



**CENTRO DE INVESTIGACIÓN Y DE ESTUDIOS AVANZADOS DEL INSTITUTO
POLITÉCNICO NACIONAL**

UNIDAD DE GENÓMICA AVANZADA

**Identificación y caracterización de ARNs largos no codificantes conservados
en Brasicáceas**

Tesis que presenta:

M.C. Evelia Lorena Coss Navarrete

Para obtener el grado de:

Doctora en Ciencias

En la especialidad de:

Biotecnología de Plantas

Directora de tesis:

Dra. Selene Lizbeth Fernández Valverde

Irapuato, Guanajuato

Enero 2023

Este trabajo se realizó en el Centro de Investigación y de Estudios Avanzados del IPN - Unidad Irapuato.

Bajo la dirección de la Dra. Selene Lizbeth Fernández Valverde en el Laboratorio de Genómica funcional y evolutiva de ARN regulatorios perteneciente a la Unidad de Genómica Avanzada, Langebio.

En colaboración con la Dra. Katarzyna Oktaba Sosin en el Laboratorio de Regulación y Topología del genoma perteneciente al Departamento de Ingeniería Genética en CINVESTAV, Unidad Irapuato; y con el Dr. Charles Stewart Gillmor III en el Laboratorio de Genética y Epigenética de Semillas perteneciente a UGA.

Gracias al apoyo del Consejo Nacional de Ciencia y Tecnología (CONACyT) a través de la beca 781634 y a la Royal Society Newton Advanced Fellowship de la Dra. Fernandez-Valverde (NAF\R1\180303) .

Este trabajo se realizó del 4 de marzo del 2019 al 1 de diciembre del 2022.

Miembros del comité tutorial

Dr. Charles Stewart Gillmor III, Unidad de Genómica Avanzada CINVESTAV-Irapuato.

Dra. Katarzyna Oktaba Sosin, Departamento de Ingeniería Genética, CINVESTAV-Irapuato.

Dr. Luis José Delaye Arredondo, Departamento de Ingeniería Genética, CINVESTAV-Irapuato.

Dra. Gertrud Lund, Departamento de Ingeniería Genética, CINVESTAV- Irapuato.

Miembro externo del comité

Dra. Catalina Arenas Huertero. Profesora investigadora de la Universidad Autónoma de San Luis Potosí (UASLP) .

Fecha del examen de grado:

AGRADECIMIENTOS

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) por su apoyo económico durante mis estudios de doctorado (CVU 781634). Así como su apoyo en las becas en formato extraordinario y obtención de grado.

Al Centro de Investigación y de Estudios Avanzados del IPN (CINVESTAV), Unidad Langebio, por las facilidades, materiales y espacio de trabajo para llevar a cabo mis estudios y experimentos.

A mi asesora de tesis la Dra. Selene Fernández Valverde, por la confianza y esperanza que depositó en mí cuando nadie más lo hizo, por su paciencia, comprensión y empatía. Sus consejos para ser una mejor persona y una mejor estudiante. Gracias por guiarme hacia un nuevo rumbo en la ciencia médica y depositar la confianza necesaria para guiarme al laboratorio de la Dra. Alejandra Medina. No tendré cómo pagarte por todo lo bueno que has generado en mí, pero espero poder ser la luz en el camino de alguien más, como tu lo fuiste en el mío.

A la Dra. Kasia Oktaba por la paciencia y dedicación por la que se sentaba conmigo a repasar los protocolos, gracias por ese tiempo y trucos que me daba para hacer más eficiente mi trabajo. Gracias por abrirme un espacio en su laboratorio y gracias por todas las aportaciones que recibía de tu parte. Gracias por crecer conmigo, alegrarte y ser parte de los logros que poco a poco iba obteniendo.

Al Dr. Luis Delaye por la paciencia y la información que me brindó para intentar entender el enfoque evolutivo del proyecto. Gracias por escucharme y apoyarme durante una de las etapas más difíciles del doctorado, gracias por impulsarme a seguir adelante.

Al Dr. Stewart Gillmor por las aportaciones y recomendaciones que me brindó a lo largo de este tiempo. Gracias por el espacio que me hizo en el laboratorio, por permitirme rodearme de un ambiente cooperativo y estar al tanto de mis necesidades.

A la Dra. Gertrud Lund, aunque fue breve el tiempo en el que pudimos interactuar, gracias por apoyarme en los trámites necesarios para continuar con mi doctorado y el tiempo en que me escuchaste durante esa difícil temporada.

A los integrantes de secretaría académica, a Diana Barbosa Sabanero y Jesús Vega Medina por todo su apoyo durante mi maestría y doctorado en este instituto, por toda la comprensión y paciencia, así como apoyo técnico y en papeleos. Por facilitarnos las becas y apoyos. Así como a Antonio Cisneros Duran por los videos que hemos hecho juntos, por su apoyo técnico y visual.

A Araceli Fernandez Cortez por su apoyo con Mazorka, gracias por todas sus enseñanzas tanto profesionales como técnicas, así como los momentos de plática e intercambio de ideas.

A mis compañeros del Laboratorio de Biología Computacional (LBC), por todas las maravillosas personas que me tendieron la mano cuando empecé este camino en la bioinformática, gracias por su tiempo, aprendizaje y por mostrarte qué tipo de compañera, estudiante y postdoc quiero llegar a ser. Muy en especial a Jose Roberto Bermudez, Victor Flores, Jose Antonio Corona, Irving Garcia y America Colmenero.

A los nuevos amigos que hice y han perdurado a lo largo de los años en esta institución, gracias por las charlas amenas y tendidas, gracias por su amor y pasión en sus proyecto porque verlas ha sido gratificante para aprender y verlas crecer en los mismos. Muy en especial a Melanie Ruvalcaba, Monica Rodriguez y Angela Juarez.

A mis padres por todo el apoyo económico y emocional que han sido para mi a lo largo de estos años, gracias por apoyarme en mis sueños y verme crecer en la ciencia. Gracias por estar conmigo en todos los proyectos de investigación en los que me involucre, gracias por viajar conmigo y ser parte fundamental de esta historia. Gracias por darme una gran familia, gracias por mis hermanos y gracias por cuidar a mis mascotas por mi y amarlos tanto.

A Alejandro por ser mi compañero, mi amigo y mi pareja, gracias amor mío por estar en las buenas y en las malas conmigo, por animarme y ayudarme a levantarme, gracias por ser un motor en mi vida. Gracias por ayudarme a recordar quien soy y apoyarme a seguir mis sueños.

Y finalmente, y muy en especial a Dios, por darme toda la fuerza y fortaleza para enfrentar las adversidades y pruebas que he tenido que pasar a lo largo de esta tesis.

CONTENIDO

Lista de abreviaturas	1
Resumen	2
Abstract	3
Capítulo	1
Introducción	4
Descubrimiento de los lncRNAs	4
Características de los lncRNAs en eucariontes	4
Clasificación de los lncRNAs	5
Mecanismos funcionales de los lncRNAs	6
Tipos de conservación empleados en lncRNAs	8
Interés en el estudio en la familia Brassicaceae	9
Antecedentes	11
Identificación de lncRNAs en <i>A. thaliana</i>	11
lncRNAs funcionales en <i>A. thaliana</i>	12
Problemas en la anotación de lncRNAs en diversas bases de datos	13
Consideraciones en la conservación de lncRNAs	14
Algunos lncRNAs conservados y funcionales	15
Los lncRNAs presentan expresión órgano o etapa específica	16
Hipótesis y objetivos	18
- Hipótesis	18
- Objetivos	18
Objetivo general	18
Objetivos específicos	18
Capítulo	2
Anotación y clasificación funcional de lncRNAs en <i>A. thaliana</i>	19
2.1 Métodos	19
2.1.1 Selección de transcriptomas públicos	19
2.1.2 Generación de nuevos transcriptomas	19
2.1.3 Limpieza, alineamiento, ensamblaje y cuantificación de los transcriptomas	20
2.1.4 Identificación de lncRNAs	21
2.1.5 Clasificación de los lncRNAs	22
2.1.6 Análisis del potencial codificante	22
2.1.7 Comparación con otras bases de datos	22
2.1.8 Cuantificación de los lncRNAs por órganos y etapas	22
2.1.9 Redes de expresión y análisis de culpable por asociación	23
2.1.10 Visualización de los datos y scripts	23
2.2. Resultados	24
2.2.1 Identificación y clasificación de lncRNAs	24

2.2.2 Similitudes y diferencias con otras bases de datos	25
2.2.3 Identificación de lncRNAs con validez experimental	26
2.2.4 Patrones de expresión de lncRNAs	28
2.2.5 Expresión de lncRNA específicos de órgano y etapa	30
2.2.6 lncRNAs con funciones específicas de órgano conocidas	30
2.2.7 Co-expresión de lncRNAs con genes codificante	31
2.3. Discusiones	34
2.3.1 Aumento en el número de lncRNAs anotados	34
2.3.2 lncRNAs con expresión órgano y etapa del desarrollo específica	34
2.3.3 Similitudes entre los organismos	35
2.3.4 Los lncRNAs se encuentran implicados en procesos biológicos	35
2.4. Conclusiones	36
Capítulo	3
Conservación de lincRNAs por sintenia y microhomología	38
3.1 Métodos	38
3.1.1 Extracción de RNA y secuenciación	38
3.1.2 Identificación y clasificación de lncRNAs	38
3.1.3 Búsqueda de proteínas ortólogas	38
3.1.4 Conservación por sintenia y microhomología de lincRNAs	39
3.1.5 Proporción de lincRNAs conservados a través de la filogenia	40
3.1.6 Porcentaje de identidad de secuencia en los lincRNAs	40
3.1.7 Identificación de lincRNAs con expresión conservada	40
3.1.8 Ganancia y pérdida de lincRNAs	40
3.1.9 Visualización de los datos y scripts	41
3.2 Resultados	41
3.2.1 Características conservadas de los lincRNAs	41
3.2.2 Identificación de los lincRNAs ortólogos en la familia Brassicaceae	42
3.2.3 Relación entre el número de parálogos identificados en lincRNAs y las duplicaciones del genoma	46
3.2.4 lincRNAs identificados como ancestrales	47
3.2.5 lincRNAs con expresión órgano específica	50
3.2.6 lincRNAs con expresión conservada	53
3.2.7 lincRNAs con expresión conservada en raíces	55
3.2.8 lincRNAs ortologos funcionales	58
3.2.9 Conservación de IPS1 y sus parálogos	60
3.2.10 Ganancia y pérdida de lincRNAs	63
3.3 Discusión	64
3.3.1 Características de los lincRNAs	65
3.3.2 lincRNAs son importantes elementos regulatorios en plantas	65
3.3.3 lincRNAs ortologos funcionales	66
3.4 Conclusiones	67
Conclusiones generales	69

Perspectivas	70
Referencias bibliográficas	71
Archivos suplementarios	94
Datos suplementarios	94
• Capítulo 2: Anotación y clasificación funcional de lncRNAs en <i>A. thaliana</i>	94
• Capítulo 3: Conservación de lincRNAs por sintenia y microhomología	95
Tablas suplementarias	96
Figuras suplementarias	112
Publicaciones	129

Lista de abreviaturas

DEG	Differential expression genes.
FDR	False Discovery Rate.
logFC	Logarithm base-2 of the fold change.
miRNA	microRNA.
lncRNA	Long non-coding RNA.
RNA-Seq	RNA Sequencing.
IPS1	Induced by Phosphate Starvation1
dpg	Días post-germinación
Pi	Fosfato
aa	Aminoácidos.
ORF	Open Reading Frame
sORF	small Open Reading Frame / Marco de lectura abierto pequeño
COOLAIR	COLD ASSISTED INTRONIC NONCODING RNA
COLDAIR	Cold Induced Long Antisense Intragenic RNA
RBP	RNA binding Protein
mya	million years ago.
AtlincRNAs	lincRNAs identificados en <i>Arabidopsis thaliana</i>
CrlicRNAs	lincRNAs identificados en <i>Capsella rubella</i>
TplincRNAs	lincRNAs identificados en <i>Thellgungiella parvula</i>
BrlicRNAs	lincRNAs identificados en <i>Brassica rapa</i>
BolincRNAs	lincRNAs identificados en <i>Brassica oleracea</i>
SAM	Shoot Apical Meristem / Meristemo apical de la parte aérea

Resumen

Los lncRNAs son reguladores potenciales de la transcripción génica, tanto en *cis* como en *trans*, relacionándose con diversas funciones como el desarrollo de órganos y la respuesta a estímulos externos. Aunque las investigaciones sobre los lncRNAs han ido en aumento en vertebrados, han sido muy pocos los caracterizados en plantas. No obstante, los estudios disponibles sobre los lncRNAs en plantas refuerzan similitudes funcionales previamente reportadas en animales, como es la modulación en la topología de la cromatina, interacción con miRNAs, modulación del splicing y algunos pocos como precursores de RNA pequeños.

Dentro de este trabajo abordamos las posibles funciones biológicas de los lncRNAs en *A. thaliana*, a partir de más de 200 transcriptomas públicos y algunos transcriptomas generados en esta especie. Encontrando cerca del 21 % presente en módulos funcionales relacionándose con funciones biológicas como fotosíntesis, desarrollo de raíces, división celular, síntesis de metabolitos, entre otros. Por su parte, el 62 % de los lncRNAs presentaba patrones de expresión órgano específico, con prevalencia en expresión en órganos relacionados con reproducción y fotosíntesis. Convirtiendo esto en evidencia importante que soporta que los lncRNAs cuentan con funciones como reguladores de genes implicados en el desarrollo vegetativo de plantas.

Asimismo, analizamos la conservación de los lincRNAs identificados en *A. thaliana*, *B. rapa*, *B. oleracea*, *C. rubella* y *T. parvula* provenientes de la parte aérea y raíces en plantas, con el fin de localizar lincRNAs ortólogos funcionales con expresión órgano específico en la familia Brassicaceae. Identificamos que más del 48 % de los lincRNAs de *A. thaliana* están conservados por sintenia y microhomología a través de las especies. Además, logramos identificar múltiples copias de cada uno de los lincRNAs identificados, relacionando la adquisición de nuevos genes con los eventos de duplicación del genoma. La identificación de estos lincRNAs conservados sugiere una relevancia funcional para la adaptación y domesticación de estas especies, así como posibles aplicaciones futuras en las especies de interés agrícola.

Abstract

lncRNAs are potential regulators of gene expression, they can regulate other genes in *cis* and *trans*, and they are related to diverse functions such as organ development and response to external stimuli. Although research on lncRNAs has been increasing in vertebrates, very few have been characterized in plants. Nevertheless, the available studies on lncRNAs in plants reinforce functional similarities previously reported in animals, such as modulation of chromatin topology, interaction with miRNAs, modulation of splicing and a few as precursors of small RNAs.

Within this work we addressed the possible biological functions of lncRNAs in *A. thaliana*, from more than 200 public transcriptomes and some transcriptomes we generated in this species. We found about 21 % present in functional modules related to biological functions such as photosynthesis, root development, cell division, metabolite synthesis, among others. In turn, 62 % of the lncRNAs presented organ-specific expression patterns, with prevalence of expression in organs related to reproduction and photosynthesis. This provides important evidence that lncRNAs have functions as regulators of genes involved in plant vegetative development.

In addition, we analyzed the conservation of lincRNAs identified in *A. thaliana*, *B. rapa*, *B. oleracea*, *C. rubella* and *T. parvula* from the shoot and roots in plants, in order to locate functional orthologous lincRNAs with specific organ expression in the Brassicaceae family. We identified more than 48 % of the lincRNAs conserved by syteny and microhomology across species. In turn, we were able to identify multiple copies of each of the lincRNAs identified, relating the acquisition of new genes to genome duplication events. However, the conservation of these lincRNAs may give us an indication of their functional importance for the adaptation of these species, as well as the beneficial contributions in species of agricultural interest and their domestication.

Capítulo 1

Introducción

Descubrimiento de los lncRNAs

En 1984, se descubrió el primer RNA largo no codificante (lncRNA, por sus siglas en inglés), el lncRNA *H19*, con un transcrito de 2.5 Kb y con alta expresión durante el desarrollo embrionario en ratones (Pachnis et al., 1984). Tanto *H19* como su gen codificante vecino *Igf2* (insulin-like growth factor 2) son genes improntados, *H19* se expresa maternalmente, mientras que *Igf2* se expresa en el alelo paterno (Bartolomei et al., 1991; Nordin et al., 2014). Lo que desató el interés en el estudio de este lncRNA fue la ausencia de traducción, aunque el gen contenía pequeños marcos de lectura abiertos (ORF, por sus siglas en inglés), presentando motivos altamente conservados tanto en mamíferos como en aves (Brannan et al., 1990). Además, presenta características similares a los mRNAs (Brannan et al., 1990). La delección de este lncRNA en ratones genera letalidad en etapas prenatales, implicando a este lncRNA con el desarrollo embrionario (Leighton et al., 1995). A partir de este descubrimiento, se han identificado numerosos lncRNAs.

Características de los lncRNAs en eucariontes

Los lncRNAs son transcritos con una longitud mayor a los 200 nucleótidos (nt) (Kapranov et al., 2007), pueden contar con características similares a los mRNAs como son 7-metilguanosina (m7G o cap) en su extremo 5' y cola poliadenilada (poli A) en su extremo 3' terminal (Di et al., 2014; Guttman et al., 2009; J. Liu et al., 2012) ([Fig. 1](#)). Estas características se relacionan con el tipo de RNA polimerasa que realiza su transcripción. La RNA Polimerasa I (RNA Pol I) transcribe genes ribosomales (rRNA), la RNA Pol II transcribe RNAs mensajeros (mRNAs), microRNAs (miRNAs), RNAs pequeños nucleares (snRNA) y nucleolares (snoRNAs), mientras que la RNA Pol III efectúa la transcripción de RNAs de transferencia (tRNAs), U6 spliceosomal RNA (U6 snRNA), 5S rRNAs, y otros RNA involucrados con el procesamiento del RNA (Paule, 2000). La mayoría de los lncRNAs identificados en mamíferos son transcritos por la RNA Polimerasa II (RNA Pol II) (Guttman et al., 2009; Ponting et al., 2009), un ejemplo es el lncRNA *H19* (Brannan et al., 1990). De igual manera, en plantas se ha reportado la transcripción de los lncRNAs por la RNA pol II en *Arabidopsis thaliana* (Di et al., 2014) y en arroz (J. Yuan et al., 2018). Además, se han reportado dos polimerasas adicionales en plantas, la RNA Polimerasa IV (RNA Pol IV) y RNA Polimerasa V (RNA Pol V), las cuales participan en la ruta de biogénesis de los siRNAs (24 nt) y la metilación *de novo* en el DNA, mecanismo conocido como metilación del DNA mediada por RNA (RdDM), implicada en el silenciamiento de elementos repetidos y transposones (Matzke & Mosher, 2014). Adicionalmente, se ha encontrado que la RNA Pol V llega a transcribir algunos de los genes de lncRNAs (Ariel et al., 2015; Di et al., 2014).

Los lncRNAs, generalmente, presentan una menor cantidad de exones, un menor tamaño de su transcrito y niveles de expresión más bajos que los genes que codifican proteínas en mamíferos (Cabili et al., 2011; Hezroni et al., 2015; Necsulea et al., 2014), insectos (M.-J. M. Chen et al., 2016; F. Liu et al., 2019) y en plantas (Deng et al., 2018). Centrándonos en el número de exones, encontramos que existen especies con interés comercial que contienen

una mayor proporción de lncRNAs monoexónicos como son algodón (*Gossypium arboreum*) (J. Zhang et al., 2022; Zou et al., 2016), maíz (Li et al., 2014), tomate (Y. Zhou et al., 2019), durazno (*Prunus persica*) (H. Zhou et al., 2022), arroz (X. Huang et al., 2021; Y.-C. Zhang et al., 2014) y naranja (*Poncirus trifoliata*) (C.-Y. Wang et al., 2017). Sin embargo, plantas como el mijo (*Setaria italica*) y *Cleistogenes songorica* presentan lncRNAs con transcritos multiexónicos (T. Wang et al., 2020). En el caso del organismo modelo en plantas *Arabidopsis thaliana*, la mayoría de los lncRNAs identificados son monoexónicos, sus transcritos son más pequeños y su expresión es más baja en términos globales en comparación con los genes codificantes de proteínas (Di et al., 2014; Palos et al., 2022; J. Wang et al., 2021). Este mismo comportamiento se ha identificado en *Brassica rapa*, otro miembro de la familia Brassicaceae (A. Wang et al., 2019). Sin embargo, a diferencia de estas dos especies antes mencionadas en *Brassica oleracea*, la especie más cercana filogenéticamente a *B. rapa* (~ 2.84 mya) (Tiempo estimado de acuerdo a <http://www.timetree.org/>), la mayoría de los transcritos de lncRNAs son multiexónicos (X. Zhu et al., 2019). De igual manera, en *Brassica napus*, la especie generada por la cruce entre *B. rapa* y *B. oleracea*, presenta una mayor proporción de lncRNAs multiexónicos (Shen et al., 2018).

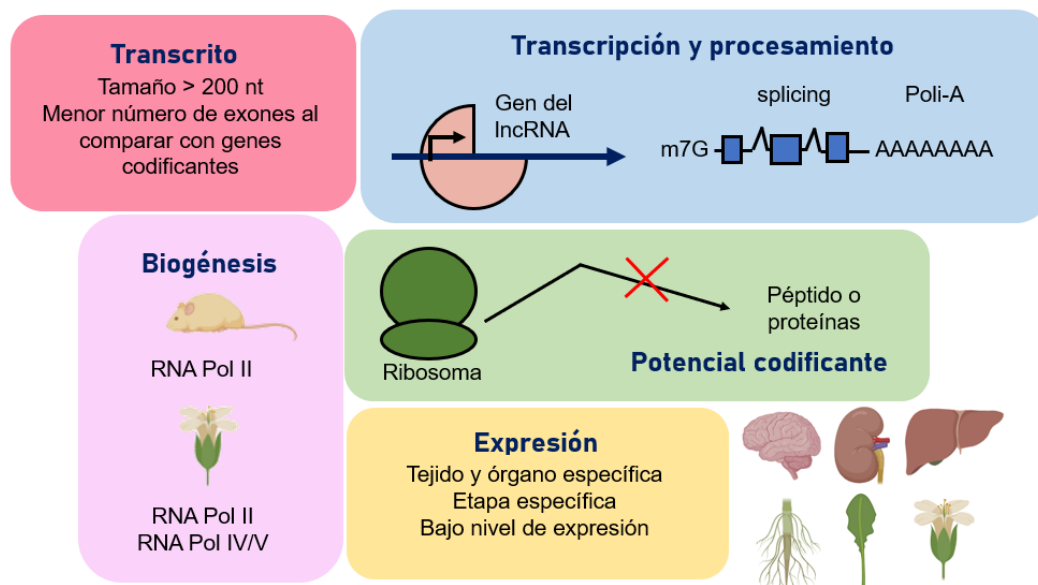


Figura 1. Características generales de los lncRNAs. Figura inspirada en (M. Sun & Kraus, 2015).

Clasificación de los lncRNAs

Los lncRNAs se pueden clasificar con base en su localización genómica respecto a los genes codificantes de proteínas. En este trabajo se emplearon las siguientes 4 categorías para la clasificación de los lncRNAs (Fig. 2), considerando las categorías previamente reportadas por la base de datos GENCODE para la clasificación de lncRNAs en humanos (Derrien et al., 2012; Mattick & Rinn, 2015).

- lncRNAs intergénicos (lincRNAs)**, todos aquellos lncRNAs provenientes de regiones intergénicas.
- lncRNA natural antisentido (NAT)**, todo aquel lncRNA que se superponen total o parcialmente al exón de otro gen en la cadena complementaria.

- c) **lncRNA exónico sentido**, todo aquel lncRNA que se superponen total o parcialmente al exón de otro gen en la misma cadena.
- d) **lncRNA intrónico**, todo lncRNA que se encuentre en el intrón de un gen codificante de proteína y no tenga superposición con ninguno de sus exones, incluyendo los que están en la misma cadena o complementaria al gen superpuesto.

