



Facultad de Ciencias Bioquímicas y Farmacéuticas - Universidad Nacional
de Rosario

Tesis de Doctorado

**Estudios sobre la regulación de la expresión génica por
microARNs en plantas mediante estrategias
bioinformáticas**

Presentada por: Uciel Pablo Chorostecki

Rosario, Argentina

2016

Estudios sobre la regulación de la expresión génica por microARNs en plantas mediante estrategias bioinformáticas

Uciel Pablo Chorostecki

Licenciado en Ciencias de la Computación

Universidad Nacional de Rosario

Esta Tesis es presentada como parte de los requisitos para optar al grado académico de Doctor en Ciencias Biológicas, de la Universidad de Rosario y no ha sido presentada previamente para la obtención de otro título en esta u otra Universidad. La misma contiene los resultados obtenidos en investigaciones llevadas a cabo en el Instituto de Biología Molecular y Celular de Rosario (IBR-CONICET), dependiente de la Facultad de Cs. Bioquímicas y Farmacéuticas, durante el período comprendido entre el ?? y el ??, bajo la dirección del Dr. Javier Palatnik.

I would like to dedicate this thesis to my loving parents . . .

Índice general

Índice de figuras	vii
-------------------	-----

Índice de tablas	ix
------------------	----

1	Introducción	1
1.1	Silenciamiento génico y ARN pequeños en plantas	1
1.1.1	Variedades de ARN pequeños en plantas	1
1.2	miARNs en plantas	3
1.3	Biogénesis de miARNs	4
1.4	Regulación de la expresión génica por miARNs	5
1.4.1	Regulación por corte del ARN blanco	5
1.4.2	Regulación de la traducción por miARNs	5
1.4.3	Generación de ta-siRNAs y siRNAs secundarios	6
1.5	Procesos biológicos regulados por miARNs en plantas	6
1.6	Predicción de genes blanco de miARNs	7
1.6.1	Métodos para predicción de genes blanco de miARNs en animales .	7
1.6.2	Métodos para predicción de genes blanco de miARNs en plantas . .	8
1.6.3	identificación de genes blanco regulados por miARNs mediante técnicas de secuenciación de alto rendimiento	9
1.7	sPARE???	10
2	Objetivos	13
2.1	Objetivo general	13
2.2	Objetivos específicos	13
3	Métodos	15
3.1	Predicción de genes regulados por miARNs en plantas	15
3.1.1	MiARN consensos	15
3.1.2	Predicción de genes regulados por miARNs	15

3.2	comTAR: una herramienta para la predicción de genes blanco regulados por miARNs en plantas	17
3.2.1	MiARN y transcriptos	18
3.2.2	Búsqueda de genes blanco	18
3.2.3	Herramienta web y almacenamiento de datos	18
4	Resultados Capítulo 1	21
4.1	Introducción	21
4.2	Resultados	21
4.2.1	Predicción de genes regulados por miARNs.	21
4.2.2	comTAR: una herramienta para la predicción de genes blanco regulados por miARNs en plantas.	34
5	Estudios genómicos sobre la biogénesis de miARN en plantas	39
5.1	Introducción	39
5.2	Resultados	40
5.2.1	40
5.2.2	Estudio de la evolución y biogénesis de miARNs en plantas	43
6	Conclusiones	45
6.1	Short title	45
6.2	Short title	46
	References	47
	Appendix A Anexo	57

Índice de figuras

4.1	Estrategia	23
5.1	Mecanismos de procesamiento	42

Índice de tablas

4.1	miARNs y sus genes blanco en plantas	24
4.2	Detection of miRNA targets using different filters	28
A.1	Especies y base de datos utilizadas para la búsquedas de genes blanco de miARNs conservados	65
A.2	My caption	66
A.3	Oligonucleotide primers used for RT-qPCR	67
A.4	Oligonucleotide primers used for 5' RACE	67

Resumen

Los microARNs (o miARNs) son ARN no codificantes que regulan la expresión génica en animales y plantas y están implicados en procesos biológicos muy variables, como el desarrollo, la diferenciación y el metabolismo. Con un largo de aproximadamente 21 nucleótidos, los miARNs reconocen secuencias parcialmente complementarias en los ARNm blanco, provocando su corte o arresto de la traducción. Los miARNs han saltado rápidamente a la primera plana del interés de la comunidad científica como un nuevo nivel en el control de la expresión génica en eucariotas. Estudios recientes han puesto de manifiesto que los miARNs están estrechamente involucrados en distintas enfermedades de importancia. Los cálculos actuales consideran que cerca del 40% de los genes de humanos se encuentran regulados por miARNs.

Está generalmente aceptado que los miARNs en plantas tienen una extensiva complementariedad con sus genes blanco y su predicción por lo general se basa en el uso de parámetros empíricos deducidos de interacciones conocidos del par miARN-gen blanco. En este trabajo, primero desarrollamos una estrategia para la identificación de genes blanco regulados por miARNs en plantas, basado en la conservación evolutiva del par miARN-gen blanco. Además, pudimos encontrar genes blanco específicos de Solanaceae y demostrar que la estrategia se puede utilizar para la búsqueda de genes blanco pertenecientes a un grupo determinado de especies.

A partir de estos resultados, desarrollamos una herramienta bioinformática para identificar genes blanco de miARNs basada principalmente en la conservación durante la evolución de la interacción del par miARN-gen blanco en distintas especies. Esta herramienta fue usada para predecir nuevas interacciones y validar experimentalmente genes blanco no conocidos anteriormente en *Arabidopsis thaliana*. Algunos de ellos podrían participar de las mismas vías que genes blanco conocidos anteriormente, sugiriendo que algunos miARNs pueden controlar diferentes aspectos de un proceso biológico.

La biogénesis de los miARNs es un proceso clave porque determina la secuencia exacta de nucleótidos del ARN pequeño funcional. Poco se sabía sobre el reconocimiento de los precursores de plantas por la maquinaria de procesamiento. En la segunda parte de este trabajo presentamos una estrategia para estudiar aspectos mecánicos de la biogénesis de

los miARNs en plantas. Tratamos de dilucidar la dirección de procesamiento en precursores de miARNs en *Arabidopsis thaliana* a partir de los patrones de evolución de los precursores.

Introducción 1

Introducción

1.1 Silenciamiento génico y ARN pequeños en plantas

El control de la expresión génica es un proceso vital para los organismos unicelulares. Este proceso se utiliza para que las células se ajusten a los cambios en su ambiente de manera de optimizar el desarrollo, la respuesta celular al estrés y al entorno, entre muchos otros procesos. En plantas, la generación de los numerosos tipos celulares diferentes que forman en conjunto un organismo multicelular depende de que los genes se activen en las células adecuadas y en los momentos precisos del desarrollo. Muchos de los mecanismos regulatorios operan a nivel genético generando un control a nivel transcripcional y post-transcripcional.

En los últimos años se ha descubierto que pequeñas moléculas de 20 a 25 nucleótidos de longitud, son reguladores críticos de la expresión génica en eucariotas. En plantas existen distintos tipos de ARN pequeños, que por diferencias en su biogénesis y modos de acción, han sido clasificados en distintas clases. Estos ARN pequeños comparten algunos elementos entre ellos, pero su función y biogénesis son distintas.

1.1.1 Variedades de ARN pequeños en plantas

Los ARN pequeños provienen del procesamiento de regiones helicoidales de precursores de ARN. Los mismos se pueden clasificar en dos grupos principales. Los derivados de precursores de doble hebra (ARNdh) formados por hibridación intermolecular de dos cadenas de ARN complementarias y los derivados de precursores de RNA de una sola hebra (ARNsh) que forman una estructura de tallo-burbuja.

Los ARN pequeños más abundantes, derivados de precursores de ARNdh, son conocidos como siARNs (small interfering RNAs). Pueden ser de origen endógeno o exógeno. A su vez, los siARN se pueden clasificar en dos grupos, aquellos que producen un silenciamiento

a nivel transcripcional (TGS, del inglés: Transcriptional Gene Silencing) [24, 115] y aquellos que lo producen a nivel post-transcripcional (PTGS, del inglés: Post Transcriptional Gene Silencing)[49]. El primer grupo se origina a partir de ciertas regiones genómicas como transposones u otras secuencias repetitivas que son transcritos por la ARN polimerasa IV. Estos ARNsh se convierten en ARNdh por acción de una RDR. A continuación los ARNdh son procesados por DCL3 para finalmente liberar siARN de 24 nt de longitud. El segundo grupo, los PTGS son originados a partir de ARNs de virus o ARNm proveniente de un transgén. Por acción de una RDR se convierten en ARNdh pero en este caso es procesado por DCL2 o DCL4 para finalmente liberar siARN de 21-22 nt que serán encargados del silenciamiento post-transcripcional [120]. Los roles biológicos del silenciamiento mediado por siARNs incluyen resistencia contra virus [114], protección del genoma contra elementos móviles de ADN y roles regulatorios sobre genes endógenos [14, 24, 115].

Existen otros ARNdh menos abundantes que han sido descubiertos posteriormente. Este grupo está constituido por un lado por los ta-siARN (del inglés, trans-acting siARNs) y por el otro lado los nat-siARNs y nat-miARNs (del inglés, natural antisense siRNAs y miRNAs) [24, 89, 110, 115].

Los genes que darán lugar a ta-siRNAs son transcritos que generalmente son denominados TAS. Este transcripto luego es identificado y cortado por acción de un miARN que posee una secuencia complementaria al mismo [4]. Los fragmentos de corte de un TAS se estabilizan y por la acción de ARN polimerasas dependientes de ARN, se convierten en un ARNdh que es procesado por DCL4, generando ARNs pequeños de 21 nucleótidos. Finalmente algunos ta-siARNs se incorporan a un complejo RISC y guían el corte de los transcritos de otros genes de manera similar a la acción de un miARN [4, 42, 118, 124].

En contraste con los otros tipos de siARNs, que dependen de un RDR para sintetizar el precursor ARNdh, se cree que los precursores de ARNdh de NAT-siARNs surgen de la hibridación de transcritos separados [20, 58]. Estos ARNs separados pueden ser complementarios debido a que fueron transcritos por hebras opuestas en el mismo locus, y se denominan cis-NAT-siARNs. Como alternativa, los ARNs hibridizados pueden surgir de genes que no se superoponen, estos se denominan trans-NAT-siARNs. Sólo cis-NAT-siARNs han sido descritos en plantas; los trans-NAT-siARNs permanecen sólo como una posibilidad hipotética. El ARNdh formado es procesado por DCL1 o DCL2 liberando siARN primarios que inducen el corte sobre el ARNm expresado constitutivamente. Posteriormente, una RDR sintetiza la hebra complementaria del ARNm cortado. Esto da lugar nuevamente a un ARNdh el cual es procesado por DCL1 para producir una sucesión de siARN ubicados en fase de 21 nt cada uno denominados nat-siARN secundarios [20, 58].

La clase más abundante de ARN pequeños derivados de precursores de ARNsh son los miARNs. Tienen entre 20 y 22 nt de longitud y son originados de largos precursores con extensa estructura secundaria en forma tallo-burbuja. Estos precursores luego son procesados por acción de DCL1, liberando el ARN maduro. Actúan como reguladores negativos de la expresión génica reprimiendo la expresión de genes endógenos principalmente a nivel post-transcripcional. Esta regulación se puede dar a través de la inhibición de la traducción o del corte de los ARN mensajeros. Tanto en animales como en plantas, existen un gran número de miARNs conservados en distintas especies. En animales se han encontrado miARNs conservados en especies muy distantes como gusanos hasta humanos [87]. Y en plantas sucede lo mismo, donde existen miARNs conservados desde musgos hasta dicotiledóneas [10, 11, 31]. No se han encontrado miARNs conservados en animales y en plantas, lo que sugiere que han aparecido en forma independientes durante la evolución [31]. Si bien los miARNs en plantas y en animales comparten componentes básicos, su biogénesis y su modo de acción difiere entre ellos.

1.2 miARNs en plantas

Los miARNs son generados a partir de loci endógenos, tanto en animales como en plantas. Controlan una gran variedad de procesos biológicos, como el desarrollo, la diferenciación, proliferación y respuesta a estrés [9, 23, 31, 101, 115]

Hasta hoy, en *Arabidopsis thaliana* se han identificado más de 300 [59] miARNs. Se han utilizado distintos enfoques para identificar los miARNs: el clonado directo de ARN pequeños, mediante secuenciación de alto rendimiento y mediante estudios genéticos y mediante predicciones bioinformáticas [31], siendo esta última la más común para la mayoría de las especies.

Los miARNs en plantas están codificados por familias de genes de 1 a 32 miembros que dan lugar a miARNs maduros idénticos o muy similares. Cada *locus* perteneciente a una familia codifica un miARN maduro idéntico o casi idéntico. Hasta el momento han sido definidas unas 42 familias de miARNs en plantas, las que regulan una amplia variedad de procesos biológicos. Doce de dichas familias tienen como blanco ARN mensajeros que codifican factores de transcripción involucrados en el desarrollo, mientras que otras están relacionadas con rutas de respuesta a señales ambientales y hormonales, entre otros, estando la mayoría de ellas conservadas entre mono y dicotiledóneas [55]. Muchos de estos pequeños ARNs han aparecido recientemente en la evolución y por lo tanto aparecen en un número pequeño de especies [9, 11]. Además está claro si tienen algún rol biológico [11, 31].

Sin embargo, existen 22 familias de miARNs que están altamente conservadas en las plantas, estando presentes en angiospermas, gimnospermas y algunas de ellas aún en plantas basales como los musgos [7, 9, 125] (ver Tabla 4.1). Estos últimos miARNs cumplen funciones esenciales para la biología de las plantas [55].

1.3 Biogénesis de miARNs

Los miARNs se diferencian de otros ARNs pequeños por su particular biogénesis que implica su escisión de un precursor con extensa estructura secundaria localizado en un largo transcripto primario. En general, la biogénesis de estos ARN pequeños comienza con la transcripción por la ARN polimerasa II [118] a partir de unidades transcripcionales propias distribuidas en el genoma [92]. Los transcriptos primarios, llamados pri-miARNs, pueden tener varias kilobases de longitud y sufrir modificaciones post-transcripcionales como ser splicing, capping y poliadenilación. Estos transcriptos contienen precursores para miARNs con extensa estructura secundaria en forma de tallo-burbuja (stem-loop) [55].

En animales, el procesamiento comienza en el núcleo por DROSHA y finaliza en el citoplasma por la acción de DICER. En plantas, los precursores son procesados completamente en el núcleo a través de la acción de una ribonucleasa llamada DCL1 [92, 98] (del inglés DICER LIKE 1) en asociación con el cofactor proteico de unión a ARN de doble hebra HYL1 [51] (del inglés HYPONASTIC LEAVES 1) y la proteína SERRATE [69].

Al parecer es la estructura secundaria por sobre la secuencia primaria del precursor la más importante en la determinación del correcto procesamiento del mismo [17]. El producto generado a partir de los cortes llevados a cabo por DCL1, es un dúplex miARN-miARN* que luego continúa siendo procesado por otros componentes enzimáticos hasta dar lugar al miARN maduro de 21 nt.

El paso final de la biogénesis de los miARN es la incorporación asimétrica, a partir del dúplex miARN-miARN*, del miARN maduro dentro de un complejo de silenciamiento. Este complejo se denomina RISC (del inglés RNAi Silencing Complex). El componente central de todos los complejos de silenciamiento es un miembro de la familia de proteínas ARGONAUTA (AGO). En *Arabidopsis* existen distintas proteínas AGO que participan en diferentes procesos biológicos [111] y la incorporación de los ARN pequeños en los distintos complejos depende de la identidad del nucleótido del extremo 5' y de la vía de biogénesis [77, 79, 108].

En la mayoría de los miARNs el nucleótido extremo 5' es una U y en general la principal efectora de la actividad es AGO1 [77, 112, 115]. Complejos RISC similares se encuentran

presentes en células animales. Más recientemente han sido identificadas proteínas adicionales que regularían la actividad de la maquinaria de procesamiento [17].

En animales, los miARNs reconocen principalmente a la región 3' no codificante de ARN mensajeros blanco inhibiendo su traducción. En plantas es más común que los miARNs se unan a secuencias complementarias en los ARNm blanco en la región codificante señalándolos para su degradación [55]. En cualquier caso, es el miARN el que proporciona la especificidad contra las moléculas de ARN blanco [13].

1.4 Regulación de la expresión génica por miARNs

1.4.1 Regulación por corte del ARN blanco

En animales existe un gran número de genes blanco mediado por miARNs y un ARNm puede estar regulado por varios miARNs. En cambio los miARNs en plantas regulan un número limitado de genes blanco que además tienen un único sitio que es altamente complementario al miARN [115].

El ARN pequeño guía al complejo RISC hacia una molécula de ARNm complementario. Luego del reconocimiento de ARNm blanco por complementariedad de bases, la proteína AGO1 del complejo RISC introduce un corte en un enlace fosfodiéster del ARNm. Este corte ocurre entre las posiciones 10 y 11 desde el extremo 5' del miARN, independientemente de la longitud del miARN [40, 68, 70, 84, 119]

Luego del corte mediado por el miARN, los fragmentos 3' son degradados mediante la actividad de la enzima citoplasmática 5'-3' EXORIBONUCLEASA4 (XRN4) en *En A. thaliana* [104]. Los fragmentos 5' también pueden ser degradados por un complejo denominado Exosoma, el cual está encargado de diferentes funciones de degradación y procesamiento de ARNs [25]. En *Arabidopsis* la degradación del fragmento 5', puede ser acelerada por uridilación en el extremo 3' por la enzima una enzima denominada HESO1 [93].

1.4.2 Regulación de la traducción por miARNs

Los miARNs en animales, en general, son parcialmente complementarios a uno o más sitios presentes en la región 3' no traducida de los ARNm blancos [23, 37, 64] y raramente sufren el tipo de corte antes mencionado. Además la limitada complementariedad de secuencia, permite que los miARNs de animales regulen generalmente la expresión de muchos genes blanco. Un mecanismo que involucra la inhibición de la traducción del ARNm blanco por el miARN explica la represión de la expresión de los blancos de miARNs en animales [37]. En otras ocasiones, los miARNs de animales disminuyen la vida media de los transcritos

a los que se unen [48]. En algunos pocos miARN de plantas también se ha demostrado la existencia de un mecanismo de represión traduccional, además del corte del transcripto [8, 22, 27, 36, 62, 100].

1.4.3 Generación de ta-siRNAs y siRNAs secundarios

El corte de un transcripto por un miARN puede inducir la generación de ARN pequeños secundarios [4, 26, 30, 71, 80, 80]. Existe un mecanismo por el cual se genera un ARN pequeño específico en plantas que se denomina ta-siARN (trans-acting short-interfering RNA). Los genes que darán lugar a ta-siRNAs son transcriptos que generalmente son denominados TAS. El transcripto es identificado y cortado por acción de un miARN que posee una secuencia complementaria al transcripto [4]. Los fragmentos de corte de un TAS se estabilizan y por la acción de ARN polimerasas dependientes de ARN, se convierten en un ARN de doble hebra. Luego el ARN doble hebra es procesado por una Ribonucleasa de tipo III, llamada DICER- LIKE4 (DCL4), generando ARNs pequeños de 21 nucleótidos, los ta-siARNs. Finalmente estos ta-siARNs se incorporan a un complejo RISC y guían el corte de los transcriptos de otros genes de manera similar a la acción de un miARN [4, 42, 118, 124].

1.5 Procesos biológicos regulados por miARNs en plantas

Los procesos biológicos regulados por miARNs en plantas son muchos, y algunos miARNs desempeñan papeles importantes en el desarrollo, otros en la trasducción de señales hormonales, respuesta a estrés y respuesta a señales del ambiente [96, 97, 115].

De las 22 familias de miARNs conservados en plantas, 11 regulan factores de transcripción y la mayoría de ellos están involucrados en procesos de desarrollo o diferenciación celular [55] (Tabla 4.1). Por ejemplo los genes blanco del miR172 pertenecen a la familia de factores de transcripción AP2 (APETALA2) que está involucrado en el tiempo de floración y en el desarrollo de la hoja [8, 27]. La familia del miR319 regula factores de transcripción TCP (TEOSINTE BRANCHED1), la familia del miR165/166 factores de transcripción HD-ZIP III (HOMEODOMAIN-LEUCINE ZIPPER de clase III) y la familia del miR396 regulan factores de transcripción GRF (GROWTH-REGULATING FACTOR). Todas estas familias intervienen en el desarrollo de la hoja regulando distintos factores de transcripción [21, 84, 95]. El miR164, participa en el establecimiento del meristema apical del tallo, en el desarrollo de la raíz y definición de los bordes de los órganos mediante la regulación de miembros de la familia de factores de transcripción NAC (NAM, ATAF1/2 and CUC2) [63].

