UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

CENTRO DE BIOTECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM BIOLOGIA CELULAR E MOLECULAR

MICRORNAS DE SEMENTES DE SOJA MADURA E EM GERMINAÇÃO E *TRANSFER RNA-DERIVED FRAGMENTS* (TRFS) ASSOCIADOS A PROTEÍNAS ARGONAUTAS DE ARABIDOPSIS

Guilherme Loss de Morais

Porto Alegre, RS, Brasil

Março, 2013

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

CENTRO DE BIOTECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM BIOLOGIA CELULAR E MOLECULAR

MICRORNAS DE SEMENTES DE SOJA MADURAS E EM GERMINAÇÃO E *TRANSFER RNA-DERIVED FRAGMENTS* (TRFS) ASSOCIADOS A PROTEÍNAS ARGONAUTAS DE ARABIDOPSIS

Guilherme Loss de Morais

Orientador: Dr. Rogerio Margis

Tese submetida ao Programa de Pós-Graduação em Biologia Celular e Molecular do Centro de Biotecnologia da UFRGS como requisito parcial para a obtenção do título de Doutor em Ciências.

Porto Alegre, RS, Brasil

Março, 2013

INSTITUIÇÕES E FONTES FINANCIADORAS

As atividades de pesquisa cujos resultados estão reunidos nesta tese de Doutorado foram desenvolvidas no Laboratório de Genomas e Populações de Plantas do Centro de Biotecnologia da Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, sob a orientação do Prof. Dr. Rogerio Margis.

Parte deste trabalho utilizou recursos financeiros do projeto GenoSoja, financiado pelo CNPq dentro do plano de transcriptômica referente à pesquisa de microRNAs e também do projeto Agroestruturante em Bioenergia, financiado pela FINEP e FAPERGS. Este trabalho utilizou recursos financeiros do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e da Coordenação de Aperfeiçoamento Pessoal de Nível Superior (CAPES).

AGRADECIMENTOS

Agradeço a CAPES por fornecer a bolsa de Doutorado, ao PPGBCM por disponibilizar de uma estrutura excelente para desenvolvimento de pesquisa, aos professores do mesmo programa que colaboraram à minha formação acadêmica, bem como minha comissão de acompanhamento de Doutorado, constituída pelos professores Dr. Guido Lenz e Dr. Arthur Germano Fett Neto, o qual também foi revisor e membro da banca de defesa. Agradeço também ao Dr. Michael Sammeth e Dra. Ana Körbes por participarem da banca de defesa.

Agradeço aos colegas do Laboratório de Genomas e Populações de Plantas (LGPP), a Ana Christoff, Andreia, Cláudia, Cordenonsi, Felipe, Fernanda, Fran, Frank, Lorrayne, Júlio, Maurício e Vanessa. Agradeço à professora Dra. Márcia Margis, por ser praticamente uma co-orientadora e ao meu orientador Rogerio Margis, um pesquisador fantástico que sempre me incentivou.

Sou grato aos meus amigos Quinho, Noble, Andersonn, Marcelo, Mateus, Adri e Thanise. Agradeço aos meus gatos “Piu” e “Nino”. Agradeço à minha família, meu irmão Victor, minha irmã Fernanda, meu sobrinho Matheus, minha noiva Josi e à minha mãe Reusa.

MICRORNAS DE SEMENTES DE SOJA MADURA E EM GERMINAÇÃO E *TRANSFER RNA-DERIVED FRAGMENTS* (TRFS) ASSOCIADOS A PROTEÍNAS ARGONAUTAS DE ARABIDOPSIS

Autor: Guilherme Loss de Morais

Orientador: Rogerio Margis

Resumo

O advento de técnicas de sequenciamento de alta eficiência possibilitou o estudo mais aprofundado de pequenos RNAs, como os microRNAs (miRNAs), a classe melhor caracterizada, e a identificação de novas classes como a dos *transfer RNA-derived Fragments* (tRFs). Os pequenos RNAs podem atuar como reguladores negativos da expressão gênica do seu transcrito alvo. Este mecanismo, denominado Silenciamento Gênico Pós-Transcricional (PTGS) ou RNA interferência (RNAi), pode ocorrer pela indução da clivagem do transcrito alvo, ou pela repressão da tradução do mesmo. Em soja, ainda não foram descritos miRNAs atuantes na germinação da semente, os quais foram abordados no primeiro capítulo desta tese. Utilizando duas bibliotecas de sequenciamento de alta eficiência, uma relativa a sementes maduras e outra composta de uma combinação de sementes em germinação (3, 5 e 7 dias), foram identificados um total de 178 microRNAs, sendo 36 inéditos. Dos 178, 8 miRNAs com alvos potencialmente relacionados à germinação da semente, às rotas de auxina, giberelina, metabolismo lipídico, de nitrogênio e homeostase de potencial redox, foram validados por análise de degradoma. O segundo capítulo aborda a caracterização de tRFs em Arabidopsis associados com proteínas Argonauta (AGO), as quais são essenciais ao RNAi. Foram utilizadas 26 bibliotecas de sequenciamento de argonautas imunoprecipitadas (AGO-IP), relativas às AGOs 1, 2, 4, 5, 7 e 9, além de 3 bibliotecas de degradoma. O mapeamento destas sequências nos tRNAs de Arabidopsis revelou que estes pequenos RNAs são majoritariamente associados a AGO1 e 2, sendo a classe 5' de 19 nucleotídeos de comprimento a mais comum. Contudo, estes não obedecem aos critérios de direcionamento a proteínas AGO relativos ao primeiro nucleotídeo do pequeno RNA, como ocorre com miRNAs. Foram identificados quatro transcritos alvos, validados por análise do degradoma, os quais possivelmente sofrem PTGS via tRFs. Ambos os capítulos apresentam uma robusta caracterização *in silico* de pequenos RNAs em plantas inferindo suas possíveis funções. Contudo, mais experimentos devem ser efetuados para confirmação de seus papeis em soja e Arabidopsis.

Tese de Doutorado, Programa de Pós-graduação em Biologia Celular e Molecular, Centro de Biotecnologia, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brasil. (142 p.) Março, 2013.

MICRO*RNA*S FROM MATURE AND GERMINATING SOYBEAN SEEDS AND *TRANSFER RNA-DERIVED FRAGMENTS* (TRFS) ASSOCIATED WITH ARGONAUTE PROTEINS IN ARABIDOPSIS

Author: Guilherme Loss de Morais

Advisor: Rogerio Margis

Abstract

The advent of the deep sequencing approach enabled a better characterization of small RNAs, such as microRNAs (miRNAs), the well-known small RNA class, and the identification of new classes like the transfer RNA-derived Fragments (tRFs). The small RNAs can act as negative regulators of gene expression of their target transcript. This mechanism, known as Post Transcriptional Gene Silencing (PTGS), involves dicing of the target transcript, or translational repression. In soybean, the microRNAs acting on seed germination are unknown. These miRNAs were described in the first chapter of this thesis. Using two deep sequencing libraries, relative to the mature seeds and a combination of germinating seeds (3, 5 and 7 days). A total of 178 miRNAs were identified, including 36 new ones. Eight miRNAs had targets potentially related to seed germination including some acting on auxin and gibberellin pathways, lipid and nitrogen metabolism and redox homeostasis, and were validated by degradome analysis. The second chapter showed the characterization of Argonaut (AGO) associated tRFs in Arabidopsis. AGO is an essential protein for PTGS. A total of 26 deep sequencing libraries from immunoprecipitated Argonauts (AGO-IP), relative to the AGO 1, 2, 4, 7, and 9, plus 3 degradome libraries were used. The tRFs were mainly associated with AGO1 and 2, and the 5' class of 19 nucleotides in length was the most common one. However the tRFs did not follow the rule for AGO loading, were the first nucleotide lead the microRNA to a specific AGO. We identified four tRF target transcripts validated by degradome analysis, which possibly undergo the PTGS pathway. Both chapters present a robust *in silico* characterization of small RNAs in plants, inferring their possible functions. However, more experiments should be performed to confirm their roles in soybean and Arabidopsis

---

Ph.D. Thesis, Graduate Program in Cell and Molecular Biology, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil. (142 p.) March, 2013.

Sumário

LISTA DE ABREVIATURAS

AGO – enzima Argonauta

Ala – Alanina

Arg – Arginina

Gly - Glicina

BLAST – *Basic Local Alignment Sequence Tool* (ferramenta básica de alinhamento local de sequências)

bp – base pair (pares de bases)

cDNA – complementary DNA (DNA complementar)

CDS – Coding Sequence (sequência codificadora)

DCL –Enzima Dicer-Like

DNA – Desoxirribonucleic acid (ácido desoxirribonucleico)

hcRNAs –*Heterocromatic* RNAs (RNAs heterocromáticos)

MFEI – Minimum Free Energy Index (índice de Energia mínima livre)

mRNA – messenger RNA (RNA mensageiro)

miRNA – microRNA

natsiRNAs –*natural anti sence small* RNAs (pequenos RNAs anti senso naturais)

nt – nucleotídeo

pre-miRNA – precursor of microRNA (precursor de microRNA)

PTGS – *Post Transcriptional Gene Silecing* (silenciamento gênico pós transcricional)

pri-miRNA – primary microRNA (microRNA primário)

RDR – RNA Polimerase dependente de RNA

RNA – Ribonucleic acid (ácido ribonucléico)

rRNA – RNA ribossomal

SILAC – *Stable Isotope Labelling with Amino acids in Cell culture* (rotulação de isótopos estáveis em aminoácidos em cultura celular)

siRNA – *small interfering* RNAs (pequenos RNAs de interferência)

SNP - Single Nucleotide Polymorfism (Polimorfismo de nucleotídeo Único)

SVM – *Support Vector Machines* (Máquinas de Suporte Vetorial)

TasiRNAs - *Trans Acting small interfering* RNAs (pequenos RNA de interferência atuantes em *trans*)

tRNA – RNA transportador

tRFs – transfer RNA-derived Fragments (fragmentos derivados de tRNA)

1. INTRODUÇÃO

O presente trabalho aborda os pequenos RNAs de duas plantas modelo, focando-se na caracterização de microRNAs em soja (*Glycine max*), planta modelo para oleaginosas e na descrição de uma classe recente de pequenos RNAs, denominada de transfer RNA-derived Fragmets (tRFs) em *Arabidopsis thaliana*. Ambos os tipos de pequenos RNAs foram analisados por ferramentas de bioinformática, em bibliotecas de sequenciamento de alta eficiência (*Deep Sequencing* ou *High Throughput Sequencing*), desenvolvidas pelo nosso grupo ou disponibilizadas em bancos de dados públicos.

1.1 DO SEQUENCIAMENTO DE ALTA EFICIÊNCIA À BIOINFORMÁTICA

Antes de abordar as classes de pequenos RNAs é interessante introduzir as tecnologias de sequenciamento de alta eficiência, suas aplicações na caracterização de pequenos RNAs, bem como as análises de bioinformática necessárias. Nos últimos anos, houve um grande avanço na maneira em que os dados oriundos de sequenciamento são produzidos, seja por diminuição de custos (Muers, 2011), ou pela grande quantidade de dados que estas tecnologias produzem, oscilando na casa de centenas de milhares de sequências (*reads*) por sequenciamento (Paszkiewicz & Studholme, 2010).

O sequenciamento de alta eficiência está revolucionando a biologia molecular (Bräutigam & Gowik, 2010), sendo aplicado com sucesso no sequenciamento de genomas de plantas (Imelfort & Edwards, 2009), estudos de metagenômica (Coetzee et al., 2010) e metatranscriptômica (Molina et al., 2012), sequenciamento de RNA (RNA seq), de cromatina imunoprecipitada (ChIP seq) (Imelfort &

Edwards, 2009), detecção de padrões de metilação no genoma (Dowen et al., 2012), na descoberta de pequenos RNAs (Zhou et al., 2010) e no sequenciamento de degradoma, o qual se baseia na técnica de 5'-Rapid Amplification of cDNA Ends (RACE) com finalidade de caracterizar padrões de degradação em transcritos (German et al., 2008), sendo bastante utilizada para confirmação de alvos de microRNAs (Eshoo et al., 2011).

Uma das primeiras tecnologias de sequenciamento de alta eficiência foi o pirosequenciamento (Ronaghi, 1998), a qual utiliza a metodologia de sequenciamento por síntese, na qual, conforme ocorre a polimerização da cadeia complementar de DNA, há emissão de luz captada por um receptor. Está técnica foi a primeira a ser incorporada em uma plataforma comercial, chamada de pirosequenciador 454. Nessa plataforma, cada fragmento é ligado a um adaptador que, por sua vez, é ligado a esferas (*beads*) de 28 micrômetros de diâmetro. Os adaptadores são utilizados como região de hibridização de oligonucleotídeos iniciadores (*primers*), necessários para início da polimerização da cadeia de DNA por PCR. A técnica se baseia na utilização das enzimas ATP sulfurilase, luciferase, apirase e na emissão luz originada por atividade enzimática (Ronaghi, 1998).

Cada ciclo de pirosequenciamento inicia com a adição de um único tipo de nucleotídeo na cadeia por uma DNA polimerase, resultando na liberação uma molécula de pirofosfato (PPi), o qual, junto com persulfato de amônio é convertido em ATP pela enzima ATP sulfurilase. O ATP é então utilizado juntamente com a luciferina pela enzima luciferase. O resultado dessa última reação é oxi-luciferina e emissão de luz, a qual é captada por uma câmera CCD. Após cada ciclo, os nucleotídeos não polimerizados na cadeia são degradados pela enzima apirase. Em seguida repete-se o processo, mas com um tipo de nucleotídeo distinto do anterior (Ronaghi, 1998).

Outra tecnologia bastante popular é a utilizada em plataformas Illumina (Solexa, Genome

Analyzer, HiSeq200, MiSeq), as quais foram utilizadas no presente trabalho. Esta tecnologia se baseia na utilização de adaptadores, ancorados em uma placa, que são utilizados como oligonucleotídeos iniciadores para uma PCR dos *amplicons* a serem sequenciados. Ao fim da reação, cada fragmento será representado inúmeras vezes, formando "colônias" de DNA. Este fato é importante, pois cada produto de PCR é sequenciado inúmeras vezes, diminuindo a taxa de erro de sequenciamento. As reações de sequenciamento se baseiam na utilização de nucleotídeos terminadores de cadeia reversíveis ligados a fluoróforos distintos (um diferente para cada nucleotídeo). Conforme a reação de PCR inicia, são adicionados os terminadores de cadeia reversíveis, que interrompem a PCR, permitindo o sinal de fluorescência a ser captado por uma câmera CCD. Em seguida, os nucleotídeos terminadores não ligados são retirados da reação e aqueles ligados ao fragmento são desbloqueados, com a retirada do fluoróforo, permitindo a PCR continuar. Este processo se repete inúmeras vezes, fazendo com que cada nucleotídeo do fragmento a ser sequenciado seja resolvido um a um (Bentley et al., 2012; Mardis, 2008).

Desde a introdução das técnicas de sequenciamento de alta eficiência, independentemente da tecnologia utilizada, as análises de bioinformática tem sido um desafio (Edwards et al., 2013). Devido ao grande número de resultados, há necessidade de robusta estrutura de *hardware* e *softwares* eficientes para estocar, processar, analisar e interpretar esses dados (Koboldt et al., 2010). Contudo, já existem algumas ferramentas que permitem a caracterização destas bibliotecas de sequenciamento. Por exemplo, a ferramenta FASTX (http://hannonlab.cshl.edu/fastx_toolkit/index.html), a qual é utilizada para processamento dos *reads*, tais como retirada de adaptadores e sequências de baixa qualidade; o *software* Velvet (Zerbino & Birney, 2008), que é utilizado para montagem de *contigs*; Bowtie (Langmead et al., 2009), ferramenta usada para mapear *reads* em um genoma ou transcriptoma e

Samtools (Heng Li et al., 2009), ferramenta utilizada para manipular arquivos de mapeamentos de *reads* e análises de SNP *call*. Para análise de pequenos RNAs existe o *UEsmall RNA toolkit* (Moxon et al., 2008), o qual pode ser utilizado na identificação destes, predição de microRNAs, análises de enovelamento de RNAs, entre outras. A disponibilidade de sequências em bancos de dados como o miRBase (Kozomara & Griffiths-Jones, 2011) e de sequenciamento de alta eficiência como o *Gene expression Omnibus* (GEO) (http://www.ncbi.nlm.nih.gov/gds/) favorecem a caracterização de pequenos RNAs.

1.2 OS MICRORNAS

Os pequenos RNAs são uma classe de RNA não codificantes, tendo nos microRNAs ou miRNAs seus representantes melhor caracterizados (Meyers et al., 2008). Os microRNAs são produtos de transcritos (pré-miRNAs) de aproximadamente 200 nt de comprimento. Estes, por sua vez, são resultantes do processamento de pri-miRNAs, os quais são transcritos por uma RNA Polimerase II, a partir de um MIR gene localizado em regiões intergênicas, éxons, ou íntrons (MIRtrons) (Axtell et al., 2011; Voinnet, 2009).

Os pré-miRNAs formam estruturas em forma de grampo, devido a possuírem sequências repetidas invertidas, as quais são clivadas pelo complexo *D-body,* formado por RNAses do tipo III, denominadas *Dicer-Like* (DCL), e outras proteínas acessórias (Voinnet, 2009). Este processamento produz pequenos RNAs entre 21 e 24 nt, posteriormente metilados na terminação 3'-OH pela enzima HEN1, protegendo os mesmos da degradação por nucleases degradadoras de pequenos RNAs (SDN) (Yu et al., 2005). Uma das fitas do miRNA é direcionada ao seu transcrito alvo por proteínas Argonautas (AGO) no

complexo RISC, formando o complexo *miRNA-induced silencing complex* (miRISC) e posteriormente agindo no silenciamento gênico (L. Ding & Han, 2007). A fita complementar é geralmente degradada por uma nuclease SDN (Axtell et al., 2011) (Figura 1). Porém, recentemente foi visto que esta fita complementar, também denominada de miRNA*, também pode ser carregada em enzimas AGO e regular a expressão de transcritos, como o MIR393*, que regulou o gene MEMB12 durante um ensaio de infecção de *Pseudomonas syringae* em *Arabidopsis thaliana* (Xiaoming Zhang et al., 2011).

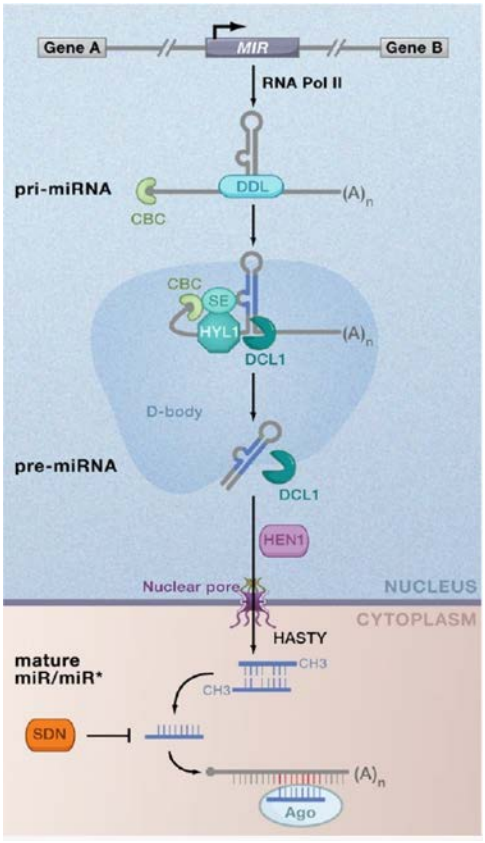


Figura 1: Biogênese de microRNAs. Os pri-miRNAs transcritos por uma RNA Polimerase II são estabilizados pela proteína *Dawdle* (DDL), até o processamento do pré-miRNA no *D-body*, composto pelas proteínas: *Serrate* (SE), *Hyponastic Leaves 1* (HYL1), *Dicer-like1* (DCL1), e *nuclear cap-binding complex* (CBC). O pré-miRNA resultante é novamente clivado por uma DCL1, gerando o miRNA maduro, que será exportado para o citoplasma pela proteína de membrana HASTY, metilado pela HEN1. Uma das fitas do miRNA é carregada pela AGO ao seu transcrito alvo, enquanto que a outra é degradada (adaptado de Voinnet, 2009).

Ao ser direcionado pela AGO ao transcrito alvo, o microRNA pode induzir o silenciamento do transcrito alvo através de *Post Transcriptional Gene Silecing* (PTGS) (Voinnet, 2009) por duas formas: i) clivagem do transcrito alvo, entre a 10-11 base no miRNA ligado ao transcrito alvo (Figura 2A); ii) repressão da maquinaria de tradução, induzida pela ligação do miRNA com o transcrito alvo (Addo-Quaye et al., 2008; Voinnet, 2009) (Figura 2B).

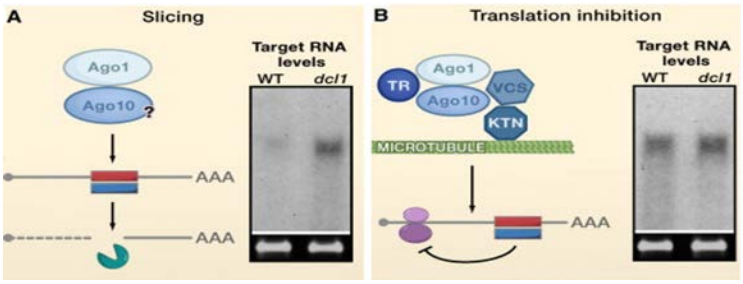


Figura 2: Mecanismos de regulação gênica por microRNAs em plantas por PTGS. **A)** regulação por clivagem (*slicing*), na qual ocorre diminuição dos níveis de transcritos no tipo selvagem (WT), mas não no mutante *dcl1*(*knockout* DCL). **B)** Inibição da tradução do transcrito alvo induzida pelo microRNA. Neste caso não ocorre diminuição no sinal da sonda no *Northern blot* referente ao WT (Adaptado de Voinet, 2009).

Os fragmentos dos transcritos alvo clivados são degradados de maneiras distintas. O fragmento referente à terminação 5' é direcionado a um complexo multiprotéico que possui atividade exonuclease 3'→ 5', denominado exossoma, e o fragmento do transcrito relativo à terminação 3' é degradado pela exonuclease XRN4, a qual possui atividade nucleolítica 5'→ 3' (Huntzinger & Izaurralde, 2011).

Das formas de regulação gênica via miRNAs, a repressão da tradução é pouco caracterizada, sendo que nos primeiros estudos acreditava-se que microRNAs de plantas regulavam majoritariamente seus transcritos alvos por clivagem de transcritos (Huntzinger & Izaurralde, 2011). Porém, a repressão da tradução foi caracterizada em *Arabidopsis thaliana*, na qual o transcrito de APETALA2 foi reprimido pelo MIR172 (X. Chen, 2004) e o transcrito do gene SBP-box foi reprimido pelo MIR156/157

(Gandikota et al., 2007).

1.3 Outras classes de pequenos RNAs

De todo os RNAs em uma célula eucariótica, somente 1% é relativo aos pequenos RNAs (Zhuang, Fuchs, & Robb, 2012), sendo os microRNAs geralmente a classe mais abundante. A distinção entre as diferentes classes é feita pela caracterização da biogênese, processamento, padrão de expressão e função destes na célula (Farazi et al., 2008). Entre as classes de pequenos RNAs encontradas em plantas, se destacam os *small interfering RNAs* (siRNAs), *Trans Acting small interfering RNAs* (TasiRNAs), *natural anti sense small RNAs* (natsiRNAs), *Heterocromatic* RNAs (hcRNAs) (Farazi et al., 2008) e *transfer RNA-derived Fragmets (tRFs)* (Hsieh et al., 2009).

Os siRNAs são a segunda maior classe de pequenos RNAs, tendo tamanho entre 21 a 24 pares de bases (Zhuang et al., 2012). Estes podem ser originados tanto a partir de regiões de transcritos sobrepostos (Ghildiyal & Zamore, 2009), quanto pela ação da enzima RNA Polimerase dependente de RNA (RDR) (Garcia-Ruiz et al., 2010) Os siRNAs tem sua biogênese tanto dependente de DCL, quanto independente, sendo que, na primeira, estes são monofosfatados na extremidade 5', diferentemente dos independentes de DCL, que são polifosfatados (Sijen et al., 2007). Os siRNAs estão envolvidos na degradação de transcritos alvos (PTGS) e controle transcricional pela metilação de DNA e realocação de histonas (Carthew & Sontheimer, 2009).

Os *Trans Acting small interfering RNAs* (TasiRNAs) são pequenos RNAs de 21-22 nt (Farazi et al., 2008) transcritos pela RNA Polimerase II em um precursor TAS, o qual é reconhecido por um microRNA associado a uma proteína AGO. Esta proteína cliva o precursor em pequenos RNAs, os

quais servem de molde para uma RDR, resultando em pequenos RNAs fita dupla processados por uma DCL, liberando pequenos RNAs de 21 nt (Lima et al., 2012) direcionados ao RISC, atuando na regulação gênica, similarmente aos miRNAs e siRNAs. A nomenclatura “Trans” é devido a estes regularem genes com baixa semelhança de sequência ao precursor TAS que lhes originou (Vazquez et al., 2004).

Os *natural anti sense small* RNAs (natsiRNAs) são pequenos RNAs de 21-24 nt de comprimento gerados de transcritos sobrepostos, atuando no controle da expressão de forma semelhante aos siRNAs (Lima et al., 2012). Os *Heterocromatic RNAs* (hcRNAs) são pequenos RNAs de 24 nt de comprimento (Farazi et al., 2008). Os precursores dos hcRNAs são originados de regiões repetitivas, transcritos pela RNA Polimerase IV, convertidos em RNA de fita dupla por uma RDR e processados em pequenos RNAs de fita dupla por uma DCL (Blevins et al., 2009). Uma das fitas é carregada por uma proteína AGO, atuando na metilação de DNA direcionada por RNA (*RNA-directed DNA methylation* -RdDM) (Pontes et al., 2006), alterando a expressão gênica durante a transcrição.

Recentemente, foi caracterizado uma nova classe de pequenos RNAs, originadas de RNAs transportadores, denominadas de *transfer RNA-derived Fragmets* (tRFs) (Lee et al., 2009). Os tRFs foram primeiramente identificados em culturas celulares humanas, e, mais recentemente caracterizados em plantas (Chen et al., 2011; Hsieh et al., 2009). Sua biogênese é associada a diferentes RNAses, incluindo as *Dicers*, as quais resultam em três classes de tRFs distintas, relativas à sua posição no tRNA. O pré-tRNA gera tRFs 3' U a partir da extremidade 3' poli-uridilada, enquanto o tRNA maduro resulta em tRFs 5' e 3' CCA, relativos às extremidades deste (Sobala & Hutvagner, 2011) (Figura 3).

Além disso, uma característica importante é que os tRFs também foram caracterizados como atuantes no PTGS em humanos (Haussecker et al., 2010) (Tabela 1). Porém, é desconhecida a possível

ação de tRFs no silenciamento gênico em plantas.

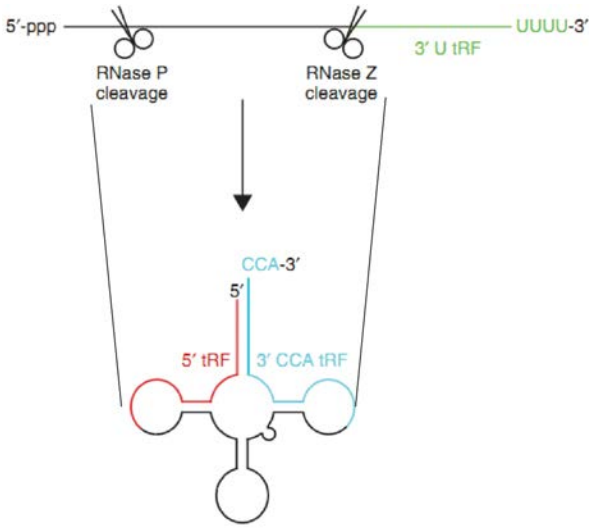


Figura 3: Modelo proposto da biogênese das três classes de tRFs. Em vermelho o 5' tRFs, em azul o 3' CCA tRF, ambos processados do tRNA maduro. Em verde o 3' U tRF, o qual é processado do pré-tRNA (adaptado de (Sobala & Hutvagner, 2011))

Tabela 1: Síntese dos resultados obtidos sobre o possível papel dos tRFs em PTGS (adaptado de (Haussecker et al., 2010))

Parameter	MicroRNA	3` CCA tRF	3` U tRF
Biogenesis	Dicer dependent	Dicer dependent	5′ RNaseZ 3′ Pol III termination
Argonaute association	Ago1,Ago2 = Ago3,Ago4	Ago1,Ago2 ≤ Ago3,Ago4 Enriches 18–20-nt species	Ago1,Ago2 < Ago3,Ago4
RNAi-type *trans*-silencing	Yes	Yes	No

1.4 BIOINFORMÁTICA DE PEQUENOS RNAS

Existem dois desafios principais para bioinformática de pequenos RNAs. O primeiro é relativo à identificação das regiões que originaram o pequeno RNA em estudo. O segundo desafio, mais complexo, refere-se à predição de transcritos alvo afetados via PTGS ou metilação de DNA.

Inicialmente uma adaptação da ferramenta *Basic Local Alignment Search Tool* (BLAST) (Altschul et al., 1990) foi feita para identificar os pequenos RNAs, como descrito em Sunkar e Zhu (2004). Na caracterização de pequenos RNAs, onde os precursores não formam estruturas secundárias em forma de grampo, como em siRNAs, TAsiRNAs, natsiRNAs, hcRNAs, essa metodologia é valida, porém, na caracterização de miRNAs, há necessidade de algoritmos um pouco mais sofisticados. Além do mais, não há uma metodologia consenso e vários grupos desenvolvem metodologias *in house* para caracterização de pequenos RNAs como em Cai et al. (2012), Guo et al. (2009), Hale et al. (2009) e Wei et al. (2009). Estas metodologias muitas vezes não são disponibilizadas pelos autores.

A identificação de microRNAs utilizou a característica destes formarem uma estrutura secundária em forma de grampo, como parâmetro na predição de pré-miRNAs (Wang et al., 2005). Estes pré-miRNAs também apresentam um índice de energia livre mínima (MFEI) menor que as demais classes de RNAs (mRNA, tRNA, rRNA) (Zhang et al., 2006), fazendo com que o MFEI fosse adotado como critério para classificação dos pre-microRNAs em vários trabalhos, como em cevada (Hackenberg et al., 2012) e milho (Kang et al., 2012).

Recentemente nosso grupo de pesquisa, em parceria com o grupo de bioinformática da Faculdade de Ciências da Computação da UFRGS, desenvolveu uma ferramenta de validação de pre-microRNAs denominada FilterPrecursors (Apêndice IX, X).

Esta ferramenta processa os arquivos de mapeamentos de *reads* em pre-microRNAs, avaliando se o padrão de mapeamento confere a formação de uma ou duas colunas de *reads*, relativas ao microRNA e seu microRNA* (Figura 4). A ferramenta FilterPrecursors foi empregada com sucesso em dois trabalhos recentes de caracterização de miRNAs em canola (Körbes et al., 2012) e pitanga (Guzman et al., 2012).

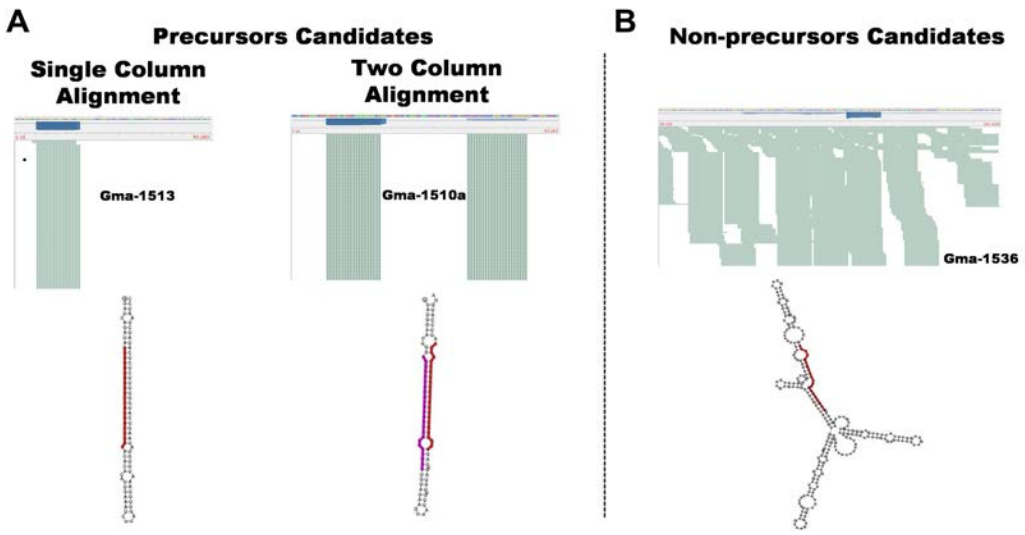


Figura 4: Perfil de mapeamentos de *reads*, processados pela ferramenta FilterPrecursors em três precursores de microRNAs (pre-miRNAs) disponíveis no banco de dados miRBase (http://www.mirbase.org/) oriundos do trabalho de Subramanian et al. (2008). A) pre-microRNAs verdadeiros que contém somente um miRNA maduro (GMA-1513) ou que apresentam miRNA e miRNA* (Gma-1510a). B) Pré-miRNA falso positivo contendo mais do que duas "pilhas" de *reads.* No detalhe as estruturas secundárias de cada pré-miRNA.

O segundo, e mais complexo, desafio de bioinformática de pequenos RNAs é a predição dos transcritos alvos de pequenos RNAs, na qual maior atenção é direcionada aos miRNAs (Ding et al., 2012). Nestes casos, não há necessidade de pareamento perfeito entre o miRNA e o transcrito alvo, podendo haver *gaps, mismatches* e *wobbles* (pareamento guanina – uracila) (Dai et al., 2011). Além do mais, existe uma região do microRNA, do segundo até o oitavo nucleotídeo, denominada "*seed*" a qual exige um pareamento mais estrito entre o miRNA e seu transcrito alvo (Brennecke et al., 2005), sendo relacionada à especificidade do miRNA ao seu alvo (Ruby et al., 2006). Essa região foi caracterizada em microRNAs humanos (Lewis, Burge, & Bartel, 2005), porém em plantas a região *seed* não é bem caracterizada ainda, razão pela qual esta região somente é considerada em predições de transcritos alvo de miRNAs humanos.

As ferramentas de predição de alvos evoluíram desde o miRanda (Enright et al., 2003), a qual fazia uma simples procura por complementaridade de sequências, até ferramentas que utilizam algoritmos de aprendizagem de máquina, tais como *Naïve Bayes* (Yousef et al., 2007), *Support Vector Machines – SVM* (Mitra & Bandyopadhyay, 2011) e *Random Forest*, como em uma ferramenta denominada RFmiRTarget, desenvolvida pelo nosso grupo de pesquisa em parceria como o grupo de bioinformática da Faculdade de Ciências da Computação da UFRGS (Mendoza et al., 2012).

Para a predição de alvos que sofreram PTGS por clivagem, a utilização de degradoma pode diminuir o número de falsos positivos da análise (Yang et al., 2011), aprimorando a mesma. Já existem ferramentas de predição que utilizam degradoma, como CleaveLand (Addo-Quaye, et al., 2009), starBase (Yang et al., 2011), SeqTar (Zheng et al., 2012) e PAREsnip (Folkes et al., 2012). Contudo, cabe ressaltar que este tipo de análise não permite diminuir o número de falsos positivos em PTGS induzido por repressão da maquinaria de tradução.

A técnica *Stable Isotope Labelling with Amino acids in Cell culture* (SILAC) (Oda, et al., 1999; Ong, 2002), é um método proteômico de alta eficiência baseada na comparação de duas culturas celulares, uma delas contendo isótopos pesados, não radioativos, nos resíduos de aminoácidos, permitindo a diferenciação entre ambas (Ong, 2002). As proteínas totais extraídas das duas populações celulares podem ser analisadas simultaneamente por espectrometria de massa, onde os pares de peptídeos idênticos, mas de composição isotópica distinta são diferenciados pela diferença das massas atômicas. Desta maneira, a abundância de cada resíduo de aminoácido (peptídeo) pode ser obtida (Ong, 2002), inferido a expressão gênica em nível da tradução (Thomson et al., 2011).

A técnica de SILAC foi empregada com sucesso na identificação de 12 alvos em células *HeLa* contendo o MIR1 de humanos super expresso (Vinther et al., 2006) e auxiliou na identificação de 10 genes reprimidos pelo MIR143 humano (Yang et al., 2010). Contudo, não existem ferramentas de predição de transcritos alvos utilizando dados de SILAC de maneira similar à utilização de sequenciamento degradoma.

1.5 A SOJA (*GLYCINE MAX*) COMO MODELO DE ESTUDO DE LEGUMINOSAS

A soja (*Glycine max* (L.) Merrill) é uma planta pertencente à família Fabaceae, originária da China, que tem sido utilizada extensivamente como fonte de óleo e proteína no grão. A cultura é considerada uma das mais importantes *commodities* no mundo, com produção de 264 milhões de toneladas de grãos, para a qual o Brasil colabora com a segunda maior produção (FAO, 2013). Além disso, é uma das culturas que mais cresceram em países tropicais no período de 1999 a 2008 (Phalan et al., 2013).

Os grãos de soja são fonte de proteínas e óleo de alta qualidade, apresentando conteúdo médio de

421 Kg^{-1} e 195 Kg^{-1} de proteínas e óleo, respectivamente (Bellaloui, 2010), sendo responsável por 60% do óleo vegetal no mundo (Lee et al., 2007). Além da importância econômica e nutritiva, a soja é considerada uma planta modelo para estudos em espécies oleaginosas (Severin et al., 2010), fatos que impulsionaram o sequenciamento do genoma (Schmutz et al., 2010), projetos de transcriptoma (Cheng & Strömvik, 2008; Severin et al., 2010), e de proteômica (Duressa et al., 2011; Mathesius et al., 2011; Ohyanagi et al., 2012).

Técnicas de engenharia genética têm sido extensivamente utilizadas para desenvolver plantas de soja resistentes ou tolerantes a estresses bióticos e abióticos (Murad & Rech, 2012) ou com maior teor de óleo (Clemente & Cahoon, 2009). Devido aos microRNAs estarem relacionados à regulação fina de rotas metabólicas, tais como aquelas relacionadas a respostas a estresses abióticos e bióticos (Lima et al., 2012; Ni et al., 2012) e desenvolvimento de órgãos (Issue, 2005), os mesmos constituem alvos interessantes para engenharia genética de plantas (Liu & Chen, 2012).

Em soja, já foram caracterizados microRNAs relacionados a estresses bióticos, tais como infecção por *Phytophthora sojae* (Guo et al., 2011), *Fusarium virguliforme* (Radwan et al., 2011), *Phakopsora pachyrhizi* (Kulcheski et al., 2011), a estresses abióticos, tais como privação de fosfato (Zeng et al., 2010) seca (Kulcheski et al., 2011; Li et al., 2011), salinidade e alcalinidade (Li et al., 2011), a interações simbióticas, como nodulação de raízes (Li et al., 2010; Subramanian et al., 2008; Wang et al., 2009), e ao desenvolvimento de tecidos e órgãos, tais como o meristema apical da parte aérea (Wong et al., 2011) e sementes (Shamimuzzaman & Vodkin, 2012; Song et al., 2011) Contudo, até o momento, não existem publicações sobre a caracterização dos microRNAs de soja envolvidos na germinação da semente.

1.6 *ARABIDOPSIS THALIANA* – PLANTA MODELO E OS PEQUENOS RNAS

Arabidopsis thaliana pertence à família Brassicaceae, possui fácil cultivo, ciclo de vida pequeno, com protocolos de transformação genética bem estabelecidos (Zhang et al., 2006). Esta planta emergiu como planta modelo há aproximadamente 25 anos, para pesquisas em biologia molecular e genética vegetal (Koornneef & Meinke, 2010), sendo a primeira planta a ter o genoma sequenciado (The Arabidopsis Genome Initiative, 2000) e ter microRNAs identificados (Reinhart et al., 2002; Rhoades et al., 2002). Atualmente existem mais de 800 publicações sobre pequenos RNAs de *A. thaliana*, 299 precursores de miRNAs depositados no miRBase (release 19) (http://www.mirbase.org) e 12 projetos de sequenciamento de alta eficiência depositados no banco de dados GEO (http://www.ncbi.nlm.nih.gov/geo). Com estudos mais aprofundados de pequenos RNAs, principalmente em *A. thaliana*, começaram a ser identificados, além das enzimas envolvidas na biogênese e PTGS, outros tipos de pequenos RNAs, como os supracitados tRFs.

2 OBJETIVO GERAL

Os avanços nas tecnologias de sequenciamento e em bioinformática permitem maior compreensão da estrutura e funcionalidade de genomas e transcriptomas de plantas. Neste cenário, com a descoberta dos miRNAs e de outras classes de pequenos RNAs, como os tRFs, uma rede de regulação gênica fina e complexa foi estabelecida em plantas.

O presente trabalho tem como objetivo identificar e caracterizar os microRNAs de soja envolvidos com o processo de germinação e os tRFs associados com proteínas Argonautas em Arabidopsis, além da identificação de seus potenciais transcritos alvos.

2.1 OBJETIVOS ESPECÍFICOS

→ Caracterizar as bibliotecas de sequenciamento de alta eficiência obtidas de semente madura e em germinação de *Glycine max L.*;

→ Identificar os microRNAs conhecidos e novos nestas bibliotecas;

→ Identificar transcritos alvos dos microRNAs conhecidos e novos;

→ Utilizar a biblioteca pública de degradoma de sementes madura para validar as clivagens dos potenciais transcritos alvo;

→ Identificar os tRFs de *Arabidopsis thaliana*;

→ Caracterizar os tRFs associados com diferentes proteínas Argonauta de *A. thaliana*;

→ Identificar os prováveis transcritos alvo dos tRFs;

→ Validar a clivagem dos transcritos utilizando bibliotecas públicas de degradoma.

CAPÍTULO I - CHARACTERIZATION OF SOYBEAN MICRORNAS FROM MATURE AND GERMINATING SEEDS

Autores: Guilherme Loss-Morais[a], Ana Paula Körbes[b], Márcia Margis-Pinheiro[b] and Rogerio Margis[a,c]

[a]Laboratório de Genomas e Populações de Plantas , Centro de Biotecnologia, Universidade Federal do Rio Grande do Sul- UFRGS.

[b]Laboratório de Genética Molecular de Plantas , Departamento de Genética, Universidade Federal do Rio Grande do Sul- UFRGS.

[c]Departamento de Biofísica Universidade Federal do Rio Grande do Sul- UFRGS.

Artigo em preparação

Title:

Characterization of soybean microRNAs from mature and germinating seeds

Authors

Guilherme Loss-Morais[a], Ana Paula Körbes[b], Márcia Margis-Pinheiro[b] and Rogerio Margis[a,c,*]

[a]Biotecnology Center, [b]Genetics Department, and [c]Biophysics Department, Federal University of Rio Grande do Sul, Brazil.

[*]Corresponding author

Email addresses:

GL: guilherme.loss@gmail.com

APK: anapkorbes@yahoo.com.br

MPM: marcia.margis@ufrgs.br

RM: rogerio.margis@ufrgs.br (Av. Bento Gonçalves, 9500. Prédio 43431- Campus do Vale - CxP. 15005. Porto Alegre, RS,Brasil. CEP 91501-970

Phone: (51) 3308-6087/3308-6074 Fax: (51) 3308-7309

Running title:

Soybean microRNAs in mature and germinating seeds

Abstract:

The microRNAs (miRNAs) are non-coding RNAs with 19 to 24 nt in length, acting as important regulators of gene expression. Soybean is a protein and oil rich crop used for animal and human feed, but despite domestication and availability of complete genome, there are few transcriptome and microRNA studies during vegetative growth and stress-conditions. Up to now no miRNAs related to germination process have been characterized. Using deep sequencing, we identified 142 known miRNAs and 36 new ones in mature and germinating seeds. Using the mature seed degradome, eight miRNAs were characterized with target transcript related with possible functions in soybean seed germination, involved in auxin and gibberellin pathways, lipid and nitrogen metabolism and redox homeostasis. This study contributes to the understanding of microRNAs functions in plants, relating them to the process of seed germination.

Keywords: *Glycine max*, soybean, seed, germination and microRNAs

Introduction:

The microRNAs (miRNAs) are a class of small non-coding RNAs of 19-24 nucleotides in length, which act as specific and negative regulators of gene expression, thus playing important roles in gene regulation (Bartel 2004; Jones-Rhoades, Bartel et al. 2006; Voinnet 2009).

The miRNA biogenesis is a well defined process, in which a MIR gene is transcribed by RNA Pol II, resulting in the pri-miRNA precursors which are capped at the 5` end and polyadenylated at the 3` end. Posteriorly, an RNAse III dicer-like enzyme (DCL) and HYL1 (Bartel 2004), a dsRNA binding protein, bind to the pri-miRNA complex and cleave the pri-miRNA, releasing the pre-miRNA. The pre-miRNA is also cleaved by a DCL enzyme, releasing a small RNA duplex, named mature miRNA. The

mature miRNA is methylated at the 3` end by HEN1 enzyme and then exported by HASTY enzyme to the cytoplasm (Namuth-Covert et al. 2009). The post-transcriptional gene silencing (PTGS), induced by miRNAs, starts when a selected strand of mature miRNA duplex is loaded into the Argonaute protein, forming the RNA-induced silencing complex (RISC), with subsequent binding to the target transcript (Chen 2005). The RISC induces the PTGS by transcript endonucleolytic cleavage or ‘’slicing’’ at the center of miRNA-target hybrids or repression of its translation (Voinnet 2009).

Soybean (*Glycine max*) is one of the most important oil-seed crops, and accounts for 48% of global oil production in the international market. The oil content generally ranges from 13% to 22% among soybean cultivars (Wang, Zhang et al. 2007). Soybean seeds also contain a high percentage of protein (40%), being considered the most nutritious crop plant (Song, Liu et al. 2011). Moreover, soybean has been adopted as a potential source of biofuels (Xu, Jiang et al. 2010; Song, Liu et al. 2011). The nutritional and economic importance of soybean stimulated the recent publication of its genome sequence (Schmutz, Cannon et al. 2010), RNA-seq Atlas (Severin, Woody et al. 2010), transcriptome (Cheng and Stromvik 2008; Kovinich, Saleem et al. 2011) and proteomic studies (Natarajan, Xu et al. 2007; Arai, Hayashi et al. 2008; Oehrle, Sarma et al. 2008).

High-throughput DNA sequencing technologies boosted the identification of miRNAs (Lister, Gregory et al. 2009). In soybean, miRNAs were characterized in vegetative tissues (Zhang, Pan et al. 2008; Joshi, Yan et al. 2010; Wong, Zhao et al. 2010), nodule related tissues (Subramanian, Fu et al. 2008; Wang, Li et al. 2009; Joshi, Yan et al. 2010; Li, Deng et al. 2010), developmental seed stages (Song, Liu et al. 2011), abiotic and biotic stresses (Kulcheski, De Oliveira et al. 2011; Radwan, Liu et al. 2011).