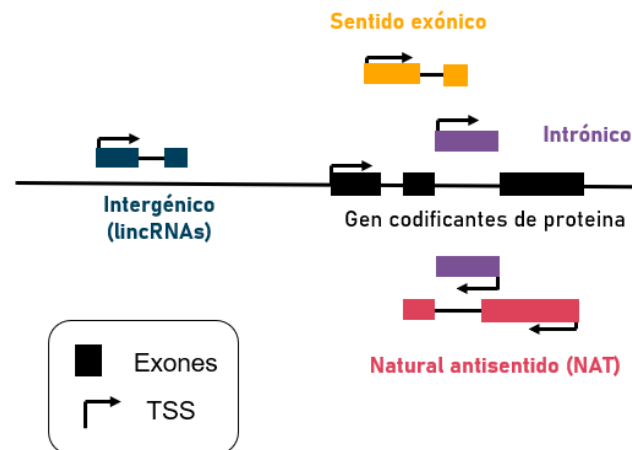


Figura 2. Clasificación de los lncRNAs de acuerdo a su localización genómica. Figura inspirada en (Derrien et al., 2012; Mattick & Rinn, 2015)

Mecanismos funcionales de los lncRNAs

Los lncRNAs pueden interactuar con DNA, otros RNAs y proteínas, a través del apareamiento de bases o a través de la unión a dominios estructurales de unión a RNA (RBP) (Ariel et al., 2020; Franco-Zorrilla et al., 2007b; Gaiti et al., 2015; Mercer & Mattick, 2013; Y. Wang et al., 2014a; Yoon et al., 2012). Se ha estimado que aproximadamente el 10 % del proteoma en los eucariontes cuentan con estos dominios de unión a RNA (RNA binding proteins, RBP) (Buenrostro et al., 2014). En animales, los lncRNAs pueden regular la transcripción de genes por parte de la unión con la Pol II (ejemplo *B2 RNA*) (T. A. Allen et al., 2004). De igual manera en plantas, se ha reportado la interacción de lncRNAs con la Pol II (lncRNA *ELENA1*) (Seo et al., 2017a), DNA (*MARS*) (Roulé, Christ, et al., 2022; Y. Wang et al., 2014a), moduladores de cromatina (*APOLLO*, *COLDIAIR*, *ENOD40*) (Campalans et al., 2004; Kim et al., 2017; Y. Wang et al., 2014a), factores de transcripción (Moison et al., 2021) y miRNAs (*IPS1*, *At4-3*, *At4-2*) (Franco-Zorrilla et al., 2007b; Shin et al., 2006a) (Fig. 3).

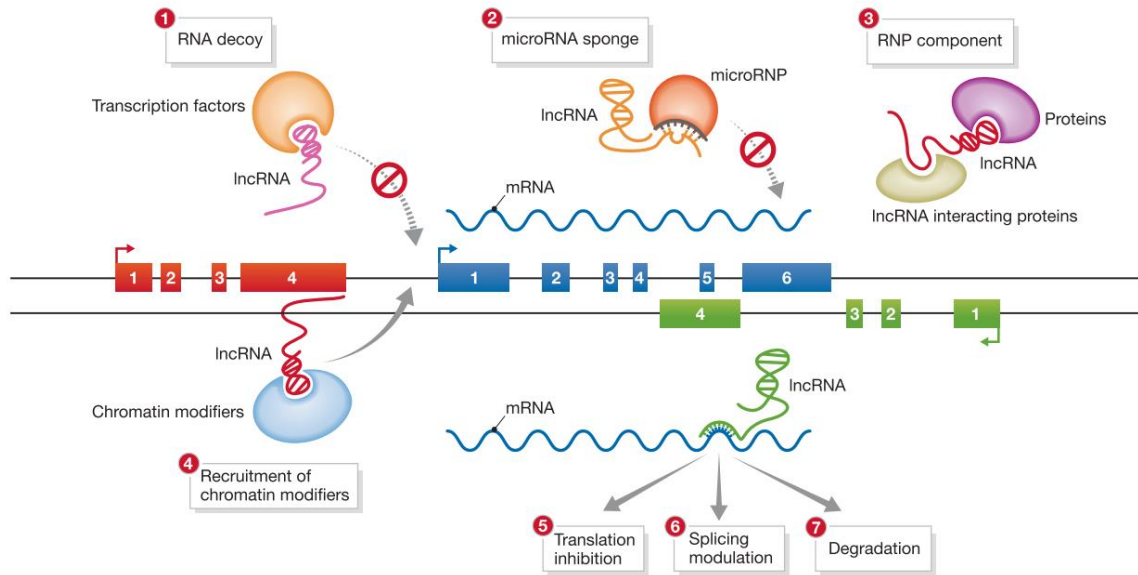


Figura 3. Mecanismos funcionales de los lncRNAs en la regulación génica. Figura proveniente de (Hu et al., 2012).

Algunos de los mecanismos funcionales de los lncRNAs descritos a la fecha son: **1) señuelo de RNA (RNA decoy)**, donde varios lncRNAs contienen secuencias de reconocimiento para la unión de los TF y de este modo alejarlos de los sitios diana en el DNA (Kino et al., 2010; Ng et al., 2012); **2) esponjas de miRNAs (microRNA sponge/ Target mimicry)**, actúan como señuelos similares a los sitios diana de los miRNAs evitando la interacción con el gen blanco (Cesana et al., 2011; Karreth et al., 2011; Salmena et al., 2011); **3) componente del complejo de ribonucleoproteínas (RNP) (RNP component)**, se unen en combinaciones específicas a proteínas reguladoras actuando como proteínas de andamiaje de los complejos de RNP (Ng et al., 2012; Tsai et al., 2010); **4) reclutamiento de complejos modificadores de la cromatina**, guían a los modificadores de la cromatina a las dianas en el DNA, ya sea en *cis* (Bertani et al., 2011; K. C. Wang et al., 2011) o *trans* (Rinn et al., 2007); además, los lncRNAs también pueden controlar la expresión génica a nivel transcripcional y postranscripcional, mediante **5) la inhibición de la traducción** (Yoon et al., 2012) y **6) la modulación del splicing** (Tripathi et al., 2010), respectivamente. Adicionalmente, **7) pueden unirse al mRNA para generar su degradación** (Gong & Maquat, 2011) (Fig. 3). Adicionalmente, en plantas se han reportado algunos lncRNAs como precursores de miRNAs (J. Liu et al., 2012), siRNAs y tasiRNAs (E. Allen et al., 2005; Yu et al., 2013).

Algunos lncRNAs pueden influir en la expresión de los genes codificantes vecinos, ya sea en el gen aledaño o regulando la expresión de los genes en el mismo cromosoma, determinando a este mecanismo como regulación en *cis* (Engreitz et al., 2016; Kung et al., 2013; Ulitsky & Bartel, 2013). Lo cual también incluye los mecanismos de regulación de lncRNAs en sus genes superpuestos (por ejemplo, XIST, roX2, COOLDAIR, COLDAIR, HID1) (Csorba et al., 2014; Heo & Sung, 2011; Rosa et al., 2016; Y. Wang et al., 2014a; W. Yuan et al., 2016). Mientras que la regulación en *trans*, se relaciona con los cambios en la expresión génica de los genes localizados en otros cromosomas (Kung et al., 2013; Ulitsky & Bartel, 2013).

Tipos de conservación empleados en lncRNAs

Mediante el empleo de predicciones a partir de datos provenientes de órganos y etapas de desarrollo específicas obtenidas por RNA-Seq, se han logrado identificar lncRNAs ortólogos evolutivamente conservados a través de las especies en mamíferos (Darbellay & Necsulea, 2020; Hezroni et al., 2015; Necsulea et al., 2014). El término ortología se refiere al gen o conjunto de genes que han evolucionado a partir de un único gen ancestral en las especies ancestrales de referencia (Fitch, 2000; Gabaldón & Koonin, 2013). A partir del empleo de alineamientos genómicos se ha logrado demostrar que los exones de los lncRNAs evolucionan más rápido que en los genes codificantes de proteínas tanto en mamíferos (Cabili et al., 2011; Chodroff et al., 2010; Darbellay & Necsulea, 2020; Guttman et al., 2009; Hezroni et al., 2015; Necsulea et al., 2014; Ulitsky et al., 2011) como en plantas (Corona-Gomez, n.d.-b; Di et al., 2014; Rymarquis et al., 2008; J. Yuan et al., 2016; Y.-C. Zhang et al., 2014).

Existen 3 tipos de conservación más empleados para la búsqueda de lncRNAs a través de las especies: **A) Conservación por sintenia**, **B) Conservación por microhomología o motivos** y **C) Conservación por estructura** (Ramírez-Colmenero et al., 2020; Ross & Ulitsky, 2022; Ulitsky, 2016) (Figura 4). En el caso de la conservación por sintenia, los lncRNAs pueden encontrarse conservados bajo este parámetro y expresarse desde el *loci* sinténico a pesar de haber perdido la mayor parte o toda la conservación de la secuencia (Ramírez-Colmenero et al., 2020). Los métodos de conservación no son mutuamente excluyentes por lo que podemos encontrar lncRNA con más de un tipo de conservación.

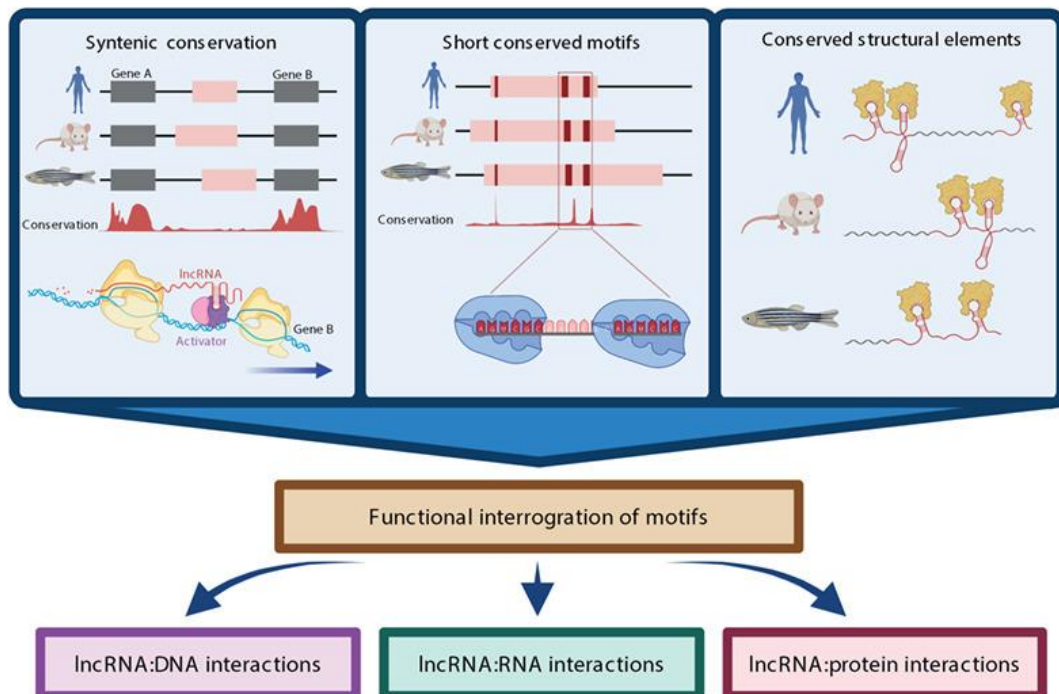


Figura 4. Tipos de conservación empleados para la búsqueda de lncRNAs ortólogos. La **conservación por sintenia** se refiere a la búsqueda del lncRNA a través de la localización de un vecindario genómico conservado por genes codificantes ortólogos en las especies. La **conservación por secuencia** (microhomología o motivos) puede ser total o parcial al lncRNA, logrando identificar pequeños motivos. Mientras que la **conservación por**

estructura se relaciona con el plegamiento secundario del lncRNA en distintas especies (Ross & Ulitsky, 2022).

Algunos lncRNAs presentan una alta conservación en la secuencia de sus promotores mientras que su región codificante se encuentra pobremente conservada, tanto a nivel genómico, como a nivel de transcrito, a comparación de los genes que codifican a proteínas tanto en mamíferos (Darbellay & Necsulea, 2020; Hezroni et al., 2015; Necsulea et al., 2014) como en plantas (Di et al., 2014; Wunderlich et al., 2014; J. Yuan et al., 2016). Al igual que en los genes codificantes, se han identificado motivos conservados y relacionados con la respuesta al estrés por frío, luz roja lejana, salinidad (11-12 nt), a fosfato (8 nt) y calor (heat shock, 10 nt) en sus promotores (Di et al., 2014; Wunderlich et al., 2014; J. Yuan et al., 2016), relacionando a los lncRNA con un órgano o condición específica (Franco-Zorrilla et al., 2007b; Y. Wang et al., 2014a).

Interés en el estudio en la familia Brassicaceae

La familia de Brassicaceae surgió aproximadamente hace 54 mya, uno de los miembros más conocidos es *A. thaliana*, al ser la especie modelo en el estudio de la genética de plantas (Beilstein et al., 2008). Además, también cuenta con miembros de interés agrícola como son *Brassica rapa*, *Brassica oleracea* y *Brassica napus* (Bancroft et al., 2011; F. Cheng et al., 2016). Contando con algunas especies particulares como son *Thellungiella parvula* y *Thellungiella salsuginea* (*Thellungiella halophita*), capaces de crecer en suelos con altas concentraciones de sal, soporta heladas y suelos con pH ácido (Griffith et al., 2007; Oh et al., 2010; Q. Zhang et al., 2013). Presentan amplia distribución de estas especies por la mayoría de los continentes ([Tabla 1](#)).

Tabla 1. Características de las especies analizadas en este trabajo. Información obtenida de (*Plants of the World Online*, n.d.)

Especies	Origen	Distribución
<i>B. oleracea</i>	Sur de Europa	América del norte y sur, Europa, Asia y África
<i>B. rapa</i>	Sur de Europa y norte de África	América, Europa, Asia y África
<i>T. parvula</i>	Sur de Europa	Norte de Europa
<i>A. thaliana</i>	Europa-Asia y África	Norte y sur de América, Asia, Europa y África
<i>C. rubella</i>	Sur de Europa y norte de África	Sur de Europa

Debido a estos cambios en el ecosistema a lo largo del tiempo, algunas especies dentro de esta familia presentan múltiples copias en sus genomas, relacionado con los eventos de duplicación del genoma ocurridos hace 124 mya (β) y hace 47 mya (α); aun con el paso del tiempo se reporta uno de los eventos de duplicación más recientes hace 7 a 12 mya, afectando en el número de copias de los genes presentes en las especies *Camelina sativa*, *B. rapa* y *B. oleracea* (Jiao et al., 2011; Kagale, Robinson, et al., 2014; S. Liu et al., 2014;

Tang & Lyons, 2012). Una de las especies con el genoma más grande en esta familia es *C. sativa*, miembro del linaje 1, la cual presenta 3 subgenomas (Cs-G1, Cs-G2 y Cs-G3), con 6, 7 y 7 cromosomas cada uno ($2n=40$) (Kagale, Koh, et al., 2014). Estos eventos de duplicación en los genomas conllevan a la localización de múltiples copias (parálogos) de un mismo gen ancestral, reportando hasta 3 copias de cada gen anotado en *A. thaliana* con respecto a *C. sativa*, *B. rapa* y *B. oleracea* (Kagale, Koh, et al., 2014; Kagale, Robinson, et al., 2014; S. Liu et al., 2014); y con más del 70 % de conservación por sintenia en los genes codificantes entre estas especies, demostrando una estrecha cercanía entre las especies (Kagale, Koh, et al., 2014; Kagale, Robinson, et al., 2014; S. Liu et al., 2014). De igual manera, se han reportado inversiones dentro del genoma en esta familia (Koch & Kiefer, 2005; Murat et al., 2015).

Estos nuevos genes obtenidos de la duplicación del genoma pueden proporcionar nuevas funciones génicas para la adaptación a diversas condiciones de crecimiento (Gehring et al., 2006; Moser, 2012; Serdyuk et al., 2022). Siendo observado en las ventajas agronómicas para la producción de aceite en la especie *C. sativa*, en las que incluyen la madurez temprana, bajo requerimiento de agua y nutrientes, la adaptabilidad a condiciones ambientales adversas y la resistencia a diversos patógenos (Gehring et al., 2006; Moser, 2012; Serdyuk et al., 2022). Analizando la selección purificadora (relación $Ka/Ks < 1$) dentro de esta especie, se ha favorecido para la generación de genes con funciones conservadas, lo que indica que algunas de las secuencias conservadas y relacionadas con los genes codificantes fueron finalmente beneficiosas para la supervivencia y el crecimiento en esta especie (Song et al., 2020). Tal es el ejemplo del factor de transcripción WRKY, identificado un total de 173 genes *CsWRKY* que tienen 65 ortólogos correspondientes de *A. thaliana*, e incluso se encontró que múltiples genes *CsWRKY* corresponden a un único ortólogo *AtWRKY* (Song et al., 2020).

Cabe destacar que estas duplicaciones no solo afectan a los genes codificantes, encontrando en el caso de lncRNAs múltiples copias a lo largo de los genomas. El primer lncRNA identificado en plantas fue *ENOD40* (early nodulin 40) en soya (*Glycine max*) (*GmENOD40*), el cual se asocia con la diferenciación y proliferación celular durante el proceso de nodulación (Charon et al., 1997; Crespi et al., 1994; Kouchi & Hata, 1993; W. C. Yang et al., 1993), se han reportado dos copias presentes en *Lotus japonicus*, siendo una de las especies más cercanas filogenéticamente (Flemetakis et al., 2000). Se encuentra presente dentro de las leguminosas, relacionándose con la nodulación (Ganguly et al., 2021) y con especies carentes de este rasgo, como son el tabaco (*Nicotiana tabacum*) (Sande et al., 1996), arroz (*Oryza brachyantha* y *Oryza sativa*) (Kouchi et al., 1999) y maíz (*Zea mays*) (*ZmENOD40*) (Compaan et al., 2003), lo cual sugiere que este lncRNA se encuentra conservado en el reino vegetal. Al igual que en leguminosas, este lncRNAs se ha reportado como un lncRNA monoexónico en arroz (Kouchi et al., 1999), lo cual difiere con los parálogos identificados en *Arachis hypogaea*, presentando dos exones en cada transcrito (Ganguly et al., 2021). Estos dos lncRNAs están relacionados con eventos de duplicación del genoma en *Arachis hypogaea* (Ganguly et al., 2021), siendo previamente reportado en *Lotus* (Flemetakis et al., 2000). El lncRNA *ENOD40* presenta conservación por sintenia y microhomología a través de las especies como son *Lotus japonicus*, *A. hypogaea*, *Glycine max*, *Medicago truncatula*, *Phaseolus vulgaris*, y *Aeschynomene evenia* (Ganguly et al., 2021). A pesar de la baja conservación en su secuencia a través de las especies (Flemetakis et al., 2000; Kouchi et al., 1999), se ha logrado identificar este lncRNA en *A. thaliana* ampliando la búsqueda de ORF menores a 100 aminoácidos (aa) en datos de perfiles ribosómicos asignando como sORF1

(AT1G10682) (Hsu et al., 2016). A pesar de contar con un ORF menor a 100 aa, como un valor sugerente de poco o nulo potencial de codificar a una proteína (Dinger et al., 2008); los dos péptidos generados de la traducción de este lncRNA, se unen a la subunidad de la sacarosa sintasa en los nódulos (nuevo órgano fijador de nitrógeno) de las raíces, lo cual relaciona a estos péptidos con el control de sacarosa durante la fijación de nitrógeno en soya (tamaño de los péptidos 12 aa y 24 aa) (GmENOD40) (Rohrig et al., 2002) y *Medicago sativa* (12-13 aa) (MtENOD40) (Charon et al., 1997). Algunos de los lncRNAs que cuentan con un ORF pequeño solo en una minoría de los eventos de traducción llegan a dar lugar a péptidos estables y funcionales (Andrews & Rothnagel, 2014; Choi et al., 2019; Housman & Ulitsky, 2016; Ji et al., 2015). Por ejemplo, el lncRNA *LINC00961*, que se ha identificado como represor de mTORC1, el cual es el regulador central del metabolismo celular, además, este lncRNA codifica a un péptido pequeño de 90 aa en humano y 75 aa en ratones conocido como SPAR (Matsumoto et al., 2017). Demostrando que la pérdida de este polipéptido promovía la regeneración muscular después de la lesión (Matsumoto et al., 2017). Por lo tanto, se especula que tanto los mRNAs como los lncRNAs pueden desempeñar funciones duales como codificación de proteínas o péptidos, así como RNAs no codificantes (Ariel et al., 2015; Housman & Ulitsky, 2016; Rymarquis et al., 2008; Ulitsky & Bartel, 2013). Resulta importante afirmar que no se ha logrado identificar al lncRNA *ENOD40* por sintenia en la familia Brassicaceae (Ganguly et al., 2021). Por lo que resulta de interés la localización de uno de los lncRNAs más conservados y funcionales en el reino vegetal.

Otro lncRNA funcionalmente conservado es el lincRNA *IPS1* identificado inicialmente en tomate (*Lycopersicon esculentum* L.) (lncRNA *TOMATO PHOSPHATE STARVATION-INDUCED 1, TPSI1*) (C. Liu et al., 1997), perteneciente a una familia de genes que contienen la secuencia complementaria (CR, complementary region) al miR-399, con un tamaño de apenas 23 nt (Rymarquis et al., 2008). Denotando a todos los lincRNA que presentarán esta CR dentro de la familia TPSI (Burleigh & Harrison, 1999; Franco-Zorrilla et al., 2007b; Shin et al., 2006a). En *A. thaliana*, se han reportado 4 genes con este motivo presente en la secuencia, el lincRNA *IPS1* y los 3 parálogos anotados como *At4* (*At4-1*, *At4-2* y *At4-3*), presentando redundancia funcional entre estos genes (Franco-Zorrilla et al., 2007b; Shin et al., 2006a). Función relacionada con la homeostasis del fosfato en la planta a través del secuestro del miR99, evitando la degradación de los transcritos del factor de transcripción PHO2 (Franco-Zorrilla et al., 2007b; Shin et al., 2006a). Sin embargo, la búsqueda de este motivo no se encuentra relacionado directamente con la localización de lncRNAs ortólogos en otras especies, tal es el caso del lncRNA *PILNCR1* en maíz, el cual presenta la CR al miR-399, pero el vecindario genómico difiere al reportado en *IPS1* de tomate, medicago, arroz y *A. thaliana* (Du et al., 2018). Pudiendo relacionar a este lncRNA dentro de la misma familia de genes, pero sin ser clasificado como un ortólogo al lincRNA *IPS1*.

Antecedentes

Identificación de lncRNAs en *A. thaliana*

En *A. thaliana*, uno de los trabajos más importantes para la anotación de lncRNA es el trabajo de (J. Liu et al., 2012), en el cual emplearon múltiples microarreglos y cuatro transcriptomas en condiciones wild type (o sin estrés) de flor, hojas y raíces. Permitiendo la identificación de 6500 lncRNAs (J. Liu et al., 2012), los cuales posteriormente fueron incorporados en la

anotación de Araport11 (C.-Y. Cheng et al., 2017), muchos de los cuales se conservan dentro de esta anotación y continúan siendo analizados hasta el día de hoy (*lncRNA FLINC*, *LINC-AP2*, *TAR-224*, *lncCOBRA1*, *lncCOBRA3*, *lncCOBRA5*, *SVALK*, *DRIR*, *MARS*) (Gao et al., 2016; Kindgren et al., 2018; Kramer et al., 2022; Qin et al., 2017a; Roulé, Christ, et al., 2022; Severing et al., 2018; Yi et al., 2015).