Por otra parte, entre los genes blanco que no corresponden a factores de transcripción, hay genes involucrados en diversos aspectos de la biología vegetal. Algunos codifican proteínas pertenecientes a la familia F-Box o enzimas E2 conjugantes de ubiquitina, las cuales están implicadas en la marcación de proteínas para la degradación por el proteosoma [96]. Otros genes con función conocida codifican para proteínas involucradas en la transducción de señales hormonales, o proteínas involucradas en el metabolismo energético, la respuesta a estrés o déficit de nutrientes.

Por otro lado, en otras especies que se han realizados estudios de ARN pequeños por técnicas de alto rendimiento, se han encontrado miARNs no conservados [31, 88, 95]. Los genes blanco regulados por estos miARNs tienen funciones más variables que los blanco de los miARNs conservados, de todos modos se ha demostrado el rol biológico de un pequeño número de estas interacciones [31].

En esta tesis nos enfocaremos principalmente en las familias de miARNs conservadas entre distintas especies de plantas (Tabla 4.1). Estas familias han sido estudiadas profundamente, por lo que su mayoría se encuentran ampliamente caracterizadas.

1.6 Predicción de genes blanco de miARNs

1.6.1 Métodos para predicción de genes blanco de miARNs en animales

Existen muchos métodos para la predicción de genes blanco de miARNs que fueron desarrollados por distintos grupos de investigación. Gran parte de ellos continúan manteniendo y actualizando sus algoritmos. El grupo de Cohen en el EMBL propuso el primer predictor de genes blanco de miARNs en 2003 [107] que luego fue actualizado en el 2005 [106]. TargetScan y TargetScanS fueron desarrollados por Bartel en el MIT y Burge en Cambridge [41, 47, 66, 67]. DIANA-microT, otra herramienta popular que fue creada por el grupo de Hatzigeorgiou, fue actualizada recientemente a la versión 5.0 [72, 73, 86]. El grupo de Rajewsky publicó su predictor de genes blanco en el 2005 y fue actualizado en el 2011 [6, 91].

Hay dos categorías de modelos predictivos: heurísticos y empíricos. Los modelos heurísticos utilizan algoritmos de detección que buscan posiciones a lo largo de la secuencia del ARNm y funciones de puntajes que filtran los blanco mediante la combinación de varios valores de entrada. Los primeros predictores aplican enfoques heurísticos, debido a la falta de una cantidad suficiente de datos para construir los modelos basados en el conocimiento empírico.

Los modelos predictivos utilizan entradas que se derivan del conocimiento de los detalles mecanísticos de las interacciones miARN-ARNm. La entrada más utilizada de los modelos predictivos, es la complementariedad de base entre el miARN y ARNm. En contraste con el apareamiento de bases casi perfecto en las plantas [94], los miARNs de animales por lo general se unen al ARNm con sólo algunas posiciones que están apareadas [61]. La complementariedad del apareamiento de bases en la región semilla (seed), que comprende los ocho primeros nucleótidos en el extremo 5' del miARN, es particularmente importante. Los parametros de entradas de accesibilidad del sitio y de conservación evolutiva se utilizan para aumentar la especificidad. La mayoría de los predictores usan softwares existentes, como el Vienna RNA package [52], mFold [127], DINAMelt [74] y mFold [35], para el cálculo de la energía libre. El uso de la conservación evolutiva de los genes blanco de miARNs está motivada por la premisa de que las especies 'similares' deben compartir miARNs comunes y genes blanco comunes. Sin embargo, esto conduce a la omisión de los genes blanco no conservados [65, 109].

1.6.2 Métodos para predicción de genes blanco de miARNs en plantas

Los algoritmos mencionados anteriormente son utilizados generalmente para predecir genes blanco de miARNs en animales. Algunos de ellos, como MiRanda, se utilizan tanto en animales como en plantas. Existen herramientas diseñadas específicamente para la predicción de genes blanco de miARNs en plantas. Algunas de ellas fueron implementadas como herramientas independientes, otras como servidores web y algunas como ambos.

La herramienta psRNATarget [32] utiliza un enfoque de programación dinámica, alineando las secuencias utilizando el algoritmo de Smith-Waterman modificado y aplicando el algoritmo 'RNAup'. Targetfinder [15] implementa un programa "FASTA" junto con un sistema de puntuación para mismatches, bulges o gaps para alinear las secuencias. TAPIR [19] está integrado con dos opciones de búsqueda, el motor de búsqueda "FASTA" (para búsquedas rápidas), y el motor de búsqueda "RNA hybrid" (para obtener resultados precisos). La herramienta Target-align [117] también emplea el método de puntuación basado en el algoritmo de Smith-Waterman para predecir las complementariedades entre los miARNs y los RNAs mensajeros. Por último, la herramienta p-TAREF [53] implementa un support vector regression (RVS) y utiliza la información de la "variación de la densidad de dinucleótido" alrededor del sitio target de conjuntos de datos de *A. thaliana*, *Oryza sativa*, *Medicago truncatula* y *Solanum lycopersicum*.

Además existen otras herramientas que son utilizadas para predecir genes blanco de miARNs pero que no fueron específicamente diseñadas para esto. Por ejemplo, la herramienta Web miRNA designer, WMD3 [83] es utilizada para el diseño de miARNs artificiales

para silenciar la expresión de genes blanco específicos. Pero esta herramienta también es empleada para la predicción de genes blanco de miARNs en plantas. RNAHybrid [60] es una herramienta utilizada para buscar la energía mínima libre de un ARN largo y uno corto. La secuencia corta se hibrida con la parte más adecuada de la secuencia larga. La herramienta está destinada principalmente para la predicción de genes blanco de miARNs, pero no es su único uso.

Patmatch [122] que está integrado en "The Arabidopsis Information Resource" (TAIR)¹ permite realizar búsqueda de secuencias de nucleótidos o de péptidos corto, como dominios pequeños o motivos. Esta herramienta puede ser adaptada, para utilizarla como herramienta de búsqueda de genes blanco de miARNs en plantas.

La mayoría de las herramientas específicas de plantas fueron desarrolladas para predecir genes blanco con una alta especificidad en el organismo modelo *A. thaliana*, si se utilizan los parámetros de predicciones óptimos. Y los sistemas de puntajes optimizados de Arabidopsis no se pueden utilizar como un umbral en el análisis de organismos no modelos [105]. Además estas herramientas en general detectan una gran cantidad de potenciales genes blanco (que incluyen los validados experimentalmente), pero su debilidad radica en que predicen una gran número de falsos positivos.

1.6.3 identificación de genes blanco regulados por miARNs mediante técnicas de secuenciación de alto rendimiento

Con los avances de los métodos de secuenciación de alto rendimiento o "next-generation sequencing", la identificación sistemática de genes blanco de miARNs específicos en un tiempo corto ya es un hecho. Un método denominado PARE (Construction of Parallel Analysis of RNA Ends) [43]. La misma consiste en una combinación de la estrategia modificada 5'RACE, para la identificación de genes blanco de miARNs y permite realizar estudios del degradoma de ARN, mediante técnicas de alto rendimiento.

A partir de este método, se han desarrollado diferentes herramientas como SeqTar [126], que es un método para la identificación de corte guiado por miARNs de degradoma de transcritos de plantas. Por otro lado, CleaveLand [2] es un pipeline computacional para la detección de cortes de miARNs en datos de degradoma. Esta herramienta toma como entrada secuencias de degradoma, ARNs pequeños y una base de datos de ARNm y devuelve los potenciales genes blanco de esos ARNs pequeños. sPARTA (small RNA-PARE target analyzer) [56] es un pipeline paralelizado, para el análisis integrado de miARNs en plantas y

¹<https://arabidopsis.org>

un conjunto de ARNm. Este pipeline incluye un software de identificación de nuevos genes blanco regulados por miARNs y utiliza datos de PARE.

Por otro lado, la herramienta PAREsnip permite realizar búsqueda de potenciales genes blanco de todos los ARN pequeños obtenidos de técnicas de secuenciación de alto rendimiento. Mediante la búsqueda de genes blanco de todo el "sRNAome" se puede facilitar identificación de genes blanco de ARN pequeños a grande escala.

1.7 sPARE???

Del Methods de Schapire et al.

Introduccion/introduccion

Objetivos 2

Objetivos

2.1 Objetivo general

Uno de los objetivos general de este trabajo de Tesis consiste en identificar a los genes regulados por miARNs y descubrir sus roles en plantas. Además, como objetivo queremos contribuir al conocimiento de la regulación del procesamiento de los miARNs en plantas. Se espera que los resultados de esta Tesis sirvan no solo para alcanzar los objetivos de investigación planteados sino también para promover el desarrollo de la Bioinformática como una disciplina que brinda una oportunidad única para que, a partir de investigaciones en las ciencias básicas, pueda hallarse el camino hacia el desarrollo de aplicaciones de interés estratégico para el país.

2.2 Objetivos específicos

1. Identificar genes regulados por miARNs en plantas.
 - Diseñar una estrategia para la identificación de genes blanco regulados por miARNs en plantas, basado en la conservación evolutiva del par miARN-gen blanco.
 - Desarrollar una herramienta web para la predicción de genes blanco de miARNs en diferentes especies de plantas.
2. Estudiar la biogénesis de los miARNs en plantas.
 - Caracterizar la relación entre la evolución de los precursores de miARNs en plantas y los mecanismos de procesamiento determinados previamente.

Materiales y Métodos 3

Métodos

3.1 Predicción de genes regulados por miARNs en plantas

En la primer parte de esta tesis diseñamos una estrategia para la identificación de genes blanco regulados por miARNs basado en la conservación evolutiva del par miARN-gen blanco. La metodología aplicada es la siguiente.

3.1.1 MiARN consensos

Las 22 familias de miARNs conservadas en angiospermas fueron consideradas para esta parte del trabajo [11, 39]. MiR319 y miR159 que codifican para miARNs similares, fueron considerados como familias diferentes ya que regulan a genes blanco distintos [85]. Consideramos todos los miembros de estas familia, obtenidos de miRBASE¹, pertenecientes a *A. thaliana*, *Populus trichocarpa* y *Oryza Sativa*. Variaciones en las posiciones 1, 20 y 21 son muy comunes en las familias de miARNs [34]. Por esto, definimos como secuencia consenso, a las secuencias más comunes (posiciones 2-19) de distintos miembros de cada familia (tabla 4.1).

3.1.2 Predicción de genes regulados por miARNs

Conjunto de datos de plantas

Los datos de las secuencias pertenecen a librerías extraídas de “Gene Index Project”², que consiste en una base de datos de ESTs ensamblados. Seleccionamos un conjunto de datos pertenecientes a Angiospermas. Además utilizamos secuencias de ARNm completos de *A.*

¹<http://mirbase.org>

²<http://compbio.dfci.harvard.edu/tgi/>

*thaliana*³ y *Oryza Sativa*⁴ (ver tabla A.1). La búsqueda la realizamos utilizando PatMatch [122], que es un programa de búsqueda de patrones de nucleótidos cortos o péptidos. El programa puede ser usado para encontrar coincidencias con un patrón de secuencia específico y permite el uso de códigos de secuencias ambiguas y expresiones regulares y por esto se puede utilizar la búsqueda con mismatches, inserciones y deleciones. Realizamos la búsqueda de potenciales genes blanco permitiendo tres mismatches con las secuencias consensos, mientras que las interacciones G:U y los bulges fueron considerados mismatches. Para realizar el alineamiento del par miARN-gen blanco, desarrollamos una versión modificada del algoritmo de programación dinámica Needleman-Wunsch [81], utilizando el lenguaje Perl⁵. Además, desarrollamos scripts para integrar los módulos de Blastx [5] utilizando el proteoma de Arabidopsis y el módulo RNAhybrid [45] que es una herramienta que permite encontrar la menor energía libre de hibridación (MFE) de dos secuencias de ARN.

Filtros

Las secuencias candidatas fueron etiquetadas con el identificador del locus (locus ID) con mejor puntuación (best hit) en *A. thaliana*, utilizando el módulo de Blastx (Corte del evalue de $10e^{-5}$). De este modo, genes blanco de distintas especies que tenían la misma etiqueta fueron agrupados juntos, ya que tendrían el mismo homólogo en *A. thaliana*. El filtro de conservación evolutiva hace referencia al número mínimo de especies donde la misma etiqueta estaba presente para un miARN particular. El filtro empírico está basado en trabajos previos [100] y hace referencia a la energía de interacción MFE (mínima energía libre de hibridación de al menos 72% del apareamiento perfecto). El otro filtro empírico requiere que entre el par miARN-gen blanco, solamente está permitido un mismatch entre la posición 2 y la 12 del miARN (1-11 de nuestra búsqueda modificada con las secuencias consenso).

Controles

Como control, realizamos las búsquedas del mismo modo que lo hicimos para los miARNs conservados, pero utilizando secuencia al azar. Para cada miARN conservado, generamos 20 secuencias al azar (scramble) dividiendo las secuencias originales de a di-nucleótidos y luego generando nuevas secuencias al azar conservando esa composición de los di-nucleótidos como fue descrito previamente [54]. De estas 20 secuencias al azar, elegimos las 10 que tenían el número más similar del total de genes blanco para el miARN real correspondiente. La relación señal/ruido fue calculada como el cociente entre el número de genes blanco para

³<http://arabidopsis.org>

⁴<http://rice.plantbiology.msu.edu>

⁵<http://perl.org>

los miARNs y el número de genes blanco del promedio obtenido para las secuencias al azar. Como un control adicional, seleccionamos dos miARNs que no están conservados durante la evolución, que son el miR158 y el miR173.

Ecotipos utilizados y condiciones de crecimiento

Las plantas de *A. thaliana* utilizadas para los experimentos en esta parte del trabajo corresponden a el ecotipo Columbia-0 Col-0. Las plantas fueron cultivadas en una cámara de crecimiento con un régimen de 16 h de luz ($100 \mu\text{E.m.}^{-2}\text{s}^{-1}$) y 8 h de oscuridad (condición día largo). La temperatura de crecimiento fue de 23°C durante el ciclo luz/oscuridad, mientras que la humedad fue mantenida en 65% de humedad relativa. Las plantas fueron regadas 2 veces por semana con agua. Para el crecimiento directo en tierra, las semillas fueron estratificadas a 4°C por 2 días en tubos de microcentrífuga con 1ml de 0,1% (p/v) agar, y luego sembradas en tierra. Las plantas de *Nicotiana tabacum* (cv Petit Havana) fueron crecidas en condición día largo durante 8 semanas y la segunda hoja fue utilizada para el análisis de ARN.

Cleavage site mapping of target mRNA and expression analysis

ARN Poly(A)+ fue extraído a partir de 50 mg de ARN total de plántulas de Col-0 utilizando el kit comercial PolyAtract®(Promega) La ligación del Oligo Adaptador de ARN, transcripción reversa y 5' RACE fueron realizadas como se describió anteriormente [85] Two nested gene-specific reverse oligonucleotides were used for 5' RACE. Los productos de la PCR fueron resueltos en geles de agarosa al 2% y se detectaron por tinción con bromuro de etidio. La PCR en tiempo real cuantitativa (qPCR) para los genes blanco del miR396 y miR159 se realizó como se ha descrito anteriormente [85, 95] La lista de los cebadores para estos ensayos están descritos en las tablas A.3 y A.4. Las plantas que sobreexpresan el miR396 y miR159 se han descrito previamente [85, 95].

3.2 comTAR: una herramienta para la predicción de genes blanco regulados por miARNs en plantas

A partir de los resultados positivos obtenidos de la estrategia descrita anteriormente, decidimos desarrollar una herramienta web y dejarla disponible para la comunidad científica denominada comTAR que está disponible en un sub-dominio de la página web institucional del IBR: <http://rnabiology.ibr-conicet.gov.ar/comtar>.

3.2.1 MiARN y transcriptos

Como las secuencias del maduro del miARN puede variar en distintas especies, especialmente en la posición 1, 20 y 21 ([28], utilizamos secuencias del 2-19 (18nt) para realizar las búsquedas. Como además existen variaciones en las secuencias en los distintos miARNs de las mismas familias, utilizamos la más representativa teniendo en cuenta los genomas de *Arabidopsis*, álamo y arroz. De este modo comTAR contiene datos pre-calculados, de potenciales genes blanco para 22 miARNs conservados en plantas (ver tabla 4.1) donde el usuario puede navegar los resultados y cambiar los parámetros de entrada. Además, el usuario puede realizar la búsqueda de nuevos ARNs pequeños teniendo en cuenta esta consideración. El cálculo se hace en el cluster del CCT-Rosario y los datos se obtienen luego de unas horas. Como la herramienta web la realizamos tiempo después de haber hecho la estrategia para predicción de genes blanco, utilizamos una nueva base de datos más actualizada y completa denominada Phytozome⁶ [46]. La misma corresponde a secuencias de transcriptos de plantas formado por archivos de nucleótidos en formato FASTA de transcriptos de ARNm (UTR, exones) con variantes de splicing.

3.2.2 Búsqueda de genes blanco

La búsqueda de genes blanco la realizamos de la misma manera que la descrita anteriormente con algunos cambios. Además de actualizar la base de datos y utilizar la de Phytozome, actualizamos la base de datos de *A. thaliana* por la del TAIR10. Las secuencias candidatas fueron etiquetadas con el mejor hit del locus ID del Arabidosis TAIR10, utilizando los archivos de anotación de Phytozome, y lo utilizamos como “TAG” (etiqueta). Por último, cada TAG de Arabidopsis fue indexado con una breve descripción funcional y computacional obtenida del TAIR10 y los genes blanco candidatos fueron agrupados por familias teniendo en cuenta la clasificación de familias del TAIR10.

3.2.3 Herramienta web y almacenamiento de datos

ComTAR fue diseñado como una aplicación web con un framework open-source en PHP denominado Codeigniter para la interfaz gráfica, pero el análisis está basado en un back-end escrito en Perl. Los datos que surgen de ese análisis fueron almacenados en una base de datos en MySQL⁷. El back-end es el encargado de realizar la búsqueda de secuencias y además ahí es donde se integraron las herramientas y scripts para aumentar la especificidad y sensibilidad

⁶<http://phytozome.jgi.doe.gov>

⁷<http://mysql.com>

3.2 comTAR: una herramienta para la predicción de genes blanco regulados por miARNs en plantas

19

de comTAR. También el back-end es el encargado de generar los resultados finales. Mientras el front-end es el responsable de mostrar los resultados (Figura 3.1). El TAG del mejor hit en *Arabidopsis* es el que determina el número de especies donde un gen blanco está presente, y el número mínimo de especie es un parámetro que es definido por el usuario.

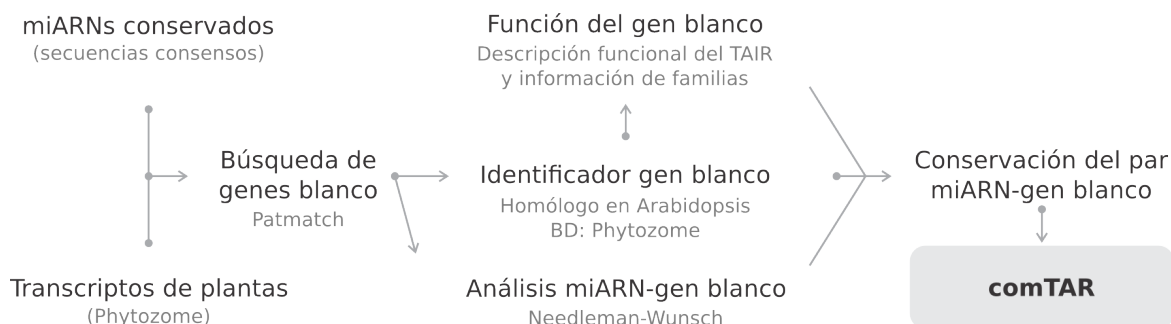


Fig. 3.1 comTAR. Diagrama de flujo que describe la herramienta

Capítulos 4

Resultados Capítulo 1

Aplicaciones bioinformáticas para el estudio de interacciones miARN-gen blanco

4.1 Introducción

patmatch del TAIR es utilizado para...

