((((.((.....)).)))))))..........((...

TGGCATCAATA**CCGGCCACCAACAAAGAAAAAAACT**AAACAATATTATGGCGAACT

..)).......)))))))).)))))))))))))))..))))).....))).))))) (-51.80)

Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)
TAGTTTTTTTCTTTGTGGTGGCCGGG (3)	16	26
AGTTTTTTTCTTTGTGGTGGCCGG (1)	17	24
AGTTTTTTTCTTTGTGGTGGCCGGG (75)	17	25
AGTTTTTTTCTTTGTGGTGGCCGGGT (1)	17	26
GTTTTTTTCTTTGTGGTGGCC (1)	18	21
GTTTTTTTCTTTGTGGTGGCCG (1)	18	22
GTTTTTTTCTTTGTGGTGGCCGG (1)	18	23
GTTTTTTTCTTTGTGGTGGCCGGG (104)	18	24
GTTTTTTTCTTTGTGGTGGCCGGGT (3)	18	25
GTTTTTTTCTTTGTGGTGGCCGGGTT (1)	18	26
TTTTTTTCTTTGTGGTGGCCGGG (1)	19	23
TTTTTTCTTTGTGGTGGCCGGG (3)	20	22
TTTTTCTTTGTGGTGGCCGGG (2)	21	21
TTTTCTTTGTGGTGGCCGGG (1)	22	20
TTTCTTTGTGGTGGCCGGG (2)	23	19
TTCTTTGTGGTGGCCGGG (4)	24	18
TCTTTGTGGTGGCCGGGTTAA (1)	25	21
TACCGGCCACCAACAAAGAA (1)	95	20
ACCGGCCACCAACAAAGA (1)	96	18
ACCGGCCACCAACAAAGAAAAAAACT (9)	96	26
CCGGCCACCAACAAAGAA (5)	97	18
CCGGCCACCAACAAAGAAA (1)	97	19
CCGGCCACCAACAAAGAAAAAAACT (19)	97	25
CCGGCCACCAACAAAGAAAAAAACTA (2)	97	26
CCACCAACAAAGAAAAAAACT (1)	101	21
CACCAACAAAGAAAAAAACT (2)	102	20

>Pre-NF06 Gm13:8,342,024..8,342,234 [+] intron_Glyma13g08090
TTTCCTTCTCTTGAGCTTCTTAAACCCTAAGCCCACAAAACCTAAGCCCCAAAATCAGCTCTTCTATTGTAACCCTAAGTTCCAAATGGTGATATGTATTTATTAATACTTCTAAT
.(((((((((.(((.((((((((((.((((((((((((((((.((((((((((.((((((((((...(((...((((((((((((.(((((((((.................))
ATATCATGATTTGGGACTTAGTACTACTGCAGAAGAGTTGA**GTTTGGGGCTTGTGTTTTGTGGGC**TTAGGCTTTAAGAAGCCCAAAAGGAGGAAA
)))))))..)))))))))))))...)))...)))))))))))..))))))))))..)))))))))))))))))..))))))))))..)))..)))))))). (-122.03)

Sequência (nº de *reads*)	Posição inicial	Tamanho (nt)	Sequência (nº de *reads*)	Posição inicial	Tamanho (nt)
TCTTGAGCTTCTTAAACCCTAAG (5)	9	23	TGGGGCTTGTGTTTTGTGGGC (12)	161	21
TCTTGAGCTTCTTAAACCCTAAGC (3)	9	24	TGGGGCTTGTGTTTTGTGGGCT (4)	161	22
TTTTGTGGGCTTAGGGTTTAA (1)	20	21	GGGGCTTGTGTTTTGTGGGCT (4)	162	21
TAAGCCCACAAAACCTAAGCCC (1)	28	22	TTGTGTTTTGTGGGCTTAGGCT (1)	167	22
TAAGCCCACAAAACCTAAGCCCC (1)	28	23	TGTGTTTTGTGGGCTTAGGCT (2)	168	21
CCACAAAACCTAAGCCCCAAAA (1)	33	22	GTTTTGTGGGCTTAGGCTTTAAGA (1)	171	24
CCACAAAACCTAAGCCCCAAAATC (1)	33	24	TTAAAGCCTAAGCCCACAAAA (1)	172	21
GTTTGGGGCTTGTGTTTTGTG (1)	158	21	TTTTGTGGGCTTAGGCTTTAAGAA (1)	172	24
GTTTGGGGCTTGTGTTTTGTGGG (1)	158	23	TTTGTGGGCTTAGGCTTTAAGA (1)	173	22
GTTTGGGGCTTGTGTTTTGTGGGC (72)	158	24	TTGTGGGCTTAGGCTTTAAGAA (4)	174	22
TTTGGGGCTTGTGTTTTGTGGG (5)	159	22	TGTGGGCTTAGGCTTTAAGAA (3)	175	21
TTTGGGGCTTGTGTTTTGTGGGC (13)	159	23	CTTAGGCTTTAAGAAGCCCAAA (1)	181	22
TTTGGGGCTTGTGTTTTGTGGGCT (15)	159	24	TTAAGAAGCCCAAAAGGAGGAA (1)	189	22
TTGGGGCTTGTGTTTTGTGGGC (7)	160	22	TAAGAAGCCCAAAAGGAGGAA (1)	190	21
TTGGGGCTTGTGTTTTGTGGGCT (2)	160	23	TAAGAAGCCCAAAAGGAGGAAA (1)	190	22

>Pre-NF07 Gm15:39,190,763..39,191,053 [-] intergênico (duplex)
TTATGGGCATGGCATAGTTGGGGTGGAGTATA**AGGGTCACCTGTTCTTGGGTTA**TGAAGCAGGTTAGGGGTGCAGGGGCAAAGCCCTTGCAGAACACGTAATATTATACACGGATT
.(((((((((.((((.(((((((((((((((.((((((((((.((((.(((((((..(((..((((.((((((.((((((((((((((((((((((((((((((((((.((.(((((.(((((
AAGATAACGCACGTGGGCTAAGATTTACACATACATTATCTTTACAGGTGCATTATTTTAATCCATGTACAACATTATGTGTTCTGCGAGGGCTTTGCCCCTACACCCCCAACCAA
(((((((.((((.((....(((((.............)))))..)).))))).)))))))))))))).))))).)).)))))))))))))))))))))))))))))).)))))).)))))..
CTTTGTA**ACCCAACAACAAGTGATCCTTA**CACTTCACTCCAACTTTGCCCTGCCCATAA
)))..)))))))).))))).)))))))))).)))))))))))))).))))).)))))))). (-188.72)

Sequência (nº de *reads*)	Posição inicial	Tamanho (nt)	Sequência (nº de *reads*)	Posição inicial	Tamanho (nt)
TTATGGGCATGGCATAGTTGG (1)	1	21	ATGGGCATGGCATAGTTGGGGTG (1)	3	23
TTATGGGCATGGCATAGTTGGG (2)	1	22	TGGGCATGGCATAGTTGGGG (1)	4	20
TTATGGGCATGGCATAGTTGGGG (2)	1	23	TGGGCATGGCATAGTTGGGGTGGA (1)	4	24
TATGGGCATGGCATAGTTGGG (1)	2	21	GGGCATGGCATAGTTGGGGTGGAG (2)	5	24

GCATAGTTGGGGTGGAGTAT (1)	12	20
CATAGTTGGGGTGGAGTATAAGGG (4)	13	24
TAGTTGGGGTGGAGTATAAGGG (3)	15	22
TAGTTGGGGTGGAGTATAAGGGT (1)	15	23
TAGTTGGGGTGGAGTATAAGGGTC (1)	15	24
GTGGAGTATAAGGGTCACCTGTT (3)	23	23
TGGAGTATAAGGGTCACCTGT (1)	24	21
TGGAGTATAAGGGTCACCTGTT (7)	24	22
CCAAGAACAGGTGACCCTTATACT (1)	27	24
TAAGGGTCACCTGTTCTTGGGT (4)	31	22
TAAGGGTCACCTGTTCTTGGGTTA (1)	31	24
AAGGGTCACCTGTTCTTGGGTTA (2)	32	23
AAGGGTCACCTGTTCTTGGGTTAT (2)	32	24
AGGGTCACCTGTTCTTGGGTTA (22)	33	22
AGGGTCACCTGTTCTTGGGTTATG (1)	33	24
GGGTCACCTGTTCTTGGGTTA (2)	34	21
GGGTCACCTGTTCTTGGGTTAT (1)	34	22
CTGTTCTTGGGTTATGAAGCAGGT (2)	41	24
CTTGGGTTATGAAGCAGGTTAGGG (1)	46	24
TGGGTTATGAAGCAGGTTAGGGG (1)	48	23
ACCCCTAACCTGCTTCATAACCCA (1)	48	24
TGGGTTATGAAGCAGGTTAGGGGT (1)	48	24
GTTATGAAGCAGGTTAGGGGTG (1)	51	22
AGGTTAGGGGTGCAGGGGCAAAGC (1)	61	24
CAAAGCCCTTGCAGAACACGT (2)	79	21
CAAAGCCCTTGCAGAACACGTA (8)	79	22
CAAAGCCCTTGCAGAACACGTAA (9)	79	23
CAAAGCCCTTGCAGAACACGTAAT (4)	79	24
AAAGCCCTTGCAGAACACGTAA (1)	80	22
AAAGCCCTTGCAGAACACGTAATA (4)	80	24
CTTGCAGAACACGTAATA (1)	86	18
ATTATACACGGATTAAGATAACGC (1)	103	24
ATTATACACGGATTAAGATAACGCA (1)	103	25
TTATACACGGATTAAGATAACGCA (3)	104	24
ATACACGGATTAAGATAACGC (1)	106	21
TACACGGATTAAGATAACGCA (1)	107	21
TGTAAAGATAATGTATGTGTAAAT (1)	139	24
CATTATGTGTTCTGCGAGGGCT (1)	189	22
TATGTGTTCTGCGAGGGCTT (1)	192	20
TATGTGTTCTGCGAGGGCTTTG (2)	192	22
TATGTGTTCTGCGAGGGCTTTGCC (2)	192	24
TGTGTTCTGCGAGGGCTTT (1)	194	19
GTTCTGCGAGGGCTTTGCC (1)	197	19
TTCTGCGAGGGCTTTGCCCCTACA (1)	198	24
CCCCCAACCAACTTTGTAACC (1)	222	21
CCCCCAACCAACTTTGTAACCC (2)	222	22
CAACCAACTTTGTAACCCAACAAC (1)	226	24
GGATCACTTGTTGTTGGGTTA (1)	238	21
TAACCCAACAACAAGTGATCCTTA (1)	238	24
AACCCAACAACAAGTGATCCTT (1)	239	22
AACCCAACAACAAGTGATCCTTA (6)	239	23
AACCCAACAACAAGTGATCCTTAC (1)	239	24
ACCCAACAACAAGTGATCCTT (1)	240	21
ACCCAACAACAAGTGATCCTTA (70)	240	22
ACCCAACAACAAGTGATCCTTACA (4)	240	24
CCAACAACAAGTGATCCTTACA (3)	242	22
ACAACAAGTGATCCTTACACTTCA (1)	245	24
AGTGATCCTTACACTTCACTCCAA (1)	251	24
CAAAGTTGGAGTGAAGTGTAAGGA (1)	256	24
TACACTTCACTCCAACTTTGCCC (1)	260	23
CACTTCACTCCAACTTTGCCCT (3)	262	22
TTCACTCCAACTTTGCCCTGCC (1)	265	22
TCACTCCAACTTTGCCCTGCCCA (1)	266	23

```
>Pre-NF08   Gm16:19,925,619..19,925,903 [+]     intergênico
CCACTACGCAGGCACGATTTTTTAGATACCTTTGTCAGGCTTTGAATACCATGTTAAATTAGAAACAGAGGAATAATTGATTGAAAATAATTTGCCCGAA
(((((..(((((((((((((((.((((...(((((((((((((((.((((((((((((..((((((((((((.((((((((((((((.....(((..(((
CTACACAACGGAGTCTGTTTTATATAGGATAAATTACATTAATTAATGTAATTAGTCTATATAAACACTCATTCGCTGTGTATTCAAGCACTCGGTTTTT
.((((((.((((((..((((((.(((((((...(((((((((....)))))))))..))))))))))))..).)))).)))))))..)))......))))
CAGTCAGTTATCCGTTTGTTTCTTGTATTTTAACATGGTATCAGGGCCTGATAGAAAGATCTCAGAAGTCGTGTCTGCCCGGTGG
))))))))))))).)))))))).....)))))))))))))))))))))))))))...))))).)))))))))))))..)))))) (-130.60)
```

Sequência (nº de *reads*)	Posição inicial	Tamanho (nt)
CAACGGAGTCTGTTTTATATAGGA (1)	106	24
CAAGAAACAAACGGATAACTGACT (1)	202	24
AACATGGTATCAGGGCCTGATAGA (17)	232	24
AACATGGTATCAGGGCCTGATAGAA (1)	232	25
TGGTATCAGGGCCTGATAGAAAGA (1)	236	24

```
>Pre-NF09a  Gm02:39,925,519..39,925,660 [-]     intergênico
TAGAAATAACTACTTTCTACATTGGTCGTTTAATAACCGATCTAGAATGTAATATTTTCGATGTAGAAAATAT
(((((..(((((((((((((((((((((((((((((((((((((((((((((((((.(((((.....
TACATTCTAGATCGAAAATATTACATTCTAGATCGGTTATTAAACGACCGATGTAGAAAGTAGTTCTTCTA
....)))))).)))))))))))))))))))))))))))))))))))))))))))))))).)))))) (-94.50)
```

Sequência (nº de *reads*)	Posição inicial	Tamanho (nt)
GAAATAACTACTTTCTACATTGGT (1)	3	24
ATAACTACTTTCTACATTGGTCGTT (1)	6	25
AACTACTTTCTACATTGGTCGTTT (1)	8	24
ATTCTAGATCGGTTATTAAACGAC (1)	25	24
CATTCTAGATCGGTTATTAAACGA (1)	26	24
TTAATAACCGATCTAGAATGT (1)	30	21
TTAATAACCGATCTAGAATGTAAT (1)	30	24
AATATTACATTCTAGATCGGTTAT (1)	33	24
AATATTACATTCTAGATCGGTTAT (1)	90	24
TTAATAACCGATCTAGAATGTAAT (1)	93	24
TTAATAACCGATCTAGAATGT (1)	96	21
CATTCTAGATCGGTTATTAAACGA (1)	97	24
ATTCTAGATCGGTTATTAAACGAC (1)	98	24
CATCGGTCGTTTAATAACCGAT (2)	105	22

Sequência (nº de *reads*)	Posição inicial	Tamanho (nt)
GGTTATTAAACGACCGATGTA (1)	108	21
GGTTATTAAACGACCGATGTAG (1)	108	22
GGTTATTAAACGACCGATGTAGAA (2)	108	24
TCTACATCGGTCGTTTAATAAC (1)	109	22
TTTCTACATCGGTCGTTTAATAAC (2)	109	24
GTTATTAAACGACCGATGTAGAAA (1)	109	24
TTATTAAACGACCGATGTAGA (1)	110	21
CTTTCTACATCGGTCGTTTAATAA (1)	110	24
TTATTAAACGACCGATGTAGAAAG (1)	110	24
TATTAAACGACCGATGTAGAA (4)	111	21
TATTAAACGACCGATGTAGAAA (2)	111	22
TATTAAACGACCGATGTAGAAAG (4)	111	23
TATTAAACGACCGATGTAGAAAGT (61)	111	24
ATTAAACGACCGATGTAGAAAGTA (4)	112	24

TTAAACGACCGATGTAGAAAG (1)	113	21	TAAACGACCGATGTAGAAAGTAGT (2)	114	24
TTAAACGACCGATGTAGAAAGT (4)	113	22	AAACGACCGATGTAGAAAGTAGTT (4)	115	24
TAAACGACCGATGTAGAAAGTA (1)	114	22	CTACTTTCTACATCGGTCGTT (1)	116	21

>Pre-NF09b Gm15:31,393,116..31,393,303 [+] intergênico

```
TAGAAAAACTACTTTCTACATCGATCGTTTAATAACTGATCTAGAATGTAATCTTTTCTACGTCGTTGCTTATGTTAGCATGACTTAGAAACAA
((((((((((((((((((((.((((((((((((((((((((((((((((.((((((((((.(((((((((((((((((((((((((..
GGGGTTTCTAAGTCGTGCTGACATAAGCAACCACGTAGAAAATATTACATTCTAGATCAGTTATTAAACGACCGATGTAGAAAGTAGTTATTTCTA
...))))))))))))))))))))))))))).)))))))))).)))))))))))))))))))))))).)))))))))))))))).)))))) (-132.50)
```

Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)	Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)
AACATAAGCAACGACGTAGAAAAG (1)	53	24	**TATTAAACGACCGATGTAGAAAGT** (61)	156	24
TCTACATCGGTCGTTTAATAAC (1)	154	22	ATTAAACGACCGATGTAGAAAGTA (4)	157	24
TTTCTACATCGGTCGTTTAATAAC (2)	154	24	TTAAACGACCGATGTAGAAAG (1)	158	21
GTTATTAAACGACCGATGTAGAAA (1)	154	24	TTAAACGACCGATGTAGAAAGT (4)	158	22
TTATTAAACGACCGATGTAGA (1)	155	21	TAAACGACCGATGTAGAAAGTA (1)	159	22
CTTTCTACATCGGTCGTTTAATAA (1)	155	24	TAAACGACCGATGTAGAAAGTAGT (2)	159	24
TTATTAAACGACCGATGTAGAAAG (1)	155	24	AAACGACCGATGTAGAAAGTAGTT (4)	160	24
TATTAAACGACCGATGTAGAA (4)	156	21	CTACTTTCTACATCGGTCGTT (1)	161	21
TATTAAACGACCGATGTAGAAA (2)	156	22	ACGACCGATGTAGAAAGTAGTTATT (1)	162	25
TATTAAACGACCGATGTAGAAAG (4)	156	23	GAAATAACTACTTTCTACATCGGT (1)	165	24

>Pre-NF09c Gm18:42,937,102..42,937,265 [+] intergênico

```
AGGATCGTAGTAGAAAAACTACTTTCTACATCGGTTGTTTAATAACCGATCTAGAATGTAATATTTTCTACGTCATTGCTTA
(((((..(..(((((((((((((((((((((((((((((((((((((((((((((((((((....(((((.(((..
TGTCAGCATGATTAAAAATGTTACATTCTAGATCGGTTATTAAACGACCGATGTAGAAAGTAGTTATTTCTAAAATAATTTT
....))))))))..)))))))))))))))))))))))))))))))))))))))))))))).)))))..)..)))))) (-90.10)
```

Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)	Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)
TTAATAACCGATCTAGAATGT (1)	39	21	ATTCTAGATCGGTTATTAAACGAC (1)	107	24
TTAATAACCGATCTAGAATGTAAT (1)	39	24	CATCGGTCGTTTAATAACCGAT (2)	114	22
AATATTACATTCTAGATCGGTTAT (1)	42	24	GGTTATTAAACGACCGATGTA (1)	117	21
AGCAATGACGTAGAAAATATTACA (1)	57	24	GGTTATTAAACGACCGATGTAG (1)	117	22
TTAATAACCGATCTAGAATGT (1)	105	21	GGTTATTAAACGACCGATGTAGAA (2)	117	24
CATTCTAGATCGGTTATTAAACGA (1)	106	24	TCTACATCGGTCGTTTAATAAC (1)	118	22

TTTCTACATCGGTCGTTTAATAAC (2)	118	24	ATTAAACGACCGATGTAGAAAGTA (4)	121	24
GTTATTAAACGACCGATGTAGAAA (1)	118	24	TTAAACGACCGATGTAGAAAG (1)	122	21
TTATTAAACGACCGATGTAGA (1)	119	21	TTAAACGACCGATGTAGAAAGT (4)	122	22
CTTTCTACATCGGTCGTTTAATAA (1)	119	24	TAAACGACCGATGTAGAAAGTA (1)	123	22
TTATTAAACGACCGATGTAGAAAG (1)	119	24	TAAACGACCGATGTAGAAAGTAGT (2)	123	24
TATTAAACGACCGATGTAGAA (4)	120	21	AAACGACCGATGTAGAAAGTAGTT (4)	124	24
TATTAAACGACCGATGTAGAAA (2)	120	22	CTACTTTCTACATCGGTCGTT (1)	125	21
TATTAAACGACCGATGTAGAAAG (4)	120	23	ACGACCGATGTAGAAAGTAGTTATT (1)	126	25
TATTAAACGACCGATGTAGAAAGT (61)	120	24	GAAATAACTACTTTCTACATCGGT (1)	129	24

>Pre-NF10 Gm12:648,893..649,030 [+] intergênico (duplex)

TGTTAGACCAGT**CAGCTGAATCCTCTTATGATC**AAGAGCTGTAATGCATATCAGTTTTGTTATCAGTTTGAA

.(((((((((((((.((((.(((((((((((.(((.((((((..(((...(((((..(((....))))).)))))).