Posterior a este trabajo se han generado múltiples bases de datos que usan herramientas bioinformáticas distintas para la identificación de lncRNAs (J. Jin et al., 2013; Paytuví Gallart et al., 2016; Szcześniak et al., 2019; Y. Zhao et al., 2016). En estos trabajos podemos destacar la base de datos CANTATAdb v2.0 por su número de lncRNAs identificados (Szcześniak et al., 2016, 2019), para los cuales no solo se tomó en cuenta una anotación predictiva, sino que se basó en 10 transcriptomas, utilizando una metodología robusta de anotación que empleaba la identificación de lncRNAs por Coding Potential Calculator (CPC) v0.9.r2 (Kong et al., 2007). Otra base de datos importante es GreeNC, la cual también emplea una anotación predictiva mediante CPC para la identificación de lncRNAs en diferentes especies a partir de los transcritos provenientes de Phytozome (Goodstein et al., 2012) y ENSEMBL (Bolser et al., 2016). Además, clasifica a algunos lncRNAs como precursores de miRNAs (Paytuví Gallart et al., 2016). Otra base de datos que resalta por ser la más usada y que sirve como referencia es Araport11 (C.-Y. Cheng et al., 2017)). La anotación de lncRNAs en Araport11 está compuesta por la anotación de (J. Liu et al., 2012) más las anotaciones de lncRNAs en TAIR10, además, no es exclusiva de lncRNAs, contiene una anotación extensa de todo tipo de genes: codificantes y RNAs reguladores. En total las bases de datos cuentan con la siguiente cantidad de genes anotados como lncRNAs para *A. thaliana*: 4,080 genes en CANTATAdb v2.0 (Szcześniak et al., 2019), 2,752 genes en GreeNC (Paytuví Gallart et al., 2016), y 3,559 genes en Araport11 (C.-Y. Cheng et al., 2017). Pero consideramos que hay espacio para mejorar el número de anotaciones en cuestión de lncRNAs, usando un análisis más exhaustivo. Además de proporcionar una clasificación más clara de los lncRNAs de acuerdo con su contexto genómico y su posible función dada su co-expresión con genes codificantes.

lncRNAs funcionales en *A. thaliana*

Existen aproximadamente 160 lncRNAs anotados como funcionales dentro de la base de datos EVlncRNAs V2.0 en *A. thaliana* (B. Zhou et al., 2021), los cuales han sido confirmados por qRT-PCR, mutantes (overexpression, knockdown), Southern blot, Northern blot, ChIP, microarreglos e hibridación *in situ* (B. Zhou et al., 2021). Hasta la fecha se han logrado caracterizar 19 lncRNAs adicionales a los reportados en esta base de datos, relacionándolos con procesos biológicos como la germinación y desarrollo (*lncCOBRA1*, *lncCOBRA3*, *lncCOBRA5*, *MARS*) (Kramer et al., 2022; Roulé, Christ, et al., 2022), aclimatación de frío (*SVALK*, *CIL1*) (Kindgren et al., 2018; G. Liu et al., 2022), sequía, salinidad (*DRIR*) (Qin et al., 2017a), dormancia (*asDOG1*) (Fedak et al., 2016), floración (*MAS*, *FLAIL*) (Y. Jin et al., 2021; X. Zhao et al., 2018) y estímulos por luz azul (*BLIL1*) (Z. Sun et al., 2020a). Además, en algunos casos se identificó su mecanismo funcional, como precursores de miRNAs (*pri-miR165a*, *pri-miR858a*, *pri-miR858a*, *pri-miR160c*) (Ben Amor et al., 2009; Hirsch et al., 2006), precursores de siRNAs (*TAS1a*, *TAS2*, *TAS3*, *TAS4*) (Ladomery & Pisignano, 2021), esponjas de miRNAs (*BLIL1*), remodeladores de la cromatina (*MARS*, *FLAIL*) (Y. Jin et al., 2021; Roulé,

Christ, et al., 2022) e interacciones con proteínas (*IncCOBRA1*) (Kramer et al., 2022). Con un total de 179 lncRNAs caracterizados como funcionales en *A. thaliana*.

Problemas en la anotación de lncRNAs en diversas bases de datos

Dentro de los 179 lncRNAs caracterizados como funcionales en *A. thaliana*, algunos de estos se encuentran clasificados como "genes codificantes" en Araport11. Dentro de los cuales encontramos al lncRNA *IPS1* (*Induced by Phosphate Starvation 1*, AT3G09922) y a su parólogo *At4* (AT5G03545) (Shin et al., 2006a), ambos lncRNA cuentan con la función de esponja del microRNA miR399 en ausencia de fosfato (Franco-Zorrilla et al., 2007b; Shin et al., 2006a), presentan conservación y funcionalidad en diversas especies de plantas (Ajmera et al., 2018; Burleigh & Harrison, 1999; Calixto et al., 2019a; C. Y. Huang et al., 2011). Específicamente en el caso de *IPS1*, existe una enzima con el mismo nombre (AtIPS1, AT4G39800, 1-L-myo-inositol-1 P synthase; E.C.5.5.1.4), siendo asignados en NCBI (Stoesser, 2004) con el mismo ID, a pesar de presentar diferente localización genómica y siendo diferentes genes (*IPS1*: AT3G09922, *AtIPS1*: AT4G39800). A causa de este error, las bases de datos como Araport11 (C.-Y. Cheng et al., 2017) y Plaza v4.0 (Van Bel et al., 2017), presentan información funcional (términos GO) con respecto a la enzima. Siendo un problema bastante grave para el análisis y clasificación de lncRNAs.

Así mismo, las bases de datos de lncRNAs en plantas GreeNC (Di Marsico et al., 2022), PlncRNADB (Bai et al., n.d.) y PlncDB v2.0 (J. Jin et al., 2020) contienen proteínas transmembranales, enzimas (ATP sintasas, peroxidasas), proteínas ribosomales, y en algunos casos muy poco frecuentes, factores de transcripción (Tabla 2). Siendo indispensable realizar una nueva anotación de los lncRNAs en *A. thaliana*, dejando de lado los programas de anotación predictiva (ejemplo CPC) y empleando un método más eficiente en la exclusión de genes codificantes, para lo cual se seleccionó el Método Estricto (ME) con base en trabajos previos para la identificación de lncRNAs en humano y esponja marina (Cabili et al., 2011; Gaiti et al., 2015).

Tabla 2. Ejemplos de genes codificantes asignados erróneamente como lncRNAs.

Gene ID	Anotación en Araport11	Otras bases de datos	link con las bases de datos
AT1G29580	Subunidad de la RNA pol II	GreeNC, PlncDB	Araport11 GreeNC
AT1G36640	Proteína transmembranal	GreeNC, PlncDB	Araport11 GreeNC
AT1G05575	Proteína transmembranal	GreeNC PlncDB PlncRNADB	Araport11 GreeNC
AT1G46336	Proteína transmembranal	GreeNC PlncDB	Araport11 GreeNC
AT1G59171	Inositol-pentakisphosphate 2-kinase family protein	GreeNC PlncDB	Araport11 GreeNC
AT1G61575	Serine/Threonine kinase	GreeNC PlncB	Araport11 GreeNC
AT2G27535	ribosomal protein L10A family protein	GreeNC PlncB PlncRNADB	Araport11 GreeNC

AT3G56390	WRKY transcription factor-like protein	GreeNC PlncB	Araport11 GreeNC
AT3G01130	ATP synthase E chain	GreeNC PlncB PlncRNADB	Araport11 GreeNC
AT5G14110	peroxidase (DUF 3339)	GreeNC PlncB	Araport11 GreeNC

Consideraciones en la conservación de lncRNAs

Se han realizado algunos intentos por encontrar lncRNAs conservados a través de las especies en plantas, sin embargo, uno de los principales filtros en su análisis ha sido la conservación por secuencia (L. Chen & Zhu, 2022; Mohammadin et al., 2015; Nelson et al., 2016; Palos et al., 2022; Sang et al., 2021). Aun en un corto tiempo de divergencia entre las especies, los lncRNAs muestran una baja conservación a nivel de su secuencia dentro de la misma familia (Nelson et al., 2016; Palos et al., 2022). Dentro de esta familia se ha empleado el concepto de sintenia para la localización de 6,500 lncRNAs de *A. thaliana* en 10 especies (J. Liu et al., 2012). Sin embargo, muchos de estos han sido descartados al seleccionar solo aquellos con una mayor conservación por secuencia al compararlos con los lncRNAs de *A. thaliana* (Nelson et al., 2016). Determinando que los lncRNAs presentan una menor conservación a nivel de secuencia a comparación con los genes codificantes conforme la escala evolutiva aumenta (Nelson et al., 2016).

A pesar de emplear múltiples comparaciones o ampliar la cantidad de especies analizadas, si se considera solo a nivel de secuencia, se llegan a perder más del 80 % de los lncRNAs a través de las especies (L. Chen & Zhu, 2022; Mohammadin et al., 2015; Nelson et al., 2016; Sang et al., 2021), debido a que algunos lncRNAs pueden contener pequeños motivos conservados en su secuencia, siendo difíciles de alinear o por no alcanzar valores de significancia estadística (Quinn et al., 2016; Ulitsky & Bartel, 2013). Aún utilizando la conservación por sintenia como segundo método de análisis, las implicaciones resultantes del primer filtro pueden afectar la detección de los lncRNA (L. Chen & Zhu, 2022; Palos et al., 2022).

Por su parte, emplear criterios muy estrictos en relación al rango de genes codificantes para definir las regiones sinténicas conservadas afecta al número de regiones identificadas (Vakirlis et al., 2020). Una explicación alternativa es que, a medida que el criterio de sintenia se hace más estricto, los genes que cumplen este criterio podrían estar sometidos a restricciones evolutivas más fuertes y, por tanto, es menos probable que diverjan, lo cual depende de la distancia evolutiva y del grado más general de conservación de la sintenia entre las especies que se comparan (Vakirlis et al., 2020). Aunado a esto, dentro de la familia Brassicaceae se ha analizado la conservación por sintenia definiendo un rango de 10 genes flanqueantes al lncRNA, obteniendo el 17.30 % (3560 de 20,416 genes) lncRNAs conservados solo por sintenia y 18.1 % (3725 de 20,416 genes) lncRNAs conservados por secuencia y sintenia dentro de la misma familia (Palos et al., 2022). En otro de los estudios más recientes se reporta el 4.84 % (199 de 4106 genes) lncRNAs conservados por secuencia y sintenia (L. Chen & Zhu, 2022). Resultando importante considerar bajos niveles en la conservación de secuencia en regiones sinténicas conservadas a pesar del corto tiempo de divergencia entre estas especies (L. Chen & Zhu, 2022; Palos et

al., 2022). Llegando a considerar la conservación de estos lincRNAs dependiente de la estructura o de ciertos motivos localizados que le permitan llegar a ser importantes en la regulación génica, y por ende, contar con expresión en una condición definida (L. Chen & Zhu, 2022; Palos et al., 2022).

Considerando que el 93 % de los genes codificantes agrupados en ortogrupos en *B. rapa* se encuentran conservados con *A. thaliana* (X. Wang et al., 2011), también se ha encontrado que no solo la secuencia se encuentra conservada, si no también las posiciones genómicas dentro de la familia Brassicaceae (Corona-Gomez et al., 2020). Resultando importante partir de un alineamiento genómico para la ubicaciones de los bloques sinténicos compartidos entre genes codificantes y lincRNAs, para una correcta ubicación de los mismos.

Algunos lincRNAs conservados y funcionales

Para estudiar la evolución de las interacciones entre los lincRNAs y el genoma, se ha empleado el sistema de compensación por dosis en *Drosophila melanogaster*, regulado por dos lincRNAs (*roX1* y *roX2*). Estos dos lincRNAs duplican la expresión del cromosoma X en los machos para igualar la expresión génica de las hembras (Quinn et al., 2016). Presentan redundancia funcional atribuida a un pequeño motivo, roXbox, constituido por 8 nt localizados en el stem-loop de la estructura de *roX1* y *roX2* (Ilik et al., 2013; Park et al., 2007). Por lo que, la búsqueda de ortólogos de *roX* en *Drosophila* por medio de la conservación por secuencia había resultado fallida (Alekseyenko et al., 2013; Park et al., 2007). Mediante un análisis de conservación de los genes *roX* en 27 especies de moscas de la fruta se encontraron 47 nuevos genes ortólogos (19 pertenecen a *roX1* y 28 *roX2*), con un aproximado de 40 mya de divergencia evolutiva entre las especies, mediante la combinación de la conservación por secuencia, por sintenia y estructura secundaria del RNA (Quinn et al., 2016). Adicionalmente, se encontró que estos lincRNAs ortólogos conservan su función a pesar de la baja conservación a nivel de secuencia que existe en especies distantes, aunque la eficiencia del rescate disminuye al aumentar la distancia evolutiva (Quinn et al., 2016).

En plantas, se ha encontrado que el lincRNA *ENOD40* (early nodulin 40) presenta una baja conservación por secuencia dentro de las leguminosas (Flemetakis et al., 2000), abarcando un tiempo de 40 mya (Lavin et al., 2005), encontrando un porcentaje de conservación del 78 y 72 % de identidad entre el lincRNA identificado previamente en soya y las dos copias presentes en *Lotus japonicus* (Flemetakis et al., 2000). Se propone que la presencia de dos copias provenientes del mismo gen se relaciona con eventos de duplicación en el genoma de esta especie (Flemetakis et al., 2000). Por su parte, se ha demostrado que el lincRNA *ENOD40* se encuentra conservado por sintenia y microhomología a través de las especies (Ganguly et al., 2021). Aunque, en términos globales su secuencia pobremente conservada, se han localizado dos regiones con una homología mayor al 50 % en arroz (Kouchi et al., 1999), ambas regiones coinciden con motivos de unión a proteínas (RNA binding protein, RBP) (Campalans et al., 2004). Este lincRNA interactúa con las proteínas y permite la relocalización del núcleo al citoplasma (Campalans et al., 2004), función similar a la reportada con el lincRNA *ASCO* con la interacción con las proteínas nuclear speckle RBP, durante el desarrollo de las raíces (Bardou et al., 2014). De igual manera, el lincRNA *lncCOBRA1* cuenta con algunos motivos RBP muy particulares, al estar altamente conservados en humanos (Gosai et al., 2015) y a través de la familia Brassicaceae (Kramer et al., 2022). Las proteínas

que interactúan con *IncCOBRA1* están implicadas en una amplia gama de funciones biológicas, incluyendo la respuesta a la citoquinina y al ácido abscísico (ABA), la gluconeogénesis y la fotorrespiración (Ran et al., 2020). Modificaciones en este lincRNA lo relacionan con la germinación y el desarrollo de plantas (Kramer et al., 2022) ([Tabla 3](#)).

Dentro de la misma ubicación genómica del lincRNA *ENOD40*, se ha identificado un transcrito antisentido a este lincRNA, conocido como *NAT-AhDONE40 (DONE40)* en *Arachis hypogaea*, el cual interactúa con el dominio SET (Su(var)3-9, enhancer of Zeste and Trithorax) de la metiltransferasa de histona AhASHR3 (ABSENT SMALL HOMEOTIC DISC1 RELATED3), un miembro de la familia Trithorax, relacionando esta interacción con la deposición de marcas de activación (H3K36me3 y H3K4me3) en el locus de *ENOD40*, en ambas copias (cromosoma 5 y cromosoma 15, genes AhENOD40-1 y AhENOD40-2, respectivamente) del gen (Ganguly et al., 2021). Por su parte, se ha relacionado a esta proteína modificadores de histonas con el control de la división celular en el meristemo de la raíz y de las células quiescentes en *A. thaliana* (Kumpf et al., 2014), lo cual cobra sentido al encontrarse la inducción de la división celular una vez iniciado el proceso de nodulación en leguminosas (Charon et al., 1997; Crespi et al., 1994; Flemetakis et al., 2000; Kouchi & Hata, 1993; W. C. Yang et al., 1993). Demostrando que la función de los lincRNAs no necesariamente recae en la secuencia del transcrito (Darbellay & Necsulea, 2020).

Tabla 3. Algunos de los AtlincRNAs conservados y con funciones conservadas a través de las especies.

lncRNA	Gene ID	Mecanismo funcional	Función biológica	Referencia
<i>IPS1</i>	AT3G09922	Esponja de miRNA	Homeostasis de fosfato	(Franco-Zorrilla et al., 2007b)
<i>HID1</i>	AT2G35747	Unión al promotor del factor de transcripción PIF3	Respuesta a luz roja, elongación del hipocotilo	(Y. Wang et al., 2014a)
<i>ENOD40</i>	AT1G10682	Translocación de proteínas	Desarrollo de la raíz y nodulación	(Charon et al., 1997; Rohrig et al., 2002)
<i>IncCOBRA1</i>	AT3G05655	Unión a proteínas	Desarrollo y germinación	(Kramer et al., 2022)

Los lncRNAs presentan expresión órgano o etapa específica

En la mayoría de las especies de plantas y animales en dónde han sido estudiados, se ha encontrado que la mayoría de los lncRNAs cuentan con patrones de expresión espacio temporal (edad biológica) y órgano muy específicos, con grupos muy particulares de células que los expresan en cerebro, hígado, riñón y testículos (Darbellay & Necsulea, 2020; Necsulea et al., 2014). Por otra parte, en *A. thaliana* se han encontrado lncRNAs expresados específicamente en flores, hojas y raíces de 4 semanas de edad en respuesta a un estrés biótico (elf18) y abiótico (choque térmico y sequía) (J. Liu et al., 2012). De igual manera, en arroz se ha observado la expresión específica de los lncRNAs en anteras, pistilo, semillas y parte aérea (Y.-C. Zhang et al., 2014). Sin embargo muy pocos de estos se encuentran caracterizados, destacando al lincRNA *ENOD40* cuenta con patrones espacio temporal y espacial, al iniciar su transcripción a las pocas horas de la inoculación con las bacterias rhizobium, y continuar su expresión en las células corticales de la raíz (Charon et al., 1997;

Flemetakis et al., 2000). Adicionalmente, se ha reportado una alta expresión de este lncRNA durante el desarrollo embrionario (Flemetakis et al., 2000). Por lo que, los lncRNAs pueden participar en diferentes etapas del desarrollo, prendiendo y apagando dependiendo del contexto regulatorio que se necesite en el organismo.

En esta tesis, abordamos la exploración de la conservación de los lncRNAs en Brasicás usando un enfoque comparativo. Comenzamos realizando una nueva anotación de los lncRNAs en *A. thaliana* a partir de los transcriptomas públicos y generados en el laboratorio, posteriormente analizar sus posibles funciones biológicas mediante los módulos de co-expresión, encontrando lncRNAs compartiendo funciones biológicas con genes codificantes relacionados con fotosíntesis, desarrollo de raíces, división celular, síntesis de metabolitos, entre otros (Capítulo 2). Posteriormente, en el Capítulo 3, analizamos la conservación de los lncRNAs identificados en *A. thaliana*, *B. rapa*, *B. oleracea*, *C. rubella* y *T. parvula* provenientes de la parte aérea y raíces en plantas, Identificando lncRNAs ortólogos funcionales con expresión conservada en raíces y parte aérea en esta etapa del desarrollo analizada. La mayoría de los lncRNAs identificados mostraron patrones globales de expresión. Asimismo, observamos que los eventos de duplicación del genoma tienen un impacto en la generación de nuevos genes de lncRNAs. La conservación y retención de los lncRNAs nos permitió generar un catálogo de lncRNAs de interés para caracterización experimental posterior, así como una posible aplicación futura en especies de interés agrícola.

Hipótesis y objetivos

- Hipótesis

Existen lncRNAs ortólogos funcionales que participan en el establecimiento de la identidad de órganos en plántulas de diversas Brasicáceas.

- Objetivos

Objetivo general

Identificar los lncRNAs ortólogos funcionales en la parte aérea y raíces de plántulas en diferentes Brasicáceas.

Objetivos específicos

OE1.- Anotar todos los lncRNAs en *A. thaliana* a partir de datos públicos e inferir sus posibles funciones biológicas (Capítulo 2).

OE2.- Secuenciar, identificar y clasificar los lncRNAs con base en su posición genómica en raíces y parte aérea de 5 especies de Brasicáceas (Capítulo 2 y 3).

OE3.- Determinar los genes diferencialmente expresados en la parte aérea y raíces en Brasicáceas (Capítulo 3).

OE4.- Identificar lncRNAs conservados por posición, secuencia (motivos) y estructura secundaria en Brasicáceas (Capítulo 3).

Capítulo 2

Anotación y clasificación funcional de lncRNAs en *A. thaliana*

Los resultados obtenidos en este capítulo deriva del artículo “**Transcriptome-guided annotation and functional classification of long non-coding RNAs in *Arabidopsis thaliana***” publicado en [Scientific Reports](#) (Coss-Navarrete, Evelia L.*; Corona-Gómez, José Antonio*; García-López, Irving J; Pérez-Patiño, Jaime Alejandro; Fernández-Valverde, Selene L, 2022. *Co-autores).

2.1 Métodos

2.1.1 Selección de transcriptomas públicos

Se emplearon 220 transcriptomas provenientes de datos públicos tomando en cuenta dos criterios: 1) una buena profundidad y calidad de secuenciación, 2) datos provenientes de una sola condición, órgano o etapa del desarrollo de *A. thaliana* (Col-0), sin incluir mutantes. Los transcriptomas fueron descargados de las bases de datos Gene Expression Atlas (GEA) (Petryszak et al., 2016) (152 transcriptomas) y Gene Expression Omnibus (GEO) (Edgar et al., 2002) (68 transcriptomas). Además, se emplearon los 4 transcriptomas generados de *A. thaliana* ([Sección 2.1.2](#)) provenientes de la parte aérea y raíces con una edad de 8 días después post-germinación (dpg). Contando con un total de 224 transcriptomas (118 *single-end* y 106 *paired-end*) ([Tabla 4](#), [Tabla S1](#), [Dataset S3](#)).

2.1.2 Generación de nuevos transcriptomas

Se germinaron semillas de *A. thaliana* Col-0 en el medio Murashige and Skoog (MS) en condiciones de días largos (21 °C, fotoperiodo 16 / 8h) y se recolectaron las plántulas con una edad aproximada a los 8 días después de la germinación [Etapa 1.02, cotiledones extendidos y 2 hojas de roseta con un tamaño de 1 mm de largo (Boyes, 2001)], seccionando las plántulas de manera individual en parte aérea (shoot) y raíces (root).

La extracción de RNA total se realizó empleando TRIzol (Invitrogen, 15596018), y de acuerdo a las especificaciones del proveedor. Se eliminó el DNA residual empleando la DNase I (TURBOTM DNase, Invitrogen, AM2238). La calidad y concentración de las muestras fue medido usando el NanoDrop 2000C (Thermo Fisher Scientific Inc). La integridad del RNA fue analizado mediante el gel de agarosa al 1.5 %, para posteriormente seleccionar RNA poliadenilado con el kit NEBNext Poly (A) mRNA Magnetic Isolation (NEBNext, E7490S). Las bibliotecas fueron generadas empleando el kit NEBNext Ultra II Directional RNA (NEBNext, E7760S) y NEBNext Multiplex oligos for Illumina (SET 1) (NEBNext, E7335). Posteriormente, la secuenciación se realizó por parte de la compañía Novogene en el equipo Hi-Seq X de Illumina, analizando 2x150 nt (PE150). Los transcriptomas obtenidos fueron cargados a la base de datos de SRA ([Tabla S1](#)).

Tabla 4. Transcriptomas empleados en la anotación de lncRNAs.

Clasificación	Experimento	Edad*	órgano	Número de transcriptomas
Línea base (Baseline)	Expresión de miRNAs en embriones	3 días después de la polinización	embriones	2
	Etapas de desarrollo del meristemo apical	7 a 16 dpg	meristemo apical (SAM)	30
	Búsqueda de lncRNAs	8 dpg	plántulas	4
	Búsqueda de lncRNAs	14 a 35 dpg	raíces, flores, frutos y hojas	4
	Atlas de tejidos	7 a 54 dpg	Múltiples órganos y tejidos	56
	Etapas de desarrollo de silicuas	0 a 20 dpg	silicuas	4
Expresión diferencial	Respuesta a la luz roja	5 dpg	hipocótilo y cotiledones	57
	Respuesta a la luz azul	3 dpg	plántulas	6
	Deficiencia de hierro	13 dpg	plántulas	6
	Deficiencia de fosfato	13 dpg	plántulas	6
	Aclimatación al frío	14 dpg	plántulas	6
	Respuesta al frío prolongado	14 dpg	plántulas	6
	Tratamiento con ABA	14 dpg	plántulas	17
	Respuesta a la sequía	14 dpg	plántulas	10

*dpg = días post-germinación

2.1.3 Limpieza, alineamiento, ensamblaje y cuantificación de los transcriptomas

Se analizó la calidad de los transcriptomas mediante los programas FastQC v0.11.2 (Andrew, 2010) y MultiQC v1.0 (Ewels et al., 2016). Se filtraron las secuencias de baja calidad y se eliminaron los adaptadores por medio del programa Trimmomatic v0.32 (HEADCROP:10-5 LEADING:5 SLIDINGWINDOW:4:15 MINLEN:30-60) (Bolger et al., 2014). Posteriormente, se realizó el alineamiento guiado con el genoma de referencia de *A. thaliana* TAIR10/Araport11 (Berardini et al., 2015) empleando STAR v2.7.2.b (--alignMatesGapMax 120000) (Dobin et al., 2013). El ensamblaje del genoma se realizó con el programa StringTie v.3.4 (-f 0.3 -m 50 -a 10 -j 15 -c 2.5) (M. Pertea et al., 2015) y empleando el archivo de anotación de Araport11 (C.-Y. Cheng et al., 2017). Los archivos obtenidos se unificaron en un solo archivo con la función --merge (-c 2.5 -f 0.3) de StringTie (-c 2.5 -f 0.3) (M. Pertea et al., 2015). La cuantificación de los transcritos se realizó con Kallisto v0.44.0 [los parámetros para los transcriptomas single-end: --single -t 8 -l (40, 67, 141 80) -s (5, 10, 20); para paired-end:

default] (Bray et al., 2016). La transformación de transcritos a genes se realizó con el programa Tximport v1.24.0 (Soneson et al., 2015a), a partir del archivo tx2gene ([Dataset S1](#)) generado con el programa ClusterBed de BEDtools (parámetros de default) (Quinlan & Hall, 2010).