Seed germination is regulated through an elaborate and interactive signaling network that

integrates diverse environmental cues into hormonal signaling pathways (Park, Kim et al. 2011). A possible role of miRNAs in seed germination process is still poorly understood, with only two works relating miRNAs to seed germination in maize (Ding, Wang et al. 2012) and Arabidopsis (Nonogaki 2008). However there is no characterization of microRNAs related to germination in soybean. The main goal of the present study was the characterization of microRNAs and their putative targets related to germination in soybean, distinguishing miRNAs profiles from mature and germinating seeds.

Materials and methods

Plant material and RNA isolation

Soybean (*Glycine max* cv. Conquista) seeds were grown for 0 (mature seed), 3, 5 and 7 days, in a greenhouse with a temperature regime of 24 ± 2°C, between two layers of moistened filter paper. At the end of each time-point period, the samples were immediately frozen in liquid nitrogen and stored at -80°C until RNA extraction. The samples were ground to a fine powder in liquid nitrogen with mortar and pestle. The total RNA was isolated using NucleoSpin (Macherey-Nagel) RNA extraction kit, following the manufacturer`s instructions. Each sample time point was composed of pools of three to four seeds.

RNA quality was evaluated by electrophoresis on a 1 % agarose gel. The amount of RNA was verified using a Quibit fluorometer and Quant- iT RNA assay kit according to the manufacturer's instructions (Invitrogen, CA, USA). Total RNA (> 10 μg) was sent to Fasteris Life Sciences SA (Plan-les-Ouates, Switzerland) for processing and sequencing using Solexa technology on the Illumina Genome Analyzer GAII. Two small RNAs libraries were constructed; the first library comprises a pool of mature seeds and the second one consists in an equimolar pool of each germinating seed time point. Quality scores are based on the relative confidence of base calls using elements of cluster generation

and image quality. Briefly, the processing by Illumina for the miRNA analyses consisted of the following successive steps: acrylamide gel purification of the RNA bands corresponding to the size range 20-30 nt, ligation of the 3' and 5' adapters to the RNA in two separate subsequent steps each followed by acrylamide gel purification, cDNA synthesis followed by acrylamide gel purification and a final step of PCR amplification to generate the cDNA colonies template library for Illumina sequencing. After removing the adapter sequences, the sequences were trimmed into different read lengths from 19 to 24 nt for further analysis. Only reads with FASTQ > 13 were used.

Accession numbers

The sequencing data are available at the NCBI Gene Expression Omnibus (GEO) ([http://www.ncbi.nlm.nih.gov/geo]). The accession number GSE38373 contains the sequence data of mature and germinating seeds libraries from smallRNA experiments.

Bioinformatics analysis

The bioinformatic approaches used in the present work are shown in supplementary figure 1. Briefly, the soybean miRNAs precursors (pre-miRNAs) from miRBase (release 19) (http://www.mirbase.org/) (Griffiths-Jones 2010) were grouped into unique precursor sequences file. Next, using the Bowtie tool (Langmead, Trapnell et al. 2009), reads (19 to 24 nt) of each library were mapped, separately, against the soybean precursors file, in order to indentify the known soybean miRNAs. The unmapped small RNAs were subjected to further precursor prediction with miRCat prediction tool (http://srna-tools.cmp.uea.ac.uk/plant/cgi-bin/srna-tools.cgi?rm=input_form&tool=mircat) (Moxon, Schwach et al. 2008) with default parameters, with exception of the minimum length of hairpin, changed to 54, which is the smallest plant precursor identified in miRBase database.

The putative new soybean microRNAs precursors were subjected to local BLAST (Altschul, Gish et al. 1990) against soybean mature microRNAs from miRBase. The precursors which showed until four mismatches in BLAST analysis were named “New loci from known precursors”. The remainder precursors were considered new miRNAs precursors. All new precursors followed the criteria for annotation of plant microRNAs (Meyers, Axtell et al. 2008).

The frequencies of each mapped read in the characterized precursors were retrieved by *in house* python scripts, and inferences of expression, based on high-throughput sequencing data, were normalized by the simple RPM method (Reads Per Million).

All mature microRNAs identified in this work were used in target prediction with psRNATarget tool (Dai and Zhao 2011) against all *G. max* transcripts. All target predictions were analyzed with the public available soybean mature seed degradome (Song, Liu et al. 2011) (GSM647200). Target predictions which showed degradome reads in the middle part of target hybridization region and without mismatches were considered degradome confirmed targets.

Results and Discussion

Overview of soybean deep sequencing libraries

In order to identify the miRNA transcriptome related to *G. max* mature and germinating seeds, small RNA libraries were constructed and sequenced by using the Illumina GAII platform. The deep sequencing yielded a total of 3.531.877 reads from mature seed and 4.107.004 reads from germinating seed libraries, respectively. The number of sequences distributed by size, from both libraries, showed that 21 and 24 nucleotides were the most abundant small RNAs (Figure 1A), corroborating previous works of microRNAs characterization in soybean seed, based on deep sequencing approach (Li, Dong

et al. 2011; Song, Liu et al. 2011). The microRNAs identification was carried out only with reads ranging from 19 to 24 nucleotides in length and considering that all reads mapped on the soybean precursors were truly microRNAs. Canonic miRNAs present 21 nt in length (Voinnet 2009), in agreement with our findings, where those of 21 nt in length were the most abundant (Figure 1B) and diverse in sequence (Figure 1C).

Characterization of new soybean miRNAs precursors

In order to identify unpublished soybean miRNAs precursors, all small RNAs unmapped on the known soybean precursors were used for precursor prediction using the miRCat tool. Our analysis revealed 44 new miRNAs, of which 8 were new *loci* of know families, relative to MIR156, 162, 164, 395, 398 and 482 (Supplementary Figure 2) and 36 *loci* of unpublished soybean precursors (Figure 2). As reported previously by Zhang et al. (2006), known plant pre-miRNA precursors have MFEI values higher than other RNAs (tRNA = 0.64, rRNA = 0.59 and mRNA = 0.65) (Zhang, Pan et al. 2006). In our analysis, 32 miRNAs had MFEI value higher than 0.7, suggesting that they are most likely miRNA precursors (Supplementary Figure 3).

Characterization of mature soybean microRNAs

In order to characterize the mature miRNAs, which were mapped to known and new precursors, we retrieved the data relative to sequence abundances and position in the precursors. The complete data are shown in the Supplementary Table 1 and are summarized in the Table 1. The miRNAs precursors can produce mature forms from opposite arms and they are denoted with a -3p or -5p suffix (Meyers, Axtell et al. 2008). In the present work we identified 58 new miRNAs (5p/3p) from known precursors

(Supplementary Table 1), revealing that there are still new miRNAs originating from known precursors.

IsomiRNAs variants of canonical microRNAs have been considered products of inaccuracies in Dicer pre-miRNA processing (Guo and Lu 2010). A total of 249 isomiRNAs were found from known precursors plus 41 isomiRNAs in the newly described miRNAs (Table 1, Supplementary Table 1), corroborating previous works which had found isomiRNAs using high-throughput sequencing in soybean (Kulcheski, de Oliveira et al. 2011; Song, Liu et al. 2011), *Medicago truncatula* (Lelandais-Briere, Naya et al. 2009) and *Arabidopsis thaliana* (Hsieh, Lin et al. 2009).

Differing from isomiRNAs, miRNAs offset (moRNAs) are small RNAs adjacent to the canonical microRNAs and were first characterized *in Ciona intestinalis* (Shi, Hendrix et al. 2009). Here we found that soybean MIR159a, 169f and 482b can produce moRNAs (Table 1, Supplementary Table 1, and Supplementary Figure 4), corroborating the recent publication of Zhang et al. (2010), which identified moRNAs in *A. thaliana*, *Oryza sativa*, *Physcomitrella patens*, *Medicago truncatula* and *Populus trichocarpa*.

High throughout sequencing can not only identify miRNAs but can also accurately measure miRNA expression (Hoen, Ariyurek et al. 2008; Moldovan, Spriggs et al. 2009). Thus, we used the abundance of mapped reads (Reads Per Million- RPM) (Curaba, Spriggs et al. 2012) to infer miRNA expression (Figure 3). Deep sequencing libraries from mature and germinating seeds were also compared with libraries from leaves and roots (Kulcheski, de Oliveira et al. 2011). Figure 3A shows the 6 more representative microRNA families in the four deep sequencing libraries. MIR159 was ubiquitous in all libraries. In Arabidopsis, MIR159 targets GAMYB-like proteins (Palatnik, Wollmann et al. 2007) and accumulates predominantly in young leaves and flowers, adult rosette leaves, cauline leaves and siliques (Achard, Herr et al. 2004) and was over expressed in Arabidopsis roots under

hypoxia treatment (Moldovan, Spriggs et al. 2009). The MIR159 family was the most representative miRNA family in the root library, (Figure 3A). This library was constructed from roots of soybean cultivated in a hydroponic system. This cultivation method might have contributed to MIR159 representativeness, since hydroponic system could lead to some degree of hypoxia stress. MIR159 was also well represented in mature and germinating seeds, indicating that this miRNA could be related to germinating process, corroborating previous works, in Arabidopsis. MIR159 regulates the expression of GAMYB-like genes in seeds, which participate in GA-induced pathways required for aleurone grain development (Alonso-Peral, Li et al. 2010) and in germinating seeds, in an ABA-induced pathway (Reyes and Chua 2007).

The MIR156, 166, 167, 169, 1507 were also abundant in mature and germinating seeds (Figure 3A). In Arabidopsis, the SQUAMOSA PROMOTER-BINDING PROTEIN (SBP)-LIKE (SPL) genes that are targeted by MIR156 encode proteins that participate in the regulation of the post-germinative switch, from the cotyledon stage to the vegetative-leaf stage (Nonogaki 2010). The MIR166 targets mRNAs coding for HD-Zip transcription factors including Phabulosa (PHB) and Phavoluta (PHV) that regulate axillary meristem initiation and leaf development (Rhoades, Reinhart et al. 2002) and showed the highest abundance during the very early stage of seed germination in maize (Wang, Liu et al. 2011) The MIR167 targets AUXIN RESPONSIVE FACTORS (ARFs) (Wu, Tian et al. 2006), which bind to auxin response promoter elements and mediate gene expression responses to the phytohormone auxin (Hagen and Guilfoyle 2002). Moreover, both MIR167 and ARF genes were related to seed development and germination in maize (Xing, Pudake et al. 2011). The MIR169 targets mRNAs coding for CCAAT binding factor (CBF)-HAP2-like proteins (Rhoades, Reinhart et al. 2002) and might be involved in seed development and germination in Arabidopsis seeds (Martin, Liu et al. 2005). The

MIR1507 is exclusively of the Fabaceae family and was found in nitrogen-fixing soybean nodules (Subramanian, Fu et al. 2008; Wang, Li et al. 2009), shoot apical meristem (Wong, Zhao et al. 2010) and was strongly expressed in soybean seed coat and cotyledons (Zabala, Campos et al. 2012); however, the functions of MIR1507 in mature and germinating seeds are currently unknown, as well as it target. Taken together, these data corroborate that this miRNAs could exert important functions in seed maintenance and germinating processes.

Eight miRNAs, which were up or down-regulated in the two deep-sequencing libraries with targets confirmed by degradome analysis, were analyzed in more detail (Figure 3B and C). The MIR408 was up-regulated in germinating seed library and targets Plantacyanin mRNA, as described in Arabidopsis (Abdel-Ghany and Pilon 2008). Plantacyanin is a plant-specific protein that contains a single copper ion (Ryden and Hunt 1993) and has been suggested as a signaling molecule in *M. truncatula* nodules (Fedorova, van de Mortel et al. 2002) and also in primary defense responses in *Spinacia oleracea* and *A. thaliana* (Nersissian, Immoos et al. 1998), but its role in germination is unknown. The MIR393 3p was also described in *P. trichocarpa* (Puzey, Karger et al. 2012) and *A. lyrata* (Fahlgren, Jogdeo et al. 2010), but no targets were found. The soybean MIR393 3p targets a long-chain acyl-CoA synthetase (LACS) mRNA, related to lipid degradation, by the conversion of fatty acids into CoA esters (Baud and Lepiniec 2010). MIR393 3p was up-regulated in comparison to the seed library (Figure 3B), but were down-regulated in comparison to the library from leaves (Supplementary Figure 5). This down-regulation of MIR393 3p could promote the up-regulation of LACS, corroborating the work of Fulda et. Al (2004), which showed that this enzyme was highly expressed during Arabidopsis germination (Fulda, Schnurr et al. 2004).

The MIR2111 5p targets a galactose oxidase/kelch repeat-containing transcript, but its role in

plants is unknown. The MIR4397 3p1 is an isomiRNA of MIR4397 3p, however, considering the position criteria for annotation of plant miRNAs (Meyers, Axtell et al. 2008), the first one seems to be more suitable as 3p than the second one (data not shown). The MIR4397 3p1 targets a nitrilase transcript, which plays a role in nitrogen metabolism, cleaving nitriles (Bork and Koonin 1994). Moreover, the knockout of nitrilases in *Z. mays* resulted in shorter radicles (Kriechbaumer, Park et al. 2007). Thus, the possible regulation of nitrilase by the MIR4397 3p1 could be related to radicle formation in soybean germinating seeds.

The MIR5041 3p is a new mature miRNA opposite to MIR5041 5p previously described by Radwan et al. (2011) that targets a transcript encoding a gibberellin regulated protein. Gibberellins act antagonistically to Abscisic Acid (ABA), breaking ABA-induced dormancy, promoting germination (Steber and McCourt 2001); thus, MIR5041 3p could be involved in soybean seed germination.

Herein we found that soybean MIR164 5p was up-regulated in germinating seeds (Figure 3B) and targets a no apical meristem (NAM) transcription factor, corroborating previous works (Song, Liu et al. 2011; Shamimuzzaman and Vodkin 2012). The NAM protein was indicated as having a role in determining positions of meristem primordia (Souer, van Houwelingen et al. 1996). Moreover, this transcription factor was related to auxin pathway in Arabidopsis seed germination (Park, Kim et al. 2011). Soybean MIR164 5p could indirectly regulate the development of meristem in germinating seeds, although more experiments are needed to confirm this possibility.

The new MIR27 5p targets a protein containing Domain of Unknown Function and currently does not have a defined role in soybean seed germination. The MIR862 3p is down-regulated in germinating seeds (Figure 3B) and targets a glutaredoxin transcript. Glutaredoxins were related to cell redox homeostasis (Foyer and Noctor 2005) and act as an early signal in the germination of seeds

(Buchanan, Schurmann et al. 1994). Moreover, the soybean seed is rich in storage proteins such as glycinin (Scott, Jung et al. 1992) and their mobilization during germination by proteolysis could need a previous step of reduction of disulfide bridges from storage proteins, which would make the storage proteins more susceptible to proteolysis. The down-regulation of MIR MIR862 3p could lead to an up-regulation of Glutaredoxins during the soybean seed germination, favoring the proteolysis, as aforementioned.

The present work described a set of soybean miRNA which could potentially regulate lipid and nitrogen metabolism, auxin and gibberellin pathways and redox maintenance. More experiments are necessary to better characterize miRNAs related to these processes. Towards this future goal, we are currently developing a set of primers for RT-qPCR miRNA and target expression analysis and target cleavage by 5` RACE techniques.

References

Abdel-Ghany, S. E. and M. Pilon (2008). "MicroRNA-mediated systemic down-regulation of copper protein expression in response to low copper availability in Arabidopsis." J Biol Chem 283(23): 15932-45.

Achard, P., A. Herr, et al. (2004). "Modulation of floral development by a gibberellin-regulated microRNA." Development 131(14): 3357-3365.

Alonso-Peral, M. M., J. Li, et al. (2010). "The microRNA159-regulated GAMYB-like genes inhibit growth and promote programmed cell death in Arabidopsis." Plant Physiol 154(2): 757-71.

Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." J Mol Biol 215(3): 403-10.

Arai, Y., M. Hayashi, et al. (2008). "Proteomic analysis of highly purified peroxisomes from etiolated soybean cotyledons." Plant Cell Physiol 49(4): 526-39.

Bartel, D. P. (2004). "MicroRNAs: genomics, biogenesis, mechanism, and function." Cell 116(2): 281-97.

Baud, S. and L. Lepiniec (2010). "Physiological and developmental regulation of seed oil production." Prog Lipid Res 49(3): 235-49.

Bork, P. and E. V. Koonin (1994). "A new family of carbon-nitrogen hydrolases." Protein Sci 3(8):

1344-6.
Buchanan, B. B., P. Schurmann, et al. (1994). "Thioredoxin: a multifunctional regulatory protein with a bright future in technology and medicine." Arch Biochem Biophys 314(2): 257-60.
Chen, X. (2005). "MicroRNA biogenesis and function in plants." FEBS Lett 579(26): 5923-31.
Cheng, K. C. and M. V. Stromvik (2008). "SoyXpress: a database for exploring the soybean transcriptome." BMC Genomics 9: 368.
Curaba, J., A. Spriggs, et al. (2012). "miRNA regulation in the early development of barley seed." BMC Plant Biol 12: 120.
Dai, X. and P. X. Zhao (2011). "psRNATarget: a plant small RNA target analysis server." Nucleic Acids Res 39(Web Server issue): W155-9.
Ding, D., Y. Wang, et al. (2012). "MicroRNA transcriptomic analysis of heterosis during maize seed germination." PLoS One 7(6): e39578.
Fahlgren, N., S. Jogdeo, et al. (2010). "MicroRNA gene evolution in *Arabidopsis lyrata* and *Arabidopsis thaliana*." Plant Cell 22(4): 1074-89.
Fedorova, M., J. van de Mortel, et al. (2002). "Genome-wide identification of nodule-specific transcripts in the model legume *Medicago truncatula*." Plant Physiol 130(2): 519-37.
Foyer, C. H. and G. Noctor (2005). "Redox homeostasis and antioxidant signaling: a metabolic interface between stress perception and physiological responses." Plant Cell 17(7): 1866-75.
Fulda, M., J. Schnurr, et al. (2004). "Peroxisomal Acyl-CoA synthetase activity is essential for seedling development in *Arabidopsis thaliana*." Plant Cell 16(2): 394-405.
Griffiths-Jones, S. (2010). "miRBase: microRNA sequences and annotation." Curr Protoc Bioinformatics Chapter 12: Unit 12 9 1-10.
Guo, L. and Z. Lu (2010). "Global expression analysis of miRNA gene cluster and family based on isomiRs from deep sequencing data." Comput Biol Chem 34(3): 165-71.
Hagen, G. and T. Guilfoyle (2002). "Auxin-responsive gene expression: genes, promoters and regulatory factors." Plant Mol Biol 49(3-4): 373-85.
Hsieh, L. C., S. I. Lin, et al. (2009). "Uncovering small RNA-mediated responses to phosphate deficiency in Arabidopsis by deep sequencing." Plant Physiol 151(4): 2120-32.
Jones-Rhoades, M. W., D. P. Bartel, et al. (2006). "MicroRNAs and their regulatory roles in plants." Annu Rev Plant Biol 57: 19-53.
Joshi, T., Z. Yan, et al. (2010). "Prediction of novel miRNAs and associated target genes in *Glycine max*." BMC Bioinformatics 11 Suppl 1: S14.
Kovinich, N., A. Saleem, et al. (2011). "Combined analysis of transcriptome and metabolite data reveals extensive differences between black and brown nearly-isogenic soybean (*Glycine max*) seed coats enabling the identification of pigment isogenes." BMC Genomics 12: 381.
Kriechbaumer, V., W. J. Park, et al. (2007). "Maize nitrilases have a dual role in auxin homeostasis and beta-cyanoalanine hydrolysis." J Exp Bot 58(15-16): 4225-33.
Kulcheski, F. R., L. F. de Oliveira, et al. (2011). "Identification of novel soybean microRNAs involved in abiotic and biotic stresses." BMC Genomics 12: 307.
Langmead, B., C. Trapnell, et al. (2009). "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." Genome Biol 10(3): R25.
Lelandais-Briere, C., L. Naya, et al. (2009). "Genome-wide *Medicago truncatula* small RNA analysis revealed novel microRNAs and isoforms differentially regulated in roots and nodules." Plant Cell 21(9): 2780-96.

Li, H., Y. Deng, et al. (2010). "Misexpression of miR482, miR1512, and miR1515 increases soybean nodulation." Plant Physiol 153(4): 1759-70.
Li, H., Y. Dong, et al. (2011). "Characterization of the stress associated microRNAs in *Glycine max* by deep sequencing." BMC Plant Biol 11(1): 170.
Lister, R., B. D. Gregory, et al. (2009). "Next is now: new technologies for sequencing of genomes, transcriptomes, and beyond." Curr Opin Plant Biol 12(2): 107-18.
Martin, R. C., P. P. Liu, et al. (2005). "Simple purification of small RNAs from seeds and efficient detection of multiple microRNAs expressed in *Arabidopsis thaliana* and tomato (*Lycopersicon esculentum*) seeds." Seed Science Research 15(4): 319-328.
Meyers, B. C., M. J. Axtell, et al. (2008). "Criteria for annotation of plant MicroRNAs." Plant Cell 20(12): 3186-90.
Moldovan, D., A. Spriggs, et al. (2009). "Hypoxia-responsive microRNAs and trans-acting small interfering RNAs in Arabidopsis." J Exp Bot 61(1): 165-77.
Moxon, S., F. Schwach, et al. (2008). "A toolkit for analysing large-scale plant small RNA datasets." Bioinformatics 24(19): 2252-3.
Natarajan, S., C. Xu, et al. (2007). "Proteomic and genetic analysis of glycinin subunits of sixteen soybean genotypes." Plant Physiol Biochem 45(6-7): 436-44.
Nersissian, A. M., C. Immoos, et al. (1998). "Uclacyanins, stellacyanins, and plantacyanins are distinct subfamilies of phytocyanins: plant-specific mononuclear blue copper proteins." Protein Sci 7(9): 1915-29.
Nonogaki, H. (2008). "Repression of transcription factors by microRNA during seed germination and postgerminaiton: Another level of molecular repression in seeds." Plant Signal Behav 3(1): 65-7.
Nonogaki, H. (2010). "MicroRNA gene regulation cascades during early stages of plant development." Plant Cell Physiol 51(11): 1840-6.
Oehrle, N. W., A. D. Sarma, et al. (2008). "Proteomic analysis of soybean nodule cytosol." Phytochemistry 69(13): 2426-38.
Palatnik, J. F., H. Wollmann, et al. (2007). "Sequence and expression differences underlie functional specialization of Arabidopsis MicroRNAs miR159 and miR319." Developmental Cell 13(1): 115-125.
Park, J., Y. S. Kim, et al. (2011). "Integration of auxin and salt signals by the NAC transcription factor NTM2 during seed germination in Arabidopsis." Plant Physiol 156(2): 537-49.
Puzey, J. R., A. Karger, et al. (2012). "Deep annotation of *Populus trichocarpa* microRNAs from diverse tissue sets." PLoS One 7(3): e33034.
Radwan, O., Y. Liu, et al. (2011). "Transcriptional analysis of soybean root response to *Fusarium virguliforme*, the causal agent of sudden death syndrome." Mol Plant Microbe Interact 24(8): 958-72.
Reyes, J. L. and N. H. Chua (2007). "ABA induction of miR159 controls transcript levels of two MYB factors during Arabidopsis seed germination." Plant J 49(4): 592-606.
Rhoades, M. W., B. J. Reinhart, et al. (2002). "Prediction of plant microRNA targets." Cell 110(4): 513-20.
Ryden, L. G. and L. T. Hunt (1993). "Evolution of protein complexity: the blue copper-containing oxidases and related proteins." J Mol Evol 36(1): 41-66.
Schmutz, J., S. B. Cannon, et al. (2010). "Genome sequence of the palaeopolyploid soybean." Nature

463(7278): 178-83.
Scott, M. P., R. Jung, et al. (1992). "A protease responsible for post-translational cleavage of a conserved Asn-Gly linkage in glycinin, the major seed storage protein of soybean." Proc Natl Acad Sci U S A 89(2): 658-62.
Severin, A. J., J. L. Woody, et al. (2010). "RNA-Seq Atlas of *Glycine max*: a guide to the soybean transcriptome." BMC Plant Biol 10: 160.
Shamimuzzaman, M. and L. Vodkin (2012). "Identification of soybean seed developmental stage-specific and tissue-specific miRNA targets by degradome sequencing." BMC Genomics 13: 310.
Shi, W., D. Hendrix, et al. (2009). "A distinct class of small RNAs arises from pre-miRNA-proximal regions in a simple chordate." Nat Struct Mol Biol 16(2): 183-9.
Song, Q. X., Y. F. Liu, et al. (2011). "Identification of miRNAs and their target genes in developing soybean seeds by deep sequencing." BMC Plant Biol 11: 5.
Souer, E., A. van Houwelingen, et al. (1996). "The no apical meristem gene of Petunia is required for pattern formation in embryos and flowers and is expressed at meristem and primordia boundaries." Cell 85(2): 159-70.
Steber, C. M. and P. McCourt (2001). "A role for brassinosteroids in germination in Arabidopsis." Plant Physiol 125(2): 763-9.
Subramanian, S., Y. Fu, et al. (2008). "Novel and nodulation-regulated microRNAs in soybean roots." BMC Genomics 9: 160.
Hoen, P. A., Y. Ariyurek, et al. (2008). "Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms." Nucleic Acids Res 36(21): e141.
Unver, T., D. M. Namuth-Covert, et al. (2009). "Review of current methodological approaches for characterizing microRNAs in plants." Int J Plant Genomics 2009: 262463.
Voinnet, O. (2009). "Origin, biogenesis, and activity of plant microRNAs." Cell 136(4): 669-87.
Wang, H. W., B. Zhang, et al. (2007). "The soybean Dof-type transcription factor genes, GmDof4 and GmDof11, enhance lipid content in the seeds of transgenic Arabidopsis plants." Plant J 52(4): 716-29.
Wang, L., H. Liu, et al. (2011). "Identification and characterization of maize microRNAs involved in the very early stage of seed germination." BMC Genomics 12: 154.
Wang, Y., P. Li, et al. (2009). "Identification and expression analysis of miRNAs from nitrogen-fixing soybean nodules." Biochem Biophys Res Commun 378(4): 799-803.
Wong, C. E., Y. T. Zhao, et al. (2010). "MicroRNAs in the shoot apical meristem of soybean." J Exp Bot 62(8): 2495-506.
Wu, M. F., Q. Tian, et al. (2006). "Arabidopsis microRNA167 controls patterns of ARF6 and ARF8 expression, and regulates both female and male reproduction." Development 133(21): 4211-8.
Xing, H., R. N. Pudake, et al. (2011). "Genome-wide identification and expression profiling of auxin response factor (ARF) gene family in maize." BMC Genomics 12: 178.
Xu, J., J. Jiang, et al. (2010). "Production of hydrocarbon fuels from pyrolysis of soybean oils using a basic catalyst." Bioresour Technol 101(24): 9803-6.
Zabala, G., E. Campos, et al. (2012). "Divergent patterns of endogenous small RNA populations from seed and vegetative tissues of *Glycine max*." BMC Plant Biol 12(1): 177.
Zeng, H. Q., Y. Y. Zhu, et al. (2010). "Analysis of phosphorus-deficient responsive miRNAs and cis-

elements from soybean (*Glycine max* L.)." J Plant Physiol 167(15): 1289-97.
Zhang, B., X. Pan, et al. (2008). "Identification of soybean microRNAs and their targets." Planta 229(1): 161-82.
Zhang, B. H., X. P. Pan, et al. (2006). "Evidence that miRNAs are different from other RNAs." Cell Mol Life Sci **63(2): 246-54.**
Zhang, W., S. Gao, et al. (2010). "Multiple distinct small RNAs originate from the same microRNA precursors." Genome Biol 11(8): R81.

Figure Legends

Figure 1: Size distribution and the frequency of small RNAs and microRNAs from mature and germinating seeds from deep sequencing libraries. **A)** Abundance of Small RNAs by read size in mature seeds (red) and germinating seeds (blue) libraries. **B)** Abundance of microRNAs ranging from 19 to 24 nt in size in comparison to raw libraries. **C)** Diversity of microRNAs ranging from 19 to 24 nt in size in comparison to raw libraries.

Figure 2: Secondary structures of the 36 new microRNAs precursors. Magenta: 5p microRNAs; red: 3p microRNAs. The ΔG is also shown for each precursor.

Figure 3: Expression analysis of miRNAs in mature and germinating seeds and degradome analysis. **A)** The six more representative microRNA families in mature and germinating seeds and from leaves and roots. **B)** Expression of MIR408, 393, 2111, 4397, 5041, 164, New-27 and 862. In red, microRNAs up regulated in germinating seeds library; in blue, microRNAs up regulated in mature seeds library. **C)** Degradome analysis showing the miRNA and target alignment. The bold letter and underlined region shows the cleavage site and the black arrowhead shows the number of degradome reads mapped in the cleavage site region (absolute numbers). In red, known miRNAs; In green, new 5p/3p microRNAs and in blue, unpublished microRNAs.

Supplementary Figure 1: Schematic representation of bioinformatics approaches used for microRNA identification and target prediction.

Supplementary Figure 2: Secondary structures of the seven new *loci* of known precursors. Magenta: 5p microRNAs; red: 3p microRNAs. The ΔG is also shown for each precursor.

Supplementary Figure 3: Minimum Free Energy Index (MFEI) of the predicted soybean precursors. The red line shows the MFEI cutoff.

Supplementary Figure 4: Secondary structures of the MIR159a, 169f and 482b precursors showing the miRNA offset (moRNAs).Magenta: 5p microRNAs; red: 3p microRNAs. The ΔG is also shown for each precursor.

Supplementary Figure 5: Expression analysis of the MIR393 3p in leaves and germinating seed libraries.

Figure 1

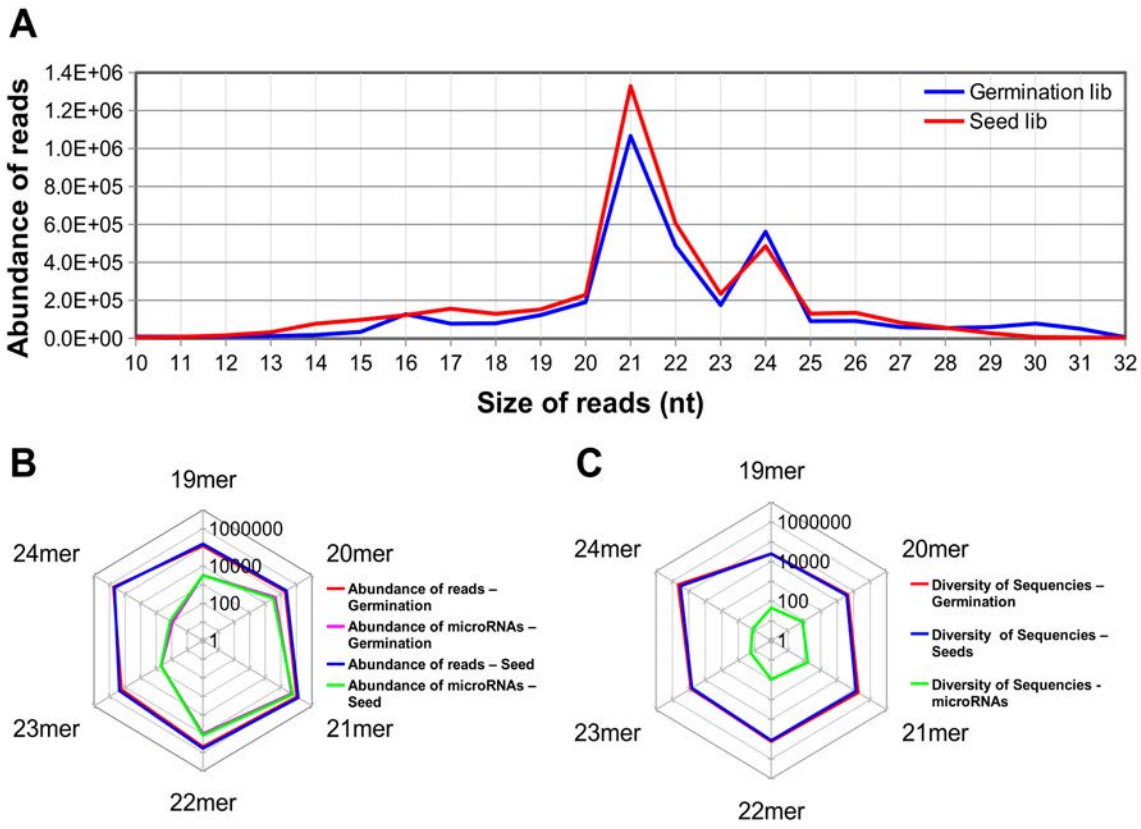

A
Abundance of reads
1.4E+06
1.2E+06
1.0E+06
8.0E+05
6.0E+05
4.0E+05
2.0E+05
0.0E+00
10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32
Size of reads (nt)
Germination lib
Seed lib
B
19mer
20mer
21mer
22mer
23mer
24mer
1000000
10000
100
1
Abundance of reads – Germination
Abundance of microRNAs – Germination
Abundance of reads – Seed
Abundance of microRNAs – Seed
C
19mer
20mer
21mer
22mer
23mer
24mer
1000000
10000
100
1
Diversity of Sequencies – Germination
Diversity of Sequencies – Seeds
Diversity of Sequencies - microRNAs

Figure 2

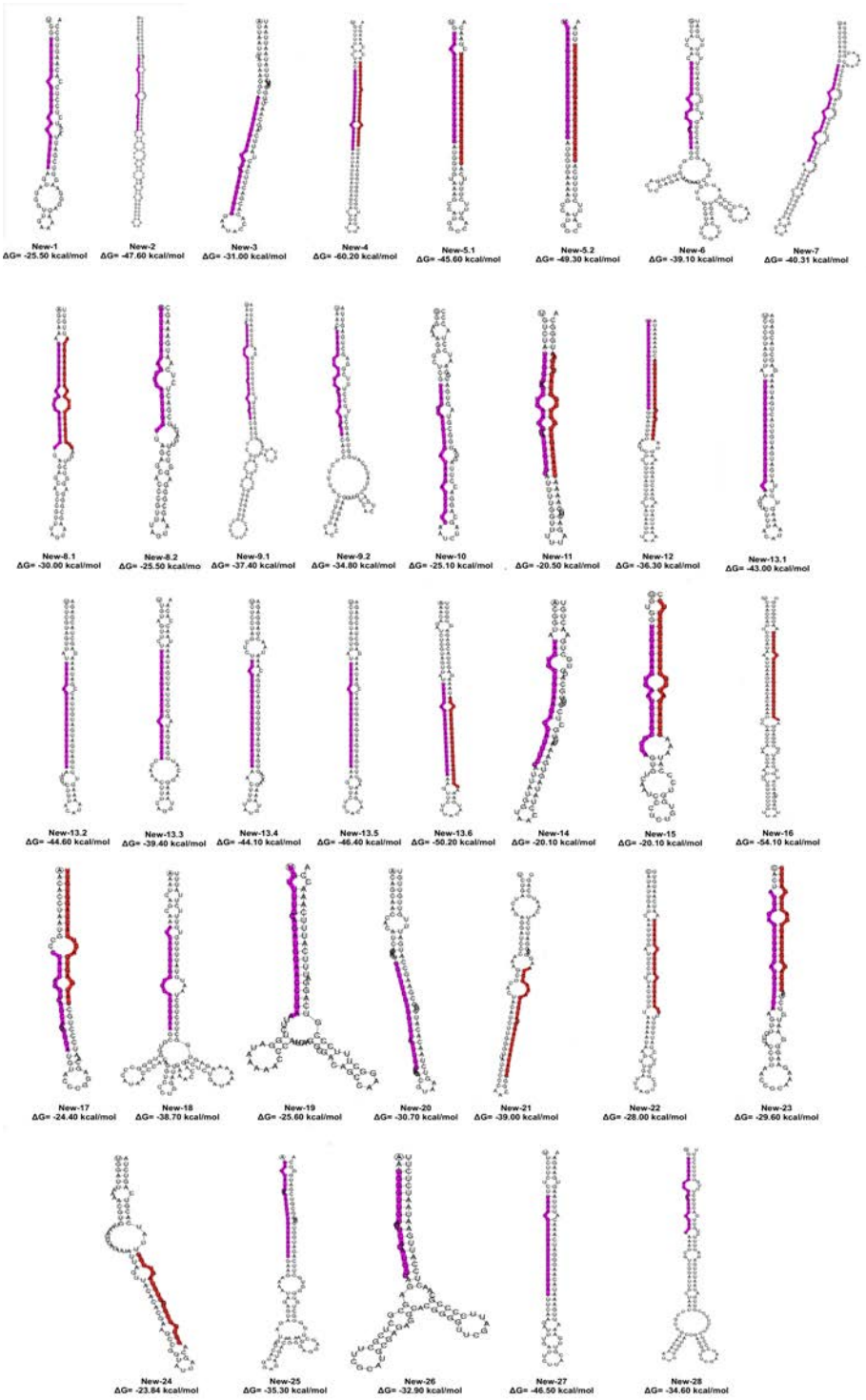
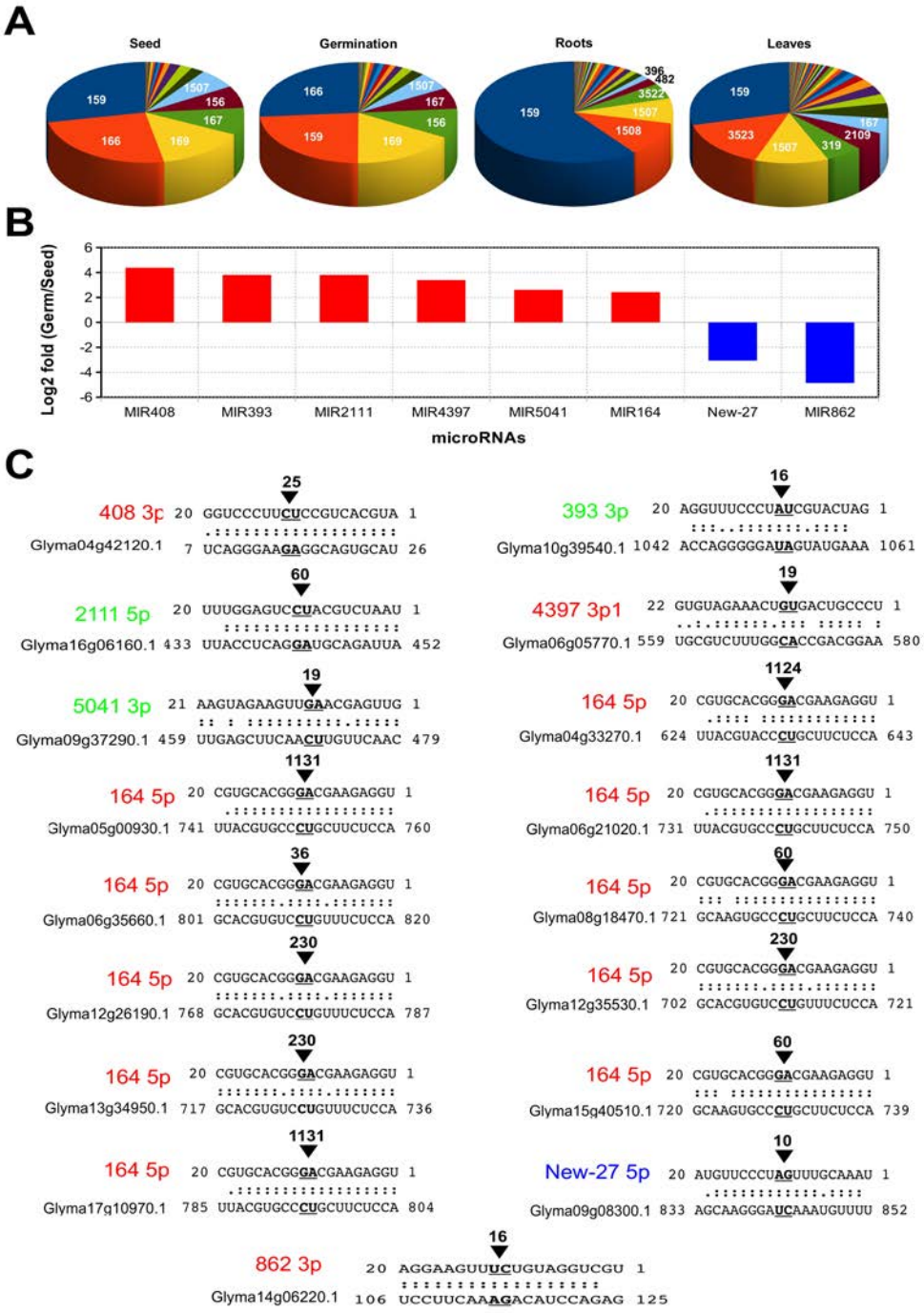
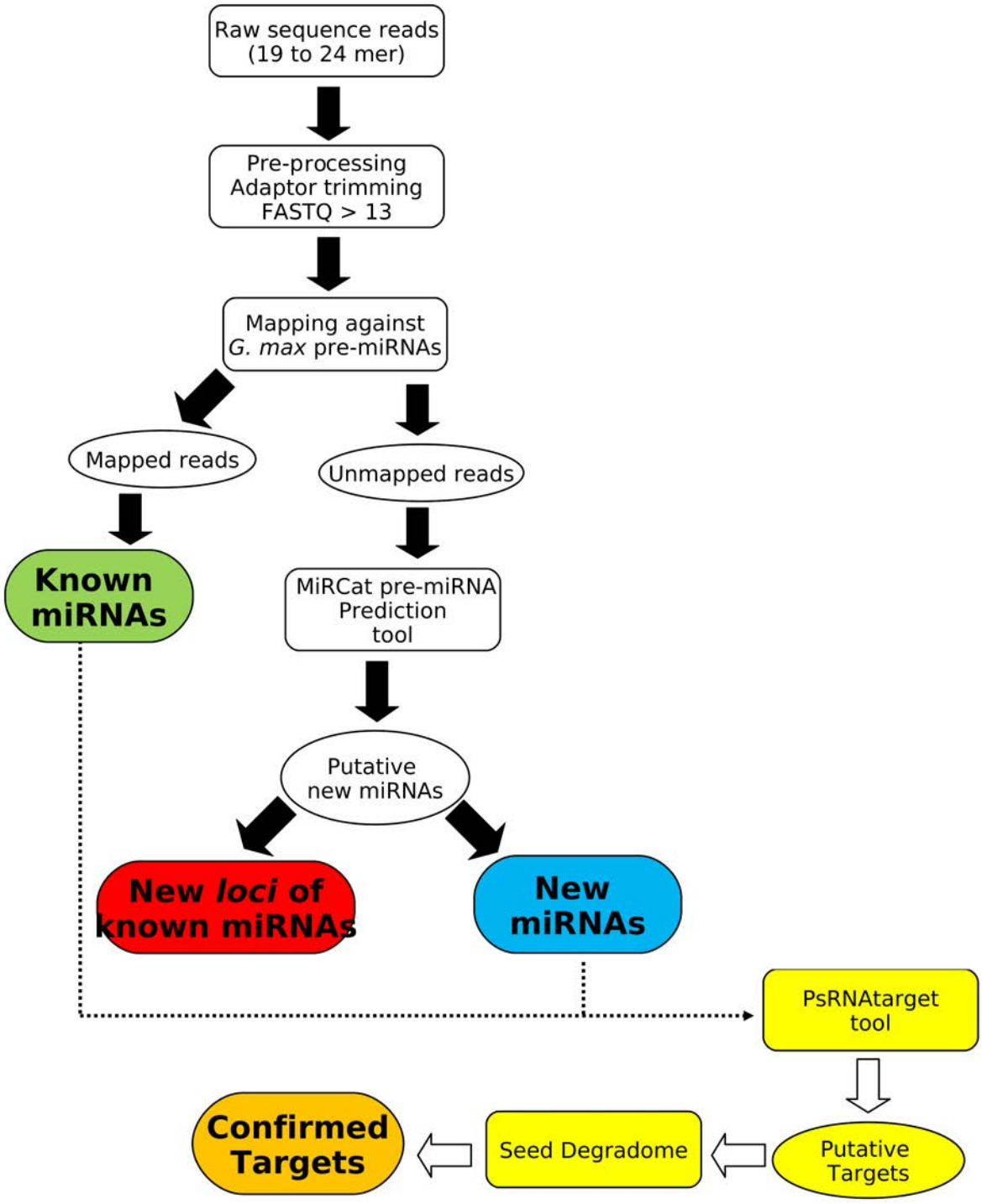

New-1
ΔG= -25.50 kcal/mol
New-2
ΔG= -47.60 kcal/mol
New-3
ΔG= -31.00 kcal/mol
New-4
ΔG= -60.20 kcal/mol
New-5.1
ΔG= -45.60 kcal/mol
New-5.2
ΔG= -49.30 kcal/mol
New-6
ΔG= -39.10 kcal/mol
New-7
ΔG= -40.31 kcal/mol
New-8.1
ΔG= -30.00 kcal/mol
New-8.2
ΔG= -25.50 kcal/mo
New-9.1
ΔG= -37.40 kcal/mol
New-9.2
ΔG= -34.80 kcal/mol
New-10
ΔG= -25.10 kcal/mol
New-11
ΔG= -20.50 kcal/mol
New-12
ΔG= -36.30 kcal/mol
New-13.1
ΔG= -43.00 kcal/mol
New-13.2
ΔG= -44.60 kcal/mol
New-13.3
ΔG= -39.40 kcal/mol
New-13.4
ΔG= -44.10 kcal/mol
New-13.5
ΔG= -46.40 kcal/mol
New-13.6
ΔG= -50.20 kcal/mol
New-14
ΔG= -20.10 kcal/mol
New-15
ΔG= -20.10 kcal/mol
New-16
ΔG= -54.10 kcal/mol
New-17
ΔG= -24.40 kcal/mol
New-18
ΔG= -38.70 kcal/mol
New-19
ΔG= -25.60 kcal/mol
New-20
ΔG= -30.70 kcal/mol
New-21
ΔG= -39.00 kcal/mol
New-22
ΔG= -28.00 kcal/mol
New-23
ΔG= -29.60 kcal/mol
New-24
ΔG= -23.84 kcal/mol
New-25
ΔG= -35.30 kcal/mol
New-26
ΔG= -32.90 kcal/mol
New-27
ΔG= -46.50 kcal/mol
New-28
ΔG= -34.60 kcal/mol