4.2 Resultados

4.2.1 Predicción de genes regulados por miARNs.

Diseño de una estrategia para la identificación de genes blanco regulados por microARNs basado en la conservación evolutiva del par microARN-gen blanco.

Enfocamos nuestro análisis en 22 miARNs que están conservados en Angiospermas [31, 34]. En general estos miARNs están codificados por pequeñas familias hasta 32 miembros. En los genomas completos de Arabidopsis, poplar y arroz es común encontrar variaciones en la secuencia de los miARNs pertenecientes a una misma familia, especialmente en el primer nucleótido y los nucleótidos 20 y 21 [34].

Sin embargo, observamos que la región entre la posición 2 y 19 está bastante conservada y pudimos encontrar una secuencia consenso presente en la mayoría de los miembros de cada familia de miARNs en esas tres especies (tabla 4.1). Curiosamente, las bases variables fuera de esta región conservada son propensas a tener mismatches con genes blanco conocidos, lo que indica que podría existir una correlación entre la interacción miARN-gen blanco y la conservación de la secuencia del miARN.

Diseñamos una estrategia para identificar nuevos pares miARN-gen blanco principalmente basada en la conservación evolutiva de la secuencia del gen blanco (Figura 4.1). Las secuencias consenso de 18 nt de cada familia de miARN fueron usadas inicialmente para realizar la búsqueda de genes blanco en contigs de ESTs, de 41 especies de plantas, obtenidos de “Gene Index Project” un proyecto mantenido y administrado por la universidad de Harvard que contiene un catálogo completo de genes en una amplia gama de organismos incluyendo plantas. Además se utilizaron ARNm completos para *A. thaliana* y *Oryza Sativa* para ver la lista completa de especies, ver tabla A.1). Utilizando las secuencias consenso de 18nt y permitiendo 3 mismatches (errores), la búsqueda de genes blanco arrojó como resultado 38.597 genes distribuidos en las 43 especies (Figura 4.1, bin 1). Las interacciones G-U y los bulges fueron considerados como mismatches en esta primera búsqueda. Todos los genes blanco de *A. thaliana* conocidos hasta ese momento fueron identificados usando esta estrategia con la excepción de CSD2, un gen blanco del miR398 que contiene 4 mismatches (tabla A.1).

Teniendo en cuenta que la mayoría de los genes blanco arrojados presentan una escasa descripción del tipo genómica funcional, realizamos un BLASTx contra el proteoma de *A. thaliana*. El “locus ID” obtenido como “best hit” se utilizó como tag (etiqueta) para identificar al candidato en distintas especies (Figura 4.1). A pesar que esta estrategia no necesariamente identifica el gen ortólogo de Arabidopsis, sirve como propósito de clasificación de cada potencial gen blanco de miARN. Aunque la mayoría de los potenciales genes blanco pudieron ser fácilmente asignados con una etiqueta, algunos pocos casos, que incluye a los genes que representan ARNs no codificantes fueron perdidos en este paso.

Este enfoque permite la selección de los mejores candidatos basándose en la presencia de los genes blanco en un número distinto de especies. Utilizando 4 especies como el mínimo de especies requeridas (ya que tiene una buena especificidad), dio como resultado 3.781 genes que corresponden a 533 tags diferentes (Figura 4.1, bin 2).

La búsqueda también se puede hacer en combinación con filtros empíricos de interacción par miARN-gen blanco que tienen en cuenta la energía de interacción y la posición de los mismatches (ver Materiales y métodos). De los 38.597 candidatos iniciales, 9.375 pasan estos filtros (Figura 4.1, bin 4). Combinando filtros de energía y filtro de conservación evolutiva, la búsqueda arrojó como resultado 563 candidatos correspondientes a 146 tags (Figura 4.1, bin 5).

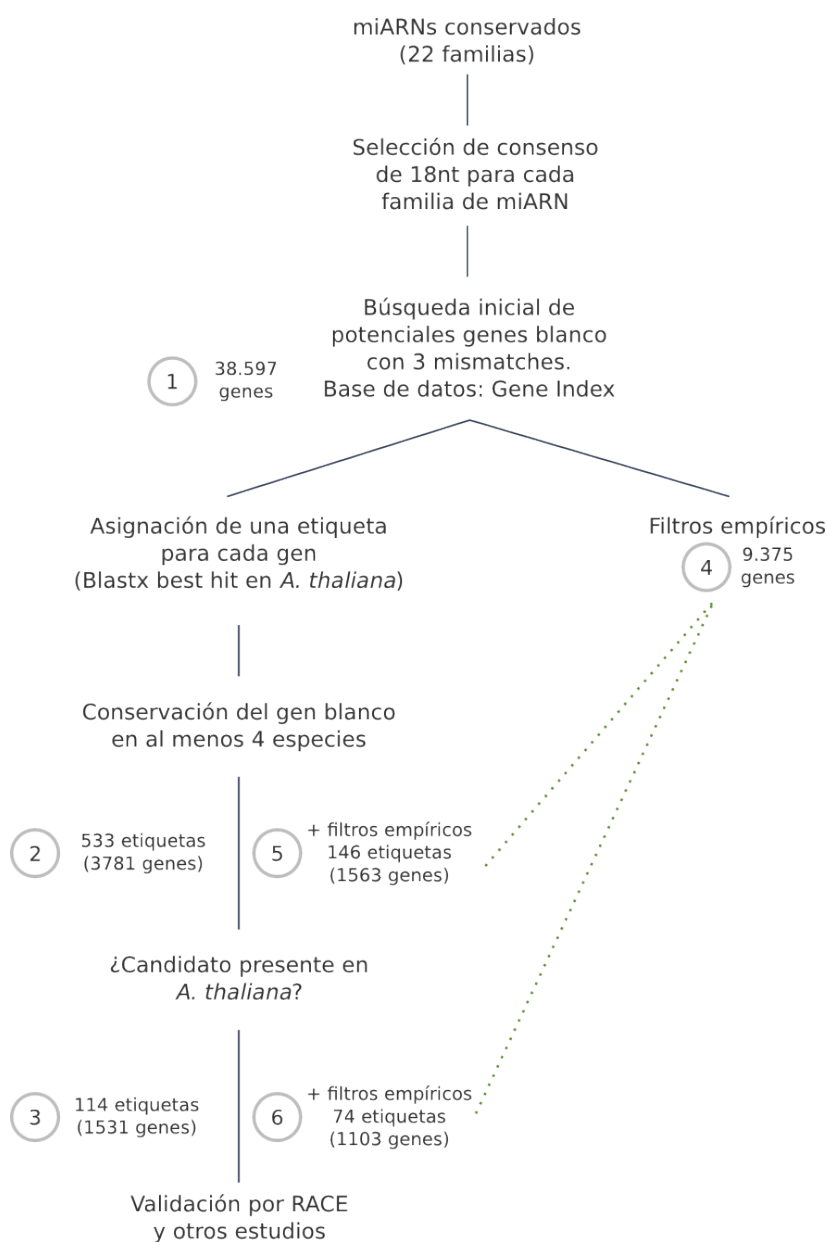


Figura 4.1 Esquema de la estrategia para la identificación de nuevos genes blanco. El número de genes blanco está identificado en cada paso. Luego de aplicar el análisis de conservación, todos los genes que tienen el mismo hit en Arabidopsis, fueron considerados como un solo gen blanco. El lado derecho muestra la búsqueda hecha con filtros empíricos: bin 5 y 6 incluyen genes blanco seleccionados con ambos filtros, empíricos y de conservación. Mientras que el bin 2 y 3 muestra los potenciales genes blanco seleccionados sólo con el filtro de conservación.

Table 4.1 miARNs y sus genes blanco en plantas

miARN	Consenso (18 nt)	Targets conocidos ^(a,b)
miR156	GACAGAAGAGAGTGAGCA	factores de transcripción SPL
miR159	TTGGATTGAAGGGAGCTC	factores de transcripción MYB, NOZZLE (NZL)
miR160	GCCTGGCTCCCTGTATGC	factores de transcripción ARF
miR162	CGATAAACCTCTGCATCC	DCL1
miR164	GGAGAAGCAGGGCACGTG	factores de transcripción NAC
miR166	CGGACCAGGCTTCTATTCC	factores de transcripción HDZip
miR167	GAAGCTGCCAGCATGATC	factores de transcripción ARF, IAA-ALANINE RESISTANT 3 (IAR3)
miR168	CGCTTGGTGCAGGTCGGG	AGO1
miR169	AGCCAAGGATGACTTGCC	factores de transcripción CCAAT-HAP2
miR171	TTGAGCCGTGCCAATATC	factores de transcripción GRAS
miR172	GAATCTTGATGATGCTGC	factores de transcripción AP2
miR319	TGGACTGAAGGGAGCTCC	factores de transcripción TCP
miR390	AGCTCAGGAGGGATAGCG	TAS RNA
miR393	CCAAAGGGATCGCATTGA	TIR1 proteins, F-BOX proteins
miR394	TGGCATTCTGTCCACCTC	proteínas F-BOX
miR395	TGAAGTGTGTGGGGGAAC	ATP-sulfurilasas, transportadores de sulfato
miR396	TCCACAGCTTCTTGAAC	factores de transcripción GRF, MMG4.7, FLUORESCENT IN BLUE LIGHT (FLU)
miR397	CATTGAGTGCAGCGTTGA	Laccases
miR398	GTGTTCTCAGGTCACCCC	Cu/Zn SODs, CytC oxidase protein subunit, Chaperona de cobre (CCS)
miR399	GCCAAGGAGATTGCCC	Enzima E2 de conjugación de ubiquitina
miR408	TGCACTGCCTCTTCCCTG	Blue copper proteins, Laccases, P-TYPE ATPase (PAA2), PAC1 (Proteasome component)
miR827	TAGATGACCATCAGCAAA	SPX proteins

a Los genes blanco fueron agrupados según sus funciones.

b Nuevos genes blanco validados experimentalmente en este estudio están indicados en negrita.

Parámetros empíricos y de conservación evolutiva pueden actuar de manera sinérgica para identificar genes blanco regulados por miARNs.

Potenciales genes blanco de miARNs fueron clasificados de acuerdo al mínimo número de especie en donde fueron detectados (Figura 2A-E). Como control para cada miARN generamos 10 secuencias “scramble” (al azar), dividiendo las secuencias originales de a di-nucleótidos y luego generando nuevas secuencias al azar conservando la composición de los di-nucleótidos. Estas secuencias al azar fueron utilizadas para realizar búsqueda de genes blanco del mismo modo que lo hicimos para las secuencias originales. La relación señal/ruido fue calculada como el cociente entre el número de genes blanco para los miARNs y el número promedio obtenido de las secuencias al azar. El radio fue de 1,2 para todos los miARNs juntos sin requerir conservación y esa relación incrementa con el número de especie en donde los genes blanco fueron detectados (Figura 4.2 A, recuadro). Los datos para todos los miARNs y sus potenciales genes blanco conservados en al menos 4 especies están incluidos en la tabla 4.2.

Luego estudiamos la selección de candidatos teniendo en cuenta los filtros empíricos. Para esto aplicamos una versión modificada de los filtros descritos anteriormente y requiriendo (i) una energía mínima de hibridación (MFE) de al menos 72% del apareamiento perfecto de cada secuencia consenso y (ii) que sólo un mismatch pudiera estar presente entre la posición 1 y la 11 de la secuencia consenso (2-12 del miARN). De la búsqueda inicial 9.375 genes pasaron estos filtros conteniendo el 97% de los genes validados anteriormente de Arabidopsis. (Figura 4.1, bin 4).

Al aplicar solamente este filtro empírico, dio como resultado una relación señal/ruido de 1,7, al agrupar todos los miARNs juntos (Figura 4.2 A). Observamos que aplicar si-

multáneamente los filtros empíricos y de conservación aumentaron significativamente la relación señal/ruido para todos los miARNs juntos (Figura 4.2 A recuadro) y también de cada miARN individualmente (Figura 4.2 B-E, recuadros y tabla 4.2). En varios casos, esta relación llega hasta 10 cuando se requiere de que el gen blanco este presente en más de 5 especies y que pase los filtros empíricos (Figure 4.2 A–D). Este efecto sinérgico indica que el filtro de conservación evolutiva y los parámetros empíricos pueden estar seleccionando aspectos diferentes de la interacción miARN-gen blanco.

Observamos que el número de genes blanco candidato y la relación señal/ruido es variable entre los distintos miARNs. El miR396 tiene la mayor cantidad de potenciales genes blanco, 92 de ellos presentes en al menos 4 especies y 26 de ellos pasan además los filtros empíricos (Tabla 4.2 y Figura 4.2 B). El miR408 y el miR398 también tienen un número alto de potenciales genes blanco y buenas relaciones de señal/ruido (Figura 4.2 C-D).

En contraste, ciertos miARNs como el miR162, miR168 y miR399 tienen un solo potencial gen blanco conservado en al menos 4 especies de acuerdo con nuestra búsqueda (Tabla 4.2 y Figura 4.2 E). Al menos en el caso del miR162 y del miR168 este resultado podría estar reflejando su rol específico en la regulación por retroalimentación de la biogénesis del miARN, ya que controlan los niveles de expresión DCL1 y AGO1 respectivamente [113, 119].

Como control adicional para nuestra estrategia hicimos la búsqueda de genes blanco del miR158 y miR173, que son miARNs presentes solamente en *A. thaliana* y especies bien cercanas (17). Como era esperado estos miARNs no generaron más candidatos que sus versiones al azar (Tabla 4.2 y Figura 4.2 F).

Luego chequeamos si los pares miARN-gen blanco altamente conservados tenían una interacción más fuerte que los que están presentes en pocas especies. Para esto calculamos la energía mínima de hibridación para cada interacción detectada en nuestro trabajo. Observamos que los pares miARN-gen blanco presentes en muchas especies tienden a tener energía de interacción mayores que los que están presentes en menos especies (Figure 4.3 A). De todos modos, la correlación no fue notoria y algunas interacciones miARN-gen blanco tuvieron una baja energía de hibridación (Figure 4.3 A). Estos resultados muestran que una alta conservación podría no ser necesariamente equivalente a una fuerte interacción, la misma podría proporcionar una explicación para los efectos sinérgicos causados por los filtros de evolución y empíricos sobre la relación señal/ruido.

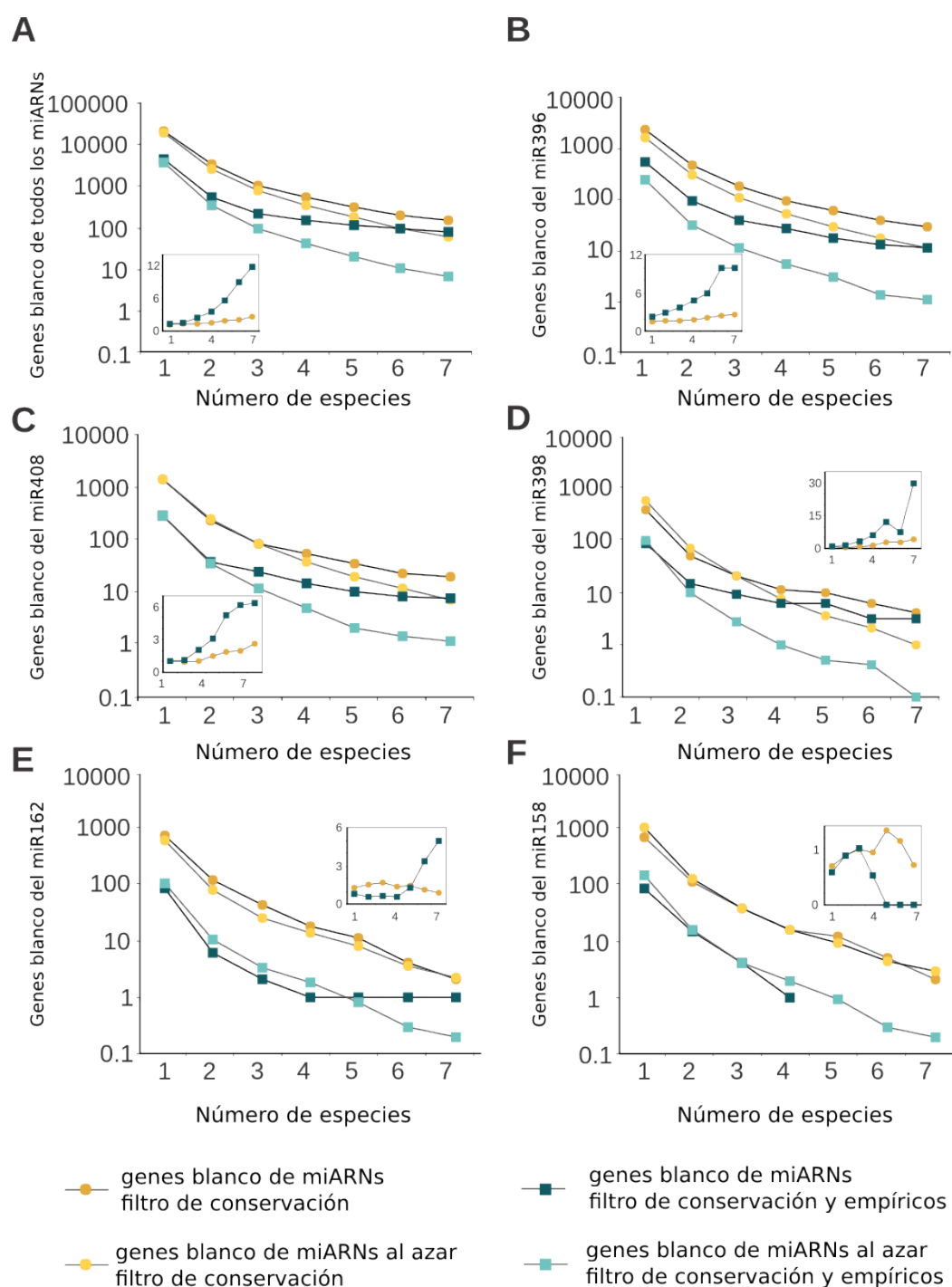


Figura 4.2 Conservación de potenciales genes blanco en distintas especies. Todos los miARNs (A), miR396 (B), miR408 (C), miR398 (D), miR162 (E), miR158 (F). Puntos naranja representan los genes blanco de miARNs usando filtro evolutivo. Puntos amarillos representan los genes blanco de las secuencias al azar usando filtro evolutivo. El cuadrado azul muestra los genes blanco de miARNs luego de aplicar filtros empíricos y evolutivos, mientras que el cuadrado celeste representa los genes blanco de las secuencias al azar en las mismas condiciones. Los recuadros muestran la relación señal/ruido.

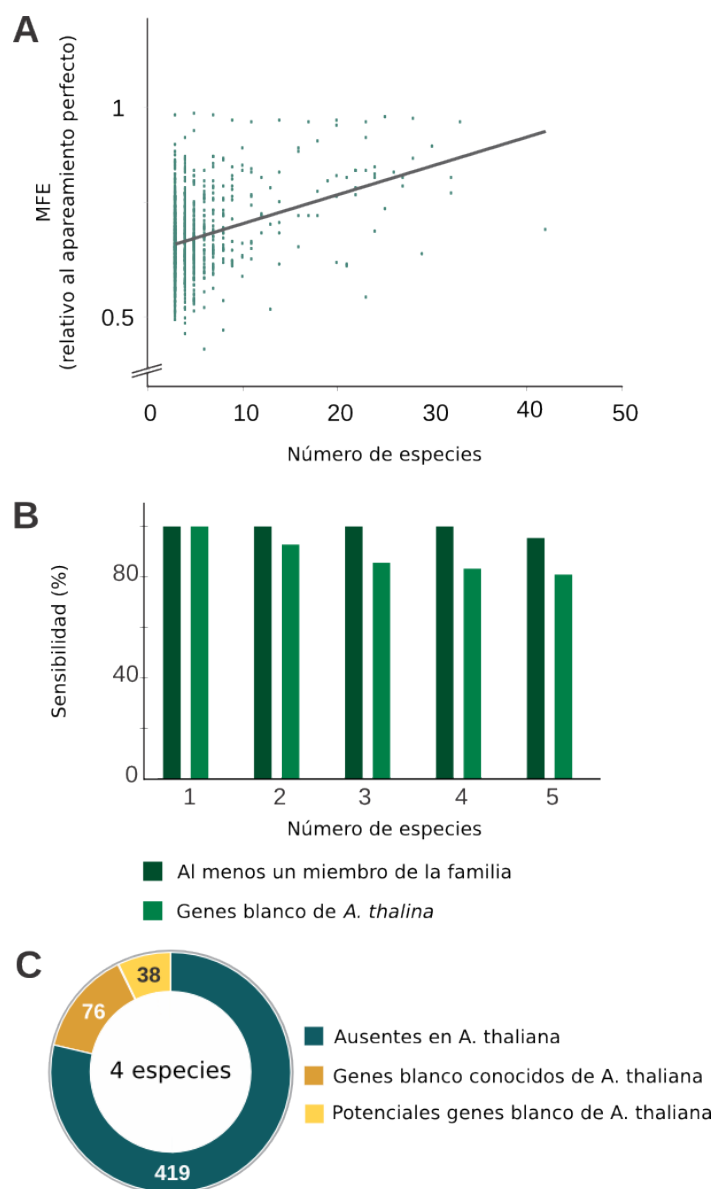


Figura 4.3 Selección de genes blanco por conservación evolutiva de la secuencia. (A) Relación entre MFE y el número de especies en donde cada gen blanco fue detectado. (B) Sensibilidad de la estrategia, analizado de dos modos distinto. Verde claro: evaluando la presencia de genes validados en *Arabidopsis* y en verde oscuro teniendo en cuenta la presencia de por lo menos un gen blanco de cada familia regulada por miARNs. (C) Clasificación de los potenciales genes blanco presentes en al menos 4 especies.