2.1.4 Identificación de lncRNAs

Se reemplazó el *gene_id* por el *ref_gene_id* del archivo merge con un pequeño script de bash. Se obtuvo la secuencia FASTA de cada transcrito a partir del genoma de referencia y el archivo merge (previamente generado y modificado) empleando el programa gffread v0.11.8 (-F --keep-genes) (G. Pertea & Pertea, 2020). Además, se predijo la secuencia de aminoácidos para cada transcrito empleando el programa TransDecoder v5.3.0 (Haas et al., 2013). Para la identificación de lncRNAs se emplearon 9 filtros, seleccionados con base en trabajos previos (Cabili et al., 2011; Gaiti et al., 2015), se le asignó el nombre de Método estricto (ME) ([Sección 2.1.10](#)). En el artículo se menciona como Strict Method (SM). Primero, (1) se seleccionaron los transcritos mayores a 200 nt provenientes únicamente del núcleo mediante el programa infseq de EMBOSS v6.6.0 (Rice et al., 2000); (2) se eliminaron las secuencias homólogas a proteínas, comparando la secuencia nucleotídica (BLASTx, $\text{evalue} \leq 1e-6$) y (3) aminoacídica (BLASTp, $\text{evalue} \leq 1e-6$, strand="plus") (Altschul et al., 1990) de cada transcrito con la base de datos de Uniprot (UniProt Consortium, 2021), respectivamente. (4) Se eliminaron las secuencias que presentaban homología con dominios presentes en la base de datos de Pfam v33.0 (Finn et al., 2014; Rice et al., 2000) empleando el programa HMMER v3.1b2 ($\text{evalue} \leq 1e-6$) (Finn et al., 2011). (5) Además, se filtraron las secuencias con presencia de péptido señal con SignalP v4.1(D-cutoff: 0.45) (Petersen et al., 2011). (6) Se removieron secuencias con ORF > 100 aa empleando el programa getorf de EMBOSS v6.6.0 (Rice et al., 2000). Adicionalmente, (7) se eliminaron las secuencias con homología a proteínas no redundantes anotados en la base de datos de NCBI nr (Pruitt, 2004) usando BLASTx ($\text{evalue} \leq 1e-6$, strand="plus", pident ≥ 70.00) (Altschul et al., 1990). Con el fin de seleccionar lncRNAs anotados previamente como proteínas hipotéticas, se seleccionaron los mejores hits y se extrajo su información mediante el programa blastdbcmd (Altschul et al., 1990). Se retuvieron todas las secuencias anotadas en NCBI como: "hypothetical protein" (in Refseq), "similar to" (NCBI's annotation pipeline), "putative protein", "unknown (unknown protein, unknown, partial, unknown)", "predicted protein" y "unnamed protein product", considerando que la gran mayoría de estas secuencias no cuentan con caracterización funcional ("Gene Frequently Asked Questions," 2018). (8) Se eliminaron todos transcritos relacionados con tRNAs y rRNAs comparando su secuencia con los anotados en *A. thaliana* empleando BLASTn ($\text{evalue} \leq 1e-6$, strand = "plus", -max_target_seqs 100, qcovs >50) (Altschul et al., 1990) y los predichos por el modelo de covarianza de la base Rfam (Nawrocki et al., 2015) empleando el programa infernal v1.1.2 (Nawrocki & Eddy, 2013). (9) Con el fin de eliminar los artefactos generados por el alineador se filtraron las secuencias con intrones mayores a 6000 bp (H.-X. Sun & Chua, 2019). Finalmente, se realizó una revisión manual comparando los transcritos obtenidos con la base de datos de Araport11, eliminando los transcritos clasificados como proteínas transmembranales, enzimas, genes codificantes de proteínas y proteínas ribosomales ([Sección 2.1.10](#)).

2.1.5 Clasificación de los lncRNAs

Los lncRNAs se clasificaron con base en su posición genómica en: **lincRNA**, **intrónico**, **NAT y exónico sentido**, categorías y definiciones señaladas por GENCODE v7 para clasificar los lncRNAs en humanos (Derrien et al., 2012), y Gaiti y colaboradores (Gaiti et al., 2015) en esponjas marinas. En todas las categorías se consideraron las isoformas de los genes codificantes para la correcta clasificación de los lncRNAs. Adicionalmente, se extrajeron lncRNAs que no presentaron dirección en su secuencia, es decir, lncRNAs a los cuales el programa StringTie les asignó un punto (.) en el sentido de la secuencia. Clasificándolos como: **intrónico sin sentido**. Se empleó el archivo de anotación de Araport11 (C.-Y. Cheng et al., 2017) y el programa intersectBed de BEDtools (sense-exonic lncRNAs [-wo -f 0.1 -s], NAT [-wo -f 0.1 -S], intronic [-wo -f 1] and lincRNAs [-wo -v]) (Quinlan & Hall, 2010) para la ubicación de los lncRNAs con relación a los genes codificantes y de esta manera poder clasificarlos en las categorías antes descritas ([Sección 2.1.10](#)). Posteriormente, se realizó la visualización de los lncRNAs clasificados mediante el buscador de genomas UCSC (Kent et al., 2002).

Por su parte, los lncRNAs se clasificaron de acuerdo con su nivel de expresión en **Low Confidence (SM LC)**, aquellos genes con expresión menor a 3 Transcripts per Million (TPM) en un solo transcriptoma; mientras que el resto fue asignado como **High Confidence (SM HC)**.

2.1.6 Análisis del potencial codificante

Como un análisis de validación adicional del ME se empleó el programa CPAT v3.04 (L. Wang et al., 2013) para estimar el potencial codificante de las secuencias obtenidas. Este programa emplea las frecuencias kmer relativas, crea un modelo de los datos con las frecuencias de los hexámeros y se obtiene un score. Este valor permite clasificar a los transcritos en cuatro categorías: codificante (score > 0.5), no codificante (score <= 0.5), alta confianza de ser una secuencia codificante (High Confidence coding, score >= 0.9) y alta confianza de ser una secuencia no codificante (High Confidence non-coding, score <= 0.1).

2.1.7 Comparación con otras bases de datos

Para conocer cuántos de los 6764 genes de lncRNAs identificados con el ME contaban con una identificación previa en otras bases de datos, se realizó una comparación entre las coordenadas de los genes anotados en GreenNC v1.12 (2752 genes) (Di Marsico et al., 2022; Paytuví Gallart et al., 2016), CANTATAdb v2.0 (4080 genes) (Szcześniak et al., 2016, 2019) y Araport11 (3559 genes) (C.-Y. Cheng et al., 2017) empleando el programa intersectBed (-wo -s -f 1 -F 1) (Quinlan & Hall, 2010). Se corroboró la intersección de cada transcrito con el visualizador UCSC y se resumieron los resultados empleando los paquetes VennDiagram v1.7.1 (H. Chen & Boutros, 2011) y UpsetR v1.4.0 (Conway et al., 2017) (R/Bioconductor).

2.1.8 Cuantificación de los lncRNAs por órganos y etapas

Se dividieron los transcriptomas de acuerdo a cada órgano analizado, y posteriormente se clasificaron los diferentes órganos en etapas del desarrollo ([Fig 6b](#)). Las dos primeras etapas

pertenecen a la fase vegetativa e incluyen: la germinación de la semilla (Etapa 1, de 3 a 5 días de edad) y el desarrollo de las hojas (Etapa 2, de 6 a 25 días de edad); el resto de las etapas forman parte de la fase reproductiva, que van desde la presencia de la primera inflorescencia (a los 26 días de edad) (Etapa 3, de 26 a 29 días de edad), la producción de flores (Etapa 4, de 30 a 47 días de edad), hasta la generación de silicuas (Etapa 5, de 48 a 51 días de edad) ([Tabla S1](#), [Fig 6b](#)).

Para identificar lncRNAs con expresión específica del órgano se empleó el estimador *Tau* (Yanai et al., 2005). El resultado tiene un rango de 0 a 1, donde los genes cercanos a 1 son considerados específicos del órgano o etapa del desarrollo ([Fig S2a](#), [Dataset S4](#)). Se empleó como valor de corte en la asignación de los lncRNAs la media del valor *Tau* para los genes codificantes (0.54) ([Dataset S4](#), [Sección 2.1.10](#)).

2.1.9 Redes de expresión y análisis de culpable por asociación

Para determinar las posibles funciones biológicas de los lncRNAs, se empleó el método culpable por asociación (*guilty-by-association*). Este enfoque identifica las anotaciones funcionales enriquecidas de genes codificadores de proteínas coexpresados con los lncRNAs, lo que permite inferir los procesos biológicos en los que estos lncRNAs pueden estar involucrados. Se construyó una red de co-expresión empleando el paquete WGCNA v1.69 (R/Bioconductor) (Langfelder & Horvath, 2008) y la tabla de normalizadas con *Variance Stabilizing Transformation* (VST) (de paquete DESeq2 v1.28.1 (R/Bioconductor) (Love et al., 2014a). Para asegurar una conectividad promedio en la red se empleó un valor de $\beta = 20$, permitiendo mantener la topología sin escala de forma estacionaria ([Fig S3](#)). Los grupos de genes co-expresados fueron nombrados módulos de co-expresión de acuerdo con la nomenclatura obtenida del programa WGCNA. La red generada fue de tipo *signed* con una correlación bicor (biweight midcorrelation), y los resultados para cada módulo se mantuvieron separados (unmerged). Los perfiles de expresión se representan por su componente principal (módulo eigengene), el cual nos indica la variación de los niveles de expresión en cada módulo, con un tamaño mínimo de 50 genes por módulo ([Sección 2.1.10](#)).

Para identificar las funciones asociadas a cada módulo de expresión, se realizó el análisis de enriquecimiento de categorías funcionales (Gene Ontology, GO, v3.11.4) con topGO v2.40.0 (R/Bioconductor) (Alexa et al., 2006), se analizaron las categorías de: componente celular (CC, Cellular Component), proceso biológico (BP, Biological Process) y función molecular (MF, Molecular Function). Finalmente, para evaluar la importancia del enriquecimiento de las categorías GO se empleó la prueba exacta de Fisher para múltiples pruebas (Benjamini-Hochberg, BH) ($q_{val.bh} < 0.01$, FDR < 1 %). Se agruparon las categorías, eliminaron términos redundantes y visualizaron los resultados mediante ReviGo v1.6.0 (rrvgo) (Supek et al., 2011) y el paquete treemap v 2.4-6 (R/Bioconductor).

2.1.10 Visualización de los datos y scripts

Todas las anotaciones de lncRNAs fueron subidas al visualizador de genomas de UCSC. Los archivos de coordenadas (formato BED) de los genes de lncRNAs identificados y clasificados se encuentran disponibles en el [Dataset S1](#).

Los scripts empleados en el análisis se encuentran en el repositorio de Github:

https://github.com/RegRNALab/Transcriptome-guided_lncRNA_annotation.

La visualización de los lncRNAs con las diferentes bases de datos se encuentra en UCSC, en el siguiente link:

https://genome.ucsc.edu/s/EveCoss/lncRNAs_mydata_SM

2.2. Resultados

2.2.1 Identificación y clasificación de lncRNAs

Se identificaron 6764 genes (7070 transcritos) como lncRNAs provenientes de los 224 transcriptomas en *A. thaliana* empleando el ME propuesto. Estos incluyeron 4354 lincRNAs (4548 transcritos), 2060 NATs (2133 transcritos), 213 exónico sentido (248 transcritos, sense-exonic), y 185 intrónico (187 transcritos, intronic) ([Fig 5a](#), [Dataset S1](#)), 78 lncRNAs intrónicos no tenían orientación transcripcional (sentido) ya que fueron identificados en transcriptomas *single-end*. Además, 33 genes de lncRNA (46 transcritos) se clasificaron como NATs y exónico sentido debido a la posición del lncRNA flanqueado por genes codificantes tanto sentido como antisentido en la cadena de DNA. Se verificó manualmente que no se trataba de UTRs 3' extendidas de genes codificantes de proteínas que se solapan. Además, 15 genes tenían isoformas que pertenecían a diferentes categorías, debido a la posición de sus isoformas ([Dataset S2](#)). Para proporcionar una medida de la expresión observada para los lncRNAs poco expresados, clasificamos aquellos que tenían menos de 3 TPMs en un único transcriptoma como de Low Confidence (SM LC) y los restantes lncRNAs como de High Confidence (SM HC) ([Dataset S2](#)). Se utilizó el umbral de un solo transcriptoma ya que hay numerosos órganos (carpelo, pedicelo de la flor, pétalo, pecíolo, polen, sépalo, septum, estambre, internodo del tallo, nodo del tallo y valvas) para los que sólo tenemos un único transcriptoma ([Dataset S1](#)). Además, evaluamos el potencial de codificación de los lncRNAs identificados por el ME (o SM) utilizando CPAT (Kern et al., 2018). Encontramos que los lncRNAs presentaban valores de potencial codificantes significativamente más bajos que los genes codificantes en Araport11 ([Fig S4a](#)) y que la gran mayoría de ellos fueron clasificados como no codificantes o no codificantes de alta confianza (High Confidence non-coding) por CPAT ([Fig S4b](#)).

Además, los lncRNAs identificados tienen menos exones por transcrito (mediana 1; media 1,23) ([Fig S5a](#)) que los genes codificantes (mediana 4; media 6). Además, sus transcritos maduros son más pequeños (media de 437,3 nt) que los de los genes codificantes (media de 1799 nt) ([Fig S5b](#)). Estas características coinciden con lo observado anteriormente en animales (M.-J. M. Chen et al., 2016; Darbellay & Necsulea, 2020; Gaiti et al., 2015; Hezroni et al., 2015), moscas (Deng et al., 2018) y otras plantas (He et al., 2021; P. Wang et al., 2020; X. Zhao et al., 2019; Y. Zhu et al., 2017).

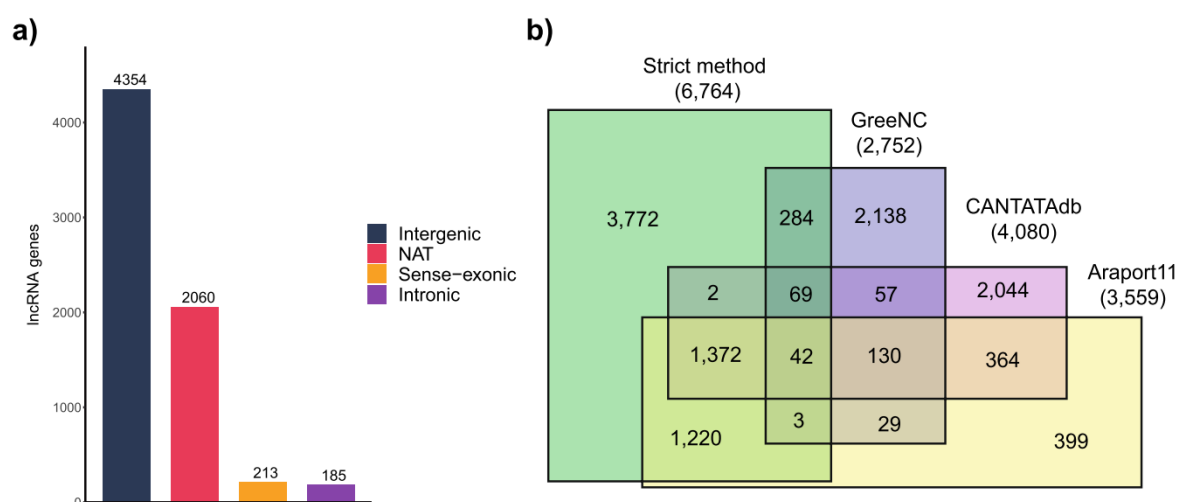


Figura 5. Anotación de los genes de lncRNAs y sus comparaciones reportadas en otras bases de datos de lncRNAs en plantas. (a) Distribución de los 6764 genes de lncRNAs predichos por el ME (SM). (b) Diagrama de Venn de las comparaciones entre las bases de datos con el ME (SM, verde), GreenNC (morado), CANTATAdb (rosa) y Araport11 (amarillo) donde se han anotado los lncRNAs de *A. thaliana*.

2.2.2 Similitudes y diferencias con otras bases de datos

El total de lncRNAs anotados por el ME (o SM) (6764 genes) supera a las bases de datos más destacadas de *A. thaliana*: GreenNC (v1.12) tiene 2752 genes (3008 transcripciones) (Paytuví Gallart et al., 2016), CANTATAdb (v2.0) 4080 genes (4373 transcripciones) (Szcześniak et al., 2019) y Araport11, 3559 genes (3970 transcripciones) (C.-Y. Cheng et al., 2017). Una comparación con estas bases de datos reveló que 3772 genes lncRNAs en nuestra anotación son nuevos y no han sido previamente reportados en ninguna de estas bases de datos (Fig 5b); los nuevos lncRNAs fueron categorizados en 2326 lincRNAs (2454 transcritos), 1218 NATs (1227 transcritos), 111 exónico sentido (124 transcritos) y 145 intrónico (146 transcritos). Estos nuevos lncRNAs representan un aumento del 93.08 % (2275 sobre 2444) en el número de lincRNAs en lincRNAs y un aumento del 134.70 % (1502 sobre 1115) en NATs, con respecto a la base de datos Araport11. Además, encontramos que 398 genes de lncRNAs son compartidos entre nuestra anotación y la base de datos GreenNC (Paytuví Gallart et al., 2016), 1485 con CANTATAdb (Szcześniak et al., 2019), y 2637 con Araport11 (C.-Y. Cheng et al., 2017), siendo la base de datos Araport11 la que mejor concuerda con nuestros datos; nuestra anotación contiene aproximadamente el 74,09 % (2637 sobre 3559) de los lncRNAs anotados en Araport11 (Fig 5b).

Sólo 130 lncRNAs son compartidos entre las bases de datos GreenNC, CANTATAdb, y Araport11, y sólo hay 42 lncRNAs compartidos entre las cuatro anotaciones (Fig 5b). Es importante señalar que es probable que hayan otros lncRNAs en *A. thaliana* que no se identifican en nuestro análisis, ya que no todas las condiciones, los órganos y las etapas de desarrollo se han estudiado utilizando RNA-Seq. Sin embargo, nuestra anotación es la primera que aprovecha la mayoría de los datos transcriptómicos disponibles para esta especie, asegurando que las secuencias obtenidas son sólo las de los lncRNAs expresados.

Esto, combinado con un método de anotación robusto, evita la redundancia con otros tipos de transcritos que no son lncRNAs.

Al comparar nuestra anotación con Araport11, observamos que nuestras anotaciones no siempre coincidían en la misma clasificación de biotipos. La clasificación más concordante entre ambas anotaciones fue entre los lncRNAs, donde 1747 genes lncRNAs corresponden a la misma anotación ([Fig. S6](#)). Sin embargo, varios lncRNAs identificados en nuestra anotación no están clasificados como lncRNAs en Araport11: 288 genes lncRNA (265 lncRNAs, 12 NATs, 6 exónico sentido, 4 lncRNA-NAT, y 1 exónico sentido lncRNA) están anotados en Araport11 como "nueva región transcrita", y 388 como "genes codificantes" (217 lncRNAs, 110 NATs, 43 exónico sentido, 13 NAT-exónico sentido, 3 exónico sentido lncRNAs, 1 lncRNA-NAT y 1 intrónico) ([Fig. S6](#)).

2.2.3 Identificación de lncRNAs con validez experimental

Revisamos manualmente los 388 lncRNAs clasificados como "genes codificantes" en Araport11 y concluimos que se tratan, de lncRNAs erróneamente anotados como "genes codificantes" en Araport11. Entre ellos, encontramos *IPS1* (*Induced by Phosphate Starvation 1*, AT3G09922), un lncRNA con una función de objetivo de imitación para el microRNA miR399 en ausencia de fósforo (Franco-Zorrilla et al., 2007b). Otro lncRNA erróneamente clasificado fue el parálogo de *IPS1*, *At4* (AT5G03545) (Shin et al., 2006a), que es funcionalmente redundante a *IPS1*. Ambos han sido previamente validados experimentalmente y se ha encontrado que están conservados en varias especies de plantas (Ajmera et al., 2018; Burleigh & Harrison, 1999; Calixto et al., 2019a; C. Y. Huang et al., 2011). De forma similar, el lncRNA *APOLLO* (*AUXIN-REGULATED PROMOTER LOOP*, AT2G34655) (Ariel et al., 2014a) está anotado como un gen codificador de proteínas. También encontramos múltiples lncRNAs erróneamente anotados como snoRNAs, nuevas regiones transcritas y otros RNAs, incluyendo los lncRNAs validados experimentalmente: *HID1* (*HIDDEN TREASURE 1*, AT2G35747) (Y. Wang et al., 2014a), *MARS* (*MARneral Silencing*, AT5G00580) (Roulé, Christ, et al., 2022), y *DRIR* (*Drought-induced RNA*, AT1G21529) (Qin et al., 2017a), respectivamente ([Tabla 5](#)).

Además de estas categorías, identificamos numerosos lncRNAs que fueron anotados como elementos transponibles (92, reclasificados como 91 lncRNAs y 1 NAT), otros RNA (83: 77 lncRNAs, 3 lncRNA-NAT, 1 NAT, 1 exónico sentido y 1 exónico sentido-lncRNA), pseudogenes (48: 43 lncRNAs, 2 NAT, 2 exónicos sentido y 1 NAT-lncRNA), snoRNA (5: 5 lncRNAs) y snRNA (1: 1 NAT) ([Fig. S6](#)). Finalmente, encontramos 3222 genes de lncRNAs que no son compartidos entre Araport11 y nuestra anotación. Estos lncRNAs comprenden 1885 lncRNAs, 1117 NATs, 143 intrónicos, 67 exónicos sentido, y 10 genes compartidos entre NATs y exónico sentido (9 genes), un gen intrónico y exónico sentido ([Fig. S6](#)). Estos últimos 10 genes tenían dos anotaciones debido a que tenían isoformas que pertenecían a dos categorías diferentes. Esta comparación muestra que la anotación de lncRNAs en Araport11, una de las bases de datos de referencia más prominentes para *A. thaliana*, tiene importantes imprecisiones que se resuelven en nuestras anotaciones, lo que resulta en una mejora en la clasificación de los genes lncRNA.

Tabla 5. Clasificaciones de algunos lncRNAs encontrados en Araport11.

Annotated lncRNAs	Araport11 annotation	Araport11 gene IDs	Other lncRNAs databases	Biological role	Biological condition	Experimental validation
<i>IPS1</i>	protein-coding	AT3G09922	GreenC, EVLncRNAs and PNRD	Phosphate homeostasis	Phosphate deficiency	(Franco-Zorrilla et al., 2007a)
<i>At4</i>	protein-coding	AT5G03545	GreenC, EVLncRNAs and PNRD	Phosphate homeostasis	Phosphate deficiency	(Shin et al., 2006b)
<i>APOLO</i>	protein-coding	AT2G34655	GreenC, EVLncRNAs and PNRD	Auxin-controlled development	Auxin	(Ariel et al., 2014b)
<i>HID1</i>	snoRNA	AT2G35747	EVLncRNAs	Photomorphogenesis	Continuous red light	(Y. Wang et al., 2014b)
<i>DRIR</i>	other RNA	AT1G21529	NONCODE (NONATHT000172.1), PNRD (NONATHT000172) and EVLncRNAs	Stomatal closure	Drought, salinity and ABA	(Qin et al., 2017b)
<i>ELENA1</i>	other RNA	AT4G16355	GreenC and CANTATAdb, NONCODE (NONATHT002899.1.1)	Defense	<i>Pseudomonas syringae</i> pv.tomato DC3000	(Seo et al., 2017b)
<i>TAS1a</i>	other RNA	AT2G27400	GreenC	Acclimation and freezing tolerance	Cold	(Calixto et al., 2019b)
<i>MARS</i>	novel transcribed region	AT5G00580	-	Germination	Response to ABA	(Roulé, Ariel, et al., 2022)
<i>BLIL1</i>	NAT	AT1G26218	CANTATAdb, GreenC and NONCODE (NONATHT000092.1)	Photomorphogenesis	Blue light	(Z. Sun et al., 2020b)
<i>COOLAIR</i>	NAT	AT5G01675	EVLncRNAs and CANTATAdb	Flowering time	Cold	(Csorba et al., 2014)
<i>FLORE</i>	NAT	AT1G69572	CANTATAdb, GreenC and NONCODE (NONATHT000834.1.1)	Circadian rhythms	Long days, short day and 12h light/12h dark	(Henriques et al., 2017a)
<i>FLINC</i>	lincRNA	AT1G08103	PlncDB (ID=ncrna9858)	Flowering time	At 16 °C	(Severing et al., 2018)

Cabe destacar que dentro de nuestra anotación, 48 genes lncRNA (120 transcritos) tienen una anotación ambigua, ya que están anotados simultáneamente como NAT y exónico sentido (33 genes lncRNA; 93 transcritos), lincRNAs y exónico sentido (5 genes; 12 transcritos), lincRNAs y NAT (9 genes; 13 transcritos), e intrónico y exónico sentido (1 gen; 2 transcritos) ([Dataset S2](#)). En concreto, en el caso de los lncRNAs anotados como NAT exónico sentido, se superponen a dos genes diferentes que codifican proteínas, adquiriendo así una anotación separada para cada gen. Del mismo modo, otros genes de lncRNAs

cuentan con isoformas en diferentes categorías, dependiendo de la localización genómica de cada isoforma.