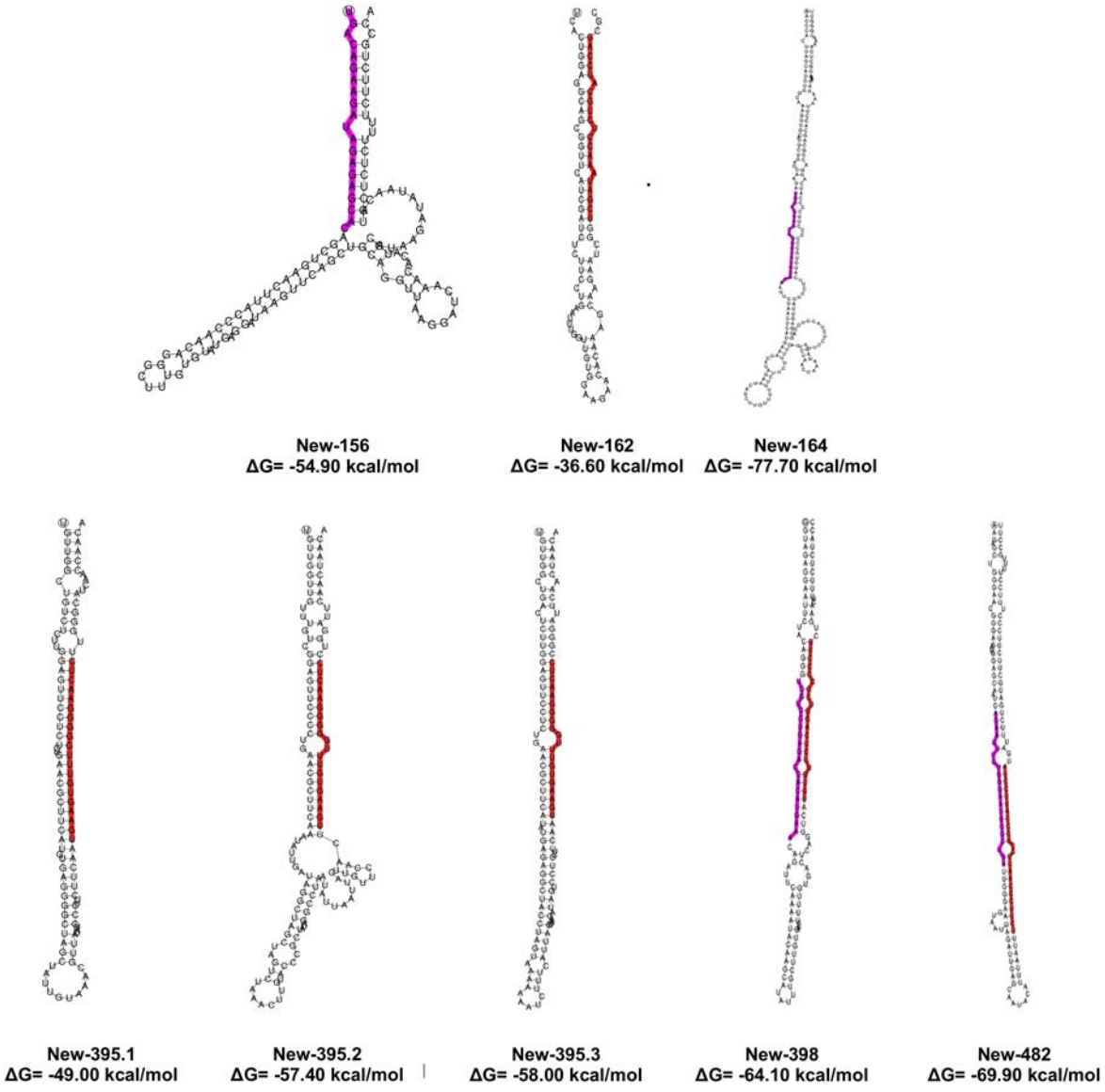
Figure 3


A
Seed
Germination
Roots
Leaves
B
Log2 fold (Germ/Seed)
MIR408
MIR393
MIR2111
MIR4397
MIR5041
MIR164
New-27
MIR862
microRNAs
C
408 3p
Glyma04g42120.1
393 3p
Glyma10g39540.1
2111 5p
Glyma16g06160.1
4397 3p1
Glyma06g05770.1
5041 3p
Glyma09g37290.1
164 5p
Glyma04g33270.1
Glyma05g00930.1
Glyma06g21020.1
Glyma06g35660.1
Glyma08g18470.1
Glyma12g26190.1
Glyma12g35530.1
Glyma13g34950.1
Glyma15g40510.1
Glyma17g10970.1
New-27 5p
Glyma09g08300.1
862 3p
Glyma14g06220.1

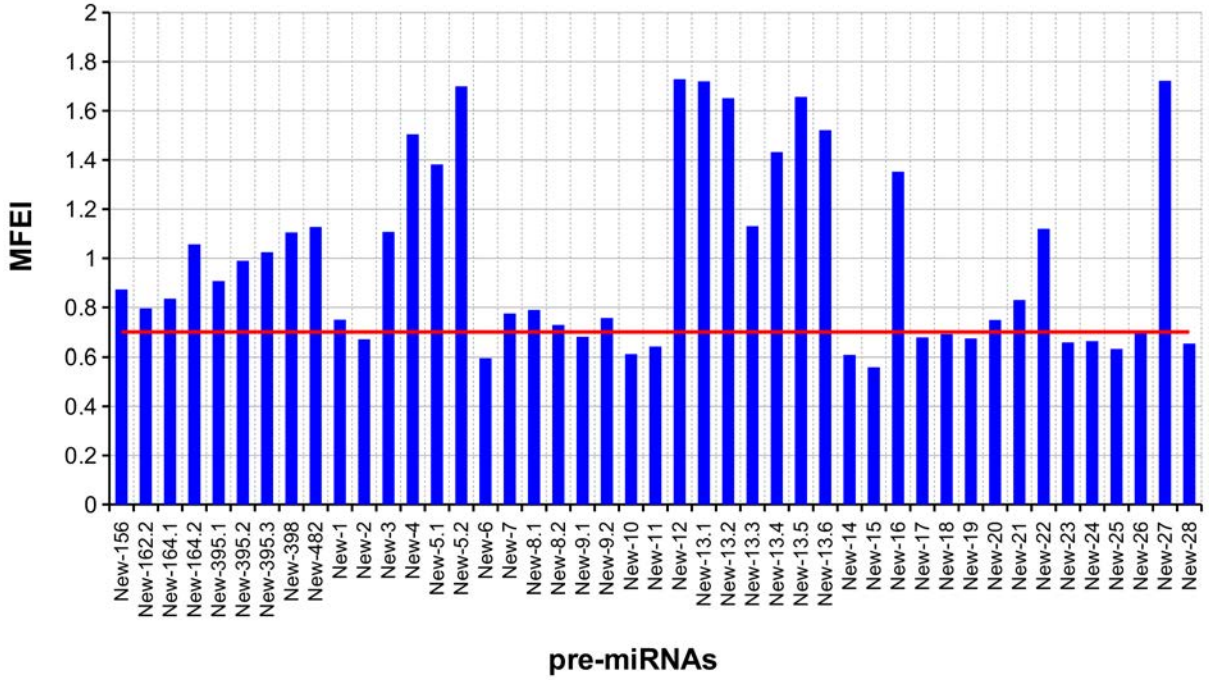
Supplementary Figure 1



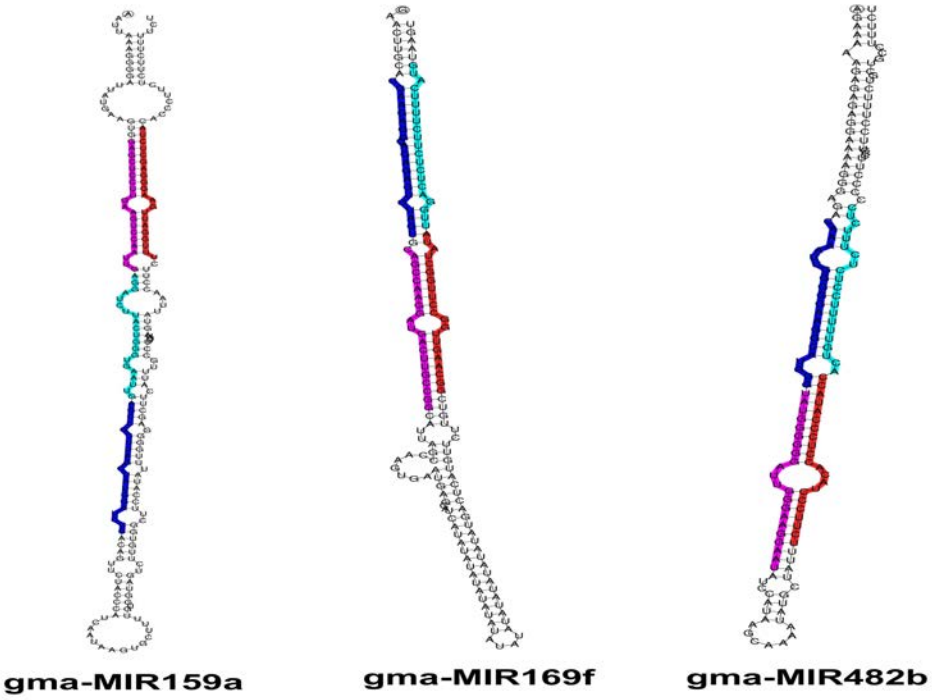
Supplementary Figure 2



Supplementary Figure 3



Supplementary Figure 4



Supplementary Figure 5

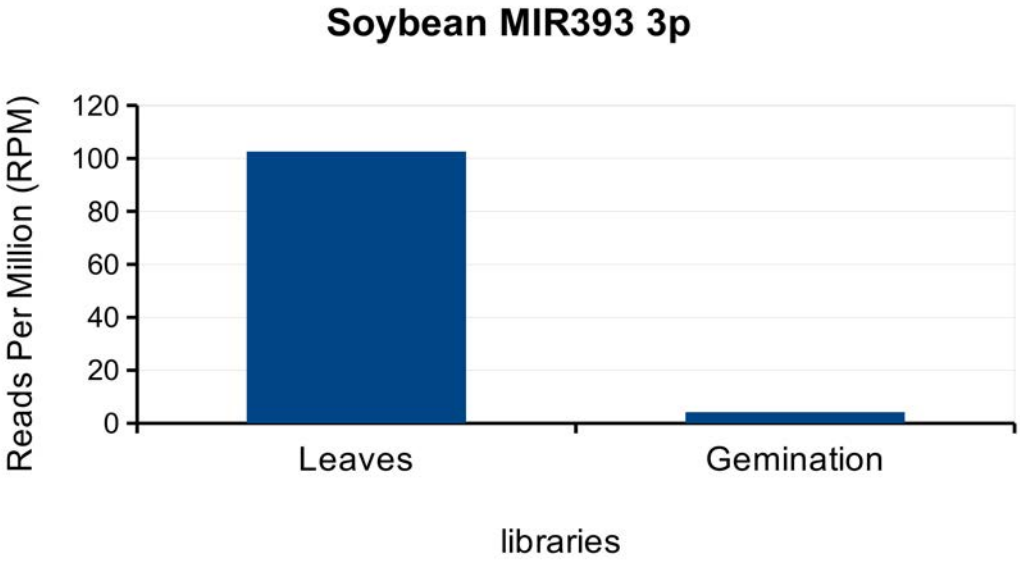


Table 1

Table1 : Number of miRNAs, isomiRNAs and moRNAs identified in both germinating and seed reads from deep sequencing libraries

Classes	Sizes						
	19	**20**	**21**	**22**	**23**	**24**	**Total**
Known miRNAs	5	31	74	24	3	5	**142**
New miRNAs (5p/3p) from known precursors	2	5	39	8	4	0	**58**
IsomiRNAs from known precursors	38	57	86	61	5	2	**249**
New miRNAs from new precursors	1	5	12	12	0	6	**36**
IsomiRNAs from new precursors	5	5	12	10	5	4	**41**
moRNAs	0	0	5	1	0	0	**6**
Total	**51**	**103**	**228**	**116**	**17**	**17**	

Supplementary table 1

Supplementary Table 1: Identified miRNAs, isomiRNAs and moRNAs identified in mature seed and germinating seeds deep sequencing library. In red, known mature miRNAs; in green, New mature miRNAs of known precursors and in blue unpublished mature miRNAs. The moRNAs are shown by underlined sequences

Loci	microRNAs acronyms	Pre-miRNAs	Position	Mature Reads	Size	Matures seed library Counts	Germination library Counts
		CACACCAGAUUG	5p	UGACAGAAGAGAGUGAGCAC	20	3524	5745
			5p1	UGACAGAAGAGAGUGAGCA	19	108	188
			5p2	UGACAGAAGAGAGUGAGCACA	21	30	39
	gma-miR156a		5p3	ACAGAAGAGAGUGAGCACA	19	5	27
			5p4	GACAGAAGAGAGUGAGCAC	19	9	16
			3p1	GCUCACUUCUCUAUCUGUCAGC	22	27	94
		UGAUGUGAGAUA	5p	UGACAGAAGAGAGAGCACA	21	20	32
			5p1	UUGACAGAAGAGAGAGCAC	21	35891	35388
			5p2	UGACAGAAGAGAGAGCAC	20	1795	3236
			5p3	UUGACAGAAGAGAGAGCA	20	1057	688
			5p4	UGACAGAAGAGAGAGCA	19	95	169
	gma-miR156b		5p5	UUGACAGAAGAGAGAGCACA	22	83	70
			5p6	GACAGAAGAGAGAGCAC	19	17	28
			5p7	UUGACAGAAGAGAGAGC	19	27	16
			3p1	GCCCUCUCUUCCUCUGUCAUC	21	54	40
		ACUUGACCACUA	3p	UUGACAGAAGAUAGAGAGCAC	21	3524	5745
			3p1	UGACAGAAGAUAGAGAGCAC	20	218	349
			3p2	UUGACAGAAGAUAGAGAGCA	20	139	77
	gma-MIR156c		3p3	UGACAGAAGAUAGAGAGCACA	21	40	29
			3p4	UUGACAGAAGAUAGAGAGCACA	22	43	20
		CUACUUGGUAAU	5p	UUGACAGAAGAUAGAGAGCAC	21	4004	5113
			5p1	UGACAGAAGAUAGAGAGCAC	20	218	349
			5p2	UUGACAGAAGAUAGAGAGCA	20	139	77
			5p3	UGACAGAAGAUAGAGAGCACA	21	40	29
	gma-miR156d		5p4	ACAGAAGAUAGAGAGCACAG	20	1	27
			5p5	UUGACAGAAGAUAGAGAGCACA	22	43	20
			5p6	UGACAGAAGAUAGAGAGCA	19	43	16
			3p1	GCUCUCUAUACUUCUGUCAUC	21	9	28
			3p2	GCUCUCUAUACUUCUGUCAUCA	22	8	27
		AGGAGGUGUUGG	5p	UUGACAGAAGAUAGAGAGCAC	21	4004	5113
			5p1	UGACAGAAGAUAGAGAGCAC	20	218	349
			5p2	UUGACAGAAGAUAGAGAGCA	20	139	77
			5p3	UUGACAGAAGAUAGAGAGCACU	22	17	17
	gma-miR156e		5p4	UGACAGAAGAUAGAGAGCA	19	42	16
			3p1	GCUCUCUAGUCUUCUGUCAUC	21	34	28
			3p2	GCUCUCUAGUCUUCUGUCAUCA	22	21	17
		AUAUCUCAUGUU	5p	UUGACAGAAGAGAGAGCACA	22	83	70
			5p1	UUGACAGAAGAGAGAGCAC	21	35891	35388
			5p2	UGACAGAAGAGAGAGCAC	20	1795	3236
			5p3	UUGACAGAAGAGAGAGCA	20	1057	688
			5p4	UGACAGAAGAGAGAGCA	19	95	169
	gma-miR156f		5p5	UGACAGAAGAGAGAGCACA	21	20	32
			5p6	GACAGAAGAGAGAGCAC	19	17	28
			5p7	UUGACAGAAGAGAGAGC	19	27	16
			3p1	GCUCUCUCUUCCUCUGUCAUC	21	207	382
			3p2	GCUCUCUCUUCCUCUGUCAUCA	22	8	17
		UGAACAAUAUCU	5p	ACAGAAGAUAGAGAGCACAG	20	1	27
			5p1	UUGACAGAAGAUAGAGAGCAC	21	4004	5113
			5p2	UGACAGAAGAUAGAGAGCAC	20	218	349
			5p3	UUGACAGAAGAUAGAGAGCA	20	139	77
	gma-MIR156g		5p4	UGACAGAAGAUAGAGAGCACA	21	40	29
			5p5	UUGACAGAAGAUAGAGAGCACA	22	43	20
			5p6	UGACAGAAGAUAGAGAGCA	19	42	16
			5p7	GACAGAAGAUAGAGAGCAC	19	5	7
		GAAAUUGACAGA	5p	UGACAGAAGAGAGUGAGCAC	20	3524	5745
			5p1	UUGACAGAAGAGAGUGAGCAC	21	217	327
			5p2	UGACAGAAGAGAGUGAGCA	19	108	188
			5p3	UGACAGAAGAGAGUGAGCACA	21	30	39
			5p4	ACAGAAGAGAGUGAGCACA	19	5	27
	gma-MIR156h		5p5	GACAGAAGAGAGUGAGCAC	19	9	16
			5p6	ACAGAAGAGAGUGAGCACACA	21	10	11
			3p1	GCUCACUUCUCUUUCUGUCAACU	23	43	115
			3p2	GCUCACUUCUCUUUCUGUCAAC	22	30	91
			3p3	GUGCUCACUUCUCUUUCUGUCA	22	6	15
			3p4	UGCUCACUUCUCUUUCUGUCAAC	23	7	12
		GUGAACUUAUUC	3p	UGACAGAAGAGAGAGCAC	20	1795	3236
	gma-MIR156h		3p1	UGACAGAAGAGAGAGCA	19	95	165
		UGACAGAAGAGA	5p	UGACAGAAGAGAGAGCAC	20	1795	3236
			5p1	UGACAGAAGAGAGAGCA	19	95	165
	gma-MIR156i		5p2	UGACAGAAGAGAGAGCACA	21	20	32
			5p3	GACAGAAGAGAGAGCAC	19	17	28
		GGUGAUGCUGUU	5p	UUGACAGAAGAUAGAGAGCAC	21	1795	3236
			5p1	UGACAGAAGAGAGAGAGCA	19	95	165
	gma-MIR156j		5p2	UGACAGAAGAGAGAGAGCACA	21	20	32
			5p3	GACAGAAGAGAGAGAGCAC	19	17	28

		GAAAUUGACAGA	5p	UGACAGAAGAGAGUGAGCAC	20	3524	5745
			5p1	UUGACAGAAGAGAGUGAGCAC	21	217	327
			5p2	UGACAGAAGAGAGUGAGCA	19	108	188
			5p3	UGACAGAAGAGAGUGAGCACA	21	30	39
			5p4	ACAGAAGAGAGUGAGCACA	19	5	27
			5p5	GACAGAAGAGAGUGAGCAC	19	9	16
	gma-MIR156k		5p6	ACAGAAGAGAGUGAGCACACA	21	10	11
			3p1	GCUCACUUCUCUUUCUGUCAACU	23	118	78
			3p2	GCUCACUUCUCUUUCUGUCAAC	22	30	91
			3p3	GUGCUCACUUCUCUUUCUGUCA	22	6	15
			3p4	UGCUCACUUCUCUUUCUGUCAAC	23	7	12
		GGAAUAUUGGGG	5p	UUGACAGAAGAUAGAGAGCAC	21	4004	5113
	gma-MIR1561		5p1	GCUCUCUAAGCUUCUGUCAUCC	22	0	10
		GUAUGUGAAGAA	5p	UUGACAGAAGAUAGAGAGCAC	21	4004	5113
			5p1	UGACAGAAGAUAGAGAGCAC	20	218	349
			5p2	UUGACAGAAGAUAGAGAGCA	20	139	77
			5p3	UGACAGAAGAUAGAGAGCACA	21	40	29
	gma-MIR156m		5p4	ACAGAAGAUAGAGAGCACAG	20	1	27
			5p5	UUGACAGAAGAUAGAGAGCACA	22	43	20
			5p6	UGACAGAAGAUAGAGAGCA	19	42	16
		UGUAUGUGUACU	5p	UGACAGAAGAGAGUGAGCAC	20	3524	5745
			5p1	UGACAGAAGAGAGUGAGCA	19	108	188
			5p2	UUGACAGAAGAGAGUGAGCAC	21	217	327
			5p3	UGACAGAAGAGAGUGAGCACA	21	30	39
	gma-MIR156n		5p4	ACAGAAGAGAGUGAGCACA	19	5	27
			5p5	GACAGAAGAGAGUGAGCAC	19	9	16
			3p1	GCUCACCACUCUUUCUGUCGGUU	23	20	45
			3p2	GCUCACCACUCUUUCUGUCGGU	22	10	26
		GUAUGUGUACUU	5p	UGACAGAAGAGAGUGAGCAC	20	3524	5745
			5p1	UUGACAGAAGAGAGUGAGCAC	21	217	327
			5p2	UGACAGAAGAGAGUGAGCA	19	108	188
			5p3	UGACAGAAGAGAGUGAGCACA	21	30	39
	gma-MIR156o		5p4	ACAGAAGAGAGUGAGCACA	19	5	27
			5p5	GACAGAAGAGAGUGAGCAC	19	9	16
			3p1	GCUCACUACUCUUUCUGUCGGUU	23	21	134
			3p2	GCUCACUACUCUUUCUGUCGGU	22	5	45
		GAGAUGACAGAA	5p	UGACAGAAGAGAGUGAGCAC	20	3524	5745
			5p1	UGACAGAAGAGAGUGAGCA	19	108	188
			5p2	UGACAGAAGAGAGUGAGCACA	21	30	39
	gma-MIR156w		5p3	ACAGAAGAGAGUGAGCACA	19	5	27
			5p4	GACAGAAGAGAGUGAGCAC	19	9	16
			5p5	ACAGAAGAGAGUGAGCACACA	21	10	11
			3p1	GCUUACUCUCUAUCUGUCAUC	21	118	78
		GGUGACAGAAGA	5p	UGACAGAAGAGAGUGAGCAC	20	3524	5745
			5p2	UGACAGAAGAGAGUGAGCA	19	108	188
			5p3	UGACAGAAGAGAGUGAGCACA	21	30	39
	gma-MIR156x		5p4	ACAGAAGAGAGUGAGCACA	19	5	27
			5p5	GACAGAAGAGAGUGAGCAC	19	9	16
			5p6	ACAGAAGAGAGUGAGCACACA	21	10	11
		UGACAGAAGAUA	5p	UGACAGAAGAUAGAGAGCAC	20	3524	5745
			5p1	UGACAGAAGAUAGAGAGCACA	21	40	29
Gm19:28,821,820..28,821,980	New-156		5p2	ACAGAAGAUAGAGAGCACAG	20	1	27
			5p3	UGACAGAAGAUAGAGAGCA	19	42	16
		AAUUAAAGGGGA	5p	GAGCUCCUUGAAGUCCAAUUG	21	0	58
			5p1	GAGCUCCUUGAAGUCCAAUU	20	73	248
			5p2	GAGCUCCUUGAAGUCCAAUUGA	22	6	52
			5p3	AGCUGCUUAGCUAUGGAUCCC	21	4	23
			5p4	AGGAUCUUACUGGGUGAAUUG	21	0	5
			3p	UUUGGAUUGAAGGGAGCUCUA	21	210734	154700
	gma-MIR159a		3p1	UUGGAUUGAAGGGAGCUCUA	20	11221	10673
			3p2	UUUGGAUUGAAGGGAGCUCU	20	2898	2283
			3p3	UUUGGAUUGAAGGGAGCUCUAC	22	380	368
			3p4	UUUGGAUUGAAGGGAGCUC	19	189	244
			3p5	UUGGAUUGAAGGGAGCUCU	19	75	85
			3p6	UUGGAUUGAAGGGAGCUCUAC	21	31	33
			3p7	UGGAUUGAAGGGAGCUCUA	19	43	29
		CAUGCAUACAUA	5p	UGCCUGGCUCCCUGUAUGCCA	21	74	60
	gma-MIR160		3p1	GCGUAUGAGGAGCCAAGCAUA	21	8	18
		AUGUGUAUGUGC	5p	UGCCUGGCUCCCUGUAUGCC	20	0	0
			5p1	UGCCUGGCUCCCUGUAUGCCA	21	74	60
	gma-MIR160b		3p1	GCGUAUGAGGAGCCAAGCAUA	21	8	18
		UGCCUGGCUCCC	5p	UGCCUGGCUCCCUGUAUGCC	20	0	0
	gma-MIR160c		5p1	UGCCUGGCUCCCUGUAUGCCA	21	74	60
		UGCCUGGCUCCC	5p	UGCCUGGCUCCCUGUAUGCC	20	0	0
	gma-MIR160d		5p1	UGCCUGGCUCCCUGUAUGCCA	21	74	60
		UACUUGGCGUGC	5p	UGCCUGGCUCCCUGUAUGCC	20	0	0
	gma-MIR160e		5p1	UGCCUGGCUCCCUGUAUGCCA	21	74	60
		GUGAAGUCACUG	3p	UCGAUAAACCUCUGCAUCCA	20	3	1
			3p1	UCGAUAAACCUCUGCAUCCAG	21	1498	724

		3p6	UCGGACCAGGCUUCAUUCCCCC	22	56	42
	AGUUGAGGGGAA	5p1	GGAAUGUCGUUUGGUUCGAGA	21	6	69
		3p	UCGGACCAGGCUUCAUUCCCC	21	63967	58528
		3p1	UUCGGACCAGGCUUCAUUCCCC	22	999	866
		3p2	UCGGACCAGGCUUCAUUCCC	20	559	485
		3p3	UUCGGACCAGGCUUCAUUCCC	21	243	174
gma-MIR166g		3p4	UCGGACCAGGCUUCAUUCC	19	60	96
		3p5	UUUCGGACCAGGCUUCAUUCC	21	46	69
		3p6	GGACCAGGCUUCAUUCCCC	19	51	49
		3p7	UCGGACCAGGCUUCAUUCCCCU	22	9	19
		3p8	UUCGGACCAGGCUUCAUUCC	20	5	10
	GGAAAUGGGAGA	5p	GGAAUGUUGUUUGGCUCGAGG	21	0	147
		3p	UCUCGGACCAGGCUUCAUUCC	21	4872	5149
		3p1	UCGGACCAGGCUUCAUUCCCG	21	128	1654
		3p2	UCGGACCAGGCUUCAUUCCC	20	559	485
gma-MIR166h		3p3	UCUCGGACCAGGCUUCAUUC	20	111	160
		3p4	UCGGACCAGGCUUCAUUCC	19	60	96
		3p5	CUCGGACCAGGCUUCAUUCC	20	37	26
	GUUGAGGGGAAU	5p	GGAAUGUUGUUUGGCUCGAGG	21	0	147
		3p	UCUCGGACCAGGCUUCAUUCC	21	4872	5149
		3p1	UCGGACCAGGCUUCAUUCCCG	21	128	1654
		3p2	UCGGACCAGGCUUCAUUCCC	20	559	485
gma-MIR166i		3p3	UCUCGGACCAGGCUUCAUUC	20	111	160
		3p4	UCGGACCAGGCUUCAUUCC	19	60	96
		3p5	CUCGGACCAGGCUUCAUUCC	20	37	26
	GGUUGAUGGGAA	5p	GGAAUGUUGUUUGGCUCGAGG	21	0	147
		3p	UCUCGGACCAGGCUUCAUUCC	21	4872	5149
		3p1	UCGGACCAGGCUUCAUUCCCG	21	128	1654
		3p2	UCGGACCAGGCUUCAUUCCC	20	559	485
gma-MIR166j		3p3	UCUCGGACCAGGCUUCAUUC	20	111	160
		3p4	UCGGACCAGGCUUCAUUCC	19	60	96
		3p5	CUCGGACCAGGCUUCAUUCC	20	37	26
	GGGGAAUGUUGC	5p1	GGAAUGUUGGCUGGCUCGAGG	21	34	97
		5p2	GGAAUGUUGGCUGGCUCGAGGC	22	19	60
		3p	UCGGACCAGGCUUCAUUCCCC	21	63697	58528
		3p1	UCGGACCAGGCUUCAUUCCC	20	559	485
gma-MIR166o		3p2	UUCGGACCAGGCUUCAUUCCC	21	243	174
		3p3	UCGGACCAGGCUUCAUUCC	19	60	96
		3p4	UUCGGACCAGGCUUCAUUCC	20	5	10
	UGGGAAUGUUGU	5p1	GGAAUGUUGUUUGGCUCGAGG	21	0	147
		3p	UCGGACCAGGCUUCAUUCCCG	21	128	1654
		3p1	UCUCGGACCAGGCUUCAUUCC	21	4872	5149
		3p2	UCGGACCAGGCUUCAUUCCC	20	559	485
gma-MIR166q		3p3	UCUCGGACCAGGCUUCAUUC	20	111	160
		3p4	UCGGACCAGGCUUCAUUCC	19	60	96
		3p5	CUCGGACCAGGCUUCAUUCC	20	37	26
	GAAGUUCGCAAA	3p	UGAAGCUGCCAGCAUGAUCUA	21	1595	1470
gma-MIR167a		3p1	UGAAGCUGCCAGCAUGAUCU	20	110	100
		3p2	UGAAGCUGCCAGCAUGAUC	19	14	15
	AAGGGUCACAAA	5p	UGAAGCUGCCAGCAUGAUCUA	21	1595	1470
		5p1	UGAAGCUGCCAGCAUGAUCU	20	110	100
gma-MIR167b		5p2	UGAAGCUGCCAGCAUGAUC	19	14	15
		3p1	GUCAUGCUGUGCUAGCCUCACU	22	6	20
	UUUGAGAGGUUG	5p	UGAAGCUGCCAGCAUGAUCUG	21	2473	2095
		5p1	UGAAGCUGCCAGCAUGAUCUGG	22	8765	6980
gma-MIR167c		5p3	UGAAGCUGCCAGCAUGAUCU	20	110	100
		5p4	UGAAGCUGCCAGCAUGAUC	19	14	15
		3p1	UCAGGUCAUCUUGCAGCUUCA	21	56	53
	AAGGGUCACAAG	5p	UGAAGCUGCCAGCAUGAUCUA	21	1595	1470
		5p1	UGAAGCUGCCAGCAUGAUCU	20	110	100
gma-MIR167d		5p2	UGAAGCUGCCAGCAUGAUC	19	14	15
		3p1	GUCAUGCUGUGCUAGCCUCACU	22	6	20
	UCAUGCACCACU	5p	UGAAGCUGCCAGCAUGAUCUU	21	16360	11044
		5p1	UGAAGCUGCCAGCAUGAUCUUA	22	39552	26125
		5p2	UUGAAGCUGCCAGCAUGAUCUU	22	161	114
		5p3	UGAAGCUGCCAGCAUGAUCU	20	110	100
gma-miR167e		5p4	AGCUGCCAGCAUGAUCUUA	19	118	90
		5p5	UGAAGCUGCCAGCAUGAUC	19	14	15
		5p6	GAAGCUGCCAGCAUGAUCUU	20	21	14
		5p7	UGAAGCUGCCAGCAUGAUCUUAA	23	11	10
		3p1	GAUCAUGUGGCAGUUUCACC	20	11	10
	UCAUGCACCACU	5p	UGAAGCUGCCAGCAUGAUCUU	21	16360	11044
		5p1	UGAAGCUGCCAGCAUGAUCUUA	22	39552	26125
		5p2	UUGAAGCUGCCAGCAUGAUCUU	22	161	114
		5p3	UGAAGCUGCCAGCAUGAUCU	20	110	100
gma-miR167f		5p4	AGCUGCCAGCAUGAUCUUA	19	118	90
		5p5	GAAGCUGCCAGCAUGAUCUU	20	21	14
		5p6	UGAAGCUGCCAGCAUGAUC	19	14	15
		5p8	UGAAGCUGCCAGCAUGAUCUUAA	23	11	10
		3p1	GAUCAUGUGGCAGUUUCACC	20	11	10

gma-miR167g	CAGCAGUUGAAG	5p	UGAAGCUGCCAGCAUGAUCUGA	22	182	208
		5p1	UGAAGCUGCCAGCAUGAUCUG	21	2473	2095
		5p2	UGAAGCUGCCAGCAUGAUCU	20	110	100
		5p3	UGAAGCUGCCAGCAUGAUC	19	14	15
		5p4	AGCUGCCAGCAUGAUCUGAGUU	22	7	21
		3p1	GAUCAUGUGGCUGCUUCACC	20	5	14
gma-MIR167j	CAGCAGUUGAAG	5p	UGAAGCUGCCAGCAUGAUCUG	21	2473	2095
		5p1	UGAAGCUGCCAGCAUGAUCUGA	22	182	208
		5p2	UGAAGCUGCCAGCAUGAUCU	20	110	100
		5p3	UGAAGCUGCCAGCAUGAUC	19	14	15
		5p4	AGCUGCCAGCAUGAUCUGAGUU	22	7	21
		3p1	GAUCAUGUGGCUGCUUCACC	20	5	14
gma-MIR167o	CAAGAUGUUGUU	3p	UGAAGCUGCCAGCAUGAUCUG	21	2473	2095
		3p1	UGAAGCUGCCAGCAUGAUCU	20	110	100
		3p2	UGAAGCUGCCAGCAUGAUC	19	14	15
gma-miR168	CACUGUGCGGUC	5p	UCGCUUGGUGCAGGUCGGGAA	21	17595	12305
		5p1	CGCUUGGUGCAGGUCGGGAA	20	14	49
		5p2	UCGCUUGGUGCAGGUCGGGAAC	22	35	35
		5p3	UCGCUUGGUGCAGGUCGGGA	20	30	24
		5p4	CGCUUGGUGCAGGUCGGGAAC	21	2	11
		3p1	CCCGCCUUGCAUCAACUGAAU	21	72	69
		3p2	UGGAUCCCGCCUUGCAUCAAC	21	11	10
gma-MIR168b	CGGUCUCUAAUU	5p	UCGCUUGGUGCAGGUCGGG	19	2	2
		5p1	UCGCUUGGUGCAGGUCGGGAA	21	17595	12305
		5p2	UCGCUUGGUGCAGGUCGGGAAC	22	35	35
		5p3	UCGCUUGGUGCAGGUCGGGA	20	30	24
		3p1	CCCGCCUUGCAUCAACUGAAU	21	72	69
gma-miR169a	AAGAGGAAGAGA	5p	CAGCCAAGGAUGACUUGCCGG	21	56824	56827
		5p1	AGCCAAGGAUGACUUGCCGG	20	3520	6378
		5p2	CAGCCAAGGAUGACUUGCCGGC	22	19	19
		5p3	AGCCAAGGAUGACUUGCCGGC	21	3	13
		3p1	GGCAAGUUGUGUUUGGCUAUG	21	0	10
gma-MIR169b	UAUGAUGCUGCA	5p	CAGCCAAGGAUGACUUGCCGA	21	50	48
		5p1	CAGCCAAGGAUGACUUGCC	19	6	8
		5p2	AGCCAAGGAUGACUUGCCGA	20	4	6
gma-MIR169c	GUUUGAGGCCAU	5p	AGCCAAGGAUGACUUGCCGG	20	3520	6378
		5p1	AAGCCAAGGAUGACUUGCCGG	21	10	26
		5p2	AGCCAAGGAUGACUUGCCGGC	21	3	13
gma-MIR169e	GUCUUGCAUGAA	5p	AGCCAAGGAUGACUUGCCGG	20	3520	6378
		5p1	GAGCCAAGGAUGACUUGCCGG	21	4	3
		5p2	AGCCAAGGAUGACUUGCCGGC	21	3	13
gma-MIR169f	GAACUUGCACGA	5p	CAGCCAAGGAUGACUUGCCGG	21	56824	56827
		5p1	AGCCAAGGAUGACUUGCCGG	20	3520	6378
		5p2	GCCAAGGAUGACUUGCCGG	19	206	205
		5p3	CGAAGAGGCAGAGAGUGUAGU	21	0	9
		3p1	GGCAAGUUGGCCUUGGCUAUA	21	2	22
		3p2	UUGGACUCUCUUCUUUUCAUG	21	0	15
gma-MIR169g	AAGUAGAGUGCA	5p	CAGCCAAGGAUGACUUGCCGG	21	56824	56827
		5p1	CAGCCAAGGAUGACUUGCCGGA	22	204	173
		5p2	GCCAAGGAUGACUUGCCGG	19	206	205
		5p3	CAGCCAAGGAUGACUUGCCGGA	22	204	173
		5p4	AGCCAAGGAUGACUUGCCGGA	21	54	65
		3p1	GGCAAGUUGUUCUUGGCUA	19	13	21
		3p2	GGCAAGUUGUUCUUGGCUACA	21	4	9
gma-MIR169h	GUGCUUUUGAGC	5p	GGCGAGACAUCUUGGCUCAUU	21	0	0
		5p1	AGCCAAGGAUGACUUGCCGG	20	3520	6378
		5p2	GCCAAGGAUGACUUGCCGG	19	206	205
gma-MIR169j	CCUCCCCCCAUU	5p	AGCCAAGGAUGACUUGCCGG	20	3520	6378
		5p1	UAGCCAAGAAUGACUUGCCGG	21	5	94
gma-MIR169k	GGUUUUAGAGUG	5p	CAGCCAAGAAUGACUUGCCGG	21	5	56
gma-MIR1691	GAGUGUUUGCAA	5p	CAGCCAAGAAUGACUUGCCGG	21	5	56
gma-MIR169m	GGUCCUAUCAAC	5p	CAGCCAAGGAUGACUUGCCGG	21	56841	56827
		5p1	AGCCAAGGAUGACUUGCCGG	20	3520	6378
		5p2	CAGCCAAGGAUGACUUGCCGGA	22	204	173
		5p3	AGCCAAGGAUGACUUGCCGGA	21	54	65
		3p1	GGCAAGUUGUUCUUGGCUA	19	13	21
gma-MIR169p	CAGCCAAGGAUG	5p	CAGCCAAGGAUGACUUGCCGG	21	56841	56827
		5p1	AGCCAAGGAUGACUUGCCGG	20	3520	6378
		5p2	CAGCCAAGGAUGACUUGCCGGC	22	19	19
		5p3	AGCCAAGGAUGACUUGCCGGC	21	3	13
		3p1	UGUCGGCAAGUUGGCCUUGGC	21	6	12
New-169s	GAUGAAGCCAAG	5p	AGCCAAGGAUGACUUGCCGG	20	3520	6378
		5p2	AGCCAAGGAUGACUUGCCGGA	21	54	65
		5p3	AAGCCAAGGAUGACUUGCCGG	21	10	26
	GUUCAACGGGAU	5p1	GGAUAUUGGUCCGGUUCAAUA	21	5	51

gma-MIR171a		5p2	GGGAUAUUGGUCCGGUUCAAUA	22	0	39
		3p	UGAGCCGUGCCAAUAUCACGA	21	0	1
		3p1	UUGAGCCGUGCCAAUAUCACGA	22	197	494
		3p2	UUGAGCCGUGCCAAUAUCACG	21	14	33
gma-MIR171b	UAGACACGGCGU	5p	ACGGCGUGAUAUUGGUACGGCUC	23	0	0
		5p1	CGUGAUAUUGGUACGGCUCAUC	22	13	16
		3p	CGAGCCGAAUCAAUAUCACUC	21	0	0
gma-miR171c	UUAAGAACUUGA	5p	AGAUAUUGGUGCGGUUCAAUC	21	0	0
		3p1	UGAUUGAGCCGUGCCAAUAUC	21	88	80
		3p2	UUGAGCCGUGCCAAUAUCACA	21	5	19
gma-MIR171e	GUUUAUAGUAAU	3p	UGAUUGAGCCGUGCCAAUAUC	21	88	80
gma-MIR171f	ACUUGUUGAUGG	3p	UGAUUGAGCCGUGCCAAUAUC	21	88	80
		3p2	UUGAGCCGUGCCAAUAUCACA	21	5	19
gma-MIR171g	UGAUGUUGGCUU	3p	UGAUUGAGCCGUGCCAAUAUC	21	88	80
gma-MIR171h	GCAAAAGUAGAC	5p1	CGUGAUAUUGGUACGGCUCAUC	22	13	16
		3p	AUUGAGACGAGCCGAAUCAAU	21	0	0
gma-MIR171i	UUAAAGGCAAGC	5p	AUAAGAAAGCAAUGCUCAAA	20	0	0
		5p1	GGAUAUUGGUCCGGUUCAAUA	21	5	51
		5p2	GGGAUAUUGGUCCGGUUCAAUA	22	0	39
		3p	UUGAGCCGUGCCAAUAUCACGA	22	197	494
		3p1	UUGAGCCGUGCCAAUAUCACG	21	14	33
gma-MIR171j	CAAGUAUAUGAA	5p	UAUUGGCCUGGUUCACUCAGA	21	55	73
		3p1	UGAUUGAGCCGUGCCAAUAUC	21	88	80
gma-MIR171k	GAGAAAGCGAUG	5p	CGAUGUUGGUGAGGUUCAAUC	21	1	0
		3p1	UUGAGCCGCGCCAAUAUCACU	21	35	24
gma-MIR171l	GAGAAAGCGAUG	5p	CGAUGUUGGUGAGGUUCAAUC	21	1	0
		3p1	UUGAGCCGCGCCAAUAUCACU	21	35	24
New-171p	UGACAAAUUAAG	3p1	UUGAGCCGCGUCAAUAUCUUA	21	81	57
gma-MIR172a	UUAACAGUCGUU	5p1	GUAGCAUCAUCAAGAUUCACA	21	1	15
		3p	AGAAUCUUGAUGAUGCUGCAU	21	1	11
gma-MIR172b	UUGACAGUCGUU	5p	GUAGCAUCAUCAAGAUUCAC	20	0	2
		5p1	GUAGCAUCAUCAAGAUUCACA	21	1	15
		3p	AGAAUCUUGAUGAUGCUGCAU	21	1	11
gma-MIR172c	AAAUCAGUCACU	5p1	GGAGCAUCAUCAAGAUUCACA	21	23	3
		5p2	GUAGCAUCAUCAAGAUUCAC	20	0	2
		3p	GGAAUCUUGAUGAUGCUGCAG	21	0	0
gma-MIR172d	AAAACAGUCGCU	5p1	GGAGCAUCAUCAAGAUUCACA	21	23	3
		3p	GGAAUCUUGAUGAUGCUGCAGCAG	24	0	0
gma-MIR172e	AAAACAGUCACU	5p1	GGAGCAUCAUCAAGAUUCACA	21	23	3
		3p	GGAAUCUUGAUGAUGCUGCAGCAG	24	0	0
gma-MIR172f	GCGGAUGUAGCA	5p1	GUAGCAUCAUCAAGAUUCACA	21	1	15
		3p	AGAAUCUUGAUGAUGCUGCA	20	0	2
		3p1	AGAAUCUUGAUGAUGCUGCAU	21	1	11
gma-MIR172g	GCAGGUGCAGCA	5p	GCAGCACCAUCAAGAUUCAC	20	0	2
		3p1	AGAAUCUUGAUGAUGCUGCAU	21	1	11
gma-MIR172h	GCAGGUGCAGCA	5p	GCAGCAGCAUCAAGAUUCACA	21	1	15
		3p	AGAAUCUUGAUGAUGCUGCAU	21	1	11
gma-MIR319a	CGUUGAAGACCC	5p1	AGAGCUUUCUUCAGUCCACU	20	8	23
		3p	UUGGACUGAAGGGAGCUCCC	20	231	220
		3p1	UUGGACUGAAGGGAGCUCCCU	21	2923	2351
		3p2	UUGGACUGAAGGGAGCUCCCUU	22	717	649
		3p3	UUGGACUGAAGGGAGCUCCC	20	231	220
		3p4	UGGACUGAAGGGAGCUCCCU	20	146	162
		3p5	UGGACUGAAGGGAGCUCCCUU	21	108	87
		3p6	UGGACUGAAGGGAGCUCCC	19	21	22
		3p7	CUUGGACUGAAGGGAGCUCCCUU	23	11	20
		3p8	CUUGGACUGAAGGGAGCUCCCU	22	8	19
gma-MIR319b	UAAGUCCUAAGC	5p1	AGAGCUUUCUUCAGUCCACU	20	8	23
		3p	UUGGACUGAAGGGAGCUCCC	20	231	220
		3p1	UUGGACUGAAGGGAGCUCCCU	21	2923	2351
		3p2	UUGGACUGAAGGGAGCUCCCUU	22	717	649
		3p3	UUGGACUGAAGGGAGCUCCC	20	231	220
		3p4	UGGACUGAAGGGAGCUCCCU	20	146	162
		3p5	UGGACUGAAGGGAGCUCCCUU	21	108	87
		3p6	UGGACUGAAGGGAGCUCCC	19	21	22
		3p7	CUUGGACUGAAGGGAGCUCCCUU	23	11	20
		3p8	CUUGGACUGAAGGGAGCUCCCU	22	8	19
gma-MIR319d	GGAGACAAAGAG	3p	UUGGACUGAAGGGAGCUCCC	20	231	220
		3p1	UGGACUGAAGGGGAGCUCCUUC	22	21	224
	AAGAUGUAUGUA	3p	UUGGACUGAAGGGAGCUCCC	20	231	220

		3p1	UUGGACUGAAGGGAGCUCCCU	21	2923	2351
		3p2	UGGACUGAAGGGAGCUCCCU	20	146	162
gma-MIR319e		3p3	UUGGACUGAAGGGAGCUCCCUC	22	12	32
		3p4	UGGACUGAAGGGAGCUCCC	19	21	22
		3p5	CUUGGACUGAAGGGAGCUCCCU	22	8	19
	UGAGGGAUAUGG	5p1	AGCUGCUGACUCGUUGGUUCG	21	2	41
		5p2	AGCUGCUGACUCGUUGGUUC	20	0	18
		3p	UUGGACUGAAGGGGCCUCUU	20	354	756
gma-MIR319f		3p1	UUGGACUGAAGGGGCCUCUUC	21	151	361
		3p2	CUUGGACUGAAGGGGCCUCUU	21	6	47
		3p3	UGGACUGAAGGGGCCUCUU	19	14	19
	GGAGGGAAAGAG	3p	UUGGACUGAAGGGAGCUCCC	20	231	220
		3p1	UGGACUGAAGGGAGCUCCUUC	21	418	632
		3p2	UUGGACUGAAGGGAGCUCCUUC	22	8	26
gma-MIR319g MI0018672		3p3	UGGACUGAAGGGAGCUCCUU	20	13	18
		3p4	UGGACUGAAGGGAGCUCCU	19	9	14
		3p5	UUGGACUGAAGGGAGCUCCU	20	30	12
	GAUUAAUAUUGA	3p	UUGGACUGAAGGGAGCUCCC	20	231	220
		3p1	UUGGACUGAAGGGAGCUCCCU	21	2923	2351
		3p2	UGGACUGAAGGGAGCUCCCU	20	146	162
gma-MIR319h		3p3	UUGGACUGAAGGGAGCUCCCUC	22	12	32
		3p4	UGGACUGAAGGGAGCUCCC	19	21	22
		3p5	CUUGGACUGAAGGGAGCUCCCU	22	8	19
	GGAGACAAAGAG	3p	UUGGACUGAAGGGAGCUCCC	20	231	220
gma-MIR319i		3p1	UGGACUGAAGGGGAGCUCCUUC	22	21	224
	AGAGCUUCCUUC	3p	UUGGACUGAAGGGAGCUCCC	20	231	220
		3p1	UUGGACUGAAGGGAGCUCCCU	21	2923	2351
		3p2	UGGACUGAAGGGAGCUCCCU	20	146	162
gma-MIR319k		3p3	UGGACUGAAGGGAGCUCCC	19	21	22
		3p4	CUUGGACUGAAGGGAGCUCCCU	22	8	19
	UCGUCCAUUUUC	3p	UUGGACUGAAGGGAGCUCCC	20	231	220
		3p1	UUUGGACUGAAGGGAGCUCCU	21	1927	316
		3p2	UGGACUGAAGGGAGCUCCUUC	21	418	632
		3p3	UUUUGGACUGAAGGGAGCUCC	21	81	10
gma-MIR319l		3p4	UUGGACUGAAGGGAGCUCCU	20	30	12
		3p5	UGGACUGAAGGGAGCUCCUU	20	13	18
		3p6	UGGACUGAAGGGAGCUCCU	19	9	14
		3p7	UUGGACUGAAGGGAGCUCCUUC	22	8	26
	AGGAAUAGUUUA	3p	UUGGACUGAAGGGAGCUCCCU	21	2923	2351
		3p1	UUGGACUGAAGGGAGCUCCC	20	231	220
		3p2	UUGGACUGAAGGGAGCUCCCUU	22	717	649
		3p3	UGGACUGAAGGGAGCUCCCU	20	146	162
gma-MIR319m		3p4	UGGACUGAAGGGAGCUCCCUU	21	108	87
		3p5	UGGACUGAAGGGAGCUCCC	19	21	22
		3p6	CUUGGACUGAAGGGAGCUCCCUU	23	11	20
		3p7	CUUGGACUGAAGGGAGCUCCCU	22	8	19
	AGAAUCUGUAAA	5p	AAGCUCAGGAGGGAUAGCGCC	21	34	71
		5p1	AGCUCAGGAGGGAUAGCGCC	20	46	40
gma-MIR390a		3p	CGCUAUCCAUCCUGAGUUUC	20	6	8
	AGAAUCUGUAAA	5p	AAGCUCAGGAGGGAUAGCACC	21	3	27
gma-MIR390b		3p1	UACUUGGCGCUAUCUAUCUUGA	22	0	0
	UGUAAAGCUCAG	5p1	AAGCUCAGGAGGGAUAGCGCC	21	34	71
		5p2	AGCUCAGGAGGGAUAGCGCC	20	46	40
gma-MIR390c		3p	CGCUAUCCAUCCUGAGUUUC	20	6	8
	GUACAGAAGAAU	5p	AAGCUCAGGAGGGAUAGCGCC	21	34	71
gma-MIR390f		5p1	AGCUCAGGAGGGAUAGCGCC	20	46	40
	GGAGGAGGCAUC	5p	UCCAAAGGGAUCGCAUUGAUC	21	0	1
gma-MIR393a		3p1	GAUCAUGCUAUCCCUUUGGAU	21	1	11
	AUCAUGAGGGUU	5p1	UUGGCAUUCUGUCCACCUCC	20	15	10
		3p	AGCUCUGUUGGCUACACUUU	20	0	0
gma-MIR394a		3p1	AGCUCUGUUGGCUACACUUUG	21	5	3
	UAACAGAGUUUA	5p1	UUGGCAUUCUGUCCACCUCC	20	15	10
gma-MIR394b		3p	AGGUGGGCAUACUGUCAACU	20	1	2
	CAGAGUUUCUUC	5p1	UUGGCAUUCUGUCCACCUCC	20	15	10
gma-MIR394c		3p	AGGUGGGCAUACUGUCAACU	20	1	2
	UCAGGUUUUCCC	3p	CUGAAGUGUUUGGGGGAACUC	21	443	178
		3p1	UGAAGUGUUUGGGGGAACUC	20	89	35
gma-MIR395a		3p2	CUGAAGUGUUUGGGGGAACU	20	44	12
	CUAUUCAGGUUU	3p	CUGAAGUGUUUGGGGGAACUC	21	443	178
		3p1	UGAAGUGUUUGGGGGAACUC	20	89	35
gma-MIR395b		3p2	CUGAAGUGUUUGGGGGAACU	20	44	12
	CUUAUUAUGUCA	3p	CUGAAGUGUUUGGGGGAACUC	21	443	178
		3p1	UGAAGUGUUUGGGGGAACUC	20	89	35
gma-MIR395c		3p2	CUGAAGUGUUUGGGGGAACU	20	44	12