CCGTATTTTCTTTGAATTACCGTAACAGCCCTTCA**TCATAGGAGGAATCAACTGGC**TGGTCTAACA

..(((.(((....))).))).)))))))).))).)))))))))))).))).)))))))))))))). (-60.30)

Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)	Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)
TTAGACCAGTCAGCTGAAT (1)	3	19	GAATCCTCTTATGATCAAGAGCT (2)	18	23
CCAGTCAGCTGAATCCTCTTA (1)	8	21	AATCCTCTTATGATCAAGAGC (2)	19	21
AGTCAGCTGAATCCTCTTATG (1)	10	21	TCATAGGAGGAATCAACTGG (1)	108	20
AGTCAGCTGAATCCTCTTATGA (2)	10	22	**TCATAGGAGGAATCAACTGGC** (61)	108	21
GTCAGCTGAATCCTCTTATGA (1)	11	21	ATAGGAGGAATCAACTGGCTGG (1)	110	22
CAGCTGAATCCTCTTATGATC (16)	13	21	TAGGAGGAATCAACTGGCTGG (4)	111	21
GAATCCTCTTATGATCAAGAGC (4)	18	22	GAATCAACTGGCTGGTCTAAC (1)	117	21

>Pre-NF11 Gm14:6,304,120..6,304,369 [-] intergênico

AGCTACATGTTTTACCATCTAAAATGCACATATAAAAAATATAAAAGGAATAAAAGATG**TAGATGAAGTTACTCTGAGCA**CTGAGAGTCATATGTATAGAAATCATTTCCTTGTCA

((((((((((.(((((((..(((((((((((((((((((....((((((..((.((((..((((((((((.(((((((((.(((((.(((((((((.(((((..(((((((...(((

TCTCTATTTATAGAGTTGAGTTGGAGATTTTCA**TACATGTGCCTCTTCGTCGCTC**AGATTAACTTCATCTTGGTCTCTTCACTCCTTACCATTTTTTGTGTGTCTGCAGTTTGTTG

.(((((....)))))..)))...)))))))))))))))))))).)))))..)).)))))))..)))))))))))..))))..)).)))...)))))...)))))))))))))..))).)))...)

ATGGTGATACATGTAGTT

)))))))..)))))))))) (-107.10)

Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)	Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)
TAGATGAAGTTACTCTGAGCA (24)	60	21	AGATGAAGTTACTCTGAGCA (8)	61	20

TGAAGTTACTCTGAGCAC (1)	64	18	CATACATGTGCCTCTTCGTCG (1)	148	21
TGAAGTTACTCTGAGCACTG (9)	64	20	CATACATGTGCCTCTTCGTCGCTCA (20)	148	25
TGAAGTTACTCTGAGCACTGA (1)	64	21	ATACATGTGCCTCTTCGTCG (1)	149	20
TGAAGTTACTCTGAGCACTGAG (16)	64	22	ATACATGTGCCTCTTCGTCGC (3)	149	21
TGAAGTTACTCTGAGCACTGAGA (5)	64	23	ATACATGTGCCTCTTCGTCGCT (4)	149	22
GAAGTTACTCTGAGCACTG (1)	65	19	ATACATGTGCCTCTTCGTCGCTC (1)	149	23
AGTTACTCTGAGCACTGAGAGTC (1)	67	23	**TACATGTGCCTCTTCGTCGCTC** (61)	150	22
TTACTCTGAGCACTGAGAGT (7)	69	20	ACATGTGCCTCTTCGTCGCT (3)	151	20
TTACTCTGAGCACTGAGAGTC (1)	69	21	ACATGTGCCTCTTCGTCGCTC (54)	151	21
CTCTGAGCACTGAGAGTCATA (2)	72	21	ACATGTGCCTCTTCGTCGCTCA (16)	151	22
TCTGAGCACTGAGAGTCATA (1)	73	20	CATGTGCCTCTTCGTCGCTCAG (3)	152	22
TGAGCACTGAGAGTCATATGT (3)	75	21	ATGTGCCTCTTCGTCGCTC (1)	153	19
AGCACTGAGAGTCATATGTAT (1)	77	21	GTGCCTCTTCGTCGCTCA (1)	155	18
AGCACTGAGAGTCATATGTATAGA (1)	77	24	TGCCTCTTCGTCGCTCAGATT (4)	156	21
GCACTGAGAGTCATATGTA (1)	78	19	TGCCTCTTCGTCGCTCAGATTA (5)	156	22
GCACTGAGAGTCATATGTATAGAAA (2)	78	25	TGCCTCTTCGTCGCTCAGATTAA (1)	156	23
ATTTTCATACATGTGCCTCTTCGT (1)	143	24	CTCTTCGTCGCTCAGATTAACT (2)	159	22
TCATACATGTGCCTCTTCGTCG (4)	147	22	TCTTCGTCGCTCAGATTAACTT (3)	160	22
TCATACATGTGCCTCTTCGTCGC (2)	147	23	TTCGTCGCTCAGATTAACTTCATC (4)	162	24
TCATACATGTGCCTCTTCGTCGCT (3)	147	24	TCGTCGCTCAGATTAACTTCATC (4)	163	23
TCATACATGTGCCTCTTCGTCGCTC (19)	147	25	GTCGCTCAGATTAACTTCATC (2)	165	21
TCATACATGTGCCTCTTCGTCGCTCA (1)	147	26	TCGCTCAGATTAACTTCATCT (1)	166	21

>Pre-NF12 Gm20:42,606,806..42,607,020 [+] inrtergenic

GATCGCC**ATTCGCACTGAAATGGATGTCCGTTGTTGTGACTAGTTAATGGGCAT**AAATCTTATTTCCCTCGACGACTTAATGTGTTGACATGTCAATTTTGCCTACAA

((((((.(((((((.((.(((..(((((((((((((((((((((((((.(.((((((((((.((((.((((((((((((.(((((((((((((((....)))

AATAAAGACTGACATGTCAGGACATACTTAAGTCGTCAAAGGGAATAGGATTTAAGGCC**ATTAACTAGTCACAACAATGGA**CACTCATCTCTGTGTGAATAGCGATC

)))......))))))))))).))))...)))))))))..)))))))))))))))).).)))))))))))))))))))))))))..))))..)).)))))))).))))))

(-125.40)

Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)	Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)
ATTCGCACTGAAATGGAT (1)	8	18	CGCACTGAAATGGATGTCCGTT (1)	11	22
ATTCGCACTGAAATGGATGTCCG (1)	8	23	GCACTGAAATGGATGTCCGTTGTT (2)	12	24
ATTCGCACTGAAATGGATGTCCGT (25)	8	24	ACTGAAATGGATGTCCGT (1)	14	18
ATTCGCACTGAAATGGATGTCCGTT (1)	8	25	ACTGAAATGGATGTCCGTTGTTGT (2)	14	24
CGCACTGAAATGGATGTCCG (1)	11	20	CTGAAATGGATGTCCGTTGTT (1)	15	21

TGAAATGGATGTCCGTTGTTGTGA (3)	16	24
AATGGATGTCCGTTGTTGTGA (1)	19	21
AATGGATGTCCGTTGTTGTGACT (2)	19	23
AATGGATGTCCGTTGTTGTGACTA (1)	19	24
ATGGATGTCCGTTGTTGTGACTAG (1)	20	24
GGATGTCCGTTGTTGTGACTAGTT (1)	22	24
TCCGTTGTTGTGACTAGTTAAT (1)	27	22
CCGTTGTTGTGACTAGTTAATGGG (1)	28	24
CGTTGTTGTGACTAGTTAATG (1)	29	21
CGTTGTTGTGACTAGTTAATGG (10)	29	22
CGTTGTTGTGACTAGTTAATGGG (15)	29	23
CGTTGTTGTGACTAGTTAATGGGC (8)	29	24
GTTGTTGTGACTAGTTAATGGG (33)	30	22
GTTGTTGTGACTAGTTAATGGGC (1)	30	23
GTTGTTGTGACTAGTTAATGGGCA (3)	30	24
TTGTTGTGACTAGTTAATGG (1)	31	20
TTGTTGTGACTAGTTAATGGG (9)	31	21
TTGTTGTGACTAGTTAATGGGC (4)	31	22
TTGTTGTGACTAGTTAATGGGCA (4)	31	23
TTGTTGTGACTAGTTAATGGGCAT (52)	31	24
TGTTGTGACTAGTTAATGGG (2)	32	20
TGTTGTGACTAGTTAATGGGCA (2)	32	22
TGTTGTGACTAGTTAATGGGCAT (4)	32	23
TGTTGTGACTAGTTAATGGGCATA (1)	32	24
GTTGTGACTAGTTAATGGGCAT (5)	33	22
CATAAATCTTATTTCCCTCGACGA (1)	52	24
ATAAATCTTATTTCCCTCGACGAC (1)	53	24
TAAATCTTATTTCCCTCGACGACT (1)	54	24
AAATCTTATTTCCCTCGACGACT (1)	55	23
AAATCTTATTTCCCTCGACGACTT (1)	55	24
AATCTTATTTCCCTCGACGACTTA (3)	56	24
TTTCCCTCGACGACTTAATGTGT (1)	63	23
ACTTAATGTGTTGACATGTCAATT (1)	75	24
ACATGTCAGGACATACTTAAGTCGT (1)	120	25
AGTCGTCAAAGGGAATAGGATTTA (1)	139	24
TCGTCAAAGGGAATAGGATTTA (1)	141	22
TCGTCAAAGGGAATAGGATTTAAG (1)	141	24
GTCAAAGGGAATAGGATTTA (2)	143	20
GTCAAAGGGAATAGGATTTAAGGC (1)	143	24
TCAAAGGGAATAGGATTTAAGGCC (1)	144	24
CAAAGGGAATAGGATTTAAGGC (1)	145	22
GAATAGGATTTAAGGCCATTAACT (1)	151	24
GGATTTAAGGCCATTAACTAGTCA (1)	156	24
CATTGTTGTGACTAGTTAATGGCC (1)	164	24
TTGTTGTGACTAGTTAATGG (1)	166	20
CCATTAACTAGTCACAACAAT (1)	166	21
CCATTAACTAGTCACAACAATGGA (3)	166	24
CATTAACTAGTCACAACAATGG (3)	167	22
CATTAACTAGTCACAACAATGGA (1)	167	23
ATTAACTAGTCACAACAATGGA (35)	168	22
TTAACTAGTCACAACAATGGA (12)	169	21
TAGTCACAACAATGGACACTCA (1)	174	22
AGATGAGTGTCCATTGTTGTGACT (1)	175	24
CACAACAATGGACACTCATCTC (1)	178	22
ACAACAATGGACACTCATCTCTGT (1)	179	24
AACAATGGACACTCATCTCTGTG (1)	181	23
AACAATGGACACTCATCTCTGTGTG (1)	181	25
AATGGACACTCATCTCTGTGTGA (1)	184	23
AATGGACACTCATCTCTGTGTGAA (1)	184	24
ACTCATCTCTGTGTGAATAGCGA (1)	191	23
CTCATCTCTGTGTGAATAGCGA (1)	192	22

>Pre-NF13 Gm16:7,236,797..7,236,894 [-] intergênico
GGTTTTTGTAG**GCCGAAGATGAAGAGCTTTGTAT**ATTCTGACACCTCTTGTATGAGATGGATCTTTCATTAAAGGTTTCATCTTTGAGCACAGAAAGC
(.(((((((..((((((((((((..(((((..........((.((((......))).)).........)))))).)))))))))))).)))))))..) (-27.05)

Sequência (nº de *reads*)	Posição inicial	Tamanho (nt)
AGGCCGAAGATGAAGAGCT (2)	10	19
AGGCCGAAGATGAAGAGCTT (1)	10	20
AGGCCGAAGATGAAGAGCTTTG (2)	10	22
AGGCCGAAGATGAAGAGCTTTGTA (3)	10	24
AGGCCGAAGATGAAGAGCTTTGTAT (8)	10	25
AGGCCGAAGATGAAGAGCTTTGTATA (1)	10	26
GGCCGAAGATGAAGAGCT (18)	11	18
GGCCGAAGATGAAGAGCTT (2)	11	19
GGCCGAAGATGAAGAGCTTTG (7)	11	21
GGCCGAAGATGAAGAGCTTTGT (1)	11	22
GGCCGAAGATGAAGAGCTTTGTA (9)	11	23
GGCCGAAGATGAAGAGCTTTGTAT (15)	11	24
GGCCGAAGATGAAGAGCTTTGTATA (15)	11	25
GGCCGAAGATGAAGAGCTTTGTATAT (3)	11	26
GCCGAAGATGAAGAGCTT (6)	12	18
GCCGAAGATGAAGAGCTTT (3)	12	19
GCCGAAGATGAAGAGCTTTG (33)	12	20
GCCGAAGATGAAGAGCTTTGT (1)	12	21
GCCGAAGATGAAGAGCTTTGTA (24)	12	22
GCCGAAGATGAAGAGCTTTGTAT (50)	12	23
GCCGAAGATGAAGAGCTTTGTATA (27)	12	24
GCCGAAGATGAAGAGCTTTGTATAT (22)	12	25

Sequência (nº de *reads*)	Posição inicial	Tamanho (nt)
GCCGAAGATGAAGAGCTTTGTATATT (10)	12	26
CCGAAGATGAAGAGCTTT (1)	13	18
CCGAAGATGAAGAGCTTTG (1)	13	19
CCGAAGATGAAGAGCTTTGTA (1)	13	21
AAGATGAAGAGCTTTGTAT (1)	16	19
AGATGAAGAGCTTTGTATA (1)	17	19
GAGCTTTGTATATTCTGACACCT (1)	24	23
GAGCTTTGTATATTCTGACACCTCT (1)	24	25
CTTTGTATATTCTGACACCTCTTG (2)	27	24
TTGTATATTCTGACACCTCTTG (1)	29	22
TTGTATATTCTGACACCTCTTGT (1)	29	23
TGTATATTCTGACACCTCT (2)	30	19
TGTATATTCTGACACCTCTTG (1)	30	21
GTATATTCTGACACCTCTTG (2)	31	20
GTATATTCTGACACCTCTTGT (1)	31	21
TATATTCTGACACCTCTTG (2)	32	19
TATTCTGACACCTCTTGTAT (1)	34	20
TATGAGATGGATCTTTCATT (1)	51	20
TAAAGGTTTCATCTTTGA (1)	70	18
TAAAGGTTTCATCTTTGAGCACAGAA (1)	70	26
AAAGGTTTCATCTTTGAGCACAG (1)	71	23
AAAGGTTTCATCTTTGAGCACAGAAA (1)	71	26
AGGTTTCATCTTTGAGCACAGAAA (1)	73	24

```
>Pre-NF14   Gm17:27,739,062..27,739,312 [+]     intron_ Glyma17g26570
TGATTTTTGTTGAGTCCCTTATAAATTTTTTCATTGTTTCGAGTCTCTGATATATTAAAAGT
```

TTTGTTCTGGATCCCTGTCGTC

```
AGTTAAGTGATGACGTGACACATCTCAATGAT
.((((((((((((..((.(((((((((((.(((((..((((..(((.((((((((((((((((((((..((.(((((((((((((((((((.(((((((((((((..((
TGATATATATTGGAAGAGTTGTTGAGATGGTGTCACATCATCACTTAACTGACGACAGAGACTCGGAACAAAACTTTTAATATACTAGGAACTCAGAACAACGGAAAAATTTATTA
(..............))).)).)))))).)))))).)))))))))))))))))))).))..)))))))))))))))))))).))))..))))..))))).)))))))))))).)
GAACTCAAAACAAAAATCA
).))))))..)))))))))). (-132.36)
```

Sequência (nº de *reads*)	Posição inicial	Tamanho (nt)
TAAATTTTTTCATTGTTTCGAGT (1)	22	23
ACTTTTAATATATCAGAGACTCGA (1)	39	24
AGTTTTGTTCTGGATCCCTGTC (2)	60	22

Sequência (nº de *reads*)	Posição inicial	Tamanho (nt)
GTTTTGTTCTGGATCCCTGTCGTC (2)	61	24
TTTGTTCTGGATCCCTGTCGT (3)	63	21
TTTGTTCTGGATCCCTGTCGTC (48)	63	22

TTTGTTCTGGATCCCTGTCGTCAG (3)	63	24
TTGTTCTGGATCCCTGTCGTC (1)	64	21
TTGTTCTGGATCCCTGTCGTCA (1)	64	22
TTGTTCTGGATCCCTGTCGTCAGT (1)	64	24
AACTGACGACAGGGATCCAGAACA (1)	65	24
TCTGGATCCCTGTCGTCAGTTAAG (1)	68	24
TGGATCCCTGTCGTCAGTTAA (1)	70	21
TGGATCCCTGTCGTCAGTTAAG (2)	70	22
ATCACTTAACTGACGACAGGG (3)	75	21
GTCAGTTAAGTGATGACGTGA (5)	82	21
CACATCATCACTTAACTGA (1)	150	19
ACATCATCACTTAACTGACGA (1)	151	21
ATCACTTAACTGACGACAGAG (1)	156	21
ATCACTTAACTGACGACAGAGA (1)	156	22
AGTCTCTGTCGTCAGTTAAGTGAT (1)	156	24
TCACTTAACTGACGACAGAGA (1)	157	21

CTTAACTGACGACAGAGACTCG (1)	160	22
TAACTGACGACAGAGACTCGGAAC (3)	162	24
AACTGACGACAGAGACTCGGAACA (1)	163	24
TGACGACAGAGACTCGGAA (1)	166	19
TGACGACAGAGACTCGGAACAAAAC (1)	166	25
GACGACAGAGACTCGGAACA (2)	167	20
GACGACAGAGACTCGGAACAA (1)	167	21
GACGACAGAGACTCGGAACAAAA (1)	167	23
GACGACAGAGACTCGGAACAAAAC (5)	167	24
ACGACAGAGACTCGGAACAAAAC (1)	168	23
CGACAGAGACTCGGAACAAAACTT (1)	169	24
CGACAGAGACTCGGAACAAAACTTTT (1)	169	26
GACAGAGACTCGGAACAAAAC (1)	170	21
ACAGAGACTCGGAACAAAACTT (1)	171	22
CAGAGACTCGGAACAAAACTTT (3)	172	22
TCGGAACAAAACTTTTAATATA (1)	179	22

>Pre-NF15   Gm03:20,886,972..20,887,185 [-]   intron_Glyma03g16440

```
CTTTTCCCTTTCTACATTTGGATATTTACCGAGTATTAGTTCTGCCATTGGTTCTCAATCCCGTCCCTAAGGGTTTTTAGATCACCCCTTCCCTCCTTTCCTTGACCGTAGACTCT
(((((((((((((((((((((.((((((((.(((((((.((((((((.((((((((((.((.(((((..(((((((...(.(((..................))).)..)))))))
GTCCTTTATGCATCATTGGGTCAATCCCATGGGACTGAGTTTGGGAACTATTGGCGGAAGTAATACTAGGTAAATAACCAGATGTAGAAAGGGAAAAG
................((((.....)))).))))).).).)))))))))).)))))))).))))))).))))))))))))))))))))))))))))) (-100.27)
```

(In the sequence above, **TTGGCGGAAGTAATACTAGGTA** is highlighted in bold.)

Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)
CTACATTTGGATATTTACCGAG (2)	12	22
TACCGAGTATTAGTTCTGCCATT (1)	27	23
ATTGGCGGAAGTAATACTAGGTAA (1)	166	24
TTGGCGGAAGTAATACTAGGT (3)	167	21
TTGGCGGAAGTAATACTAGGTA (46)	167	22
TTGGCGGAAGTAATACTAGGTAA (1)	167	23
TTGGCGGAAGTAATACTAGGTAAA (3)	167	24
TGGCGGAAGTAATACTAGGTA (3)	168	21

Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)
CGGAAGTAATACTAGGTA (1)	171	18
AGGTAAATAACCAGATGTAGAA (1)	184	22
AGGTAAATAACCAGATGTAGAAAG (1)	184	24
AAATAACCAGATGTAGAAAGGG (1)	188	22
AAATAACCAGATGTAGAAAGGGA (5)	188	23
AAATAACCAGATGTAGAAAGGGAA (1)	188	24
AATAACCAGATGTAGAAAGGG (1)	189	21
AATAACCAGATGTAGAAAGGGA (9)	189	22
TAACCAGATGTAGAAAGGGA (1)	191	20

\>Pre-NF16   Gm17:14,170,467..14,170,651 [-]   intron_Glyma17g17360

ATGTCTATTAGAGGATATCAATGTTCAATTGGGCGTATTAACTGAGAGTGTGAGAGGTGAAGGAAGCTAATCTTGACCATTCAATCTTAA
((((..(((.(((((((..((((.((....((((.(((((..((((((((.(((((((((((((((((((((.(((.(((..((((((..
TAGGGATTCGGTGTTCGTGA**TTAGCTTCTTTCACCTTTCCC**ACTTTCACCTAATAGGCTCTCCCAGATTCTATTTACTGTCTTTAAAGTTTACAT
..))))))..))).))).))))))))))))))))))))).))))))))..))))).)))).....))...))))...)))))))..)))..)))) (-71.50)

Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)
AATTGGGCGTATTAACTGAGAGTGT (3)	27	25
TTAGCTTCTTTCACCTTTCC (1)	111	20
TTAGCTTCTTTCACCTTTCCC (40)	111	21

\>Pre-NF17   Gm03:46,862,468..46,862,679 [+]   intron_Glyma03g41350

TTATTTTTAAGGTTCAATTTGATTTTCTTA**TTTTTAAAAGGTTCAGTTAGGT**CCTTTATTTTTTAAAATGGGTTCAATTTGGTCCCACTTAAACTCATTTTTTGAAA
.(((((((..(.((((((.((.(((..(((((((((((((((((((((((((.((((((((((((.(((((......((((((((((.(((.....((((((.
ATGTTCGAAGCTGTTGGAAGAAAATGGGACCAACCCAATTTAAAAAATAAAAGACCTAATTGAACCTTTTAAAAATAAAGGGACCATATTGAAGCCAAAAAATAA
...)))))))..)))..)))....))))))))))))).))))))))))))).)))))))))))))))))))))))))..))).)).)))))).)..))))))). (-84.40)

Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)
ACTGAACCTTTTAAAAATAAGA (1)	26	22
TTTTTAAAAGGTTCAGTTAG (1)	31	20
TTTTTAAAAGGTTCAGTTAGGT (32)	31	22
TTTTAAAAGGTTCAGTTAGGT (1)	32	21
AAAAAATAAAAGACCTAATTGAAC (1)	149	24

\>Pre-NF18   Gm13:34,948,270..34,948,416 [-]   intron_Glyma13g32940

CCGATTTTTGTTTTGAATTTTCGTTGT**TCTGAGTCCATGATATATTAAA**AGTTTTGTTCCTTGTCCCTGCAGTCAATTAAGTGATAACGTGGTACCATCCTAACA
((((((((((((((.((((((((((.(((((((.(((((((((((((((((((((.(..((.((((..((.(.((((((((...(((..(..(((.((((.
TATATTGAAAGAGCCGTTGAGATAATGTCACATCATCACTTAACTCACAACAGGAACTTGAAACAAAACTTTTAATATA**TCAGGGATTCAAAACAACGAAA**AAAT
....((....))...)))).)))..)..)))...))))))))).).))..)))).))..).))))))))))))))))))))).))))))).)))))))))))...
TTATCAAGGACTCAACAAAAATTGG
...))))......))))))))))) (-84.50)

Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)
TTTCGTTGTTCTGAGTCCATGA (3)	19	22
TTTCGTTGTTCTGAGTCCATGATA (1)	19	24

Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)
TTGTTCTGAGTCCATGATATATT (1)	24	23
TGTTCTGAGTCCATGATA (1)	25	18

TGTTCTGAGTCCATGATATATT (12)	25	22
GTTCTGAGTCCATGATATATTAAA (3)	26	24
TCTGAGTCCATGATATAT (1)	28	18
TCTGAGTCCATGATATATTAA (2)	28	21
TCTGAGTCCATGATATATTAAA (20)	28	22
CTGAGTCCATGATATATTAAA (3)	29	21
ATTAAAAGTTTTGTTCCTTGTCCC (1)	44	24
TTAAAAGTTTTGTTCCTTGTC (1)	45	21
TTAAAAGTTTTGTTCCTTGTCCCT (2)	45	24
TAAAAGTTTTGTTCCTTGTCCC (1)	46	22
TAAAAGTTTTGTTCCTTGTCCCTG (1)	46	24
AAAAGTTTTGTTCCTTGTCCCTGC (4)	47	24
AAAGTTTTGTTCCTTGTCCCTG (1)	48	22
AGTTTTGTTCCTTGTCCCTGCAGT (2)	50	24
AGTCAATTAAGTGATAACGTGGTA (1)	71	24
ACAAAACTTTTAATATATCAGGGA (1)	168	24
CAAAACTTTTAATATATCAGGGA (1)	169	23
CTTTTAATATATCAGGGATTCA (1)	174	22
TTTTAATATATCAGGGATTCAAA (1)	175	23
TTAATATATCAGGGATTCAAA (1)	177	21

TAATATATCAGGGATTCAAAACA (1)	178	23
TAATATATCAGGGATTCAAAACAA (6)	178	24
ATATATCAGGGATTCAAAACAA (1)	180	22
ATATATCAGGGATTCAAAACAACG (2)	180	24
TATATCAGGGATTCAAAACAA (4)	181	21
TATATCAGGGATTCAAAACAACG (2)	181	24
ATATCAGGGATTCAAAACAACGAA (4)	182	24
TATCAGGGATTCAAAACAACGA (12)	183	22
TATCAGGGATTCAAAACAACGAA (4)	183	23
TATCAGGGATTCAAAACAACGAAA (4)	183	24
ATCAGGGATTCAAAACAACGAA (7)	184	22
ATCAGGGATTCAAAACAACGAAA (1)	184	23
TCAGGGATTCAAAACAACGA (1)	185	20
TCAGGGATTCAAAACAACGAA (7)	185	21
TCAGGGATTCAAAACAACGAAA (30)	200	21
AACGAAAAAATTTATCAAGGAC (2)	200	22
AACGAAAAAATTTATCAAGGACT (7)	200	23
AAAAATTTATCAAGGACTCAACAA (1)	205	24
AAAATTTATCAAGGACTCAACAAA (1)	206	24
ATTTATCAAGGACTCAACAAAAAT (1)	209	24

>Pre-NF19 Gm14:4,584,761..4,585,058 [-] intron_ Glyma14g06340 (duplex)

TCTAGTTACATATTGCTTGATGACATGTGG**AAGGCACCACTTCAGCAATGGA**AGATGAGCCAACAGAAGATGCCACAAGGGAGATGGACGGTCCCTGATAAA
(((((((((..((((((((((((((.(((((((((.(((((((((((((((((((((((((((..((((((((((((((((.(((((..(((((((((((..
TCCGAACAGGTTCAGCCTGGAACCTAGAGAGCCTCACAATACACAAGTATTTGTATGAAAAGTTCAAGTCTGTCTATTCCATTGCAATTTCGTGTATCAGGG
..((((.((((((......))))))...((((......(((((........))))).....))))......................))))..)))))))))
ACTATCCATCAACCCTTGTGGCGTCTTCCTTTGGCTTATCTTCCA**TTGCTGGACGTGGCGCCTTCCA**CTTGTCATCAAGCGATCCGTAACTAGA
))..))))))..))))))))))))))))..))))))))))))))))))))))).)))).))))))))).))))))))))))))..))))))))) (-177.70)

Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)
ATATTGCTTGATGACATGTGGA (10)	10	22
ATATTGCTTGATGACATGTGGAA (1)	10	23
ATATTGCTTGATGACATGTGGAAG (2)	10	24
TATTGCTTGATGACATGTGGA (3)	11	21
TATTGCTTGATGACATGTGGAAGG (6)	11	24
ATTGCTTGATGACATGTGGA (1)	12	20

Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)
ATTGCTTGATGACATGTGGAAGGC (1)	12	24
TGCTTGATGACATGTGGAAGGC (4)	14	22
TTGATGACATGTGGAAGGCAC (3)	17	22
TGATGACATGTGGAAGGCACC (2)	18	21
TGACATGTGGAAGGCACCACT (2)	21	21
GAAGGCACCACTTCAGCAATGGA (10)	30	23

GAAGGCACCACTTCAGCAATGGAA (1)	30	24
AAGGCACCACTTCAGCAATGGA (29)	31	22
AGGCACCACTTCAGCAATGGA (12)	32	21
AGGCACCACTTCAGCAATGGAA (1)	32	22
AGGCACCACTTCAGCAATGGAAGA (2)	32	24
TGAGCCAACAGAAGATGCCACAAG (1)	56	24
CCAACAGAAGATGCCACAAGGG (2)	60	22
AGATGCCACAAGGGAGATGGACGG (1)	68	24
TGCCACAAGGGAGATGGACGG (1)	71	21
CACAAGGGAGATGGACGGTCCC (1)	74	22
CAAGTCTGTCTATTCCATTGCAAT (1)	167	24
AGGGACTATCCATCAACCCTTGTGG (1)	201	25
GACTATCCATCAACCCTTGTGG (1)	204	22
GGCGTCTTCCTTTGGCTTATCTTC (1)	224	24
CTTCCTTTGGCTTATCTTCCA (1)	229	21
TTCCTTTGGCTTATCTTCCAT (1)	230	21
TTTGGCTTATCTTCCATTGCTGGA (2)	234	24

TATCTTCCATTGCTGGACGTGGCGC (1)	241	25
CATTGCTGGACGTGGCGCCTT (2)	248	21
CATTGCTGGACGTGGCGCCTTC (6)	248	22
CATTGCTGGACGTGGCGCCTTCC (5)	248	23
CATTGCTGGACGTGGCGCCTTCCA (7)	248	24
ATTGCTGGACGTGGCGCCTTC (1)	249	21
ATTGCTGGACGTGGCGCCTTCC (1)	249	22
ATTGCTGGACGTGGCGCCTTCCA (2)	249	23
TTGCTGGACGTGGCGCCTTCCA (13)	250	22
TGCTGGACGTGGCGCCTTCCA (3)	251	21
TGCTGGACGTGGCGCCTTCCAC (4)	251	22
TGCTGGACGTGGCGCCTTCCACT (1)	251	23
TGCTGGACGTGGCGCCTTCCACTT (4)	251	24
GCTGGACGTGGCGCCTTCCA (1)	252	20
GCTGGACGTGGCGCCTTCCACT (2)	252	22
CTGGACGTGGCGCCTTCCACT (1)	253	21
TGGACGTGGCGCCTTCCACTTGT (1)	254	23

>Pre-NF20 Gm20:44,450,038..44,450,191 [-] intergênico

GTTGTTGCTTGGTCTGCTTTGATATATGATTGACAATCCGAATAGCCTAT**ACAGATTGACAATCCATGTGAGCTA**TAGGGA

....((((...((((((...(((((........(((((((.((((((((((..((((((.((((.(((((((((((.(((.

ATCCGTATAGTTCACATAGATTATCAATCCATATAGGCTATACGGATTGATATTTCGCAGATATGTAGAAAAC

..)).)))))))))))).))))..))))))..))))))))))..)))))))))))..))))))..))))..... (-66.40)

Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)
TCCGAATAGCCTATACAGATTG (1)	37	22
CGAATAGCCTATACAGATTGA (1)	39	21
ACAGATTGACAATCCATGTGA (2)	51	21
ACAGATTGACAATCCATGTGAGCT (1)	51	24
ACAGATTGACAATCCATGTGAGCTA (28)	51	25
CAGATTGACAATCCATGTGAGCTA (1)	52	24

Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)
TTATCAATCCATATAGGCT (1)	102	19
TTATCAATCCATATAGGCTAT (1)	102	21
TATCAATCCATATAGGCTATA (2)	103	21
TATCAATCCATATAGGCTATACGG (1)	103	24
ATCAATCCATATAGGCTATACGGA (2)	104	24
CAATCCATATAGGCTATACGG (1)	106	21
AATCCATATAGGCTATACGGA (8)	107	21

>Pre-NF21 Gm20:13,840,325..13,840,460 [+] intron_Glyma20g10030
TTTTTTAGTTTACTACCTTTTTTTTATTTTACTACCTCACATTTTCAAATTTTTACTTTTGATATCCGCCGTTAATCTAAGTCTATTAAGTAATGATGTGACACTCTGTTAGCCTA
.(((((((((((((((.((((((((((((((((((.((.((((((((((((((((.(((((((((.(.(((((((.(((..(((..(((((((.((((((((((....(((((((.((
TCACATCATCACTTAACCGTTGATGTGATACTTTGTTAGCTTGTCATGTCATCACTTAACGGACCGAGACTAACGGC**AGTTATCAAAAGCAAAAGTTTGGA**AATGTCAGATAGTAA
((((((((.(........).)))))))))....))))))..)))))))))).))))))..)))..)))..)))))))..).)))))))).)))))))))))))))))).)).))))))
AATGAAAAAAAAATATTGTTTAGTAGTAAACTGAAAAA
))))))))))...........)))))))))))))). (-115.20)

Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)	Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)
TTTTTACTTTTGATATCCGCCGTT (2)	50	24	GCAAAAGTTTGGAAATGTCAGATA (1)	205	24
ACTTTTGATATCCGCCGTTAAT (1)	55	22	AAGTTTGGAAATGTCAGATAGTAA (1)	209	24
ACTTTTGATATCCGCCGTTAATCT (3)	55	24	AGTTTGGAAATGTCAGATAGTAAA (2)	210	24
TGTCATGTCATCACTTAACGGA (1)	158	22	TTTGGAAATGTCAGATAGTAA (1)	212	21
TCATCACTTAACGGACCGAGACT (1)	165	23	TTGGAAATGTCAGATAGTAAAA (1)	213	22
ACTAACGGCAGTTATCAAAAGCA (1)	185	23	GAAATGTCAGATAGTAAAATGAA (1)	216	23
AACGGCAGTTATCAAAAGCAAAAG (1)	188	24	AAATGTCAGATAGTAAAATGAAAA (1)	217	24
AGTTATCAAAAGCAAAAGTTTGGA (28)	194	24	AATGTCAGATAGTAAAATGAAAAA (8)	218	24

>Pre-NF22 Gm16:6,256,970.. 6,257,129 [+] intron_Glyma16g06930
GGAGAAGACATTATTATATAGTTTGGGACCATGTGGTTGTGTCCAATCTCTATGTGACACAAAATTATATTGTTTA
((((((....((((((((((((..(((.(((((((((.((((((.((((((((((((((((.((((...((((...
ATTTATAAAAAGA**GTTGATGTGTCACATGGAGATGGA**TAGGAAGTGAGGACCACATGATTTCTGGACTATATAATAAATTCTCC
....))))....))))..)))))))))))))))))).......).)))))))))..))))..))))))))))))).)))))) (-68.40)

Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)	Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)
ACATAGAGATTGGACACAACCACA (1)	32	24	AGATGGATAGGAAGTGAGGACCAC (1)	107	24
ATGTGACACATCAACTCTTTTTAT (1)	81	24	ATGGATAGGAAGTGAGGACCACAT (1)	109	24
AAAAAGAGTTGATGTGTCACATGGA (1)	83	25	GGATAGGAAGTGAGGACCACATGA (1)	111	24
AAAAGAGTTGATGTGTCACATGGA (3)	84	24	GATAGGAAGTGAGGACCACATGAT (1)	112	24
CATCTCCATGTGACACATCAACT (1)	89	23	ATAGGAAGTGAGGACCACATGATT (1)	113	24
GTTGATGTGTCACATGGAGATGGA (27)	90	24	GGAAGTGAGGACCACATGATTTCT (1)	116	24
TGATGTGTCACATGGAGATGGAT (1)	92	23	AAGTGAGGACCACATGATTTCTGG (1)	118	24
TGATGTGTCACATGGAGATGGATA (1)	92	24	AGTGAGGACCACATGATTTCTGGA (4)	119	24
ACATGGAGATGGATAGGAAGTGAG (1)	101	24	GTGAGGACCACATGATTTCTGGAC (2)	120	24
GGAGATGGATAGGAAGTGAGGAC (1)	105	23	GTCCAGAAATCATGTGGTCCTCAC (1)	120	24
GGAGATGGATAGGAAGTGAGGACC (1)	105	24	GTGAGGACCACATGATTTCTGGACT (1)	120	25

TGAGGACCACATGATTTCTGGACT (1)	121	24	AGGACCACATGATTTCTGGACTAT (4)	123	24
GAGGACCACATGATTTCTGGACTA (2)	122	24	ATGATTTCTGGACTATATAATAAAT (1)	131	25
GAGGACCACATGATTTCTGGACTAT (1)	122	25	ATTTCTGGACTATATAATAAATT (1)	134	23

>Pre-NF23 Gm10:37,397,199..37,397,385 [-] intergênico
GAGGAGAGATCAAGTTACACCAAGAGTAACTATAAGAAAGTTACTCTTCACCCTCACCATTCATTTTCTTGAAATCTAATGGTTGCAATGAGT
((((((.((((((((((...((.(((((((.(((((((((((((.((((.(((((((((((((((.(((((((.(((((((((....((..
AACTAATCTTAACCATTAAATTTCAACAAAAT**GAATGGTGAGGATGAAAAGTAACT**TTCTTAGAGTTACTCATGATTTAACTTGATCCCTCCTC
......)).)))))))))).)))))))..)))))))))))))))).))))..))))))))))))))).))))))))..)...)))))))))).)))))) (-99.20)

Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)	Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)
GAAAGTTACTCTTCACCCTCAC (1)	36	22	ATTAAATTTCAACAAAATGAATGG (1)	108	24
AAGTTACTCTTCACCCTCACCATT (1)	38	24	TTTCAACAAAATGAATGGTGA (1)	114	21
TTACTCTTCACCCTCACC (1)	41	18	CAAAATGAATGGTGAGGATGA (2)	120	21
CAAGAAAATGAATGGTGAGG (2)	52	20	CAAAATGAATGGTGAGGATGAA (2)	120	22
TCAAGAAAATGAATGGTGAGG (1)	52	21	**GAATGGTGAGGATGAAAAGTAACT** (27)	126	24
CAAGAAAATGAATGGTGAG (1)	54	21	AATGGTGAGGATGAAAAGTAACT (2)	127	23
CATTTTCTTGAAATCTAATGG (1)	62	21	TGGTGAGGATGAAAAGTAACT (1)	129	21
			AACTTTCTTAGAGTTACTCATGA (1)	146	23