2.2.4 Patrones de expresión de lncRNAs

Además de anotar los lncRNAs, aprovechamos la información transcriptómica para explorar cuántos lncRNAs se expresaban entre los órganos, etapas de desarrollo y condiciones de *A. thaliana* (Fig 6). Encontramos más lncRNAs expresados en la flor, la raíz, la plántula y el silicua (Fig 6a).

Los órganos con mayor diversidad de tipos celulares, como las flores, silicuas, raíces, plántulas y hojas tenían un mayor número de lncRNAs (Fig 6a). Esta tendencia se ha observado previamente en animales, donde los órganos con más diversidad de tipos celulares, como el cerebro, expresan más lncRNAs (Grote & Herrmann, 2015; Mattick, 2011; Mercer et al., 2008). También se sabe que los órganos reproductivos albergan una mayor diversidad de lncRNAs. Del mismo modo, en nuestros datos, las flores tienen más lncRNAs que otros órganos (Fig 6a). Curiosamente, el número de lncRNAs expresados en la flor es mucho mayor que en sus partes individuales (estambre, sépalo, pétalo, carpelo y pedicelo), sugiriendo además que la alta diversidad de órganos y tipos celulares de este órgano puede deberse a los múltiples tejidos que lo componen. Un enriquecimiento de lncRNAs en los órganos reproductivos ha sido reportado previamente en múltiples especies de plantas como la soja, el maíz y el arroz (Golicz, Singh, et al., 2018; Li et al., 2014; Y.-C. Zhang et al., 2014) y en testículos de animales (Brown et al., 2014; Cabili et al., 2011; Golicz, Bhalla, et al., 2018; Necsulea et al., 2014). Las plántulas presentan un gran número de lncRNAs expresados es la de las plántulas (Fig 6a), compuestas por una mezcla de órganos en una determinada etapa de desarrollo. Como la mayoría de los transcriptomas de condiciones de estrés abiótico utilizados en este estudio eran de plántulas, muchos lncRNAs expresados en respuesta a estos estreses se asignan a la categoría de plántulas (Fig 6a). Además, el número de transcriptomas y la profundidad de secuenciación en cada categoría se correlaciona positivamente con el número de lncRNAs encontrados (Fig 6a).

En términos de desarrollo, la fase de germinación (etapas 1 y 2) tiene el mayor número de lncRNAs (Fig 6c), seguido de la etapa 4 (desarrollo de la flor) (Fig 6), y no parece correlacionarse con el número de transcriptomas en cada etapa de desarrollo (Fig 6c). Las etapas de desarrollo en las que se produce la diferenciación de órganos o la formación de órganos tienden a expresar múltiples lncRNAs tanto en plantas (Bhatia et al., 2019; S. Liu et al., 2018; “LncRNA/circRNA–miRNA–mRNA Networks Regulate the Development of Root and Shoot Meristems of Populus,” 2019; M. Zhu et al., 2017) como en animales (Darbellay & Necsulea, 2020; Grote & Herrmann, 2015; Necsulea et al., 2014; Sarropoulos et al., 2019). Desafortunadamente, las primeras etapas de diferenciación de órganos no están representadas en nuestro conjunto de datos, lo que podría ayudarnos a identificar lncRNAs que participan en la formación de órganos.

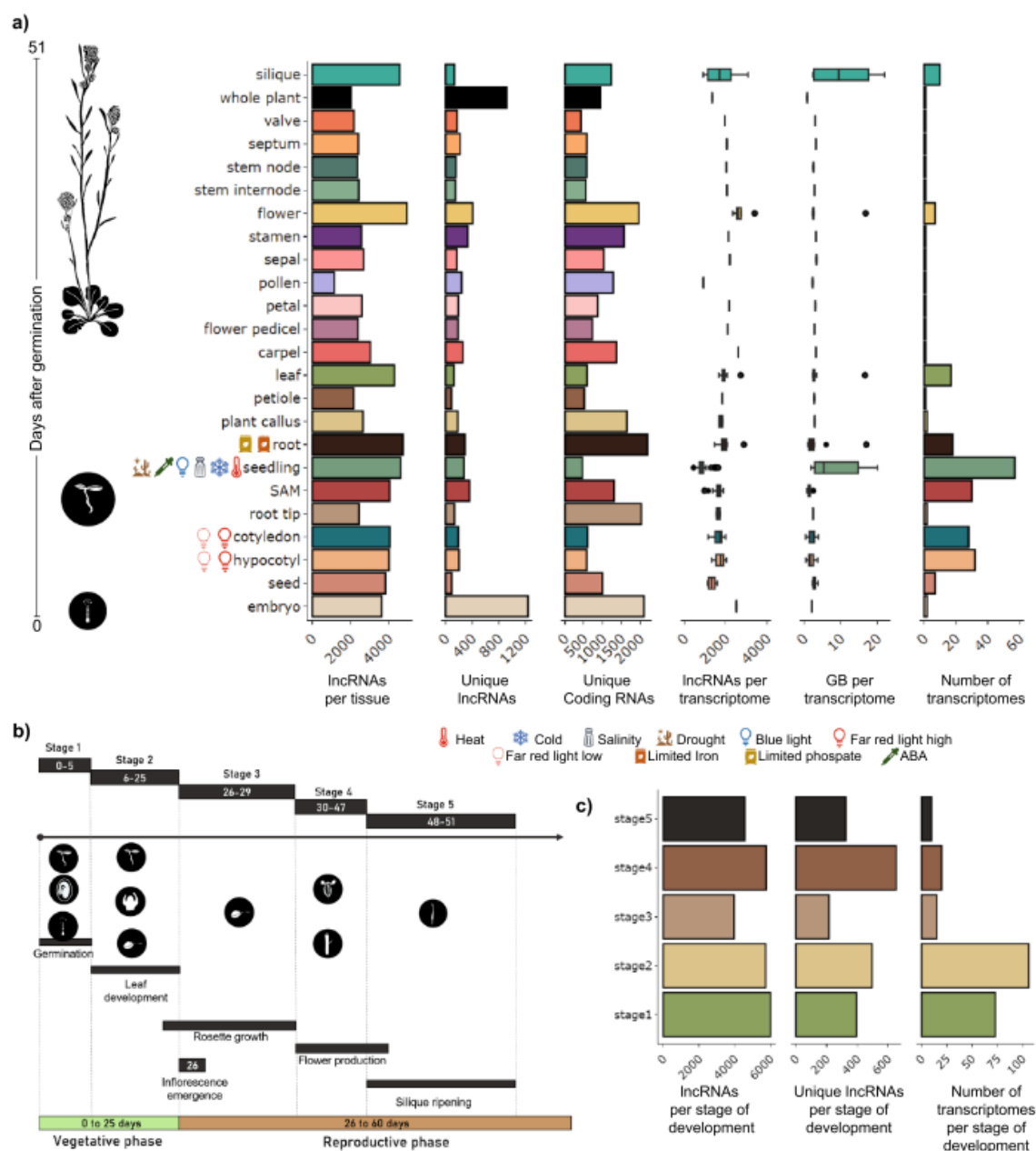


Figura 6. Expresión de lncRNAs por órgano y etapa del desarrollo en *A. thaliana*. (a) Histogramas del número de lncRNAs únicos, RNAs codificantes únicos, número de lncRNAs por transcriptoma, GB por transcriptoma y número de transcriptomas que encontrados en cada órgano y etapa del desarrollo. Descripciones de izquierda a derecha de la gráfica, el orden asignado fue de acuerdo con los días correspondientes a cada etapa del desarrollo. (b) Los órganos se agrupan en cinco etapas de vida de *A. thaliana* utilizando su posición dentro de la progresión del desarrollo de *A. thaliana*. (c) Número de lncRNAs por etapa del desarrollo. El histograma contiene el número de lncRNAs identificados en cada una de estas etapas, mientras que el segundo histograma contiene el número de lncRNAs únicos para cada una de estas etapas.

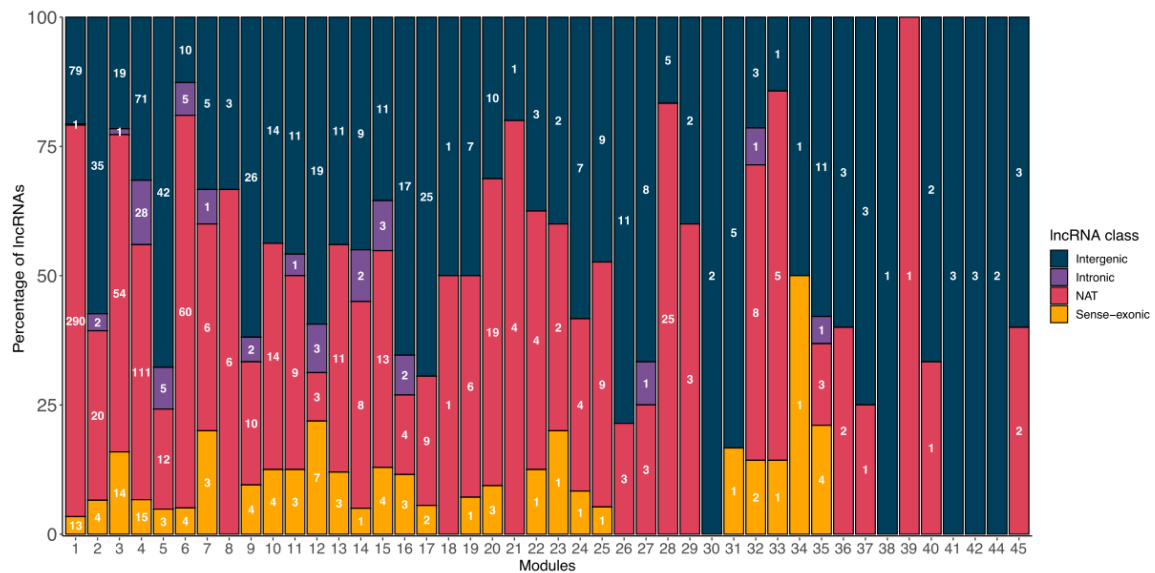


Figura 7. lncRNAs por módulo de co-expresión. Número de lncRNAs en cada módulo de co-expresión y clasificados por biotipo: Intergénico: (lincRNA, azul), Intrónico (morado), NAT (rosa), Sentido-exónico (amarillo).

2.2.5 Expresión de lncRNA específicos de órgano y etapa

Los genes que se expresan específicamente en un órgano o etapa de desarrollo concreto pueden ser importantes para establecer la identidad de ese órgano o etapa del desarrollo. Encontramos que los lncRNAs en *A. thaliana*, como en la mayoría de los organismos, se expresan de una manera más específica en un órgano en comparación con los genes codificantes (Fig S2b). El embrión y toda la planta adulta contienen la mayor cantidad de lncRNAs únicos, mientras que la raíz y el embrión expresan más genes codificantes únicos (Fig 6a). El embrión tiene el mayor número de lncRNAs únicos, a pesar de no ser el más abundante en lncRNAs (Fig 6a). Además, la raíz expresó la mayoría de los genes codificantes y lncRNAs únicos (Fig 6a). Curiosamente, había muchos más genes codificantes únicos en la raíz y casi ningún lncRNA único expresado (Fig 6a). No observamos un aumento de genes únicos en los órganos con diversas condiciones de estrés. Además, la mayoría de los lncRNAs únicos se expresan en la fase reproductiva de la planta más que en la fase vegetativa (Fig 6c). Dividiendo los lncRNAs por biotipo, encontramos que el 67.1 % (143 de 213) de los exónico sentido, el 64 % (1318 de 2060) de los NATs, el 59.8 % (2604 de 4354) de los lincRNAs, y el 48.1 % (89 de 185) de los lncRNAs intrónicos pertenecen a un solo órgano o fase. Estos resultados indican una alta especificidad de los lncRNAs en los diferentes órganos y etapas del desarrollo.

2.2.6 lncRNAs con funciones específicas de órgano conocidas

Algunos lncRNAs con funciones conocidas muestran una alta especificidad tisular por la medida de *Tau* que concuerda con su órgano funcional reportado (Tabla S2). Entre estos, encontramos los lncRNAs *IPS1* y *At4*, que tienen funciones relacionadas con la inanición de fósforo (Franco-Zorrilla et al., 2007b; Shin et al., 2006a), y el lncRNA *MARS*, que está involucrado en los cambios de la conformación de la cromatina en respuesta al ABA (Roulé,

Christ, et al., 2022). Como se esperaba, estos tres lncRNAs tienen altos valores de especificidad tisular en los órganos de la raíz ([Dataset S4](#)). Además, el lncRNA *FLINC*, relacionado con la regulación de la floración (Severing et al., 2018), está específicamente enriquecido en el meristemo apical (SAM). Por otro lado, la expresión específica de órgano de algunos lncRNAs funcionales conocidos corresponden con el órgano en el que se ha reportado su función. Tal es el caso de *HID1*, un lncRNA implicado en la elongación del hipocótilo (Y. Wang et al., 2014a), que tiene una alta especificidad tisular en el SAM ([Tabla S2](#)), a pesar de que se ha encontrado previamente que se expresa de forma ubicua (Y. Wang et al., 2014a). Del mismo modo, *APOLO*, que participa en el desarrollo de las raíces laterales en respuesta a la auxina (Ariel et al., 2014a; Bhatia et al., 2019), tiene una alta especificidad tisular en el pecíolo ([Tabla S2](#)). Esta discrepancia se debe probablemente a la falta de raíces tratadas con auxina en nuestro conjunto de datos, que es donde esperábamos ver los valores más altos de *APOLO*.

2.2.7 Co-expresión de lncRNAs con genes codificante

Para inferir una posible función para todos los lncRNAs anotados, utilizamos un enfoque llamado culpable por asociación. Para ello, construimos una red de co-expresión que incluía todos los genes codificantes y no codificantes utilizando WGCNA (Love et al., 2014a). Un total de 224 transcriptomas con 34,937 genes fueron analizados para construir esta red.

Obtuvimos un total de 45 módulos de co-expresión ([Fig 7](#)). Se encontraron 1425 (21 %) genes lncRNA en 44 de los 45 módulos de co-expresión. En total, 516 lincRNAs, 746 NAT, 104 sentido-exónico y 59 intrónico fueron co-expresados. El módulo 1 albergó la mayor cantidad de lncRNAs, con 383 de ellos, principalmente NATs (290), seguidos por lincRNAs (79), 13 sense, y un lncRNA intrónico ([Fig 7](#)). De acuerdo con el enriquecimiento funcional para procesos biológicos, este módulo destacó por procesos relacionados con la fotosíntesis, la organización de los cloroplastos y la respuesta a la luz. Los siguientes módulos con mayor número de lncRNAs son los módulos 4 y 3; estos módulos están relacionados con el procesamiento y la transcripción del RNA. En total, el 91 % (40/44) de los módulos que albergaban lncRNAs presentaron un enriquecimiento funcional para procesos biológicos. De manera interesante, 746 nuevos lncRNAs se co-expresaron con genes codificantes y se distribuyeron entre 40 módulos. Los módulos 1, 3, 4 y 6 fueron los que más lncRNAs anotaron, la mayoría de ellos lncRNAs NAT ([Fig. S7](#)). Cabe destacar que la mayoría de los nuevos lncRNAs (3026, 80.2 %) en nuestra anotación no se co-expresaron con genes codificantes.

Encontramos que los 40 módulos de co-expresión que albergaban lncRNAs y tenían un enriquecimiento funcional podían ser agrupados en 9 categorías ([Fig S8](#) a [S16](#)) principalmente por su función seguida de sus valores de eigengene (expresión de genes representativos) en los diferentes órganos o etapas de desarrollo (Langfelder & Horvath, 2008). Estas categorías funcionales son la organización del cloroplasto y la fotosíntesis (4 módulos con 409 lncRNAs) ([Fig S8](#)), la regulación del RNA y la transcripción (4 módulos con 375 lncRNAs) ([Fig S9](#)), el desarrollo de la raíz y la respuesta al estrés relacionado con la raíz (5 módulos con 125 lncRNAs) ([Fig S10](#)), el etiquetado de proteínas y el transporte con (5 módulos con 117 lncRNAs) ([Fig. S11](#)), división celular (5 módulos con 112 lncRNAs) ([Fig. S12](#)), lípidos y membranas (3 módulos con 97 lncRNAs) ([Fig. S13](#)), respuesta a patógenos

(4 módulos con 72 lncRNAs) ([Fig. S14](#)), reparación del ADN (2 módulos con 61 lncRNAs) ([Fig. S15](#)) y respuesta al estrés (5 módulos con 17 lncRNAs) ([Fig. S16](#)).

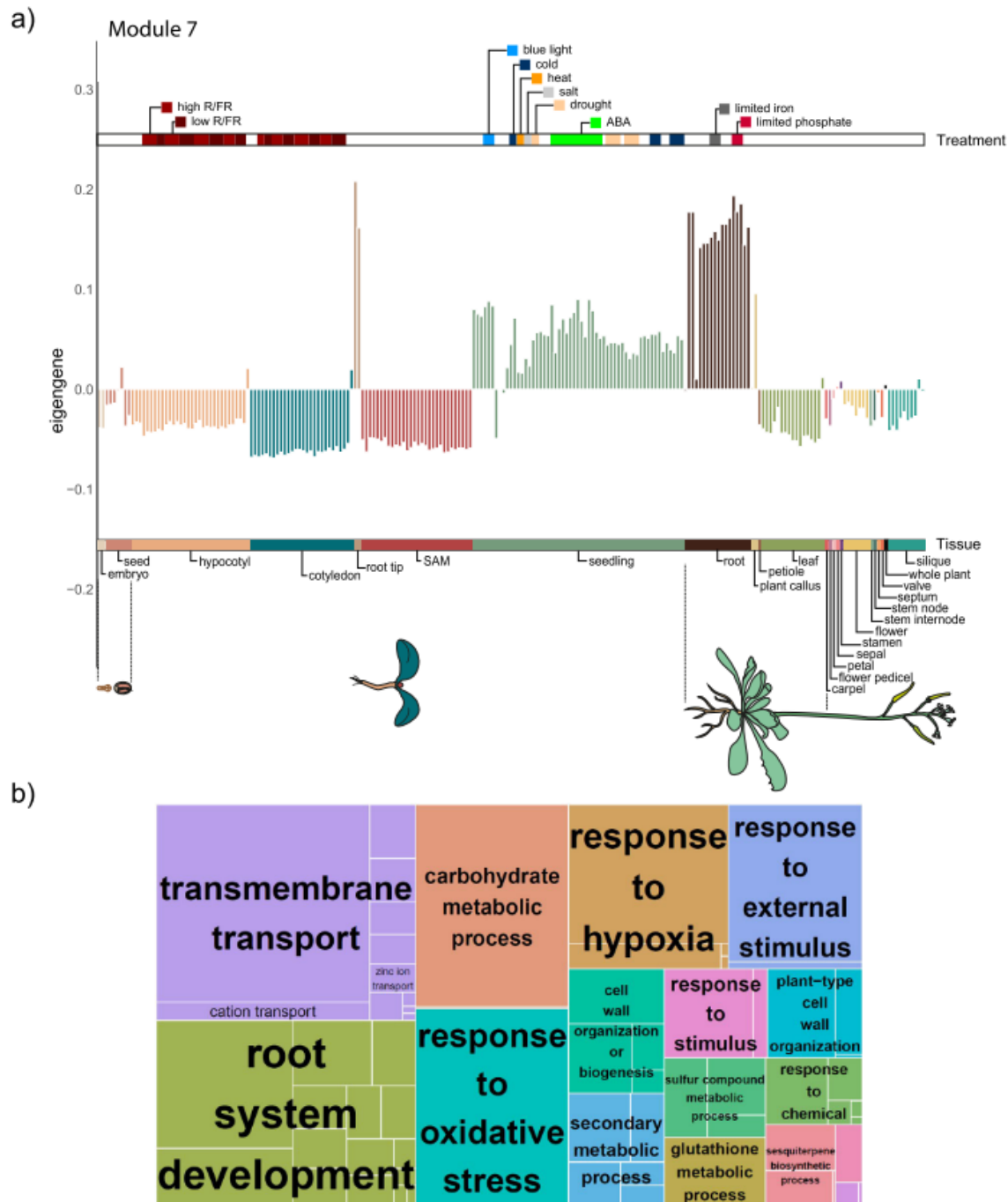


Figura 8. Patrones de expresión y Funciones sobresalientes localizadas en el Módulo 7. (a) Diagrama de barras de la expresión representativa (eigengene) por transcriptoma en el módulo 7, genes enriquecidos con alta expresión y que participan en el desarrollo de la raíz. Cada barra representa la expresión en un transcriptoma, organizada por órgano y ordenada por edad de menor (izquierda) a mayor (derecha). La barra inferior describe a qué órgano pertenece cada color, mientras que la superior los diferentes tratamientos experimentales de cada transcriptoma. (b) Treemap de funciones biológicas enriquecidas en módulo 7.

Encontramos que el mayor número de lncRNAs coexpresados se encuentran en la categoría funcional enriquecida para la organización del cloroplasto y la fotosíntesis, con valores positivos de eigengene de expresión en órganos relacionados con la fotosíntesis como hojas, cotiledones e hipocótilos. Estos lncRNAs se dividen en cuatro módulos ([Fig. S8](#)) relacionados con funciones más específicas como la respuesta a la radiación (respuesta a la luz roja, a la luz de alta intensidad) (módulo 1), la organización del cloroplasto y de los plastos (módulos 1, 32 y 33), la respuesta al frío (módulos 1 y 36) y el desarrollo de semillas y embriones (módulos 32, 33 y 36). La siguiente categoría funcional en la que encontramos numerosos lncRNAs está relacionada con la regulación del RNA y su transcripción. Esta categoría comprende cuatro módulos ([Fig. S9](#)) con funciones como el proceso metabólico del mRNA (módulo 4), el procesamiento del RNA (módulo 3) y la regulación de la expresión génica (módulos 16, 17). Los genes de estas categorías funcionales se expresan en mayor medida en los embriones, las muestras y los callos de las plantas. El perfil de expresión en esta categoría funcional es muy similar al de la categoría funcional de la división celular (5 módulos) ([Fig. S12](#)), que tiene valores de expresión positivos en embriones, semillas, sam y callo de la planta. El perfil de expresión en esta categoría funcional es muy similar a la categoría de función de la división celular (5 módulos) ([Fig. S12](#)), que tiene valores de expresión positivos en embriones, semillas, SAM y raíces. El grupo de módulos con menos lncRNAs está enriquecido en genes que participan en la respuesta a condiciones abióticas ([Fig. S16](#)). Sin embargo, muchos módulos (módulos 1, 2, 5, 6, 14, 19, 23, 36 y 45) están enriquecidos en genes que participan en otras respuestas al estrés (como la sequía y el frío). Aún así, se clasificaron en otros grupos funcionales, como el desarrollo de la raíz ([Fig. S10](#)).

Por ejemplo, en el Módulo 7, identificamos 15 lncRNAs altamente expresados en la raíz, la punta de la raíz, el callo de la planta y las plántulas (que incluyen los órganos de la raíz) y parecen estar regulados al alza en respuesta a condiciones de fosfato limitado ([Fig. 7](#)). Entre los genes de este módulo encontramos ERF71, un activador transcripcional implicado en el desarrollo de la raíz (Lee et al., 2015); FRO2, que está implicado en el crecimiento de la raíz en respuesta a la falta de hierro (Satbhai et al., 2017); NRT21, un represor de la iniciación de la raíz lateral en respuesta a condiciones de bajo nitrato o alta sacarosa (Little et al., 2005); MYB93, un factor de transcripción que actúa como regulador negativo de la formación de raíces laterales (Gibbs et al., 2014), las proteínas Aux/IAA, que funcionan como represores de los genes de respuesta temprana a la auxina (Liscum & Reed, 2002); MiZ1, que desempeña un papel en el desarrollo de las raíces laterales manteniendo los niveles de auxina y regula negativamente la sensibilidad a las citoquininas (Moriwaki et al., 2011), entre otros ([Dataset S5](#)). De hecho, este módulo está altamente enriquecido en procesos de transporte transmembrana y de desarrollo del sistema radicular.