Table 4.2 Detection of miRNA targets using different filters

	Sin filtros			Filtros empíricos			Conservación 4 especies			Todos los filtros		
	miARN	scramble	ratio	miARN	scramble	ratio	miARN	scramble	ratio	miARN	scramble	ratio
miR156	3915	3994.4	± 149.9	± 890	704.7	± 45.2	1.3	34	39.7	± 3.1	5.4	± 1.1
miR159	1663	1283.7	± 47.8	± 472	254.9	± 21.9	1.9	20	10.1	± 1.1	1.5	± 0.5
miR160	793	695.6	± 30.5	± 277	157.5	± 28.8	1.8	5	4.4	± 0.9	1.5	± 0.5
miR162	1191	930.2	± 139.5	± 108	164.7	± 24.1	0.7	18	13.5	± 3.5	1.8	± 0.5
miR164	2486	1480.2	± 60.4	± 678	333.2	± 32.2	2.0	39	12.4	± 1.9	1.5	± 0.5
miR166	879	815.5	± 45.0	± 231	129	± 14.5	1.8	16	10.6	± 1.4	0.9	± 0.4
miR167	1777	1364.2	± 146.6	± 478	214.8	± 27.5	2.2	22	20.2	± 3.6	1.8	± 0.5
miR168	962	797.5	± 48.5	± 209	185	± 14.2	1.1	6	4.4	± 0.8	1.1	± 0.5
miR169	1540	1047.2	± 69.7	± 464	181.4	± 15.6	2.6	26	11.1	± 2.1	1.2	± 0.2
miR171	884	723.4	± 32.1	± 202	113.8	± 13.4	1.8	7	6.6	± 1.4	0.7	± 0.3
miR172	3007	1693.7	± 124.7	± 540	288.1	± 40.3	1.9	34	17.7	± 1.7	2.2	± 0.6
miR319	1363	1274.2	± 113.6	± 324	249.2	± 22.3	1.3	18	15	± 2.8	1.8	± 0.5
miR390	873	814.4	± 64.3	± 335	173	± 22.5	1.9	8	4.7	± 1.2	0.7	± 0.5
miR393	986	844.6	± 58.7	± 276	124.6	± 11.1	2.2	14	7.1	± 1.2	0.5	± 0.2
miR394	1569	1531.4	± 57.5	± 188	237.1	± 25.0	0.8	26	21.4	± 2.2	3	± 0.5
miR395	1472	1226.7	± 66.7	± 426	217.6	± 16.5	2.0	11	8.8	± 1.3	2.9	± 0.3
miR396	4641	2979.3	± 246.6	± 1246	390.5	± 38.8	3.2	92	51.4	± 5.9	5.4	± 1.0
miR397	1426	1050.9	± 27.9	± 368	236.5	± 23.5	1.6	26	9.7	± 0.8	10	± 0.3
miR398	935	834	± 34.5	± 376	144	± 18.1	2.6	11	7.5	± 1.6	1	± 0.3
miR399	1192	1137.6	± 72.0	± 275	207.8	± 24.9	1.3	5	13.6	± 1.7	1	± 0.7
miR408	2782	2502.9	± 103.6	± 695	468.7	± 50.8	1.5	51	35.1	± 3.0	14	± 0.8
miR837	2261	2000.1	± 119.8	± 317	297.1	± 45.0	1.1	44	23.4	± 3.9	4	± 0.8
Total	38597	31021.7	± 1859.8	± 9375	5473.2	± 576.3	1.7	533	348.4	± 47.0	146	± 11.3
Control												
miR158	1364	1462.8	± 69.1	± 170	208.7	± 15.8	0.8	15	16	± 1.7	1	± 0.4
miR173	1386	1232.1	± 101.7	± 243	215.6	± 23.4	1.1	11	12	± 2.4	1	± 0.4

a Sin filtros, búsqueda inicial utilizando los miARN consenso de 18nt y 3 mismatches.

b Filtros empíricos, energía de al menos 72% del apareamiento perfecto y 1 mismatch en la posición 2-12 del par miARN-gen blanco.

c Conservación del ID tag en al menos cuatro especies.

d Todos los filtros, combinación de los filtros empíricos y de conservación en al menos cuatro especies.

e miARN, genes blanco para cada miARN específico.

f scramble, promedio de los genes blanco de 10 versiones al azar de cada miARN ± error estándar.

Identificación de nuevos genes blanco en *A. thaliana* por conservación de la secuencia del gen blanco.

Para encontrar nuevos genes blanco nos enfocamos en los genes potenciales que fueron seleccionados de nuestra estrategia utilizando solamente conservación evolutiva, debido a que los parámetros empíricos ya fueron utilizados extensamente en trabajos anteriores. [4, 54, 100]. En primer lugar, analizamos la detección de genes blanco validados previamente en *A. thaliana* [basado en [39]] usando nuestra estrategia y encontramos que el 84% de ellos estaban presentes en al menos 4 especies (Figura 4.3 B). Consideramos esto como un buen resultado ya puede ser que no todos los genes blanco de *Arabidopsis* estén conservados evolutivamente.

Generalmente los miARNs en plantas regulan genes que codifican para proteínas de la misma familia, es por esto que evaluamos si por lo menos un miembro de cada familia era detectado en nuestro enfoque. Encontramos genes blanco pertenecientes a casi todas las familias de genes codificantes para proteínas presentes en cuatro especies (Figura 4.3 B), con la excepción de TAS3, que es regulado por el miR390, al ser un ARN no codificante no es detectado por Blastx.

Para encontrar nuevos genes blanco regulados por miARNs, nos enfocamos únicamente en los potenciales genes blanco conservados en 4 especies, donde una de ellas es *A. thaliana* (Figura 1, bin 1). Genes blanco de miARNs que no están presentes en *A. thaliana* podrían incluir genes que perdieron su regulación durante la evolución o genes que hayan adquirido control por un miARN conservado más reciente en otras especies. La conservación en cuatro especies fue elegida como un filtro evolutivo porque provee buena sensibilidad para genes blanco conocidos.

Identificamos 114 potenciales genes que satisfacen este criterio. Donde 76 de ellos son genes validados anteriormente o genes muy relacionados (Figura 4.3 C). Curiosamente encontramos 38 genes que no tienen relación con genes blanco conocidos de miARNs y decidimos estudiar este grupo con mayor detalle. Nos enfocamos primero en los genes que estaban presentes en un gran número de especies para tener mejor especificidad (Figura 4.2) e intentamos validarlos utilizando 5' RACE PCR modificada [57, 68].

Un potencial gen blanco del MiR408 era At5g21930 que codifica para P-TYPE ATPase OF ARABIDOPSIS 2 (PAA2) y estaba presente en 22 especies distintas incluido monocotiledóneas y dicotiledóneas. MiR408 es inusual debido a que tiene un 5'-A, sin embargo >30% de las secuencias maduras del miR408 corresponden a una variante corrida 1 nt que empieza con 5'-U [76] (Figura 4.4 A). La validación experimental reveló fragmentos de ARNm compatible con este último sitio de corte (Figura 4.4 A). PAA2 es necesaria para

el transporte de iones de cobre a plastocianina [82], y su regulación por el miR408 está relacionada con el rol de este miARN en la homeostasis de cobre [121].

Otro potencial candidato del miR408 era At3g22110 que codifica para PROTEASOME ALPHA SUBUNIT C1 (PAC1) y estaba presente en 20 especies. Por medio de 5' RACE PCR demostramos que este gen es gen blanco del miR408 (Figura 4.4 A). Curiosamente la interacción del par miARN-gen blanco tiene 3 mismatches en la región 5', y se hubiera perdido como potencial gen blanco si se aplicaban solamente los filtros empíricos.

Luego estudiamos los genes blanco del miR396, donde los genes SVP y SUI1 estaban presentes en 29 y 19 especies respectivamente. Pero en ambos casos fallamos al obtener producto de la PCR utilizando 5' RACE PCR modificada. La falta de regulación de este gen por el miR396 podría estar relacionado a la débil energía de hibridación del par miARN-gen blanco, aunque no podemos descartar que el miR396 esté controlando su traducción.

Otros dos potenciales genes blanco del miR396 eran At5g43060 y At3g14110 que codifican para la proteasa MMG4.7 y FLUORESCENT IN BLUE LIGHT (FLU), respectivamente. Y en ambos casos pudimos detectar el corte (Figura 4.4 C y D).

En contraste con el miR408 y miR396, donde tienen varios potenciales genes blanco, obtuvimos un solo potencial gen blanco para el miR159, un factor de transcripción MYB que regula desarrollo del estambre y polen [78]. El otro potencial gen blanco era At4g27330, conocido como NOZZLE/SPOROXYLESS. Este factor de transcripción, que participa en desarrollo del estambre y óvulo [99, 123], fue también validado por 5' RACE PCR (Figura 4.4 E). Es interesante notar que al menos las funciones de NOZZLE y PAA2 pueden estar directamente relacionadas con el rol de genes blanco, ya descritos anteriormente, del miR159 y miR408 respectivamente.

PAA2, FLU y NOZZLE fueron detectados en mono y dicotiledóneas mientras que PAC1 y MMG4.7 fueron detectadas solamente en dicotiledóneas (Figura 4.4 A-E). Las posiciones del sitio de unión del miARN-gen blanco están altamente conservadas y muchas de las posiciones variables corresponden a mismatches con el miARN o variaciones del tipo G-C/G-U. Además este método no requiere que el sitio del gen blanco esté conservada, sino más bien que haya una interacción predicha con el miARN en distintas especies. De esta manera el sitio de NOZZLE, donde la secuencia cambia en diferentes especies (Figura 4.4 E), pudo ser detectado por este enfoque.

Identificación de nuevos genes blanco permitiendo interacciones G-U.

Los genes blanco identificados utilizando la estrategia descrita anteriormente, tienen varios mismatches y bulges con sus miARNs, lo que puede ayudar a explicar por que se perdieron en trabajos anteriores. También notamos que muchas de estas nuevas interacciones miARN-gen

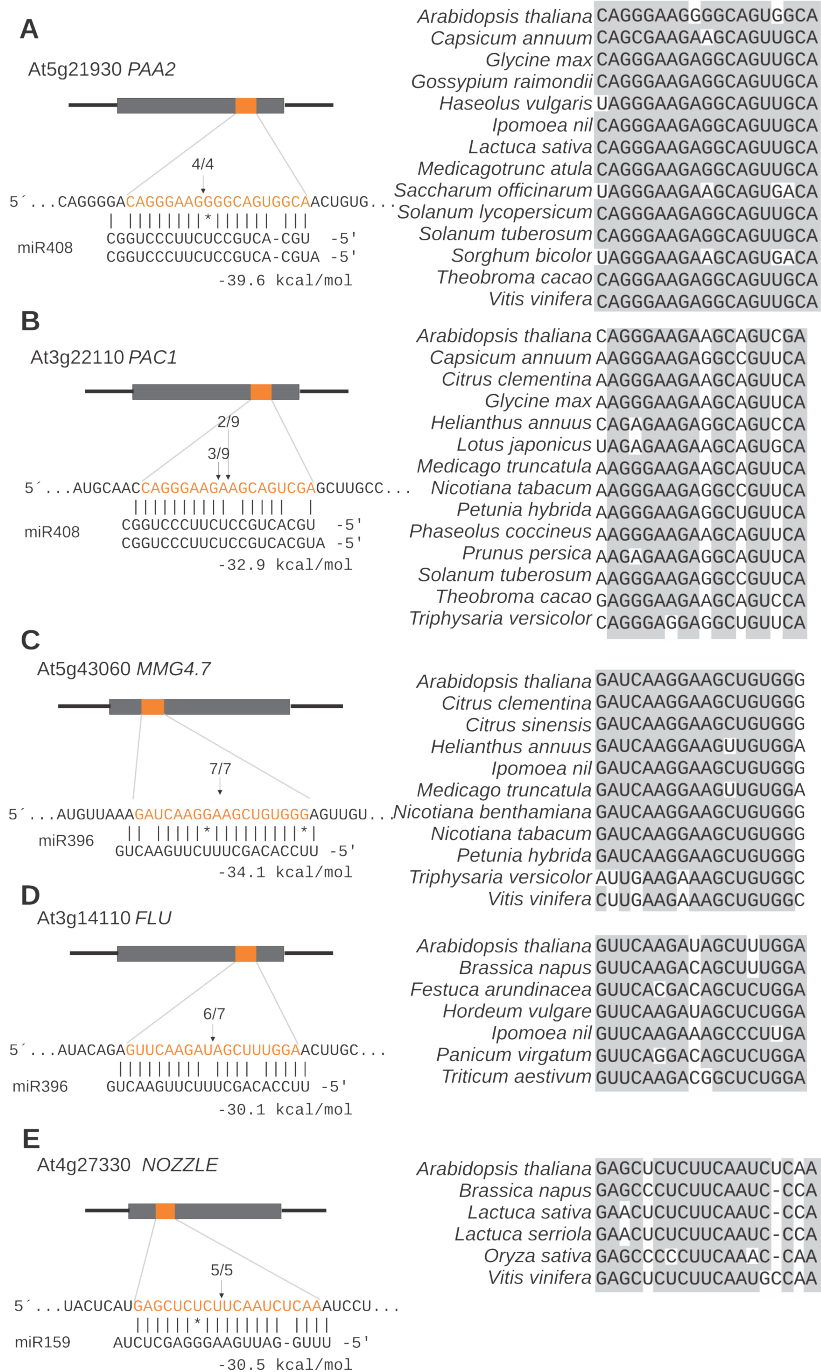


Figura 4.4 Nuevos genes blanco validados en *A. thaliana*. El alineamiento entre el miARN y los nuevos genes blanco identificados se muestran en la izquierda. La conservación evolutiva de la secuencia del sitio reconocido por el miARN en las especies seleccionadas se muestra a la derecha. La figura muestra las interacciones del miR408 con PAA2 (A), miR408 con PAC1 (B), miR396 con MMG4.7 (C), miR396 con FLU (D), miR159 con NOZZLE (E). Las flechas marcan el sitio de corte determinado por 5'RACE-PCR y los números indican la frecuencia de clonado de cada fragmento.

blanco contenían posiciones que variaban alternadamente entre G-C y G-U en distintas especies. Como consideramos G-U como mismatch en nuestra búsqueda inicial, decidimos realizar nuevamente la búsqueda con los miARNs consenso de 18nt pero permitiendo ahora 4 mismatches, donde al menos uno de ellos tiene que ser del tipo G-U. Esta búsqueda permitiría interacciones miARN-gen blanco con sólo 14 bases apareadas perfectamente.

Para compensar el uso de estos parámetros relajados en términos de mismatches, queremos que el gen blanco aparezca en al menos 10 especies distintas para aumentar la especificidad (Figura 4.5 A). Encontramos 125 potenciales genes blanco en *A. thaliana* teniendo en cuenta este criterio (Figura 4.5 A) y 34 de ellos no aparecían en las búsquedas anteriores. El gen blanco CSD2 regulado por el miR398, que no apareció anteriormente, fue detectado con estos parámetros.

Luego examinamos el último grupo de potenciales genes regulados por miARNs que estaban realizando funciones auxiliares a los genes blanco ya descritos para cada miARN. Y encontramos que el miR167 que regula factores de respuesta a auxina (ARFs), también regulaba potencialmente a un gen denominado IAA-ALANINE RESISTANT 3 (IAR3) (Figura 4.5 B y C), que está involucrado en el control de niveles libre de auxina [33, 90].

IAR3 en *Arabidopsis* tiene 3 mismatches con respecto al miR167, pero en la posición 12 de la interacción miARN-gen blanco, hay una interacción G-U en varias especies (Figura 4.5 B y C). La técnica de 5' RACE PCR confirmó que el gen realmente era gen blanco del miR167 (Figura 4.5 C).

Identificación de genes blanco específicos de *Solanaceae*.

Pensamos que la estrategia mostrada también se puede utilizar para encontrar genes blanco presentes específicamente en un grupo de especies relacionadas. Por lo tanto intentamos demostrar esto, encontrando potenciales genes blanco específicos de la familia de *Solanaceae*.

Elegimos esta familia en particular, ya que 6 especies estaban bien representadas en la biblioteca utilizada. La relación señal/ruido entre los genes blanco y las secuencias al azar era más de 2 cuando el filtro empírico o de conservación (en al menos 3 de las 6 especies *Solanaceae*) fueron aplicados (Figura 4.6 A). Curiosamente, al aplicar ambos filtros dio como resultado una relación señal/ruido por encima de 6 (Figura 4.6 A), confirmando nuestros previos hallazgos de que ambos filtros mejoran la detección de genes blanco de miARNs.

Encontramos 132 potenciales genes blanco presentes en al menos 3 especies *Solanaceae*. De este grupo, 41 genes no fueron detectados en otras especies (Figura 6B). El gen blanco más común fue la metalotioneína MT2A, presente en las 6 *Solanaceae*, como potencial gen

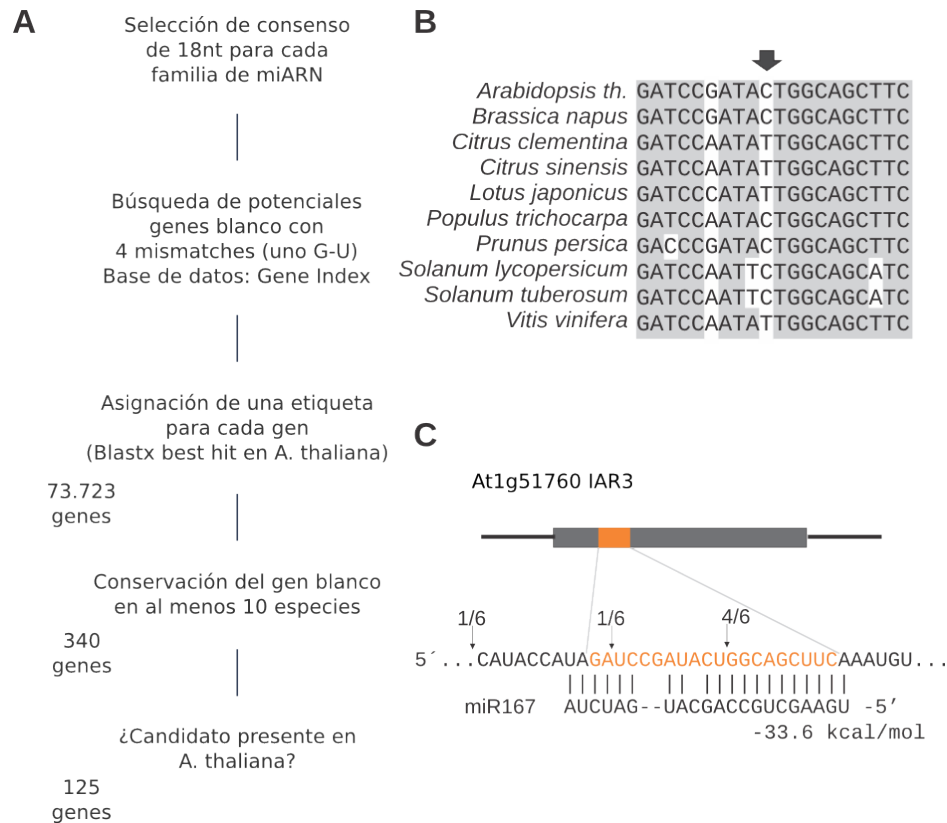


Figura 4.5 Identificación de un nuevo gen blanco de miARN, relajando los parámetros de interacción pero incrementando el parámetro de conservación evolutiva. **(A)** Esquema de la estrategia modificada para identificar genes blanco de miARNs. **(B)** Conservación del sitio blanco reconocido por el miARN en distintas especies. La flecha indica una variación de G-C o G-U con el miARN dependiendo de la especie. **(C)** Alineamiento en *Arabidopsis thaliana* del gen blanco IAR3 con el miR167. La flecha indica la posición del corte indicada por 5'RACE-PCR y el número indica la frecuencia de clonado de cada fragmento.

blanco del miR398, mientras que MT2B, homólogo de este gen, fue detectado en 5 especies (Figura 4.6 B-D).