>Pre-NF24a Gm07:44,231,008..44,231,088 [+] intergênico
TTTCTAATGTCAAAA**GTTGGACTCAAGGAACCTA**TGCTTTCATCAGGGTTCTTTTGAGTTCCAACACCTTTGCTTTGGAAA
.(((((((.(.((((.((((((((((((((((((((....)))..)))))))..)))))))..)))))))...))))).))))))). (-27.70)

Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)
CAAAAGTTGGACTCAAGGAACCTA (1)	11	24
GTTGGACTCAAGGAACCT (7)	16	18
GTTGGACTCAAGGAACCTA (11)	16	19
GTTCCAACACCTTTGCTTTGGA (1)	58	22

>Pre-NF24b Gm17:493,663..493,743 [-] intergênico
TTTCTAATGTCAAAA**GTTGGACTCAAGGAACCTA**TGCTTTCATCAGGGTTCTTTTGAGTTCCAACACCTTTGCTTTGGAAA
.(((((((.(.((((.((((((((((((((((((((....)))..)))))))..)))))))..)))))))...))))).))))))). (-27.70)

Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)
CAAAAGTTGGACTCAAGGAACCTA (1)	11	24

GTTGGACTCAAGGAACCT (7)	16	18
GTTGGACTCAAGGAACCTA (11)	16	19
GTTCCAACACCTTTGCTTTGGA (1)	58	22

>Pre-NF25 Gm03:2,812,304..2,812,468 [+] intergênico

```
TTCAGTTTAGTATTTAAAATATCAAGCTGTTTACGGTTATCCAATTGAGATTCAAGCACTTTTCTATGTGGCTTTAAGGTGTGTTT
.((((((...........(((((((.((((((.(((((((.(((((((((((((((((((((.(((((.((((((((((....
GAATGGAACACCTTAAATCCATGAAGAAAAGTGTTTGAATCTCAATTAGATAACCATAGACAACTTGATATGACATGAA
.......))))))))).)))))).)))))))))))))))))))).)))))))).)))))).)))))))))).))). (-77.60)
```

Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)
CTGTTTACGGTTATCCAATTGAGA (1)	27	24
TTTACGGTTATCCAATTGAGA (5)	30	21
TTACGGTTATCCAATTGAGAT (1)	31	21
TTATCCAATTGAGATTCAAGCACT (1)	37	24
AAAAGTGCTTGAATCTCAATTGGA (1)	40	24
CCAATTGAGATTCAAGCACT (1)	41	20
CCAATTGAGATTCAAGCACTTT (2)	41	22
CCAATTGAGATTCAAGCACTTTT (1)	41	23
CAATTGAGATTCAAGCACTTTTCT (1)	42	24
AATTGAGATTCAAGCACTT (1)	43	19
TTGAGATTCAAGCACTTTT (1)	45	19
TTGAGATTCAAGCACTTTTCTATG (1)	45	24

Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)
TGAGATTCAAGCACTTTTCTATGT (1)	46	24
AGATTCAAGCACTTTTCTAT (1)	48	20
AGATTCAAGCACTTTTCTATG (4)	48	21
TAAATCCATGAAGAAAAGTGTT (1)	100	22
AAGAAAAGTGTTTGAATCTCA (1)	110	21
AAAAGTGTTTGAATCTCAATTAG (1)	113	23
AAAAGTGTTTGAATCTCAATTAGA (2)	113	24
AAAGTGTTTGAATCTCAATT (1)	114	20
AAAGTGTTTGAATCTCAATTAGA (18)	114	23
AAAGTGTTTGAATCTCAATTAGAT (18)	114	24
TGTTTGAATCTCAATTAGAT (1)	118	20
ATGGTTATCTAATTGAGATTCAAA (1)	120	24
ATTAGATAACCATAGACAACTTGA (1)	131	24

>Pre-NF26 Gm04:5,821,402..5,821,618 [+] intron_Glyma04g07460

```
GTGTTTTTTTTTGTTTTTCATCTTTAGCAAAATTATTTTTGTTTTTAGTTCTTGTAAATTATGTTTGTTTTGTTTTTTGTCCTTATAGCACTTTAGATAATGTTTTTT
...((((((((((((.....(((((.(.(((((((((((((((((((.(((((((((((((((((((((((..(((((((.(((((.(((((((.((((....
TCATTGTTCACAACACTGTCTAATGTGCTTTAAGGACTAAAACAAAACAAACATAATTTATAGGATTAAAAGCAAAAAAATAATTTTCCAAGGAAGCAAAAAAAAAAAC
..........))))..)))))))..)))))..)))))))..)))))))))))))))))))))))))))))))))..)))))))))))).)))))))))))))))...
(-83.06)
```

Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)
ATGTGCTTTAAGGACTAAAACAAA (2)	131	24
TTAAGGACTAAAACAAAACAAA (1)	138	22
TTAAGGACTAAAACAAAACAAACA (13)	138	24

```
>Pre-NF27   Gm10:47,656,000..47,656,302 [+]     intergênico
CATTTCGTTTCTTGATTGGATGTCACTCCTTGAGCCTTGCAGCAGTTATGACAACAATTATGGGGTTTTGTGTTGTAATCATGACTATTTAAGCTTTTGGCTGAGGTTCCATTTTT
(((.((((((((((((((((((..(((.(((((((((((((((((((((.((((((.(((((((..((((.(((((((...(((((((..((((.(((((.((..((.((((
ACTCTTGGTTCCAAGGTGGAAACATTGGTTTTGCCCGTAAAAATGTCTATTTACACCTTGGAATCAGAGGTGAAATTGACACATCGGCAAAAAAATTAAATTGGCATGATTGGAAT
((..((((((((((((((.........(((((....)))))).........)))))))))))))..))))))..))..)).)))))..))))).))))..)))))))...)))))))).)))
GTGAAATTCCTTGATTGCTGTCATAACTGCTGCAAGGCTTGTGGATGGACATCCAATCAGGAAATGACATG
)..)))))))..))))))..)))))))))))))))))))))))..)))..)))))))))))))))))))))..))) (-154.50)
```

Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)	Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)
TGTCACTCCTTGAGCCTTGCA (1)	21	21	GTCATAACTGCTGCAAGGCTT (1)	252	21
TGAGCCTTGCAGCAGTTATGACAA (10)	31	24	ACTGCTGCAAGGCTTGTGGATGGA (1)	258	24
CACATTCCAATCATGCCAA (1)	217	19	TGCTGCAAGGCTTGTGGATGGA (1)	260	22
GATTGCTGTCATAACTGCTGC (1)	245	21	TGCAAGGCTTGTGGATGGACATC (1)	263	23
TGCTGTCATAACTGCTGCAAGGCT (1)	248	24	TGCAAGGCTTGTGGATGGACATCC (1)	263	24
			CAAGGCTTGTGGATGGACATC (1)	265	21

```
>Pre-NF28   Gm08:11,432,998..11,433,200 [-]     intergênico
TTTTCAATTTTGATCCTTGTAAGATATTTTGCTCATTTTTAGTCCTGTAAGTTTGTGTTTTTCAAATTTTGGTTCTTGTAAGCACGGACTTATGGGGACTGTA
.(..(((((((..(.((((((((((((((((((((((((((((((((((((((((((.((((((((((..((((......))))..))))))...
AATGAAAAAATTAGAATCTTACAAGTTCAAAAATTGGAAAAATGCAAACTTACATGGACTAAAAATGAGCAAAATATCTTACAGAGTCCAAAATTGGGAA
.............((((.......)))).)))))).)))))))))))))))).))))))))))))))))))))))))))).)..)))))))..). (-91.40)
```

Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)	Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)
TTTGCTCATTTTTAGTCCTGTA (1)	28	22	AACTTACATGGACTAAAAATGAGCA (4)	150	25
TGCTCATTTTTAGTCCTGTAA (1)	30	21	ACTTACATGGACTAAAAATGAG (5)	151	22
TGCTCATTTTTAGTCCTGTAAGT (10)	30	23	CTTACATGGACTAAAAATGAGCA (5)	152	23
TGCTCATTTTTAGTCCTGTAAGTT (1)	30	24	CTTACATGGACTAAAAATGAGCAAA (3)	152	25
CTCATTTTTAGTCCTGTAAGTT (1)	32	22	TTACATGGACTAAAAATGAGCA (5)	153	22
TCATTTTTAGTCCTGTAAGTT (1)	33	21	TTACATGGACTAAAAATGAGCAA (1)	153	23
CATTTTTAGTCCTGTAAGTTTGT (1)	34	23	**TTACATGGACTAAAAATGAGCAAA** (81)	153	24
GGGGACTGTAAATGAAAAAAT (1)	94	21	TACATGGACTAAAAATGAGC (1)	154	20
TGGAAAAATGCAAACTTACATGGAC (2)	138	25	TACATGGACTAAAAATGAGCAA (1)	154	22
GGAAAAATGCAAACTTACATGGA (1)	139	23	TACATGGACTAAAAATGAGCAAA (13)	154	23
GGAAAAATGCAAACTTACATGGACT (2)	139	25	TACATGGACTAAAAATGAGCAAAA (2)	154	24
AAACTTACATGGACTAAAAATGAG (1)	149	24	ACATGGACTAAAAATGAGCAAA (1)	155	22
AACTTACATGGACTAAAAATGAG (9)	150	23	ATGGACTAAAAATGAGCAAA (1)	157	20

>Pre-NF29 Gm13:41,358,321..41,358,477 [+] intron_Glyma13g40930
TGAACTGCATACACCGCCGCCTGGCCCGACTGATCTGCATACATATCGTGTTGCCAAGCCGTCTCCATCAAAACAG
.((((((((((.(.((.((.((..(.((((((((((.((((((((((.(((((..((.(((((((........(
GGACCAAGCGTGCCTGGAGACAGC**AAGGCAGAACGATATGTACGCAGA**TCGGTTCGATCATGCAGCAGCGTATGCAGTTCA
 (.((.....)).)))))))))..))..)))))..)))))))))).)))))))))))..)..)).)).)).).)))))))))))). (-85.40)

Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)	Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)
GGCCCGACTGATCTGCATACATATCG (1)	23	26	AAGGCAGAACGATATGTACGCAG (1)	101	23
ATACATATCGTGTTGCCAAGC (2)	39	21	**AAGGCAGAACGATATGTACGCAGA** (19)	101	24
AAGGCAGAACGATATGTACG (2)	101	20	AGGCAGAACGATATGTACGCAGA (1)	102	23
AAGGCAGAACGATATGTACGC (2)	101	21	GGCAGAACGATATGTACGC (1)	103	19

>Pre-NF30 Gm13:11,972,958..11,973,052 [-] intergênico
TTGTCTGGGTTGTGTTGTGAGCTCGATCAATTGCTCATGAAAATGGATTTGTGGCGA**TGATTATGAGGTCTGACACAAA**AATTGGTATTAGAGGA
((.((((((.(((((((((.((.(((((((((.((((.((..((......))..))))))))))))..)))).)))))))))).........)))))).)) (-26.10)

Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)	Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)
TGGATTTGTGGCGATGATTATGAG (1)	44	24	**TGATTATGAGGTCTGACACAAA** (67)	58	22
TTTGTGTCAGACCTCATAATCAT (1)	57	23	TTTGTGTCAGACCTCATAATC (1)	59	21
TGATTATGAGGTCTGACACAA (1)	58	21	TGACACAAAAATTGGTATTAGAGG (1)	71	24
TTTGTGTCAGACCTCATAATCA (1)	58	22			

>Pre-NF31 Gm12:37,542,998..37,543,242 [-] intron_Glyma12g34410
CCTCTTTGAATAAATTTCTCGATGAATAC**TCATAGGAGAGAAAAATAGGAAGG**TAAAATGAATTTGAACTTTTCCCTTTAGTTAAAATCAATTTATGCACCTAACTTGTATACAAG
(((.((((...(((((((((.((((.((((.(((((((((..(((((((((((((((((((((...((((.((((((((...(((((((((((((((((((...(((((((....))))
TTATTTATTCTAACTTCTTCAAAAGCTAAGATGCATAAGTTGATTTTAGCTTATGGGAAAAGCTCAGCGCATTTTACCTTCTTATTTTCTTCTCCTATAAGTAGTTATAGAGAAGT
)))............((((........)))).)))))))))))))))))))...)))))))).))))..)))))))))))))))))..)))))))).))))..))))..)))))))
TTCTCCAAATAGG
))...)))).))) (-105.00)

Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)
TCATAGGAGAGAAAAATAGGAAGG (18)	30	24
CATAGGAGAGAAAAATAGGAAG (1)	31	22
GCATAAGTTGATTTTAGC (1)	149	18

>Pre-NF32 Gm01:8,573,823..8,573,906 [+] intron_Glyma01g07770
TAACTACAATGATCCAA**TATTGTGTAAGCTTCCTAAAGAGA**TTACTCTTTAGGAAGCTTACACAGTATTGGATCATTGTAGTTA
((((((((((((((((((((((((((((((((((((....)))))))))))))))))))))))))))))))))))) (-64.60)

Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)	Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)
TCCAATATTGTGTAAGCTTCC (1)	13	21	TTGTGTAAGCTTCCTAAAGAGA (1)	20	22
AATATTGTGTAAGCTTCCTAAAGA (1)	16	24	TTAGGAAGCTTACACAGTATT (1)	49	24
TATTGTGTAAGCTTCCTAAAGA (1)	18	22	TACACAGTATTGGATCATTGT (1)	59	21
TATTGTGTAAGCTTCCTAAAGAGA (20)	18	24	CACAGTATTGGATCATTGTAG (2)	61	21
ATTGTGTAAGCTTCCTAAAGAGA (2)	19	23	CAGTATTGGATCATTGTAGTT (1)	63	21

>Pre-NF33 Gm02:6,233,520..6,233,676 [+] intergênico
TGGCACACTACTCAAATTGTTAGATGCTCAAATCTTTGTTACCACAGAGAGGCCAATGTGTTTCAAGTTTCAACAGAATTGAAACACA**TATTGGTCTTTTTGTAGTGAC**GAAGATT
CAAGCATCCAATGCCAATTTGGATATAAAGGTGGAGGGTCA
.(((.(.(((((((((((.(((((((..(((((((((((.(((((((((((((((((((((..(((....)))))))))))))..))))))))))))).)))))))))))
..))))))...)).)))))))........)))))).).))). (-64.30)

Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)	Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)
ACTACTCAAATTGTTAGATGCTCAAA (1)	7	26	ACACATATTGGTCTTTTTGTAGTGAC (1)	84	26
TCAAATCTTTGTTACCACAGA (1)	28	21	**TATTGGTCTTTTTGTAGTGAC** (25)	89	21
TTTGTTACCACAGAGAGGCCA (1)	35	21	TTGGTCTTTTTGTAGTGAC (1)	91	19
TTGTTACCACAGAGAGGCCAA (2)	36	21	TTGGTCTTTTTGTAGTGACGA (1)	91	21
TACCACAGAGAGGCCAATGT (1)	40	20	TTTGTAGTGACGAAGATTCAA (1)	99	21
TACCACAGAGAGGCCAATGTG (1)	40	21	TTGTAGTGACGAAGATTCAAG (1)	100	21
TACCACAGAGAGGCCAATGTGT (1)	40	22	TTGTAGTGACGAAGATTCAAGC (3)	100	22
AAACACATATTGGTCTTTTTGTAGTG (1)	82	26	TGTAGTGACGAAGATTCAAGC (1)	101	21
			GTAGTGACGAAGATTCAAG (1)	102	19

>Pre-NF34 Gm02:48,299,297..48,299,447 [+] intron_Glyma02g43540
TGTCCCAATGCTACCAAACACTGGCT**CAACCCTCCTCAGTTAGATCTC**GTCAAACAGCGAGGCGTGAATTCGCTCC
.(.(((((((((((((((.((((((((((((((((((((.(((((.(.(((.(((((((((..((.(((....
TTGACTTACTTGCCTTGCTATTTAAGGAGATCCAATTGAGGAGGGTTGAGCCAGTTTTTGGTAGCATTGGTGTCA
.)))).))...)))))))))).)))).).)))))).)))))))))))))))))))).)))))))))))))).).). (-95.40)

Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)	Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)
CTCAACCCTCCTCAGTTAGAT (1)	25	21	CAACCCTCCTCAGTTAGATC (1)	27	20

CAACCCTCCTCAGTTAGATCT (13)	27	21	AGATCCAATTGAGGAGGGTTGAGC (1)	104	24
CAACCCTCCTCAGTTAGATCTC (22)	27	22	GATCCAATTGAGGAGGGTTGA (1)	105	21
CAACCCTCCTCAGTTAGATCTCG (1)	27	23	GATCCAATTGAGGAGGGTTGAG (1)	105	22
CAACCCTCCTCAGTTAGATCTCGT (1)	27	24	ATCCAATTGAGGAGGGTTGAG (4)	106	21
AACCCTCCTCAGTTAGATCTC (1)	28	21	AGGGTTGAGCCAGTTTTTGGT (1)	118	21
CAGTTAGATCTCGTCAAACAGC (1)	37	22			

>Pre-NF35 Gm10:43,308,727..43,308,855 [-] intergênico
AGAGGGAAGAGAAAATC**TAATAGAGGGAAGAAGATGAA**GAGAATAATGATTTTTTCTTTAGACATGCA
((((((((((((((((.(((.(((((((.(((((((.(((((((.((....((((.....((((..
CGCATGGTAGAAAGCATGTTTTCTGCATCTTCCTTCTTCTCTTATGTTTTCTCTTCTTTCT
..)))))..)))))..)))))))))).)))))))..)))))))..)))..)))))))))))))))) (-53.20)

Sequência (n° de *reads*)	Posição inicial	Tamanho
TAATAGAGGGAAGAAGATGAA (27)	18	21

>Pre-NF36 Gm07:12,248,009..12,248,194 [+] intergênico
ATCGGTTGTTT**TAAAATCGATGTAGAAAGTGCC**CTTATAACATCGGTTGTTTTAAAACCGATGTAGAAAGTGTCCTTATAACATCGATTGTGA
((((((..((((((((((((((((((.((((((((((((((((((.......((((((((((((((((...(((((((((...(((..
CAAAAACGGATGTTGTATATGCACTTTCTACATCGGTTAGGCCTAAAACCGATGTTATAAGGGCACTTTTACATCGGTTTTAAGAATACCGAT
.....))))))))))))...)))))))))))))))))).......)))))))))))))))))))))))))))))))))))))..)))))) (-103.66)

Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)	Sequência (n° de *reads*)	Posição inicial	Tamanho
TTTTAAAATCGATGTAGAAAGTGCC (1)	9	25	(nt)AATCGATGTAGAAAGTGCCCTTAT (1)	15	24
TTTAAAATCGATGTAGAAAGT (1)	10	21	ATCGATGTAGAAAGTGCCCTT (2)	16	21
TTTAAAATCGATGTAGAAAGTG (2)	10	22	ATCGATGTAGAAAGTGCCCTTA (3)	16	22
TTTAAAATCGATGTAGAAAGTGCC (1)	10	24	ATCGATGTAGAAAGTGCCCTTATA (2)	16	24
TTAAAATCGATGTAGAAAGTGC (2)	11	22	ATCGATGTAGAAAGTGCCCTTATAA (1)	16	25
TAAAATCGATGTAGAAAGTGCC (26)	12	22	TCGATGTAGAAAGTGCCCTTAT (2)	17	22
TAAAATCGATGTAGAAAGTGCCC (1)	12	23	TCGATGTAGAAAGTGCCCTTATA (1)	17	23
AAAATCGATGTAGAAAGTGCC (1)	13	21	CGATGTAGAAAGTGCCCTTATA (1)	18	22
AAAATCGATGTAGAAAGTGCCC (6)	13	22	GATGTAGAAAGTGCCCTT (1)	19	18
AAATCGATGTAGAAAGTGCCC (3)	14	21	GTAGAAAGTGCCCTTATAA (1)	22	19
AAATCGATGTAGAAAGTGCCCT (1)	14	22	GTAGAAAGTGCCCTTATAACATCG (2)	22	24
AATCGATGTAGAAAGTGCCCTT (13)	15	22	GTGCCCTTATAACATCGG (1)	29	18

TGTTTTAAAACCGATGTAGAAAGT (1)	48	24	CAATCGATGTTATAAGGACACT (1)	69	22
TAAAACCGATGTAGAAAGTGTCCT (1)	53	24	ATAACATCGATTGTGACAAAAACG (1)	78	24
AAACCGATGTAGAAAGTGTCCTT (2)	55	23	GTAGAAAGTGCATATACAACA (1)	104	21
AACCGATGTAGAAAGTGTCCTT (1)	56	22	GTGCCCTTATAACATCGG (1)	142	18
AACCGATGTAGAAAGTGTCCTTAT (1)	56	24	ATAAGGGCACTTTTACATCGGTTT (3)	150	24
TAAGGACACTTTCTACATCGGT (1)	57	22	TAAGGGCACTTTTACATCGGTTT (2)	151	23

>Pre-NF37a Gm14:25,154,100..25,154,318 [+] intergênico

ATGATATATCATGTGAGAGAAAAATATCTTTTATGAGAGTAAGAAGATTAAATGGACCTTATAATCATACATGAAGGATCAAGACTAAATGTCATGTGACTAGTTA

(((.(((((((.(.(((((((.(((((((.(((((.(((.(((.((((((..((((.(.(((((((.(((((((.(((((((.((((.(((((((((((.....