	Gma-MIR3951	UCCUUUGGAGUU	3p1	UGAAGUGUUUGGGGGAACUC	20	89	35
Gm02:1,759,304..1,759,420	New-395.1	UGUUGGCUGUCU	3p1	UGAAGUGUUUGGGGGAACUC	20	89	35
Gm02:1,766,370..1,766,510	New-395.2	UGUUGGUUGUUU	3p1	UGAAGUGUUUGGGGGAACUC	20	89	35
Gm18:16,316,064..16,316,178	New-395.3	AUCAGGUGUCAC	3p1	UCUGAAGUGUUUGGGGGAACC	21	0	1
	gma-MIR396a	UCAUGGCUCUCU	5p	UUCCACAGCUUUCUUGAACUG	21	14	50
			5p1	GUUCAAUAAAGCUGUGGGAAG	21	971	2114
			5p2	UUCCACAGCUUUCUUGAACU	20	75	220
			3p	UUCAAUAAAGCUGUGGGAAG	20	45	160
			3p2	CGGUUCAAUAAAGCUGUGGGA	21	13	39
			3p3	GUUCAAUAAAGCUGUGGGAA	20	10	17
	gma-MIR396b	CUCAAGUCCUGG	5p	UUCCACAGCUUUCUUGAACUU	21	770	2088
			5p1	UUCCACAGCUUUCUUGAACU	20	75	220
			3p	GCUCAAGAAAGCUGUGGGAGA	21	48	1719
			3p2	CUCAAGAAAGCUGUGGGAGA	20	44	159
	gma-MIR396c	CAACAAGUCCUG	5p	UUCCACAGCUUUCUUGAACUU	21	770	2088
			5p1	UUCCACAGCUUUCUUGAACU	20	75	220
	gma-MIR396d	GGUCAUGCUUUU	5p1	UUCCACAGCUUUCUUGAACUU	21	770	2088
			5p2	UUCCACAGCUUUCUUGAACU	20	75	220
			3p	AAGAAAGCUGUGGGAGAAUAUGGC	24	1	2
			3p1	GCUCAAGAAAGCUGUGGGAGA	21	48	1719
			3p2	CUCAAGAAAGCUGUGGGAGA	20	44	159
	gma-MIR396e	GUGAUCUUCCAC	5p	UUCCACAGCUUUCUUGAACUGU	22	14	50
			5p2	UUCCACAGCUUUCUUGAACU	20	75	220
			5p3	UUCCACAGCUUUCUUGAACUG	21	14	50
			3p1	AUUCAAGAUAGCUGUGGAAAA	21	14	37
	gma-MIR396f	UAGCUUCUUCAG	5p	AGCUUUCUUGAACUUCUUAUGCCUU	25	0	0
			5p1	UUCCACAGCUUUCUUGAACUU	21	770	2088
			5p2	UUCCACAGCUUUCUUGAACU	20	75	220
	gma-MIR396g	AUGCUGUGUGUG	5p	UUCUUGAACUUCUUAUGCAUC	21	0	0
			5p1	UUCCACAGCUUUCUUGAACUU	21	770	2088
			5p2	UUCCACAGCUUUCUUGAACU	20	75	220
			3p1	GCUCAAGAAAGCUGUGGGAGA	21	48	1719
			3p2	CUCAAGAAAGCUGUGGGAGA	20	44	159
	gma-MIR396h	GAAUGGUCUUUU	5p	UCCACAGCUUUCUUGAACUG	20	0	0
			5p1	UUCCACAGCUUUCUUGAACU	20	75	220
			5p2	UUCCACAGCUUUCUUGAACUG	21	14	50
			3p1	AUUCAAGAUAGCUGUGGAAAA	21	14	37
	gma-MIR396i	UGGCCCUCUUUG	5p	UUCCACAGCUUUCUUGAACUG	21	14	50
			5p1	UUCCACAGCUUUCUUGAACU	20	75	220
			3p	GUUCAAUAAAGCUGUGGGAAG	21	971	2114
			3p1	UUCAAUAAAGCUGUGGGAAG	20	45	160
			3p2	CGGUUCAAUAAAGCUGUGGGA	21	13	39
			3p3	GUUCAAUAAAGCUGUGGGAA	20	10	17
	gma-MIR397a	AUGGAUCUCCUA	5p	UCAUUGAGUGCAGCGUUGAUG	21	1	2
			5p1	AUUGAGUGCAGCGUUGAUGA	20	27	34
			5p2	CAUUGAGUGCAGCGUUGAUGA	21	10	14
			5p3	UUGAGUGCAGCGUUGAUGA	19	1	13
			5p4	AUUGAGUGCAGCGUUGAUGAA	21	5	12
	gma-MIR397b	GAGUACUGUGAG	5p	UCAUUGAGUGCAGCGUUGAUG	21	1	2
			5p1	AUUGAGUGCAGCGUUGAUGA	20	27	34
			5p2	CAUUGAGUGCAGCGUUGAUGA	21	10	14
			5p3	UUGAGUGCAGCGUUGAUGA	19	1	13
			5p4	AUUGAGUGCAGCGUUGAUGAA	21	5	12
	gma-MIR398b	AGUCCAAAUGGU	3p	UGUGUUCUCAGGUCACCCCUU	21	0	3
	gma-MIR398c	GUAGGUAGAGGG	3p	UGUGUUCUCAGGUCGCCCCUG	21	155	79
			3p2	UGUGUUCUCAGGUCGCCCCU	20	76	28
Gm02:49,180,001..49,180,13	New-398	GGUAGAGGAAUU	5p1	UCGUCCUGAGACCACAUGAAA	21	28	16
			3p1	UGUGUUCUCAGGUCGCCCCUG	21	155	79
			3p2	UGUGUUCUCAGGUCGCCCCU	20	76	28
	gma-MIR403a	AGAGCCAAAUUG	3p	UUAGAUUCACGCACAAACUUG	21	9	69
	gma-MIR403b	UGAGAGACAGAG	3p	UUAGAUUCACGCACAAACUUG	21	9	69
	gma-MIR408a	GAUGUUGUCGAC	3p	AUGCACUGCCUCUUCCCUGGC	21	1	18
	gma-MIR408c	UGAGAAUGAGAU	3p	AUGCACUGCCUCUUCCCUGGC	21	1	18
	gma-MIR482a	UCAGAAUUUGUG	5p	AGAAUUUGUGGGAAUGGGCUGA	22	4	1
			5p2	GGAAUGGGCUGAUUGGGAAGC	21	1380	1739
			5p3	GGAAUGGGCUGAUUGGGAAG	20	86	60
			5p4	GGAAUGGGCUGAUUGGGAA	19	119	30
			5p5	UGGGAAUGGGCUGAUUGGGA	20	19	26
			5p6	AAUGGGCUGAUUGGGAAGCAA	21	12	21
			5p7	UGGGAAUGGGCUGAUUGGGAA	21	24	16
			3p	UCUUCCCAAUUCCGCCCAUUCCUA	24	2	0

			known_and_n				
			3p1	UUCCCAAUUCCGCCCAUUCCUA	22	46	55
		AGAAAAAGAGAG	5p	UAUGGGGGGAUUGGGAAGGAAU	22	157	575
			5p1	UAUGGGGGGAUUGGGAAGGA	20	48	680
			5p2	UAUGGGGGGAUUGGGAAGGAA	21	32	410
			5p3	AUGGGGGGAUUGGGAAGGA	19	11	40
			5p4	UAUGGGGGGAUUGGGAAGG	19	0	17
	gma-MIR482b		5p5	AUGGGGGGAUUGGGAAGGAAU	21	5	16
			5p6	GGGAAGGCAUGGGUAUGGGGG	21	20	24
			5p7	CAAAAUUGGGGGAAGGCAUGGG	22	1	17
			3p	UCUUCCCUACACCUCCCAUACC	22	1090	2283
			3p1	UCUUCCCUACACCUCCCAUAC	21	5	14
			3p2	ACUGUUUUUCCUCUCUUUCUC	21	0	6
		AGAAUUUGUGGG	5p	AUUUGUGGGAAUGGGCUGAUUGG	23	1	0
			5p1	GGAAUGGGCUGAUUGGGAAGU	21	1704	669
	gma-MIR482c		5p2	GGAAUGGGCUGAUUGGGAAG	20	86	60
			5p3	GGAAUGGGCUGAUUGGGAA	19	119	30
			3p	UUCCCAAUUCCGCCCAUUCCU	21	3	0
		GGGGGAAGACAU	5p	UAUGGGGGGAUUGGGAAGGAAU	22	157	575
			5p1	UAUGGGGGGAUUGGGAAGGA	20	48	680
	gma-MIR482d		5p2	UAUGGGGGGAUUGGGAAGGAA	21	32	410
			3p	UCUUCCCUACACCUCCCAUACC	22	1090	2283
		AAGAGCUGGGAA	5p	AGAAUUUGUGGGAAUGGGCUGA	22	4	1
			5p1	GGAAUGGGCUGAUUGGGAAGU	21	1704	669
			5p2	GGAAUGGGCUGAUUGGGAAG	20	86	60
			5p3	UGGGAAUGGGCUGAUUGGGAA	21	24	16
Gm18:50,969,490..50,969,635	New-482		5p4	UGGGAAUGGGCUGAUUGGGA	20	19	26
			5p5	GGAAUGGGCUGAUUGGGAA	19	119	30
			3p	UCUUCCCAAUUCCGCCCAUUCCUA	24	2	0
			3p1	UUCCCAAUUCCGCCCAUUCCUA	22	46	55
		UUGCCUUUAUCU	5p	UGCAUUUGCACCUGCACUUU	20	0	0
	gma-MIR530a		3p1	AGGUGCAGGUGCAUCUGCAGG	21	36	25
	New-530d	CCUGCAUUUGCA	3p1	AGGUGCAGGUGCAUCUGCAGG	21	36	25
	New-530e	GUUGAUGCCGAU	3p1	AGGUGCAGGUGCAUCUGCAGG	21	36	25
		AAGAAGUCUUCG	5p1	UCCCUCAAAGGCUUCCAGUAUU	22	5	0
	gma-MIR862a		3p	UGCUGGAUGUCUUUGAAGGAAU	22	51	1
		GGAGAACUCUUC	5p1	UCCCUCAAAGGCUUCCAGUAUU	22	5	0
	gma-MIR862b		3p	GCUGGAUGUCUUUGAAGGA	19	12	1
		CAGUGUUUGGCA	3p	UCUCAUUCCAUACAUCGUCUGA	22	50378	39145
			3p1	UCUCAUUCCAUACAUCGUCU	20	168	221
			3p2	UCAUUCCAUACAUCGUCUGA	20	183	170
			3p3	UCUCAUUCCAUACAUCGUCUG	21	118	110
	gma-MIR1507a		3p4	UCUCAUUCCAUACAUCGUC	19	37	37
			3p5	CAUUCCAUACAUCGUCUGACGA	22	0	23
			3p6	CAUUCCAUACAUCGUCUGA	19	19	11
		GUUUGACAGAGA	5p	UCUCAUUCCAUACAUCGUCUG	21	118	110
			5p1	UCUCAUUCCAUACAUCGUCUGA	22	50378	39145
			5p2	UCUCAUUCCAUACAUCGUCU	20	168	221
			5p3	UCAUUCCAUACAUCGUCUGA	20	183	170
	gma-MIR1507b		5p4	UCUCAUUCCAUACAUCGUC	19	37	37
			5p5	CAUUCCAUACAUCGUCUGACGA	22	0	23
			5p6	CAUUCCAUACAUCGUCUGA	19	19	11
			3p1	AGAGAUGUAUGGAGUGAGAGA	21	27	133
			3p2	GAGAUGUAUGGAGUGAGAGA	20	1	18
		UGGUUUCUAAUC	5p	GAGGUGUUUGGGAUGAGAGAA	21	6	56
	gma-MIR1507c		3p	CCUCAUUCCAAACAUCAUCU	20	0	0
		AAUUGCUAUCCA	5p1	ACUGCUAUUCCCAUUUCUAAA	21	14	39
			3p	UCUAGAAAGGGAAAUAGCAGUU	22	0	0
			3p1	UAGAAAGGGAAAUAGCAGUUG	21	244	300
	gma-MIR1508a		3p2	CUAGAAAGGGAAAUAGCAGUUG	22	46	123
			3p3	UCUAGAAAGGGAAAUAGCAGU	21	25	39
			3p4	UAGAAAGGGAAAUAGCAGUU	20	10	12
		AAUUGCUAUUCA	5p1	ACUGCUAUUCCCAUUUCUAAA	21	14	39
	gma-miR1508b		3p1	CGAGCAUCUUGAUCAAUGGUC	21	9	17
			3p	UCUAGAAAGGGAAAUAGCAGUU	22	0	0
		CUACUCAACUGC	3p	UAGAAAGGGAAAUAGCAGUUG	21	244	300
	gma-miR1508c		3p1	UAGAAAGGGAAAUAGCAGUU	20	10	12
			3p2	UAGAAAGGGAAAUAGCAGU	19	3	9
		CUGCAUCUUCUU	5p1	UUAAUCAAGGAAAUCACGGUCG	22	11660	13221
			5p2	UUAAUCAAGGAAAUCACGGU	20	804	942
			5p3	UUAAUCAAGGAAAUCACGG	19	479	553
	gma-MIR1509a		5p4	UUAAUCAAGGAAAUCACGGUC	21	687	532
			5p5	UAAUCAAGGAAAUCACGGUCG	21	32	39
			5p6	AAUCAAGGAAAUCACGGUCGCG	22	8	23
		CUGCAUCUUUUU	5p1	UUAAUCAAGGAAAUCACGGUU	21	214	215
	gma-MIR1509b		5p2	UUAAUCAAGGAAAUCACGGU	20	804	942
			5p3	UUAAUCAAGGAAAUCACGG	19	479	553

gma-MIR1510a	UUAUGGAACUGG	5p	AGGGAUAGGUAAAACAAUGACUGC	24	0	8
		5p1	AGGGAUAGGUAAAACAAUGAC	21	740	1786
		5p2	AGGGAUAGGUAAAACAAUGACU	22	65	258
		5p3	AGGGAUAGGUAAAACAAUG	19	97	192
		5p4	AGGGAUAGGUAAAACAAUGA	20	59	156
		5p5	UGGAGGGAUAGGUAAAACAAUG	22	14	25
		5p6	GGGAUAGGUAAAACAAUGAC	20	10	16
		5p7	UGGAGGGAUAGGUAAAACAAU	21	8	10
		3p1	UGUUGUUUUACCUAUUCCACC	21	139	512
		3p2	UGUUGUUUUACCUAUUCCACCC	22	139	346
		3p3	UUGUUUUACCUAUUCCACCCAU	22	52	48
		3p4	UGUUGUUUUACCUAUUCCAC	20	17	26
gma-MIR1510b	UUUAUGGAAGUG	5p	AGGGAUAGGUAAAACAACUACU	22	661	2068
		5p1	AGGGAUAGGUAAAACAACUA	20	765	2463
		5p2	AGGGAUAGGUAAAACAACU	19	359	399
		5p3	AGGGAUAGGUAAAACAACUAC	21	62	318
		5p4	GGGAUAGGUAAAACAACUACU	21	7	19
		5p5	GGGAUAGGUAAAACAACUA	19	2	17
		5p6	AGGGAUAGGUAAAACAACUACUU	23	2	12
		3p	UGUUGUUUUACCUAUUCCACC	21	139	512
		3p1	UGUUGUUUUACCUAUUCCACCA	22	96	181
		3p2	UGUUGUUUUACCUAUUCCAC	20	17	26
gma-MIR1511	UCAGCCGUGGUA	5p1	GUGGUAUCAGGUCCUGCUUCA	21	32	110
		3p1	AACCAGGCUCUGAUACCAUG	20	95	113
		3p2	ACCAGGCUCUGAUACCAUG	19	36	43
		3p3	AACCAGGCUCUGAUACCAU	19	32	14
gma-MIR1512b	GUCUUGAUACCU	5p	UAACUGGAAAUUCUUAAAGCAU	22	7	9
gma-MIR1512c	UUCCUUCCUUCC	5p	UAACUGAACAUUCUUAGAGCAU	22	92	13
gma-MIR1513b	GUUUCUAUGCGU	5p1	AAAUCAUGACUUUCUCUUGUA	21	10	74
		5p2	AAUCAUGACUUUCUCUUGUA	20	5	13
		3p	UGAGAGAAAGCCAUGACUUAC	21	35	71
gma-MIR1513c	GAUAUGAGAGAA	5p	UAUGAGAGAAAGCCAUGAC	19	0	0
		5p1	UGAGAGAAAGCCAUGACUUAC	21	35	71
		5p2	AAAGCCAUGACUUACACACGC	21	12	9
gma-miR1514a	CUUUGCUAUUUU	5p	UUCAUUUUUAAAAUAGGCAUU	21	0	0
		5p1	UUCAUUUUUAAAAUAGGCAUUGGG	24	15	13
		5p2	UUCAUUUUUAAAAUAGGCAUUG	22	9	17
gma-MIR1515	UAAAUGUUCAUC	5p	UCAUUUUGCGUGCAAUGAUCUG	22	74	48
		5p1	UCAUUUUGCGUGCAAUGAUCU	21	190	115
		5p2	UUUUGCGUGCAAUGAUCUGAAC	22	15	33
		5p3	UUUUGCGUGCAAUGAUCUGAA	21	3	11
gma-MIR1516	CAUGUUUGGAUA	5p	CAAGUUAUAAGCUCUUUUGAGAG	23	0	0
		5p1	UUGGAUACAAGUUAUAAGCUCU	22	3	11
		3p	CAAAAGAGCUUAUGGCUUGUA	21	0	0
gma-MIR1523	AGGACCCAUUAU	5p	AUGGGAUAAAUGUGAGCUCA	20	2	1
		5p1	UAUGGGAUAAAUGUGAGCUCA	21	50	77
		5p2	UAUGGGAUAAAUGUGAGCUC	20	13	24
gma-MIR1535	UAAGAGGCCUAG	3p	CUUGUUUGUGGUGAUGUCU	19	0	0
		3p1	CUUGUUUGUGGUGAUGUCUAG	21	4	33
		3p2	UGAUCUUGUUUGUGGUGAUGUC	22	1	9
		3p3	UCUUGUUUGUGGUGAUGUCUAG	22	2	4
gma-MIR1535b	UCUUUGGUGACU	5p	CUUGUUUGUGGUGAUGUCUAG	21	4	33
gma-MIR2109	GGUGCGAGUGUC	5p	UGCGAGUGUCUUCGCCUCUG	20	0	0
		5p1	UGCGAGUGUCUUCGCCUCUGA	21	1114	1190
		5p2	GCGAGUGUCUUCGCCUCUGA	20	48	46
		3p	GGAGGCGUAGAUACUCACACC	21	2615	2776
		3p1	GGAGGCGUAGAUACUCACAC	20	209	56
		3p2	GGAGGCGUAGAUACUCACA	19	141	40
gma-MIR2111	AUUGCUGGGUCC	5p1	UAAUCUGCAUCCUGAGGUUUA	21	0	9
		3p	GUCCUUGGGAUGCAGAUUACG	21	1	3
gma-MIR2118a	AAGAGCUUGAGG	5p1	GGAGAUGGGAGGGUCGGUAAAG	22	253	1242
		5p2	GGAGAUGGGAGGGUCGGUAAA	21	88	279
		5p3	GGAGAUGGGAGGGUCGGUA	19	9	41
		5p4	GGGAGAUGGGAGGGUCGGUAAA	22	5	19
		5p5	GGAGAUGGGAGGGUCGGUAA	20	4	12
		3p	UUGCCGAUUCCACCCAUUCCU	21	7	15
		3p1	UUGCCGAUUCCACCCAUUCCUA	22	2675	2913
		3p2	UUGCCGAUUCCACCCAUUCC	20	16	29
		3p4	UUGCCGAUUCCACCCAUUC	19	7	12
gma-MIR2118b	AGAGUGAGAAAG	5p1	GGAGAUGGGAGGGUCGGUAAAG	22	253	1242
		5p2	GGAGAUGGGAGGGUCGGUAAA	21	88	279
		5p3	GGAGAUGGGAGGGUCGGUA	19	9	41
		5p4	GGGAGAUGGGAGGGUCGGUAAA	22	5	19
		5p5	GGAGAUGGGAGGGUCGGUAA	20	4	12
		3p	UUGCCGAUUCCACCCAUUCCU	21	7	15
		3p1	UUGCCGAUUCCACCCAUUCCUA	22	2675	2913

		3p2	UUGCCGAUUCCACCCAUUCC	20	16	29
		3p3	UUGCCGAUUCCACCCAUUC	19	7	12
gma-MIR3522	GCAAAGACAUUU	5p	AGACCAAAUGAGCAGCUGA	19	5	2
		5p1	UGAGACCAAAUGAGCAGCUGA	21	15861	5933
		5p2	UGAGACCAAAUGAGCAGCU	19	39	18
		5p3	UCCUGAGACCAAAUGAGCAGC	21	17	22
		5p4	UGAGACCAAAUGAGCAGCUG	20	14	10
		3p1	AGCUGCUCAUCUGUUCUCAGG	21	286	46
gma-MIR4382	GUACAUAAAUUC	3p	UAUGUUAACUGAUUUCAUGGAU	22	18	30
gma-MIR4387b	ACAAAUGGAGUG	3p	UGUUAGUGAUAAGGCGUGAUG	21	5	7
gma-MIR4397	UUUUCUAGUGGU	5p	CAUCGUUGACGCUGACUGUACG	22	0	7
		3p	UGUCAAAGAUGUGGCGAAUACU	22	0	1
		3p1	UCCCGUCAGUGUCAAAGAUGUG	22	2	10
gma-MIR4411	ACGACGAUUAAU	3p	UUAUUGUAACUAAUUUGUCGGU	22	1	3
		3p1	UAUUGUAACUAAUUUGUCGGUA	22	7	7
		3p2	UUGUAACUAAUUUGUCGGUACC	22	2	3
gma-MIR4412	AACUGUUGCGGG	5p1	UGUUGCGGGUAUCUUUGCCUC	21	76	108
gma-MIR4413	UCAUCAAUAAGA	5p	AAGAGAAUUGUAAGUCACUG	20	0	2
		5p1	UAAGAGAAUUGUAAGUCACUG	21	185	196
		5p2	UAAGAGAAUUGUAAGUCACU	20	24	29
gma-MIR4414	GAGUAGGGUUGU	5p	AGCUGCUGACUCGUUGGCUC	20	0	2
		3p1	UUGGACUGAAGGGGCCUCUU	20	354	756
		3p2	UUGGACUGAAGGGGCCUCUUC	21	151	361
		3p3	CUUGGACUGAAGGGGCCUCUU	21	6	47
		3p4	UGGACUGAAGGGGCCUCUU	19	14	19
		3p5	CUUGGACUGAAGGGGCCUCUUC	22	1	18
gma-MIR4415a	UGGCUGCAGCAA	5p	GGUUAGCUCAAGGAUCUCAC	20	0	0
		3p	UUGAUUCUCAUCACAACAUGG	21	6	15
gma-MIR4415b	GGCUGCAUCAAG	3p	UUGAUUCUCAUCACAACAUGG	21	6	15
gma-MIR4416	CUUUGAUCUGGG	3p	ACGGGUCGCUCUCACCUAGG	20	0	0
		3p1	UACGGGUCGCUCUCACCUAGG	21	14	72
		3p2	AUACGGGUCGCUCUCACCUAGG	22	1	52
gma-MIR4994	CGUGAUAGUGGG	5p	GGUUAGCUCAAGGAUCUCAC	20	0	0
		5p1	GGUUAGCUCAAGGAUCUCACA	21	1	18
		5p2	UUAGCUCAAGGAUCUCACAUG	21	6	8
		5p3	UAGCUCAAGGAUCUCACAUGA	21	3	10
		5p4	AUAGUGGGUUAGCUCAAGGAU	21	6	7
		3p1	UGAUAUCCUUGAGCUAAUACA	21	34	37
gma-MIR4995	GGAGCCGUAGUA	5p	AGGCAGUGGCUUGGUUAAGGG	21	1	2
gma-MIR4996	UUUUCAACUAUU	5p	UAGAAGCUCCCCAUGUUCUC	20	1	1
gma-MIR5035	CUAUCCUAUGAG	5p	CUUCUAAACAUUUUUUCCCUUA	22	0	0
		5p1	UAUCCUAUGAGCUUCUAAACAUU	23	15	13
		3p1	UGUUUAGAAGCUCAUAGAAUAG	22	8	7
		3p2	UGAGGGAAAAAAUGUUUAGAAGCU	24	5	6
Gma-MIR5037	UUGAAAGGAGAA	5p	AACCCUCAAAGGCUUCCUAG	20	0	0
		3p1	CGGAAGUAACUUUGGGGGUUAAC	23	60	1
gma-MIR5038a	UGGGCAUCCAUU	5p	UGAGAAUUUGGCCUCUGUCCA	21	46	47
gma-MIR5038b	CUUGAGAAUUUG	5p	UGAGAAUUUGGCCUCUGUCCA	21	46	47
gma-MIR5041	AGAAACUUAAAG	5p	UUUCAUCUUCAACUUGCUCAA	21	1	1
		3p1	GUUGAGCAAGUUGAAGAUGAA	21	10	57
gma-MIR5042	AAAGAUGCAAAA	5p	UAUCUUGGAUCACAGCCCCAUU	22	1	1
		3p1	UGGGGCUUGAUCCAAGAUAGG	21	3	6
gma-MIR5043	CGACGACGACCG	5p	UGUCCCCUUCUCUGCACCACC	21	5	8
gma-MIR5368	UGUUUCCUGGGA	5p	CCUGGGAUUGGCUUUGGGC	19	3	15
gma-MIR5372	UUGCAUGGUUGU	5p	UUGUUCGAUAAAACUGUUGUG	21	1938	2864
		5p1	UUGUUCGAUAAAACUGUUGUGA	22	80	103
		5p2	UGUUCGAUAAAACUGUUGUGA	21	158	228
		5p3	UCGAUAAAACUGUUGUGAUAAU	22	27	30
gma-MIR5373	UCAUAGAGUCUA	3p	UCUCUUGAUUCUAGAUGAUGU	21	14	32
		3p1	CUUGAUUCUAGAUGAUGUUGA	21	51	66
gma-miR5374	CAUCGAGAUUUU	5p	UUAUAGUCUGACAUCUGGAAU	21	37	97
gma-MIR5375	GUAUUUAGCAAA	5p	ACUAUAGAAGUACUUGUGGAGC	22	0	0
		5p1	UACUAUAGAAGUACUUGUGGAGCU	24	13	24
gma-miR5667	GUUGGAACAGAG	3p	AAACAGAUCUAAAUGGAUUCC	21	7	3

				known_and_new			
	gma-MIR5668	GGAUGCUUUGGA	3p	AGCAAUGGAAUUAUAGACUGC	21	2	8
	gma-MIR5671	GGGCCAUGCCAU	5p	CAUGGAAGUGAAUCGGGUGAC	21	21	26
			5p1	CAUGGAAGUGAAUCGGGUGA	20	26	16
	gma-MIR5674	CACUUGGGUUGA	3p1	UAAUUGUGUUGUACAUUAUCA	21	31	58
	gma-MIR5679	CGAUGACCCUUC	3p1	UUGGUGACCCAGAAGAAGUUGA	22	3	7
	gma-MIR5376	UGAAGAUUUGAA	5p	UGAAGAUUUGAAGAAUUUGGGA	22	246	233
			5p1	UGAAGAUUUGAAGAAUUUGGG	21	17	19
	gma-MIR5379	CACGUCAUCACA	5p1	UCAUCACAGACAUCAUAAUGAA	22	28	42
			3p	AUGAAAAUCAUUCAUUAUGAUAUC	24	0	0
	gma-MIR5671	AUUGAUUCAUCA	5p	CAUGGAAGUGAAUCGGGUGAC	21	21	26
			5p2	CAUGGAAGUGAAUCGGGUGA	20	26	16
			5p3	CAUGGAAGUGAAUCGGGUGACU	22	16	8
			5p4	UGGAAGUGAAUCGGGUGACUC	21	5	5
	gma-MIR5781	CCUCAGACUUCG	3p	CUGAGACUGCAUCUGGCUGAAG	22	7	8
			3p2	AACUGAGACUGCAUCUGGCUGA	22	3	9
Gm05:14,208,279..14,208,353	New-1	UGGUAUUUGUCG	5p	UAUUUGUCGGGGAAAGAAUCGA	22	12	3
Gm19:26,531,292..26,531,389	New-2	AUGGGUGCAGGC	5p	CGCUCUAGGCUGUAUGGGUGCU	22	67	50
			5p2	UGUAUGGGUGCUGAAUUGGA	20	32	40
			5p3	CUGUAUGGGUGCUGAAUUGGA	21	38	29
			5p4	GGCGCUCUAGGCUGUAUGGGUG	22	11	8
			5p5	UCUAGGCUGUAUGGGUGCUGAA	22	14	5
			5p6	CGCUCUAGGCUGUAUGGGUGC	21	8	4
			5p7	GCUCUAGGCUGUAUGGGUGCUG	22	6	5
Gm01:42,369,285..42,369,375	New-3	AUUAAUUUAUAA	5p	AAUUGCUGAAUUGUUGAAGUCGUG	24	4	9
Gm02:48,422,531..48,422,635	New-4	UGUUUUAUCUAC	5p	CCGAAGUCCAAAGACAAGAA	20	0	1
			3p	UCUUGACUUUGGACUUUUGGGU	22	5	8
Gm04:869,312..869,392	New-5.1	UGUUAAAGUGCU	5p	UUAAAGUGCUUCACUUUGUGG	21	10	7
			5p2	UAAAGUGCUUCACUUUGUGGA	21	7	7
			3p	UCAUCCACAAAGUGAAGCACU	21	1	0
					0		
Gm06:878,298..878,373	New-5.2	UUAAAGUGCUUC	5p	UUAAAGUGCUUCACUUUGUGG	21	10	7
			5p2	UAAAGUGCUUCACUUUGUGGA	21	7	7
			3p	UCAUCCACAAAGUGAAGCACU	21	1	0
Gm04:47,053,005..47,053,130	New-6	GUCAUCAACGGA	5p	GGAUCCGUGGCGCAAUGGUA	20	84	16
			5p2	GGAUCCGUGGCGCAAUGGU	19	42	18
Gm05:18,310,293..18,310,402	New-7	UAUCCAAUGUGG	5p	UGGGCGUUAGAGCAUUGAGAGG	22	8	8
			5p2	UGGGCGUUAGAGCAUUGAGAGGA	23	5	9
			5p3	UGGGCGUUAGAGCAUUGAGAG	21	5	8
			5p4	GUGGGCGUUAGAGCAUUGAGAGGA	24	7	4
			5p5	GGCGUUAGAGCAUUGAGAGGA	21	6	4
Gm07:35,762,469..35,762,553	New-8.1	AGCAAAGCUUUC	5p	GCUUUCAUAGCUCAGUUGGU	20	9	8
			5p2	GCUUUCAUAGCUCAGUUGGUUAG	23	6	3
			5p3	GCUUUCAUAGCUCAGUUGGUUA	22	4	5
			3p	GUUCGACUCUCAAUGAAAGCA	21	1	0
			3p2	AGGUCUUGAGUUCGACUCUCA	21	1	0
Gm07:14,680,065..14,680,137	New-8.2	GCUUUCAUAGCU	5p	GCUUUCAUAGCUCAGUUGGU	20	9	8
			5p2	GCUUUCAUAGCUCAGUUGGUUAG	23	6	3
			5p3	GCUUUCAUAGCUCAGUUGGUUA	22	4	5
Gm09:17,884,360..17,884,472	New-9.1	UAACAUUAGCCU	5p	UUAGCCUCGGGAUGGAUAGACU	22	8	2
Gm17:14,285,394..14,285,494	New-9.2	UAACAUUAGCCU	5p	UUAGCCUCGGGAUGGAUAGACU	22	8	2
Gm09:30,681,310..30,681,391	New-10	GGGAAAAGGGCU	5p	UCAUGGACGUUGAUAAGAUCCUUC	24	1	10
Gm10:50,693..50,774	New-11	UGUCUAUCGGUU	5p	UCGGUUAGCUGAAGGGUAUUGUGU	24	0	1
			3p	UCACAGUUGUUUGGUACUCGAA	22	12	7
		''A''''UUUAGUCC	5p	UAUUUUUAGUCCUUAUAGUUU	21	0	1
			3p	AGAAUCAUGAAAUUAUAGGGA	21	13	23
Gm12:33,055,363..33,059,443	New-13.1	UCUCGUAGUUAU	5p	UUAUCAGUAACAUCAUCAUCA	21	8	14
Gm14:13,819,005..13,820,432	New-13.2	UCUCGUAGUUAU	5p	UUAUCAGUAACAUCAUCAUCA	21	8	14
Gm18:23,768,522..23,768,550	New-13.3	UUGUAGUUCUUU	5p	UUAUCAGUAACAUCAUCAUCA	21	8	14
Gm20:9,730,126..9,730,211	New-13.4	UCUCCUAGCUCU	5p	UUAUCAGUAACAUCAUCAUCA	21	8	14
Gm07:10,001,913..10,004,610	New-13.5	UCUCGUAGUUAU	5p	UUAUCAGUAACAUCAUCAUCA	21	8	14
Gm14:13,818,998..13,820,438	New-13.6	AAACAGAUCUCG	5p	UUAUCAGUAGCAUCAUCAUCA	21	34	47
			3p	GUUUGAUGAUGAUGUUACCGA	21	20	13
[illegible]	[illegible]	ACGGUAUAGUGC	5p	UAGUGCUGUAAGACCAUUUCC	21	6	6

				known_and_new			
		GGUGGUCGAGAA	5p	UCGAGAAGUGAAUGUAGCGCUA	22	13	4
			5p2	UGGUCGAGAAGUGAAUGUAGCG	22	6	4
Gm14:1,330,619..1,330,695	New-15		5p3	GUGGUCGAGAAGUGAAUGUAG	21	1	9
			3p	GUUAUGAUUGUUUUUCGGCUUC	22	0	1
		UCAACCAUCCCU	3p	AAUUUGUGUUGCUUAAUCUAUAGG	24	5	10
Gm14:49,068,379..49,073,377	New-16		3p2	UGUGUUGCUUAAUCUAUAGGA	21	6	7
		AACACCUAAUGC	5p	CCAUUAGGUCAGCAAUGGGA	20	1	0
Gm16:18,787,311..18,787,386	New-17		3p	UAUGAGUGACUUAUUAGGUGUU	22	8	7
		AAAUCAGCAAAU	5p	AUCAGAGUGGCGCAGCGGAA	20	306	115
Gm18:10,423,905..10,424,028	New-18		5p2	AUCAGAGUGGCGCAGCGGA	19	227	74
		UGUUUUGCAGAU	5p	UGUUUUGCAGAUGGAACCUGAA	22	20	23
Gm18:17,721,223..17,721,309	New-19		5p2	UUUGCAGAUGGAACCUGAAUUC	22	10	15
			5p3	UGCAGAUGGAACCUGAAUUCUG	22	10	11
		ACAGCAACCACA	5p	GGUUUCGUGGUGUAGUUGGUUA	22	61	17
			5p2	GGUUUCGUGGUGUAGUUGG	19	40	10
Gm18:18,995,461..18,995,549	New-20		5p3	GGUUUCGUGGUGUAGUUGGUUAU	23	43	5
			5p4	GGUUUCGUGGUGUAGUUGGUUAUC	24	23	3
			5p5	GGUUUCGUGGUGUAGUUGGU	20	15	4
Gm18:4,737,836..4,737,905	New-21	UCUGAUCAGAGG	3p	GGAUACCCAGGUAUGUGAAAGGCA	24	10	6
Gm19:37,100,979..37,101,075	New-22	CAUAUUGAUCAA	3p	UUAAAGUAUAGGGACCAAAUU	21	3	7
		CACUAUGGGAGC	5p	AUGGGAGCUGGCCAUGCCCGA	21	0	2
			3p	GGCAGGGCUAGUGACUGGAGUG	22	0	36
Gm19:44,377,618..44,377,700	New-23		3p2	GAAGGCAGGGCUAGUGACUGGA	22	5	9
			3p3	GCAGGGCUAGUGACUGGAGUG	21	3	9
		UGGAUUAAAACG	3p	GGUGUCGUGGUGUAGUUGG	19	439	215
			3p2	GGUGUCGUGGUGUAGUUGGUU	21	151	53
			3p3	GGUGUCGUGGUGUAGUUGGUUAU	23	84	23
			3p4	GGUGUCGUGGUGUAGUUGGU	20	71	24
Gm20:38,400,661..38,400,749	New-24		3p5	GGUGUCGUGGUGUAGUUGGUUA	22	20	9
			3p6	GGUGUCGUGGUGUAGUUGGUUAUC	24	18	7
			3p7	UGUCGUGGUGUAGUUGGUUA	20	17	5
			3p8	GUCGUGGUGUAGUUGGUUAUC	21	9	3
Scaffold_70:93,145..93,259	New-25	AGAGUAUCGUAG	5p	GAGUAUCGUAGACGUUGAUCAUCU	24	8	4
		AAGGGGUUGUAG	5p	GGGGUUGUAGCUCAUAUGGU	20	111	58
			5p2	GGGGUUGUAGCUCAUAUGGUA	21	75	38
Gm06:10,518,738..10,518,826	New-26		5p3	GGGGUUGUAGCUCAUAUGGUAGA	23	55	12
			5p4	GGGGUUGUAGCUCAUAUGG	19	10	5
Gm13:40,760,284..40,760,374	New-27	UUCUUCUCUUAA	5p	UAAACGUUUGAUCCCUUGUAU	21	127	13
		GGGGAGAGAUGG	5p	GGAGAGAUGGCUGAGUGGACU	21	47	56
			5p2	GGAGAGAUGGCUGAGUGGA	19	55	54
Gm15:46,628,479..46,628,608	New-28		5p3	GGAGAGAUGGCUGAGUGGACUA	22	32	23
			5p4	GGAGAGAUGGCUGAGUGGACUAAA	24	43	21
			5p5	GGAGAGAUGGCUGAGUGGAC	20	16	19

Supplementary Table 2

Supplentary Table 2: Predicted targets of identified moRNAs confirmed by degradome analyses

moRNAs	Position	Target *Loci*	moRNA-Target alignments	Putative Function
gma-MIR169f	3p2	Glyma01g37370.1	miRNA 20 UACUUUUCU**UC**UCUCAGGUU 1 :::::::. .:::::::::::: Target 190 AUGAAAGUG**AG**AGAGUCCAA 209	Pfam:00643 B-box zinc finger GO:0005622 intracellular GO:0008270 zinc ion binding
gma-MIR169f	5p3	Glyma09g01810.1	miRNA 21 UGAUGUGAGA**GA**CGGAGAAGC 1 :.::::::::::::: .::: Target 1231 AUUACACUUU**CU**GCCAUUUCA 1251	Pfam:00702 haloacid dehalogenase-like hydrolase Panther:19288 4-NITROPHENYLPHOSPHATASE-RELATED KOG:2882 p-Nitrophenyl phosphatase

CAPÍTULO II - Description of plant tRNA-derived RNA fragments (tRFs) associated with Argonaute and identification of their putative targets

Autores: Guilherme Loss-Morais[1], Peter M. Waterhouse[2] and Rogerio Margis[1§]

[1]Laboratório de Genomas e Populações de Plantas, Centro de Biotecnologia, Universidade Federal do Rio Grande do Sul

[2] Waterhouse Laboratory, The University of Sydney

Trabalho publicado na revista *Biology Direct* (I.F. 4.02)

Loss-Morais *et al. Biology Direct* 2013, **8**:6
http://www.biology-direct.com/content/8/1/6

BIOLOGY DIRECT

DISCOVERY NOTES Open Access

Description of plant tRNA-derived RNA fragments (tRFs) associated with argonaute and identification of their putative targets

Guilherme Loss-Morais[1], Peter M Waterhouse[2] and Rogerio Margis[1*]

Abstract

tRNA-derived RNA fragments (tRFs) are 19mer small RNAs that associate with Argonaute (AGO) proteins in humans. However, in plants, it is unknown if tRFs bind with AGO proteins. Here, using public deep sequencing libraries of immunoprecipitated Argonaute proteins (AGO-IP) and bioinformatics approaches, we identified the *Arabidopsis thaliana* AGO-IP tRFs. Moreover, using three degradome deep sequencing libraries, we identified four putative tRF targets. The expression pattern of tRFs, based on deep sequencing data, was also analyzed under abiotic and biotic stresses. The results obtained here represent a useful starting point for future studies on tRFs in plants.

Keywords: tRNAs, Small RNA, tRFs, tRNA-derived RNA fragments, Argonaute and Arabidopsis

Findings

Small RNAs are usually ~20 nucleotides long. Regardless of their genomic origin, small RNAs can regulate gene expression by acting as siRNAs to direct DNA methylation [1] or by acting as microRNAs to direct post transcriptional gene silencing (PTGS) [2]. microRNAs are the most studied class of small RNAs [3]. Moreover, the key enzymes related to small RNA biogenesis, such as Dicer-Like (DCL) and AGO proteins, and their roles in PTGS have been well described [2].

The recent development of high-throughput sequencing technology has improved the identification of other types of small RNAs [4], like tRNA-derived RNA fragments (tRFs) [3]. The proposed nomenclature of tRFs is based on the regions of tRNA cleavage, including 3' U tRFs that are processed from pre-tRNAs and consist of the sequence between the cleavage site and the RNA PolIII run-off poly(U) tract [5]. Mature tRNA can generate two main types of tRFs: one processed from the 5' end (5' tRFs) and another from the 3' end, harboring the added CCA sequence (3' CCA tRFs) [5].

The tRFs were first discovered in cultured *Hela* cells [6]. Subsequent work in other animal tissues showed that tRF biogenesis may involve RNAse Z [5] as well as Dicer processing [6-8].

Recently, it has been suggested that there might be cross-talk between tRFs and the canonical small RNA pathway, which includes the microRNAs [5]. Another exciting finding was that of the association of tRFs with AGO proteins [6,7] and the demonstration of a RNAi-type trans-silencing induced by a 3' CCA tRF using a reporter gene [7].

At present, only three works show the existence of tRFs in plants. In *Arabidopsis thaliana*, the 5' tRF of AspGTC and the 5' and 3' CCA tRFs of GlyTCC tRNAs were found to be overexpressed in root tissues treated with phosphate deprivation [9]. In rice, the 5' AlaAGC and ProCGG tRFs demonstrated differential expression in the callus and leaves [4]; in barley, the HisGTG tRF was the most abundant of all the small RNAs [10]. However, the possible association of tRFs to AGO proteins and their potential contribution to the RNAi pathway were not analyzed in either of the previous studies.