Luego, aprovechamos las plantas transgénicas de tabaco que contienen un transgén 35S.mir398 (A.F. Lodeyro, N. Carrillo y J.F. Palatnik resultados no publicados) y chequeamos la expresión de estos genes. Encontramos que CSD2, un gen blanco conservado del miR398, disminuía su expresión > 10 veces en las plantas transgénicas 35S:miR398 comparadas con la planta salvaje (Figura 4.6 E). Curiosamente, observamos que tanto MT2A como MT2B disminuyeron sus niveles de transcripción > 5 veces en estas plantas (Figura 4.6 E). Estos resultados concuerdan con la regulación de MT2A y MT2B por el miR398, aunque no necesariamente demuestra una interacción directa. Además, estos resultados demuestran que los genes blanco presentes en un grupo específico de especies pueden ser encontrados utilizando esta estrategia.

4.2.2 comTAR: una herramienta para la predicción de genes blanco regulados por miARNs en plantas.

A partir de la estrategia descrita en el capítulo anterior, que fue utilizada para encontrar y validar experimentalmente genes blanco regulados por miARNs en *Arabidopsis thaliana*, desarrollamos una herramienta web denominada comTAR¹ (Conserved plant miRNA target prediction tool) [29]. La misma se puede utilizar para predecir potenciales genes blanco regulados por miARNs en plantas y está basada en la conservación evolutiva del par miARN-gen blanco con un número relajado de mismatches. ComTAR permite distintas opciones/parámetros de búsqueda que pueden ser modificados por el usuario:

- Filtro de mismatch: Solamente un mismatch está permitido entre la posición 1 y la 11 de la secuencia del miARN consenso. (Sí/No).
- Corte por energía de hibridación: Se define que un gen blanco es predicho si la mínima energía de hibridación está por debajo del corte elegido.
- El número mínimo de especies donde un mismo TAG está presente para un miARN particular.

Buscar potenciales genes blanco de miARN

Esta es la búsqueda por defecto. El usuario puede realizar la búsqueda de genes blanco de miARNs conservados. En la primer pantalla se muestra los potenciales genes blanco para un

¹<http://rnabiology.ibr-conicet.gov.ar/comtar>

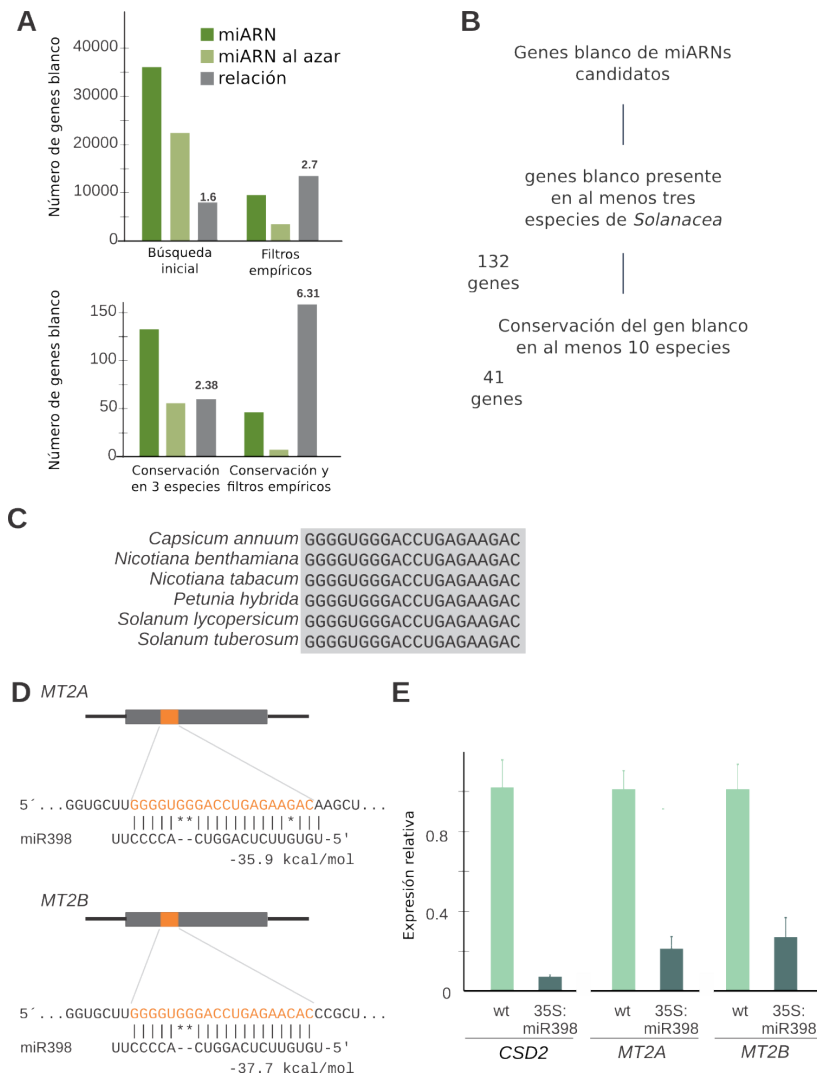


Figura 4.6 Identificación de genes blanco de miARN, específicos de *Solanaceae*. **(A)** Predicción de genes blanco de miARN en cinco especies de *Solanaceae*. El número de genes blanco de todos los miARNs conservados juntos se muestra luego de aplicar distintos filtros. También se muestran los genes blanco obtenidos a partir de las secuencias al azar. **(B)** Esquema que muestra la estrategia para identificar genes blanco específicos de *Solanaceae*. **(C)** Conservación del sitio reconocido por el miR398 con MT2A específico de *Solanaceae*. **(D)** Esquema que muestra el sitio de unión entre el miR398 y MT2A y MT2B. **(E)** Niveles de transcritos de CSD2, MT2A y MT2B en plantas salvajes y plantas transgénicas de tabaco (cv Petit havana) que sobreexpresan el miR398.

miARN dado (Figura 4.7), con una breve descripción del gen, la familia a la que pertenece y además en cuantas y cuáles especies está presente. También, para cada especie que está presente, se tiene acceso por pantalla al alineamiento del miARN-gen blanco, la energía de hibridación y los filtros empíricos de interacciones conocidas del par miARN-gen blanco (Figura 4.8).

Buscar familias de potenciales genes blanco de miARN

Debido a que los miARNs en plantas en general regulan genes que codifican a proteínas de las misma familias, la herramienta tiene otra funcionalidad donde permite la búsqueda de genes agrupados por familias en vez de agruparlos por TAG. De este modo genes en distintas especies con diferentes TAG, pero que pertenecen a la misma familia pueden ser detectados como familias de potenciales genes blanco.

¿Es este gen un potencial gen blanco de algun miARN conservado?

El usuario puede introducir un locus TAG en particular (tanto de Arabidopsis como el 'gene ID' del Phytozome) y se identifica si este gen en particular puede ser un potencial gen blanco de algun miARN y en cuantas especies aparece. En Arabidopsis se utiliza el LocusID como identificador, mientras que en Phytozome este identificador varía según la especie y se puede ver la precedencia de cada especie en el sitio de Phytozome.

Buscar tu secuencia particular

En esta parte del programa el usuario puede realizar la búsqueda de nuevos ARNs pequeños teniendo en cuenta que la secuencia introducida tiene que ser de 18nt de largo (posiciones 2-19). Luego de la búsqueda, se da un link al usuario y después de unas horas, cuando haya sido procesado el cálculo, el usuario puede entrar a ese link y navegar los resultados por pantalla.

comTAR
conserved plant miRNA target prediction tool

Find potential miRNA targets

Find potential miRNA target families

Is this gene a potential miRNA target?

Targets

miRNA: miR398_GTGTTCTCAGGTCACCCC
MFE cutoff: -29.232 kcal/mol
MM Filter: Yes
Species: All

Go Back

Arabidopsis Tag [?]	Count [?]	Species [?]	Target description [?]	Gene family [?]	Alignments [?]
AT1G12520	16	Show/Hide	copper chaperone for SOD1		View
AT1G08830	9	Show/Hide	copper/zinc superoxide dismutase 1		View
AT3G15640	4	Show/Hide	Rubredoxin-like superfamily protein		View
AT1G80230	3	Show/Hide	Rubredoxin-like superfamily protein		View
AT2G26975	3	Show/Hide	Ctr copper transporter family		View
AT2G40340	3	Show/Hide	Integrase-type DNA-binding superfamily protein	AP2-EREBP Transcription Factor Family	View
AT3G27200	3	Show/Hide	Cupredoxin superfamily protein	Miscellaneous Membrane Protein Families	View
AT3G43860	3	Show/Hide	glycosyl hydrolase 9A4	Glycoside Hydrolase Gene Families	View
AT4G00050	3	Show/Hide	basic helix-loop-helix (bHLH) DNA-binding superfamily protein	basic Helix-Loop-Helix (bHLH) Transcription Factor	View
AT5G12220	3	Show/Hide	las1-like family protein		View

Figura 4.7 Resultados de la búsqueda con parámetros por defecto para el miR398

Sequence ID	Species	5'-target-3' Alignment 3'-miRNA-5' [?]	MFE [?]
Aqua_013_00504.1	Aquilegia coerulea	TGGGCGACCTGGGAACAT * * * * * CCCCACTGGACTCTTGTG	-31.7
471402	Arabidopsis lyrata	TGGGAGACCTGGGAACAC * * * * * CCCCACTGGACTCTTGTG	-32.1
AT1G12520.1	Arabidopsis thaliana	TGGGAGACCTGGGAACAC * * * * * CCCCACTGGACTCTTGTG	-32.1
Bradi5g18900.3	Brachypodium distachyon	TTGGTGACCTGGGAACGC ** * * * * CCCCACTGGACTCTTGTG	-33.5
Bra026968	Brassica rapa	TGGGCGACCTGGGAACAC * * * * * CCCCACTGGACTCTTGTG	-32.5
Carubv10011816m	Capsella rubella	TGGGAGACCTGGGAACAC * * * * * CCCCACTGGACTCTTGTG	-32.1
evm.model.supercontig_29.47	Carica papaya	TAGGTGACCTGAGAACAT ** * * * * CCCCACTGGACTCTTGTG	-34.2
Ciclev10021134m	Citrus clementina	TTGGTGACCTGGGAACAC ** * * * * CCCCACTGGACTCTTGTG	-33.9
orange1.1g020436m	Citrus sinensis	TTGGTGACCTGGGAACAg ** * * * * CCCCACTGGACTCTTGTG	-32.7

Figura 4.8 Parte de la salida de comTAR mostrando el par miR398/SOD1 (At1g12520) en diferentes especies

Capítulos 5

Estudios genómicos sobre la biogénesis de miARN en plantas

5.1 Introducción

Los miARNs son una clase de ARNs de 20-22nt de largo que son originados de genes endógenos y regulan otros ARNs por complementariedad de bases [115]. Se distinguen de otros ARNs pequeños por su biogénesis única que involucra el corte preciso del tallo del precursor con estructura de tallo-burbuja [?] (stem-loop). Aunque la evidencia actual indica que los miARNs han surgido y especializada de forma independiente en animales y las plantas, su biogénesis depende del reconocimiento de claves estructurales ubicadas en los precursores de miARN [12, 17, 31].

En nuestro grupo actualmente se está estudiando la biogenesis de miARNs, específicamente como los precursores son reconocidos y procesados en plantas [18] Estos precursores tienen una estructura de tallo-burbuja característica [?], que se cree que proporciona las claves para el procesamiento del mismo y la liberación de los ARN pequeños de 21 nt. Mientras que los precursores de miARN en animales tienen estructuras homogéneas, los precursores de miARNs en plantas constituyen una amplia gama de estructuras, y sus longitud pueden variar entre 50 y 900 nucleótidos [18, 31]. En particular, la biogénesis de los miARNs es un proceso clave porque determina la secuencia exacta de nucleótidos del ARN pequeño funcional. Si bien en el caso de animales está claro cuáles elementos estructurales son reconocidos en los precursores durante su procesamiento, poco se sabía sobre el reconocimiento de los precursores de plantas por la maquinaria de procesamiento.

Muchos precursores en plantas tienen un tallo de ~15 nt debajo del duplex miARN/miARN* seguido por un loop interno, que sirve como una señal estructural de reconocimiento

por la maquinaria de procesamiento [50, 75, 103, 116]. Sin embargo, este determinante de procesamiento no se encuentra en todos los precursores [?]. Además, la biogénesis de los miARNs conservados evolutivamente como ser el miR319 y miR159 comienzan con un corte al lado del loop interno y continúa con 3 cortes adicionales en una dirección de burbuja a base hasta que finalmente el miARN es liberado [3, 18]. Se ha demostrado que otros precursores de plantas liberan otros ARNs pequeños además del miARN [? ?], aunque los mecanismos de procesamiento subyacentes eran desconocidos.

5.2 Resultados

5.2.1

En esta parte del proyecto de tesis y en el marco de una colaboración con el grupo del Dr. Blake Meyers (Delaware, USA), el cual se especializa en secuenciación y análisis de ARN pequeños, nos propusimos entender cómo se procesan los precursores de miARNs plantas. Colegas del laboratorio realizaron una estrategia para analizar sistemáticamente intermediarios de procesamiento de miARNs y caracterizar la biogénesis de la mayoría de los miARNs conservados presentes en *Arabidopsis thaliana* mediante técnicas de secuenciación de alto rendimiento, utilizando los equipos de última generación disponibles en Delaware (USA). Esta técnica desarrollada en el laboratorio se conoce como sPARE [?] (del inglés Specific Parallel Amplification of RNA Ends). Utilizando esta técnica encontramos que los miARNs son procesados por cuatro mecanismos, dependientes de la dirección secuencial de la maquinaria de procesamiento y del número de cortes requeridos para liberar el miARN. La clasificación de los precursores, teniendo en cuenta los mecanismos de procesamiento, reveló determinantes estructurales específicos para cada grupo. Se encontró que la complejidad de las vías de procesamiento de miARN se produce tanto en precursores jóvenes como en conservados y que los miembros de la misma familia pueden ser procesados de diferentes maneras. Además hemos observado que diferentes determinantes estructurales compiten por la maquinaria de procesamiento y que miRNAs alternativos pueden ser generados a partir de un único precursor. Los resultados ofrecen una explicación para la diversidad estructural de los genes de precursores de miARN en plantas y nuevas perspectivas hacia la comprensión de la biogénesis de los ARNs pequeños (Bologna et al., 2013).

Análisis de datos y precursores detectados

Mediante la cantidad de cortes detectados la técnica de SPARE permite definir si el mecanismo es base a loop o loop a base. Esta técnica arroja una gran cantidad de datos producto de

la secuenciación de alto rendimiento, por lo que se necesita de un enfoque bioinformático para la interpretación de los resultados. Por la gran cantidad de precursores a estudiar y el número de bibliotecas se necesitó un análisis previo de los datos y una forma de presentarlos. Para esto construimos e implementamos un pipeline bioinformático utilizando "in-house" scripts y datos disponibles de miRBASE para poder analizar los datos de las bibliotecas de deep-sequencing obtenidos a partir de la técnica de SPARE. Además desarrollamos una herramienta web para visualizar estos datos utilizando MySQL como base de datos para almacenar los datos procesados (Figura 8).

Un precursor fue considerado como detectado si más de tres lecturas corresponden a la secuencia de ese precursor. De esta manera encontramos fragmentos de ARN que corresponden a 129 precursores, 71 de ellos de miARNs conservados y 58 de miARNs jóvenes (Figura 9). Y con la ayuda de la herramienta pudimos definir la dirección de procesamiento de la mayoría de los precursores en *A. thaliana*. De los cuales 32 de ellos fueron definidos como procesados por un mecanismo de base a loop, ya que se encontraron los cortes en la parte proximal del duplex miARN/miARN* sin detectar cortes en la parte de arriba del duplex, como en el caso del miR168a, miR172b y el miR395b (Figura 10). Además encontramos 16 precursores de miARNs conservados con cortes detectados (>5%) en el lado distal del miARN/miARN* los cuales fueron definidos como loop a base (Figura 11).

Estructura secundaria de los precursores

Para ver si había diferencias estructurales para los precursores con diferentes mecanismos de procesamiento, determinamos la estructura secundaria de precursores detectados que se procesan en dirección base a loop (Figura 12) y los que se procesan loop a base (Figura 13). Obtuvimos las estructuras secundarias para cada precursor calculada a partir de la herramienta mfold (33) con los parámetros por default a 37°C de temperatura. Definimos a una coincidencia en cada posición con un 0, mientras que bulges y mismatches los consideramos como 1. El lado proximal del duplex miARN/miARN* fue definido como la posición +1 y analizamos desde la posición -25 a la posición +40 (Figura 12 y 13).

De esta manera los precursores pueden padecer cuatro mecanismos de procesamiento con distintas características (Figura 5.1).

- En los precursores con un mecanismo **corto de base a loop**, un loop interno seguido por un tallo inferior de ~15nt especifica la posición del primer corte. Esta estructura se encuentra en la mayoría de familias de miARNs [75, 102, 116]. A pesar de que el tallo puede contener bulges, la transición de un loop interno (simple hebra) al tallo inferior es bastante marcada, y tres pares de bases apareadas generalmente definen el comienzo

del stem inferior del precursor [18]. El segundo corte procede a una distancia fija de ~ 21 nt desde la posición del primer corte.

- En los precursores con un mecanismo **secuencial de base a loop** (ej: familia del miR169), el primer corte procede como en los cortos de base a loop, pero luego son necesario dos cortes más para liberar el miARN, generando en el proceso niveles bajos de RNA pequeños adicionales [18].
- En los precursores con un mecanismo **cortos de loop a base** (ej: familia del miR156 y miR160), el procesamiento es guiado por un tallo superior, y son necesarios dos cortes para liberar el miARN maduro. La región terminal de estos precursores tienen una largo conservado de ~ 42 y un loop pequeño [18].
- En los precursores con un mecanismo **secuencial de loop a base** (ej: familia del miR319 y miR159), cuatro cortes secuenciales por DCL1 son los encargados de procesar los precursores de miARNs. En general muestran un tallo largo superior, del cual otros ARNs pequeños son generados [3, 16, 18]

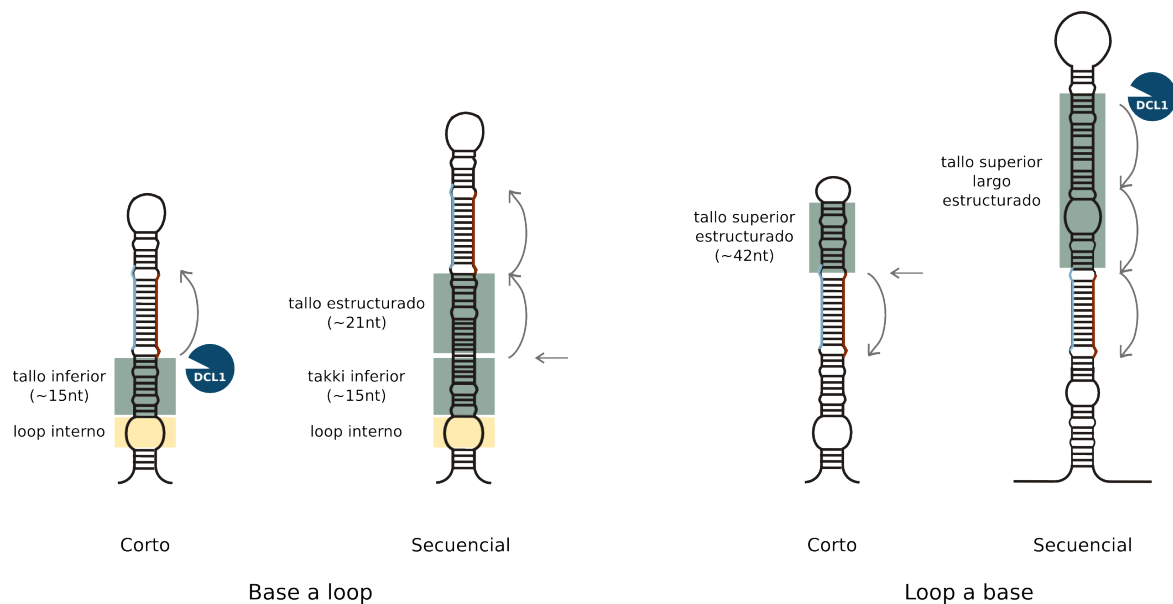


Figura 5.1 Distintos mecanismos de procesamiento de miARNs en plantas

Que los precursores de miARNs en plantas sean procesados de diferentes maneras, nos llevó a especular que su patrón de evolución también puede ser diferente y podrían estar vinculados a su mecanismo de procesamiento.