AAAAGTCACATGACTTTAA**ATCTTGATCGTTCATGTTTGGT**TATACGATTCAAATAATCATCTCATTGTCATATAAGATATCTTCATCTCTCTCTCTCTCTCTTGATATACAT

...)))))))))))))))..)))))))).(((((((.)))))))).).))))..))))))..)))..)))..))))).)))))))..))..........))))))..).).)))))))))

(-85.00)

Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)	Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)
AAATCTTGATCGTTCATGTTTGGT (20)	124	24	**ATCTTGATCGTTCATGTTTGGT** (27)	126	22
ACCAAACATGAACGATCAAGATTT (1)	124	24	TCTTGATCGTTCATGTTTGGT (6)	127	21
AATCTTGATCGTTCATGTTT (1)	125	20	TCTTGATCGTTCATGTTTGGTT (3)	127	22
AACCAAACATGAACGATCAAGATT (1)	125	24	TCTTGATCGTTCATGTTTGGTTAT (1)	127	24
AATCTTGATCGTTCATGTTTGGTT (2)	125	24	TCTTGATCGTTCATGTTTGGTTATA (1)	127	25

>Pre-NF37b Gm19:40,112,991..40,113,140 [+] intergênico

GAGTGAGAGAATTAAATTTGGA**CCGTATAACCAAACATGGAGGGT**TAAGATTTAACGTTAAAAAGTCACA

((((((.(..((((((((((((((((((((((((((((.((((((((((((..((((((((((...

TGACTTTTTAAGTGGTTACATGACATTAA**ATCTTGATCGTTCATGTTTGGT**TATACGGTCCAGATTTAATCTTGTTACTC

.))))))))))...(((....))).))))))))))))).)))))))))))))))))))))))))))))..).)))))) (-80.20)

Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)	Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)
AGAATTAAATTTGGACCGTATAAC (1)	8	24	CGTATAACCAAACATGGAGGGT (3)	24	22
TAAATTTGGACCGTATAACCAA (1)	13	22	GTATAACCAAACATGGAGGGT (2)	25	21
AAATTTGGACCGTATAACCAAA (1)	14	22	TAACCAAACATGGAGGGTTAAGA (2)	28	23
CCGTATAACCAAACATGGAGG (4)	23	21	TAACCAAACATGGAGGGTTAAGAT (2)	28	24
CCGTATAACCAAACATGGAGGG (2)	23	22	CCAAACATGGAGGGTTAAGATT (1)	31	22
CCGTATAACCAAACATGGAGGGT (22)	23	23	CAAACATGGAGGGTTAAGATT (3)	32	21
CCGTATAACCAAACATGGAGGGTT (1)	23	24	CAAACATGGAGGGTTAAGATTT (2)	32	22

AACGATCAAGATTTAATGTCATGT (1)	88	24	TCTTGATCGTTCATGTTTGGT (6)	101	21
ATGACATTAAATCTTGATCGTTCA (1)	90	24	TCTTGATCGTTCATGTTTGGTT (3)	101	22
GACATTAAATCTTGATCGTTCATG (1)	92	24	TCTTGATCGTTCATGTTTGGTTAT (1)	101	24
TGAACGATCAAGATTTAATGT (1)	93	21	TCTTGATCGTTCATGTTTGGTTATA (1)	101	25
AAACATGAACGATCAAGATTTAAT (1)	95	24	CGTTCATGTTTGGTTATACGGT (9)	108	22
AAATCTTGATCGTTCATGTTTGGT (20)	98	24	CGTTCATGTTTGGTTATACGGTC (3)	108	23
ACCAAACATGAACGATCAAGATTT (1)	98	24	GTTCATGTTTGGTTATACGGT (1)	109	21
AATCTTGATCGTTCATGTTT (1)	99	20	GTTCATGTTTGGTTATACGGTC (4)	109	22
AACCAAACATGAACGATCAAGATT (1)	99	24	TTCATGTTTGGTTATACGGT (2)	110	20
AATCTTGATCGTTCATGTTTGGTT (2)	99	24	TGGTTATACGGTCCAGATTTAA (2)	118	22
ATCTTGATCGTTCATGTTTGGT (27)	100	22	GGTTATACGGTCCAGATTTAAT (1)	119	22

>Pre-NF38 Gm15:12,954,693..12,954,874 [+] cds-intron-cds_Glyma15g16650

CCATGCTCGCTGCGTCCGGCGTCGTCCAG**GTATTAGATGCATACTCAGTGGAT**TTAATTACTTAGTTTGTTTTTCTGTTATCACATTATTA

((((.(((((((.((..((....(((.(((((((((((.(((.(((((((((((((((((.((((((.(((...(((...(((((((....

CTATAAGTGTGAATTCAGCATTTAATTAATTGATTAATTAAGTCTACTGATTATATATCTGATACCAGGAGTTTGAAGCGGACGAGGGTGG

.....)))))))...))).....))).))))))).))))))))))))))))).))).))))))))))).))))))..)).))))..))))) (-71.60)

Sequência (nº de *reads*)	Posição inicial	Tamanho (nt)
GTATTAGATGCATACTCAGTGGAT (22)	30	24
CTACTGATTATATATCTGATACC (1)	135	23
CTGATTATATATCTGATACCAG (1)	138	22
TGATTATATATCTGATACCAG (1)	139	21

>Pre-NF39a Gm18:42,254,378..42,254,587 [+] intergênico

TTCTAAGACGGTTGTCTCCACGACTGACTTAGAATGTAACACATTCTACATCGGTCTTATTCATGCAACTGTCTCAGAATGTGTTATATTCTAAGACGGTCCTCG

.((((((((((((((((.(((((((.((((((((((((((((((((((((((((((((((((.((((((((((((.((((((((((((.((((((((((....

CGACGACTGTCTTAGTATGTAACACATTCTAAGACGGTTGTGTG**TATAGGACCGATGTAGAATGTGTT**ACATTCTAAGATGGTCGTAGAGACAACCGTCTTAGAA

....))))))))))).))))))))))))).))))))))))))).))))))))))))))))))))))))))))))))))))))).)))))))).)))))))))))))))). (-152.00)

Sequência (nº de *reads*)	Posição inicial	Tamanho (nt)	Sequência (nº de *reads*)	Posição inicial	Tamanho (nt)
GTCGTGGAGACAACCGTCTTA (1)	4	21	TGTAACACATTCTACATCGGTCTT (1)	35	24
TGTCTCCACGACTGACTTAGAATG (1)	13	24	AATAAGACCGATGTAGAATGTGTT (1)	38	24
CCACGACTGACTTAGAATGTA (2)	18	21	TAAGACCGATGTAGAATGTGT (1)	39	21
CCACGACTGACTTAGAATGTAA (4)	23	19	ACACATTCTACATCGGTCTTATT (1)	39	23

CACATTCTACATCGGTCTTAT (1)	40	21
TGAATAAGACCGATGTAGAATGTG (1)	40	24
CACATTCTACATCGGTCTTATTCA (1)	40	24
ACATTCTACATCGGTCTTATTCA (1)	41	23
ATGAATAAGACCGATGTAGAATGT (2)	41	24
ATCGGTCTTATTCATGCAACTGTC (2)	50	24
TCGGTCTTATTCATGCAACTGTC (1)	51	23
CGGTCTTATTCATGCAACTG (2)	52	20
CGGTCTTATTCATGCAACTGTC (3)	52	22
GGTCTTATTCATGCAACTGTC (1)	53	21
GGTCTTATTCATGCAACTGTCTC (1)	53	23
TGCAACTGTCTCAGAATGTGT (1)	64	21
TATATTCTAAGACGGTCCTCGCGA (1)	85	24
ATACTAAGACAGTCGTCGCGAGG (1)	101	23
GAATGTGTTACATACTAAGACAGT (1)	111	24
CATTCTAAGACGGTTGTGTGTA (2)	130	22
ATTCTAAGACGGTTGTGTGTA (1)	131	21

TCTAAGACGGTTGTGTGTATAGGA (1)	133	24
TGTGTGTATAGGACCGATGTAGAA (1)	144	24
TGTATAGGACCGATGTAGAATGTG (1)	148	24
GTATAGGACCGATGTAGAATGTGT (1)	149	24
TATAGGACCGATGTAGAAT (2)	150	19
TATAGGACCGATGTAGAATGT (2)	150	21
TATAGGACCGATGTAGAATGTG (11)	150	22
TATAGGACCGATGTAGAATGTGTT (94)	150	24
ATAGGACCGATGTAGAATG (1)	151	19
ATAGGACCGATGTAGAATGTGTT (6)	151	23
ATAGGACCGATGTAGAATGTGTTA (2)	151	24
TAGGACCGATGTAGAATGTGTT (10)	152	22
AGGACCGATGTAGAATGTGT (1)	153	20
AGGACCGATGTAGAATGTGTT (5)	153	21
AGGACCGATGTAGAATGTGTTACA (4)	153	24
TTCTAAGATGGTCGTAGAGACA (1)	177	22

>Pre-NF39b Gm19:11,095,283..11,095,442 [-] intergênico

GTCTTAGAATGTAACACATTCTACATCGGTCCTATACACACAACCATCTTAAAATGTATTTCATACTAAGACGGTTGTTG

((((.((((((((((((((((((((((((((((((.((((.(((((.(((((((...(((.(((((((.(((..(..

CAATGATCATCTTAGAATGTGTTACATTCTAAGACGGTTATGTG**TATAGGACCGATGTAGAATGTGTT**ACATTCTCAGAC

..)..))).)))))).)))...)))))).)))))..)))).))))))))))))))))))))))))))))))).)))) (-95.00)

Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)
GTCTTAGAATGTAACACAT (1)	1	19
GTCTTAGAATGTAACACATT (1)	1	20
AATGTGTTACATTCTAAGACGGTT (1)	96	24
ACCGTCTTAGAATGTAACACA (3)	98	21
ATAACCGTCTTAGAATGTAA (1)	102	20
TGTATAGGACCGATGTAGAATGTG (1)	123	24
GTATAGGACCGATGTAGAATGTGT (1)	124	24
TATAGGACCGATGTAGAAT (2)	125	19
TATAGGACCGATGTAGAATGT (2)	125	21

Sequência (n° de *reads*)	Posição inicial	Tamanho (nt)
TATAGGACCGATGTAGAATGTG (11)	125	22
TATAGGACCGATGTAGAATGTGTT (94)	125	24
ATAGGACCGATGTAGAATG (1)	126	19
ATAGGACCGATGTAGAATGTGTT (6)	126	23
ATAGGACCGATGTAGAATGTGTTA (2)	126	24
TAGGACCGATGTAGAATGTGTT (10)	127	22
AGGACCGATGTAGAATGTGT (1)	128	20
AGGACCGATGTAGAATGTGTT (5)	128	21
AGGACCGATGTAGAATGTGTTACA (4)	128	24

```
Pre-NF40    Gm16:1,163,892..1,164,088 [+] intron_Glyma16g01590
TGCATTATAGGTTTAATTACTCATTTGGTCTCTATGCTTGCACAATCTTTATGTTTTAGTCCTTACACTTAGAAATTACTCATTTTAGTTCCTATACACACACTTACCTGTTTTTG
((((.((.((((.(((((((((((((((((((((((((((((((((((((((.(((((((((((..((((((((((((((....((((((((((.(((((.((((((...(((((((.(
CCCTTACCATCCAAAATTGTAGGGACTAAAACGAATTAAAAAGTGCTTGTGTAGCGACTACCACGAGTTAAAAAGTGTATGTATTGGGACTAAAACAAATAGTTTCTAAGTGTAGC
.(((((((.((.....)).)))))).).))))))...........((((((((((...))).)))))))....))))))).)))))).)))))))))))....))))))))))))))).
AATTAAAACATAACGATTGTGCAAGTATAGGGACCAAATGAGTGATTAATCCTTTATTGTA
.))))))))))))).)))))))))))))))))))))))))))))))))))).))).))).)))) (-136.80)

Sequência (n° de reads) Posição inicial   Tamanho (nt)
AAGGACTAAAACATAAAGATTGTGC (1)       40    25
ATAACGATTGTGCAAGTATAGG (1)          242   22
ATAACGATTGTGCAAGTATAGGGA (13)       242   24
TAACGATTGTGCAAGTATAGG (1)           243   21
TAACGATTGTGCAAGTATAGGG (1)          243   22
ACGATTGTGCAAGTATAGGGA (1)           245   21
GTGCAAGTATAGGGACCAAATGAG (1)        251   24
CAAGTATAGGGACCAAATGA (1)            254   20
CAAGTATAGGGACCAAATGAG (1)           254   21
CAAGTATAGGGACCAAATGAGT (1)          254   22
```

Anexo 3: Artigo aceito para publicação no periódico Genetics and Molecular Biology (Molina *et al*, 2012)

Metatranscriptomic analysis of small RNAs present in soybean deep sequencing libraries

Lorrayne Gomes Molina[1,2], Guilherme Cordenonsi da Fonseca[1], Guilherme Loss de Morais[1], Luis Felipe Valter de Oliveira[1,2], Joseane Biso de Carvalho[1], Franceli Rodrigues Kulcheski[1] and Rogerio Margis[1,2,3]

[1]Centre of Biotechnology and PPGBCM, Laboratory of Genomes and Plant Population, Federal University of Rio Grande do Sul (UFRGS), Porto Alegre, RS, Brazil.
[2]Programa de Pós-Graduação em Genética e Biologia Molecular, Federal University of Rio Grande do Sul (UFRGS) Porto Alegre, RS, Brazil.
[3]Department of Biophysics, Federal University of Rio Grande do Sul (UFRGS) Porto Alegre, RS, Brazil.

Short title: Metatranscriptomic analysis of small RNAs
Keywords: next generation sequencing, small RNA, siRNA, molecular markers

Send correspondence to Rogerio Margis. Centre of Biotechnology and PPGBCM, Laboratory of Genomes and Plant Population, Building 43431, Federal University of Rio Grande do Sul (UFRGS) P.O. Box 15005, 91501-970, Porto Alegre, RS, Brazil. E-mail: rogerio.margis@ufrgs.br.

Abstract

A large number of small RNAs unrelated to the soybean genome were identified after deep sequencing of soybean small RNA libraries. A metatranscriptomic analysis was carried out to identify the origin of these sequences. Comparative analyses of small interference RNAs (siRNAs) present in samples collected in open areas corresponding to soybean field plantations and samples from soybean cultivated in greenhouses under a controlled environment were made. Different pathogenic, symbiotic and free-living organisms were identified from samples of both growth systems. They included viruses, bacteria and different groups of fungi. This approach can be useful not only to identify potentially unknown pathogens and pests, but also to understand the relations that soybean plants establish with microorganisms that may affect, directly or indirectly, plant health and crop production.

Introduction

Until recently, analysis of the microbial diversity in environmental samples was conducted only after isolation, culture and identification of microorganisms and subsequent sequencing of cloned libraries (Cardenas and Tiedje, 2008). However, these conventional methods are limited to the minority of species that can be cultured (Chistoserdova, 2010). New culture-independent methods, such as metagenomics and metatranscriptomics, have been developed (Xu, 2006; Cardenas and Tiedje, 2008; Adams *et al.*, 2009; Warnecke and Hess, 2009; Chistoserdova, 2010). These methods refer to studies of the collective set of genomes and transcriptomes of mixed microbial communities and may be applied to the exploration of all microorganisms that reside in marine environments, soils, human and animal clinical samples, sludge, polluted environment, and plants (Kent *et al.*, 2007; Zoetendal *et al.*, 2008; Adams *et al.*, 2009; Al Rwahnih *et al.*, 2009; Poretsky *et al.*, 2009; Shi *et al.*, 2009; Desai *et al.*, 2010; Gifford *et al.*, 2010; Roossinck *et al.*, 2010). The metagenomic approaches, including metatranscriptomics, involve the sequencing of random DNA or RNA-derived complementary DNA (cDNA) profiles and subsequent determination of taxonomic diversity and prospective genes related to response to environmental conditions (Rosen *et al.*, 2009). With the advent of new sequencing technologies (high-throughput sequencing), more data can be generated in a relatively short time, in a practical and cost-effective way (Creer *et al.*, 2010). Moreover, this approach allows the direct sequencing of DNA or cDNA, avoiding any cloning bias and leading to large-scale studies (Adams *et al.*, 2009).

Metagenomic analysis of RNA deep-sequencing data has been used in plant disease diagnostics, such as in grapevine (Al Rwahnih *et al.*, 2009; Coetzee *et al.*, 2010), sweet potato (Kreuze *et al.*, 2009), tomato and *Liatris spicata* (Adams *et al.*, 2009). Coupling of metagenomics with pyrosequencing in these studies has also allowed the detection of bacterial and fungal RNA, suggesting that this approach can contribute to massive identification of crop-associated microbiota, including pathogenic, symbiotic, and free-living organisms.

In addition to other sequencing methodologies, deep sequencing libraries of small RNA (sRNA) can contribute to microbial identification studies as they may include non-coding RNAs (e.g. rRNA), small regulatory RNAs, such as microbial sRNAs (Shi *et al.*, 2009) and host small interfering RNAs (Kreuze *et al.*, 2009), as well as mRNA fragments that normally would not be represented in libraries enriched for polyA-tailed mRNA commonly used in metatranscriptomic studies. As an example of the application of high-throughput sequencing of plant sRNAs, Kreuze *et al.* (2009), using short read sequences of approximately 24 base pairs (bp), successfully identified viruses infecting sweet potato, even those present in extremely low titer symptomless infections.

With this in mind, the current study describes the use of high-throughput sequencing of sRNAs to identify potential pathogens and other microorganisms from samples of soybeans grown in the field and in controlled-environment conditions.