The work described here was designed to identify putative AGO-associated tRFs in *Arabidopsis thaliana* by analyzing public small RNA deep sequencing libraries, including those from AGO immunoprecipitation (AGO-IP) assays. Putative tRF target sequences were also found by examining *Arabidopsis* public degradome sequencing libraries. The expression patterns of tRFs under abiotic

* Correspondence: rogerio.margis@ufrgs.br
[1]Universidade Federal do Rio Grande do Sul, Centro de Biotecnologia, Predio 43431, Sala 213, POBox 15005, Porto Alegre, RS, Brazil
Full list of author information is available at the end of the article

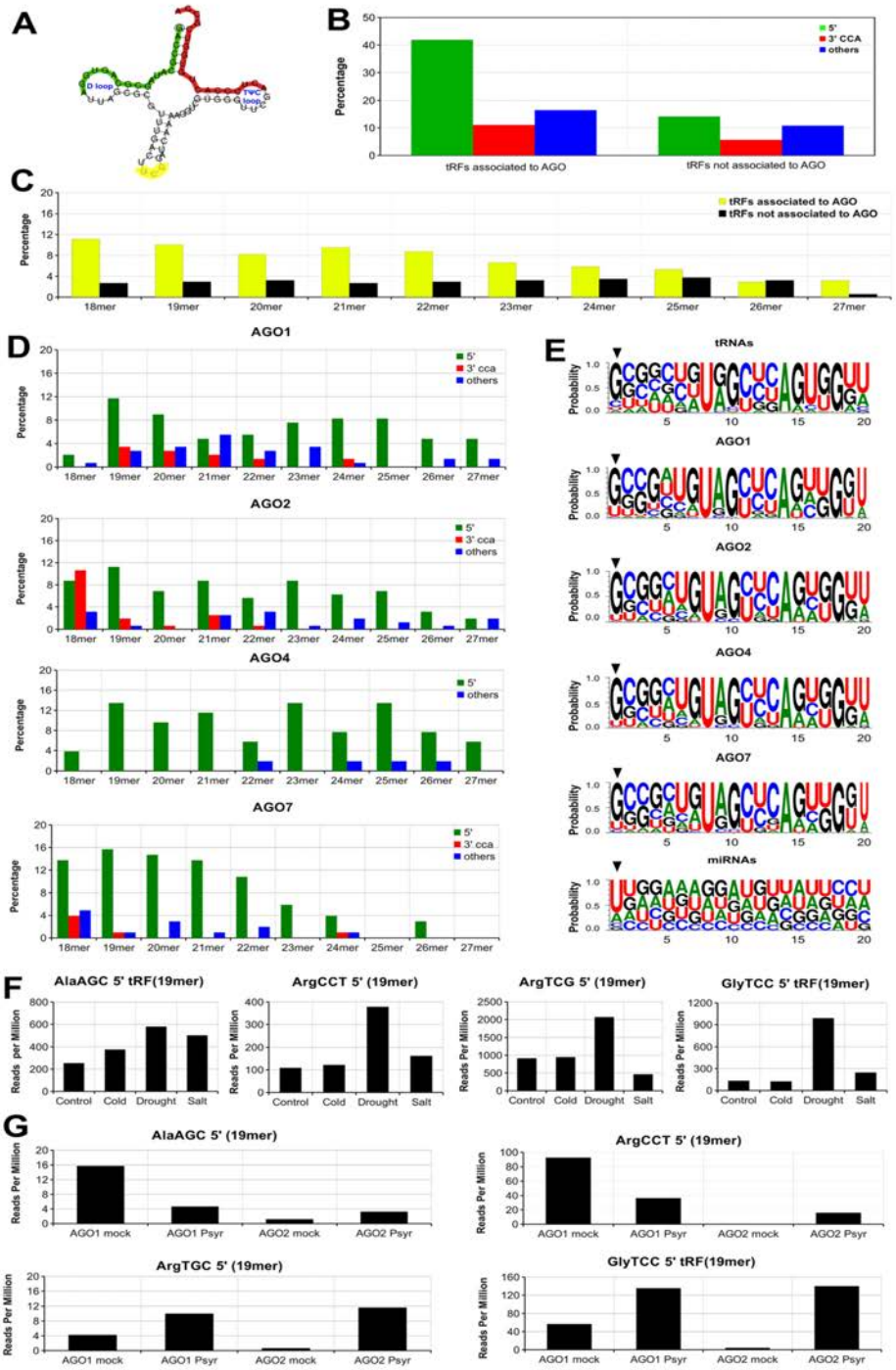


Figure 1 (See legend on next page.)

(See figure on previous page.)
Figure 1 tRNA-derived RNA fragments (tRFs) from *Arabidopsis thaliana* associated with AGO. A) Schematic representation of ArgTCG tRNA showing the 5' tRF in green, the 3' CCA tRF in red and the anti-codon in yellow. The D and TΨC loops are also shown. **B)** tRF class diversity of AGO-associated tRFs and unassociated ones. **C)** Length diversity of AGO-associated tRFs and unassociated ones. **D)** tRF length diversity of AGO1, 2, 4 and 7 IP deep sequencing libraries. **E)** Logo representation of the first 20 nucleotides of tRNAs, AGO1-IP tRFs, AGO2-IP tRFs, AGO4-IP tRFs, AGO7-IP tRFs and the *A. thaliana* microRNAs (miRBase v. 18). The black arrowheads indicate the first nucleotide at the 5' end. **F)** Expression pattern of AlaAGC, ArgCCT, ArgTCG and GlyTCC 5' tRFs in control (untreated), drought (40-50% relative water content), cold (5°C for 24 hours), and salt (200 mM of NaCl for 5 hours) conditions. The expression patterns are shown in reads per million, where the tRF frequency was divided by the total number of reads and multiplied by one million. **G)** Expression pattern of AlaAGC, ArgCCT, ArgTCG and GlyTCC 5' tRFs in biotic stress. The expression patterns are also shown in reads per million. The leaves were inoculated with mock solution (10 mM $MgCl_2$) or *Pseudomonas syringae* (2 x 107 cfu/ml). The inoculated leaves were collected 14 hours after inoculation.

and biotic stresses were also analyzed. The present work focused on 5' and 3' CCA tRFs in *A. thaliana*, but sequences derived from the central regions of the tRNA were also searched (see methods) (Figure 1A).

We inspected AGO1, 2, 4, 6, 7 and 9 IP libraries [See Additional file 1: Table S1] and found tRFs in the AGO1, 2, 4 and 7 IP libraries (Figure 1B,-D) [See Additional file 2: Table S2]. Both, 5' and 3' CCA *Arabidopsis* tRFs were associated with AGO, mirroring previous results in mammalian systems [6,7]. Interestingly, tRFs from the central part of the tRNA were also detected (Figure 1B,-D), although 5' tRFs formed the most abundant class [4,6,9] and showed the highest sequence diversity (Figure 1B,-D).

Examining the AGO-associated and unassociated tRFs (Figure 1C) [See Additional file 3: Figure S1] revealed a bias in size distribution, with the AGO-associated tRFs being predominantly 18-22 (nt) in length (Figure 1C) and the AGO-associated 5'tRFs being predominantly 19 mers (Figure 1D) [See Additional file 3: Figure S1]. This is very similar to the situation in *Hela* cells [6].

The predominant 5' terminal nucleotide of microRNAs is a uracil [11], and this first base is thought to be a major determinant for loading onto AGO1. AGO2 and AGO4 preferentially recruit small RNAs with a 5' terminal A [12,13]. However, the most common 5' nucleotide of 5' tRFs is G (Figure 1E). Takeda et al. (2008) suggested that *Arabidopsis* may have an AGO gene with a preference for microRNAs starting with guanine [12]; however, it does not seem to be applicable to tRFs.

Further, to investigate if the 5' tRFs associated with AGOs act in the RNAi pathway in plants, as has been suggested in animals [7], we looked for tRF targets in *Arabidopsis* using a well-known plant microRNA target prediction tool coupled with degradome analyses. This analysis identified four possible target genes [See Additional file 4: Table S3]. However, this method assumes that the mechanism and characteristics of tRF target recognition are similar to those for microRNAs, which remains to be demonstrated. Indeed, it is possible that tRFs may play a role in DNA and chromatin modification because we found that tRFs associated with AGO4 (Figure 1D), which is known to be involved in this process [12].

In order to inspect the expression pattern of tRFs in abiotic stress treatments, we conducted an analysis of the AlaAGC, ArgCCT, ArgTCG and GlyTCC 5' tRFs, using the available deep sequencing data (Figure 1F). Drought conditions enhanced the expression of the four tRFs, including the GlyTCC 5' tRF, which is already known to be up-regulated in response to phosphate deprivation [9]. Hsieh et al. (2009) discussed that tRFs accumulate in a developmentally regulated manner and become dominant in specific tissues or under specific stress conditions [9]. Thus, the 5' GlyTCC seems to be dominant in both phosphate deprivation and drought treatment.

The expression pattern of tRFs under biotic stress in plants is currently unknown. In order to identify tRFs that respond to biotic stress, we conducted an expression analysis of the same four 5' tRFs in AGO1 and AGO2 immunoprecipitated deep sequencing libraries from *Arabidopsis* infected with *Pseudomonas syringae* or mock solution (Figure 1G). The four 5' tRFs showed increased expression in infected AGO2-IP libraries (Figure 1G). AGO2 is a protein of unknown function [2]; however, this protein was recently characterized as being strongly induced by *P. syringae* infection [14]. This work also investigated the microRNA pathway and showed that the expression levels of miR393*, which associated with AGO2-IP and targets a transcript related to exocytosis, was enhanced in *P. syringae* infection assay [14]. Here, we found an increase in expression of 5' tRFs in the AGO2-IP, indicating a possible role for 5' tRFs in *P. syringae* infection. However, more experiments should be performed.

Conclusions

Small RNAs are important regulators of gene expression, and recent advances in sequencing and bioinformatics techniques have stimulated the discovery of new classes of small RNAs. Here, we report for the first time that tRNA-derived RNA fragments (tRFs) associate with AGO proteins in plants. The first nucleotide does not

seem to determine which 5' tRF is directed to which AGO protein, as observed in microRNAs. However, there is some enrichment of uridine at the 5' end. Moreover, we identified putative tRF targets and analyzed the expression of tRFs under abiotic and biotic stresses. The results presented in this study can be considered as valuable support for future studies on the complex networks involved in tRF-mediated gene regulation in plants.

Methods

In order to find tRFs associated with AGO, 34 deep sequencing libraries were retrieved from the GEO database (http://www.ncbi.nlm.nih.gov/geo/) [15], including 25 libraries of AGO-IP and three degradome libraries [See Additional file 1: Table S1]. We identified a third tRF class, corresponding to tRFs originating from the internal sequences of the tRNA. These reads did not map to the very first nucleotide of 5' tRFs or the very last nucleotide of 3' CCA tRFs.

The bioinformatics approaches used to identify tRFs associated with AGO were shown in Additional file 5: Figure S2. Briefly, reads from a control (GSM647184) library were mapped against all mature *Arabidopsis* tRNAs previously obtained from the TAIR database (http://www.arabidopsis.org), resulting in putative tRFs. Further, the putative tRFs were used as a query to inspect the AGO-IP libraries. The putative tRFs, which were found in the AGO-IP and had a frequency of more than 10 reads, were retrieved and considered AGO-associated tRFs. Later, the AGO-associated tRFs were used for target prediction against all *Arabidopsis* transcripts using the psRNATarget tool (http://plantgrn.noble.org/psRNATarget/). The degradome libraries were used to confirm possible target cleavage, lowering the false positive rate in the tRF target prediction.

Additional files

Additional file 1: Table S1. Details of the deep sequencing libraries used in the present analyses.

Additional file 2: Table S2. List and details of the tRFs identified in the present work.

Additional file 3: Figure S1. Raw read frequencies of AGO1, 2, 4 and 7 immunoprecipitated libraries. Raw frequency of the tRFs is also shown. The most expressed reads or tRFs of each AGO-IP library are underlined.

Additional file 4: Table S3. Report the predicted tRFs targets validated by degradome analyses.

Additional file 5: Figure S2. Fluxogram showing the bioinformatics approaches for identification and tRF target prediction of AGO-associated tRFs. The putative targets were used as a reference to screen degradome libraries. The degradome reads, which were mapped to the approximate central portion of the tRF target recognition site and show at least one match and one wobble in tRF:target pairing, were retrieved. So far, putative targets were validated by degradome analyses.

Competing interests
The authors declare that they have no competing interests.

Authors' contribution
GLM conceived the idea, performed the computational work and wrote the paper. PMW and RM contributed to the interpretation of the results and the preparation of manuscript. All authors read and approved the final manuscript.

Authors' response
The main criticism made by both referees concerned the necessity of experimental validation of predicted targets of Arabidopsis tRFs in order to demonstrate the reliability of the predictions made.
As stated, the target prediction was performed using psRNAtarget. The putative hibrydization site between each tRF and its corresponding target transcript were searched in public Arabidopsis DEGRADOME sequencing libraries. The authors consider that the presence of the corresponding sequences in the DEGRADOME libraries provides reliability, in a first instance, to the *in silico* predicted targets. Authors agree that RACE and other experiments would be required to assure the exactitude and extent of tRFs targets, but also consider that these experiments would be out of the scope of the present work. It is important to remark that along the reviewing process of this work, a third paper was published about the identification of tRFs in plants, but without any comments about their association to AGO proteins. This work was incorporated in our list of references.

Reviewer number: 1

Report form:
The authors screened existing Arabidopsis databases of the small RNAs associated with AGO to find tRNA-derived small RNAs (tRFs) and to identify their potential targets. The work is rather modest in scope and would greatly benefit from experimental validation of the tRF targets. The outcome of the work could be useful for those studying RNAi pathways in plants.

Reviewer number: 2

Report form:
This is the first report of tRNA fragments associated with Argonaute in plants and accordingly of interest. The manuscript would be improved if the authors were explicit about the reliability of the tRF target prediction.

Acknowledgements
This work was supported by CNPq (Conselho Nacional de DesenvolvimentoCientífico e Tecnológico) grant number 400790/2012-2. G. Loss-Morais has a Ph.D. fellowship from CAPES, and R. Margis is a recipient of CNPq research fellowship number 307868/2011-7.

Author details
[1]Universidade Federal do Rio Grande do Sul, Centro de Biotecnologia, Predio 43431, Sala 213, POBox 15005, Porto Alegre, RS, Brazil. [2]The University of Sydney, Sydney, NSW, Australia.

Received: 25 October 2012 Accepted: 7 February 2013
Published: 12 February 2013

References

1. Mosher R, Melnyk CW: **siRNAs and DNA methylation: seedy epigenetics.** *Trends Plant Sci* 2010, **15:**204–210.
2. Voinnet O: **Origin, Biogenesis, and Activity of Plant MicroRNAs.** *Cell* 2009, **136**(4):669–687.
3. Lee YS, Shibata Y, Malhotra A, Dutta A: **A novel class of small RNAs: tRNA-derived RNA fragments (tRFs).** *Genes Dev* 2009, **23:**2639–2049.
4. Chen CJ, Liu Q, Zhang YC, Qu LH, Chen YQ, Gautheret D: **Genome-wide discovery and analysis of microRNAs and other small RNAs from rice embryogenic*callus*.** *RNA Biol* 2011, **8:**538–547.
5. Sobala A, Hutvagner G: **Transfer RNA-derived fragments: origins, processing, and functions.** *Wiley Interd Reviews RNA* 2011, **2:**853–62.
6. Cole C, Sobala A, Lu C, Thatcher SR, Bowman A, Brown JWS, Green PJ, Barton GJ, Hutvagner G: **Filtering of deep sequencing data reveals the existence of abundant Dicer-dependent small RNAs derived from tRNAs.** *RNA* 2009, **15:**2147–2160.

7. Haussecker D, Huang Y, Lau A, Parameswaran A, Fire AZ, Kay M: **Human tRNA-derived small RNAs in the global regulation of RNA silencing.** *RNA* 2010, **16:**673–695.
8. Babiarz JE, Ruby JG, Wang Y, Bartel DP, Blelloch R: **Mouse ES cells express endogenous shRNAs, siRNAs, and other Microprocessor-independent, Dicer-dependent small RNAs.** *Genes Dev* 2008, **22**(20):2773–2785.
9. Hsieh LC, Lin SI, Shih ACC, Chen JW, Lin WY, Tseng CY, Li WH, Chiou TJ: **Uncovering small RNA-mediated responses to phosphate deficiency in Arabidopsis by deep sequencing.** *Plant Phys* 2009, **151:**2120–2132.
10. Hackenberg M, Huang PJ, Huang CY, Shi BJ, Gustafson P, Langridge P: **A Comprehensive Expression Profile of MicroRNAs and Other Classes of Non-Coding Small RNAs in Barley Under Phosphorous-Deficient and -Sufficient Conditions.** *DNA Research* 2012, **19:**1–17.
11. Zhang B, Pan X, Cannon CH, Cobb GP, Anderson T: **Conservation and divergence of plant microRNA genes.** *Plant J* 2006, **46:**243–259.
12. Takeda A, Iwasaki S, Watanabe T, Utsumi M, Watanabe Y: **The mechanism selecting the guide strand from small RNA duplexes is different among argonaute proteins.** *Plant Cell Phys* 2008, **49:**493–500.
13. Mi S, Cai T, Hu Y, Chen Y, Hodges E, Ni F, Li S, Zhou H, Long C, Chen C, Hannon GJ: **Sorting of Small RNAs into Arabidopsis Argonaute Complexes Is Directed by the 5′ Terminal Nucleotide.** *Cell* 2008, **133:**116–127.
14. Zhang X, Zhao H, Gao H, Wang H, Katiyar-Agarwal S, Huang H, Raikhe N, Jin N: **Arabidopsis Argonaute 2 Regulates Innate Immunity via miRNA393*-Mediated Silencing of a Golgi-Localized SNARE Gene, MEMB12.** *Mol Cell* 2011, **42:**356–366.
15. Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall K, Phillippy KH, Sherman PM, Muertter RN, Holko MK, Ayanbule O, Yefanov A, Soboleva A: **NCBI GEO: Archive for functional genomics data sets–10 years on.** *Nucleic Acids Res* 2011, **39:**1005–1010.

doi:10.1186/1745-6150-8-6
Cite this article as: Loss-Morais *et al.*: Description of plant tRNA-derived RNA fragments (tRFs) associated with argonaute and identification of their putative targets. *Biology Direct* 2013 **8**:6.

Additional files provided with this submission:

Additional file 1: Additional_File_1.xls, 25K
http://www.biology-direct.com/imedia/1536052740834380/supp1.xls
Additional file 2: Additional_File_2.xls, 52K
http://www.biology-direct.com/imedia/1807406939834380/supp2.xls
Additional file 3: Additional_File_3.tif, 2276K
http://www.biology-direct.com/imedia/1068203603834381/supp3.tiff
Additional file 4: Additional_File_4.pdf, 779K
http://www.biology-direct.com/imedia/1408578600834381/supp4.pdf
Additional file 5: Additional_File_5.tif, 1481K
http://www.biology-direct.com/imedia/1356876180834381/supp5.tiff

Additional File 3: Raw read frequencies of AGO1, 2, 4 and 7 immunoprecipitated libraries. Raw frequency of the tRFs is also shown. The most expressed reads or tRFs of each AGO-IP library are underlined.

Additional File 5: Fluxogram showing the bioinformatics approaches for identification and tRF target prediction of AGO-associated tRFs. The putative targets were used as a reference to screen degradome libraries. The degradome reads, which were mapped to the approximate central portion of the tRF target recognition site and show at least one match and one wobble in tRF:target pairing, were retrieved. So far, putative targets were validated by degradome analyses.

Additional File 1

Additional File 1: GEO Accessions, small RNA library titles, tissue or organ, Illumina platform and *A. thaliana* ecotypes of deep sequencing libraries used for tRFs identification. The GEO Accessions with gray background indicates experiments without Argonaute Immune-Precipitation and the underlined ones indicates degradome libraries. The dash represents lack of information.

GEO Accessions	Small RNA Library Title	Tissue / organ	Illumina Platform	Ecotypes
GSM647184	Arabidopsis_control	shoot	GA	Col gl1 RD29:LUC
GSM647185	Arabidopsis_drought	shoot	GA	Col gl1 RD29:LUC
GSM647186	Arabidopsis_cold	shoot	GA	Col gl1 RD29:LUC
GSM647187	Arabidopsis_salt	shoot	GA	Col gl1 RD29:LUC
GSM253622	Small RNAs that are associated with Arabidopsis AGO1 complex	-	GA	Col-0
GSM642335	AGO1-associated sRNA_mock	leaves	GA	Col-0
GSM642336	AGO1-associated sRNA_R2	leaves	GA	Col-0
GSM707682	Two-step_purification_AGO1-associated_sRNAs_flower	flower	GAII	-
GSM707683	Two-step_purification_AGO1-associated_sRNAs_leaf	leaves	GAII	-
GSM707684	Two-step_purification_AGO1-associated_sRNAs_root	root	GAII	-
GSM707685	Two-step_purification_AGO1-associated_sRNAs_seedling	seedlings	GAII	-
GSM707690	IP_AGO1-associated_sRNAs_flower	flower	GAII	-
GSM707691	IP_AGO1-associated_sRNAs_root	root	GAII	-
GSM253623	Small RNAs that are associated with Arabidopsis AGO2 complex	-	GA	Col-0
GSM642337	AGO2-associated sRNA_mock	leaves	GA	Col-0
GSM642338	AGO2-associated sRNA_R2	leaves	GA	Col-0
GSM304283	HA-AGO2 Co-IP Fraction Small RNA	Inflorescence tissue: flower stages 1-12	GAII	Col-0
GSM253624	Small RNAs that are associated with Arabidopsis AGO4 complex	-	GA	Col-0
GSM415787	antiAGO4 associated sRNAs R2	mixed stage of inflorescence	GAII	Col-0
GSM415788	antiAGO4 associated sRNAs R1	mixed stage of inflorescence	GAII	Col-0
GSM707686	Two-step_purification_AGO4-associated_sRNAs_flower	flower	GAII	-
GSM707687	Two-step_purification_AGO4-associated_sRNAs_leaf	leaves	GAII	-
GSM707688	Two-step_purification_AGO4-associated_sRNAs_root	root	GAII	-
GSM707689	Two-step_purification_AGO4-associated_sRNAs_seedling	seedlings	GAII	-
GSM253625	Small RNAs that are associated with Arabidopsis AGO5 complex	-	GA	Col-0
GSM415789	FLAG AGO6 associated sRNAs (SL9)	mixed stage of inflorescence	GAII	Col-0
GSM415790	FLAG AGO6 associated sRNAs (SL10)	mixed stage of inflorescence	GAII	Col-0
GSM304285	HA-AGO7 Co-IP Fraction Small RNA	Inflorescence tissue: flower stages 1-12	GAII	Col-0
GSM415791	antiAGO9 associated sRNAs (SL11)	mixed stage of inflorescence	GAII	Col-0
GSM415792	antiAGO9 associated sRNAs (SL12)	mixed stage of inflorescence	GAII	Col-0
GSM278334	dT primed, pool-amplified	Inflorescence	GA	Col-0
GSM278335	random primed, primer extension	Inflorescence	GA	Col-0
GSM278370	random primed, primer extension	Seedlings	GA	Col-0

Additional File 2: tRFs associated to Argonaute proteins. The tRNAs, sequences, lengths and classes are shown.

tRNAs	Sequences	TRFs Length	Class
TrpCCA	TTCACGTCGGGTTCACCA	18	3' CCA
TyrGTA	AAATCCAGCTCGGCCCACCA	20	3' CCA
LeuTAA	AACCCCACAGCCTGCACCA	19	3' CCA
SerTGA	AACCCTGCTGTCGACGCCA	19	3' CCA
SerTGA	AACCCTGCTGTTGACGCCA	19	3' CCA
LeuCAG	AATCCCACTCTTGACACCA	19	3' CCA
LeuTAA	ACCCCACAGCCTGCACCA	18	3' CCA
ValCAC	ACCCGGGCTCAGACACCA	18	3' CCA
SerTGA	ACCCTGCTGTCGACGCCA	18	3' CCA
SerTGA	ACCCTGCTGTTGACGCCA	18	3' CCA
GlnTTG	ACTCCCGGTAGGACCTCCA	19	3' CCA
LeuCAG	ATCCCACTCTTGACACCA	18	3' CCA
LeuAAG	ATCCCACTGTCAACACCA	18	3' CCA
AspGTC	ATCCCCGGCAACGGCGCCA	19	3' CCA
SerCGA	ATCCTGCTGTTGACGCCA	18	3' CCA
SerGCT	ATCTCTCAGGCGACGCCA	18	3' CCA
TrpCCA	ATTCACGTCGGGTTCACCA	19	3' CCA
GlyGCC	ATTCCCGGCTGGTGCACCA	19	3' CCA
TyrGTA	CAAATCCAGCTCGGCCCACCA	21	3' CCA
LeuCAG	CAAATCCCACTCTTGACACCA	21	3' CCA
GlnCTG	CAATTCTCGGTAGAACCTCCA	21	3' CCA
AsnGTT	CCCCTCCTTCTAGCGCCA	18	3' CCA
SerAGA	CGAATCCTGCCGTTCACGCCA	21	3' CCA
AspGTC	CGATCCCCGGCAACGGCGCCA	21	3' CCA
GlyGCC	CGATTCCCGGCTGGTGCACCA	21	3' CCA
GlyTCC	CTCCCGGCAGACGCACCA	18	3' CCA
GlnTTG	CTCCCGGTAGAACCTCCA	18	3' CCA
GlnTTG	CTCCCGGTAGGACCTCCA	18	3' CCA
TyrGTA	CTGGTTCAAATCCAGCTCGGCCCACCA	27	3' CCA
GluTTC	GATTCCCGGCATCGGAGCCA	20	3' CCA
GlyGCC	GATTCCCGGCTGGTGCACCA	20	3' CCA
TyrGTA	GTTCAAATCCAGCTCGGCCCACCA	24	3' CCA
LeuTAG	GTTCGAGTCCGAGTAGCGGCACCA	24	3' CCA
AspGTC	GTTCGATCCCCGGCAACGGCGCCA	24	3' CCA
GlnCTG	TCAATTCTCGGTAGAACCTCCA	22	3' CCA
PheGAA	TCCACGCTCACCGCACCA	18	3' CCA
LysTTT	TCCCCACAGACGGCGCCA	18	3' CCA

AspGTC	TCCCCGGCAACGGCGCCA	18	3' CCA
AlaAGC	TCGATACCCCGCATCTCCACCA	22	3' CCA
GlyGCC	TCGATTCCCGGCTGGTGCACCA	22	3' CCA
GlyGCC	TTCCCGGCTGGTGCACCA	18	3' CCA
ProAGG ProTGG	TTCTCGGAACGCCCCCCA	18	3' CCA
GlnCTG	TTCTCGGTAGAACCTCCA	18	3' CCA
TyrGTA	CCGACCTTAGCTCAGTTGG	19	5'
TyrGTA	CCGACCTTAGCTCAGTTGGT	20	5'
TyrGTA	CCGACCTTAGCTCAGTTGGTA	21	5'
TyrGTA	CCGACCTTAGCTCAGTTGGTAG	22	5'
TyrGTA	CCGACCTTAGCTCAGTTGGTAGA	23	5'
TyrGTA	CCGACCTTAGCTCAGTTGGTAGAGC	25	5'
LeuTAG	GACAGTTTGGCCGAGTGG	18	5'
ArgTCG	GACCGCATAGCGCAGTGG	18	5'
ArgTCG	GACCGCATAGCGCAGTGGA	19	5'
ArgTCG	GACCGCATAGCGCAGTGGAT	20	5'
ArgTCG	GACCGCATAGCGCAGTGGATTAGCG	25	5'
ArgCCG	GACCGCGTGGCCTAATGGA	19	5'
GlyGCC	GCACCAGTGGTCTAGTGGTA	20	5'
GlyGCC	GCACCAGTGGTCTAGTGGTAGA	22	5'
GlyGCC	GCACCAGTGGTCTAGTGGTAGAA	23	5'
GlyGCC	GCACCAGTGGTCTAGTGGTAGAAT	24	5'
GlyGCC	GCACCAGTGGTCTAGTGGTAGAATA	25	5'
GlyGCC	GCACCAGTGGTCTAGTGGTAGAATAG	26	5'
GlyGCC	GCACCAGTGGTCTAGTGGTAGAATAGT	27	5'
ThrTGT	GCCCGTATAGCTCAGTGGT	19	5'
ThrTGT	GCCCGTATAGCTCAGTGGTA	20	5'
ThrTGT	GCCCGTATAGCTCAGTGGTAGAGC	24	5'
LysCTT	GCCCGTCTAGCTCAGTTGG	19	5'
LysCTT	GCCCGTCTAGCTCAGTTGGT	20	5'
LysCTT	GCCCGTCTAGCTCAGTTGGTA	21	5'
LysCTT	GCCCGTCTAGCTCAGTTGGTAG	22	5'
LysCTT	GCCCGTCTAGCTCAGTTGGTAGA	23	5'
LysCTT	GCCCGTCTAGCTCAGTTGGTAGAGC	25	5'
LeuTAG	GCCGCTATGGTGAAATTGG	19	5'
LeuTAG	GCCGCTATGGTGAAATTGGT	20	5'
LeuTAG	GCCGCTATGGTGAAATTGGTA	21	5'
LeuTAG	GCCGCTATGGTGAAATTGGTAGA	23	5'
LysTTT	GCCGTCTTAGCTCAGTGG	18	5'
LysTTT	GCCGTCTTAGCTCAGTGGT	19	5'
LysTTT	GCCGTCTTAGCTCAGTGGTA	20	5'
LysTTT	GCCGTCTTAGCTCAGTGGTAGAG	23	5'

LysTTT	GCCGTCTTAGCTCAGTGGTAGAGC	24	5'
ArgCCT	GCGCCTGTAGCTCAGTGG	18	5'
ArgCCT	GCGCCTGTAGCTCAGTGGA	19	5'
ArgCCT	GCGCCTGTAGCTCAGTGGATAG	22	5'
ArgCCT	GCGCCTGTAGCTCAGTGGATAGAGC	25	5'
PheGAA	GCGGGGATAGCTCAGTTG	18	5'
PheGAA	GCGGGGATAGCTCAGTTGG	19	5'
PheGAA	GCGGGGATAGCTCAGTTGGG	20	5'
PheGAA	GCGGGGATAGCTCAGTTGGGA	21	5'
PheGAA	GCGGGGATAGCTCAGTTGGGAG	22	5'
PheGAA	GCGGGGATAGCTCAGTTGGGAGA	23	5'
PheGAA	GCGGGGATAGCTCAGTTGGGAGAG	24	5'
PheGAA	GCGGGGATAGCTCAGTTGGGAGAGC	25	5'
GlyTCC	GCGTCTGTAGTCCAACGG	18	5'
GlyTCC	GCGTCTGTAGTCCAACGGT	19	5'
GlyTCC	GCGTCTGTAGTCCAACGGTT	20	5'
GlyTCC	GCGTCTGTAGTCCAACGGTTA	21	5'
GlyTCC	GCGTCTGTAGTCCAACGGTTAG	22	5'
GlyTCC	GCGTCTGTAGTCCAACGGTTAGG	23	5'
GlyTCC	GCGTCTGTAGTCCAACGGTTAGGA	24	5'
GlyTCC	GCGTCTGTAGTCCAACGGTTAGGAT	25	5'
GlyTCC	GCGTCTGTAGTCCAACGGTTAGGATA	26	5'
GlyTCC	GCGTCTGTAGTCCAACGGTTAGGATAA	27	5'
AsnGTT	GCTGGAATAGCTCAGTTGGT	20	5'
AsnGTT	GCTGGAGTAGCTCAGTTGGT	20	5'
ThrAGT	GCTTTCATAGCTCAGTTGGTTAGAG	25	5'
TrpCCA	GGATCCGTGGCGCAATGG	18	5'
TrpCCA	GGATCCGTGGCGCAATGGTA	20	5'
IleAAT	GGCCTATTAGCTCAGTTGGT	20	5'
IleAAT	GGCCTATTAGCTCAGTTGGTTA	22	5'
ProTGG ProAGG ProCGG	GGGCATTTGGTCTAGTGGTATGATTCT	27	5'
ProAGG ProTGG	GGGCGTTTGGTCTAGTGGTATGATTCT	27	5'
AlaAGC AlaTGC	GGGGATGTAGCTCAAATG	18	5'
AlaAGC AlaTGC	GGGGATGTAGCTCAAATGG	19	5'
AlaAGC AlaTGC	GGGGATGTAGCTCAAATGGT	20	5'
AlaAGC AlaTGC	GGGGATGTAGCTCAAATGGTA	21	5'
AlaAGC AlaTGC	GGGGATGTAGCTCAAATGGTAG	22	5'
AlaAGC AlaTGC	GGGGATGTAGCTCAAATGGTAGA	23	5'
AlaAGC AlaTGC	GGGGATGTAGCTCAAATGGTAGAG	24	5'
AlaAGC AlaTGC	GGGGATGTAGCTCAAATGGTAGAGC	25	5'
AlaAGC	GGGGATGTAGCTCAGATG	18	5'
AlaAGC	GGGGATGTAGCTCAGATGG	19	5'

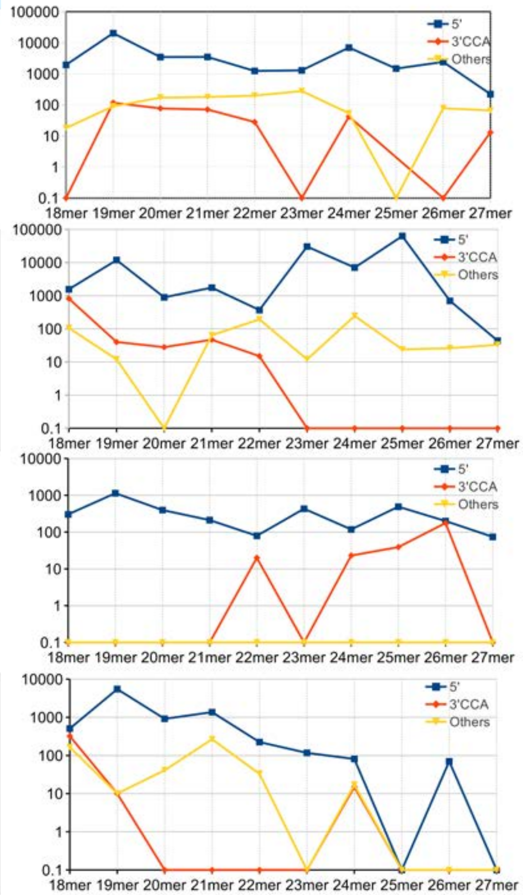
AspGTC	GTCGTTGTAGTATAGTGGTAAGTA	24	5'
AspGTC	GTCGTTGTAGTATAGTGGTAAGTAT	25	5'
AspGTC	GTCGTTGTAGTATAGTGGTAAGTATT	26	5'
AspGTC	GTCGTTGTAGTATAGTGGTAAGTATTC	27	5'
ValCAC_	GTCTGGGTAGTGTAGTCGG	19	5'
ValCAC_	GTCTGGGTAGTGTAGTCGGTTATCA	25	5'
SerAGA	GTGGACGTGCCGGAGTGGT	19	5'
SerAGA	GTGGACGTGCCGGAGTGGTT	20	5'
SerAGA	GTGGACGTGCCGGAGTGGTTA	21	5'
SerAGA	GTGGACGTGCCGGAGTGGTTAT	22	5'
SerAGA	GTGGACGTGCCGGAGTGGTTATC	23	5'
SerAGA	GTGGACGTGCCGGAGTGGTTATCGGG	26	5'
HisGTG	GTGGCTGTAGTTTAGTGGT	19	5'
LeuAAG	GTTGATATGGCCGAGTTG	18	5'
LeuAAG	GTTGATATGGCCGAGTTGG	19	5'
GluTTC	TCCGATGTCGTCCAGCGG	18	5'
GluTTC	TCCGATGTCGTCCAGCGGT	19	5'
GluTTC	TCCGATGTCGTCCAGCGGTTAGG	23	5'
GluTTC	TCCGATGTCGTCCAGCGGTTAGGA	24	5'
GluTTC	TCCGATGTCGTCCAGCGGTTAGGAT	25	5'
GluTTC	TCCGATGTCGTCCAGCGGTTAGGATA	26	5'
GluCTC	TCCGTCGTAGTCTAGCTGG	19	5'
GluCTC	TCCGTCGTAGTCTAGCTGGTT	21	5'
GluCTC	TCCGTCGTAGTCTAGCTGGTTAG	23	5'
GluCTC	TCCGTCGTAGTCTAGCTGGTTAGG	24	5'
GluCTC	TCCGTCGTAGTCTAGCTGGTTAGGA	25	5'
GluCTC	TCCGTCGTAGTCTAGCTGGTTAGGATA	27	5'
GluTTC	TCCGTTATCGTCCAGCGGTTAGG	23	5'
GluTTC	TCCGTTATCGTCCAGCGGTTAGGATA	26	5'
GluCTC	TCCGTTGTAGTCTAGCTG	18	5'
GluCTC	TCCGTTGTAGTCTAGCTGG	19	5'
GluCTC	TCCGTTGTAGTCTAGCTGGTC	21	5'
GluCTC	TCCGTTGTAGTCTAGCTGGTCA	22	5'
GluCTC	TCCGTTGTAGTCTAGCTGGTCAG	23	5'
GluCTC	TCCGTTGTAGTCTAGCTGGTCAGG	24	5'
GluCTC	TCCGTTGTAGTCTAGCTGGTCAGGA	25	5'
GluCTC	TCCGTTGTAGTCTAGCTGGTCAGGAT	26	5'
GluCTC	TCCGTTGTAGTCTAGCTGGTCAGGATA	27	5'
AsnGTT	TCCTCAGTAGCTCAGTGG	18	5'
AsnGTT	TCCTCAGTAGCTCAGTGGTA	20	5'
AsnGTT	TCCTCAGTAGCTCAGTGGTAG	21	5'
AsnGTT	TCCTCAGTAGCTCAGTGGTAGA	22	5'

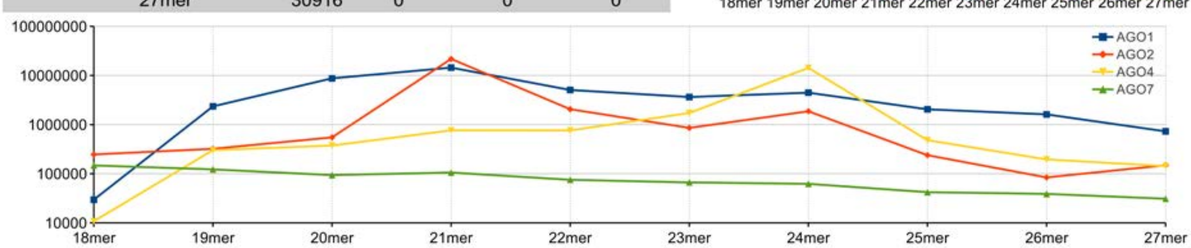
AsnGTT	TCCTCAGTAGCTCAGTGGTAGAG	23	5'
LeuTAG	GTTCGAGTCCGAGTAGCGGC	20	other
ArgCCT	AAGCAGAAGGTCGTAGGT	18	other
TyrGTA	AATGGGGACGGACTGTAAATTCGT	24	other
GlnCTG	ACATTGGACTCTGAATCCAGT	21	other
ArgTCG	ACCGCATAGCGCAGTGGA	18	other
TyrGTA	ACCTTAGCTCAGTTGGTA	18	other
GlnCTG	AGGACATTGGACTCTGAATCCAGT	24	other
GluCTC	AGGATACTCGGCTCTCACCC	20	other
GluTTC	ATGTCGTCCAGCGGTTAGGATATCTGG	27	other
GluTTC	CAGCGGTTAGGATATCTGGCT	21	other
GluCTC	CAGGATACTCGGCTCTCACCC	21	other
LysCTT	CCCGTCTAGCTCAGTTGGTAGAGC	24	other
GluTTC	CCGATGTCGTCCAGCGGTTAGGAT	24	other
LysTTT	CCGTCTTAGCTCAGTGGTAGAGC	23	other
AsnGTT	CCTCAGTAGCTCAGTGGTAGAGC	23	other
GlyTCC	CCTTCCAAGCAATAGACCCG	20	other
ArgCCT	CGCCTGTAGCTCAGTGGA	18	other
GlyTCC	CGTCTGTAGTCCAACGGTTAGGA	23	other
GlnCTG	CTAGCGGTTAGGACATTGGACT	22	other
GluCTC	CTCGGCTCTCACCCGAGAGA	20	other
GlnCTG	GACATTGGACTCTGAATCCA	20	other
GlnCTG	GACATTGGACTCTGAATCCAGT	22	other
TyrGTA	GACCTTAGCTCAGTTGGTA	19	other
GluTTC	GATGTCGTCCAGCGGTTAGGAT	22	other
ProTGG ProAGG ProCGG	GCATTTGGTCTAGTGGTATGATTCTC	26	other
GluCTC	GCTGGTTAGGATACTCGGCTCT	22	other
HisGTG	GCTGTAGTTTAGTGGTAAGAATTCC	25	other
HisGTG	GGCTGTAGTTTAGTGGTAAGAATTCC	26	other
AlaAGC AlaTGC	GGGATGTAGCTCAAATGGTAGAG	23	other
AlaAGC	GGGATGTAGCTCAGATGG	18	other
AlaAGC	GGGATGTAGCTCAGATGGT	19	other
AlaAGC	GGGATGTAGCTCAGATGGTA	20	other
AlaAGC	GGGATGTAGCTCAGATGGTAGA	22	other
AlaAGC	GGGATGTAGCTCAGATGGTAGAGC	24	other
GluCTC	GTAGTCTAGCTGGTTAGGA	19	other
LeuTAG LeuAAG LeuCAG	GTCCGAAAGGGCGTGGGTTCA	21	other
GlyGCC	GTCTAGTGGTAGAATAGTACCC	22	other
GlyTCC	GTCTGTAGTCCAACGGTTAGGA	22	other
ProAGG ProTGG ProCGG	GTGGTATGATTCTCGCTT	18	other
GluTTC	GTTATCGTCCAGCGGTTAGGA	21	other
AspGTC	GTTGTAGTATAGTGGTAAGTATTCCC	26	other

GluCTC	GTTGTAGTCTAGCTGGTCAGGA	22	other
GluCTC	GTTGTAGTCTAGCTGGTCAGGAT	23	other
GluCTC	GTTGTAGTCTAGCTGGTCAGGATACTC	27	other
ArgCCT	TAAGCAGAAGGTCGTAGGT	19	other
LeuTAA	TAAGGGGGAAGACTTAAGTTC	21	other
GlyGCC	TAGAATAGTACCCTGCCACGGTACA	25	other
GlnCTG	TAGGACATTGGACTCTGAATCCAGT	25	other
GluTTC	TAGGATATCTGGCTTTCACCC	21	other
ProAGG ProTGG ProCGG	TAGTGGTATGATTCTCGCTT	20	other
GlnTTG	TAGTGGTTAGCACTCTGGA	19	other
GluTTC	TCCAGCGGTTAGGATATCTGGC	22	other
AspGTC	TCGTTGTAGTATAGTGGTAAGTATTCC	27	other
GlnCTG	TCTAGCGGTTAGGACATTGGA	21	other
GlyGCC	TCTAGTGGTAGAATAGTACCC	21	other
ProAGG ProTGG ProCGG	TCTAGTGGTATGATTCTCGCTT	22	other
AlaAGC	TGCGAGAGGTACGGGGATC	19	other
HisGTG	TGGCTGTAGTTTAGTGGTAAGAATTCC	27	other
GluTTC	TGTCGTCCAGCGGTTAGGATATCTGG	26	other
GluTTC	TTAGGATATCTGGCTTTCACC	21	other
GluTTC	TTAGGATATCTGGCTTTCACCC	22	other
ValAAC	TTCGTGGTGTAGTTGGTTATC	21	other
SerTGA	TTGGGCTTCGCCCGCGCAGGTTCG	24	other
ProAGG ProTGG ProCGG	TTTGGTCTAGTGGTATGATTCTC	23	other

Additional File 3

AGO-IP	Size	Raw library reads	tRFs		
			5'	3'CCA	Others
AGO1	18mer	29581	1940	0	18
	19mer	2349250	20382	117	90
	20mer	8739952	3460	77	171
	21mer	14455789	3498	71	179
	22mer	5077062	1235	28	199
	23mer	3636069	1296	0	278
	24mer	4462020	7000	42	54
	25mer	2045131	1477	0	0
	26mer	1610270	2409	0	78
	27mer	728765	222	13	66
AGO2	18mer	246607	1561	815	106
	19mer	321508	11889	40	12
	20mer	548963	893	28	0
	21mer	21692644	1750	47	64
	22mer	2059962	369	15	193
	23mer	855813	30493	0	12
	24mer	1869846	7041	0	245
	25mer	238237	62945	0	24
	26mer	83985	700	0	26
	27mer	148992	44	0	33
AGO4	18mer	10814	302	0	0
	19mer	302348	1141	0	0
	20mer	375469	397	0	0
	21mer	764413	212	0	0
	22mer	760535	79	20	0
	23mer	1722373	430	0	0
	24mer	14289038	118	23	0
	25mer	481709	490	39	0
	26mer	195706	199	176	0
	27mer	142875	74	0	0
AGO7	18mer	147611	508	322	164
	19mer	122785	5482	10	10
	20mer	93549	913	0	41
	21mer	105456	1364	0	266
	22mer	75338	225	0	33
	23mer	66623	117	0	0
	24mer	62369	81	15	17
	25mer	42012	0	0	0
	26mer	38826	70	0	0
	27mer	30916	0	0	0



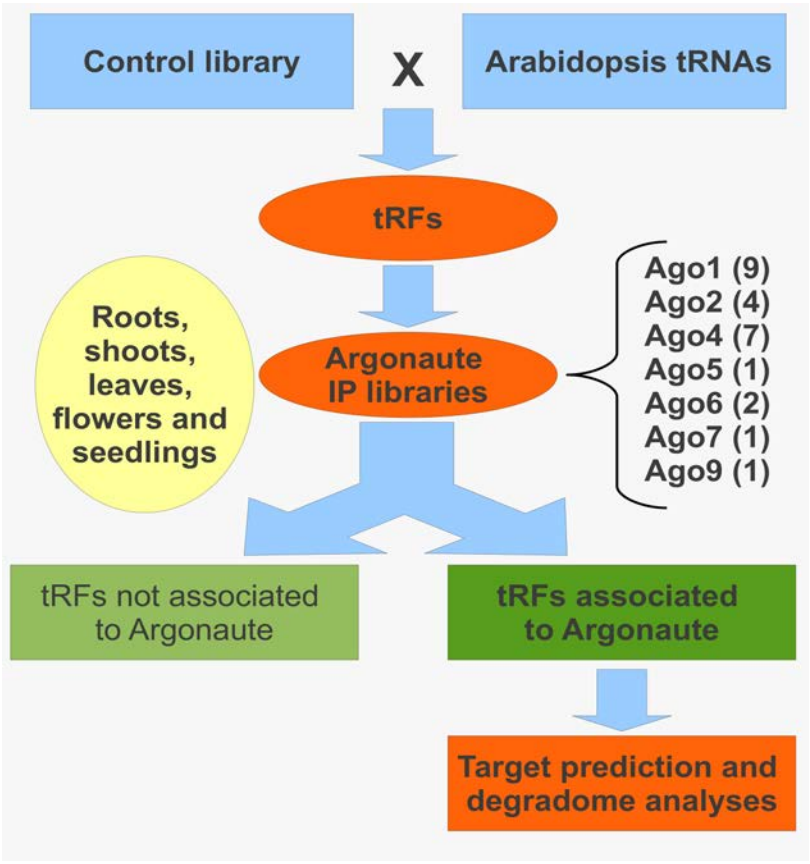


Additional File 4

Additional File 4: Degradome analyses of predicted tRFs targets. The tRFs acronyms, target loci, tRFs:target alignments, target descriptions, GO biological processes and GO molecular functions. The region of predicted target cleavage is showed by underlined regions.

tRFs	Target loci	tRFs:target alignments	Target Descriptions	GO Biological Processes	GO Molecular Functions
AlaAGC (19mer)	AT3G61060.1	tRF 17 UAGACUCGAUGUAGGGG 1 :::::.:::.:::.:.: Target 1206 AUCUGGGCUGCAUUCUC 1222	Phloem protein 2-A13 (PP2-A13)	Involved in response to wounding	Functions in carbohydrate binding
ArgCCT (19mer)	AT3G05050.1	tRF 19 AGGUGACUCGAUGUCCGCG 1 :.::::::: ::::::. Target 1164 UUCACUGAGCAACAGGUUA 1182	Protein kinase superfamily protein	Involved in protein phosphorylation	Funtions in serine/threonine kinase activity
ArgTCG (19mer)	AT2G24790.1	tRF 18 GGUGACGCGAUACGCCAG 1 ::.::::::.:::::: Target 261 CCGCUGCGUUAUGCGUCA 278	COL3 (CONSTANS-LIKE 3) transcription factor	Involved in regulation of photomorphogenesis, regulation of transcription	Functions in protein binding, zinc ion binding
GlyTCC (19mer)	AT3G57280.1	tRF 18 GGCAACCUGAUGUCUGCG 1 :.:::::.::: :..:. Target 925 CUGUUGGAUUACUGGUGU 942	Transmembrane protein 14C	unknown	unknown

Additional File 5



3.0 Considerações Finais e Perspectivas

3.1 Considerações Finais

O papel dos pequenos RNAs como reguladores negativos ao longo do desenvolvimento de órgãos e tecidos (Guleria et al., 2011), bem como em respostas a estresses bióticos e abióticos (Khraiwesh, et al., 2012), está bem documentada em inúmeros trabalhos científicos, nos quais, no mínimo, uma relação de expressão entre o microRNA e seu transcrito alvo foi estabelecida.