5.2.2 Enfoque bioinformático para el estudio de la evolución y biogénesis de miARNs en plantas

Conclusiones 6

Conclusiones

6.1 Primera parte

En cuanto a la primera parte, a través de diferentes estrategias y estudios, hemos alcanzado las siguientes conclusiones:

- Diseñamos una estrategia para identificar genes blanco regulados por miARNs en plantas, basado en la conservación evolutiva del par microARN-gen blanco.
- El enfoque requiere que la interacción miARN-gen blanco, pueda ocurrir en el contexto de un conjunto mínimo de parámetros que interactúan en diferentes especies. Pero la secuencia del gen blanco en sí, no necesariamente tiene que estar conservada.
- Además, nuestro enfoque permite ajustar el número de especies requeridas como un filtro para realizar la búsqueda con diferentes sensibilidades y relaciones señal/ruido.
- Utilizando esta estrategia identificamos y validamos experimentalmente nuevos genes blanco en *A. thaliana*, a pesar de que este sistema ya había sido estudiado en detalles en distintos enfoques genómicos a gran escala ([1, 4, 44, 54, 88, 100]).
- Tres de los nuevos genes blanco validados tienen bulges. Parámetros empíricos usualmente le otorgan una gran penalidad a ellos, que puede llegar a ser el doble que un mismatch regular [54], sin embargo es probable que genes blancos con bulges asimétricos sean más frecuente de lo que se pensaba previamente en plantas.
- El enfoque ofrece una estrategia alternativa a otras predicciones que se basan en parámetros empíricos del par miARN-gen blanco [4, 31, 38, 54].

- Una ventaja de la estrategia presentada es que las interacciones miARN-gen blanco conservadas probablemente participen en procesos biológicos relevantes.
- Además, esta estrategia puede ser fácilmente modificada para incorporar datos de otras bibliotecas, y/o para realizar la búsqueda de genes blanco presentes en un grupo específico de especies de plantas.

6.2 Segunda parte

Recientemente, en el laboratorio hemos desarrollado un enfoque genómico a gran escala para estudiar el procesamiento de miARNs en plantas y hemos determinado de esta manera el mecanismo de procesamiento de la mayoría de los miARNs de *A. thaliana* conservados evolutivamente [18]. Hemos encontrado que los miARNs en plantas pueden ser procesados por cuatro mecanismos, dependientes de la dirección secuencial de la maquinaria de procesamiento y del número de cortes requeridos para liberar el miARN (Figura 5.1) [18]. Precusores procesados en el mismo mecanismo comparten determinantes estructurales, explicando el gran rango de tamaño y forma observado en precusores de miARNs en plantas [18].

En la segunda parte de esta Tesis,

References

- [1] Addo-quaye, C., Eshoo, T. W., Bartel, D. P., and Axtell, M. J. (2009). Endogenous siRNA and microRNA targets identified by sequencing of the Arabidopsis degradome. *NIH Public Access*, 18(10):758–762.
- [2] Addo-Quaye, C., Miller, W., and Axtell, M. J. (2009a). CleaveLand: a pipeline for using degradome data to find cleaved small RNA targets. *Bioinformatics*, 25(1):130–131.
- [3] Addo-Quaye, C., Snyder, J. A., Park, Y. B., Li, Y. F., Sunkar, R., and Axtell, M. J. (2009b). Sliced microRNA targets and precise loop-first processing of MIR319 hairpins revealed by analysis of the *Physcomitrella patens* degradome. *RNA*, 15(12):2112–2121.
- [4] Allen, E., Xie, Z., Gustafson, A. M., and Carrington, J. C. (2005). microRNA-Directed Phasing during Trans-Acting siRNA Biogenesis in Plants. *Cell*, 121(2):207–221.
- [5] Altschup, S. F., Gish, W., Pennsylvania, T., and Park, U. (1990). Basic Local Alignment Search Tool. *Journal of Molecular Biology*, pages 403–410.
- [6] Anders, G., Mackowiak, S. D., Jens, M., Maaskola, J., Kuntzagk, A., Rajewsky, N., Landthaler, M., and Dieterich, C. (2012). doRiNA: a database of RNA interactions in post-transcriptional regulation. *Nucleic Acids Res.*, 40(Database issue):D180–186.
- [7] Arazi, T., Talmor-Neiman, M., Stav, R., Riese, M., Huijser, P., and Baulcombe, D. C. (2005). Cloning and characterization of micro-RNAs from moss. *The Plant journal : for cell and molecular biology*, 43(6):837–48.
- [8] Aukerman, M. J. and Sakai, H. (2003). Regulation of flowering time and floral organ identity by a MicroRNA and its APETALA2-like target genes. *Plant Cell*, 15(11):2730–2741.
- [9] Axtell, M. J. (2008). Evolution of microRNAs and their targets: are all microRNAs biologically relevant? *Biochimica et biophysica acta*, 1779(11):725–34.
- [10] Axtell, M. J. and Bartel, D. P. (2005). Antiquity of microRNAs and their targets in land plants. *Plant Cell*, 17(6):1658–1673.
- [11] Axtell, M. J. and Bowman, J. L. (2008). Evolution of plant microRNAs and their targets. *Trends in Plant Science*, 13(7):343–349.
- [12] Axtell, M. J., Westholm, J. O., and Lai, E. C. (2011). Vive la différence: biogenesis and evolution of microRNAs in plants and animals. *Genome Biol.*, 12(4):221.

- [13] Bartel, D. P., Lee, R., and Feinbaum, R. (2004). MicroRNAs: Genomics , Biogenesis , Mechanism , and Function Genomics : The miRNA Genes. *Cell*, 116:281–297.
- [14] Baulcombe, D. (2004). RNA silencing in plants. *Nature*, 431(7006):356–363.
- [15] Bo, X. and Wang, S. (2005). TargetFinder: a software for antisense oligonucleotide target site selection based on MAST and secondary structures of target mRNA. *Bioinformatics*, 21(8):1401–1402.
- [16] Bologna, N. G., Mateos, J. L., Bresso, E. G., and Palatnik, J. F. (2009). A loop-to-base processing mechanism underlies the biogenesis of plant microRNAs miR319 and miR159. *The EMBO journal*, 28(23):3646–56.
- [17] Bologna, N. G., Schapire, A. L., and Palatnik, J. F. (2012). Processing of plant microrna precursors. *Briefings in Functional Genomics*.
- [18] Bologna, N. G., Schapire, A. L., Zhai, J., Chorostecki, U., Boissbouvier, J., Meyers, B. C., and Palatnik, J. F. (2013). Multiple RNA recognition patterns during microRNA biogenesis in plants. *Genome research*, 23(10):1675–89.
- [19] Bonnet, E., He, Y., Billiau, K., and Van de Peer, Y. (2010). TAPIR, a web server for the prediction of plant microRNA targets, including target mimics. *Bioinformatics*, 26(12):1566–1568.
- [20] Borsani, O., Zhu, J., Verslues, P. E., Sunkar, R., and Zhu, J. K. (2005). Endogenous siRNAs derived from a pair of natural cis-antisense transcripts regulate salt tolerance in Arabidopsis. *Cell*, 123(7):1279–1291.
- [21] Bowman, J. L. (2004). Class III HD-Zip gene regulation, the golden fleece of ARGONAUTE activity? *Bioessays*, 26(9):938–942.
- [22] Brodersen, P., Sakvarelidze-Achard, L., Bruun-Rasmussen, M., Dunoyer, P., Yamamoto, Y. Y., Sieburth, L., and Voinnet, O. (2008). Widespread translational inhibition by plant miRNAs and siRNAs. *Science*, 320(5880):1185–1190.
- [23] Carrington, J. C. and Ambros, V. (2003). Role of microRNAs in plant and animal development. *Science*, 301(5631):336–338.
- [24] Chapman, E. J. and Carrington, J. C. (2007). Specialization and evolution of endogenous small RNA pathways. *Nat. Rev. Genet.*, 8(11):884–896.
- [25] Chekanova, J. A., Gregory, B. D., Reverdatto, S. V., Chen, H., Kumar, R., Hooker, T., Yazaki, J., Li, P., Skiba, N., Peng, Q., Alonso, J., Brukhin, V., Grossniklaus, U., Ecker, J. R., and Belostotsky, D. A. (2007). Genome-wide high-resolution mapping of exosome substrates reveals hidden features in the Arabidopsis transcriptome. *Cell*, 131(7):1340–1353.
- [26] Chen, H. M., Chen, L. T., Patel, K., Li, Y. H., Baulcombe, D. C., and Wu, S. H. (2010). 22-Nucleotide RNAs trigger secondary siRNA biogenesis in plants. *Proc. Natl. Acad. Sci. U.S.A.*, 107(34):15269–15274.

- [27] Chen, X. (2004). A microRNA as a translational repressor of APETALA2 in Arabidopsis flower development. *Science*, 303(5666):2022–2025.
- [28] Chorostecki, U., Crosa, V. A., Lodeyro, A. F., Bologna, N. G., Martin, A. P., Carrillo, N., Schommer, C., and Palatnik, J. F. (2012). Identification of new microRNA-regulated genes by conserved targeting in plant species. *Nucleic Acids Research*.
- [29] Chorostecki, U. and Palatnik, J. F. (2014). comTAR: a web tool for the prediction and characterization of conserved microRNA targets in plants. *Bioinformatics (Oxford, England)*, 30(14):2066–7.
- [30] Cuperus, J. T., Carbonell, A., Fahlgren, N., Garcia-Ruiz, H., Burke, R. T., Takeda, A., Sullivan, C. M., Gilbert, S. D., Montgomery, T. A., and Carrington, J. C. (2010). Unique functionality of 22-nt miRNAs in triggering RDR6-dependent siRNA biogenesis from target transcripts in Arabidopsis. *Nat. Struct. Mol. Biol.*, 17(8):997–1003.
- [31] Cuperus, J. T., Fahlgren, N., and Carrington, J. C. (2011). Evolution and functional diversification of mirna genes. *The Plant cell*, 23(2):431–442.
- [32] Dai, X. and Zhao, P. X. (2011). psRNATarget: a plant small RNA target analysis server. *Nucleic Acids Res.*, 39(Web Server issue):W155–159.
- [33] Davies, R. T., Goetz, D. H., Lasswell, J., Anderson, M. N., and Bartel, B. (1999). IAR3 Encodes an Auxin Conjugate Hydrolase from Arabidopsis. *Plant Cell*, 11(March):365–376.
- [34] Debernardi, J. M., Rodriguez, R. E., Mecchia, M. A., and Palatnik, J. F. (2012). Functional Specialization of the Plant miR396 Regulatory Network through Distinct MicroRNA–Target Interactions. *PLoS Genet*, 8(1):e1002419.
- [35] Ding, Y., Chan, C. Y., and Lawrence, C. E. (2004). Sfold web server for statistical folding and rational design of nucleic acids. *Nucleic Acids Res.*, 32(Web Server issue):W135–141.
- [36] Dugas, D. V. and Bartel, B. (2008). Sucrose induction of Arabidopsis miR398 represses two Cu/Zn superoxide dismutases. *Plant Mol. Biol.*, 67(4):403–417.
- [37] Fabian, M. R., Sonenberg, N., and Filipowicz, W. (2010). Regulation of mRNA Translation and Stability by microRNAs. *Annual Review of Biochemistry*, (79):351–79.
- [38] Fahlgren, N. and Carrington, J. (2010). mirna target prediction in plants. In Meyers, B. C. and Green, P. J., editors, *Plant MicroRNAs*, volume 592 of *Methods in Molecular Biology*, pages 51–57. Humana Press.
- [39] Fahlgren, N., Jogdeo, S., Kasschau, K. D., Sullivan, C. M., Chapman, E. J., Laubinger, S., Smith, L. M., Dasenko, M., Givan, S. a., Weigel, D., and Carrington, J. C. (2010). MicroRNA gene evolution in Arabidopsis lyrata and Arabidopsis thaliana. *The Plant cell*, 22(4):1074–89.
- [40] Floyd, S. K. and Bowman, J. L. (2004). Gene regulation: ancient microRNA target sequences in plants. *Nature*, 428(6982):485–486.

- [41] Friedman, R. C., Farh, K. K., Burge, C. B., and Bartel, D. P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.*, 19(1):92–105.
- [42] Gasciolli, V., Mallory, A. C., Bartel, D. P., and Vaucheret, H. (2005). Partially redundant functions of Arabidopsis DICER-like enzymes and a role for DCL4 in producing trans-acting siRNAs. *Curr. Biol.*, 15(16):1494–1500.
- [43] German, M. A., Luo, S., Schroth, G., Meyers, B. C., and Green, P. J. (2009). Construction of Parallel Analysis of RNA Ends (PARE) libraries for the study of cleaved miRNA targets and the RNA degradome. *Nat Protoc.*, 4(3):356–362.
- [44] German, M. a., Pillay, M., Jeong, D.-H., Hetawal, A., Luo, S., Janardhanan, P., Kannan, V., Rymarquis, L. a., Nobuta, K., German, R., De Paoli, E., Lu, C., Schroth, G., Meyers, B. C., and Green, P. J. (2008). Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends. *Nature biotechnology*, 26(8):941–6.
- [45] Giegerich, R., Rehmsmeier, M., Steffen, P., and Ho, M. (2004). Fast and effective prediction of microRNA / target duplexes. (2003):1507–1517.
- [46] Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., and Rokhsar, D. S. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic acids research*, 40(Database issue):D1178–86.
- [47] Grimson, A., Farh, K. K., Johnston, W. K., Garrett-Engele, P., Lim, L. P., and Bartel, D. P. (2007). MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell*, 27(1):91–105.
- [48] Guo, H., Ingolia, N. T., Weissman, J. S., and Bartel, D. P. (2010). Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*, 466(7308):835–840.
- [49] Hamilton, A. J. and Baulcombe, D. C. (1999). A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science*, 286(5441):950–952.
- [50] Han, J., Lee, Y., Yeom, K. H., Nam, J. W., Heo, I., Rhee, J. K., Sohn, S. Y., Cho, Y., Zhang, B. T., and Kim, V. N. (2006). Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. *Cell*, 125(5):887–901.
- [51] Han, M.-H., Goud, S., Song, L., and Fedoroff, N. (2004). T@Articlepmid14691535, Author="Stark, A. and Brennecke, J. and Russell, R. B. and Cohen, S. M. ", Title="Identification of Drosophila MicroRNA targets", Journal="PLoS Biol.", Year="2003", Volume="1", Number="3", Pages="E60", Month="Dec" he Arabidopsis double-stranded RNA-binding protein HYL1 plays a role in microRNA-mediated gene regulation. *Proceedings of the National Academy of Sciences of the United States of America*, 101(4):1093–8.
- [52] Hofacker, I. L. (2003). Vienna RNA secondary structure server. *Nucleic Acids Res.*, 31(13):3429–3431.
- [53] Jha, A. and Shankar, R. (2011). Employing machine learning for reliable miRNA target identification in plants. *BMC Genomics*, 12:636.

- [54] Jones-Rhoades, M. W. and Bartel, D. P. (2004). Computational identification of plant micrnas and their targets, including a stress-induced mirna. *Molecular Cell*, 14(6):787 – 799.
- [55] Jones-Rhoades, M. W., Bartel, D. P., and Bartel, B. (2006). MicroRNAs and their regulatory roles in plants. *Annual review of plant biology*, 57:19–53.
- [56] Kakrana, A., Hammond, R., Patel, P., Nakano, M., and Meyers, B. C. (2014). sPARTA: a parallelized pipeline for integrated analysis of plant miRNA and cleaved mRNA data sets, including new miRNA target-identification software. *Nucleic Acids Res.*, 42(18):e139.
- [57] Kasschau, K. D., Xie, Z., Allen, E., Llave, C., Chapman, E. J., Krizan, K. a., and Carrington, J. C. (2003). P1/HC-Pro, a Viral Suppressor of RNA Silencing, Interferes with Arabidopsis Development and miRNA Function. *Developmental Cell*, 4(2):205–217.
- [58] Katiyar-Agarwal, S., Morgan, R., Dahlbeck, D., Borsani, O., Villegas, A., Zhu, J. K., Staskawicz, B. J., and Jin, H. (2006). A pathogen-inducible endogenous siRNA in plant immunity. *Proc. Natl. Acad. Sci. U.S.A.*, 103(47):18002–18007.
- [59] Kozomara, A. and Griffiths-Jones, S. (2014). miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic acids research*, 42(Database issue):D68–73.
- [60] Krüger, J. and Rehmsmeier, M. (2006). Rnahybrid: microrna target prediction easy, fast and flexible. *Nucleic Acids Research*, 34(suppl 2):W451–W454.
- [61] Lai, E. C. (2004). Predicting and validating microRNA targets. *Genome Biol.*, 5(9):115.
- [62] Lanet, E., Delannoy, E., Sormani, R., Floris, M., Brodersen, P., Crete, P., Voinnet, O., and Robaglia, C. (2009). Biochemical evidence for translational repression by Arabidopsis microRNAs. *Plant Cell*, 21(6):1762–1768.
- [63] Laufs P, Peaucelle A, M. H. T. J. (2004). MicroRNA regulation of the CUC genes is required for boundary size control in Arabidopsis meristems. *Development*, 12(5):622–627.
- [64] Lee, R. C., Feinbaum, R. L., and Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5):843–854.
- [65] Lekprasert, P., Mayhew, M., and Ohler, U. (2011). Assessing the utility of thermodynamic features for microRNA target prediction under relaxed seed and no conservation requirements. *PLoS ONE*, 6(6):e20622.
- [66] Lewis, B. P., Burge, C. B., and 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.
- [67] Lewis, B. P., Shih, I. H., Jones-Rhoades, M. W., Bartel, D. P., and Burge, C. B. (2003). Prediction of mammalian microRNA targets. *Cell*, 115(7):787–798.
- [68] Llave, C., Xie, Z., Kasschau, K. D., and Carrington, J. C. (2002). Cleavage of Scarecrow-like mRNA Targets Directed by a Class of Arabidopsis miRNA. *Science*, 297(September):2053–2056.