Material and Methods

Plant material

The sRNA sequences analyzed in this study were obtained from deep-sequencing libraries from different projects related to the soybean transcriptome (Genosoja and Genoprot). The libraries were constructed from root samples of soybean cultivars grown under greenhouse conditions, and from flower, seed and pod samples from soybean plants grown under field conditions.

Root samples were obtained from 'Embrapa 48' and 'BR 16' soybean cultivars grown in a greenhouse at Embrapa-Soja in Londrina, Brazil. Plants were cultivated using a hydroponic system composed of plastic containers (30 liters) and an aerated 6.6 pH-balanced nutrient solution. Seeds were pre-germinated on moist filter paper in the dark at 25°C ± 1°C and 65% ± 5% rhr. Plantlets were then placed in polystyrene supports with the roots of the seedlings completely immersed in the nutrient solution. Each tray containing the seedlings was maintained in a greenhouse at 25°C ± 2°C and 60% ± 5% rh under natural daylight (photosynthetic photon flux density (PPFD) =1.5 x 10^3 µmoles m^{-2} s^{-1}, equivalent to 8.93 x 10^4 lux) and a 12:12 h photoperiod. Roots of seedlings with the first trifoliate leaf fully developed (V2 developmental stage) were frozen in liquid nitrogen and stored at -80°C until RNA extraction.

Three stages of seed germination, pods, and mature seeds from the soybean cultivar 'Conquista' were also used in RNA extraction. In a growth chamber, seeds were incubated for 3, 5 and 7 days on rolls of moistened filter paper at 26°C. Pods (R3-R5) were collected from field plants grown at the Federal University of Rio Grande do Sul (UFRGS), Porto Alegre, Brazil.

Flower samples (R2-R3) were collected from the soybean cultivar 'Urano' grown at the experimental field of the University of Passo Fundo (UPF), in Passo Fundo, Brazil. Collected flowers were immediately powdered in Trizol (Invitrogen, CA, USA) and stored until RNA extraction.

RNA extraction and sequencing

Total RNA was isolated from roots, seeds, seedlings, pods and flowers using Trizol (Invitrogen, CA, USA), and following the manufacturer's instructions. RNA quality was evaluated by electrophoresis in 1.0% agarose gels, and the amount checked using a Quibit fluorometer and Quant-iT RNA assay kit (Invitrogen, CA, USA) according to the manufacturer's instructions. Approximately 10 µg of total RNA were sent to Fasteris Life Sciences SA (Plan-les-Ouates, Switzerland) for processing and sequencing using Solexa technology on an Illumina Genome Analyzer GAII. Briefly, the processing by Illumina consisted of the following successive steps: acrylamide gel purification of RNA bands corresponding to the size range 20–30 nucleotides (nt), ligation of 3′ and 5′ adapters to the RNA in two separate subsequent steps, each followed by acrylamide gel purification,

cDNA synthesis, and a final step of PCR amplification to generate a DNA colony template library for Illumina sequencing. After removing vector sequences, reads of 19 to 24 nt were used for further analysis.

Sequence analysis

The detailed analysis of endogenous sRNAs, including microRNAs, obtained from the soybean libraries described above is the topic of a separate study and will not be discussed here. However, since in addition a large number of sRNAs unrelated to the soybean genome were identified, a metagenomic analysis was indicated to identify the origin of these sequences. To this end, all reads were assembled into contigs using the Velvet 0.7.31 *de novo* assembly algorithm (Zerbino and Birney, 2008) with the following parameters: hash length of 23, coverage cut-off of 50, expected coverage of 1,000, and a minimum scaffold length of 100.

Assembled contigs matching the soybean genome were eliminated from further analysis using BLAST (BLASTn) "stand alone" version 2.2.20 against the soybean genome database available at Phytozome with the following parameters: expectation value (-e) of $1e^{-10}$, cost to open a gap (-G) of -6, cost to extend a gap (-E) of -6, penalty for a nucleotide mismatch (-q) of -5. The remaining contigs were used to search the NCBI database using BLASTn (nucleotide blast) with default parameters and an expectation value of 10^{-5}. The contigs were classified according to the sequence with the highest hit score found with BLASTn. The number of reads aligning to the contigs, coverage and average depth were determined with the SOAP tool (Li *et al.*, 2009). Default parameters were used and only filtered data (reads aligning to the references with a high confidence) are reported.

Results

Sequence analysis

The sRNA libraries analyzed in this study contained 5,627,802 reads (each consisting of 19-24 bp) from root, 8,610,347 from seeds and pods, and 9,314,206 from flower (Figure 1). The sRNA sequences were assembled into contigs ranging from 40 to 300 nucleotides, approximately. Sequence assembly produced 2,646, 15,521 and 28,382 contigs from root, seed and pod, and flower, respectively. After elimination of the soybean sequences, 253 (root), 2,574 (flower) and 1,959 (seed and pod) contigs remained for further analysis. These contigs were used in BLASTn searches against the NCBI database.

After BLASTn annotation, a large number of sequences remained unidentified (Figure 1), accounting for 73.4% of the field samples (Figure 2A) and 37.2% of the controlled environmental samples (Figure 3A). Contigs that corresponded to soybean sequences, but could not be filtered by the local BLASTn, represented 17.4% of the total contigs from field samples (Figure 2A) and 5.1% from controlled environment samples (Figure 3A) . There were also contigs corresponding to sequences from other plant species deposited in NCBI (Figures 2A and 3A).

Apart from the contigs mentioned above, 134 contigs from the controlled environment (root) and 335 from the field (flower, seed and pod) had hits to previously sequenced microorganisms and viruses (Figure 1) and provided the results shown in the

following topics. These contigs were distributed in different taxonomic groups based on their best BLASTn hits (Figures 2B and 3B). Contigs showing similarities to sequences from different taxonomic groups (multiple affiliations) were classified at the taxonomic level immediately above.

Taxonomic origin of sRNAs from controlled environmental samples

The most represented taxon in the controlled environment samples belongs to the domain Eukaryota (49.3%), including unicellular eukaryotes (46.3%) and fungi (3.0%), followed by the domain Bacteria (32.8%). In addition, 17.9% of the microbial contigs could only be classified to the taxonomic level of the domain Eukaryota (Figure 3B).

Unicellular eukaryote organisms were found only in controlled environment samples. The amoeba genus *Naegleria* was well represented (12 contigs) within the kingdom Excavata, with one contig classified to species level, viz. *Naegleria fowleri* (Table 1, Figures 3C and 4C). Contigs corresponding to the phylum Kinetoplastida were distributed at different levels of taxonomic classification through the family Bodonidae, including *Neobodo designis* and *Rhynchomonas nasuta.* The family Trypanosomatidae was also identified within this phylum, including the genus *Trypanosoma.* Within the kingdom Chromalveolata (group Heterokontophyta), the class Chrysophyceae was identified through two contigs. The genera *Chrysolepidomonas* and *Spumella* (class Chrysophyceae) were represented by one and seven contigs, respectively. One contig was assigned and classified as pertaining to *Spumella elongata*. The genus *Mallomonas* was also identified within the kingdom Chromalveolata (group Heterokontophyta).

Microalgae belonging to kingdom Plantae were represented by the genera *Chlorococcum* and *Pyramimonas*. The phylum Cercozoa belonging to kingdom Rhizaria was represented by six contigs, with four pertaining to the genus *Cercomonas* and one to the genus *Gymnophrys.*

The kingdom Fungi was poorly represented in the controlled environment libraries (Figures 3B, C). One contig was assigned to the genus *Rozella* and another one to the genus *Acaulospora* (Figure 4B). We also found a contig with 94% of identity to an uncultured soil fungus and one with multiple affiliations within the subkingdom Dikarya (Table 1).

The domain Bacteria was the most diverse taxon in controlled environment samples (Figures 3B,C). Contigs affiliated with multiple taxa could be classified only to the domain Bacteria, phyla Bacterioidetes and Proteobacteria, class Alphaproteobacteria and families Sphingomonadaceae, Bradyrhizobiaceae, Caulobacteriaceae, Burkholderiaceae and Comamonadaceae.

The classes Alpha- and Gammaproteobacteria were equally represented, followed by Betaproteobacteria. Within the class Alphaproteobacteria, contigs corresponded to the order Rhizobiales, including the genera *Bosea*, *Rhizobium* and *Mesorhizobium*. There were contigs assigned to families Caulobacteriaceae and Sphingomonadaceae, including the genera *Phenylobacterium* and *Sphingobium*, respectively (Figure 4A).

Within the class Gammaproteobacteria, contigs were derived from the order Burkholderiales, families Burkholderiaceae (including *Burkholderia* and *Ralstonia solanacearum*) and Comamonadaceae (including the genus *Verminephrobacter* and the species *Delftia acidovorans*) and genus *Leptothrix.* The group Oceanospirillales was represented by a single contig. Within the group Pseudomonadales, the genus *Pseudomonas* was the only one to be identified with a single contig. The group Xanthomonadales was

represented by the species *Ignatzschineria larvae* (Figure 4), and the phylum Bacterioidetes by the species *Cytophaga hutchinsoni* and genus *Flectobacillus*, both members of the family Cytophagaceae (Figure 4).

Further contigs were classified at the genus leve and shared high identity (at least 87%) to nucleotide sequences of species deposited in the NCBI database (Supplementary Material, Table S1). These may belong to the same or a closely related species, viz. the unicellular eukaryotes *Naegleria gruberi*, *Mallomonas asmundae* and *Chrysolepidomonas dendrolepidota*, the bacteria *Phenylobacterium lituiforme*, *Laribacter hongkongensis* and *Leptothrix cholodnii*, and the fungus *Acaulospora scrobiculata*.

Taxonomic origin of sRNAs from field samples

The BLAST analysis revealed that 87.2% of the contigs assembled from field samples were related to BPMV (Bean Pod Mottle Virus) (Figure 2B). These contigs were assembled from seed and pod samples and showed identities ranging from 94 to 100% for both RNA1 and RNA2 sequences from the virus (Table 2). The remaining sequences were fungi (9.9%) and bacteria (3.0%).

Within the kingdom Fungi, contigs were classified according to the subkingdom Dikarya and phylum Basidiomycota. One contig was assigned within the phylum Basidiomycota, subphylum Agaricomycetes, and two contigs were assigned to the order Tremellales. Within the order Tremellales, two contigs represented the genus *Dioszegia* and one contig the genus *Trichosporon* (Figure 4B). In addition, one contig was not identical but showed high identity (94%) to the species *Xanthophyllomyces dendrorhous* and could be classified only in the genus *Xanthophyllomyces* (Table 2) which belongs to the class Tremellomycetes.

The phylum Ascomycota subphylum Pezizomycotina was the most represented fungal taxon (Figure 2C). It was identified through nine contigs that had multiple affiliations within this taxon, these being with the classes Sordariomycetes and Dothideomycetes. Within this latter there were contigs affiliated to the order Capnodiales, including the family Mycosphaerellaceae and the genus *Cladosporium* (Figure 4B).

A single contig was classified only at the domain level Bacteria. There were contigs to the class Gammaproteobacteria, one to the genus *Burkholderia* and four to the family Enterobacteriaceae, including one to *Buchnera aphidicola*. There were also contigs associated with the class Alphaproteobacteria (order Rhizobiales) and the phyla Firmicutes (order Bacillales) and Actinobacteria from the genus *Streptomyces* (Figures 2 and 4A, Table 2).

Discussion

High-throughput sequencing allows the direct sequencing of DNA or cDNA from the environment, avoiding any cloning bias and also being less costly and faster than the Sanger method (Cardenas and Tiedje, 2008). The Illumina (Solexa) technology is capable of generating 36 million reads with average length of 35 bp within 4 days, which is several times higher than the output of traditional cloning libraries. The short read lengths are compensated by massive output, speed, simplicity and coverage, including regions recalcitrant to cloning.

To obtain ribosomal sequences (commonly used in the identification of species) from complex microbial communities, classical polymerase chain reaction (PCR) has been used with primers complementary to highly conserved regions of rRNA Rosen et al., "Signal processing for metagenomics: extracting information from the soup.", representing the bias of being a targeted approach (Bailly *et al.*, 2007).

Because the new sequencing technologies have the potential to sequence technically difficult regions, such as those that form firm secondary structures, they are useful in the analysis of rRNA (Cardenas and Tiedje, 2008).

The metagenomic analysis of sRNA high-throughput libraries, as done in this study, is a non-targeted approach that avoids amplification steps and the design of the 'universal primers'. In this way, the present study provided several rRNA sequences that could be used in taxonomic identification (Tables 1 and 2).

Here, we demonstrate the feasibility of metatranscriptomics/metagenomics to study the microbial communities present in soybean plants cultivated in field and controlled environment. In this work, high-throughput sequencing generated several records for bacteria, fungi, unicellular eukaryotes, and viruses at different taxonomic levels (species, genus, family, order, class, phylum, kingdom or domain).

In root tissues of soybean plants cultivated in a controlled environment, three groups of organisms were detected, these being unicellular eukaryotes, bacteria and fungi. The unicellular eukaryotes were the most abundant group of organisms in root tissues of soybean cultivated in a controlled environment. Unicellular eukaryotes included some subgroups of eukaryote microorganisms, such as microalgae (genera *Pyramimonas* and *Chlorococcum*, these belonging to the division Chlorophyta, kingdom Plantae, and the genus *Mallomonas* belonging to the phylum Heterokontophyta or Stramenopiles, kingdom Chromalveolata), flagellates (including *Neobodo designis*, *Rhynchomonas nasuta* and the genus *Trypanosoma*, belonging to the class Kinetoplastida, kingdom Excavata, and the genus *Spumella* belonging to the phylum Heterokontophyta or Stramenopiles, kingdom Chromalveolata), amoeba (genus *Naegleria*, including *N. loweri*, belonging to the class Heterolobosea, kingdom Excavata), and ameboflagellates (genera *Cercomonas* and *Gymnophrys* belonging to the phylum Cercozoa, kingdom Rhizaria), which are free-living organisms in freshwater, soil and marine habitats (Bailly *et al.*, 2007). The fact that plants grown in a controlled environment were cultivated in a hydroponic system with nutritive solution could explain the high percentage of unicellular eukaryotes found in this habitat, since these organisms feed on the sediment of organic matter present in water.

The domain Bacteria was the second highest group in terms of occurrence in root samples, including endophytic and epiphytic bacteria. Species, genera, families, and orders related to growth promotion and nodulation through endosymbiosis with rhizobia (Juteau *et al.*, 2004; Kuklinsky-Sobral *et al.*, 2004; Delmotte *et al.*, 2009; Ikeda *et al.*, 2009, 2010; Okubo *et al.*, 2009) were found in this work (Table S1). They include the genera *Rhizobium, Mesorhizobium, Sphingobium, Burkholderia, Bosea and Pseudomonas*, the families Bradyrhizobiaceae, Sphingomonadaceae and Rhizobiaceae, and the order Rhizobiales. We also found some species and genera involved in the degradation of cellulose in plant debris and that of organic matter in humid and freshwater habitats, such as the species/genera *Cytophaga hutchinsoni*, *Delftia acidovorans, Ignatzschineria larvae*, *Laribacter* (Woo *et al.*, 2009), *Leptothrix* and *Phenylobacterium.*

Ralstonia solanacearum is considered a bacterial pathogen. It causes wilt in important crops (including soybean) in other countries. Nevertheless, populations of

Ralstonia solanacearum may occur as free-living microorganisms in watercourses or in a latent form in plants without causing disease (Grey and Steck, 2001; Mole *et al.*, 2007). The genus *Pseudomonas* identified in this study includes members that are pathogenic to soybean, such as *P. syringae* pv. *glycinea*, or non-pathogenic ones, surviving as saprophytes, epiphytes or endophytes (Kuklinsky-Sobral *et al.*, 2004).

Endophytic and epiphytic bacteria can contribute to the health, growth and development of plants. Promotion of plant growth by these bacteria may result from indirect effects, such as the biocontrol of soilborne diseases through competition for nutrients, siderophore-mediated competition for iron, antibiosis, or the induction of systemic resistance in the host plant. Direct effects, such as the production of phytohormones, providing the host plant with fixed nitrogen, or solubilization of phosphorus and iron present in soil, may also be of relevance (Kuklinsky-Sobral *et al.*, 2004).

The bacterial families Commamonadaceae and Caulobacteriaceae have members involved in sediment degradation in freshwater. Other groups of bacteria, such as members of the family Methylophilaceae and the order Oceanospirillales can survive in soil and humid habitats.

Fungi present in the roots of soybean plants were found to belong to the genus *Rozella,* these being involved in the biological control of other fungi in plants. The genus *Acaulospora* forms arbuscular mycorhiza in plants, thus increasing nutrient absorption from soil, mainly phosphorus.

In seeds, seedlings, pods and flowers from soybean plants cultivated in field (crop), three groups of organisms were identified: virus, fungi and bacteria. The virus group was represented by a single member only, the *Bean Pod Mottle Virus* (BPMV). This virus is widespread in the major soybean growing areas throughout Brazil and the world (Anjos *et al.*, 2000; Giesler *et al.*, 2002). Mottling originates at the hilum and is also referred to as "bleeding hilum" since the hilum color seems to bleed from its normal zone. From then on, the mottling of the seed has similar coloration as the hilum (Giesler *et al.*, 2002). BPMV was detected mainly in the mature seed library of sRNA, and seeds used in the RNA extraction procedure showed symptoms of mottling (data not shown), predicting the occurrence of BPMV infection.

The identified fungi were classified at species, genus, family, order, class, subphylum, and phylum levels. The yeast genera *Xanthophyllomyces*, *Dioszegia* and *Thichosporon* may occur in flower, seed, stem and leaf surfaces, and are essential for biological control of pathogens (Kucsera *et al.*, 1998; Wang *et al.*, 2008; Weber *et al.*, 2008). The genus *Cladosporium* found in this study in flower and seed tissues of soybean is known to contain entomopathogenic species with potential for biocontrol of pest insects (Pimentel *et al.*, 2006). The family Mycosphaerellaceae was detected in this work. It includes various genera, especially *Cercospora*, *Pseudocercospora*, *Mycosphaerella*, *Septoria*, *Ramularia*, etc. that represent more than 10,000 species. The genera *Cercospora*, *Mycosphaerella* and *Septoria* contain some species considered as pathogenic to soybean (Crous *et al.*, 2009). The subkingdom Dikarya, phylum Basidiomycota, subphylum Pezizomycotina, classes Dothideomycetes, Agaricomycetes and Sordariomycetes, and order Tremellales contain both phytopathogenic and nonpathogenic fungi species related to soybean, including species with potential for biocontrol of pest insects and diseases.

Moreover, in the field environment many bacteria were found associated with soybean tissue. The bacterium *Buchnera aphidicola* colonizes insects and may be present

on plant surfaces. The genus *Burkholderia* includes endophytic species present in soil and plant (Kuklinsky-Sobral *et al.*, 2005). The family Enterobacteriaceae includes many endophytes, such as the genus *Pantoea,* which occur mainly in leguminous plants involved in biological control of phytopathogens (Delmotte *et al.*, 2009; Ikeda *et al.*, 2010). The orders Rhizobiales and Bacillales contain endophytes and species that colonize the rhizosphere and phyllosphere, and are related to soybean nodulation and biological control of pest insects, fungi and bacteria.

Several contigs could not be classified deeper into the eukaryote domain due to multiple affiliations or hitting to sequences from environmental samples (many from uncultured freshwater eukaryotes) (Table S1).

The high percentage of unknown sequences in this study could correspond to artifacts or to transcript fragments from poorly known taxa. Similarly, in many samples of a metagenomic study of plant viruses (Roossinck *et al.*, 2010), sequences without similarity to GenBank sequences represented more than half of the contigs. Furthermore, in a metatranscriptomic analysis of microbial communities from watercourse, half of the possible protein-encoding sequences from pyrosequencing had no significant hits to previously sequenced genes. The length of the contigs influenced this frequency, as the analysis of larger (>200 bp) sequences resulted in twice the frequency of annotated sequences when compared to shorter (<100 bp) reads (Poretsky *et al.*, 2009).