De forma geral, se aceita que exista uma correlação de expressão inversa entre o microRNA e seu transcrito alvo (Ritchie et al., 2009; Voinnet, 2009). Este mecanismo é denominado de miRNA "*switch*" (Alonso-Peral et al., 2012), devido ao papel de forte repressor deste sobre a expressão do transcrito alvo, quando aumentada a expressão do microRNA que o regula.

Contudo, alguns trabalhos têm demonstrado que existem exceções desta correlação inversa. Recentemente, Alonso-Peral e colaboradores (2012) demonstraram que a abundância do MIR159 de *A. thaliana* foi constante entre folhas (rosetas), meristema apical caulinar e sementes em germinação, diferentemente do seu conhecido transcrito alvo (MYB33), o qual variou entre estes tecidos. O autor sugere que além do mecanismo de *switch*, o qual ocorre em folhas e no meristema apical caulinar, existe o mecanismo de "*tuning*" da expressão gênica induzido por miRNAs, como ocorre na germinação da semente. A regulação por *tuning* seria mais fina, implicando em pequenas mudanças de expressão entre microRNA e transcrito alvo (Alonso-Peral et al., 2012).

Além disso, existe a regulação espacial e a regulação temporal da expressão de miRNAs e transcritos alvos (Voinnet, 2009). Entende-se por regulação espacial quando o microRNA e seu alvo

são transcritos em um mesmo tempo, mas em células distintas, mesmo que vizinhas. Por outro lado, a regulação de expressão temporal é quando um microRNA e seu alvo são expressos em uma mesma célula, mas em períodos de tempo distintos (Voinnet, 2009).

Admite-se que a regulação espacial possa resultar na correlação positiva, como exemplificado na expressão do MIR395 e seu transcrito alvo SULTR2 em Arabidopsis, durante privação de enxofre, onde o MIR395 tem expressão aumentada no floema, células companheiras e mesófilo, enquanto que o SULTR2 tem a expressão aumentada em células do xilema e parênquima e ausência de expressão no floema, células companheiras e mesófilo (Kawashima et al., 2009).

A análise da expressão do MIR395 e SULTR2 tecido específica não resultaria em nenhuma correlação, diferentemente ao se analisar amostras destes tecidos juntamente, resultando em uma correlação positiva, indicando regulação espacial entre miRNA e transcrito alvo (Gomollon et al., 2012; Kawashima et al., 2009). Logo, assumindo que microRNAs são exclusivamente reguladores negativos de seus transcritos alvo, uma correlação positiva pode ser um indicativo de ocorrência de uma regulação espacial (Gomollon et al., 2012).

Surpreendentemente, Gomollon et al. (2012), estudando o desenvolvimento do fruto de tomate, identificou que os números de correlações positivas e negativas eram similares, indo contra o modelo apresentando no trabalho seminal de Voinnet (2009), em que a co-regulação negativa seria dominante. Esta similaridade de correlações pode ser o motivo que levou Alonso-Peral et al. (2012) a afirmar que a eficácia no RNAi de microRNAs pode variar com o tecido analisado. Além disso, é importante ressaltar que os microRNAs representam somente uma camada de regulação gênica para a qual a regulação transcricional e estabilidade de RNA podem contribuir ainda mais para um distanciamento do esperado perfil de correlação negativa (Gomollon et al., 2012).

O estabelecimento da relação dos pequenos RNAs e seu transcrito alvo baseado no possível perfil de expressão de correlação negativa pode representar uma subestimativa do número real destes. Portanto, talvez seja mais prudente, antes da análise de expressão de ambos, a confirmação se realmente ocorre PTGS entre o pequeno RNA e o seu transcrito alvo em estudo. Porém, esta abordagem ainda não é muito difundida, sendo que a maioria dos trabalhos científicos publicados, ainda foca na identificação de alvos baseando-se na possível correlação negativa.

A identificação de um transcrito que possa ser alvo de um pequeno RNA por PTGS, pode ser analisada através de PCR 5' RACE, análise de degradoma ou SILAC, sendo a última semelhante à análise de expressão, só que em nível protéico. Cabe frisar que o RNAi em nível protéico também pode ser dependente de regulação espacial e temporal, apesar de nenhum trabalho científico ter considerado este fato até o momento. Por consequência, a análise de SILAC pode ter a mesma problemática supracitada.

Os genes MIR geralmente são pequenos, conservados e em alguns casos pertencem a uma família extensa, tornando estratégias de *knockout* gênico para genes MIRs laboriosas e demoradas (Eamens & Wang, 2011). Estratégias de construção de mutantes *knockdown*, seja pelo método anti-miRNA (Eamens et al., 2011), *target mimicry* (Franco-Zorrilla et al., 2007), ou silenciamento gênico transcricional do miRNA (Vaistij et al., 2010), podem auxiliar na confirmação de RNAi. A utilização da técnica de hibridização *in situ* de pequenos RNAs (Wheeler et al., 2007) pode também auxiliar na caracterização a regulação espacial e temporal dos pequenos RNAs.

Apesar da metodologia de sequenciamento de alta eficiência ser extremamente útil para identificar pequenos RNAs, ressalvas devem ser feitas nas análises de expressão de pequenos RNAs, devido a estes não serem, geralmente, efetuados em replicatas biológicas, impedindo análises estatísticas

robustas e uma confiável análise de expressão (Tarazona et al., 2011). Os trabalhos de identificação de microRNAs em sementes maduras e em germinação de soja e da caracterização de tRFs associados a proteínas AGO em Arabidopsis, bem como a identificação e validação de seus respectivos transcritos alvos são um primeiro passo na caracterização do papel de reguladores gênicos destes pequenos RNAs em plantas.

A idéia inicial do presente projeto de Doutorado era caracterizar microRNAs envolvidos com o metabolismo lipídico durante o desenvolvimento da semente de soja, mas devido a duas publicações de outros grupos de pesquisa abordando os microRNAs durante o desenvolvimento da semente em *G. max*, mudamos o enfoque da pesquisa para a caracterização dos miRNAs durante a germinação da soja. Apesar desta dificuldade inicial, nosso grupo continua estudando microRNAs durante o desenvolvimento da semente de outras oleógenas, como canola (Körbes et al., 2012), pinhão-manso e tungue (dados não publicados).

Todo este esforço resume-se em possíveis aplicações futuras dos microRNAs como alvo para estratégias de engenharia genética, com objetivo de aumentar a qualidade e/ou quantidade de lipídeos nas sementes, bem como a possível utilização do padrão de expressão dos microRNAs como marcador molecular do conteúdo lipídico.

O propósito de estudar tRFs veio da observação que existiam muitos pequenos RNAs, oriundos de sequenciamento de alta eficiência, que eram mapeados em tRNAs. Ao revisar a bibliografia, vimos que estes pequenos RNAs tinham sido caracterizados em humanos, mas fracamente caracterizados em plantas. Deste fato, surgiu o segundo trabalho desta tese, o qual se baseou em uma simples, mas robusta, análise de bioinformática de dados públicos de sequenciamento de alta eficiência.

3.2 PERSPECTIVAS

Os microRNAs de soja identificados na sementes madura e em germinação poderão ser analisados da seguinte forma:

- Validação de clivagem dos transcritos alvo através de PCR 5' RACE;
- Análise da expressão dos miRNAs e genes alvo por RT-qPCR durante diferentes estádios de desenvolvimento da semente;
- Hibridização *in situ* dos microRNAs.

Tendo em vista que tRFs são menos caracterizados do que microRNAs, o nosso grupo de pesquisa investigará nos seguintes experimentos:

- Validação de clivagem dos transcritos alvo através de PCR 5' RACE;
- Análise da expressão de tRFs e transcritos alvo por RT-qPCR em diferentes tecidos, órgãos e estresses;
- Construção de plantas transgênicas de Arabidopsis super-expressando os tRFs;
- Avaliação fenotípica e molecular das plantas transgênicas sob diferentes estresses;
- Análise de uma possível translocação de tRFs no floema (Laboratório do Dr. Peter Waterhouse).
- Hibridização *in situ* dos tRFs.

4.0 Referências Bibliográficas da Introdução e Considerações finais

Addo-Quaye, C., Eshoo, T. W., Bartel, D. P., Axtell, M. J. (2008). Endogenous siRNA and miRNA targets identified by sequencing of the Arabidopsis degradome. Current Biology, 18(10), 758–762. doi:10.1016/j.cub.2008.04.042.

Addo-Quaye, C., Miller, W., Axtell, M. J. (2009). CleaveLand: a pipeline for using degradome data to find cleaved small RNA targets. Bioinformatics, 25(1), 130–131. doi:10.1093/bioinformatics/btn604.

Alonso-Peral, M. M., Sun, C., Millar, A. a. (2012). MicroRNA159 can act as a switch or tuning microRNA independently of its abundance in Arabidopsis. PloS One, 7(4), e34751. doi:10.1371/journal.pone.0034751.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J. (1990). Basic local alignment search tool. Journal of Molecular Biology, 215(3), 403–410. doi:10.1016/S0022-2836(05)80360-2.

Axtell, M. J., Westholm, J. O., Lai, E. C. (2011). Vive la différence: biogenesis and evolution of microRNAs in plants and animals. Genome Biology, 12(4), 221-234. doi:10.1186/gb-2011-12-4-221.

Bellaloui, N. (2010). Soybean seed protein, oil, fatty acids, and mineral composition as influenced by soybean-corn rotation. Agricultural Sciences, 1(3), 102–109. doi:10.4236/as.2010.13013.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., (2012). Europe PMC Funders Group Accurate Whole Human Genome Sequencing using Reversible Terminator Chemistry. Nature, 456(7218), 53–59. doi:10.1038/nature07517.

Blevins, T., Pontes, O., Pikaard, C. S., Meins, F. (2009). Heterochromatic siRNAs and DDM1 independently silence aberrant 5S rDNA transcripts in Arabidopsis. PloS One, 4(6), e5932. doi:10.1371/journal.pone.0005932.

Brennecke, J., Stark, A., Russell, R. B., Cohen, S. M. (2005). Principles of microRNA-target recognition. PLoS Biology, 3(3), e85. doi:10.1371/journal.pbio.0030085

Bräutigam, A. Gowik, U. (2010). What can next generation sequencing do for you? Next generation sequencing as a valuable tool in plant research. Plant Biology, 12(6), 831–841. doi:10.1111/j.1438-8677.2010.00373.x.

Cai, Y., Zhou, Q., Yu, C., Wang, X., Hu, S., Yu, J., Yu, X. (2012). Transposable-element associated small RNAs in *Bombyx mori* genome. PloS One, 7(5), e36599. doi:10.1371/journal.pone.0036599.

Carthew, R. W. Sontheimer, E. J. (2009). Review Origins and Mechanisms of miRNAs and siRNAs. Cell, 136(4), 642–655. doi:10.1016/j.cell.2009.01.035.

Chen, C.-J., Liu, Q., Zhang, Y.-C., Qu, L.-H., Chen, Y.-Q., Gautheret, D. (2011). Genome-wide discovery and analysis of microRNAs and other small RNAs from rice embryogenic *callus*. RNA Biology, 8(3), 538–547. doi:10.4161/rna.8.3.15199.

Chen, X. (2004). A microRNA as a translational repressor of APETALA2 in Arabidopsis flower development. Science, 303(5666), 2022–2025. doi:10.1126/science.1088060.

Cheng, K. C. C. Strömvik, M. V. (2008). SoyXpress: a database for exploring the soybean transcriptome. BMC Genomics, 9, 368. doi:10.1186/1471-2164-9-368.

Clemente, T. E., Cahoon, E. B. (2009). Soybean oil: genetic approaches for modification of functionality and total content. Plant Physiology, 151(3), 1030–1040. doi:10.1104/pp.109.146282.

Coetzee, B., Freeborough, M.J., Maree, H. J., Celton, J.-M., Rees, D. J. G., Burger, J. T. (2010). Deep sequencing analysis of viruses infecting grapevines: Virome of a vineyard. Virology, 400(2), 157–163. doi:10.1016/j.virol.2010.01.023.

Dai, X., Zhuang, Z., Zhao, P. X. (2011). Computational analysis of miRNA targets in plants: current status and challenges. Briefings in bioinformatics, 12(2), 115–121. doi:10.1093/bib/bbq065.

Ding, J., Zhou, S., Guan, J. (2012). Finding microRNA targets in plants: current status and perspectives. Genomics, Proteomics & Bioinformatics, 10(5), 264–275. doi:10.1016/j.gpb.2012.09.003.

Ding, L., Han, M. (2007). GW182 family proteins are crucial for microRNA-mediated gene silencing. Trends in Cell Biology, 17(8), 411–416. doi:10.1016/j.tcb.2007.06.003.

Dowen, R. H., Pelizzola, M., Schmitz, R. J., Lister, R., Dowen, J. M., Nery, J. R., Dixon, J. E. (2012). Widespread dynamic DNA methylation in response to biotic stress. Proceedings of the National Academy of Sciences of the United States of America, 109(32), E2183–91. doi:10.1073/pnas.1209329109.

Duressa, D., Soliman, K., Taylor, R., Senwo, Z. (2011). Proteomic Analysis of Soybean Roots under Aluminum Stress. International Journal of Plant Genomics. doi:10.1155/2011/282531

Eamens, A. L., Agius, C., Smith, N. A., Waterhouse, P. M.,, Wang, M.-B. (2011). Efficient silencing of endogenous microRNAs using artificial microRNAs in *Arabidopsis thaliana*. Molecular Plant, 4(1), 157–170. doi:10.1093/mp/ssq061.

Eamens, A. L., Wang, M.-B. (2011). Alternate approaches to repress endogenous microRNA activity in *Arabidopsis thaliana*. Plant Signaling & Behavior, 6(3), 349–359. doi:10.4161/psb.6.3.14340.

Edwards, D., Batley, J., Snowdon, R. J. (2013). Accessing complex crop genomes with next-generation sequencing. TAG. Theoretical and Applied Genetics, 126(1), 1–11. doi:10.1007/s00122-012-1964-x.

Enright, A. J., John, B., Gaul, U., Tuschl, T., Sander, C., Marks, D. S. (2003). MicroRNA targets in Drosophila. Genome Biology, 5(1), R1. doi:10.1186/gb-2003-5-1-r1

FAO. Disponível em: <www.fao.org>. Acesso em: fevereiro 2013.

Farazi, T. A, Juranek, S. A, Tuschl, T. (2008). The growing catalog of small RNAs and their association with distinct Argonaute/Piwi family members. Development, 135(7), 1201–1014. doi:10.1242/dev.005629.

Folkes, L., Moxon, S., Woolfenden, H. C., Stocks, M. B., Szittya, G., Dalmay, T., Moulton, V. (2012). PAREsnip: a tool for rapid genome-wide discovery of small RNA/target interactions evidenced through degradome sequencing. Nucleic Acids Research, 40(13), e103. doi:10.1093/nar/gks277.

Franco-Zorrilla, J. M., Valli, A., Todesco, M., Mateos, I., Puga, M. I., Rubio-Somoza, I., Leyva, A. (2007). Target mimicry provides a new mechanism for regulation of microRNA activity. Nature Genetics, 39(8), 1033–1037. doi:10.1038/ng2079.

Gandikota, M., Birkenbihl, R. P., Höhmann, S., Cardon, G. H., 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. The Plant Journal, 49(4), 683–693. doi:10.1111/j.1365-313X.2006.02983.x.

Garcia-Ruiz, H., Takeda, A., Chapman, E. J., Sullivan, C. M., Fahlgren, N., Brempelis, K. J., Carrington, J. C. (2010). Arabidopsis RNA-Dependent RNA Polymerases and Dicer-Like Proteins in Antiviral Defense and Small Interfering RNA Biogenesis during Turnip Mosaic Virus Infection. Plant Cell, 22(2), 481–496. doi:10.1105/tpc.109.073056.

German, M. a, Pillay, M., Jeong, D.-H., Hetawal, A., Luo, S., Janardhanan, P., Kannan, V. (2008). Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends. Nature Biotechnology, 26(8), 941–946. doi:10.1038/nbt1417.

Ghildiyal, M., Zamore, P. D. (2009). Small silencing RNAs: an expanding universe. Nature Reviews. Genetics, 10(2), 94–108. doi:10.1038/nrg2504.

Lopez-Gomollon S., Mohorianu I., Szittya G., Moulton V., Dalmay T. (2012). Diverse correlation patterns between microRNAs and their targets during tomato fruit development indicates different modes of microRNA actions Planta, 236(6), 1875-1887. doi: 10.1007/s00425-012-1734-7.

Guleria, P., Mahajan, M., Bhardwaj, J., Yadav, S. K. (2011). Plant small RNAs: biogenesis, mode of action and their roles in abiotic stresses. Genomics, Proteomics & Bioinformatics, 9(6), 183–199. doi:10.1016/S1672-0229(11)60022-3.

Guo, N., Ye, W.-W., Wu, X.-L., Shen, D.-Y., Wang, Y.-C., Xing, H., Dou, D.-L. (2011). Microarray profiling reveals microRNAs involving soybean resistance to *Phytophthora sojae* Genome, 54(11), 954–958. doi:10.1139/g11-050.

Guo, X., Zhang, Z., Gerstein, M. B., Zheng, D. (2009). Small RNAs originated from pseudogenes: cis- or trans-acting? PLoS Computational Biology, 5(7), e1000449. doi:10.1371/journal.pcbi.1000449.

Guzman, F., Almerão, M. P., Körbes, A. P., Loss-Morais, G., Margis, R. (2012). Identification of microRNAs from *Eugenia uniflora* by high-throughput sequencing and bioinformatics analysis. PloS One, 7(11), e49811. doi:10.1371/journal.pone.0049811.

Hackenberg, M., Huang, P.-J., Huang, C.-Y., Shi, B.-J., Gustafson, P., Langridge, P. (2012). A Comprehensive Expression Profile of MicroRNAs and Other Classes of Non-Coding Small RNAs in Barley Under Phosphorous-Deficient and Sufficient Conditions. DNA Research, epub 1–17. doi:10.1093/dnares/dss037.

Hale, C. J., Erhard, K. F., Lisch, D., Hollick, J. B. (2009). Production and processing of siRNA precursor transcripts from the highly repetitive maize genome. PLoS Genetics, 5(8), e1000598. doi:10.1371/journal.pgen.1000598.

Haussecker, D., Huang, Y., Lau, A., Parameswaran, P., Fire, A. Z., Kay, M. a. (2010). Human tRNA-derived small RNAs in the global regulation of RNA silencing. RNA, 16(4), 673–695. doi:10.1261/rna.2000810.

Hsieh, L.-C., Lin, S.-I., Shih, A. C.-C., Chen, J.-W., Lin, W.-Y., Tseng, C.-Y., Li, W.-H. (2009). Uncovering small RNA-mediated responses to phosphate deficiency in Arabidopsis by deep sequencing. Plant Physiology, 151(4), 2120–2132. doi:10.1104/pp.109.147280.

Huntzinger, E., Izaurralde, E. (2011). Gene silencing by microRNAs: contributions of translational repression and mRNA decay. Nature Reviews. Genetics, 12(2), 99–110. doi:10.1038/nrg293.

Imelfort, M., Edwards, D. (2009). De novo sequencing of plant genomes using second-generation technologies. Briefings in Bioinformatics, 10(6), 609–618. doi:10.1093/bib/bbp039.

Eckardt, N.A. (2005). MicroRNAs Regulate Auxin Homeostasis and Plant Development, 17(5), 1335–1338. doi: 10.1105/tpc.105033159.

Kang, M., Zhao, Q., Zhu, D., Yu, J. (2012). Characterization of microRNAs expression during maize seed development, BMC Genomics, 13(360), 1-11. doi:10.1186/1471-2164-13-360.

Kawashima, C.G., Yoshimoto, N., Maruyama-Nakashita, A., Tsuchiya, Y.N., Saito, K., Takahashi, H., Dalmay, T. (2009) Sulphur starvation induces the expression of microRNA-395 and one of its target. Plant Journal, 57(2),313–321 doi: 10.1111/j.1365-313X.2008.03690.x.

genes but in different cell types. Plant Journal 57:313–321Khraiwesh, B., Zhu, J.-K., Zhu, J. (2012). Role of miRNAs and siRNAs in biotic and abiotic stress responses of plants. Biochimica et Biophysica Acta, 1819(2), 137–148. doi:10.1016/j.bbagrm.2011.05.001.

Koboldt, D. C., Ding, L., Mardis, E. R., Wilson, R. K. (2010). Challenges of sequencing human genomes. Briefings in Bioinformatics, 11(5), 484–498. doi:10.1093/bib/bbq016.

Koornneef, M., Meinke, D. (2010). The development of Arabidopsis as a model plant. The Plant Journal, 61(6), 909–921. doi:10.1111/j.1365-313X.2009.04086.x.

Kozomara, A., Griffiths-Jones, S. (2011). miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Research, 39(Database issue), D152–7. doi:10.1093/nar/gkq1027.

Kulcheski, F. R., De Oliveira, L. F., Molina, L. G., Almerão, M. P., Rodrigues, F. A., Marcolino, J., Barbosa, J. F. (2011). Identification of novel soybean microRNAs involved in abiotic and biotic stresses. BMC Genomics, 12, 307. doi:10.1186/1471-2164-12-307.

Körbes, A. P., Machado, R. D., Guzman, F., Almerão, M. P., De Oliveira, L. F. V., Loss-Morais, G., Turchetto-Zolet, A. C. Pinheiro-Margis, M. Margis, R. (2012). Identifying conserved and novel microRNAs in developing seeds of *Brassica napus* using deep sequencing. PloS One, 7(11), e50663. doi:10.1371/journal.pone.0050663.

Langmead, B., Trapnell, C., Pop, M., Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology, 10(3), R25. doi:10.1186/gb-2009-10-3-r25.

Lee, J., Bilyeu, K. D., Shannon, J. G. (2007). Genetics and breeding for modified fatty acid profile in soybean seed oil. Journal of Crop Science Biotechnology, 10(4), 201–210.

Lee, Y. S., Shibata, Y., Malhotra, A., Dutta, A. (2009). A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). Genes & Development, 23(22), 2639–2649. doi:10.1101/gad.1837609.

Lewis, B. P., Burge, C. B., Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell, 120(1), 15–20. doi:10.1016/j.cell.2004.12.035.

Li, Haiyan, Dong, Y., Yin, H., Wang, N., Yang, J., Liu, X., Wang, Y. (2011). Characterization of the stress associated microRNAs in *Glycine max* by deep sequencing. BMC Plant Biology, 11, 170. doi:10.1186/1471-2229-11-170.

Li, Heng, Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16), 2078–2079. doi:10.1093/bioinformatics/btp352.

Li, Hui, Deng, Y., Wu, T., Subramanian, S., Yu, O. (2010). Misexpression of miR482, miR1512, and miR1515 increases soybean nodulation. Plant Physiology, 153(4), 1759–1770. doi:10.1104/pp.110.156950.

Lima, J. C. De, Loss-morais, G., Margis, R. (2012). MicroRNAs play critical roles during plant development and in response to abiotic stresses, Genetics and Molecular Biology, 4, 1069–1077. doi: 10.1590/S1415-47572012000600023.

Liu, Q., Chen, Y.-Q. (2012). The potential roles of microRNAs in molecular breeding. Methods in Molecular Biology , 877, 303–11. doi:10.1007/978-1-61779-818-4_23.

Mardis, E. R. (2008). Next-generation DNA sequencing methods. Annual Review of Genomics and Human Genetics, 9, 387–402. doi:10.1146/annurev.genom.9.081307.164359.

Mathesius, U., Djordjevic, M. A., Oakes, M., Goffard, N., Haerizadeh, F., Weiller, G. F., Singh, M. B., (2011). Comparative proteomic profiles of the soybean (*Glycine max*) root apex and differentiated root zone. Proteomics, 11(9), 1707–1719. doi:10.1002/pmic.201000619.

Mendoza, M. R.; Fonseca, G. C.; Morais, G. L.; Alves, R.; Bazzan, A. L. C.; Margis, R. (2012). RFMirTarget: A Random Forest Classifier for Human miRNA Target Gene Prediction. Lecture Notes in Computer Science, 7409, 97-108. doi: 10.1007/978-3-642-31927-3_9.

Meyers, B. C., Axtell, M. J., Bartel, B., Bartel, D. P., Baulcombe, D., Bowman, J. L., Cao, X. (2008). Criteria for annotation of plant MicroRNAs. The Plant Cell, 20(12), 3186–3190. doi:10.1105/tpc.108.064311

Mitra, R., Bandyopadhyay, S. (2011). MultiMiTar: a novel multi objective optimization based miRNA-target prediction method. PloS One, 6(9), e24583. doi:10.1371/journal.pone.0024583.

Molina, L.G., Fonseca, G.C.,Morais, G.L.,Oliveira, L.F.V.,Carvalho, J.B, Kulcheski F.R., Margis, R. Metatranscriptomic analysis of small RNAs present in soybean deep sequencing libraries. Genetics and Molecular Biology 35(1), 292-303. doi: 10.1590/S1415-47572012000200010.

Moxon, S., Schwach, F., Dalmay, T., Maclean, D., Studholme, D. J., Moulton, V. (2008). A toolkit for analysing large-scale plant small RNA datasets. Bioinformatics, 24(19), 2252–2253. doi:10.1093/bioinformatics/btn428.

Muers, M. (2011). Technology: Getting Moore from DNA sequencing. Nature Reviews, 12(9), 586. doi:10.1038/nrg3059.

Murad, A. M., Rech, E. L. (2012). NanoUPLC-MSE proteomic data assessment of soybean seeds using the Uniprot database. BMC Biotechnology, 12(82). doi:10.1186/1472-6750-12-82,

Ni, Z., Hu, Z., Jiang, Q., Zhang, H. (2012). Overexpression of gma-MIR394a confers tolerance to drought in transgenic *Arabidopsis thaliana*. Biochemical and Biophysical Research Communications, 427(2), 330–335. doi:10.1016/j.bbrc.2012.09.055.

Oda, Y., Huang, K., Cross, F. R., Cowburn, D., Chait, B. T. (1999). Accurate quantitation of protein expression and site-specific phosphorylation. Proceedings of the National Academy of Sciences of the United States of America, 96(12), 6591–6596.

Ohyanagi, H., Sakata, K., Komatsu, S. (2012). Soybean Proteome Database 2012: update on the comprehensive data repository for soybean proteomics. Frontiers in Plant Science, 3(110). doi:10.3389/fpls.2012.00110.

Ong, S.E. (2002). Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics. Molecular & Cellular Proteomics, 1(5), 376–386. doi:10.1074/mcp.M200025-MCP200.

Paszkiewicz, K., Studholme, D. J. (2010). De novo assembly of short sequence reads. Briefings in Bioinformatics, 11(5), 457–472. doi:10.1093/bib/bbq020

Phalan, B., Bertzky, M., Butchart, S. H. M., Donald, P. F., Scharlemann, J. P. W., Stattersfield, A. J., Balmford, A. (2013). Crop expansion and conservation priorities in tropical countries. PloS One, 8(1), e51759. doi:10.1371/journal.pone.0051759.

Pontes, O., Li, C. F., Costa Nunes, P., Haag, J., Ream, T., Vitins, A., Jacobsen, S. E. (2006). The Arabidopsis chromatin-modifying nuclear siRNA pathway involves a nucleolar RNA processing center. Cell, 126(1), 79–92. doi:10.1016/j.cell.2006.05.031.

Radwan, O., Liu, Y., Clough, S. J. (2011). Transcriptional analysis of soybean root response to *Fusarium virguliforme*, the causal agent of sudden death syndrome. Molecular Plant-Microbe Interactions, 24(8), 958–972. doi:10.1094/MPMI-11-10-0271.

Reinhart, B. J., Weinstein, E. G., Rhoades, M. W., Bartel, B., Bartel, D. P. (2002). MicroRNAs in plants. Genes & Development, 16(13), 1616–1626. doi:10.1101/gad.1004402.

Ritchie W, Rajasekhar M, Flamant S, Rasko JEJ (2009) Conserved Expression Patterns Predict microRNA Targets. PLoS Computational Biology 5(9): e1000513. doi:10.1371/journal.pcbi.1000513.

Rhoades, M. W., Reinhart, B. J., Lim, L. P., Burge, C. B., Bartel, B., Bartel, D. P. (2002). Prediction of plant microRNA targets. Cell, 110(4), 513–520. doi:10.1016/S0092-8674(02)00863-2.

Ronaghi, M. (1998). DNA Sequencing:A Sequencing Method Based on Real-Time Pyrophosphate. Science, 281(5375), 363–365. doi:10.1126/science.281.5375.363.

Ruby, J. G., Jan, C., Player, C., Axtell, M. J., Lee, W., Nusbaum, C., Ge, H. (2006). Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. Cell, 127(6), 1193–1207. doi:10.1016/j.cell.2006.10.040.

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D. L. (2010). Genome sequence of the palaeopolyploid soybean. Nature, 463(7278), 178–183. doi:10.1038/nature08670.

Severin, A. J., Woody, J. L., Bolon, Y.-T., Joseph, B., Diers, B. W., Farmer, A. D., Muehlbauer, G. J. (2010). RNA-Seq Atlas of *Glycine max*: a guide to the soybean transcriptome. BMC Plant Biology, 10(160). doi:10.1186/1471-2229-10-160.

Shamimuzzaman, M., Vodkin, L. (2012). Identification of soybean seed developmental stage-specific and tissue-specific miRNA targets by degradome sequencing. BMC Genomics, 13(310). doi:10.1186/1471-2164-13-310.

Sijen, T., Steiner, F. a, Thijssen, K. L., Plasterk, R. H. (2007). Secondary siRNAs result from unprimed RNA synthesis and form a distinct class. Science, 315(5809), 244–247. doi:10.1126/science.1136699.

Sobala, A., Hutvagner, G. (2011). Transfer RNA-derived fragments: origins, processing, and functions. Wiley interdisciplinary reviews. RNA, 2(6), 853–862. doi:10.1002/wrna.96

Song, Q.-X., Liu, Y.-F., Hu, X.-Y., Zhang, W.-K., Ma, B., Chen, S.-Y., Zhang, J.-S. (2011). Identification of miRNAs and their target genes in developing soybean seeds by deep sequencing. BMC Plant Biology, 11(5). doi:10.1186/1471-2229-11-5.

Subramanian, S., Fu, Y., Sunkar, R., Barbazuk, W. B., Zhu, J.-K., Yu, O. (2008). Novel and nodulation-regulated microRNAs in soybean roots. BMC Genomics, 9(160). doi:10.1186/1471-2164-9-160.

Sunkar, R., Zhu, J. (2004). Novel and Stress-Regulated MicroRNAs and Other Small RNAs from Arabidopsis. The Plant Cell, 16(8), 2001–2019. doi:10.1105/tpc.104.022830.

Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A., Conesa, A. (2011). Differential expression in RNA-seq: a matter of depth. Genome Research, 21(12), 2213–2223. doi:10.1101/gr.124321.111.

The Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature, 408(6814), 796–815. doi:10.1038/35048692.

Thomson, D. W., Bracken, C. P., Goodall, G. J. (2011). Experimental strategies for microRNA target identification. Nucleic Acids Research, 39(16), 6845–6853. doi:10.1093/nar/gkr330.

Vaistij, F. E., Elias, L., George, G. L., Jones, L. (2010). Suppression of microRNA accumulation via RNA interference in *Arabidopsis thaliana*. Plant Molecular Biology, 73(4-5), 391–397. doi:10.1007/s11103-010-9625-4.

Vazquez, F., Vaucheret, H., Rajagopalan, R., Lepers, C., Gasciolli, V., Mallory, A. C., Hilbert, J.-L. (2004). Endogenous trans-acting siRNAs regulate the accumulation of Arabidopsis mRNAs. Molecular Cell, 16(1), 69–79. doi:10.1016/j.molcel.2004.09.028.

Vinther, J., Hedegaard, M. M., Gardner, P. P., Andersen, J. S., Arctander, P. (2006). Identification of miRNA targets with stable isotope labeling by amino acids in cell culture. Nucleic Acids Research, 34(16), e107. doi:10.1093/nar/gkl590.

Voinnet, O. (2009). Origin, Biogenesis, and Activity of Plant MicroRNAs, Cell, 136(4), 669–687

Wang, X., Zhang, J., Li, F., Gu, J., He, T., Zhang, X., Li, Y. (2005). MicroRNA identification based on

sequence and structure alignment. Bioinformatics, 21(18), 3610–3614. doi:10.1093/bioinformatics/bti562.

Wang, Y., Li, P., Cao, X., Wang, X., Zhang, A., Li, X. (2009). Identification and expression analysis of miRNAs from nitrogen-fixing soybean nodules. Biochemical and Biophysical Research communications, 378(4), 799–803. doi:10.1016/j.bbrc.2008.11.140.

Wei, B., Cai, T., Zhang, R., Li, A., Huo, N., Li, S., Gu, Y. Q. (2009). Novel microRNAs uncovered by deep sequencing of small RNA transcriptomes in bread wheat (*Triticum aestivum* L.) and *Brachypodium distachyon* (L.) Beauv. Functional & Integrative Genomics, 9(4), 499–511. doi:10.1007/s10142-009-0128-9.

Wheeler G., Valoczi A., Havelda Z., Dalmay T. (2007). *In situ* detection of animal and plant microRNAs. DNA Cell Biology, 26(4), 251-255. doi:10.1089/dna.2006.0538.

Wong, C. E., Zhao, Y.-T., Wang, X.-J., Croft, L., Wang, Z.-H., Haerizadeh, F., Mattick, J. S. (2011). MicroRNAs in the shoot apical meristem of soybean. Journal of Experimental Botany, 62(8), 2495–2506. doi:10.1093/jxb/erq437.

Yang, J.-H., Li, J.-H., Shao, P., Zhou, H., Chen, Y.-Q., Qu, L.-H. (2011). StarBase: a database for exploring microRNA-mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data. Nucleic Acids Research, 39(Database issue), D202–9. doi:10.1093/nar/gkq1056.

Yang, Y., Chaerkady, R., Kandasamy, K., Huang, T.-C., Selvan, L. D. N., Dwivedi, S. B., Kent, O. A. (2010). Identifying targets of miR-143 using a SILAC-based proteomic approach. Molecular Biosystems, 6(10), 1873–1882. doi:10.1039/c004401f.

Yousef, M., Jung, S., Kossenkov, A. V, Showe, L. C., Showe, M. K. (2007). Naïve Bayes for microRNA target predictions-machine learning for microRNA targets. Bioinformatics, 23(22), 2987–2992. doi:10.1093/bioinformatics/btm484.

Yu, B., Yang, Z., Li, J., Minakhina, S., Yang, M., Padgett, R. W., Steward, R. (2005). Methylation as a crucial step in plant microRNA biogenesis. Science, 307(5711), 932–935. doi:10.1126/science.1107130.

Zeng, H. Q., Zhu, Y. Y., Huang, S. Q., Yang, Z. M. (2010). Analysis of phosphorus-deficient responsive

miRNAs and cis-elements from soybean (*Glycine max* L.). Journal of Plant Physiology, 167(15), 1289–1297. doi:10.1016/j.jplph.2010.04.017.

Zerbino, D. R., Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Research, 18(5), 821–829. doi:10.1101/gr.074492.107.

Zhang, B. H., Pan, X. P., Cox, S. B., Cobb, G. P., Anderson, T. a. (2006). Evidence that miRNAs are different from other RNAs. Cellular and Molecular Life Sciences, 63(2), 246–254. doi:10.1007/s00018-005-5467-7.

Zhang, Xiaoming, Zhao, H., Gao, S., Wang, W.-C., Katiyar-Agarwal, S., Huang, H.-D., Raikhel, N. (2011). Arabidopsis Argonaute 2 regulates innate immunity via miRNA393(*)-mediated silencing of a Golgi-localized SNARE gene, MEMB12. Molecular Cell, 42(3), 356–366. doi:10.1016/j.molcel.2011.04.010.

Zhang, Xiuren, Henriques, R., Lin, S.-S., Niu, Q.-W., Chua, N.-H. (2006). Agrobacterium-mediated transformation of *Arabidopsis thaliana* using the floral dip method. Nature Protocols, 1(2), 641–646. doi:10.1038/nprot.2006.97.

Zheng, Y., Li, Y.-F., Sunkar, R., Zhang, W. (2012). SeqTar: an effective method for identifying microRNA guided cleavage sites from degradome of polyadenylated transcripts in plants. Nucleic Acids Research, 40(4), e28. doi:10.1093/nar/gkr1092.

Zhou, X., Ren, L., Meng, Q., Li, Y., Yu, Y., Yu, J. (2010). The next-generation sequencing technology and application. Protein & Cell, 1(6), 520–536. doi:10.1007/s13238-010-0065-3.

Zhuang, F., Fuchs, R. T., Robb, G. B. (2012). Small RNA expression profiling by high-throughput sequencing: implications of enzymatic manipulation. Journal of Nucleic Acids. doi:10.1155/2012/360358.

APÊNDICE

Apêndice I

Guilherme Loss de Morais

Curriculum Vitae

Fevereiro/2013

Guilherme Loss de Morais

Curriculum Vitae

Dados pessoais

Nome Guilherme Loss de Morais
Nascimento 22/01/1983 - Porto Alegre/RS - Brasil
CPF 434.685.600-44

Formação acadêmica/titulação

2010 Doutorado em Biologia Celular e Molecular.
Universidade Federal do Rio Grande do Sul, UFRGS, Porto Alegre, Brasil
Título: microRNAs presentes em sementes de plantas oleógenas
Orientador: Rogerio Margis
Bolsista do(a): Coordenação de Aperfeiçoamento de Pessoal de Nível Superior

2008 - 2010 Mestrado em Programa de Pós-Graduação em Biologia Celular e Molecular.
Universidade Federal do Rio Grande do Sul, UFRGS, Porto Alegre, Brasil
Título: Radiação e diversidade molecular das Proteínas Inativadoras de Ribossomos (RIPs) de Ricinus communis L., Ano de obtenção: 2010
Orientador: Rogerio Margis
Bolsista do(a): Conselho Nacional de Desenvolvimento Científico e Tecnológico

2002 - 2007 Graduação em Ciências Biológicas.
Pontifícia Universidade Católica do Rio Grande do Sul, PUCRS, Porto Alegre, Brasil

---

Formação complementar

2006 - 2006 Extensão universitária em Monitoria Em Disciplina de Graduação.
Pontifícia Universidade Católica do Rio Grande do Sul, PUCRS, Porto Alegre, Brasil

2006 - 2006 Curso de curta duração em Biologia molecular, Genética e Biotecnologia.
Pontifícia Universidade Católica do Rio Grande do Sul, PUCRS, Porto Alegre, Brasil

2005 - 2005 Curso de curta duração em Delineamentos Experimentais em Biologia.
Pontifícia Universidade Católica do Rio Grande do Sul, PUCRS, Porto Alegre, Brasil

2005 - 2005 Extensão universitária em Monitoria Em Disciplina de Graduação.
Pontifícia Universidade Católica do Rio Grande do Sul, PUCRS, Porto Alegre, Brasil

2005 - 2005 Extensão universitária em Monitoria no Projeto Recreare.
Pontifícia Universidade Católica do Rio Grande do Sul, PUCRS, Porto Alegre, Brasil

2004 - 2004 Extensão universitária em Monitoria Em Disciplina de Graduação.
Pontifícia Universidade Católica do Rio Grande do Sul, PUCRS, Porto Alegre, Brasil

2004 - 2004 Curso de curta duração em Fronteiras da Biologia Celular.

Atuação profissional

1. Universidade Federal do Rio Grande do Sul - UFRGS

Vínculo institucional

2008 - Atual Vínculo: Bolsista , Enquadramento funcional: Aluno Pós-Graduação , Carga horária: 40, Regime: Dedicação exclusiva

Atividades

03/2008 - Atual Pesquisa e Desenvolvimento, Centro de Biotecnologia
Linhas de pesquisa:
Análise de Expressão gênica , Bioinformática de pequenos RNAs , Análise filogenética de famílias gênicas

2. Pontifícia Universidade Católica do Rio Grande do Sul - PUCRS

Vínculo institucional

2004 - 2006 Vínculo: estagiário bolsista , Enquadramento funcional: Outro (bolsista) , Carga horária: 20, Regime: Parcial

2003 - 2004 Vínculo: estagiário voluntário , Enquadramento funcional: Outro (estagiário voluntário) , Carga horária: 16, Regime: Parcial

2002 - 2004 Vínculo: monitor de exposição , Enquadramento funcional: Outro (aluno de graduação) , Carga horária: 27, Regime: Parcial

Atividades

04/2004 - 02/2006 Estágio, Faculdade de Biociências, Departamento de Biologia
Estágio:
Laboratóriode Biotecnologia Vegetal

04/2004 - 02/2006 Pesquisa e Desenvolvimento, Faculdade de Biociências, Departamento de Biologia
Linhas de pesquisa:
Cultura de tecidos vegetais, micropropagação, formação de multiplas brotações e enraizamento de brotos in vitro. , Cultura de bactérias fitopatgênicas, respostas sistêmicas de plantas à bactérias fitopatogênicas.

06/2003 - 03/2004 Estágio, Faculdade de Biociências, Departamento de Biologia
Estágio:
Laboratório de Biotecnologia Vegetal

03/2002 - 03/2004 Estágio, Museu de Ciências e Tecnologia
Estágio:
Monitoria na exposição do MCT

Linhas de pesquisa

1. Análise de Expressão gênica

Objetivos:Análise de expressão genica de proteínas inativadoras de ribossomos durante o desenvolvimento de sementes de mamona (ricinus communis)Análise de expressão gênica de genes envolvidos no metabolismo lipidico em mamona, soja, canola e pinhão-mansoAnálise de expressão de pequenos RNAs (small RNAs), focando em microRNAs e Transfer-RNA derived Fragments (tRFs)

2. Análise filogenética de famílias gênicas

Objetivos:Analisar a evolução de fitocistatinasAnalisar a evolução de proteínas inativadoras de ribossomos em plantas, utilizando mamona como modelo evolutivoAnalisar a evolução de diacilglicerol transferases em plantas, utilizando soja como modelo

3. Bioinformática de pequenos RNAs

Objetivos:Caracterizar in silico pequenos RNAs oriundos de sequenciamentos de alta eficiência (deep sequencing), como foco em microRNAs e Transfer-RNA derived Fragments (tRFs)

Projetos

Projetos de pesquisaProjetos de pesquisa

2010 - Atual Agroenergia e metabolismo lipídico em oleaginosas

Descrição: Identificação de genes vegetais com potencial para uso em processos envolvendo a produção de biocombustíveis e melhoramento de espécies vegetais passiveis de serem usados em projetos de agroenergia.
Situação: Em andamento Natureza: Projetos de pesquisa
Alunos envolvidos: Doutorado (2);
Integrantes: Guilherme Loss de Morais Andreia Carina Turchetto Zolet; Marcia Margis-Pinheiro; Rogerio Margis (Responsável); Alexandro Cagliari; ana paula korbes; felipe santos maraschin

2010 - Atual MicroRNAs e silenciamento gênico pós-transcricional em plantas

Descrição: Identificação do papel de diferentes miRNAs e de suas famílias na regulação pós-transcricional da expressão gênica em arroz, soja, outras oleaginosass e espécies nativas. Emprego da metodologia de RNAi na produção de plantas transgênicas em estudos de genômica funcional.
Situação: Em andamento Natureza: Projetos de pesquisa
Alunos envolvidos: Mestrado acadêmico (1); Doutorado (7);
Integrantes: Guilherme Loss de Morais Marcia Margis-Pinheiro; Rogerio Margis (Responsável); ana paula korbes; Ana Paula Christoff; Andréia Caverzan ; Franceli Rodrigues Kulcheski ; Frank Lino Guzman Escudero ; Guilherme Cordenonsi da Fonseca ; João Braga de Abreu Neto ; Vanessa Galli

2008 - Atual Proteinases cisteínicas de origem vegetal e seus inibidores: as fitocistatinas

Descrição: Estudo da atividade de proteinases cisteínicas no processamento de proteínas vegetais. Ênfase no processamento de RIPs em mamona. Estudo da diversidade evolutiva e identificação do perfil de expressão gênico em diferentes tecidos e situações de estresse biótico e abiótico e interação com as proteinases alvo.

Situação: Em andamento Natureza: Projetos de pesquisa
Alunos envolvidos: Mestrado acadêmico (1); Doutorado (1);
Integrantes: Guilherme Loss de Morais Marcia Margis-Pinheiro; Rogerio Margis (Responsável); Ana Paula Christoff

Producão

Produção bibliográfica

Artigos completos publicados em periódicos

1. **LOSS-MORAIS, GUILHERME**, TURCHETTO-ZOLET, ANDREIA CARINA, ETGES, MATHEUS, Cagliari, Alexandro, KÖRBES, ANA PAULA, MARASCHIN, FELIPE DOS SANTOS, MARGIS-PINHEIRO, MÁRCIA, MARGIS, ROGERIO
Analysis of castor bean ribosome-inactivating proteins and their gene expression during seed development. Genetics and Molecular Biology (Impresso). 2013.

2. **LOSS-MORAIS, GUILHERME**, WATERHOUSE, PETER M, Margis, Rogerio
Description of plant tRNA-derived RNA fragments (tRFs) associated with Argonaute and identification of their putative targets. BIOL DIRECT. , v.8, p.6 - , 2013.

3. GUZMAN, FRANK, ALMERÃO, MAURICIO P., KÖRBES, ANA P., **LOSS-MORAIS, GUILHERME**, Margis, Rogerio, RAHMAN, ABIDUR
Identification of MicroRNAs from *Eugenia uniflora* by High-Throughput Sequencing and Bioinformatics Analysis. Plos One. , v.7, p.e49811 - , 2012.

4. KÖRBES, ANA PAULA, MACHADO, RONEI DORNELES, GUZMAN, FRANK, ALMERÃO, MAURICIO PEREIRA, DE OLIVEIRA, LUIZ FELIPE VALTER, **LOSS-MORAIS, GUILHERME**, TURCHETTO-ZOLET, ANDREIA CARINA, Cagliari, Alexandro, DOS SANTOS MARASCHIN, FELIPE, Margis-Pinheiro, Marcia, Margis, Rogerio
Identifying Conserved and Novel MicroRNAs in Developing Seeds of *Brassica napus* Using Deep Sequencing. Plos One. , v.7, p.e50663 - , 2012.

5. MOLINA, LORRAYNE GOMES, CORDENONSI DA FONSECA, GUILHERME, **MORAIS, GUILHERME LOSS DE**, DE OLIVEIRA, LUIZ FELIPE VALTER, CARVALHO, JOSEANE BISO DE, KULCHESKI, FRANCELI RODRIGUES, Margis, Rogerio
Metatranscriptomic analysis of small RNAs present in soybean deep sequencing libraries. Genetics and Molecular Biology (Impresso). , v.35, p.292 - 303, 2012.