Varios lncRNAs caracterizados funcionalmente pertenecen a categorías funcionales específicas. Por ejemplo, los lncRNAs *DRIR*, *At4* y *APOLO* se encuentran en el grupo de módulos relacionados con la función de la raíz y la respuesta al estrés. Se sabe que *DRIR* regula el cierre de los estomas en la sequía (Qin et al., 2017a), *At4* está asociado con la respuesta a la deficiencia de fosfato (Sánchez-Calderón et al., 2006), y *APOLO* es un lncRNA regulador que controla directamente su gen vecino *PID* y un gran número de genes independientes por asociación de DNA en respuesta a la auxina (Ariel et al., 2014b, 2020). Las funciones de estos lncRNAs encajan con lo que hemos observado en el enriquecimiento funcional de los módulos donde se encuentran. Además de estos ejemplos, tenemos algunos

otros en el grupo de los cloroplastos y la fotosíntesis, como *FLORE* (Henriques et al., 2017b). Este lncRNA ha sido identificado como un factor importante en el fotoperiodo. El lncRNA *FLINC*, identificado como mediador de la floración en respuesta a la temperatura (Severing et al., 2018), se encuentra en el grupo de funciones de regulación y transcripción del RNA ([Dataset S5](#)).

2.3. Discusiones

En este trabajo, generamos una nueva y mejorada anotación de lncRNAs en *A. thaliana*, con el empleo de 224 transcriptomas ([Dataset S3](#)) obtenidos de 24 órganos (partes de la planta), 11 condiciones y 5 etapas de desarrollo (20 puntos de tiempo) ([Dataset S3](#)). Encontramos 6764 genes lncRNAs (7070 transcripciones), incluyendo 3772 nuevos lncRNAs ([Fig 5b](#)). Entre lncRNAs anotados, identificamos 58 genes (86 transcritos) de lncRNAs validados experimentalmente en *A. thaliana* a partir de la base de datos EvlncRNAs (B. Zhou et al., 2021), lo que apoya nuestra capacidad para identificar lncRNAs funcionalmente relevantes aprovechando los transcriptomas existentes disponibles públicamente.

2.3.1 Aumento en el número de lncRNAs anotados

Comparando los lncRNAs anotados por el método estricto contra las bases de datos encontramos pocos lncRNAs compartidos, identificando 398 lncRNAs genes compartidos con GreenNC (Di Marsico et al., 2022; Paytuví Gallart et al., 2016), 1485 lncRNAs genes compartidos con CANTATadb (Szcześniak et al., 2019), y alrededor del 74.09 % de los lncRNAs en Araport11 se encuentran en nuestra anotación (C.-Y. Cheng et al., 2017). De manera importante, nuestro enfoque de curación nos ayudó a identificar varios lncRNAs que fueron erróneamente anotados como genes codificantes, incluyendo el lncRNA IPS1, un lncRNA validado experimentalmente con múltiples sitios objetivo para miR399, inducido en ausencia de fosfato (Franco-Zorrilla et al., 2007b). Otro ejemplo es su parálogo *At4*, que presenta redundancia funcional con *IPS1* (Shin et al., 2006a). Aunque estos dos lncRNAs están funcionalmente conservados en tomate (*Lycopersicon esculentum* L.) (lncRNA *TPSI1*) (C. Liu et al., 1997), *Medicago truncatula* (*Mt4*) (Burleigh & Harrison, 1999), arroz (*OsIPS1*) (Ajmera et al., 2018) y cebada (*HvIPS1*) (C. Y. Huang et al., 2011). Del mismo modo, el bien caracterizado lncRNA *APOLO*, que regula el desarrollo de las raíces laterales (Ariel et al., 2014a), está anotado como un gen codificante de proteínas. Los lncRNAs caracterizados experimentalmente *HID1* (Y. Wang et al., 2014a), *MARS* (Roulé, Christ, et al., 2022) y *DRIR* (Qin et al., 2017a) también fueron clasificados erróneamente ([Tabla 5](#)). Dado el amplio uso de la base de datos Araport11, recomendamos una revisión de sus anotaciones a partir de nuestros resultados.

2.3.2 lncRNAs con expresión órgano y etapa del desarrollo específica

Los cientos de conjuntos de datos transcriptómicos que utilizamos nos permitieron analizar la abundancia de lncRNAs en los diferentes órganos y en las etapas de desarrollo. Nuestro análisis reveló que los órganos con más diversidad de tipos celulares muestran el mayor número de lncRNAs en *A. thaliana* ([Fig. 6a](#)). Este patrón es particularmente prominente en los órganos relacionados con la reproducción (flor y silicua), de forma similar a informes anteriores en múltiples animales (Brown et al., 2014; Cabili et al., 2011; Golicz, Bhalla, et al.,

2018; Necsulea et al., 2014) y especies de plantas (Golicz, Singh, et al., 2018; Li et al., 2014; Y.-C. Zhang et al., 2014).

Encontramos que la profundidad y el número de los transcriptomas son los factores experimentales que más afectan a nuestra capacidad para identificar nuevos lncRNAs en cualquier muestra, de forma similar a los esfuerzos de anotación previos en varias especies (Golicz, Singh, et al., 2018; Hon et al., 2017; Iyer et al., 2015). Por lo tanto, recomendamos tener una mayor profundidad de secuenciación para acelerar el descubrimiento de lncRNAs. Una de las limitaciones de nuestro estudio es la falta de datos de las etapas en las que se produce la diferenciación de los órganos en la planta, incluyendo la formación de la flor y las etapas embrionarias y la formación de los gametos-el estudio de estas condiciones biológicas sería esencial para ayudar a completar el catálogo de lncRNAs de *A. thaliana* y avanzar en nuestra comprensión del papel de los lncRNAs en la formación de las estructuras de la planta.

2.3.3 Similitudes entre los organismos

En los animales, la formación y diferenciación de órganos se produce principalmente en la fase embrionaria, mientras que en las plantas, no sólo ocurre en la fase embrionaria, sino también en la germinación y el desarrollo de las flores. Se ha demostrado que los lncRNAs ampliamente expresados y conservados se expresan durante el desarrollo del órgano, que tiene la mayor probabilidad de ser funcional. A medida que el órgano madura, un número creciente de lncRNAs específicos de la especie y del órgano tienen más probabilidades de ser no funcionales (Sarropoulos et al., 2019).

Encontramos que la expresión de los lncRNAs es significativamente más específica que la expresión de los genes codificantes. Casi el 62 % (4188) de los lncRNAs tienen perfiles de expresión restringidos a un órgano o estadio específico, mientras que sólo el 45.6 % (12,638) de las proteínas son específicas de un órgano o estadio concreto. Este hallazgo coincide con informes anteriores en *A. thaliana* (J. Liu et al., 2012) y otras especies (Cabili et al., 2011; Derrien et al., 2012; Guttman et al., 2010; Ma et al., 2013). Además, la mayoría de los sense-exonic, NATs y lincRNAs mostraron una alta especificidad tisular, mientras que los lncRNAs intrónicos tuvieron la menor especificidad tisular, en general muy similar a la especificidad tisular de los genes codificantes de proteínas.

2.3.4 Los lncRNAs se encuentran implicados en procesos biológicos

Encontramos 1241 lncRNAs co-expresados, que pudimos asociar con nuestras amplias categorías funcionales (Fig. 7, S8-S16). Usando este enfoque, encontramos categorías funcionales que involucran a los lncRNAs similares a las reportadas previamente tanto en *A. thaliana* como en otras especies de plantas. Por ejemplo, encontramos 70 lncRNAs de *A. thaliana* distribuidos en los módulos 5, 21, 18 y 39, todos ellos funcionalmente enriquecidos en genes codificantes asociados a la sequía. Numerosos lncRNAs están involucrados en esta respuesta en las plantas (Gelaw & Sanan-Mishra, 2021), incluyendo 664 lncRNAs en el maíz (W. Zhang et al., 2014), 51 en la yuca, 1096 y 126 en una variedad de *Brassica napus* resistente a la sequía (Tan et al., 2020). De manera similar, identificamos cinco módulos con 72 lncRNAs relacionados con la respuesta a patógenos (Fig. S14). Se ha encontrado previamente que los lncRNAs se expresan diferencialmente en respuesta a la infección en el tomate (Cui et al., 2017) y el maíz (Han et al., 2019). Sin embargo, las funciones que podemos

asignar a los lncRNAs están limitadas por nuestro conjunto de transcriptomas; sólo podemos identificar funciones biológicas enriquecidas en los órganos y condiciones disponibles en nuestro panel. Este análisis podría mejorarse incluyendo más transcriptomas en el futuro.

Trabajos anteriores ya han establecido la relación entre los lncRNAs y los procesos relacionados con la fotosíntesis en *A. thaliana* y el arroz (Shi et al., 2020), así como en la respuesta a diferentes tipos de luz (Kohnen et al., 2016; Z. Sun et al., 2020a). La fotosíntesis es posiblemente la vía biológica más importante en las plantas. Nuestros resultados muestran que la función con el mayor número de lncRNAs está relacionada con la organización del cloroplasto y la respuesta a la luz ([Fig. S8](#)); esto indica que un gran número de lncRNAs pueden estar involucrados en estos procesos. Cabe destacar que la mayoría de nuestros datos se obtuvieron de órganos fotosintéticos y plántulas, lo que puede explicar por qué nuestros módulos más grandes, que comprenden la mayoría de los lncRNAs, están asociados a procesos fotosintéticos.

También identificamos lncRNAs co-expresados con genes implicados en el desarrollo de la raíz y la respuesta de la raíz a múltiples estímulos ambientales ([Fig. 8](#)). Se ha demostrado previamente que los lncRNAs participan en la diferenciación de la raíz y en la respuesta a diferentes condiciones de estrés en *A. thaliana* (Ariel et al., 2014a, 2020; Bazin & Bailey-Serres, 2015; Ben Amor et al., 2009; Stauffer & Maizel, 2014), *Medicago truncatula*, donde 5561 lncRNAs cambian su expresión en la raíz debido al estrés osmótico (T.-Z. Wang et al., 2015), y en *Populus*, donde 295 lncRNAs cambian su expresión durante el desarrollo de la raíz ("lncRNA/circRNA-miRNA-mRNA Networks Regulate the Development of Root and Shoot Meristems of *Populus*," 2019).

Sorprendentemente, varios lncRNAs previamente caracterizados, incluyendo *ELENA1*, *MARS*, *COOLAIR*, *IPS1* y *HID1*, no están asociados con ningún módulo en particular. Esto podría deberse en parte a que algunos de estos lncRNAs realizan su función reguladora en condiciones ambientales específicas (por ejemplo, frío prolongado en el caso de *COOLAIR*), poco representadas en nuestro panel transcriptómico (Iyer et al., 2015). Además, la asociación funcional a través de la co-expresión no es un método a prueba de fallos; sólo identifica lncRNAs expresados en la mayoría de los órganos muestreados y que tienen una fuerte asociación de expresión con genes con funciones similares. Por lo tanto, muchos otros nuevos lncRNAs reportados en nuestra anotación sin asociación funcional pueden tener funciones importantes que este enfoque no pudo identificar.

Esperamos que esta anotación de lncRNAs altamente curada e informada por la transcriptómica con asociaciones funcionales a través de la co-expresión en *A. thaliana* se convierta en un recurso valioso para la comunidad de *A. thaliana* y de lncRNAs de plantas. En el futuro, queremos evaluar si las relaciones de asociación funcional entre lncRNAs y otros RNAs se conservan en diferentes especies y cómo su pérdida o ganancia podría estar asociada con la pérdida o ganancia de rasgos particulares en esta y otras familias de plantas.

2.4. Conclusiones

En este estudio se identificaron 6764 genes de lncRNAs en *A. thaliana*, los cuales se clasificaron en intrónicos, lincRNAs, exonico-sentido y NATs, siendo la gran proporción lincRNAs.

Solo el 21 % (1425 genes) presentaba co-expresión con otros genes en los módulos funcionales. Encontrando diferentes distribuciones en el contenido de lncRNAs en cada módulo, algunas de las funciones relacionadas con estos lncRNAs, se encuentran módulos de genes que regulan fotosíntesis, respuesta celular al daño del DNA, desarrollo de raíces, división celular, síntesis de metabolitos, entre otros.

Relacionamos al 62 % (4188) de los lncRNAs con perfiles de expresión restringidos a un órgano o estadio específico. Los órganos con un mayor número de lncRNAs únicos y expresados fueron los órganos maduros, relacionado con la reproducción y fotosíntesis. Siendo un punto importante para la exploración y búsqueda de funciones conservadas en otras especies, ya que podemos considerarlos los mejores candidatos para el estudio y caracterización funcional de lncRNAs.

Capítulo 3

Conservación de lincRNAs por sintenia y microhomología

El artículo referente a los resultados obtenidos en este capítulo se encuentra en proceso de escritura, teniendo como título tentativo **“Syntenic and microhomology conservation of plant lincRNAs”** (Coss-Navarrete, Evelia Lorena, Corona-Gomez José Antonio, Gillmor Stewart, Hurst Laurence, Oktaba Katarzyna, y Fernandez-Valverde Selene L.).

3.1 Métodos

3.1.1 Extracción de RNA y secuenciación

Se realizó la germinación de las plántulas para las especies *Capsella rubella* (CS22697), *Thellungiella parvula* (CS22663), *Brassica rapa* (CS29001) y *Brassica oleracea* (CS29002) en el medio MS, bajo condiciones de días largos (21 °C, 16 / 8h). Se extrajo el RNA total de la parte aérea y las raíces de cada plántula, de acuerdo a la [Sección 2.1.2](#). Las bibliotecas se secuenciaron en Hi-Seq X de Illumina (PE150). En el caso de *A. thaliana* (Col-0, CS70000), se emplearon los transcriptomas generados en la [Sección 2.1.2](#) (Corona-Gomez et al., n.d.) ([Tabla S1](#)).

3.1.2 Identificación y clasificación de lncRNAs

Se realizó el alineamiento y ensamblaje guiado con el genoma de referencia de cada especie empleando STAR v2.7.2.b (Dobin et al., 2013) y StringTie v.3.4 (M. Pertea et al., 2015), de acuerdo con la [sección 2.1.3](#) ([Tabla S3](#) y [Tabla S4](#)). En el caso de *T. parvula*, se generó un archivo de anotación empleando el programa LiftOver (de UCSC, kentUtils), la versión 1 del genoma y el archivo de anotación de la versión 2 del genoma ([Table S5](#)) (Script LiftOver_complete.sh empleando como referencia el pipeline propuesto en <https://iamphioxus.org/tag/liftover/>).

Para la identificación de lncRNAs en cada especie se empleó el pipeline de la [Sección 2.1.4](#) (Corona-Gomez et al., n.d.). Posteriormente, los lncRNAs fueron clasificados de acuerdo con su localización genómica y con base a la cercanía con sus genes codificantes vecinos en: lncRNAs, intrónicos, exónico sentido y NATs, como se describe en la [Sección 2.1.5](#). Los archivos obtenidos se convirtieron a formato BED mediante los programas gtfToGenePred y genePredToBed (de KentUtils, UCSC) (Kent et al., 2002).

3.1.3 Búsqueda de proteínas ortólogas

Para identificar las proteínas ortólogas se empleó OrthoFinder [-t 8 -a 8 -S blast -l 1.7 -M msa -A mafft -T fasttree] (Emms & Kelly, 2019), se descargaron las secuencias de aminoácidos únicas (primaria) de las bases de datos de Ensembl Plants (Bolser et al., 2016), Phytozome v13 (Goodstein et al., 2012) y (T. Zhao & Schranz, 2019) ([Table S6](#)) para las 11 especies analizadas de la familia Brassicaceae. En el caso de las secuencias descargadas en Phytozome V13, se empleó el script primary_transcript.py de OrthoFinder (Emms & Kelly, 2019) para obtener la secuencia primaria de cada gen.

3.1.4 Conservación por sintenia y microhomología de lincRNAs

Para analizar la conservación por sintenia y microhomología de los lincRNAs identificados en *A. thaliana*, *C. rubella*, *T. parvula*, *B. rapa* y *B. oleracea* provenientes de la parte aérea y las raíces se propuso la siguiente estrategia. La estrategia empleada se dividió en 3 pasos: (1) Localización por posición de los lincRNAs y genes codificantes de las especies de referencia (*A. thaliana*, *C. rubella*, *T. parvula*, *B. rapa* y *B. oleracea*) con respecto a las demás especies de la familia Brassicaceae; (2A) Identificación de los lincRNAs conservados por sintenia; (2B) Búsqueda de los motivos conservados en los lincRNAs (microhomología) y (3) Identificación de lincRNAs conservados por sintenia y microhomología.

Primero, localizamos las coordenadas de los lincRNAs y genes codificantes de cada una de las especies de referencia (*A. thaliana*, *C. rubella*, *T. parvula*, *B. rapa* y *B. oleracea*) empleando el alineamiento genómico (generado previamente en el laboratorio, (Corona-Gomez et al., 2020)) y el programa *halLiftover* (de *haltools*) (Hickey et al., 2013). Para identificar los genes codificantes ortólogos y con posición conservada a través de las especies, se empleó la metodología señalada en *01_alignement_HallLifover.md* partir de los resultados obtenidos con *OrthoFinder* y *halLifover*.

A continuación, para la identificación de los lincRNAs conservados por sintenia se seleccionó una de las especies de referencia y se extrajeron los genes codificantes localizados a 10 kb (upstream y downstream) de cada uno de los lincRNAs identificados en dicha especie. Después, comparamos los genes codificantes vecinos de cada lincRNA con los resultados obtenidos en el paso previo, con el fin de localizar todos los genes codificantes conservados por ortología y posición en las otras especies. Este paso se repitió de manera independiente para cada una de las especies de referencia empleando la metodología de *02_protein_orthology_OrthoFinder.md* y el programa *BEDtools* (Quinlan & Hall, 2010).

Simultáneamente, se realizó el análisis de conservación por microhomología de los lincRNAs, para lo cual se extrajeron las secuencias FASTA de todos los lincRNAs conservados por posición (Paso 1) empleando la metodología de *03_synteny_analysis.md* a partir del genoma y el resultado de *halLiftover*. Se descartaron las secuencias menores a 10 nt y lincRNAs detectados como especie específica. Posteriormente, se realizó la búsqueda de motivos conservados empleando el programa *MEME suite* [1 a 5 motivos por secuencia, tamaño de 8 a 50 nt, pvalue ≤ 0.05] (Bailey et al., 2015).

Finalmente, se seleccionaron los lincRNAs conservados por sintenia y microhomología a través de las especies con los resultados obtenidos y empleando la metodología de *04_MEME_suite.md*. En el cual:

- Se identificaron el número de motivos presentes en cada especie por cada lincRNA (E value -3), por lo menos conservado en dos especies.
- Se obtuvieron las matrices binarias empleando el paquete *memes* (v 1.6.0, en R) (Nystrom & McKay, 2021), [rJava](#) (v1.0-6, en R).

Todos los scripts se encuentran en el Github: https://github.com/EveliaCoss/Anexos_tesis.git

3.1.5 Proporción de lincRNAs conservados a través de la filogenia

Para estimar la proporción de lincRNAs encontrados en cada nodo de la filogenia se generaron matrices binarias a partir de los resultados obtenidos de la conservación por sintenia y microhomología, posteriormente empleamos la metodología propuesta por (Vakirlis et al., 2020). Los tiempos de divergencia fueron tomados de [TimeTree.org](https://www.timetree.org/) (Hedges et al., 2006; Kumar et al., 2017) (estimated times).

3.1.6 Porcentaje de identidad de secuencia en los lincRNAs

Para analizar el porcentaje de identidad en la secuencia de los lincRNAs a través de las especies, primero se extrajeron las coordenadas de los lincRNAs de cada especie de referencia con respecto a las 11 especies analizadas de la familia Brassicaceae. Para lo cual se empleó el programa intersectBed (de BEDtools) (Quinlan & Hall, 2010), el archivo de coordenadas de los lincRNAs identificados en cada especie (formato BED) ([Sección 3.1.2](#)) y los resultados de halLiftover (previamente obtenidos en la [Sección 3.1.4](#)). En el caso de las especies en las cuales no contábamos con el archivo de coordenadas de los lincRNAs, se optó por emplear su archivo de anotación, con el fin de localizar la ubicación de los lincRNAs con respecto a los transcritos anotados en la especie. Posteriormente, se extrajeron las secuencias en formato FASTA empleando el programa faSomeRecords (de KentUtils, UCSC) (Kent et al., 2002) y se analizó el porcentaje de identidad de las secuencias mediante Clustal Omega (v1.2.4) [-t DNA --full --percent-id --distmat-out --output-order=tree-order --outfmt=clu] (Sievers & Higgins, 2018).

3.1.7 Identificación de lincRNAs con expresión conservada

Se realizó la cuantificación de la expresión en cada transcriptoma mediante el pseudo alineamiento empleando Kallisto (v0.44.0) (Bray et al., 2016) para cada una de las especies. Posteriormente, se importaron los datos a R (v4.1.2) mediante el paquete tximport (v1.20) (Soneson et al., 2015b), obteniendo los valores de expresión en cuentas (counts) y Transcript per million (TPM). Para el análisis de expresión diferencial, se emplearon las cuentas obtenidas y se compararon ambos órganos (shoot vs root) en cada especie empleando el paquete DESeq2 (v1.32) (Love et al., 2014b). Adicionalmente, se analizaron los niveles de expresión de los lincRNAs, clasificándolos de acuerdo a su nivel de conservación en: 1) conservados por sintenia y microhomología, 2) conservados por sintenia (sin motivos conservados) y 3) especie específica; y de acuerdo con sus patrones de expresión en: A) parte aérea, B) raíces o C) con expresión global en la plántula (expresión en ambos órganos). Las cuentas fueron normalizadas empleando la función rlog (regularized log) del paquete DESeq2 (Love et al., 2014b).

3.1.8 Ganancia y pérdida de lincRNAs

Para entender la dinámica evolutiva de los lincRNAs dentro de la familia Brassicaceae, se empleó el programa DOLLOP (de PHYLIP v3.698) (Feisenstein, 1989), con el fin de localizar la ganancia y la pérdida de cada uno de los lincRNAs identificados en las 5 especies (*A. thaliana*, *C. rubella*, *T. parvula*, *B. rapa* y *B. oleracea*).

3.1.9 Visualización de los datos y scripts

Los archivos de coordenadas (formato BED) de los genes de lncRNAs identificados y clasificados se encuentran disponibles en el [Dataset S1](#) (*A. thaliana*), [S6](#) (*C. rubella*), [S7](#) (*T. parvula*), [S8](#) (*B. rapa*) y [S9](#) (*B. oleracea*).

Los scripts empleados en el análisis se encuentran en el repositorio de Github:

https://github.com/EveliaCoss/Anexos_tesis.git

3.2 Resultados

3.2.1 Características conservadas de los lincRNAs

Para realizar una comparación directa de los lincRNAs a través de las especies, primero se identificaron todos los lncRNAs presentes en los transcriptomas generados para cada una de las especies de referencia (*A. thaliana*, *C. rubella*, *T. parvula*, *B. rapa* y *B. oleracea*) empleando el Método Estricto (ME, o Strict Method [SM]) (Corona-Gomez et al., n.d.), basado en métodos previamente reportados (Cabili et al., 2011; Gaiti et al., 2015). Todos los transcriptomas provienen de transcritos poliadenilados, *paired-end* y cadena específica (strand specific). Cuentan con un tamaño de 150 pares de bases de longitud (150 PE) ([Tabla S1](#)). Se identificaron 3579 genes de lncRNAs (3805 transcritos) en *A. thaliana*, 451 (466 transcritos) *C. rubella*, 1096 (1120 transcritos) *T. parvula*, 2546 (2628 transcritos) *B. rapa* y 2460 (2559 transcritos) *B. oleracea* provenientes de la parte aérea y raíces (plántulas de ~ 8 días después de la germinación) ([Fig 9A](#)). Los lncRNAs fueron clasificados de acuerdo con su localización genómica, asignando una gran proporción de estos como lincRNAs (~57 %), excepto en el caso de *B. oleracea* con un 35 %. Se identificaron 2490 genes de lincRNAs (2634 transcritos) de *A. thaliana*, 420 (436 transcritos) *C. rubella*, 644 (661 transcritos) *T. parvula*, 1840 (1886 transcritos) *B. rapa* y 926 (926 transcritos) *B. oleracea* ([Fig 9A](#)). Aproximadamente más del 70 % de los lincRNAs identificados eran monoexónicos en las especies analizadas, mientras que los genes codificantes contenían un mayor número de exones en sus transcritos ([Fig 9C](#) y [Fig S17A](#)). En el caso de la especie *B. oleracea*, la gran mayoría de los lincRNAs identificados eran multiexónicos y solo el 30 % eran monoexónicos ([Fig 9C](#) y [Fig S17A](#)). Cerca del 50 % de los lincRNAs presentaban un tamaño de transcrito en un rango entre 400 nt a 500 nt. Además, se identificaron lincRNAs con transcritos mayores a 1000 nt, teniendo una mayor prevalencia (~ 25 %) en los lincRNAs identificados en *C. rubella* y *B. oleracea* ([Fig 9D](#) y [Fig S17B](#)). En términos globales de expresión, los lincRNAs tienen una menor expresión a comparación de los genes codificantes, exceptuando el caso de la especie *B. oleracea* ([Fig 9B](#)).