- [69] Lobbes, D., Rallapalli, G., Schmidt, D. D., Martin, C., and Clarke, J. (2006). SERRATE: a new player on the plant microRNA scene. *EMBO reports*, 7(10):1052–8.
- [70] Mallory, A. C., Reinhart, B. J., Jones-rhoades, M. W., Zamore, P. D., Kathryn, M., and Bartel, D. P. (2004). MicroRNA control of PHABULOSA in leaf development : importance of pairing to the microRNA 5' region. *Embo journal*, 23(16):3356–3364.
- [71] Manavella, P. A., Koenig, D., and Weigel, D. (2012). Plant secondary siRNA production determined by microRNA-duplex structure. *Proc. Natl. Acad. Sci. U.S.A.*, 109(7):2461–2466.
- [72] Maragkakis, M., Alexiou, P., Papadopoulos, G. L., Reczko, M., Dalamagas, T., Giannopoulos, G., Goumas, G., Koukis, E., Kourtis, K., Simossis, V. A., Sethupathy, P., Vergoulis, T., Koziris, N., Sellis, T., Tsanakas, P., and Hatzigeorgiou, A. G. (2009). Accurate microRNA target prediction correlates with protein repression levels. *BMC Bioinformatics*, 10:295.
- [73] Maragkakis, M., Vergoulis, T., Alexiou, P., Reczko, M., Plomaritou, K., Gousis, M., Kourtis, K., Koziris, N., Dalamagas, T., and Hatzigeorgiou, A. G. (2011). DIANA-microT Web server upgrade supports Fly and Worm miRNA target prediction and bibliographic miRNA to disease association. *Nucleic Acids Res.*, 39(Web Server issue):W145–148.
- [74] Markham, N. R. and Zuker, M. (2005). DINAMelt web server for nucleic acid melting prediction. *Nucleic Acids Res.*, 33(Web Server issue):W577–581.
- [75] Mateos, J. L., Bologna, N. G., Chorostecki, U., and Palatnik, J. F. (2010). Identification of microRNA processing determinants by random mutagenesis of Arabidopsis MIR172a precursor. *Current Biology*, 20(1):49–54.
- [76] Maunoury, N. (2011). AGO1 and AGO2 Act Redundantly in miR408-Mediated Plantacyanin Regulation. 6(12).
- [77] Mi, S., Cai, T., Hu, Y., Chen, Y., Hodges, E., Ni, F., Wu, L., Li, S., Zhou, H., Long, C., Chen, S., Hannon, G. J., and Qi, Y. (2008). Sorting of small RNAs into Arabidopsis argonaute complexes is directed by the 5' terminal nucleotide. *Cell*, 133(1):116–127.
- [78] Millar, A. a. and Gubler, F. (2005). The Arabidopsis GAMYB-like genes, MYB33 and MYB65, are microRNA-regulated genes that redundantly facilitate anther development. *The Plant cell*, 17(3):705–21.
- [79] Montgomery, T. A., Howell, M. D., Cuperus, J. T., Li, D., Hansen, J. E., Alexander, A. L., Chapman, E. J., Fahlgren, N., Allen, E., and Carrington, J. C. (2008a). Specificity of ARGONAUTE7-miR390 Interaction and Dual Functionality in TAS3 Trans -Acting siRNA Formation. pages 128–141.
- [80] Montgomery, T. A., Yoo, S. J., Fahlgren, N., Gilbert, S. D., Howell, M. D., Sullivan, C. M., Alexander, A., Nguyen, G., Allen, E., Ahn, J. H., and Carrington, J. C. (2008b). AGO1-miR173 complex initiates phased siRNA formation in plants. *Proc. Natl. Acad. Sci. U.S.A.*, 105(51):20055–20062.

- [81] Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443 – 453.
- [82] Niyogi, K. K., Pilon, M., Shikanai, T., Abdel-ghany, S. E., and Mu, P. (2005). Two P-Type ATPases Are Required for Copper Delivery in *Arabidopsis thaliana* Chloroplasts. 17(April):1233–1251.
- [83] Ossowski, S., Schwab, R., and Weigel, D. (2008). Gene silencing in plants using artificial microRNAs and other small RNAs. *Plant J.*, 53(4):674–690.
- [84] Palatnik, J. F., Allen, E., Wu, X., Schommer, C., Schwab, R., Carrington, J. C., and Weigel, D. (2003). Control of leaf morphogenesis by microRNAs. *Nature*, 425(6955):257–263.
- [85] Palatnik, J. F., Wollmann, H., Schommer, C., Schwab, R., Boisbouvier, J., Rodriguez, R., Warthmann, N., Allen, E., Dezulian, T., Huson, D., Carrington, J. C., and Weigel, D. (2007). Sequence and expression differences underlie functional specialization of *Arabidopsis* microRNAs miR159 and miR319. *Developmental cell*, 13(1):115–25.
- [86] Paraskevopoulou, M. D., Georgakilas, G., Kostoulas, N., Vlachos, I. S., Vergoulis, T., Reczko, M., Filippidis, C., Dalamagas, T., and Hatzigeorgiou, A. G. (2013). DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. *Nucleic Acids Res.*, 41(Web Server issue):W169–173.
- [87] Pasquinelli, A. E., Reinhart, B. J., Slack, F., Martindale, M. Q., Kuroda, M. I., Maller, B., Hayward, D. C., Ball, E. E., Degan, B., Muller, P., Spring, J., Srinivasan, A., Fishman, M., Finnerty, J., Corbo, J., Levine, M., Leahy, P., Davidson, E., and Ruvkun, G. (2000). Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, 408(6808):86–89.
- [88] Rajagopalan, R., Vaucheret, H., Trejo, J., and Bartel, D. P. (2006). A diverse and evolutionarily fluid set of microRNAs in *Arabidopsis thaliana*. *Genes & development*, 20(24):3407–25.
- [89] Ramachandran, V. and Chen, X. (2008). Small RNA metabolism in *Arabidopsis*. *Trends Plant Sci.*, 13(7):368–374.
- [90] Rampey, R. A., Leclere, S., Kowalczyk, M., Ljung, K., Bartel, B., Biology, C., and Texas, R. A. R. (2004). A Family of Auxin-Conjugate Hydrolases That Contributes to Free Indole-3-Acetic Acid Levels during *Arabidopsis* Germination 1. 135(June):978–988.
- [91] Rehmsmeier, M., Steffen, P., Hochsmann, M., and Giegerich, R. (2004). Fast and effective prediction of microRNA/target duplexes. *RNA*, 10(10):1507–1517.
- [92] Reinhart, B. J., Weinstein, E. G., Rhoades, M. W., Bartel, B., and Bartel, D. P. (2002). MicroRNAs in plants. pages 1616–1626.
- [93] Ren, G., Xie, M., Zhang, S., Vinovskis, C., Chen, X., and Yu, B. (2014). Methylation protects microRNAs from an AGO1-associated activity that uridylates 5' RNA fragments generated by AGO1 cleavage. *Proc. Natl. Acad. Sci. U.S.A.*, 111(17):6365–6370.

- [94] Rhoades, M. W., Reinhart, B. J., Lim, L. P., Burge, C. B., Bartel, B., and Bartel, D. P. (2002). Prediction of plant microRNA targets. *Cell*, 110(4):513 – 520.
- [95] Rodriguez, R. E., Mecchia, M. A., Debernardi, J. M., Schommer, C., Weigel, D., and Palatnik, J. F. (2010). Control of cell proliferation in *Arabidopsis thaliana* by microRNA miR396. 112:103–112.
- [96] Rubio-Somoza, I., Cuperus, J. T., Weigel, D., and Carrington, J. C. (2009). Regulation and functional specialization of small RNA-target nodes during plant development. *Curr. Opin. Plant Biol.*, 12(5):622–627.
- [97] Rubio-Somoza, I. and Weigel, D. (2011). MicroRNA networks and developmental plasticity in plants. *Trends Plant Sci.*, 16(5):258–264.
- [98] Schauer, S. E., Jacobsen, S. E., Meinke, D. W., and Ray, A. (2002). DICER-LIKE1: blind men and elephants in *Arabidopsis* development. *Trends Plant Sci.*, 7(11):487–491.
- [99] Schiefthaler, Balasubramanian, Sieber, Chevalier, Wisman, and Schneitz (1999). Molecular analysis of NOZZLE , a gene involved in pattern formation and early sporogenesis during sex organ development in *Arabidopsis thaliana*. *Proc. Natl Acad. Sci.*, 96(September):11664–11669.
- [100] Schwab, R., Palatnik, J. F., Riester, M., Schommer, C., Schmid, M., and Weigel, D. (2005). Specific effects of micrnas on the plant transcriptome. *Developmental Cell*, 8(4):517 – 527.
- [101] Shenoy, A. and Blelloch, R. H. (2014). Regulation of microRNA function in somatic stem cell proliferation and differentiation. *Nat. Rev. Mol. Cell Biol.*, 15(9):565–576.
- [102] Song, L., Axtell, M. J., and Fedoroff, N. V. (2010). RNA secondary structural determinants of miRNA precursor processing in *Arabidopsis*. *Curr. Biol.*, 20(1):37–41.
- [103] Song, L., Han, M. H., Lesicka, J., and Fedoroff, N. (2007). *Arabidopsis* primary microRNA processing proteins HYL1 and DCL1 define a nuclear body distinct from the Cajal body. *Proc. Natl. Acad. Sci. U.S.A.*, 104(13):5437–5442.
- [104] Souret, F. F., Kastenmayer, J. P., and Green, P. J. (2004). AtXRN4 degrades mRNA in *Arabidopsis* and its substrates include selected miRNA targets. *Mol. Cell*, 15(2):173–183.
- [105] Srivastava, P. K., Moturu, T. R., Pandey, P., Baldwin, I. T., and Pandey, S. P. (2014). A comparison of performance of plant miRNA target prediction tools and the characterization of features for genome-wide target prediction. *BMC Genomics*, 15:348.
- [106] Stark, A., Brennecke, J., Bushati, N., Russell, R. B., and Cohen, S. M. (2005). Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. *Cell*, 123(6):1133–1146.
- [107] Stark, A., Brennecke, J., Russell, R. B., and Cohen, S. M. (2003). Identification of *Drosophila* MicroRNA targets. *PLoS Biol.*, 1(3):E60.
- [108] Takeda, A., Iwasaki, S., and Watanabe, T. (2008). The Mechanism Selecting the Guide Strand from Small RNA Duplexes is Different Among Argonaute Proteins. 49(4):493–500.

- [109] Thadani, R. and Tammi, M. T. (2006). MicroTar: predicting microRNA targets from RNA duplexes. *BMC Bioinformatics*, 7 Suppl 5:S20.
- [110] Vaucheret, H. (2006). Post-transcriptional small RNA pathways in plants: mechanisms and regulations. *Genes Dev.*, 20(7):759–771.
- [111] Vaucheret, H. (2008). Plant ARGONAUTES ´. *Cell*, (May).
- [112] Vazquez, F., Cre, P., and Bartel, D. P. (2004a). The action of ARGONAUTE1 in the miRNA pathway and its regulation by the miRNA pathway are crucial for plant development. pages 1187–1197.
- [113] Vazquez, F., Gascioli, V., Cr  t  , P., and Vaucheret, H. (2004b). The nuclear dsRNA binding protein HYL1 is required for microRNA accumulation and plant development, but not posttranscriptional transgene silencing. *Current biology : CB*, 14(4):346–51.
- [114] Voinnet, O. (2001). RNA silencing as a plant immune system against viruses. *Trends Genet.*, 17(8):449–459.
- [115] Voinnet, O. (2009). Origin, Biogenesis, and Activity of Plant MicroRNAs. *Cell*, 136(4):669–687.
- [116] Werner, S., Wollmann, H., Schneeberger, K., and Weigel, D. (2010). Structure determinants for accurate processing of miR172a in *Arabidopsis thaliana*. *Curr. Biol.*, 20(1):42–48.
- [117] Xie, F. and Zhang, B. (2010). Target-align: a tool for plant microRNA target identification. *Bioinformatics*, 26(23):3002–3003.
- [118] Xie, Z., Allen, E., Fahlgren, N., Calamar, A., Givan, S. A., and Carrington, J. C. (2005). Expression of *Arabidopsis* MIRNA Genes . *Plant Physiology*, 138(August):2145–2154.
- [119] Xie, Z., Kasschau, K. D., and Carrington, J. C. (2003). Negative Feedback Regulation of Dicer-Like1 in *Arabidopsis* by microRNA-Guided mRNA Degradation. *Current Biology*, 13(9):784–789.
- [120] Xie, Z. and Qi, X. (2008). Diverse small RNA-directed silencing pathways in plants. *Biochim. Biophys. Acta*, 1779(11):720–724.
- [121] Yamasaki, H., Abdel-Ghany, S. E., Cohu, C. M., Kobayashi, Y., Shikanai, T., and Pilon, M. (2007). Regulation of copper homeostasis by micro-RNA in *Arabidopsis*. *The Journal of biological chemistry*, 282(22):16369–78.
- [122] Yan, T., Yoo, D., Berardini, T. Z., Mueller, L. A., Weems, D. C., Weng, S., Cherry, J. M., and Rhee, S. Y. (2005). Patmatch: a program for finding patterns in peptide and nucleotide sequences. *Nucleic Acids Research*, 33(suppl 2):W262–W266.
- [123] Yang, W.-c., Ye, D., Xu, J., and Sundaresan, V. (1999). The SPOROCTELESS gene of *Arabidopsis* is required for initiation of sporogenesis and encodes a novel nuclear protein. pages 2108–2117.
- [124] Yoshikawa, M., Peragine, A., Park, M. Y., and Poethig, R. S. (2005). A pathway for the biogenesis of trans-acting siRNAs in *Arabidopsis*. *Genes Dev.*, 19(18):2164–2175.

-
- [125] Zhang, B., Pan, X., Cannon, C. H., Cobb, G. P., and Anderson, T. A. (2006). Conservation and divergence of plant microRNA genes. *Plant J.*, 46(2):243–259.
- [126] Zheng, Y., Li, Y. F., Sunkar, R., and Zhang, W. (2012). SeqTar: an effective method for identifying microRNA guided cleavage sites from degradome of polyadenylated transcripts in plants. *Nucleic Acids Res.*, 40(4):e28.
- [127] Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, 31(13):3406–3415.

Appendix A

Anexo

Predicción de genes regulados por microARNs.

```
1  #!/usr/bin/perl -w
2  use DBI;
3  use FindBin;
4  use strict;
5
6  # Constants
7  package Constants;
8  use constant MM => 4;
9  use constant MM_TYPE => 'is';
10 use constant DB => 'patmatch_2013';
11 use constant PLANTDB => 'phytozome';
12
13 # MySQL
14 my $host = "localhost";
15 my $userid;
16 my $passwd;
17
18 # Arguments
19 my $pattern = $ARGV[0] || die "Must give pattern";
20 my $table_name = $ARGV[1] || die "Must give table name";
21
22 $pattern = uc($pattern);
23 $pattern =~ tr/uu/tT/;
24
25 my $mismatches = MM;
26 my $mismatch_types = MM_TYPE;
27
28 # Check pattern
29 my $$SYNTAX_CHECKER_BIN = "perl " . $FindBin::Bin . '/patmatchPatternChecker.pl';
30 my $patStatus = $$SYNTAX_CHECKER_BIN 'dna' $pattern;
31 chomp($patStatus);
32
33 # Variables
34 my ($gen_name, $hit_start, $hit_end, $target, $mirna) = '';
35 my $deltaG = 0;
36
37 # Files RNAhybrid target and microRNA 5'3' y blast
38 my $target_file = $FindBin::Bin . "/extra_files/target_rnahybrid.txt";
39 my $mirna_file = $FindBin::Bin . "/extra_files/mirna_rnahybrid.txt";
40 my $blast_file_sequence = $FindBin::Bin . "/extra_files/" . "seq_blast.txt";
41 my $blast_database = $FindBin::Bin . "/extra_files/blast/TAIR10_pep_20101214_updated";
42
43 my $sequence_file = '';
44 my $tab = $table_name . "_" . $pattern;
45
46 hybrid_mirna_file($pattern, $mirna_file);
47 fill_table_mirnas($pattern, $table_name, $tab);
```

```

48
49 my $stab_db = create_table($stab);
50 my @specie_db = species(PLANTDB);
51
52 foreach my $file (@specie_db){
53
54     my $fasta_file = $file->{'fasta'};
55     my $specie = $file->{'specie'};
56
57     $sequence_file = $FindBin::Bin . '/databases/' . PLANTDB . "/" . $fasta_file;
58     print $fasta_file . " - ";
59     print $specie . "\n";
60
61     my ($gen_sv, $target_sv, $align_sv, $mir_sv, $deltaG_sv, $nro_mm_sv, $ins_sv) = '';
62     my ($del_sv, $sust_mm_sv, $gu_mm_sv) = '';
63     my ($filtro_mm, $family, $sub_family, $alias) = '';
64
65     my %res_blast;
66     my @res_ned, @res_mm;
67     my @res_family;
68
69     if ($patStatus eq "OK") { # syntax OK, run PatMatch
70         my $SCAN_PIPELINE = "perl " . $FindBin::Bin . '/scan_pipeline.pl';
71         open(OUTPUT, "$SCAN_PIPELINE -c '$pattern' '$sequence_file' '$mismatches' '$mismatch_types' |");
72         open(INFO, $sequence_file);
73         while (my $line = <OUTPUT>) {
74             if ( $line =~ />(.*?):\[(\d*?),(\d*?)\]/ ) {
75                 #keep output parameters
76                 $gen_name = $1;
77                 $hit_start = $2;
78                 $hit_end = $3;
79                 $gen_sv = $gen_name;
80             }
81             elsif ( $line =~ /(.*)(\d*?)/ ) {
82                 $target = trim($1);
83                 $mirna = $pattern;
84                 $mirna = reverse complement($mirna);
85                 # Needleman
86                 @res_ned = Needleman($mirna, $target);
87                 $target_sv = $res_ned[0];
88                 $align_sv = $res_ned[1];
89                 $mir_sv = $res_ned[2];
90                 # Mismatches
91                 @res_mm = mismatches($target_sv, $align_sv, $mir_sv);
92                 $nro_mm_sv = trim($res_mm[0]);
93                 $ins_sv = trim($res_mm[1]);
94                 $del_sv = trim($res_mm[2]);
95                 $sust_mm_sv = trim($res_mm[3]);
96                 $gu_mm_sv = trim($res_mm[4]);
97
98                 hybrid_target_file($target, $target_file);
99
100                 #RNAhybrid
101                 my $RNAhybrid_BIN = 'RNAhybrid -d 0 -m 12000';
102                 my $RNAhybrid_Status = '$RNAhybrid_BIN -t' $target_file '-q' $mirna_file;
103                 chomp($RNAhybrid_Status);
104                 if ( $RNAhybrid_Status =~ m/mfe: (.*) kcal \/ mol/ ){
105                     $deltaG = $1;
106                     $deltaG =~ s/ //g;
107                     $deltaG_sv = $deltaG;
108                 }
109
110                 $filtro_mm = mm_position($align_sv);
111                 insert($stab_db, $specie, $gen_sv, $target_sv, $align_sv, $mir_sv, $nro_mm_sv, $ins_sv, $del_sv, $sust_mm_sv, $gu_mm_sv, $filtro_mm, $deltaG_sv);
112             }
113         }
114         close(OUTPUT);
115         close(INFO);
116         blast_from_db_annotation($fasta_file, $specie, $stab_db);
117         family_from_db($fasta_file, $stab_db);
118     }
119
120     else
121     {
122         print "Invalid pattern syntax";

```

```

123     }
124 }
125
126 sub fill_table_mirnas {
127     my ($sequence, $name, $table_reference) = @_;
128     $target = reverse(complement($sequence));
129     hybrid_target_file($target, $target_file);
130     my $hyb_perf;
131
132     #RNAhybrid
133     my $RNAhybrid_BIN = 'RNAhybrid -d 0 -m 12000';
134     my $RNAhybrid_Status = '$RNAhybrid_BIN -t' $target_file '-q' $mirna_file;
135     chomp($RNAhybrid_Status);
136     if ( $RNAhybrid_Status =~ m/mfe: (.*) kcal/mol/ ){
137         $deltaG = $1;
138         $deltaG =~ s/ //g;
139         $hyb_perf = $deltaG;
140     }
141
142     my $sth;
143     my $db = DB;
144     my $connectionInfo = "dbi:mysql:$db;$host";
145     my $dbh = DBI->connect($connectionInfo, $userid, $passwd);
146
147     my $query_insert = "INSERT INTO mirnas
148         (name, sequence, table_reference, hyb_perf)
149         values
150         ('$name',
151          '$sequence',
152          '$table_reference',
153          '$hyb_perf'
154         );";
155
156     $sth = $dbh->prepare($query_insert);
157     $sth->execute();
158 }
159
160
161 sub species {
162     (my $db_plants) = @_;
163
164     my $db = DB;
165     my $connectionInfo = "dbi:mysql:$db;$host";
166     my $sth;
167     my $dbh = DBI->connect($connectionInfo, $userid, $passwd);
168
169     my $specie;
170     my $fasta_file;
171     my $specie_data = {};
172
173     my $ref;
174     my @res;
175
176     my $query = "SELECT fasta, specie from plants where db = '$db_plants'";
177
178     $sth = $dbh->prepare($query);
179     $sth->execute();
180     $sth->bind_columns(\ $fasta_file, \ $specie);
181
182     if ($sth->rows > 0) {
183         while ($ref = $sth->fetchrow_hashref()) {
184             push(@res, $ref);
185         }
186     }
187
188     return @res;
189 }
190
191 sub insert {
192     my ($table_db, $specie, $gen, $target, $align, $mirna, $mm, $ins, $del, $sust, $gu, $filtro_mm, $deltag) = @_;
193     my $sth;
194     my $db = DB;
195     my $connectionInfo = "dbi:mysql:$db;$host";
196     my $dbh = DBI->connect($connectionInfo, $userid, $passwd);
197