6. LIMA, JÚLIO CÉSAR DE, **LOSS-MORAIS, GUILHERME**, Margis, Rogerio
MicroRNAs play critical roles during plant development and in response to abiotic stresses. Genetics and Molecular Biology (Impresso). , v.35, p.1069 - 1077, 2012.

7. TURCHETTO-ZOLET, ANDREIA C, Maraschin, Felipe S, **DE MORAIS, GUILHERME L**, Cagliari, Alexandro, ANDRADE, CLAUDIA MB, Margis-Pinheiro, Marcia, Margis, Rogerio
Evolutionary view of acyl-CoA diacylglycerol acyltransferase (DGAT), a key enzyme in neutral lipid biosynthesis. BMC Evolutionary Biology (Online). , v.11, p.263 - , 2011.

8. CAGLIARI, A, Cagliari, Alexandro, MARGIS-PINHEIRO, M, **Loss, G**, MARIATH, JORGE ERNESTO DE ARAUJO, MARGIS, R, Mastroberti, Alexandra Antunes, de Araujo Mariath, Jorge Ernesto, Margis-Pinheiro, MÃ¡rcia, LOSS, GUILHERME, Margis, Rogério
Identification and expression analysis of castor bean (*Ricinus communis*) genes encoding enzymes from the

triacylglycerol biosynthesis pathway. Plant Science (Limerick). , v.179, p.499 - 509, 2010.

9. MARGIS-PINHEIRO, MÁRCIA, Margis-Pinheiro, Marcia, ZOLET, A, **Loss, G**, PASQUALI, G, MARGIS, R, Margis, Rogerio, ZOLET, ANDREIA CARINA TURCHETTO, LOSS, GUILHERME, PASQUALI, Giancarlo
Molecular evolution and diversification of plant cysteine proteinase inhibitors: New insights after the poplar genome. Molecular Phylogenetics and Evolution (Print). , v.49, p.349 - 355, 2008.

Trabalhos publicados em anais de eventos (resumo)

1. Loss, G, ETGES, M. F., CAGLIARI, A., MARGIS-PINHEIRO, M., MARGIS, R.
Phylogenetic and comparative analysis of castor bean type I and II RIPS In: II Simpósio Brasileiro de Genética Molecular de Plantas, 2009, Búzios, RJ.
Anais do II Simpósio Brasileiro de Genética Molecular de Plantas. , 2009.

2. ZOLET, A. C. T., Loss, G, MARGIS-PINHEIRO, M., MARGIS, R.
Análise evolutiva dos genes P5CS e indentificação de polimorfismo associado ao estresse hídrico em populações nativas de Schizolobium parahyba (Vell.) Blake(guapuruvu) através de marcadores moleculares (SNPs) In: XI Congresso Brasileiro de Fisiologia Vegetal, 2007, Gramado.
Anais do XI Congresso Brasileiro de Fisiologia Vegetal. , 2007.

3. **DALMAS, Fernando Rostirolla**, MORAIS, Guilherme Loss, ASTARITA, Leandro Vieira
Cultivo in vitro de segmentos caulinares de araucaria angustifolia (CONIFERAE) In: 56° Congresso Nacional de Botânica, 2005, Curitiba.
. , 2005.

Trabalhos publicados em anais de eventos (resumo expandido)

1. Loss, G, DALMAS, Fernando Rostirolla, ASTARITA, Leandro Vieira
Microestaquia in vitro de Araucaria angustifolia In: BAIRESBIOTEC2005, 2005, Buenos Aires.
Libro de resúmenes. , 2005.

Produção técnica

Programa de computador sem registro

1. MENDOZA, M. R., da Fonseca, G. C., **Loss-Morais, G.**, ALVES, R., BAZZAN, A. L. C., MARGIS, R.
RFMirTarget: A Random Forest Classifier for Human miRNA Target Gene Prediction, 2012

2. MENDOZA, M. R., **Loss-Morais, G.**, da Fonseca, G. C., OLIVEIRA, L. F. V., ALVES, R., BAZZAN, A. L. C., MARGIS, R.
FilterPrecursors: An alignment-based tool for the identification of potential pre-miRNAs, 2011

3. **Loss-Morais, G.**, MARGIS, R.
StemLooper - A simple tool for design stem-loop primers, 2011

Apêndice II

Capítulo de Livro

Título livro: BIOINFORMÁTICA: DA BIOLOGIA A FLEXIBILIDADE MOLECULAR

Título do capítulo: BIOINFORMÁTICA ASSOCIADA AO SEQUENCIAMENTO DE ALTA EFICIÊNCIA

Organizador: Hugo Verli

Autores: Guilherme Loss-Morais e Rogerio Margis

trabalho em processo de edição

BIOINFORMÁTICA ASSOCIADA AO SEQUENCIAMENTO DE ALTA EFICIÊNCIA

1 Resumo gráfico

2 Introdução

A obtenção da sequências de ácidos nucléicos (DNA e RNA) foi e é de imensa importância para o estudo e entendimento da função de genes e outros componentes de natureza derivada de ácidos nucléicos (i.e, transposons, heterocromatina, mRNA, rRNA, tRNA, pequenos RNAs, etc.).
Em paralelo ao desenvolvimento das técnicas de sequenciamento de DNA e RNA, surgiram os desafios para a análise das sequências, com uma necessidade de criação de estratégias (algoritmos) para analisar e caracterizar as sequências obtidas.
No presente capítulo serão apresentadas algumas das técnicas de sequenciamento de ácidos nucléicos e as ferramentas para análise de sequências.

2.1 O início do sequenciamento de ácidos nucléicos e as ferramentas de bioinformática

A metodologia tradicional de sequenciamento de ácidos nucléicos foi denominada de sequenciamento pelo método dideoxy, sequenciamento por terminação de cadeia, ou sequenciamento de Sanger, em referência a um de seus criadores (Figura 1) Este método se baseia na utilização do DNA molde, de um primer específico do início da região que se deseja ser sequenciado, desoxinucleosídeos trifosfatos (dNTPs) e didesoxinucleosídeos trifosfatos (ddNTPs), esses últimos não possuindo a hidroxila no carbono 3' da desoxiribose, que incorporados a sequência, impossibilitam a formação de novas ligações fosfodiéster, impedindo a DNA polimerase de prosseguir a síntese da fita complementar ao encontrar o ddNTP. O ddNTP era inicialmente marcado com fósforo radioativo sendo posteriormente utilizados grupamentos fluorescentes acoplados aos ddNTPs para detecção.
No princípio do desenvolvimento desta metodologia, para cada fragmento de DNA a ser sequenciado, eram necessário quatro reações em paralelo, referentes à combinação dos quatro didesoxinucleosídeos trifosfatos (dATP, dGTP, dCTP e dTTP), com cada um dos didesoxinucleosídeos trifosfato (ddATP, ddGTP, ddCTP, or ddTTP) adicionados em tubos separados. Ao término das reações, cada produto de reação, contendo os vários produtos de terminação do sequenciamento era resolvido por gel de policrilamida (Figura 2).
Com o contínuo aumento no número das sequências de DNA, os primeiros desafios de bioinformática associada ao sequenciado foram sendo colocados:

Como e onde guardar as sequências e a anotação das informações contidas nelas?

O National Center for Biotechnology Information – NCBI (www.ncbi.nlm.nih.gov) é um meta banco de dados desenvolvido em 1988 para disponibilizar acessos de trabalhos científicos (http://www.ncbi.nlm.nih.gov/pubmed), de sequências de ácidos nucléicos e proteínas anotadas de diversos organismos através do GenBank (http://www.ncbi.nlm.nih.gov/genbank) e de quaisquer

informação relevante para a biotecnologia.
O Genbank tem dobrado o seu tamanho a cada 18 meses, sendo que a versão ou *release* de agosto de 2012, continha aproximadamente 143 bilhões de nucleotídeos, relativos a mais de 156 milhões de sequências depositadas. Um banco de dados contendo um número imenso de sequências deve padronizar o formato das sequências depositadas neste.

Como padronizar as sequências de DNA obtidas?

O formato escolhido foi o FASTA (Figura 3), o qual se tornou ubíquo como representação de sequências de ácidos nucléicos e proteínas em bioinformática.

A partir de uma sequência de DNA obtida, como saber o que ela contém e representa (sequência codificante, íntron/éxon, etc.)?

O método mais fácil é utilizar uma ferramenta de busca por similaridade de sequências. As ferramentas mais comuns utilizam o algoritmo de Smith–Waterman, o qual efetua uma busca local entre as sequências. Este algoritmo, ao invés de procurar no total da sequência, compara segmentos de todo os possíveis tamanhos e aperfeiçoa a medida de similaridades para o maior tamanho possível entre a sequência depositada no banco de dados e a sequência de interesse. Essa característica é computacionalmente favorável, pois diminui muito a dimensionalidade do problema, ou seja, este tipo de procura despende menos recursos computacionais, consequentemente sendo mais rápida.

A primeira ferramenta de procura de sequências em um banco de dados foi a FASTA, que além de ser um formato de representação de sequências de ácidos nucléicos e proteínas, também é o nome de um programa de procura de sequências de proteínas e DNA, baseado na similaridade de sequências. Este foi desenvolvido por David Lipman e William Person em 1985, tendo como legado o formato FASTA anteriormente citado.
Atualmente a ferramenta de busca de sequências mais utilizada é o *Basic Local Alignment Search Tool* (*BLAST*). Essa foi desenvolvida por Stephen Frank Altschul e colaboradores em 1990 e assim como a ferramenta FASTA promove uma procura por sequências baseada na similaridade de sequências de forma local. Sua grande aceitação é devido principalmente à rapidez e baixo consumo de recursos computacionais, sendo a escolha de vários banco de dados, entre ele o NCBI (http://www.ncbi.nlm.nih.gov/blast).

2.2 Avanços da tecnologia de sequenciamento:

Em paralelo com o surgimento destas ferramentas de bioinformática, surgiram plataformas de sequenciamento automático, ainda baseadas no método Sanger, mas apresentando melhorias, tais como:
- reações utilizando marcações com nucleotídeos acoplados a fluoróforos e não mais a bases radioativas;
- eliminação da necessidade de efetuar-se quatro reações em paralelo, pois cada um dos quatro ddNTP sendo marcado com um fluoróforo diferente permite a migração de todos em paralelo;

- separação dos produtos das reações de sequenciamento em capilares, aferindo maior precisão e rapidez ao sequenciamento;

A automação do sequenciamento Sanger (Figura 4) deu início ao processo de execução de vários projetos genomas, entre eles o de *Arabidopsis thaliana, Danio rerio, Mus musculus e Homo sapiens,* tendo esse demorando 13 anos até seu término. Entretanto, mesmo com o uso de sequenciadores de DNA automáticos, houve a necessidade de um esforço conjunto de vários laboratórios, dispondo de sequenciadores trabalhando em paralelo (Figura 5).
Um fator limitante no processo de sequenciamento era o tamanho dos fragmentos que estes sequenciadores processavam, em média contendo 500 pares de base, sendo que um genoma completo possui um tamanho milhares de vezes maior.
Para pode lidar com esse fator limitante, foi desenvolvido uma metodologia de fragmentação randômica do DNA, denominada de *DNA shotgun* (Figura 6). A metodologia de sequenciamento de DNA *shotgun* potencializou a utilização de sequeciadores Sanger para projetos genoma, contudo um novo desafio de bioinformática surgiu:
Como montar e ordenar os fragmentos sequenciados? Como reconstituir a sequência de DNA original a partir dos fragmentos gerados?

No sequenciamento por *shotgun*, "*N*" pequenos fragmentos são gerados randomicamente com um tamanho de "*l*" pares de bases de comprimento que são determinadas pelo sequenciamento. Todos estes fragmentos estão contidos no genoma, mas a ordem relativa deles é desconhecida. Além disso, existem outras complicações.

Cada fragmento sequenciado torna-se conhecido, mas a sua orientação no genoma é desconhecida. Ou seja, há 50% de chance de um fragmento sequenciado ser senso (5'→ 3') e 50% de ser anti-senso (3'→ 5'). Outro fator complicador advém do fato de muitas sequências possuírem regiões exatamente iguais ou com alto grau de sobreposição parcial, conhecidas como regiões repetitivas (comuns em telômeros, por exemplo). Se uma região repetitiva for muito grande, não será possível posicionar uma região repetitiva corretamente em relação às demais (Figura 7).

A maneira encontrada para tentar resolver esse problema era usar as informações de sobreposição entre o extremo da direita de um fragmento com o extremo da esquerda de outro.
Um dos programas mais utilizados para essa montagem (*assembling*) de fragmentos (*contigs*) e montagens de contigs (unitig) foi o *Phragment Assembly Program* ou simplesmente *Phrap* (http://www.phrap.org/).

2.2.1 Sequenciamento de segunda geração

O sequenciamento de segunda geração ou de alta eficiência se difundiu na metade dos anos 2000, com o objetivo de aumentar a quantidade de fragmentos sequenciado e diminuir o custo por sequenciamento. Existem várias plataformas e metodologias de sequenciamento, contudo, no presente capítulo serão abordadas somente algumas plataformas e as tecnologias as quais estas são baseadas. Algumas aplicações do sequenciamento de alta eficiência estão demonstradas na Tabela 1. Nas plataformas de sequenciamento de alta eficiência ocorre a criação de bibliotecas de sequenciamento, as quais são coleções de fragmentos (*reads*) de uma genoma (DNA) ou transcriptoma (RNA).

2.2.1.1 Pirosequenciamento

Está técnica se baseia no princípio de sequenciamento por síntese, onde conforme ocorre polimerização da cadeia complementar de DNA há emissão de luz captada por um receptor. Está técnica foi a primeira a ser incorporada em uma plataforma comercial, chamada de pirosequenciador 454.
Nessa plataforma cada fragmento é ligado a um adaptador que por sua vez são ligados a esferas (*beads*) de 28 μM. Os adaptadores são utilizados como região de hibridização de oligonucleotídeos iniciadores (*primers*), necessários para início da polimerização da cadeia de DNA por PCR. A técnica se baseia na utilização das enzimas ATP sufurilase, luciferase, apirase e na emissão luz originado por atividade enzimática.
Cada ciclo de pirosequenciamento inicia com a adição de um único tipo de nucleotídeo na cadeia por uma DNA polimerase, resultando na liberação uma molécula de pirofosfato (Ppi), o qual, junto com persulfato de amônio é convertido em ATP pela enzima ATP sufurilase. O ATP é então utilizado juntamente com a luciferina pela enzima luciferase. O resultado dessa última reação é oxi-luciferina e emissão de luz, a qual é captada por uma câmera CCD. Após cada ciclo, os nucleotídeos não polimerizados na cadeia são degradados pela enzima apirase. Em seguida repete-se o processo, mas com um tipo de nucleotídeo distinto do anterior. A técnica de pirosequenciamento está esquematizada na Figura 8.

O pirosequenciamento possui a limitação de não ser fidedigna no sequenciamento de regiões homopolimétricas, ou seja, o sequenciamento de um fragmento com muitas repetições de um mesmo nucleotídeo pode resultar em uma má interpretação do sinal luminoso pela câmera CCD induzindo erros na sequência, geralmente por inserção/deleção de nucleotídeos nesta.

2.2.1.2 Sequenciamento em plataformas Ilumina (Solexa, Genome Analyzer, HiSeq 2000 e MiSeq)

Esta tecnologia se baseia na utilização de adaptadores ancorados em uma placa que são utilizados como oligonucleotídeos iniciadores para uma PCR dos fragmentos a serem sequenciados. Ao fim da reação cada fragmento será representado inúmeras vezes, formando "Colônias de DNA". Este fato é importante, pois cada fragmento é sequenciado inúmeras vezes, diminuindo a taxa de erro de sequenciamento (Figura 9).
As reações de sequenciamento destas plataformas se baseiam na utilização de nucleotídeos terminadores de cadeia reversíveis ligados a fluoróforos distintos (um diferente para cada nucleotídeo). Conforme a reação de PCR inicia, são adicionados os terminadores de cadeia reversíveis, que interrompem a PCR, permitindo o sinal de fluorescência ser captado por uma câmera CCD. Após, os nucleotídeos terminadores não ligados são retirados da reação e aqueles ligados ao fragmento são desbloqueados, com a retirada do fluoróforo, permitindo a PCR continuar. Este processo se repete inúmeras vezes, fazendo com que cada nucleotídeo do fragmento a ser sequenciado seja resolvido um a um.

2.2.1.3 Sequenciamento pela tecnologia da plataforma Ion Torrent

Esta plataforma de sequenciamento se baseia na detecção de íons de Hidrogênio durante síntese de

DNA, ou seja, esta plataforma tem a capacidade de medir alterações no pH da reação, aferindo a sequência do fragmento. A incorporação de dNTPs na cadeia em polimerização depende da formação de ligações covalentes e da liberação de pirofosfato e um íon carregado de Hidrogênio. Cada nucleotídeo é adicionado separadamente na reação, logo, se este for complementar ao fragmento em dada posição, resultará na liberação do íon de Hidrogênio e sequenciado alteração de pH, a qual é captada por um detector. Após, um novo ciclo se repete, com um nucleotídeo diferente, repetindo-se por várias vezes, ocasionando, ao término da reação, na sequência do fragmento de interesse. Uma representação esquemática deste tipo de tecnologia de sequenciamento se encontra na Figura 10.

2.2.2 Sequenciamento de Terceira geração

As plataformas mais avançadas de sequenciamento se enquadram nesta geração, que possuem como base a obtenção de dados em tempo real e não serem baseados em PCR.
Uma metodologia que se enquadra nesta categoria é a "*Single-molecule real-time*" (SMRT) a qual se utiliza de uma enzima modificada, a qual cliva fluoróforos ancorados nos nucleotídeos que são adicionados ao fragmentos de DNA a ser sequenciado. O sinal de fluorescência é captado por uma câmera e um 'filme' relativo ao sequenciamento, o qual é transmitido em tempo real ao usuário.

Outra metodologia de sequenciamento de terceira geração é a Nanopore, a qual utilizando um poro em escala nanométrica, similar aos canais encontrados nas membranas celulares.
O conceito central desta metodologia baseia-se na utilização de um nanoporo com uma molécula de DNA simples permeando este com um fluxo iônico contínuo passando pelo poro o que gera uma corrente detectável. Cada nucleotídeo adicionado promove uma alteração de corrente distinta, com isso conforme a fita complementar de DNA é polimerizada, diferentes alterações de correntes são criadas e detectas, revelando a sequência de DNA em tempo real. As principais vantagens da utilização desta metodologia é o tamanho do *read* obtido o qual pode ser maior que 5 mil pares de base (kpb) e a velocidade de 1 par de base por nanosegundo (1x10-9 segundos).

3 Análises de bioinformática associadas à análise de bibliotecas de sequenciamento de alta eficiência

A escolha de como será analisado um sequenciamento de alta eficiência, depende do objetivo da pesquisa. As alternativas de montagem de *contigs* e/ou mapeamentos de *reads* devem levar em consideração se existe informação prévia de genoma ou sequências disponíveis do organismo em estudo, ou seja, uma montagem de contigs de um organismo que já tem seu genoma montado e anotado somente fará sentido se estivermos procurando polimorfismos entre o genoma disponível e o que estamos analisando. Se não for este o objetivo do estudo, não há a necessidade de montar *contigs* novamente de um genoma já completo.

Antes de entrar na análise de ácidos nucléicos oriundos de sequenciamento de alta eficiência, é importante se familiarizar com os formatos de arquivos que a ferramentas geram, tais como FASTQ, SAM e BAM

FASTQ: O formato FASTQ foi desenvolvido pelo *Wellcome Trust Sanger Institute*, sendo uma variação

no formato FASTA, mas que contém informações de qualidade de sequência, descritos por caracteres ASCII – *American Standard Code for Information Interchange* (Tabela 2) a qual é utilizada pela maior parte da indústria de computadores para a troca de informações.
Usualmente este formato possui quatro linhas, onde a primeira linha começa com "@" seguido por um identificador e opcionalmente uma descrição. A segunda linha é a sequência propriamente dita, a terceira linha inicia com um "+" e opcionalmente pode conter o identificador e uma descrição da sequência. A quarta linha descreve a qualidade do sequenciamento de cada nucleotídeo da sequência, expresso por caracteres ASCII, como citado anteriormente.
As plataformas Illumina, apresentam pequenas diferenças quanto ao formato FASTQ supracitado, entre elas esta a primeira linha que apresenta informação do sequenciador e da reação que resultou aquela sequência e na fórmula de calcular a qualidade a partir dos caracteres da terceira linha.
Um exemplo de sequência no Formato FASTQ está descrita na Figura 11.

Existem duas formas de calcular a qualidade de uma sequência, sendo a primeira a desenvolvida pelo *Wellcome Trust Sanger Institute* (Figura 12ª) e a uma adaptada pelo grupo que desenvolveu a plataforma Solexa (Illumina). O valor "Q" ,relativo à qualidade, é obtido dada certa probabilidade ("p") de aquela base estar incorreta, a partir dos dados de fluorescência oriundos do sequenciamento (*base calling*). Ambas as fórmulas resultam em um resultado idêntico até um p = 0.05 (Q =13) (Figura 12C), desta forma o valor mínimo de Q utilizado nas análises de bioinformática é 13.

Algumas ferramentas de mapeamento de sequências, ou seja, ferramentas que procuram regiões semelhantes dos *reads* com uma sequência de referência resultam em arquivos tabulares com as informações do mapeamento. Entre os formatos conhecidos, um dos mais utilizados está o "*Sequence Alignment/Map format*" (SAM).

Arquivos de mapeamento neste formato são tabulados, onde cada coluna corresponde a informações sobre a sequência referência ou do *read*. O formato SAM possui um cabeçalho, o qual é opcional, iniciado pelo símbolo arroba (@), contendo duas linhas, a primeira contendo informações sobre a sequência de referência e a segunda sobre a sequência mapeada.
As demais linhas são relativas a informações de identificadores de sequência, posição da sequência mapeada, tamanho desta, número de "*mismatches*", orientação da sequência mapeada e qualidade desta. O formato SAM fornece todas a s informações de mapeamento e por consequência é extremamente grande, ocupando muito espaço no disco rígido do computador. Para contornar isso, há a possibilidade de transformação do formato SAM para um formato BAM o qual é mais compacto.

BOX
O "mismatch" é uma inacurácia que pode ocorrer entre o mapeamento do *read* com a sequência referência. Este parâmetro pode ser definido pelo usuário das ferramentas de mapeamento de DNA, podendo ser, geralmente de zero a três. O *mismatch* pode ser originado a partir de erros do sequenciamento ou mesmo polimorfismo entre as sequências comparadas.

3.1 Ferramentas para edição de arquivos de sequenciamento

A grande maioria dos programas para análise de sequências utiliza o sistema operacional Linux como

padrão, por isso ressaltamos que importante saber um pouco de execução de programas através do terminal ou *Shell.*

Uma distribuição de Linux muito estável, amigável e totalmente gratuita é o Ubuntu (http://www.ubuntu-br.org/). Um guia de iniciantes em Linux pode ser adquirido gratuitamente (http://www.baixaki.com.br/linux/download/ubuntu-guia-do-iniciante.htm), sendo de grande ajuda a usuários sem experiência em Linux.
Como os arquivos analisados são geralmente muito grandes, os programas utilizados são geralmente operados por um terminal, pois a utilização de um modo gráfico poderia consumir desnecessariamente recursos de processamento e memória RAM do computador, os quais não seriam utilizados para a análise.

Ao recebermos um sequenciamento, este geralmente vem no formato FASTQ, algumas vezes contendo além da sequência de interesse uma sequência relativa ao adaptador, necessário para o sequenciamento. Para retirar os adaptadores das sequências de interesse (*Clipping*) pode se utilizar a ferramenta FASTX, (http://hannonlab.cshl.edu/fastx_toolkit/download.html).
Após a instalação, indicada no manual, o *Clipping* pode ser efetuado com o seguinte comando:

fastx_clipper –a adaptador –i arquivo de entrada (FASTA/Q) –o arquivo de saída (FASTA/Q)

onde, “adaptador” se refere à sequência do adaptador, “arquivo de entrada” se refere ao arquivo a ser analisado e “arquivo de saída” se refere ao arquivo resultante dessa análise.

O arquivo ainda pode ter sequências com baixa qualidade (FASTQ < 13), para removê-las pode se executar o comando:

fastq_quality_filter –q N –p M –i arquivo de entrada (FASTA/Q) –o arquivo de saída (FASTA/Q)

“N” representa a qualidade mínima que permite não filtrar dada sequência.
“M” representa o número mínimo de porcentagem de nucleotídeos que devem apresentar a qualidade mínima na sequência

Caso exista a necessidade de transformar o arquivo FASTQ para o formato FASTA, executar o comando:

fastq_to_fasta –i arquivo de entrada (FASTQ) –o arquivo de saída (FASTA)

As plataformas de sequenciamento de alta eficiência podem resultar no sequenciamento de “N” vezes um mesmo fragmento, fato que as tornam muito úteis ao analisar um transcriptoma, pois infere a expressão do gene. Para obter a contagem de sequências repetidas, executar o comando:

fastx_collapser arquivo de entrada (FASTA/Q) –o arquivo de saída (FASTA/Q)
O formato SAM, anteriormente citado, possui todas as informações do mapeamento e por consequência ocupa muito espaço no disco rígido do computador. Para transformá-lo em um formato mais compacto,

pode se utilizar a ferramenta Samtools (http://samtools.sourceforge.net/), através do comando:

samtools –b –S arquivo de entrada.sam > arquivo de saída.bam

onde "arquivo de entrada.sam" é o arquivo no formato SAM e "arquivo de saída.bam" é o nome do arquivo em formato BAM.

Após essas pré-análises dos arquivos de sequenciamento de alta eficiência, esses podem ser utilizados para montagem de genomas ou em mapeamento de sequências. A primeira abordagem é utilizada para sequenciamento de uma espécie com genoma desconhecido, ou para sequenciamento *de novo*, o qual tem finalidade localizar polimorfismos e outras mutações.
A segunda abordagem é muito comum no estudo de transcriptoma, pois pode inferir a expressão de genes e identificar *splicing* alternativos.

3.2 Montagens de genomas utilizando a ferramenta Velvet

Esta análise é uma que consome muitos recursos computacionais, pois executa a tarefa mais complexa em bioinformática, sendo necessário um computador com configurações robustas para montagem de um genoma. Para ter uma idéia, um sequenciador de segunda geração pode gerar de 2-3 bilhões de *reads* com sequenciad de 100 cópias para cada sequência, resultando em um imenso quebra-cabeça para o computador resolver. De forma geral os algoritmos de montagem de genomas, procuram achar regiões na pontas dos *reads* (tamanho definidos pelo usuário) que possam ter sobreposição (Figura 13). Estes *reads* com sobreposição são posteriormente unidos em sequências contíguas (*contigs)* que por sua vez podem ser agrupados em *scaffolds*

3.2.1 Utilizando o Velvet

A escolha desta ferramenta é devido a esta ser gratuita e bastante robusta, mas novamente há a necessidade de ser utilizada em ambiente UNIX (Linux). E como ressaltado anteriormente o computador deverá ter muita memória RAM, pois ao tentar montar os fragmentos, *contigs* e *scaffolds*, os dados serão mandados para essa memória, que deve ser no mínimo de 12 Gb, mas muito provavelmente será necessário mais, principalmente para análises de genomas de eucariotos.A ferramenta está disponível no link (http://www.ebi.ac.uk/~zerbino/velvet/velvet_latest.tgz)

após a instalação, descrita no manual, execute os comandos:

velveth diretório_de_saída Kmer –formato_do_arquivo –tamanho_dos_reads arquivo_de_entrada

onde, "velveth" é o programa que reconhece os arquivos de entrada e produz dois arquivos (*readmaps* e Sequences) necessários para a segunda parte da montagem (velvetg). Estes arquivos ficarão dentro de uma pasta criada (diretório_de_saída), a qual poderá ter o nome que o usuário definir.
O "kmer" é um parâmetro numérico, referente a um padrão de repetições de nucleotídeos utilizado para cálculos estatísticos na ferramenta. Aconselha-se testar mais de um *kmer*, contudo com o cuidado que estas variações influenciam muito na desempenho da análise.
O formato do arquivo deve ser informado, onde o *default* é FASTA, mas a ferramenta também aceita

FASTQ.

“O tamanho_dos_reads” é o nome da opção que informa ao programa se os *reads* são pequenos (-short) ou longos (-long). Os tamanhos de reads podem variar de acordo com a plataforma de sequenciamento utilizada e alteram a forma que o algoritmo do programa opera, desta forma é mandatário informar o tamanho dos *reads*. O “arquivo_de_entrada” é o próprio arquivo contendo os *reads* de sequenciamento.

Em seguida um segundo programa será utilizado, o velvetg, o qual criará os contigs. Digite no terminal

velvetg arquivo_de_saída –cov_cutoff X –min_contig_lgth Y

“Arquivo_de_saída” se refere ao mesmo diretório criado anteriormente com a ferramenta velveth. O parâmetro “-cov_cutoff” requer um valor numérico (X) que informa o número mínimos de *reads* utilizados para cobertura na criação de *contigs*. Um número muito baixo ocasionará em bastantes *contigs*, mas com grande possibilidade de falso positivos, um número muito grande resultará em poucos *contigs*, somente os mais representativos serão montados. O “-min_contig_lgth” requer um valor numérico (Y) que informa o tamanho mínimo que o *contig* resultante deverá possuir. Dentro do diretório denominado “arquivo_de_saída” estará um arquivo denominados “contigs.fas”, contendo todos contigs criados pela ferramenta no formato FASTA.

3.3 Anotação de *contigs*

Estes *contigs* não possuirão uma anotação, ou seja, é desconhecida a informação se o *contig* montado é referente a uma região codificante, intrônica, ribossomal, intergênica, etc. Para isso podemos utilizar o BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi). Esta ferramenta disponível na web suporta análise aproximadamente 100 sequências distintas por análise. Outra opção é a ferramenta BLAST2GO (http://www.blast2go.com/b2ghome) que pode anotar milhares de sequências por análise. Esta ferramenta é em Java e pode ser executada em qualquer sistema operacional. O programa é baseado nas anotações funcionais do *Gene Ontology* (GO), sendo bastante intuitivo e possuindo modo gráfico. Contudo, para utilizar esse programa será necessário conexão com internet para acesso aos bancos de dados do NCBI e GO.

Para utilização do BLAST2GO, abra o programa, e direcione o cursor até “File” e após em “load FASTA File”, selecionando o arquivo de interesse.
Direcione o cursores até BLAST e após “Make BLAST”, selecionando o tipo de BLAST a ser utilizado (BLASTx, para nucleotídeos). Clique na seta posicionada na parte superior da tela, para iniciar a análise. Esta parte da análise pode demorar bastante para execução, dependendo do número de *contigs* a serem utilizados.
Ao encerrar esta etapa da análise direcione o cursor em “Statistics” e após “BLAST Statistics”. Esta janela mostrará as estatísticas de seu BLAST.
Direcione o seu cursor em “Mapping”, após em “GO-Mapping Step” e clique na seta posicionada na parte superior da tela, para iniciar a análise.
Direcione o seu cursor em “annotation” e após em “Run Annotation Step” e clique na seta posicionada na parte superior da tela, para iniciar a análise.

Ao encerrar esta etapa da análise direcione o cursor em “Statistics” e após “Annotation Statistics”. Esta janela mostrará as estatísticas de seu mapeamento

Essa ferramenta resultará em uma tabela contendo as informações de anotação para cada *contig* utilizado na análise.

3.4 Mapeamentos de *reads* utilizando o Bowtie

Outra abordagem que pode ser utilizada com um arquivo de sequenciamento é o mapeamento de *reads* (Figura 14) em uma região de referência, que pode ser uma região do DNA ou transcrito de RNA. A utilização de mapeamento de reads em DNA é importante para identificação de polimorfismos e mutações utilizados no estudo de populações e doenças, respectivamente. Contudo, o mapeamento de sequências é muito utilizado no estudo de transcriptoma, devido ao sequenciamento de alta eficiência resultar em milhões de reads que são utilizados como inferência de expressão gênica. A abundância pode ser diretamente relacionada à expressão gênica, onde quanto mais *reads* mapeados uma referência (transcrito) tiver, mais expresso ele pode ser considerado. Como exemplo, ao construir duas bibliotecas de sequenciamento de RNA (transcriptoma), uma relativa a um tratamento e outra ao controle. Comparando-se as abundâncias de *reads* de um mesmo transcrito entre as bibliotecas, podemos inferir se o transcrito em questão é responsivo ao tratamento, caso este apresente um aumento de *reads* na biblioteca de tratamento.
Uma característica desta abordagem é que devemos possuir um genoma ou transcriptoma anotado, o qual será utilizado como referência para o mapeamento.

Existem várias ferramentas disponíveis para mapeamento, entre elas a ferramenta Bowtie (http://bowtie-bio.sourceforge.net/index.shtml) destaca-se por ser gratuita e pelo fácil uso e pouco requerimento de recursos computacionais. De acordo com o grupo que desenvolveu a ferramenta é capaz de mapear 25 milhões de reads (35 nt de comprimento) por hora de análise, utilizando somente 2.2 GB de memória RAM.

A utilização da ferramenta também é bastante simples, onde inicialmente precisa-se formatar o arquivo contendo as sequências referências, através do comando

bowtie-build arquivo_referência.fas arquivo_referência.fas

Onde, “bowtie-build” formatará o arquivo referência (no formato FASTA) para criar o índice. Deve-se repetir o nome do arquivo de referência duas vezes, uma referente ao arquivo de entrada e outra relativa ao arquivo de saída.

Em seguida proceder o mapeamento propriamente dito através dos comandos

bowtie arquivo_referência.fas reads_para_mapear arquivo_de_saída.sam –S

Onde “bowtie” é o comando que executará a análise, utilizando o arquivo de referência previamente formatado (arquivo_referência.fas) e mapeará os *reads* (reads_para_mapear) resultando no arquivo de

saída no formato SAM, o qual é especificado pelo comando "-S". Opcionalmente, pode ser adicionado os comandos "- v" que designará o número de *mismatches* (0-3) e o comando "-f" quando os *reads* estiverem no formato FASTA.

O arquivo resultante será no formato SAM o qual poderá ser formatado para BAM como mencionado anteriormente.

3.5 Visualização de mapeamentos de reads

Tendo os *reads* mapeados na sequência referência pode ser necessário a visualização destes. Para isto existem ferramentas para tal propósito, como o Tablet (http://bioinf.scri.ac.uk/tablet/), o qual é multiplataforma, funcionando em MS Windows, Macintosh e Linux. A ferramenta é bastante útil, pois informações importantes, como cobertura de *reads* e sentidos destes em relação à sequência referência estão acessíveis podendo ser exportadas para uma tabela.

Após a instalação, o usuário pode visualizar o mapeamento de reads, ao iniciar o programa e clicar no ícone "open assembling". Uma segunda janela abrirá, e no campo superior o usuário deverá indicar o arquivo de mapeamento (BAM ou SAM). No campo inferior deverá conter o arquivo de referência (FASTA ou FASTQ), após clicar em "open" e aguarda os arquivos serem carregados no computador, esta etapa tende a demorar de acordo com o tamanho dos arquivos e com a quantidade de memória RAM disponível para a análise.
Após o carregamento dos arquivos, a ferramenta disponibilizará uma tabela, a esquerda da tela, contendo os nomes das sequências referências (*contigs*), tamanhos destas (*lenght*), quantidade de *reads* mapeados nesta e o número de *mismatches*.

No presente capítulo demonstramos uma visão global sobre a evolução dos métodos de sequenciamento, bem como ferramentas de análise. Estas áreas estão sob constante evolução, logo as metodologias, plataformas de sequenciamento e análises de bioinformática discutidas aqui, serão em breve superadas.

4 Conceitos chave

O sequenciamento de ácidos nucléicos revolucionou o entendimento da biologia, fornecendo informações de genes a genomas.

O advento de novas tecnologias de sequenciadores, ditas de alta eficiência popularizam ainda mais a utilização destas técnicas para caracterização de genes, genomas e transcriptomas dos organismos.

Para análises de bioinformática de arquivo oriundos de sequenciadores de alta eficiência, há necessidade de conhecimentos básico de UNIX (Linux) e terminal de comando (*Shell*)

5- Referências

Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Daniel R. Zerbino and Ewan Birney Genome Res. 2008. 18: 821-829 – doi:10.1101/gr.074492.107

Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.
Ben Langmead, Cole Trapnell, Mihai Pop and Steven L Salzberg Genome Biology 2009, 10:R25 – doi: 10.1186/gb-2009-10-3-r25

Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Ana Conesa, Stefan Götz, Juan Miguel García-Gómez, Javier Terol1, Manuel Talón and Montserrat Robles Bioinformatics 2005 21: 3674-3676- doi: 10.1093/bioinformatics/bti610

Legendas:

Figura 1: Frederick Sanger, em 1958.

Figura 2: Representação esquemática de um gel de policrilamida, utilizado para sequenciamento de DNA. PCRs que são bloqueadas no inicio da reação, resultam em fragmentos menores que migram mais rápido que os maiores.

Figura 3: Representação esquemática de uma sequência no formato FASTA. Em vermelho o sinal de ‘maior que’, o qual designa a linha do cabeçalho da sequência. Em verde a descrição da sequência e em azul a própria sequência. No detalhe, um exemplo da sequência da cadeia Kappa da imunoglobulina humana no formato FASTA.

Figura 4: Representação de um sequenciador Sanger automático. O fragmentos imersos em policrilamida em percorrem um fino capilar. Cada ddNTP é marcado com um fluoróforo distinto que ao ser excitado por uma fonte de luz (laser) emite fluorescência, posteriormente coletada por um detector. Os dados são interpretados por um computador que relaciona cada tipo de fluorescência como seu nucleotídeo.

Figura 5: Fotografia de sequenciadores automáticos Sanger, trabalhando em paralelo
.

Figura 6: Representação esquemática de sequenciamento baseado na técnica *shotgun*. O DNA é fragmentado mecânica ou quimicamente e posteriormente clonado em vetores e transformados em bactérias. A partir dos clones são construídas bibliotecas de sequenciamento, as quais podem ser utilizadas em sequenciadores automáticos Sanger.

Figura 7: Representação do problema de montagem de fragmentos em regiões repetitivas
.

Figura 8: Metodologia do pirosequenciamento. Os fragmentos de ácidos nucléicos são ancorados em *beads* e utilizados como moldes em uma PCR, o Ppi resultante desta reação é utilizado como substrato na primeira etapa de uma cadeia de enzimas, que resultam ao fim um sinal de luz, o qual pode ser capturado, resultando na identificação do nucleotídeo complementar a sequência amplificada por PCR.

Figura 9: Fluxograma de sequenciamento das plataformas Illumina.

Figura 10: Representação esquemática de um sequenciamento na plataforma Ion Torrent. Durante a PCR, somente bases complementares à sequência molde liberam um íon de Hidrogênio, o qual induz uma pequena oscilação no pH, captada por um sensor, revelando o nucleotídeo complementar ao molde.

Figura 11: Exemplo de uma sequência no formato FASTQ.

Figura 12: Obtenção da qualidade da sequência no formato FASTQ. A) Fórmula original desenvolvido pelo grupo Sanger. B) Fórmula modificada, aplicada em sequenciadores Illumina. C) Diferença de valores Q entre as duas fórmulas, onde ocorre sobreposição de valores até um Q=13.

Figura 13: Representação esquemática de uma montagem de genomas/transcriptomas.

Figura 14: Representação esquemática de um mapeamento de *reads* em uma sequência referência.

Tabela 1: Aplicações do sequenciamento de alta eficiência.

Tabela 2: Representação de uma Tabela ASCII, demonstrando os valores decimais e seus caracteres relativos.