The sequencing of small RNA molecules in the size range of 19 to 25 was chosen as it corresponds to the sizes of small interfering RNAs (siRNAs) and virus derived interfering RNAs (vsiRNAs). These are normally produced when RNA interference mechanisms are activated in order to degrade endogenous or pathogen-derived RNAs. The size range of the sequenced small RNAs could be enlarged up to 100 nucleotides, in order to include other small RNA molecules originated by other processes of degradation. Sizes under 19 nucleotides should be avoided since they would have ambiguous matches in different genomes and would not be useful in species prediction.

The challenges that remain in metagenomics and metatranscriptomics include DNA and RNA extraction, low stability, abundance and proportion of mRNAs in total RNA extracts (Cardenas and Tiedje, 2008). In addition, many of the difficulties encountered in microbial diversity studies are due to the complexity of the microbial community and its unevenness (few populations are of high frequency and many populations of low abundance). These difficulties can be reduced in metatranscriptomics by focusing on the active populations in a sample (Morales and Holben, 2010). When analyzing a soil microbial community, Urich *et al.* (2008) demonstrated that the deep-sequencing data from total RNA is naturally enriched in both functionally (such as mRNA libraries) and taxonomically relevant molecules, i.e. mRNA and rRNA, respectively.

Nevertheless, upon using deep-sequencing data, significant differences have been found in the taxonomic distribution of cDNAs derived from total RNA compared to DNA libraries from soil samples (Bailly *et al.*, 2007). Differences were also found when comparing cDNA derived from mRNA-enriched libraries with DNA libraries from marine microbial communities (Gilbert *et al.*, 2008). This suggests that both DNA and RNA sequences should be analyzed complementarily when investigating the diversity of species from environmental samples. In this sense, sRNA sequencing libraries can contribute to microbial identification with other types of sequences, viz. non-coding RNAs and mRNA fragments that would not normally be present in libraries enriched for polyA-tailed mRNA.

Acknowledgments

LGM was sponsored by a M.Sc. grant; GCF, GLM, LFVO and FRK by a PhD grant, JBC by a DTI-RHAE and RM by a Productivity and Research grant from the National Council for Scientific and Technological Development (CNPq, Brazil). This work was financially supported by GenoSoja consortium (CNPq 5527/2007-8) and GenoProt (CNPq 559636/2009-1).

References

Adams IP, Glover RH, Monger WA, Mumford R, Jackeviciene E, Navalinskiene M, Samuitiene M and Boonham N (2009) Next-generation sequencing and metagenomic analysis: A universal diagnostic tool in plant virology. Mol Plant Pathol 10:537-545.

Al Rwahnih M, Daubert S, Golino D and Rowhani A (2009) Deep sequencing analysis of RNAs from a grapevine showing Syrah decline symptoms reveals a multiple virus infection that includes a novel virus. Virology 387:395-401.

Anjos JRN, Charchar MJA and Gomes AC (2000) Identificação do vírus do mosqueado do feijoeiro ("Bean Pod Mottle Virus") em soja no Brasil. Documentos Embrapa Cerrados 19:1-15.

Bailly J, Fraissinet-Tachet L, Verner MC, Debaud JC, Lemaire M, Wesolowski-Louvel M and Marmeisse R (2007) Soil eukaryotic functional diversity, a metatranscriptomic approach. ISME J 1:632-642.

Cardenas E and Tiedje JM (2008) New tools for discovering and characterizing microbial diversity. Curr Opin Biotechnol 19:544-549.

Chistoserdova L (2010) Recent progress and new challenges in metagenomics for biotechnology. Biotechnol Lett 32:1351-1359.

Coetzee B, Freeborough MJ, Maree HJ, Celton JM, Rees DJ and Burger JT (2010) Deep sequencing analysis of viruses infecting grapevines: Virome of a vineyard. Virology 400:157-163.

Creer S, Fonseca VG, Porazinska DL, Giblin-Davis RM, Sung W, Power DM, Packer M, Carvalho GR, Blaxter ML, Lambshead PJ, *et al.* (2010) Ultrasequencing of the meiofaunal biosphere: Practice, pitfalls and promises. Mol Ecol 19 (Suppl 1):4-20.

Crous PW, Summerell BA, Carnegie AJ, Wingfield MJ and Groenewald JZ (2009) Novel species of Mycosphaerellaceae and Teratosphaeriaceae. Persoonia 23:119-146.

Delmotte N, Knief C, Chaffron S, Innerebner G, Roschitzki B, Schlapbach R, von Mering C and Vorholt JA (2009) Community proteogenomics reveals insights into the physiology of phyllosphere bacteria. Proc Natl Acad Sci U S A 106:16428-16433.

Desai C, Pathak H and Madamwar D (2010) Advances in molecular and "-omics" technologies to gauge microbial communities and bioremediation at xenobiotic/anthropogen contaminated sites. Bioresour Technol 101:1558-1569.

Giesler LJ, Ghabrial SA, Hunt TE and Hill JH (2002) Bean Pod Mottle Virus: A threat to U.S. soybean production. Plant Disease 86:1280-1289.

Gifford SM, Sharma S, Rinta-Kanto JM and Moran MA (2010) Quantitative analysis of a deeply sequenced marine microbial metatranscriptome. ISME J 5:461-472.

Gilbert JA, Field D, Huang Y, Edwards R, Li W, Gilna P and Joint I (2008) Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. PLoS One 3:e3042.

Grey BE and Steck TR (2001) The viable but nonculturable state of *Ralstonia solanacearum* may be involved in long-term survival and plant infection. Appl Environ Microbiol 67:3866-3872.

Ikeda S, Kaneko T, Okubo T, Rallos LE, Eda S, Mitsui H, Sato S, Nakamura Y, Tabata S and Minamisawa K (2009) Development of a bacterial cell enrichment method and its application to the community analysis in soybean stems. Microb Ecol 58:703-714.

Ikeda S, Okubo T, Anda M, Nakashita H, Yasuda M, Sato S, Kaneko T, Tabata S, Eda S, Momiyama A, *et al.* (2010) Community- and genome-based views of plant-associated bacteria: plant-bacterial interactions in soybean and rice. Plant Cell Physiol 51:1398-1410.

Juteau P, Tremblay D, Villemur R, Bisaillon JG and Beaudet R (2004) Analysis of the bacterial community inhabiting an aerobic thermophilic sequencing batch reactor (AT-SBR) treating swine waste. Appl Microbiol Biotechnol 66:115-122.

Kent AD, Yannarell AC, Rusak JA, Triplett EW and McMahon KD (2007) Synchrony in aquatic microbial community dynamics. ISME J 1:38-47.

Kreuze JF, Perez A, Untiveros M, Quispe D, Fuentes S, Barker I and Simon R (2009) Complete viral genome sequence and discovery of novel viruses by deep sequencing of small RNAs: A generic method for diagnosis, discovery and sequencing of viruses. Virology 388:1-7.

Kucsera J, Pfeiffer I and Ferenczy L (1998) Homothallic life cycle in the diploid red yeast *Xanthophyllomyces dendrorhous (Phaffia rhodozyma)*. Antonie Van Leeuwenhoek 73:163-168.

Kuklinsky-Sobral J, Araujo WL, Mendes R, Geraldi IO, Pizzirani-Kleiner AA and Azevedo JL (2004) Isolation and characterization of soybean-associated bacteria and their potential for plant growth promotion. Environ Microbiol 6:1244-1251.

Kuklinsky-Sobral J, Araújo WL, Mendes R, Pizzirani-Kleiner AA and Azevedo JL (2005) Isolation and characterization of endophytic bacteria from soybean (*Glycine max*) grown in soil treated with glyphosate herbicide. Plant Soil 273:91-99.

Li R, Yu1 Y, Li Y, Lam TW, Yiu SM, Kristiansen K and Wang J (2009) SOAP2: An improved ultrafast tool for short read alignment. Bioinformatics 25:1966-1967.

Mole BM, Baltrus DA, Dangl JL and Grant SR (2007) Global virulence regulation networks in phytopathogenic bacteria. Trends Microbiol 15:363-371.

Morales SE and Holben WE (2010) Linking bacterial identities and ecosystem processes: can "omic" analyses be more than the sum of their parts? FEMS Microbiol Ecol 75:2-16.

Okubo T, Ikeda S, Kaneko T, Eda S, Mitsui H, Sato S, Tabata S and Minamisawa K (2009) Nodulation-dependent communities of culturable bacterial endophytes from stems of field-grown soybeans. Microbes Environ 24:253-258.

Pimentel IC, Glienke-Blanco C, Gabardo J, Stuart RM and Azevedo JL (2006) Identification and colonization of endophytic fungi from soybean (*Glycine max* (L.) Merril) under different environmental conditions. Braz Arch Biol Technol 49:705-711.

Poretsky RS, Hewson I, Sun S, Allen AE, Zehr JP and Moran MA (2009) Comparative day/night metatranscriptomic analysis of microbial communities in the North Pacific subtropical gyre. Environ Microbiol 11:1358-1375.

Roossinck MJ, Saha P, Wiley GB, Quan J, White JD, Lai H, Chavarria F, Shen G and Roe BA (2010) Ecogenomics: Using massively parallel pyrosequencing to understand virus ecology. Mol Ecol 19 (Suppl 1):81-88.

Rosen GL, Sokhansanj BA, Polikar R, Bruns MA, Russell J, Garbarine E, Essinger S and Yok N (2009) Signal processing for metagenomics: extracting information from the soup. Curr Genomics 10:493-510.

Shi Y, Tyson GW and DeLong EF (2009) Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column. Nature 459:266-269.

Urich T, Lanzen A, Qi J, Huson DH, Schleper C and Schuster SC (2008) Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. PLoS One 3:e2527.

Wang QM, Jia JH and Bai FY (2008) Diversity of basidiomycetous phylloplane yeasts belonging to the genus *Dioszegia* (Tremellales) and description of *Dioszegia athyri* sp. nov., *Dioszegia butyracea* sp. nov. and *Dioszegia xingshanensis* sp. nov. Antonie Van Leeuwenhoek 93:391-399.

Warnecke F and Hess M (2009) A perspective: Metatranscriptomics as a tool for the discovery of novel biocatalysts. J Biotechnol 142:91-95.

Weber RW, Becerra J, Silva MJ and Davoli P (2008) An unusual Xanthophyllomyces strain from leaves of Eucalyptus globulus in Chile. Mycol Res 112:861-867.

Woo PC, Lau SK, Tse H, Teng JL, Curreem SO, Tsang AK, Fan RY, Wong GK, Huang Y, Loman NJ *et al.* (2009) The complete genome and proteome of *Laribacter hongkongensis* reveal potential mechanisms for adaptations to different temperatures and habitats. PLoS Genet 5:e1000416.

Xu J (2006) Microbial ecology in the age of genomics and metagenomics: Concepts, tools, and recent advances. Mol Ecol 15:1713-1731.

Zerbino DR and Birney E (2008) Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Res 18:821-829.

Zoetendal EG, Rajilic-Stojanovic M and de Vos WM (2008) High-throughput diversity and functionality analysis of the gastrointestinal tract microbiota. Gut 57:1605-1615.

Internet Resources

Phytozome, ftp://ftp.jgi-psf.org/pub/JGI_data/phytozome/v6.0/Gmax/assembly/sequence/

Supplementary Material

The following online material is available for this article:

Tables S1 - Relationship among metatranscriptomic sequencing data and species sequences in public data banks

This material is available as part of the online article from http://www.scielo.br/gmb.

Figure legends

Figure 1 - Chart flow of sequence analysis. The central column identifies the methods. The number of reads obtained at each step is indicated for samples obtained from a controlled environment (left) and from field samples (right).

Figure 2 - Comparative percentages of the contigs assembled from sRNA (19-24 nt) libraries of the field (crop soybean) for each sequence classification according to best hit with BLASTn. (A) All groups of organisms, (B) all groups excluding unidentified and plants and (C) best identified taxonomic groups. Bean Pod Mottle Virus (BPMV)

Figure 3 - Comparative percentages of the contigs assembled from sRNA (19-24 nt) libraries of the soybean cultivated in a controlled environment for each sequence classification according to best hit with BLASTn. (A) All groups of organisms, (B) all groups excluding unidentified and plants and (C) best identified taxonomic groups.

Figure 4 - Taxonomic levels of microorganisms distributed among the libraries of small RNAs in soybean cultivated in a controlled environment and in the field. (A) Bacteria (B) Fungi and (C) other Eukaryota.

Table 1 - Taxonomic affiliation of the 134 contigs assembled from libraries obtained from plants grown in a controlled environment (root samples).

Domain	Kingdom	Organism	Environment	Nr. of contigs	Identity (%)	Sequence type
Eukaryota		Fungi/Unicellular eucaryotes	Soil, plants, water, humans and animals	25	81-100	18S rRNA, 26S rRNA, 28S rRNA and Genomic DNA
Eukaryota	Chromalveolata	Stramenopiles	Water and soil	5	98-100	28S rRNA and 18S rRNA
Eukaryota	Chromalveolata	*Mallomonas* sp. (algae)	Water	2	94-100	ITS1-5.8 rRNA-ITS2-28S rRNA
Eukaryota	Chromalveolata	Chrysophyceae (phytoplankton)	Marine, lake and river water	2	100	28S rRNA
Eukaryota	Chromalveolata	*Spumella* sp.(flagellate)	Water	6	98-100	18S rRNA
Eukaryota	Chromalveolata	*Spumella elongate* (flagellate)	Water	1	100	18S rRNA
Eukaryota	Chromalveolata	*Chrysolepidomonas* sp. (flagellate)	Water	1	98	Genomic DNA
Eukaryota	Excavata	Kinetoplastida	Water	2	100	28S rRNA
Eukaryota	Excavata	Trypanosomatidae	Marine, lake and river sediments	3	90-100	24S α rRNA and genomic DNA
Eukaryota	Excavata	*Trypanosoma* sp.	Marine, lake and river sediments	1	100	18S rRNA, ITS1, 5.8S
Eukaryota	Excavata	Bodonidae	Marine, lake and river sediments	3	100	18S rRNA
Eukaryota	Excavata	*Rhynchomonas nasuta* (flagellate)	Water and soil	1	100	18S rRNA
Eukaryota	Excavata	*Neobodo* sp. (flagellate)	Marine, lake and river sediments	2	95-98	18S rRNA
Eukaryota	Excavata	*Neobodo designis* (flagellate)	Marine, lake and river sediments	8	100	18S rRNA
Eukaryota	Excavata	*Naegleria* sp. (amoeba)	Water and soil	12	91-100	18S rRNA, 28S rRNA,

						extrachromosomal rRNA plasmid DNA
Eukaryota	Excavata	*Naegleria fowleri* (amoeba)	Water and soil	1	100	18S rRNA
Eukaryota	Rhizaria	Cercozoa (ameboflagellates)	Soil and water sediments	2	89-100	18S rRNA and 28S rRNA
Eukaryota	Rhizaria	*Cercomonas* sp. (ameboflagellates)	Soil and water sediments	4	100	28S rRNA
Eukaryota	Rhizaria	*Gymnophrys* sp. (ameboflagellates)	Soil and water sediments	1	100	28S rRNA
Eukaryota	Viridiplantae	*Pyramimonas* sp. (microalgae)	Marine, lake and river water	3	95-96	28S rRNA
Eukaryota	Viridiplantae	*Chlorococcum* sp. (microalgae)	Marine, lake and river water	1	100	28S rRNA
Eukaryota	Fungi	Fungi	Soil, plant, human, animal, insect	1	94	18S rRNA
Eukaryota	Fungi	*Acaulospora* sp. (arbuscular mycorrhiza)	Plant roots	1	87	18S rRNA
Eukaryota	Fungi	*Rozella* sp.	Endoparasite of other fungi	1	92	28S rRNA
Eukaryota	Fungi	Dikarya	Soil, plant, human, animal, insect	1	98	26S rRNA
Bacteria		Bacteria	Soil, plants, water, humans and animals	10	92-100	16S rRNA and 23S rRNA
Bacteria		Proteobacteria	Soil, plants, water	2	97-100	16S rRNA and 23S rRNA
Bacteria		Alphaproteobacteria	Soil, plants, water	2	100	16S rRNA
Bacteria		Rhizobiales	Soil and plant root nodules	1	96	Genomic DNA
Bacteria		*Mesorhizobium* sp.	Soil and plant root nodules	1	98	23S rRNA
Bacteria		*Rhizobium* sp.	Soil and plant root nodules	1	100	16S rRNA
Bacteria		Bradyrhizobiaceae	Soil and plant root nodules	1	100	16S rRNA
Bacteria		*Bosea* sp.	Soil and plants	1	100	16S rRNA
Bacteria		Sphingomonadaceae	Soil	1	100	16S rRNA
Bacteria		*Sphingobium* sp.	Soil	1	100	23S rRNA
Bacteria		Caulobacteriaceae	Water	1	100	23S rRNA

Bacteria	*Phenylobacterium* sp.	Water	1	98	23S rRNA
Bacteria	Betaproteobacteria	Soil, plants, water	4	100	16S rRNA and genomic DNA
Bacteria	Methylophilaceae	Soil	1	98	23S rRNA
Bacteria	*Laribacter* sp.	Water	1	95	Genomic DNA
Bacteria	Oceanospirillales	Lakes, rivers and sea hypersaline and highly alkaline.	1	97	Genomic DNA
Bacteria	Burkholderiaceae	Soil and plants	1	100	23S rRNA
Bacteria	*Burkholderia* sp.	Soil and plants	1	96	Genomic DNA
Bacteria	*Ralstonia solanacearum*	Soil and water	1	100	Genomic DNA
Bacteria	Comamonadaceae	Water	1	100	16S rRNA and 23S rRNA
Bacteria	*Verminephrobacter* sp.	Lumbricid earthworms (*Lumbricidae nephridia*)	1	96	Genomic DNA
Bacteria	*Delftia acidovorans*	Soil and water	1	100	16S rRNA
Bacteria	*Leptothrix* sp.	Aquatic environments with sufficient organic matter	1	90	23S rRNA
Bacteria	*Pseudomonas* sp.	Plant leaves	1	100	16S rRNA
Bacteria	*Ignatzschineria larvae*	Insect mid gut (adult flesh fly)	1	100	23S rRNA
Bacteria	Bacterioidetes	Soil, plants, water	1	100	Genomic DNA
Bacteria	*Flectobacillus* sp.	Water	1	98	16S rRNA
Bacteria	*Cytophaga hutchinsoni*	Soil and plant debris (cellulose degradation)	3	100	Genomic DNA
Total			134		

Table 2 - Taxonomic affiliation of the 335 contigs assembled from libraries obtained from soybean plants grown in the field (flowers, pods, mature seeds and samples of germinating seeds).