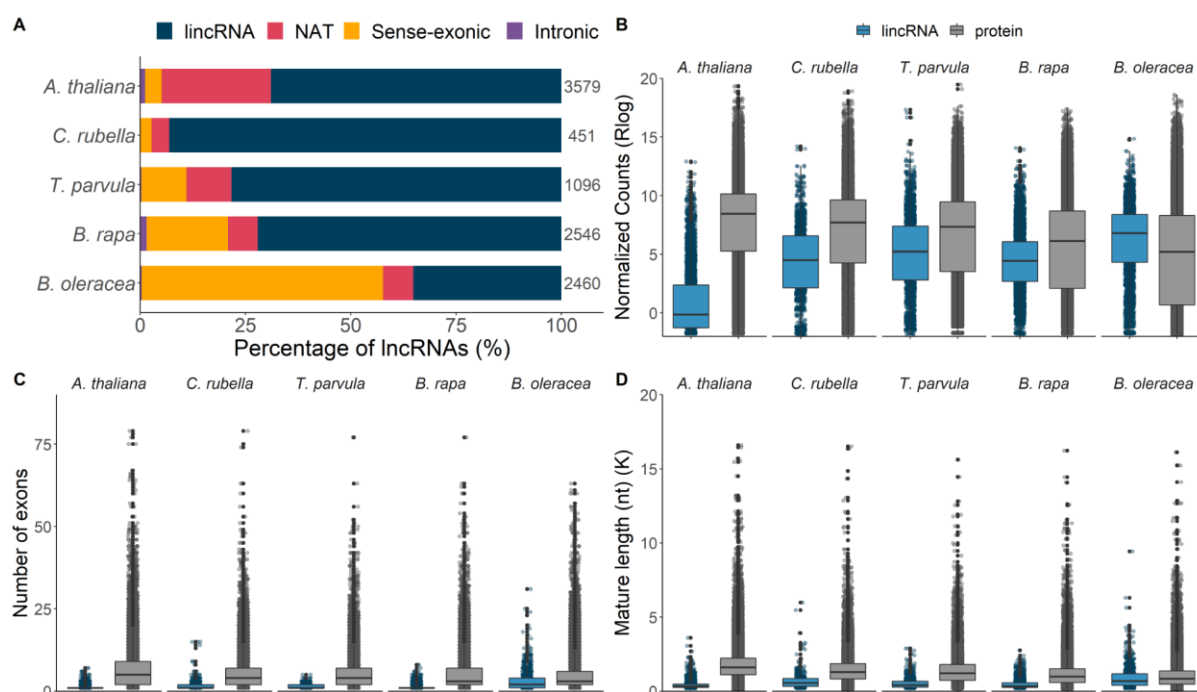


Figura 9. Características de los lincRNAs identificados en las diferentes especies. (A) Clasificación de los lincRNAs identificados de acuerdo a su contexto genómico. (B) Patrones de expresión, (C) número de exones y (D) tamaño del transcrito maduro de los lincRNAs y genes codificantes en las diferentes especies. Se analizaron 2490 genes de lincRNAs (2634 transcritos) y 27055 genes codificantes (47643 transcritos) en *A. thaliana*, 420 lincRNAs (436 transcritos) y 27641 genes codificantes (27641 transcritos) en *C. rubella*, 644 lincRNAs (661 transcritos) y 25032 genes codificantes (25032 transcritos) en *T. parvula*, 1840 lincRNAs (1886 transcritos) y 39607 genes codificantes (39613 transcritos) en *B. rapa*, 926 lincRNAs (926 transcritos) y 58163 genes codificantes (58163 transcritos) *B. oleracea*. Se eliminaron los IDs de los genes codificantes que coincidían con lincRNAs anotados en cada especie. En todos los casos existieron diferencias significativas entre los genes codificantes y los lincRNAs con un p value menor a $2.2e-16$ bajo la prueba estadística Wilcoxon test. *A. thaliana* = *Arabidopsis thaliana*, *C. rubella* = *Capsella rubella*, *T. parvula* = *Thellungiella paravula*, *B. rapa* = *Brassica rapa* y *B. oleracea* = *Brassica oleracea*.

3.2.2 Identificación de los lincRNAs ortólogos en la familia Brassicaceae

Para la búsqueda de genes ortólogos de los lincRNAs dentro de la familia Brassicaceae, se seleccionaron solo los lincRNAs de cada especie, con el fin de analizar la conservación propia de los lincRNAs, sin que la secuencia conservada de los genes codificantes influya en su conservación. Al encontrar cerca del 90 % de los genes codificantes compartidos entre las especies (Fernandez-Pozo et al., 2021; X. Wang et al., 2011). En el caso de *A. thaliana*, se emplearon los 4354 genes de lincRNAs descritos en el capítulo previo ([Sección 2.2.1](#)), dentro de los cuales se encuentran los 2490 genes de lincRNAs identificados de la parte aérea y de las raíces. Con el fin de localizar las coordenadas genómicas de los diferentes lincRNAs y genes codificantes en las 11 especies analizadas de la familia Brassicaceae, se empleó el programa halLiftover de haltools (Hickey et al., 2013) y el alineamiento genómico (previamente generado en el laboratorio, (Corona-Gomez et al., 2020)) ([Fig 10](#), step 1). Para

la ubicación de los lincRNAs conservados por sintenia, se emplearon sólo los genes codificantes conservados por posición genómica y con ortología localizados a 10 kb upstream y downstream (Fig 10, step 2). Posteriormente, se realizó la búsqueda de pequeños motivos conservados (microhomología) con el programa MEME suite (Bailey et al., 2015) [8 a 50 nt], a partir de las secuencias obtenidas del alineamiento genómico (Fig 10, step 3).

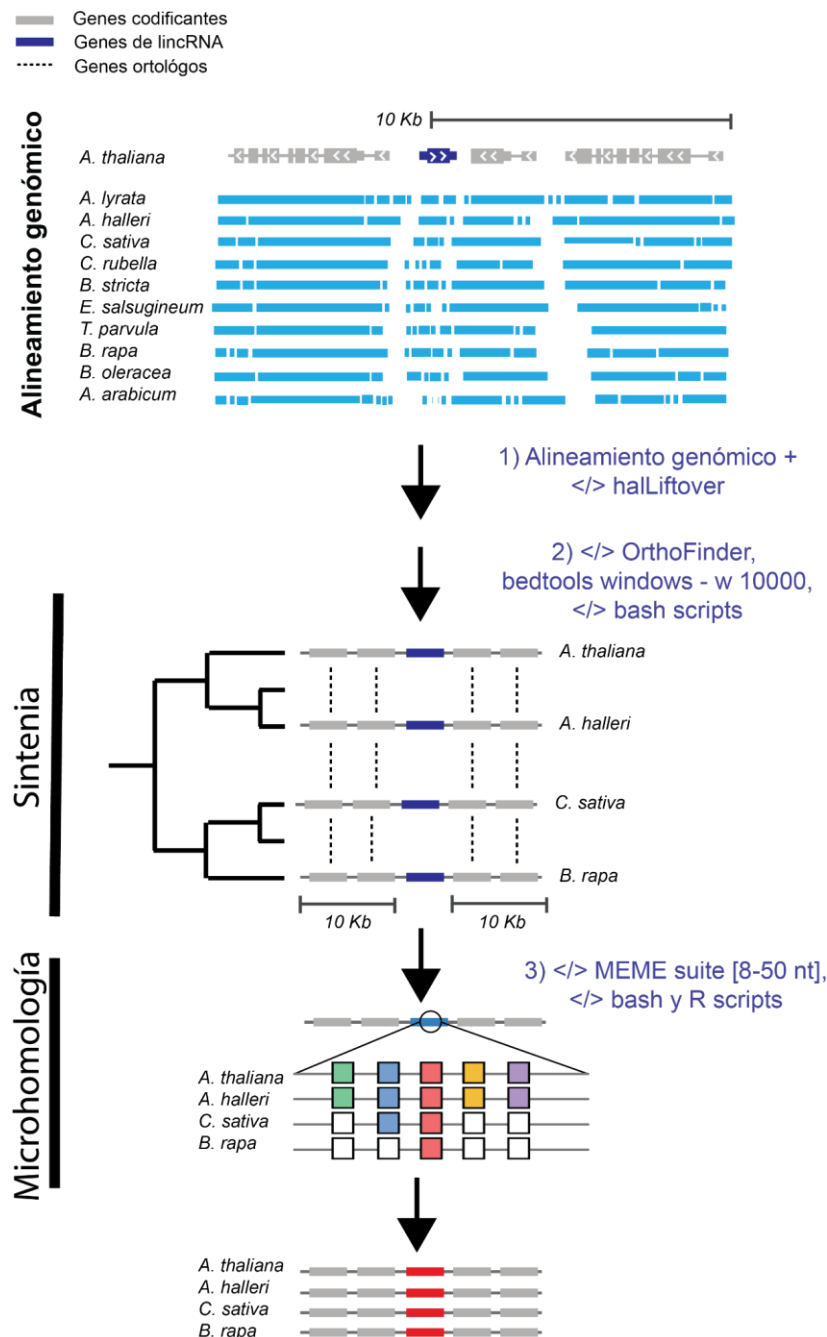


Figura 10. Pipeline del análisis de conservación por sintenia y secuencia en lincRNAs. (1) Identificar las coordenadas de los lincRNAs y genes codificantes en cada una de las especies empleando el el alineamiento genómico y el programa hallLiftover. (2) Análisis de conservación por sintenia, partiendo de la localización de los genes codificantes ortólogos a 10 kb del lincRNA en cada especie. (3) Identificación de regiones conservadas por

microhomología empleando el programa MEME suite. Las especies de referencia empleadas fueron *A. thaliana*, *C. rubella*, *T. parvula*, *B. rapa* y *B. oleracea*.

Se realizó la identificación de genes codificantes ortólogos entre las especies analizadas empleando el programa OrthoFinder (Emms & Kelly, 2019), obteniendo un total de 19,019 ortogrupos (OGs) de genes codificantes en *A. thaliana*, solo 92 ortogrupos (0.48 %) se asignaron como especie específica (Fig 11A, Tabla S6). Mientras que 18,927 OGS (99.52 %) se encontraban compartidos entre las 11 especies analizadas en la familia Brassicaceae. Encontrando más del 75 % de los ortogrupos compartidos entre las especies de referencia (*C. rubella*, *T. parvula*, *B. rapa* y *B. oleracea*).

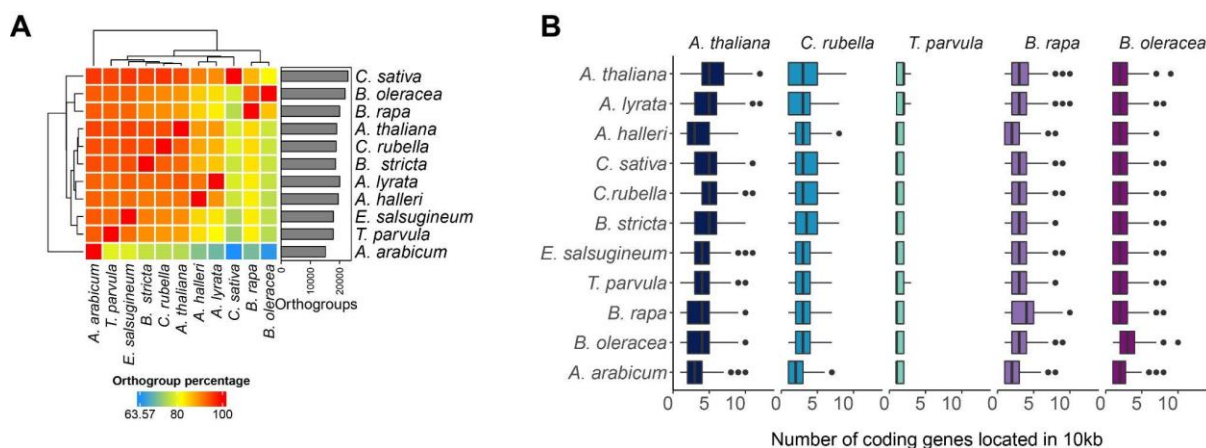


Figura 11. Conservación de genes codificantes en las 11 especies de la familia Brassicaceae. (A) Distribución porcentual de los ortogrupos compartidos entre las especies de la familia Brassicaceae. (B) Distribución de los genes codificantes localizados a 10 kb de distancia de los lincRNAs identificados en cada especie de referencia. En el eje Y tenemos las 11 especies evaluadas en la familia Brassicaceae, el orden se encuentra de acuerdo a la filogenia con respecto a *A. thaliana*. En el eje X el número de genes codificantes localizados a 10 kb con respecto a los lincRNAs identificados a partir de los transcriptomas de cada una de las especies. El rango de genes codificantes localizados fue de 1 a 12 en *A. thaliana*, 1 a 10 en *B. oleracea* y *B. rapa*, 1 a 9 en *C. rubella* y 1 a 3 en *T. parvula*.

Para la localización de los lincRNAs por medio de los bloques sinténicos se analizaron los genes codificantes a una distancia de 10 kb (upstream y downstream) de los lincRNAs. Además, se seleccionaron solo los genes codificantes que presentaran ortología y conservación por posición en el alineamiento genómico. Localizando de 1 a 12 genes codificantes ortólogos a través de la filogenia en los AtlincRNAs (promedio: 5.30, mediana: 5). Por su parte en *B. rapa* y *B. oleracea*, se localizaron de 1 a 10 genes codificantes (promedio: 3.74 y 3.36, mediana: 4 y 3, respectivamente), 1 a 9 en *C. rubella* (promedio: 3.5, mediana: 3), y finalmente, de 1 a 3 genes en *T. parvula* (promedio: 1.47, mediana: 1) (Fig 11B). Lo cual permite la localización de lincRNAs que presenten por lo menos un gen codificante ortólogo vecino dentro del bloque sinténico detectado a través de la filogenia.

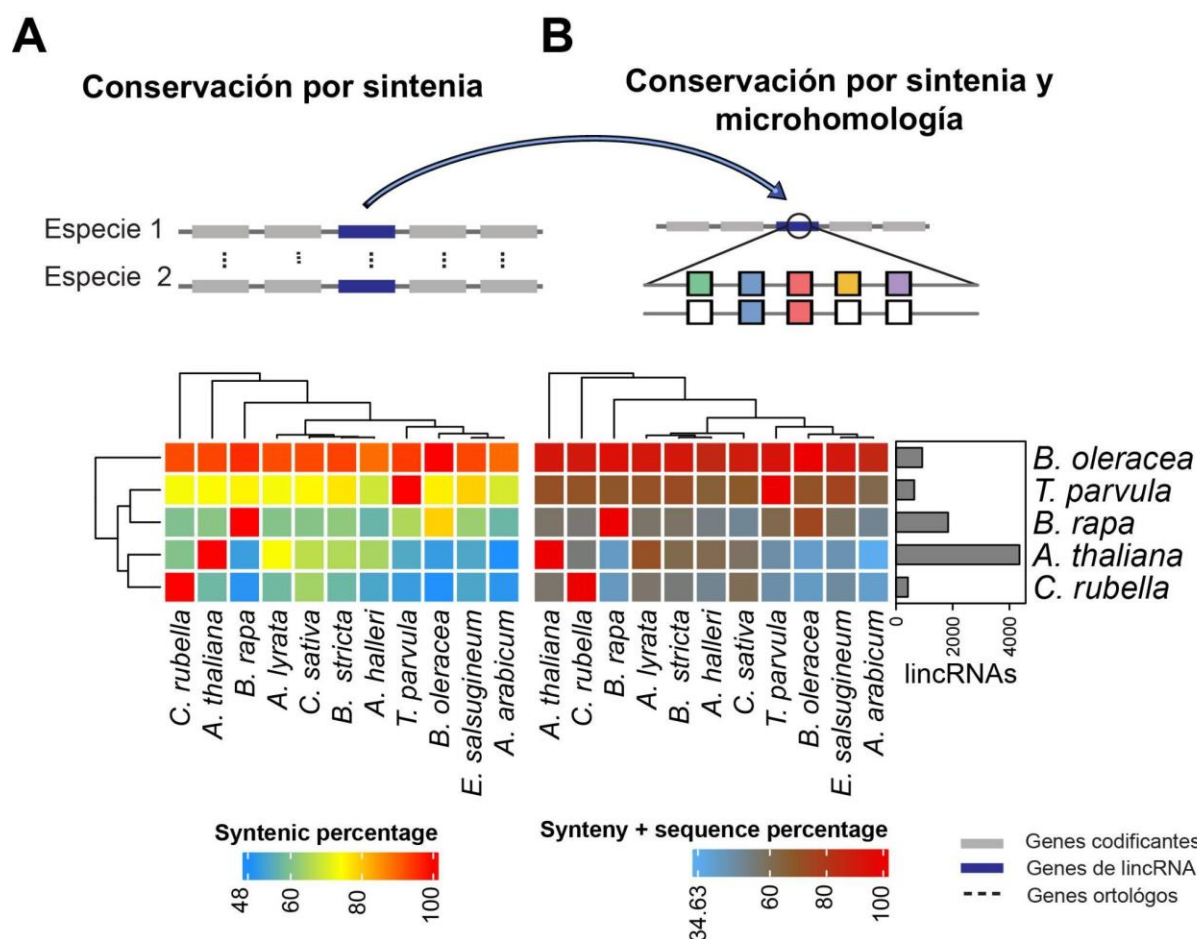


Figura 12. Porcentaje de los lincRNAs conservados por sintenia y microhomología. (A) Comparación pareada (pairwise) de la conservación por sintenia a través de la familia Brassicaceae. La escala de colores representa el porcentaje de conservación por sintenia. (B) Comparación pareada de la conservación por sintenia y microhomología para cada uno de los lincRNAs identificados en cada especie. La escala de colores representa el porcentaje de conservación por sintenia y microhomología. Las especies fueron ordenadas de acuerdo al análisis de distribución de Pearson en el paquete ComplexHeatmap.

A través de múltiples comparaciones pareadas entre las especies (pairwise), encontramos como mínimo el 48 % (2090 genes de 4354) de los lincRNAs conservados por sintenia entre *A. thaliana* y la especie basal (*A. arabicum*) (Fig 12A). Mientras que dentro de la misma familia, los lincRNAs de *A. thaliana* se encontraban conservados por sintenia en un rango del 50.92 % (2217) (con *B. oleracea*) y 73.59 % (3204 genes) (*A. lyrata*). En el caso de las otras especies, sorpresivamente los lincRNAs identificados en *B. oleracea* se encontraban ampliamente localizados por sintenia a través de las especies en un rango del 88.88 % (823 genes de 926) (*A. halleri*) y 95.57 % (885 genes) (*B. rapa*) dentro de la misma familia, y con un 89.31 % (827) hasta la especie basal. Por lo que, a pesar de la baja cantidad de lincRNAs identificados en las otras especies a comparación de los identificados en *A. thaliana*, más de la mitad de los lincRNAs se encuentran conservados por sintenia.

Posteriormente, se identificaron los motivos conservados a través de las especies en un rango de 8 a 50 nt en los lincRNAs conservados dentro de los bloques sinténicos. Los motivos más prevalentes en los lincRNA se encontraban situados en tamaños de 8, 11, 15, 21, 29, 41 y 50

nt en todas las especies, excepto en *B. oleracea* (Figura S18). El rango de conservación de los lincRNAs identificados en *A. thaliana* fue de 40.54 % (1765 genes de 4354, con respecto a *B. oleracea*) a 71.27 % (1765 genes de 4354, con respecto a *A. lyrata*). Mientras que en la especie *B. oleracea* el rango se encontró entre 85.21 % (789 de 926, con *A. halleri*) y 94.38 % (874 genes, *B. rapa*). Lo cual difiere con su especie filogenéticamente más cercana *B. rapa* con un rango de lincRNAs conservados por sintenia y secuencia de 49.46 % (910 genes de 1840, con *C. sativa*) y 74.13 % (1364 genes, *B. oleracea*) (Fig 12B). Los resultados obtenidos se ven afectados por la presencia de lincRNAs especie específica, es decir, lincRNAs solamente presentes en la especie de procedencia. Tal es el caso de la especie *C. rubella* con un 33.33 % (140 genes de 420) de los lincRNAs clasificados como especie específica. Obteniendo un rango de conservación de 41.43 % (174 genes, *B. oleracea*) a 59.76 % (251 genes, *C. sativa*) de los lincRNAs (Tabla 6).

Tabla 6. Número de lincRNAs especie específica identificados por cada especie.

Especies	Número de genes de lincRNAs	Especie específica	Solo sintenia	Sintenia + microhomología
<i>A. thaliana</i> (224 transcriptomas)	4354	856	99	3399
<i>A. thaliana</i>	2490	500	58	1932
<i>C. rubella</i>	420	133	8	279
<i>T. parvula</i>	644	87	220	337
<i>B. rapa</i>	1840	345	70	1425
<i>B. oleracea</i>	926	61	1	864

3.2.3 Relación entre el número de parálogos identificados en lincRNAs y las duplicaciones del genoma

A partir de los resultados obtenidos se localizaron los parálogos de los lincRNAs conservados por sintenia y microhomología en cada una de las especies de referencia, encontrando el 74.22 % (1342 de 1808) y el 78.35 % (1383 de 1765) de los AtlincRNAs presentan múltiples copias en *B. rapa* y *B. oleracea*, respectivamente (Fig 13A). En el caso de la especie *C. sativa* se encontró el 97.58 % (2465 genes de 2526) de los AtlincRNAs con múltiples copias (Fig 13A). De igual manera, este mismo patrón de duplicación lo encontramos en los lincRNAs identificados en las otras especies de referencia (*C. rubella*, *T. parvula*, *B. rapa* y *B. oleracea*) con más del 96 % de duplicación con relación a la especie *C. sativa*, por su parte, encontramos el 64 % de duplicación de los lincRNAs con relación a las especies *B. rapa* y *B. oleracea* (Fig 13). Relacionando el número de parálogos identificados en los lincRNAs con los eventos de duplicación del genoma ocurridos hace 124 mya (β) y hace 47 mya (α), teniendo lugar uno de los eventos de duplicación más recientes entre 7 a 12 mya, afectando en el número de copias de los genes presentes en las especies *C. sativa*, *B. rapa* y *B. oleracea* (Kagale, Robinson, et al., 2014). Siendo la duplicación uno de los escenarios evolutivos relacionados con el surgimiento de nuevos lincRNAs (Ulitsky & Bartel, 2013).

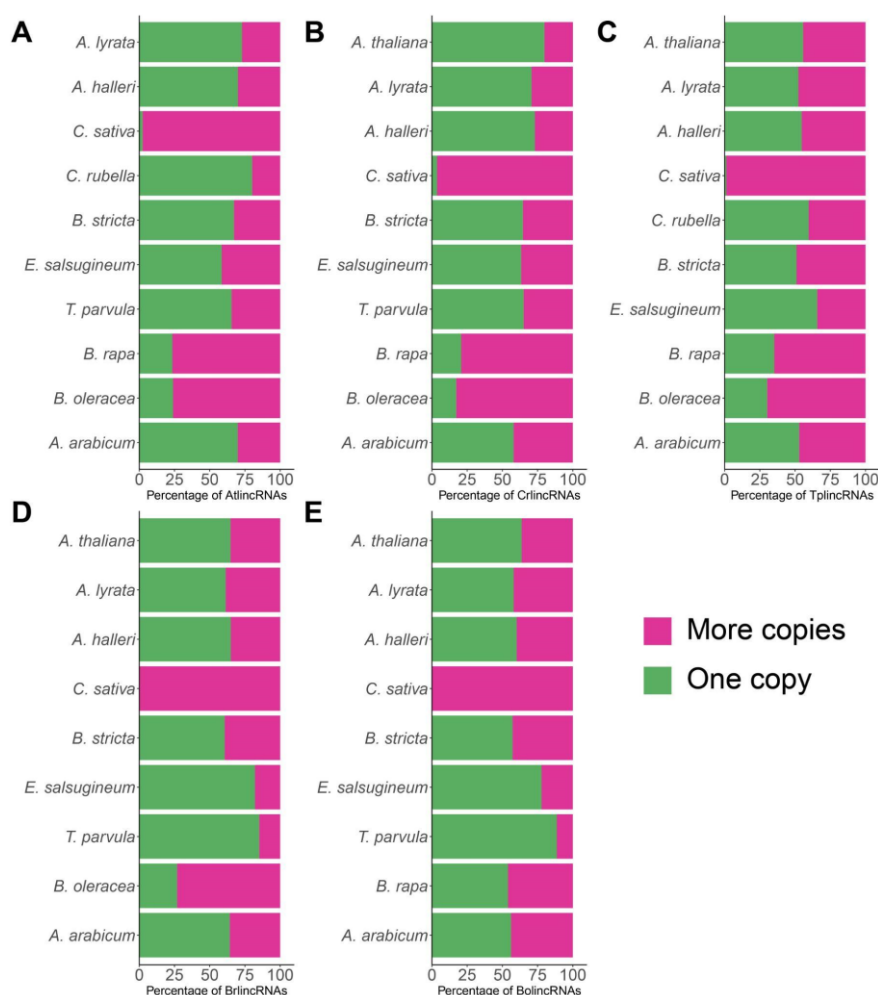


Figura 13. Parálogos identificados de los lincRNAs conservados por sintenia y microhomología. (A) Número de parálogos identificados en *A. thaliana* (B) *C. rubella*, (C) *T. parvula*, (D) *B. rapa*, y en (E) *B. oleracea*.