```

```

198 my $query_insert = "INSERT INTO $table_db
199     (file ,gen ,target ,align ,mirna ,mm ,ins ,del ,sust ,gu ,filtro_mm ,deltag )
200     values
201     ( '$specie' , '$gen' , '$target' , '$align' , '$mirna' , '$mm' , '$ins' , '$del' , '$sust' , '$gu' , '$filtro_mm' , '$deltag' );";
202
203 $sth = $dbh->prepare($query_insert);
204 $sth->execute();
205
206 }
207 sub create_table {
208
209     my ($stabmiRNA) = @_;
210     my $sth;
211     my $db = DB;
212     my $connectionInfo = "dbi:mysql:$db;$host";
213     my $dbh = DBI->connect($connectionInfo , $userid , $passwd);
214
215     my $query_create = "CREATE TABLE $stabmiRNA (id INT AUTO_INCREMENT PRIMARY KEY,
216         file VARCHAR(60) NOT NULL,
217         gen VARCHAR(30) NOT NULL,
218         target VARCHAR(30) NOT NULL,
219         align VARCHAR(30) NOT NULL,
220         mirna VARCHAR(30) NOT NULL,
221         mm INT NOT NULL,
222         ins INT NOT NULL,
223         del INT NOT NULL,
224         sust INT NOT NULL,
225         gu INT NOT NULL,
226         similar_ath VARCHAR(20) NOT NULL,
227         similar_osa VARCHAR(20) NOT NULL,
228         filtro_mm INT NOT NULL,
229         family TEXT(1000) NOT NULL,
230         sub_family VARCHAR(20) NOT NULL,
231         alias VARCHAR(20) NOT NULL,
232         deltag FLOAT NOT NULL
233     );";
234
235     $sth = $dbh->prepare($query_create);
236     $sth->execute();
237     $sth->finish();
238     $dbh->disconnect;
239     return $stabmiRNA;
240 }
241
242
243 sub blast_from_db_annotation {
244     my ($file , $specie , $mirna) = @_;
245
246     my $db= DB;
247     my $connectionInfo="dbi:mysql:$db;$host";
248     my $sth;
249     my $dbh = DBI->connect($connectionInfo , $userid , $passwd);
250     my $annotation_table = 'annotation_' . $file;
251     my $annotation;
252     my $query = "SELECT annotation from plants where fasta = '$file'";
253
254     $sth = $dbh->prepare($query);
255     $sth->execute();
256     $sth->bind_columns(\ $annotation);
257
258
259     $sth->fetch();
260
261     unless($annotation eq 'empty') {
262         my $query_update = "UPDATE $mirna miR
263             LEFT JOIN $annotation_table a
264             ON miR.gen = a.gen
265             SET miR.similar_ath = SUBSTRING(a.similar_ath,1,9) ,
266                 miR.similar_osa = SUBSTRING(a.similar_osa,1,14)
267             WHERE file = '$specie'";
268
269         $sth = $dbh->prepare($query_update);
270         $sth->execute();
271     }
272 }

```



```

273
274
275 sub family{
276     (my $gen) = @_;
277
278     my $db = DB;
279     my $connectionInfo="dbi:mysql:$db;$host";
280     my $sth;
281     my $dbh = DBI->connect($connectionInfo,$userid,$passwd);
282     my $family;
283     my $sub_family;
284     my $alias;
285     my $query = "SELECT family ,sub_family ,gene_name from gene_families where locus_tag = '$gen'";
286
287     $sth = $dbh->prepare($query);
288     $sth->execute();
289     $sth->bind_columns(\ $family ,\ $sub_family ,\ $alias );
290
291     if ( $sth->rows > 0 ){
292         while( $sth->fetch() ) {
293             my @results = ( $family , $sub_family , $alias );
294             return @results;
295         }
296     }
297     else{
298         my @results = ( '', '', '' );
299         return @results;
300     }
301 }
302
303 sub family_from_db{
304     (my $file , $mirna) = @_;
305
306     my $db = DB;
307     my $connectionInfo="dbi:mysql:$db;$host";
308     my $sth;
309     my $dbh = DBI->connect($connectionInfo,$userid,$passwd);
310     my $gen;
311
312     my $query_update = "UPDATE $mirna miR
313         LEFT JOIN gene_families f
314         ON miR.similar_ath = f.locus_tag
315         SET miR.family      = f.family ,
316             miR.sub_family   = f.sub_family ,
317             miR.alias       = f.gene_name" ;
318
319     $sth = $dbh->prepare($query_update);
320     $sth->execute();
321 }
322
323 sub hybrid_mirna_file{
324     my ($mirna,$file_mirna) = @_;
325     open(MIRNA, ">$file_mirna") || die;
326     print MIRNA ">mirna\n$mirna\n";
327     close(MIRNA);
328 }
329
330 sub hybrid_target_file {
331     my ($target,$file_target) = @_;
332     open(OUT, ">$file_target") || die;
333     print OUT ">target\n$target\n";
334     close(OUT);
335 }
336
337 sub mm_position{
338     (my $align) = @_;
339     my @align = split (//,$align);
340     my $i = 0;
341     my $j = 0;
342     @align = reverse(@align);
343     foreach (@align) {
344         if ($align[$i] eq "*"){
345             if ($i+2 < 13 ){
346                 $j++;
347             }

```

```

348     }
349     $i++;
350 }
351 if ($j > 1 ){
352     return 0;
353 }
354 else{
355     return 1;
356 }
357 }
358 }
359
360 sub mismatches{
361     my ($target_mm,$align_mm,$mirna_mm) = @_;
362     my $pos = 0;
363     my $nro_mm = 0;
364     my $del_mm = 0;
365     my $ins_mm = 0;
366     my $sust_mm = 0;
367     my $gu_mm = 0;
368     my @align_mm = split(//, $align_mm);
369     my @target_mm = split(//, $target_mm);
370     my @mirna_mm = split(//, $mirna_mm);
371     # mismatches
372     foreach (@align_mm) {
373         if ($_ =~ /\*/) {
374             if (($target_mm[$pos] eq 'G' and $mirna_mm[$pos] eq 'T') or ($target_mm[$pos] eq 'T' and $mirna_mm[$pos] eq 'G')){
375                 $gu_mm++;
376             }
377             else{}
378             $nro_mm ++;
379         }
380         $pos++;
381     }
382     # deletions
383     foreach (@target_mm) {
384         if ($_ =~ /-/ ) {
385             $del_mm ++;
386         }
387     }
388     # insertions
389     foreach (@mirna_mm) {
390         if ($_ =~ /-/ ) {
391             $ins_mm ++;
392         }
393     }
394 }
395 # substitutions
396 $sust_mm = $nro_mm - ($del_mm + $ins_mm + $gu_mm);
397
398 my @results = ($nro_mm,$ins_mm,$del_mm,$sust_mm,$gu_mm);
399 return @results;
400 }
401 }
402
403 sub Needleman {
404     my ($seq1,$seq2) = @_;
405     # scoring scheme
406     my $MATCH = 1; # +1 for letters that match
407     my $MISMATCH = -1; # -1 for letters that mismatch
408     my $GAP = -1; # -1 for any gap
409
410     # initialization
411     my @matrix;
412     $matrix[0][0]{score} = 0;
413     $matrix[0][0]{pointer} = "none";
414
415     for(my $j = 1; $j <= length($seq1); $j++) {
416         $matrix[0][$j]{score} = $GAP * $j ;
417         $matrix[0][$j]{pointer} = "left";
418     }
419     for (my $i = 1; $i <= length($seq2); $i++) {
420         $matrix[$i][0]{score} = $GAP * $i ;
421         $matrix[$i][0]{pointer} = "up";
422     }

```

```

423
424 # fill
425 for(my $i = 1; $i <= length($seq2); $i++) {
426     for(my $j = 1; $j <= length($seq1); $j++) {
427         my ($diagonal_score, $left_score, $sup_score);
428
429         # calculate match score
430         my $letter1 = substr($seq1, $j-1, 1);
431         my $letter2 = substr($seq2, $i-1, 1);
432         if ($letter1 eq $letter2) {
433             $diagonal_score = $matrix[$i-1][$j-1]{score} + $MATCH;
434         }
435         else {
436             $diagonal_score = $matrix[$i-1][$j-1]{score} + $MISMATCH;
437         }
438
439         # calculate gap scores
440         $sup_score = $matrix[$i-1][$j]{score} + $GAP;
441         $left_score = $matrix[$i][$j-1]{score} + ($GAP*10);
442
443         # choose best score
444         if ($diagonal_score >= $sup_score) {
445             if ($diagonal_score >= $left_score) {
446                 $matrix[$i][$j]{score} = $diagonal_score;
447                 $matrix[$i][$j]{pointer} = "diagonal";
448             }
449             else {
450                 $matrix[$i][$j]{score} = $left_score;
451                 $matrix[$i][$j]{pointer} = "left";
452             }
453         }
454         else {
455             if ($sup_score >= $left_score) {
456                 $matrix[$i][$j]{score} = $sup_score;
457                 $matrix[$i][$j]{pointer} = "up";
458             }
459             else {
460                 $matrix[$i][$j]{score} = $left_score;
461                 $matrix[$i][$j]{pointer} = "left";
462             }
463         }
464     }
465 }
466
467 # trace-back
468
469 my $align1 = "";
470 my $align2 = "";
471 my $align3;
472
473 # start at last cell of matrix
474 my $j = length($seq1);
475 my $i = length($seq2);
476
477 while (1) {
478     last if $matrix[$i][$j]{pointer} eq "none"; # ends at first cell of matrix
479     my $letter1 = substr($seq1, $j-1, 1);
480     my $letter2 = substr($seq2, $i-1, 1);
481
482     if ($matrix[$i][$j]{pointer} eq "diagonal") {
483         if ($letter1 eq $letter2){
484             $align3 .= "|";
485         }
486         else
487         {
488             $align3 .= "*";
489         }
490         $align1 .= substr($seq1, $j-1, 1);
491         $align2 .= substr($seq2, $i-1, 1);
492         $i--;
493         $j--;
494     }
495
496     }
497     elsif ($matrix[$i][$j]{pointer} eq "left") {

```

```

498     $align1 .= substr($seq1, $j-1, 1);
499     $align2 .= "-";
500     $align3 .= "*";
501     $j--;
502
503 }
504 elseif ($matrix[$i][$j]{pointer} eq "up") {
505     $align1 .= "-";
506     $align2 .= substr($seq2, $i-1, 1);
507     $align3 .= "*";
508     $i--;
509 }
510 }
511 }
512
513 $align1 = reverse $align1;
514 $align2 = reverse $align2;
515 $align3 = reverse $align3;
516 $align1 = complement($align1);
517
518 my @results = ($align2, $align3, $align1);
519 return @results;
520 }
521 sub complement
522 {
523     my $sequence = shift;
524     $sequence =~ tr/atcgATCG/tagcTAGC/;
525     return $sequence;
526 }
527
528 sub trim
529 {
530     (my $string) = @_;
531     $string =~ s/^\s+//;
532     $string =~ s/\s+$//;
533     $string =~ s/\n//g;
534     return $string;
535 }

```

Table A.1 Especies y base de datos utilizadas para la búsquedas de genes blanco de miARNs conservados

Specie	Database
Allium_cepa	http://compbio.dfci.harvard.edu/tgi/
Aquilegia	http://compbio.dfci.harvard.edu/tgi/
Arabidopsis_thaliana	http://arabidopsis.org/
Beta_vulgaris	http://compbio.dfci.harvard.edu/tgi/
Brassica napus	http://compbio.dfci.harvard.edu/tgi/
Capsicum_annuum	http://compbio.dfci.harvard.edu/tgi/
Citrus_clementina	http://compbio.dfci.harvard.edu/tgi/
Citrus_sinensis	http://compbio.dfci.harvard.edu/tgi/
Coffea_canephora	http://compbio.dfci.harvard.edu/tgi/
Euphorbia_esula	http://compbio.dfci.harvard.edu/tgi/
Festuca_arundinacea	http://compbio.dfci.harvard.edu/tgi/
Glycine_max	http://compbio.dfci.harvard.edu/tgi/
Gossypium	http://compbio.dfci.harvard.edu/tgi/
Gossypium_raidmondii	http://compbio.dfci.harvard.edu/tgi/
Haseolus_vulgaris	http://compbio.dfci.harvard.edu/tgi/
Helianthus_annuus	http://compbio.dfci.harvard.edu/tgi/
Hordeum_vulgare	http://compbio.dfci.harvard.edu/tgi/
Ipomoea_nil	http://compbio.dfci.harvard.edu/tgi/
Lactuca_sativa	http://compbio.dfci.harvard.edu/tgi/
Lactuca_serriola	http://compbio.dfci.harvard.edu/tgi/
Lotus_japonicus	http://compbio.dfci.harvard.edu/tgi/
Malus_x_domestica	http://compbio.dfci.harvard.edu/tgi/
Medicago_truncatula	http://compbio.dfci.harvard.edu/tgi/
Mesembryanthemum_crystallinum	http://compbio.dfci.harvard.edu/tgi/
Nicotiana_benthamiana	http://compbio.dfci.harvard.edu/tgi/
Nicotiana_tabacum	http://compbio.dfci.harvard.edu/tgi/
Oryza_sativa	http://www.jcvi.org/
Panicum_virgatum	http://compbio.dfci.harvard.edu/tgi/
Petunia_hybrida	http://compbio.dfci.harvard.edu/tgi/
Phaseolus_coccineus	http://compbio.dfci.harvard.edu/tgi/
Populus	http://compbio.dfci.harvard.edu/tgi/
Prunus_persica	http://compbio.dfci.harvard.edu/tgi/
Saccharum_officinarum	http://compbio.dfci.harvard.edu/tgi/
Secale_cereale	http://compbio.dfci.harvard.edu/tgi/
Solanum_lycopersicum	http://compbio.dfci.harvard.edu/tgi/
Solanum_tuberosum	http://compbio.dfci.harvard.edu/tgi/
Sorghum_bicolor	http://compbio.dfci.harvard.edu/tgi/
Theobroma_cacao	http://compbio.dfci.harvard.edu/tgi/
triphysaria	http://compbio.dfci.harvard.edu/tgi/
Triphysaria_versicolor	http://compbio.dfci.harvard.edu/tgi/
Triticum_aestivum	http://compbio.dfci.harvard.edu/tgi/
Vitis_vinifera	http://compbio.dfci.harvard.edu/tgi/
Zea_mays	http://compbio.dfci.harvard.edu/tgi/

Table A.2 My caption

microRNA	Target	ID
miR156/miR157	SPL	At1g27370
miR156/miR157	SPL	At1g53160
miR156/miR157	SPL	At2g33810
miR156/miR157	SPL	At3g15270
miR156/miR157	SPL	At5g43270
miR156/miR157	SPL	At1g69170
miR156/miR157	SPL	At2g42200
miR156/miR157	SPL	At3g57920
miR156/miR157	SPL	At5g50670
miR159/miR319	TCP	At1g30210
miR159/miR319	TCP	At1g53230
miR159/miR319	TCP	At2g31070
miR159/miR319	MYB	At3g11440
miR159/miR319	TCP	At3g15030
miR159/miR319	TCP	At4g18390
miR159/miR319	MYB	At5g06100
miR159/miR319	MYB	At2g26950
miR159/miR319	MYB	At2g32460
miR159/miR319	MYB	At5g55020
miR160	ARF	At1g77850
miR160	ARF	At2g28350
miR160	ARF	At4g30080
miR162	DCL	At1g01040
miR164	NAC	At1g56010
miR164	NAC	At3g15170
miR164	NAC	At5g07680
miR164	NAC	At5g53950
miR164	NAC	At5g61430
miR164	NAC	At3g12977
miR164	NAC	At5g39610
miR166/miR165	HD-ZIPIII	At1g30490
miR166/miR165	HD-ZIPIII	At1g52150
miR166/miR165	HD-ZIPIII	At2g34710
miR166/miR165	HD-ZIPIII	At5g60690
miR166/miR165	HD-ZIPIII	At4g32880
miR167	ARF	At1g30330
miR167	ARF	At5g37020
miR168	AGO	At1g48410
miR169	HAP2	At1g17590
miR169	HAP2	At1g54160
miR169	HAP2	At1g72830
miR169	HAP2	At3g05690
miR169	HAP2	At3g20910
miR169	HAP2	At5g06510
miR170/miR171	SCL	At2g45160
miR170/miR171	SCL	At3g60630
miR170/miR171	SCL	At4g00150
miR172	AP2	At2g28550
miR172	AP2	At4g36920
miR172	AP2	At5g60120
miR172	AP2	At5g67180
miR172	AP2	At2g39250
miR172	AP2	At3g54990
miR390/miR391	TAS3	At3g17185
miR390/miR391	TAS3	At5g49615
miR390/miR391	TAS3	At5g57735
miR393	TIR1/AFB	At1g12820
miR393	bHLH	At3g23690
miR393	TIR1/AFB	At3g26810
miR393	TIR1/AFB	At3g62980
miR393	TIR1/AFB	At4g03190
miR394	F-Box	At1g27340
miR395	APS	At3g22890
miR395	AST	At5g10180
miR395	APS	At5g43780
miR395	APS	At4g14680
miR396	GRF	At2g22840
miR396	GRF	At2g36400
miR396	GRF	At2g45480
miR396	GRF	At4g24150
miR396	GRF	At4g37740
miR396	GRF	At5g53660
miR396	GRF	At3g52910
miR397	LAC	At2g29130
miR397	LAC	At2g38080
miR397	LAC	At5g60020
miR398	CSD	At1g08830
miR398	CSD	At2g28190
miR398	CytC oxidase	At3g15640
miR399	E2-UBC	At2g33770
miR399	E2-UBC	At2g33770
miR408	LAC	At2g30210
mir408	PLC	At2g02850
miR827	SPX	At1g02860

Table A.3 Oligonucleotide primers used for RT-qPCR

Gene	Locus ID	Forward primer	Reverse Primer
PAA2	At5g21930	GTCCTCTTATCAGGGGACAGG	CATAGTTGCTTGTGCAAGACTCAG
MYB33	At5g06100	CTATGGAAACCGACATTCACCTG	CTTGGCTTCCAGAAGCAACATATCG
NZZ	At4g27330	TCGGGTCAGGTTATGATCGA	AGGGTTTCCTTCCATGTAGCTCC
PP2A	At1g13320	CCTGCGGTAATAACTGCATCT	CTTCACTTAGCTCCACCAAGCA
tMT2A	tobacco	TACCCAGATTGAGCTACAACGAG	GCAGGAGATTCACCCATTTCATA
tMT2B	tobacco	TACCCAGATTGAGCTACAACGAA	AGGGGATTACCCATTTCATT

Table A.4 Oligonucleotide primers used for 5' RACE

Gen	Locus ID	5' RACE	5' RACE nested
General		CGACTGGAGCACGAGGACACTGA	GGACACTGACATGGACTGAAGGAGTA
PAA2	At5g21930	GACTTATGGAGCTGCAGAAAGTAATG	CATAGTTGCTTGTGCAAGACTCAG
IAR3	At1g51760	ATCTTCTGATCCCATTAATGGTTGCATCTCG	CATATTACGCTCGCTTGCCTTGTGATAACC
NZZ	At4g27330	CAITTTAAAGCTTCAAGGACAAATCAATGGTATTAGG	AGGGTTTCCTTCCATGTAGCTCC
MMG4.7	At5g43060	ATGGTAACAACCTTAGCATTTTTC	CITTCGGTATCAATACCWCCATT
UDP	At2g47650	AATGGGCCGACATGTTCTCC	CCTCGGTGATAGTCCATGGT
SVP	At2g22540	GCAACTTTCCTTCAITCATC	TTTCATCTGCCTCAGCTCAC
loricrin-related	At5g64550	ACCATGAGCTTTGCAGTAGT	CCTCAGCACTTCGTGTACAG
	At3g14110	CGGAAGGATCAGTCAGTCTC	CCCAGCTCGGTATAACAGTC
	At3g22110	GTTTCATCGCCAAAGGTAAC	CCAGGCGAATAAGACTAGAG
AVA-P2	At1g19910	CTCTAGACTGACCAGCTCGA	GGATGATACCAACAATGAGA