Domain	Kingdom	Organism	Environment	Nr. of contigs	Identity (%)	Sequence type
Eukaryota	Fungi		Soil, dung, plant, decaying wood, fungal, insect, human	1	100	18S rRNA
Eukaryota	Fungi	Dikarya	Soil, dung, plant, rock, decaying wood, fungal, insect, human	3	100	18S, ITS1, 5.8S rRNA, ITS 2, 28S rRNA
Eukaryota	Fungi	Basidiomycota	Plant, humans, animals, water, soil	4	100	18S rRNA and 28S rRNA
Eukaryota	Fungi	Agaricomycetes	Plant, plant debris	1	100	28S rRNA
Eukaryota	Fungi	*Xanthophyllomyces* sp. (yeast)	Plant	1	94	28S rRNA
Eukaryota	Fungi	Tremellales	Human, parasites of other fungi	2	100	5.8S rRNA gene, ITS2, 26S rRNA
Eukaryota	Fungi	*Trichosporon* sp. (yeasts)	Soil, human	1	100	18S rRNA
Eukaryota	Fungi	*Dioszegia* sp. (yeasts)	Plant	2	100	26S rRNA, 18S rRNA
Eukaryota	Fungi	Pezizomycotina	Plants, soil, debris plant, rock	9	100	18S rRNA gene, ITS1, 5.8S rRNA, ITS 2, 28S rRNA
Eukaryota	Fungi	Sordariomycetes	Soil, dung, plant, decaying wood, fungal, insect, human	1	100	28S rRNA
Eukaryota	Fungi	Dothideomycetes	Plants, soil, debris plant, rock	1	100	28S rRNA
Eukaryota	Fungi	Capnodiales	Plants, soil, debris plant, rock	2	97-100	18S rRNA and 28S rRNA

Eukaryota	Fungi	Mycosphaerellaceae	Plants, soil, debris plant	3	100	18S rRNA and 28S rRNA
Eukaryota	Fungi	*Cladosporium* sp.	Plants, soil, debris plant	2	98-100	18S rRNA
Bacteria		Bacteria	Plant, humans, animals, water, soil	1	100	16S rRNA
Bacteria		Rhizobiales	Soil and root plant nodules	1	100	16S rRNA
Bacteria		Enterobacteriaceae	Plant, humans, animals, water, soil	4	100	16S rRNA, 23S rRNA and genomic DNA
Bacteria		*Buchnera aphidicola*	Insect	1	100	Genomic DNA
Bacteria		*Burkholderia* sp.	Soil, plant, human	1	100	Chromosome 2
Bacteria		*Streptomyces* sp.	Plant	1	100	16S rRNA
Bacteria		Bacillales	Plant, humans, animals, water, soil	1	100	23S rRNA
Aphanabionta		Bean pod mottle virus	Plant	292	94-100	RNA1, soy RNA1-b polyprotein, RNA2, putative defective coat protein, capsid polyprotein
Total				335		

Fig. 1

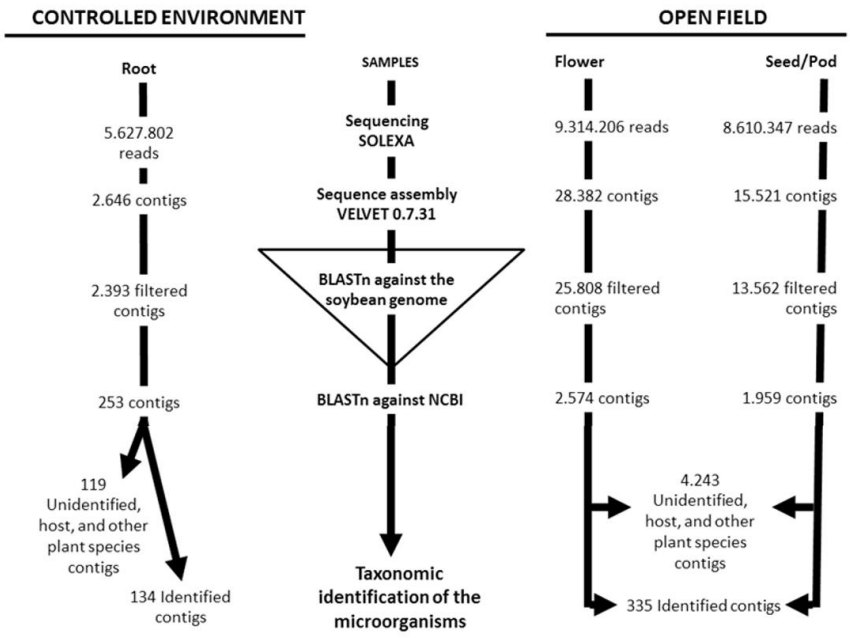

CONTROLLED ENVIRONMENT
OPEN FIELD
Root
SAMPLES
Flower
Seed/Pod
5.627.802 reads
Sequencing SOLEXA
9.314.206 reads
8.610.347 reads
2.646 contigs
Sequence assembly VELVET 0.7.31
28.382 contigs
15.521 contigs
2.393 filtered contigs
BLASTn against the soybean genome
25.808 filtered contigs
13.562 filtered contigs
253 contigs
BLASTn against NCBI
2.574 contigs
1.959 contigs
119 Unidentified, host, and other plant species contigs
4.243 Unidentified, host, and other plant species contigs
134 Identified contigs
Taxonomic identification of the microorganisms
335 Identified contigs

Fig. 2

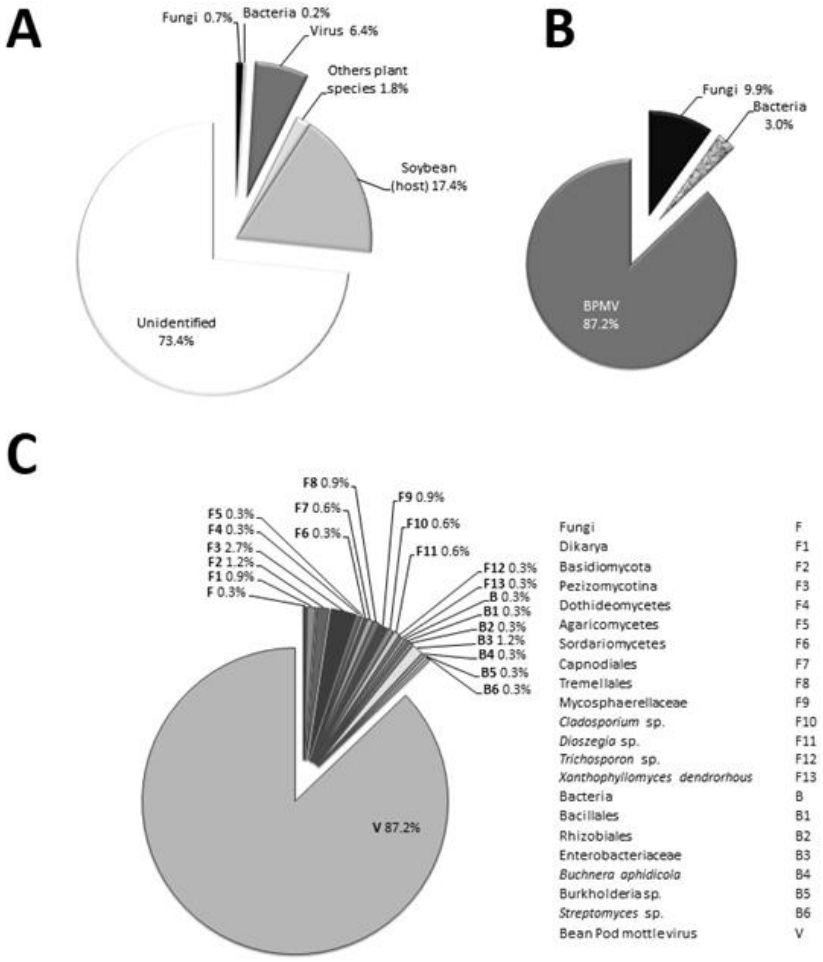

A
Fungi 0.7%
Bacteria 0.2%
Virus 6.4%
Others plant species 1.8%
Soybean (host) 17.4%
Unidentified 73.4%
B
Fungi 9.9%
Bacteria 3.0%
BPMV 87.2%
C
F8 0.9%
F9 0.9%
F7 0.6%
F10 0.6%
F5 0.3%
F6 0.3%
F4 0.3%
F11 0.6%
F3 2.7%
F12 0.3%
F2 1.2%
F13 0.3%
F1 0.9%
B 0.3%
F 0.3%
B1 0.3%
B2 0.3%
B3 1.2%
B4 0.3%
B5 0.3%
B6 0.3%
V 87.2%
Fungi F
Dikarya F1
Basidiomycota F2
Pezizomycotina F3
Dothideomycetes F4
Agaricomycetes F5
Sordariomycetes F6
Capnodiales F7
Tremellales F8
Mycosphaerellaceae F9
Cladosporium sp. F10
Dioszegia sp. F11
Trichosporon sp. F12
Xanthophyllomyces dendrorhous F13
Bacteria B
Bacillales B1
Rhizobiales B2
Enterobacteriaceae B3
Buchnera aphidicola B4
Burkholderia sp. B5
Streptomyces sp. B6
Bean Pod mottle virus V

Fig. 3

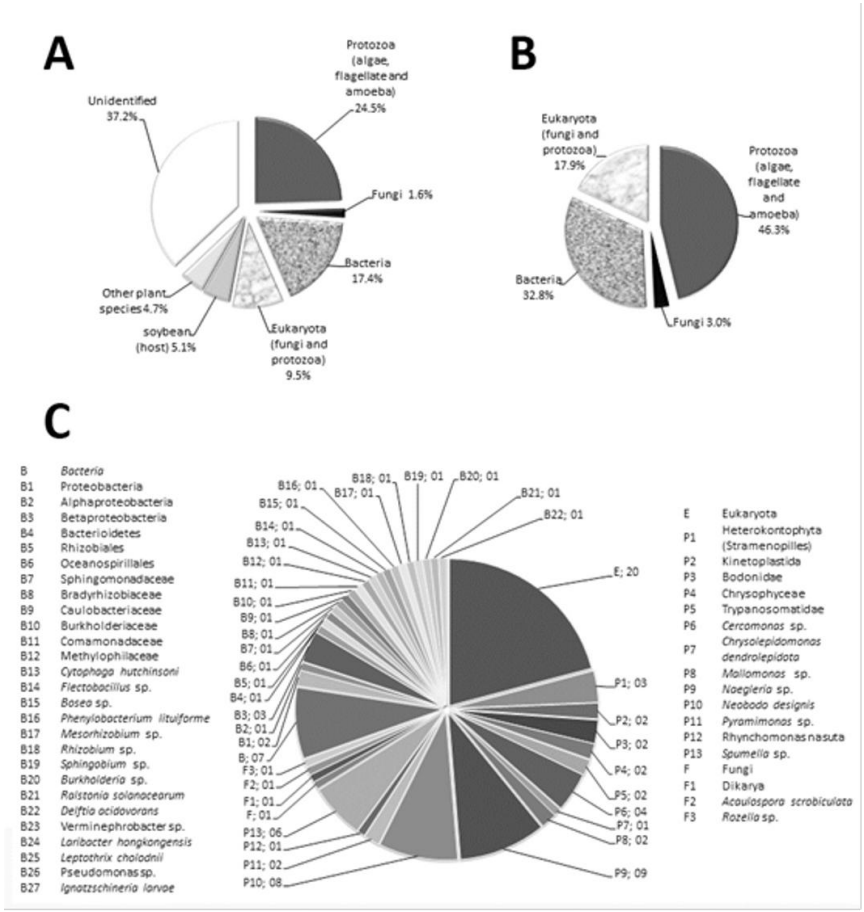
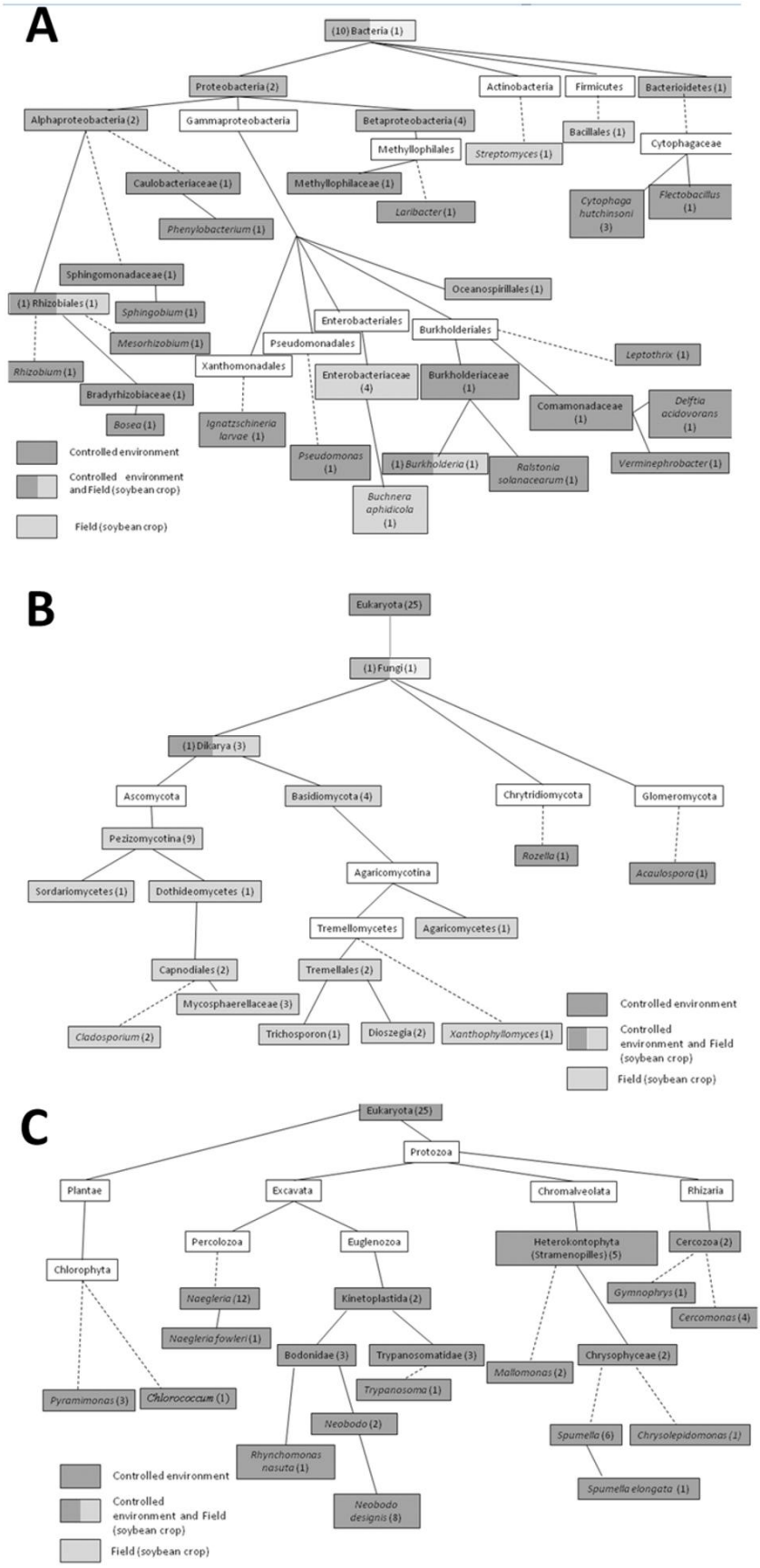

A
Protozoa (algae, flagellate and amoeba) 24.5%
Fungi 1.6%
Bacteria 17.4%
Eukaryota (fungi and protozoa) 9.5%
soybean (host) 5.1%
Other plant species 4.7%
Unidentified 37.2%
B
Protozoa (algae, flagellate and amoeba) 46.3%
Fungi 3.0%
Bacteria 32.8%
Eukaryota (fungi and protozoa) 17.9%
C
B Bacteria
B1 Proteobacteria
B2 Alphaproteobacteria
B3 Betaproteobacteria
B4 Bacterioidetes
B5 Rhizobiales
B6 Oceanospirillales
B7 Sphingomonadaceae
B8 Bradyrhizobiaceae
B9 Caulobacteriaceae
B10 Burkholderiaceae
B11 Comamonadaceae
B12 Methylophilaceae
B13 Cytophaga hutchinsoni
B14 Flectobacillus sp.
B15 Baseo sp.
B16 Phenylobacterium lituiforme
B17 Mesorhizobium sp.
B18 Rhizobium sp.
B19 Sphingobium sp.
B20 Burkholderia sp.
B21 Ralstonia solanacearum
B22 Deiftia acidovorans
B23 Verminephrobacter sp.
B24 Laribacter hongkongensis
B25 Leptothrix cholodnii
B26 Pseudomonas sp.
B27 Ignatzschineria larvae
B16; 01
B18; 01
B19; 01
B20; 01
B17; 01
B21; 01
B15; 01
B22; 01
B14; 01
B13; 01
E; 20
B12; 01
B11; 01
B10; 01
B9; 01
B8; 01
B7; 01
B6; 01
B5; 01
B4; 01
B3; 03
B2; 01
B1; 02
B; 07
F3; 01
F2; 01
F1; 01
F; 01
P13; 06
P12; 01
P11; 02
P10; 08
P1; 03
P2; 02
P3; 02
P4; 02
P5; 02
P6; 04
P7; 01
P8; 02
P9; 09
E Eukaryota
P1 Heterokontophyta (Stramenopiles)
P2 Kinetoplastida
P3 Bodonidae
P4 Chrysophyceae
P5 Trypanosomatidae
P6 Cercomonas sp.
P7 Chrysolepidomonas dendrolepidota
P8 Mallomonas sp.
P9 Naegleria sp.
P10 Neobodo designis
P11 Pyramimonas sp.
P12 Rhynchomonas nasuta
P13 Spumella sp.
F Fungi
F1 Dikarya
F2 Acaulospora scrobiculata
F3 Rozella sp.

Fig. 4


A
(10) Bacteria (1)
Proteobacteria (2)
Actinobacteria
Firmicutes
Bacteroidetes (1)
Alphaproteobacteria (2)
Gammaproteobacteria
Betaproteobacteria (4)
Bacillales (1)
Cytophagaceae
Methyllophilales
Streptomyces (1)
Caulobacteriaceae (1)
Methyllophilaceae (1)
Flectobacillus (1)
Laribacter (1)
Cytophaga hutchinsoni (3)
Phenylobacterium (1)
Sphingomonadaceae (1)
Oceanospirillales (1)
(1) Rhizobiales (1)
Sphingobium (1)
Enterobacteriales
Burkholderiales
Mesorhizobium (1)
Pseudomonadales
Leptothrix (1)
Rhizobium (1)
Xanthomonadales
Enterobacteriaceae (4)
Burkholderiaceae (1)
Bradyrhizobiaceae (1)
Comamonadaceae (1)
Delftia acidovorans (1)
Bosea (1)
Ignatzschineria larvae (1)
Pseudomonas (1)
(1) Burkholderia (1)
Ralstonia solanacearum (1)
Verminephrobacter (1)
Buchnera aphidicola (1)
Controlled environment
Controlled environment and Field (soybean crop)
Field (soybean crop)
B
Eukaryota (25)
(1) Fungi (1)
(1) Dikarya (3)
Ascomycota
Basidiomycota (4)
Chrytridiomycota
Glomeromycota
Pezizomycotina (9)
Rozella (1)
Acaulospora (1)
Agaricomycotina
Sordariomycetes (1)
Dothideomycetes (1)
Tremellomycetes
Agaricomycetes (1)
Capnodiales (2)
Tremellales (2)
Mycosphaerellaceae (3)
Cladosporium (2)
Trichosporon (1)
Dioszegia (2)
Xanthophyllomyces (1)
Controlled environment
Controlled environment and Field (soybean crop)
Field (soybean crop)
C
Eukaryota (25)
Protozoa
Plantae
Excavata
Chromalveolata
Rhizaria
Percolozoa
Euglenozoa
Heterokontophyta (Stramenopiles) (5)
Cercozoa (2)
Chlorophyta
Naegleria (12)
Kinetoplastida (2)
Gymnophrys (1)
Cercomonas (4)
Naegleria fowleri (1)
Bodonidae (3)
Trypanosomatidae (3)
Chrysophyceae (2)
Mallomonas (2)
Pyramimonas (3)
Chlorococcum (1)
Trypanosoma (1)
Neobodo (2)
Spumella (6)
Chrysolepidomonas (1)
Rhynchomonas nasuta (1)
Spumella elongata (1)
Neobodo designis (8)
Controlled environment
Controlled environment and Field (soybean crop)
Field (soybean crop)