Resumo gráfico

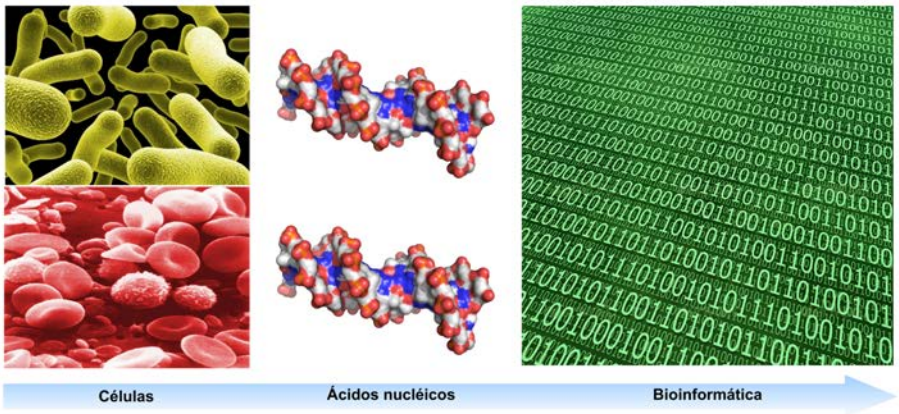


Figura 1

Figura 2

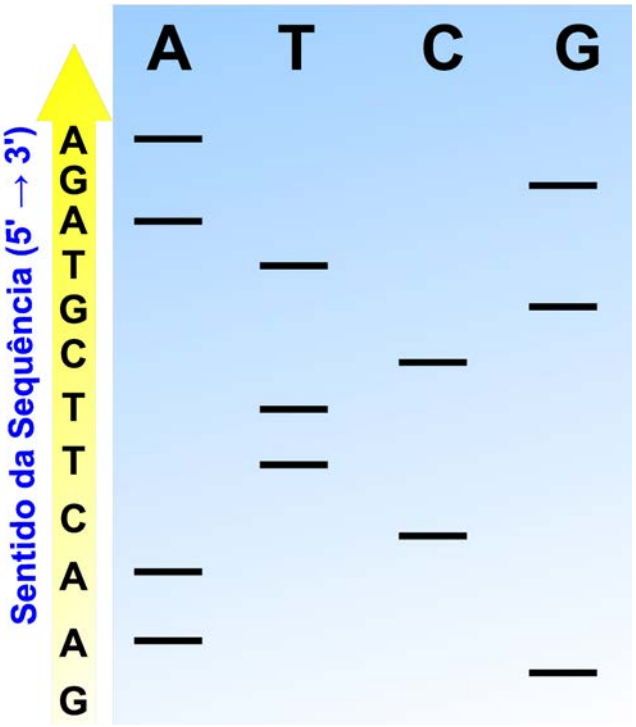


Figura 3

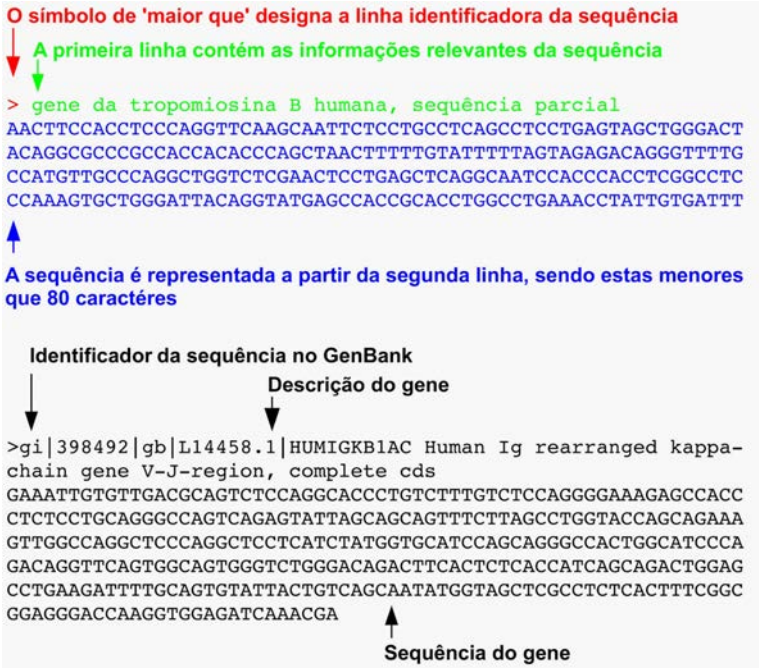


Figura 4

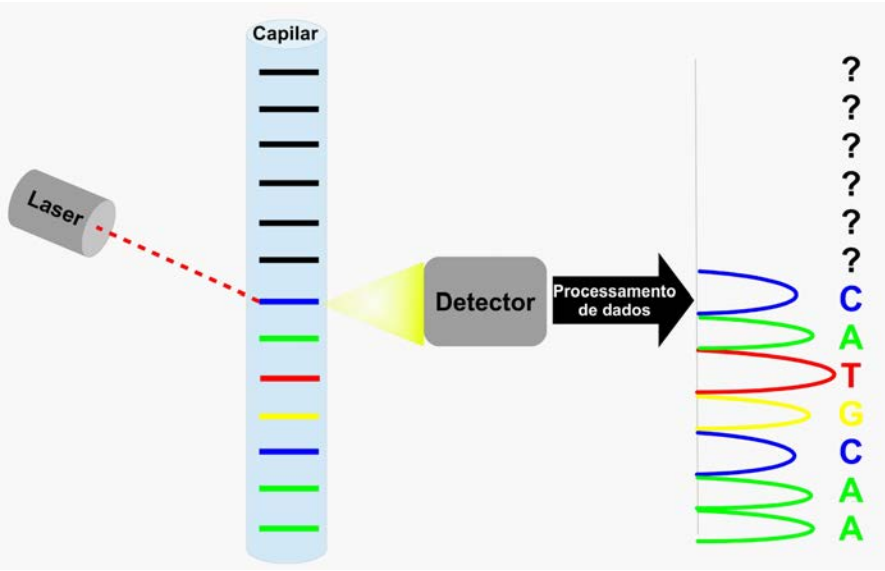


Figura 5

Figura 6

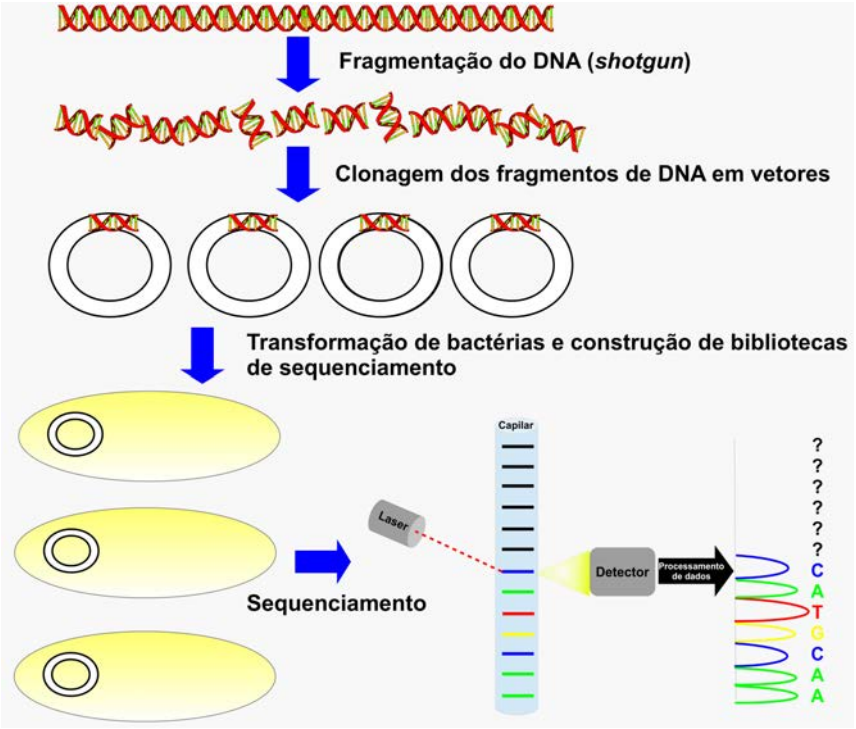


Figura 7

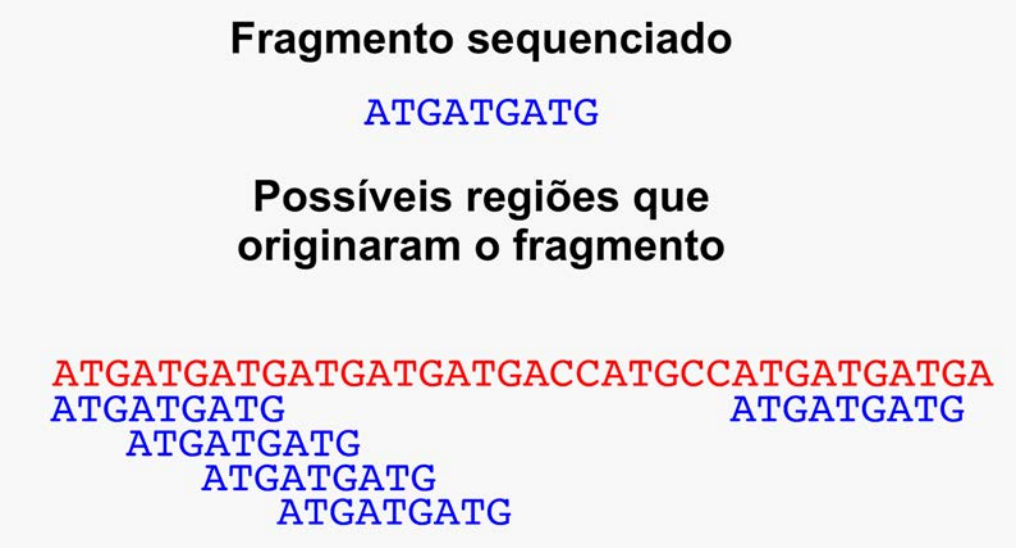


Figura 8

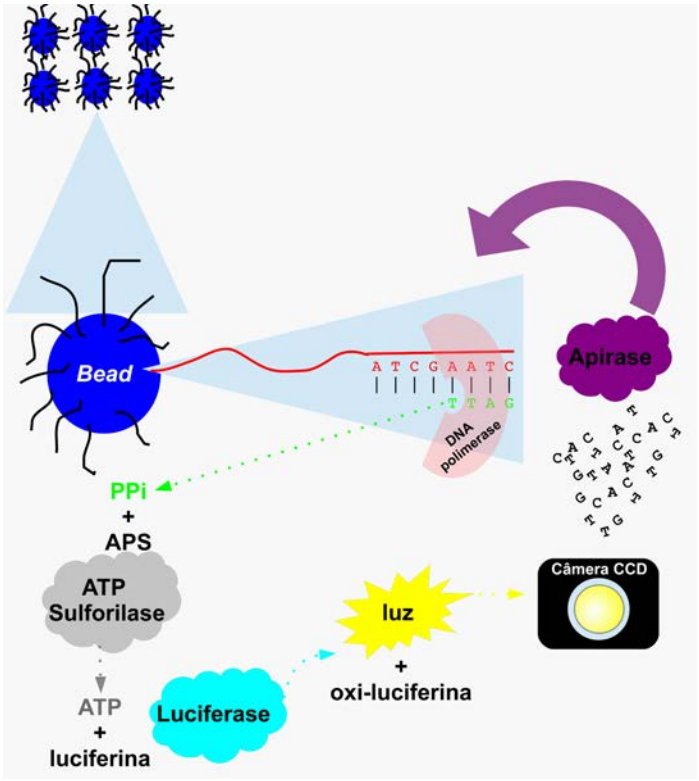


Figura 9

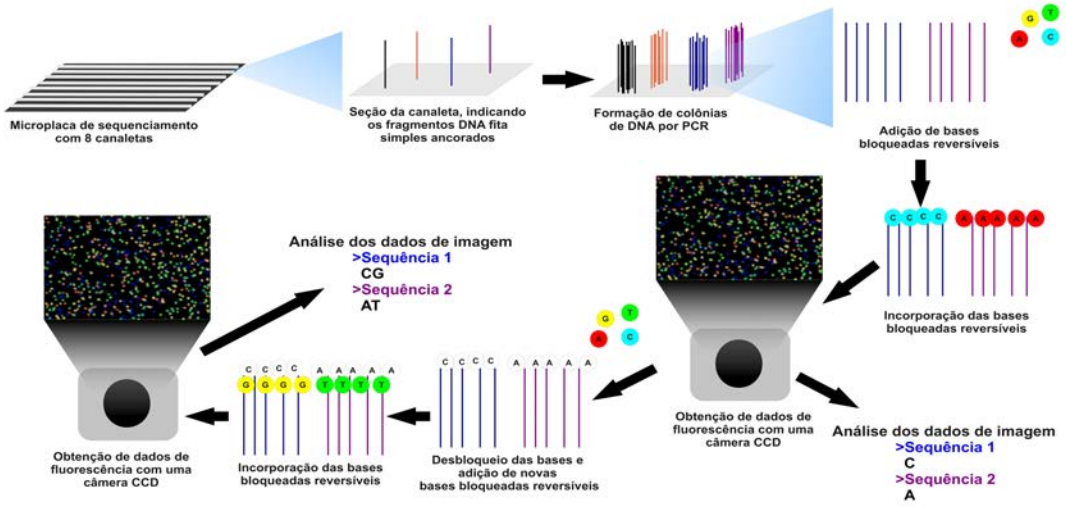


Figura 10

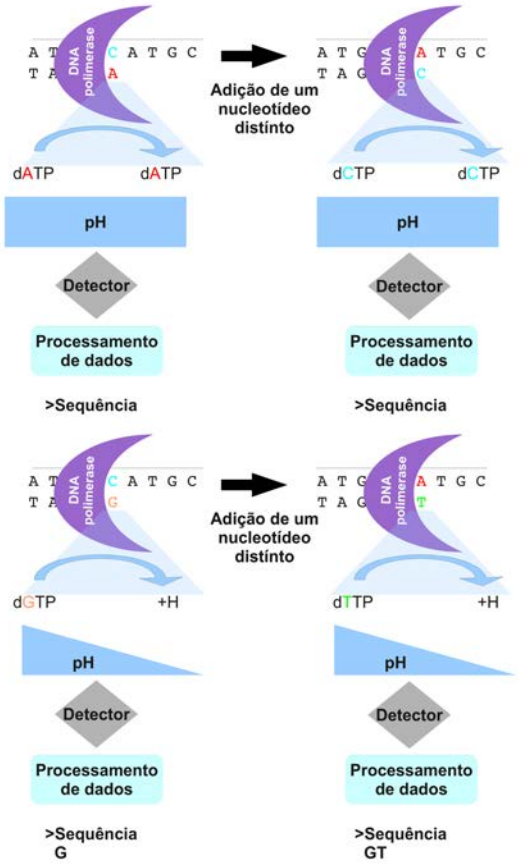


Figura 11

```
@Sequencia_1
CCTTGCTTGGAATTCGAGTTGGAGCACGGTTTCGTGTACCGTGAGCACAA
+
ggggggggcgfgcgcgf^ffccedeegggeefcfffdfff^_ccL`eaceU
```

Figura 12

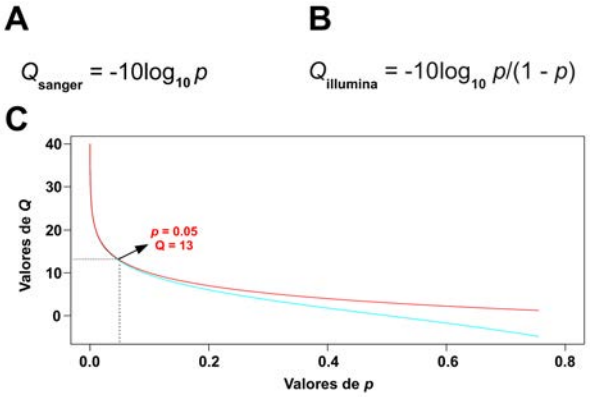


Figura 13

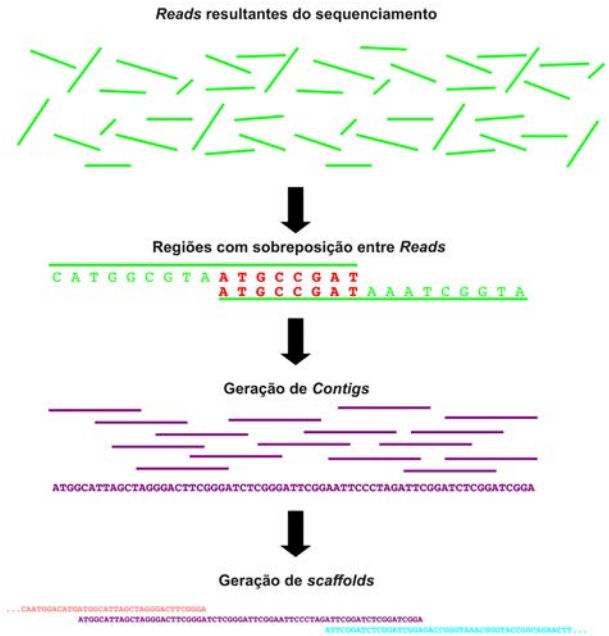


Figura 14

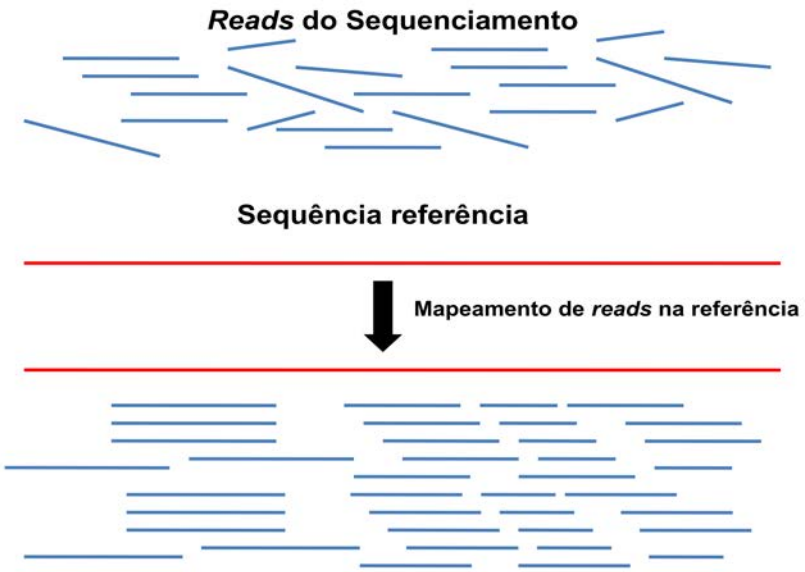


Tabela 1

Categoria	Exemplos de Aplicações
Ressenquenciamento genômico completo	Descoberta de polimorfismos e mutações em genomas de indivíduos
Sequenciamento 'Paired end'	Descobertas de variações adiquiridas e herdadas
Sequenciamento Metagenomico	Descoberta de flora comensal e infecciosa
Sequenciamento de Transcriptoma	Quantificação de expressão gênica e splicing alternativo
Sequenciamento de pequenos RNAs	Caracterização de microRNA siRNAs e outras classes de pequenos RNAs
Sequenciamento de DNA tratado com bisulfeto	Determinação de padrões de metilações de citosinas no DNA
Sequenciamento de Cromatina immunoprecipitada (ChIP-Seq)	Mapeamento em escala genômica de interações DNA-proteina
Sequeciamento de DNA tratado com Nuclease	Posicionamento de Nucleosomos

Tabela 2

Caracter	Decimal	Caracter	Decimal	Caracter	Decimal
NUL	0	+	43	V	86
SOH	1	,	44	W	87
STX	2	-	45	X	88
ETX	3	.	46	Y	89
EOT	4	/	47	Z	90
ENQ	5	0	48	[	91
ACK	6	1	49	\	92
BEL	7	2	50	]	93
BS	8	3	51	^	94
HT	9	4	52	_	95
LF	10	5	53	`	96
VT	11	6	54	a	97
FF	12	7	55	b	98
CR	13	8	56	c	99
SO	14	9	57	d	100
SI	15	:	58	e	101
DLE	16	;	59	f	102
D1	17	<	60	g	103
D2	18	=	61	h	104
D3	19	>	62	i	105
D4	20	?	63	j	106
NAK	21	@	64	k	107
SYN	22	A	65	l	108
ETB	23	B	66	m	109
CAN	24	C	67	n	110
EM	25	D	68	o	111
SUB	26	E	69	p	112
ESC	27	F	70	q	113
FS	28	G	71	r	114
GS	29	H	72	s	115
RS	30	I	73	t	116
US	31	J	74	u	117
Espaço	32	K	75	v	118
!	33	L	76	w	119
"	34	M	77	x	120
#	35	N	78	y	121
$	36	O	79	z	122
%	37	P	80	{	123
&	38	Q	81	|	124
'	39	R	82	}	125
(	40	S	83	~	126
)	41	T	84	DELETE	127
*	42	U	85		

TRABALHOS CIENTÍFICOS PUBLICADOS

Apêndice III

OPEN ACCESS Freely available online

PLOS ONE

Identifying Conserved and Novel MicroRNAs in Developing Seeds of *Brassica napus* Using Deep Sequencing

Ana Paula Körbes[1,2], Ronei Dorneles Machado[2], Frank Guzman[1], Mauricio Pereira Almerão[2], Luiz Felipe Valter de Oliveira[1], Guilherme Loss-Morais[2], Andreia Carina Turchetto-Zolet[1,2], Alexandro Cagliari[1], Felipe dos Santos Maraschin[1,3], Marcia Margis-Pinheiro[1,2], Rogerio Margis[1,2,4]*

1 PPGGBM, Departamento de Genética, Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre, Rio Grande do Sul, Brazil, **2** PPGBCM, Centro de Biotecnologia, Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre, Rio Grande do Sul, Brazil, **3** Departamento de Botânica, Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre, Rio Grande do Sul, Brazil, **4** Departamento de Biofísica, Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre, Rio Grande do Sul, Brazil

Abstract

MicroRNAs (miRNAs) are important post-transcriptional regulators of plant development and seed formation. In *Brassica napus*, an important edible oil crop, valuable lipids are synthesized and stored in specific seed tissues during embryogenesis. The miRNA transcriptome of *B. napus* is currently poorly characterized, especially at different seed developmental stages. This work aims to describe the miRNAome of developing seeds of *B. napus* by identifying plant-conserved and novel miRNAs and comparing miRNA abundance in mature versus developing seeds. Members of 59 miRNA families were detected through a computational analysis of a large number of reads obtained from deep sequencing two small RNA and two RNA-seq libraries of (i) pooled immature developing stages and (ii) mature *B. napus* seeds. Among these miRNA families, 17 families are currently known to exist in *B. napus*; additionally 29 families not reported in *B. napus* but conserved in other plant species were identified by alignment with known plant mature miRNAs. Assembled mRNA-seq contigs allowed for a search of putative new precursors and led to the identification of 13 novel miRNA families. Analysis of miRNA population between libraries reveals that several miRNAs and isomiRNAs have different abundance in developing stages compared to mature seeds. The predicted miRNA target genes encode a broad range of proteins related to seed development and energy storage. This work presents a comparative study of the miRNA transcriptome of mature and developing *B. napus* seeds and provides a basis for future research on individual miRNAs and their functions in embryogenesis, seed maturation and lipid accumulation in *B. napus*.

Citation: Körbes AP, Machado RD, Guzman F, Almerão MP, de Oliveira LFV, et al. (2012) Identifying Conserved and Novel MicroRNAs in Developing Seeds of *Brassica napus* Using Deep Sequencing. PLoS ONE 7(11): e50663. doi:10.1371/journal.pone.0050663

Editor: Michael Schubert, Ecole Normale Supérieure de Lyon, France

Received May 24, 2012; **Accepted** October 24, 2012; **Published** November 30, 2012

Funding: This work was supported by CAPES, CNPq, CNPq-Universal 472575/2011-2, Genoprot-CNPq-MCT 559636/2009-1, Agroestruturante-FAPERGS-FINEP. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: rogerio.margis@ufrgs.br

Introduction

Eukaryotic gene expression is regulated at the transcriptional and post-transcriptional levels. An important post-transcriptional mechanism that has recently been discovered is controlled by endogenous, noncoding small RNAs (sRNAs), primarily small interfering RNAs (siRNAs) and microRNAs (miRNAs) [1–4]. In plants, miRNA genes, called primary miRNAs (pri-miRNAs), are typically encoded in intergenic regions and are transcribed by RNA Polymerase II as long polyadenylated transcripts, similar to protein-coding genes [5]. These primary sequences contain an imperfect stem-loop structure that is recognized by DICER-Like1 (DCL1) for sequential cleavage, which converts the pri-miRNAs into the precursor sequences (pre-miRNAs) that are further processed to generate 18–24 nucleotide (nt)-long sequences called mature miRNAs [6]. The imperfect complementary strand to the most abundant miRNA is often called miRNA*, and both strands are originated from the 5p and 3p arms of the pre-miRNA hairpin structure. These sRNAs play critical roles during plant development, regulating a variety of processes, such as embryogenesis, seed germination, organ formation, and developmental timing and patterning [7–13]. The binding of the miRNA to mRNA targets leads to gene silencing by endonucleolytic cleavage or translational inhibition, depending on the degree of complementarity between the miRNA and its target transcript [14–18].

Brassica napus, known as Oilseed Rape, is the third most important edible oil crop worldwide (www.faostat.fao.org). During embryogenesis, *B. napus* seeds build up storage reserves in specific tissues. The vast majority of these reserves are made up of lipids (40–45%) and proteins (17–26%) that are almost exclusively stored in the cotyledons of the maturing embryo [19]. Biogenesis of oil bodies (lipid-containing structures) begins as early as the heart stage in embryogenesis and lipid accumulation rapidly increases during weeks 4–8 after anthesis [20,21]. These developmental stages are correlated with high synthetic lipid activity and a decline in the expression of genes coding for oil biosynthetic and glycolytic enzymes but not of the genes involved in the later steps of oil accumulation [22].

Apêndice IV

OPEN ACCESS Freely available online

PLOS ONE

Identification of MicroRNAs from *Eugenia uniflora* by High-Throughput Sequencing and Bioinformatics Analysis

Frank Guzman[1,2], Mauricio P. Almerão[2], Ana P. Körbes[1], Guilherme Loss-Morais[2], Rogerio Margis[1,2,3]*

1 PPGGBM, Departamento de Genética, Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre, Rio Grande do Sul, Brazil, **2** PPGBCM, Centro de Biotecnologia, Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre, Rio Grande do Sul, Brazil, **3** Departamento de Biofisica, Universidade Federal do Rio Grande do Sul - UFRGS, Porto Alegre, Rio Grande do Sul, Brazil

Abstract

Background: microRNAs or miRNAs are small non-coding regulatory RNAs that play important functions in the regulation of gene expression at the post-transcriptional level by targeting mRNAs for degradation or inhibiting protein translation. *Eugenia uniflora* is a plant native to tropical America with pharmacological and ecological importance, and there have been no previous studies concerning its gene expression and regulation. To date, no miRNAs have been reported in Myrtaceae species.

Results: Small RNA and RNA-seq libraries were constructed to identify miRNAs and pre-miRNAs in *Eugenia uniflora*. Solexa technology was used to perform high throughput sequencing of the library, and the data obtained were analyzed using bioinformatics tools. From 14,489,131 small RNA clean reads, we obtained 1,852,722 mature miRNA sequences representing 45 conserved families that have been identified in other plant species. Further analysis using contigs assembled from RNA-seq allowed the prediction of secondary structures of 25 known and 17 novel pre-miRNAs. The expression of twenty-seven identified miRNAs was also validated using RT-PCR assays. Potential targets were predicted for the most abundant mature miRNAs in the identified pre-miRNAs based on sequence homology.

Conclusions: This study is the first large scale identification of miRNAs and their potential targets from a species of the Myrtaceae family without genomic sequence resources. Our study provides more information about the evolutionary conservation of the regulatory network of miRNAs in plants and highlights species-specific miRNAs.

Citation: Guzman F, Almerão MP, Körbes AP, Loss-Morais G, Margis R (2012) Identification of MicroRNAs from *Eugenia uniflora* by High-Throughput Sequencing and Bioinformatics Analysis. PLoS ONE 7(11): e49811. doi:10.1371/journal.pone.0049811

Editor: Abidur Rahman, Iwate University, Japan

Received June 18, 2012; **Accepted** October 17, 2012; **Published** November 15, 2012

Funding: This research was supported by grants from FAPERGS (Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul) and CNPq (Conselho Nacional de Desenvolvimento Cientifico e Tecnológico). FG received a PhD fellowship from CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior). APK, MPA and RM have PNPD/CAPES, PDJ/CNPq Research/CNPq fellowships, respectively. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: rogerio.margis@ufrgs.br

Introduction

Eugenia uniflora is a tropical fruit tree native to South America. The shrubby tree produces edible cherry-like fruits, which are locally known as pitanga or the Brazilian cherry. This species belongs to the Myrtaceae family, which is characterized by the presence of tannins, flavonoids, monoterpenes and sesquiterpenes whose presence and concentration varies between specimens from different geographical locations [1–3]. Extracts from pitanga leaves contain interesting biological properties that have been reported in several studies, and pitanga juice is used in folk medicine as a diuretic, antirheumatic, antipyretic, antidiarrhetic and antidiabetic [4–9]. *E. uniflora* is also an important ecological model to study because it grows in areas of medium and large levels of rainfall and can also be found in different vegetation types and ecosystems [10]. The variation in the metabolite concentration and the adaptability to different environments observed in *E. uniflora* indicating that these are the result of the transcriptional regulation of many genes involved in metabolic and signaling pathways.

MicroRNAs (miRNAs) are small non-coding regulatory RNAs widely found in unicellular and multicellular organisms that act as regulators of gene expression at the post-transcriptional level on genes containing miRNA target sites [11]. Mature miRNAs are single-stranded RNA molecules of approximately 21 nucleotides (nt) in length processed from a precursor molecule (pre-miRNA) [12]. To regulate protein-coding genes, the mature miRNA binds with perfect or imperfect complementarity to sites in the 5′ or 3′ untranslated regions (UTR) or coding sequences (CDS) of genes, which leads to mRNA degradation or translation inhibition [13–14]. In plants, miRNAs have diverse biological functions and are involved in the regulation of optimal growth and development as well as other physiological processes, including abiotic and biotic stress responses [15]. Several studies showed that many miRNAs are conserved across different plant families [16–17]. However, family- and species-specific miRNAs that are expressed in lower

Apêndice V

Genetics and Molecular Biology, 35, 4 (suppl), 1069-1077 (2012)

www.sbg.org.br

Review Article

MicroRNAs play critical roles during plant development and in response to abiotic stresses

Júlio César de Lima[1,2,4], Guilherme Loss-Morais[1] and Rogerio Margis[1,3,4]

[1]Laboratório de Genomas e Populações de Plantas, Centro de Biotecnologia, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil.
[2]Laboratório de Fisiologia Vegetal, Departamento de Botânica, Instituto de Biologia, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil.
[3]Departamento de Biofísica, Instituto de Biologia, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil.
[4]Programa de Pósgraduação em Genética e Biologia Molecular, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil.

Abstract

MicroRNAs (miRNAs) have been identified as key molecules in regulatory networks. The fine-tuning role of miRNAs in addition to the regulatory role of transcription factors has shown that molecular events during development are tightly regulated. In addition, several miRNAs play crucial roles in the response to abiotic stress induced by drought, salinity, low temperatures, and metals such as aluminium. Interestingly, several miRNAs have overlapping roles with regard to development, stress responses, and nutrient homeostasis. Moreover, in response to the same abiotic stresses, different expression patterns for some conserved miRNA families among different plant species revealed different metabolic adjustments. The use of deep sequencing technologies for the characterisation of miRNA frequency and the identification of new miRNAs adds complexity to regulatory networks in plants. In this review, we consider the regulatory role of miRNAs in plant development and abiotic stresses, as well as the impact of deep sequencing technologies on the generation of miRNA data.

Keywords: miRNAs, development, abiotic stress, nutrients, deep sequencing.

MicroRNAs, Their Synthesis and Processing

Gene transcription is a key mechanism regulated by transcription factors and also by distinct small RNAs of 21 to 24 nucleotide of length that can act at the transcriptional and post-transcriptional level (Jamalkandi and Masoudi-Nejad, 2009; Voinnet, 2009). In plants, the regulation of gene expression mediated by small RNAs initiates after the generation of double stranded RNAs and/or single strand RNAs that are folded into stem-loop/hairpin structures in the cells. These are recognized by RNase III-like enzymes called Dicer-Like (DCL), processed into small interfering RNAs, and loaded into protein complexes (RISC) to effectuate gene silencing after the recognition of different complementary target RNAs and or DNA. Distinct biochemical pathways generate different classes of small RNAs: short interfering RNAs (siRNAs), piwi-interacting RNAs occurring exclusively in animals (piRNAs), trans-acting siRNAs (TAS), naturally anti-sense siRNAs (NAT) and microRNAs (miRNAs) (Ramachandran and Chen, 2008; Chen, 2009; Jamalkandi and Masoudi-Nejad, 2009; Liu and Paroo, 2010). TAS pathway - RNA Pol II transcribes TAS genes into a TAS precursor, which is recognized by a complementary siRNA and sliced by Argonaute (AGO) proteins into small RNA which serves as a template for RNA Dependent RNA Polymerases (RDR) to make dsRNAs. This siRNA duplex originated by Dicer-Like directs cleavage of the TAS precursor in *cis* or another target mRNAs in *trans*. MicroRNA pathway: a MIR gene is transcribed by RNA Pol II into a precursor pri-microRNA which is stabilized and cleaved by a protein complex composed of DCL and Hyponastic Leaves (HYL) into a pre-microRNA, which is further processed into a mature microRNA. The HUA Enhancer (HEN) methylates the resulting mature microRNA form in the 2'-hydroxy termini of both strands. This methylated mature form is exported to cytoplasm through HASTY protein (HST).

Send correspondence to: Rogerio Margis, Laboratório de Genomas e Populações de Plantas, Centro de Biotecnologia, Universidade Federal do Rio Grande do Sul, Caixa Postal 15005, 91501-970 Porto Alegre, Brazil. E-mail: rogerio.margis@ufrgs.br.

Apêndice VI

Turchetto-Zolet *et al. BMC Evolutionary Biology* 2011, **11**:263
http://www.biomedcentral.com/1471-2148/11/263

BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

Evolutionary view of acyl-CoA diacylglycerol acyltransferase (DGAT), a key enzyme in neutral lipid biosynthesis

Andreia C Turchetto-Zolet[1,2], Felipe S Maraschin[1], Guilherme L de Morais[2], Alexandro Cagliari[1], Cláudia MB Andrade[2], Marcia Margis-Pinheiro[1] and Rogerio Margis[1,2,3*]

Abstract

Background: Triacylglycerides (TAGs) are a class of neutral lipids that represent the most important storage form of energy for eukaryotic cells. DGAT (acyl-CoA: diacylglycerol acyltransferase; EC 2.3.1.20) is a transmembrane enzyme that acts in the final and committed step of TAG synthesis, and it has been proposed to be the rate-limiting enzyme in plant storage lipid accumulation. In fact, two different enzymes identified in several eukaryotic species, DGAT1 and DGAT2, are the main enzymes responsible for TAG synthesis. These enzymes do not share high DNA or protein sequence similarities, and it has been suggested that they play non-redundant roles in different tissues and in some species in TAG synthesis. Despite a number of previous studies on the DGAT1 and DGAT2 genes, which have emphasized their importance as potential obesity treatment targets to increase triacylglycerol accumulation, little is known about their evolutionary timeline in eukaryotes. The goal of this study was to examine the evolutionary relationship of the DGAT1 and DGAT2 genes across eukaryotic organisms in order to infer their origin.

Results: We have conducted a broad survey of fully sequenced genomes, including representatives of Amoebozoa, yeasts, fungi, algae, musses, plants, vertebrate and invertebrate species, for the presence of DGAT1 and DGAT2 gene homologs. We found that the DGAT1 and DGAT2 genes are nearly ubiquitous in eukaryotes and are readily identifiable in all the major eukaryotic groups and genomes examined. Phylogenetic analyses of the DGAT1 and DGAT2 amino acid sequences revealed evolutionary partitioning of the DGAT protein family into two major DGAT1 and DGAT2 clades. Protein secondary structure and hydrophobic-transmembrane analysis also showed differences between these enzymes. The analysis also revealed that the MGAT2 and AWAT genes may have arisen from DGAT2 duplication events.

Conclusions: In this study, we identified several DGAT1 and DGAT2 homologs in eukaryote taxa. Overall, the data show that DGAT1 and DGAT2 are present in most eukaryotic organisms and belong to two different gene families. The phylogenetic and evolutionary analyses revealed that DGAT1 and DGAT2 evolved separately, with functional convergence, despite their wide molecular and structural divergence.

Background

Triacylglycerols (TAGs), fatty acyl ester derivatives of glycerol, are a class of neutral lipids that represent the most important storage form of energy for eukaryotic cells [1,2]. In a number of plant species, TAGs are major storage lipids that accumulate in developing seeds, petals, pollen grains, and fruits [3]. Plant oils have been used for human consumption and have become important renewable resources as biofuels [4,5]. The enzymatic machinery for the formation of TAGs is located in the endoplasmic reticulum (ER). TAGs can then accumulate as oil droplets in the cytoplasm or in specialized oil storage bodies [6], which are generated through budding of the outer ER membrane [7]. A substantial part of TAG synthesis is performed by enzymes of the Kennedy pathway, which sequentially transfer acyl chains from acyl-CoAs to sn-1, -2 and -3 positions

* Correspondence: rogerio.margis@ufrgs.br
[1]Programa de Pós-Graduação em Genética e Biologia Molecular, Departamento de Genética, Universidade Federal do Rio Grande do Sul, Brazil
Full list of author information is available at the end of the article

Apêndice VII

Lecture Notes in Computer Science. 1ed.: Springer Berlin Heidelberg, 2012, v. 7409, p. 97-108 doi:10.1007/978-3-642-31927-3_9

RFMirTarget: A Random Forest Classifier for Human miRNA Target Gene Prediction

Mariana R. Mendoza[1], Guilherme C. da Fonseca[2], Guilherme L. de Morais[2], Ronnie Alves[3], Ana L.C. Bazzan[1], and Rogerio Margis[2]

[1] PPGC, UFRGS, P.O. Box 15064, Porto Alegre, RS, Brazil
{mrmendoza,bazzan}@inf.ufrgs.br
[2] PPGBCM, UFRGS, P.O. Box 15005, Porto Alegre, RS, Brazil
{guicf13,guilherme.loss}@gmail.com, rogerio.margis@ufrgs.br
[3] Vale Technological Institute Sustainable Development, Belém, PA, Brazil
ronnie.alves@vale.com

Abstract. MicroRNAs (miRNAs) are key regulators of eukaryotic gene expression whose fundamental role has been already identified in many cell pathways. The correct identification of miRNAs targets is a major challenge in bioinformatics. So far, machine learning-based methods for miRNA-target prediction have shown the best results in terms of specificity and sensitivity. However, despite its well-known efficiency in other classifying tasks, the random forest algorithm has not been employed in this problem. Therefore, in this work we present RFMirTarget, an efficient random forest miRNA-target prediction system. Our tool analyzes the alignment between a candidate miRNA-target pair and extracts a set of structural, thermodynamics, alignment and position-based features. Experiments have shown that RFMirTarget achieves a Matthew's correlation coefficient nearly 48% greater than the performance reported for the MultiMiTar, which was trained upon the same data set. In addition, tests performed with RFMirTarget reinforce the importance of the seed region for target prediction accuracy.

Keywords: miRNA, target prediction, random forest, gene regulation.

1 Introduction

MicroRNAs (miRNAs) are non-coding RNAs of ∼22 nucleotides in length that act as negative regulators of gene expression, thus playing an important role in gene regulation by targeting mRNAs with cleavage or translational repression [1]. The miRNA biogenesis is similar in both animals and plants. Mature miRNAs are formed from longer primary transcripts by two sequential processing steps mediated by a nuclear and a cytoplasmic RNase III endonuclease. In animals the responsible enzymes are Drosha and Dicer, respectively, while in plants both cleavages are performed by a Dicer homolog, DCL [1]. These cleavages generate a 60−70 nt stem-loop miRNA precursor (pre-miRNAs) and a mature miRNA duplex, respectively. Further, the mature miRNA duplex is assembled into an

M.C.P. de Souto and M.G. Kann (Eds.): BSB 2012, LNBI 7409, pp. 97–108, 2012.

TRABALHOS CIENTÍFICOS EM PREPARO OU SUBMETIDOS

Apêndice VIII

Identification of microRNAs and transcript targets in *Jatropha* seeds

Running title: MicroRNAs from *Jatropha* seeds

Vanessa Galli[1,2], Frank Guzman[3], Luiz Felipe Valter de Oliveira[1], Guilherme Loss-Morais[1], Ana Paula Körbes[3]Sérgio Delmar dos Anjos e Silva[2], , Márcia Margis-Pinheiro[3], Rogério Margis[1,3§]

[1] Center of Biotechnology and PPGBCM, Laboratory of Genomes and Plant Population, building 43431, Federal University of Rio Grande do Sul - UFRGS, P.O. Box 15005, CEP 91501-970, Porto Alegre, RS, Brazil.

[2] Brazilian Agricultural Research – EMBRAPA, P.O. Box 403, CEP 96010-971, Pelotas, RS, Brazil.

[3] PPGGBM at Federal University of Rio Grande do Sul - UFRGS, P.O. Box 15005, CEP 91501-970, Porto Alegre, RS, Brazil.

[§]Corresponding author

Rogerio Margis

Center of Biotechnology, Laboratory of Genomes and Plant Population, building 43431, Federal University of Rio Grande do Sul, Porto Alegre, RS, Brazil, CEP 91501-970.

Email: rogerio.margis@ufrgs.br

Tel: 55- 51 33087766

Apêndice IX

Title

FilterPrecursors: An alignment-based tool for theidentification of potential pre-miRNAs

Authors:

Mariana R. Mendoza[1], Guilherme Losss-Morais[2] , Guilherme Cordenonsi da Fonseca[2], Luiz Felipe. Valter de Oliveira[3], Ronnie Alves[1] , Ana Lucia C. Bazzan[1] , Rogerio Margis[2,3,§]

[1]Instituto informática Centro de Biotecnologia, [2]Centro de Biotecnologia, [3]Departamento de Genética, [4]Departamento de Biofísica, Universidade Federal do Rio Grande do Sul, Brasil. [§]Corresponding author

[§]Corresponding author:

rogerio.margis@ufrgs.br

Av. Bento Gonçalves, 9500. Prédio 43431- Campus do Vale - CxP. 15005. Porto Alegre, RS, Brasil. CEP 91501-970

Phone: (51) 3308-6087/3308-6074 Fax: (51) 3308-7309

Abstract

MicroRNAs (miRNAs) are short (21−24nt) non-coding RNAs which are known to play an important role in post-transcriptional gene expression regulation. Since the experimental identification of miRNAs and their precursors (pre-miRNAs) is difficult, the development of alternative approaches has become essential. Despite the wide application of machine learning methods, issues such as class imbalance and poor definition of features may result in pre-miRNAs misclassification. In the current work we introduce a novel and simpler approach to identify potential pre-miRNAs, as well as to refine predictions performed by popular tools, uniquely upon properties extracted from the alignment between high-throughput sequencing reads and candidate precursors. Our tool is freely available under the open

SCRIPTS DESENVOLVIDOS PARA AUXILIO NA CARACTERIZAÇÃO DE PEQUENOS RNAS

Apêndice X - FiltePrecursor
linguagem – perl 5
objetivo – analisar um arquivo de mapeamento no formato SOAP, identificando padrão de mapeamento correspondente ao microRNAs

```
# Programmed by Mariana Mendoza
# Finished:   04/05/2011
# Modified:   06/05/2011 : included tolerance to accept some reads mapped out of
the $maxColumns-columns profile
#               30/05/2011 : included parameters passing by command line

#               19/10/2011 : included the cut-off as a parameter, require at least
75% of reads mapped into the $maxColumns-columns profile, save         candidate
precursors with low frequence of mapped reads (< $cut-off) in a separate file
#!/usr/bin/perl -w
use strict;
use warnings;
use Time::HiRes;
use Getopt::Std;

my $start_time = [Time::HiRes::gettimeofday()];
my $usage = "
Usage:

 perl filterPrecursors.pl [soap output] [options]
  Process SOAP output, filtering pre-mRNAs into precursors and non-precursors
classes according to reads mapping profile over the pre-mRNAs sequences. The
optional parameters are:

  -c [integer]     Minimum number of mapped reads in the candidate precursors
(default 10)
 -o [integer] Maximum offset allowed for a single read (default 5)
 -t [integer] Maximum percentage of reads mapped out of colums (default 25)
 -p [integer] Maximum number of columns in the mapping profile (default 2)
";
# program variables
my $fileIN = shift or die $usage;
my %hashRef_Reads;          # hash: IDSeq => num_reads

my %hashRef_InitPosRef;      # hash of hashes: IDSeq => (Initial mapping position
in reference => Frequence)
my %hashPrecursors;         # hash: IDSeq => 1 of precursors
my %hashNotPrecursors;      # hash: IDSeq => 1 of non precursors
my %hashIgnoredQueries;      # hash: IDSeq => num_reads of not valid queries (less
than the minimum number of reads mapped on it)
my %options=();             # hash: algorithm's parameters set by user
my @arrayPositions;           # array of arrays: each index contains initial map
position and frequence
my @arrayTempMapInfo;         # temporary array with map position and respective
frequence
my $reference;              # temp variable for reading array
my $posInit;                # temp variable for reading array
```

```
my $ignoredCounter = 0;      # temp variable for counting candidate precursors with
less than 10 reads mapped
# user-configurable variables
my $cutoff = 10;               # minimum number of mapped reads in the candidate
precursors
my $offset = 5;              # maximum offset allowed for a single read
my $percentage = 25;         # maximum percentage of reads mapped out of columns
my $maxColumns = 2;          # maximum number of columns in the mapping profile
# check if parameters were assigned by command line. If not, keep standard values.
cutoff, offset, tolerance (%), profile (#columns)
getopts("c:o:t:p:",\%options);
if(exists $options{c}) {
    $cutoff = $options{c};

}
if(exists $options{o}) {
    $offset = $options{o};
}
if(exists $options{t}) {
    $percentage = $options{t};
}
if(exists $options{p}) {
    $maxColumns = $options{p};
}
print "\n>> Reading file $fileIN.\n";
# count the number of reads mapped in each reference sequence
open(FASTA, $fileIN);
while(<FASTA>)
{
    #      1       2      3      4      5      6      7      8      9     10
    if (/(\S+)\t(\S+)\t(\S+)\t(\S+)\t(\S+)\t(\S+)\t(\S+)\t(\S+)\t(\S+)\t(\S+)/) {
        chomp;
        if (!exists $hashRef_Reads{$8}) {
             $hashRef_Reads{$8} = 1;
        } else {
             $hashRef_Reads{$8}++;
        }
    }
}
close(FASTA);
print ">> Searching for precursors...\n";
# For all the reference sequences with a minimum number of reads mapped on it,
count the number of reads
#  per mapping start position. Finally, save on file .hifreq file reference
sequences that are potential precursors.
# For those reference sequences with less than $cutoff reads mapped on it, save it
to .lowfreq file.
open(FASTA, $fileIN);
open(OUT, ">$fileIN.hifreq");
open(LOWFREQ,">$fileIN.lowfreq");
open(LIST, ">$fileIN.list");
while(<FASTA>)
```

```
{     #     1      2      3      4      5      6      7      8      9     10
    if (/(\S+)\t(\S+)\t(\S+)\t(\S+)\t(\S+)\t(\S+)\t(\S+)\t(\S+)\t(\S+)\t(\S+)/) {
        chomp;
        if ($hashRef_Reads{$8} >= $cutoff) {
            # count number of reads mapped in this same initial position
            if (!exists $hashRef_InitPosRef{$8}{$9}) {
                $hashRef_InitPosRef{$8}{$9} = 1;
            } else {
                $hashRef_InitPosRef{$8}{$9}++;
            }
            # print in file only reference sequences with more than $cutoff reads
mapped on it
            print OUT "$1\t$2\t$3\t$4\t$5\t$6\t$7\t$8\t$9\t$10\n";
        }
        else {
            # save in file 'list' result of prediction for ALL queries, including
those with less than 10 reads mapped
            # print LIST "$8: is not a precursor\n";
              if (!exists $hashIgnoredQueries{$8}) {
                  $hashIgnoredQueries{$8} = $hashRef_Reads{$8};
              }
              print LOWFREQ "$1\t$2\t$3\t$4\t$5\t$6\t$7\t$8\t$9\t$10\n";
        }
    }
}
close(FASTA);
close(OUT);
close(LOWFREQ);
my $query;
# counter number of ignored precursors because of low mapping frequence
foreach $query (keys %hashIgnoredQueries ) {
    #print "$query\t$hashIgnoredQueries{$query}\n";
    $ignoredCounter++;
}
foreach $reference (keys %hashRef_InitPosRef ) {
    # Define a tolerance for reads not in the $maxColumns-columns profile
   my $tolerance = ($percentage/100) *  $hashRef_Reads{$reference};
    #print $hashRef_Reads{$reference}."\n";
   # For each possible precursor, save mapping profile into an array of arrays
   # each position of the array @arrayPositions has [Map Position, Frequence] info
   # Length of array @arrayPositions = number of initial mapping positions
   for $posInit (sort { $a <=> $b } keys %{ $hashRef_InitPosRef{$reference} } ) {
       $arrayTempMapInfo[0] = $posInit;
       $arrayTempMapInfo[1] = $hashRef_InitPosRef{$reference}{$posInit};
       push @arrayPositions, [@arrayTempMapInfo];
   }
   my $columns=0;    # number of columns in the mapping profile of the current
precursor
                     # there will be $maxColumns main frequences (e.g. columns) at
most!
   while ($columns < $maxColumns) {
       # check if mapping has $maxColumns-columns profile
```

```
       my $length = @arrayPositions;
       # if array is not empty...
       if ($length > 0){
           # find max frequence and its respective position
           my $max_Pos = (sort { $b->[1] <=> $a->[1] } @arrayPositions)[0]->[0];
           my $max_Freq = (sort { $b->[1] <=> $a->[1] } @arrayPositions)[0]->[1];
           # keep in array only elements outside interval [maxPos - var, maxPos +
var]
           @arrayPositions = grep { $_->[0] < ($max_Pos - $offset) || $_->[0] >
($max_Pos + $offset) } @arrayPositions;
       }
       $columns++;
   }
   my $length = 0;
   for my $i (0 .. $#arrayPositions) {
       $length = $length + $arrayPositions[$i]->[1];
   }
     #maybe we can modify this $tolerance to be 25% at most. This would imply to
have at least 75% of reads in the $maxColumns-columns profile
   print LIST $reference;
   # if there are still elements in the array, check if quantity is within
accepted tolerance
   if ($length > 0){
       if ($length eq 1 || $length < $tolerance) {
           print LIST ": is a precursor\n";
           if (!exists $hashPrecursors{$reference}) {
               $hashPrecursors{$reference} = 1;
           }
       }
       # if not, then is not a precursor!
        else {
           print LIST ": is not a precursor\n";
           if (!exists $hashNotPrecursors{$reference}) {
                  $hashNotPrecursors{$reference} = 1;
           }
       }
   }
   else {
       print LIST ": is a precursor\n";
       if (!exists $hashPrecursors{$reference}) {
           $hashPrecursors{$reference} = 1;
       }
   }
   # reset array of mapping positions
   @arrayPositions = ();
}
close(LIST);
# save in separated files reads mapped into valid and invalid precursors
open(IN, "$fileIN.hifreq");
open(OUT_PREC,">$fileIN.precursors");
open(OUT_NOTPREC,">$fileIN.nonprecursors");
while(<IN>)
```

```
{     #     1      2      3      4      5      6      7      8      9     10
    if (/(\S+)\t(\S+)\t(\S+)\t(\S+)\t(\S+)\t(\S+)\t(\S+)\t(\S+)\t(\S+)\t(\S+)/) {
        chomp;
        if (exists $hashNotPrecursors{$8}) {
            print OUT_NOTPREC "$1\t$2\t$3\t$4\t$5\t$6\t$7\t$8\t$9\t$10\n";
        }
        else {
            if (exists $hashPrecursors{$8}) {
                print OUT_PREC "$1\t$2\t$3\t$4\t$5\t$6\t$7\t$8\t$9\t$10\n";
            }
        }
    }
}
close(IN);
close(OUT_NOTPREC);
close(OUT_PREC);
# display main results
my $count_Prec = keys %hashPrecursors;
my $count_NoPrec = keys %hashNotPrecursors;
print ">> Results of search...\n";
print "\n## $ignoredCounter candidate precursors with less than $cutoff reads
mapped on it: not valid for mapping profile analysis.";
print "\n## $count_Prec real precursors and $count_NoPrec false precursors
identified.";
print "\n## Results saved in files:\n";
print "\t$fileIN.precursors\t\t\treal precursors\n";
print "\t$fileIN.nonprecursors\t\tfalse precursors\n";
print "\t$fileIN.list\t\t\tlist of all candidate precursor and their respective
possible class'\n";
`rm -r $fileIN.hifreq`;
my $diff = Time::HiRes::tv_interval($start_time);
print "\n## Elapsed time: $diff seconds\n\n";
```

Apêndice XI - StemLooper
linguagem –python 2
objetivo – projeta primers *stem loop* e forward utilizados na reações de cDNAe PCR de pequenos RNAs, respectivamente.

```
# -*- coding: utf-8 -*-
#programed by guilherme loss
import sys
from Bio.Seq import Seq
from Bio import SeqIO
from Bio.Alphabet import IUPAC
miRNA_query = sys.argv[1]
handle1 = open(miRNA_query, 'r')
for mirnas in SeqIO.parse(handle1, "fasta"):
     m_id = mirnas.id
     m_seq = mirnas.seq
     loop_name = mirnas.seq[-6:]
     m_revComp = mirnas.seq.reverse_complement()[:6]
```

```
     stlp_seq = 'gtcgtatccagtgcagggtccgaggtattcgcactggatacgac'
     stlp_primer = stlp_seq + m_revComp
     print m_id+'_Fwd'+'\t'+m_seq
     print m_id+'_stem_loop'+'\t'+loop_name+'\t'+stlp_primer
handle1.close()
```

Apêndice XII - seq_retrieved_by_index
linguagem –python 2
objetivo – a partir de um arquivo FASTA, retira sequencias delimitadas pela posição e senso

```
# -*- coding: utf-8 -*-
#programed by guilherme loss
import sys
from Bio import SeqIO
fl = sys.argv[1]
fasta_id = sys.argv[2]
range_5 = sys.argv[3]
range_3 = sys.argv[4]
mir = sys.argv[5]
sense = sys.argv[6]
handle = open(fl, 'r')
r5=int(range_5)
r3=int(range_3)
for seq_record in SeqIO.parse(handle, "fasta"):
 if fasta_id in seq_record.id:
     if sense == 'minus':
print
'>'+mir,seq_record.id,r5,r3,sense,'\n',seq_record.seq[r5:r3].reverse_complement()
     else:
         print '>'+mir,seq_record.id,r5,r3,sense,'\n',seq_record.seq[r5:r3]
handle.close()
```