3.2.4 lincRNAs identificados como ancestrales

Para comparar directamente los lincRNAs de diferentes especies y obtener grupos que probablemente comparten una ascendencia en común, generamos matrices binarias a partir de los resultados obtenidos de la conservación por sintenia y microhomología, posteriormente empleamos la metodología propuesta por (Vakirlis et al., 2020). La cual ha sido empleada previamente en la localización de genes codificantes con conservación por sintenia y baja conservación de secuencia (Vakirlis et al., 2020). A través de los scripts de R proporcionados en el artículo ([Figure6 fig6-suppl2.R](#)) (Vakirlis et al., 2020), se comparan ambas matrices binarias: [1] conservación por sintenia; 2) conservación por sintenia y microhomología] con el fin de obtener la proporción de genes conservados en base a la filogenia (tiempos estimados de acuerdo a TimeTree.org). El análisis se realizó en todas las especies de referencia. En el caso de *A. thaliana*, empleamos todos los lincRNAs identificados en el capítulo previo (4354 genes de lincRNAs) y los provenientes de los transcriptomas de raíces y parte aérea (2490 lincRNAs). Identificando 48 % (2090 de 4354 genes) de AtlinRNAs de diferentes órganos con ancestría común con la especie basal *A. arabicum* ([Fig 14A](#)), mientras que los lincRNAs

identificados de raíces y parte aérea se identificó el 46.18 % (1150 de 2490 genes) con un tiempo de divergencia calculado de 37.8 millones de años (mya) (de acuerdo a TimeTree.org). Recordando que dentro de los 4354 lincRNAs identificados previamente se encuentran presentes los 2490 lincRNAs. Por lo que, enfocándonos en los lincRNAs provenientes de estos dos órganos, encontramos 79.96 % (1991 de 2490 genes) conservados en las especies más cercanas filogenéticamente a *A. thaliana* (*A. lyrata* y *A. halleri*, 5.97 mya). Dentro del linaje 1. encontramos el 65.86 % (1640 genes) (11.91 mya) conservados por sintenia y microhomología. Compartiendo entre ambos linajes el 56.83 % (1415 genes) (25.97 mya) (Fig 14B), lo cual resulta en una alta proporción de lincRNAs conservados a comparación de la reportada previamente por (Palos et al., 2022), en el cual obtuvieron un 18.1 % (3725 de 20416 genes) de los lincRNAs conservados al emplear primero la conservación por secuencia, y posteriormente la conservación por sintenia para la localización de lincRNAs en la misma familia (Palos et al., 2022).

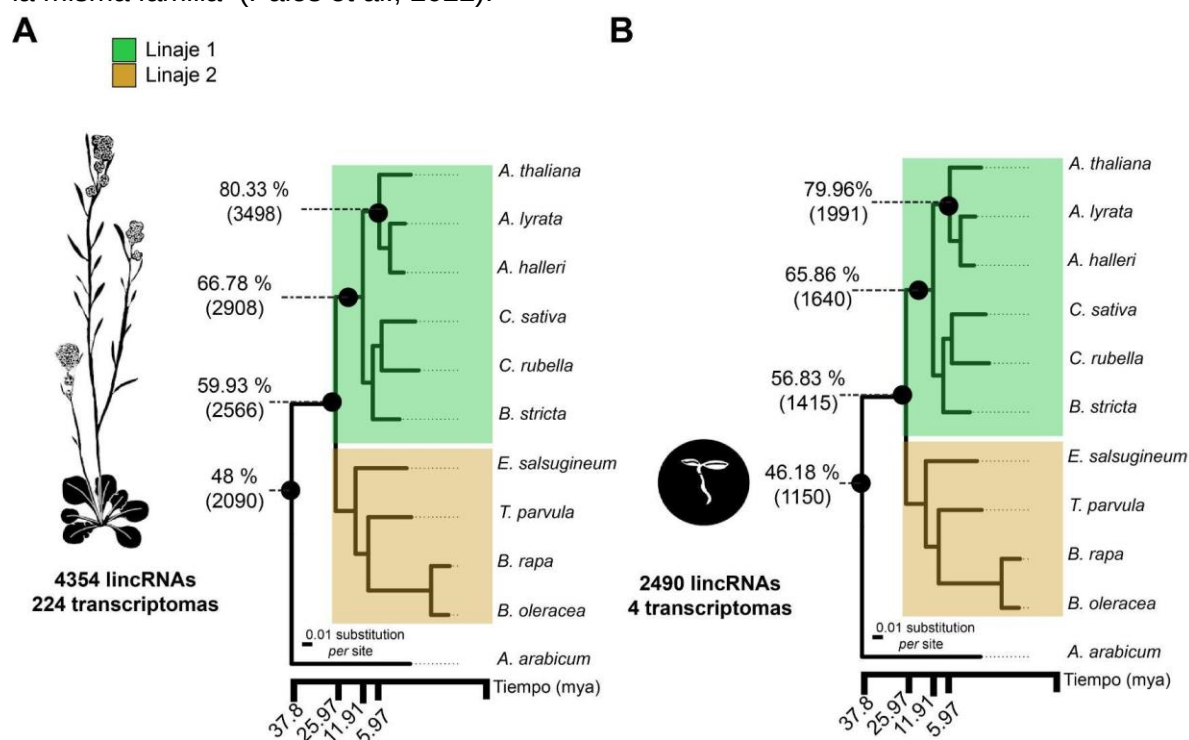


Figura 14. AtlinRNAs conservados y con ancestría en común dentro de la familia *Brassicaceae*. (A) Distribución de los AtlinRNAs (224 transcriptomas, múltiples órganos) conservados a través de la filogenia. (B) Distribución de los AtlinRNAs (4 transcriptomas, raíces y parte aérea) conservados a través de la filogenia. AtlinRNAs: lincRNAs identificados en *A. thaliana*. Tiempo de divergencia estimado de acuerdo con TimeTree.org.

Por su parte, en las especies *B. oleracea* y *B. rapa* encontramos el 93.41 % (865 de 926 genes) y el 81.25 % (1495 de 1840 genes) de los lincRNAs conservados dentro del mismo nodo con un tiempo de 2.84 mya, respectivamente (Fig 15A). Dentro del mismo linaje encontramos el 92.76 % (859 de 926 genes) y el 66.30 % (1220 genes de 1840 genes) de los lincRNAs identificados en *B. oleracea* y *B. rapa*, respectivamente (Fig 15A). En el caso de *B. oleracea* la mayoría de los lincRNAs se encontraban presentes dentro de la familia y hasta la especie basal con un 91.25 % (845 de 926 genes) y 87.68 % (812 de 926 genes), respectivamente (Fig 15A). En el caso de *B. rapa*, encontramos el 61.90 % (1139 de 1840

genes) de los lincRNAs dentro de la familia y el 56.96 % (1048 de 1840 genes) hasta la especie basal. Además de identificar un mayor número de lincRNAs en *B. rapa* a comparación con *B. oleracea*, también encontramos un mayor número de lincRNAs especie específica (*B. rapa* 18.75 %, 845 de 1840; *B. oleracea*, 6.58 %, 61 de 926), lo cual se vuelve un factor importante en las proporciones encontradas de conservación ([Fig 15A](#)). Tal es el caso de la especie *C. rubella*, en la cual encontramos el 31.66 % (133 de 420 genes) lincRNAs clasificados como especie específica. Cabe destacar que todos los lincRNAs identificados en las 5 especies de referencia presentaban más del 50 % de conservación dentro de la familia Brassicaceae ([Fig 15C, D](#)).

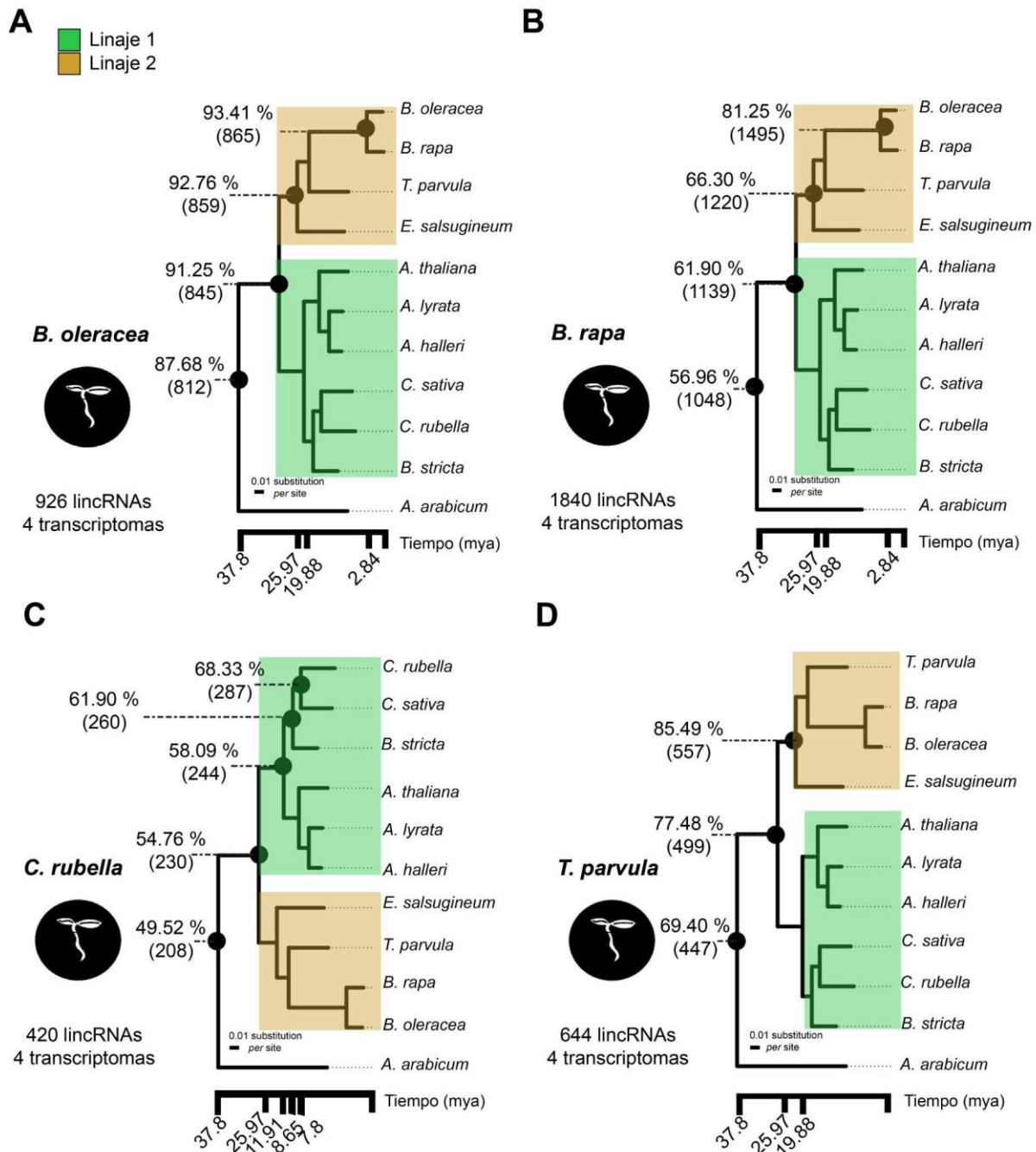


Figura 15. lincRNAs conservados y con ancestría en común dentro de la familia Brassicaceae. (A) Distribución de los lincRNAs de *B. oleracea*, (B) *B. rapa*, (C) *C. rubella* y

(D) *T. parvula* conservados por sintenia y microhomología a través de la filogenia. Tiempo de divergencia estimado de acuerdo con TimeTree.org. Todos los lincRNAs se identificaron de los transcriptomas provenientes de las raíces y parte aérea.

3.2.5 lincRNAs con expresión órgano específica

Con la finalidad de identificar los lincRNAs que presentan expresión conservada en los diferentes órganos empleados en la generación de los transcriptomas (parte aérea y raíces) en las 5 especies. Primero se analizaron los lincRNAs conservados por sintenia y microhomología, con expresión órgano específico en parte aérea o raíces en las diferentes especies, empleando los resultados obtenidos del análisis de expresión diferencial (DEG) ($\text{padj} < 0.05$ & $\log_2\text{FoldChange} \geq 0.5$). Además, se analizaron los lincRNAs con expresión global en las plántulas, a partir de los resultados obtenidos con la normalización de cuentas con la función `rlog` [regularized log] de DESeq2 (Love et al., 2014a). Los lincRNAs conservados por sintenia y microhomología tienen una menor expresión a comparación de los genes codificantes ($p\text{value } 4.41\text{e-}85$, $\text{padj } 2.205\text{e-}84$, wilcoxon test, prueba Bonferroni), exceptuando el caso de la especie *B. oleracea* ($p\text{value } 2.00\text{e-}03$, $\text{padj } 1.000\text{e-}02$, wilcoxon test, prueba Bonferroni) (Fig 16A), siendo el mismo patrón de expresión que encontramos en términos globales (Fig 16B). Además, una baja proporción de los lincRNAs presentan patrones de expresión específica en algún órgano (raíces y parte aérea), por lo que, la mayoría de los lincRNAs cuentan con expresión global a nivel de toda la plántula (entre el 49.64 al 89.40 %), encontrando este mismo comportamiento en los genes codificantes (49.69 al 60.73 %) (Fig 16B). Sin embargo, los lincRNAs con expresión diferencial en raíces y en la parte aérea presentaron niveles más altos de expresión a comparación de los lincRNAs con expresión global en *A. thaliana* y *B. oleracea* (Fig 16C). De igual manera, encontrando esta misma situación en los genes codificantes con expresión diferencial en algún órgano (Fig 16C). La alta expresión proveniente de los genes codificantes en todas las especies puede relacionarse con los factores de transcripción, enzimas y complejos proteicos relacionados con el desarrollo de las raíces y de la parte aérea, fotosíntesis y respuesta a estímulos (Fig 17). Debido a esto podemos relacionar a los lincRNAs conservados y expresados en los diferentes órganos con alguno de estos procesos, como parte de la regulación génica en los mismos (Fig 17).

Cabe destacar que los lincRNAs conservados solo por sintenia y los lincRNAs especie específica presentaron patrones de expresión definida en alguno de los órganos analizados, presentando expresión prevalente en términos globales (Fig 18). Reforzando la idea de que aun con un baja conservación por secuencia (menor al rango analizado de 8 a 50 nt), estos lincRNAs conservados solo por sintenia pueden ser importantes en la regulación génica, lo cual ha sido propuesto en los trabajos previos dentro de esta familia (L. Chen & Zhu, 2022; Palos et al., 2022).

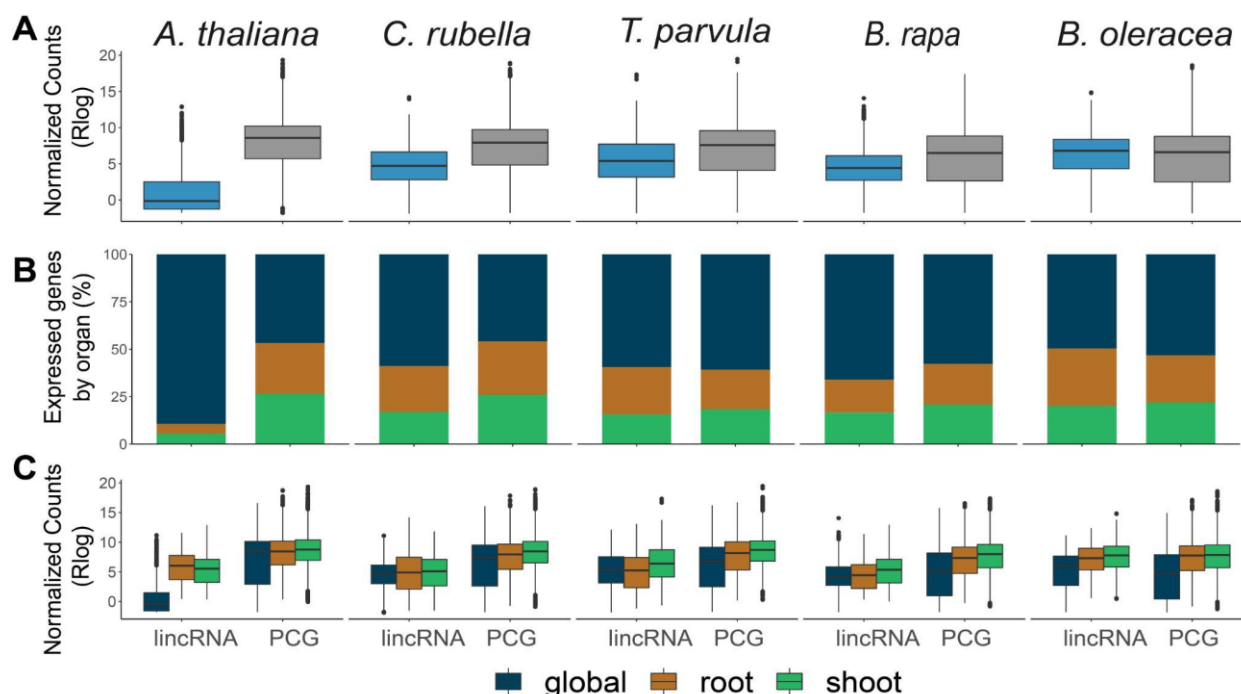


Figura 16. Patrones de expresión de los lincRNAs conservados por sintenia y microhomología en las 5 especies. (A) Expresión global de los lincRNAs conservados por sintenia y microhomología vs los genes codificantes ortólogos en cada una de las especies. (B) Distribución de los genes expresados de manera global en las plantas, así como los expresados diferencialmente en raíces y parte aérea. (C) lincRNAs y genes codificantes conservados con expresión global y órgano específica (raíces y parte aérea). Los genes codificantes ortólogos fueron obtenidos a través del programa OrthoFinder ([Sección 3.1.3](#)). PCG : Genes codificantes de proteínas (Protein-coding Genes). El análisis estadístico se realizó con la prueba de Wilcoxon (unpaired; BH correction), encontrando diferencias significativas entre la expresión de los lincRNAs vs los genes codificantes en todas las especies. En el caso de la expresión de los lincRNAs y genes codificantes en los diferentes órganos, se encontraron diferencias significativas entre todas las comparaciones excepto en: *A. thaliana* [lincRNA_root vs lincRNA_shoot], *C. rubella* [lincRNA_global vs lincRNA_root; lincRNA_global vs lincRNA_shoot; lincRNA_root vs lincRNA_shoot], *T. parvula* [lincRNA_global vs lincRNA_root; lincRNA_shoot vs protein_global], *B. rapa* [lincRNA_global vs lincRNA_root], *B. oleracea* [lincRNA_shoot vs protein_root; lincRNA_shoot vs protein_shoot].

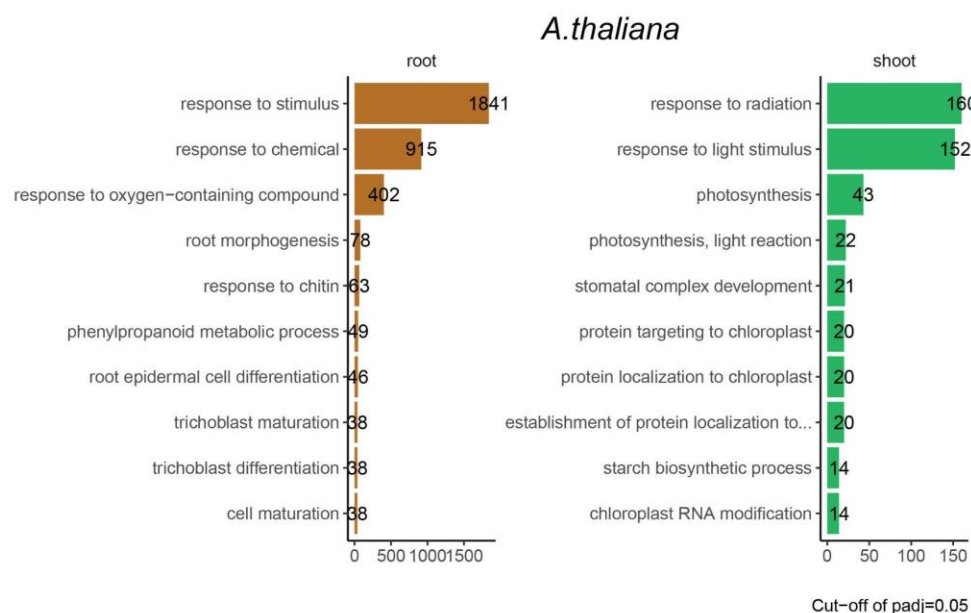


Figura 17. Funciones biológicas asociadas a los genes codificantes ortólogos y diferencialmente expresados en las raíces y parte aérea en *A. thaliana*.

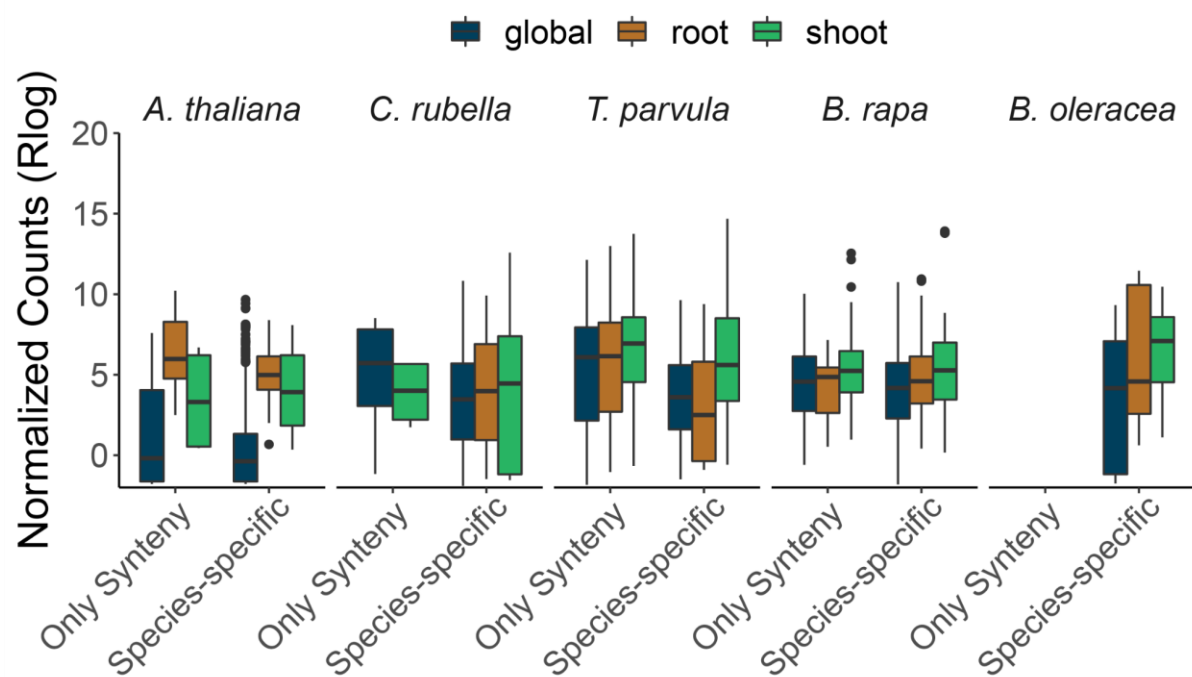


Figura 18. Patrones de expresión de los lincRNAs conservados solo por sintenia y especie específica en las 5 especies.

3.2.6 lincRNAs con expresión conservada

Con el fin de analizar los patrones de expresión de los lincRNAs conservados por sintenia y microhomología en las diferentes especies, localizamos las coordenadas de los lincRNAs conservados en cada una de las especies empleando el programa intersectBed (de BEDtools) (Quinlan & Hall, 2010). Encontrando el 39.75 % (990 de 2490 genes) AtlincRNAs expresados (Fig S20, S21), de los cuales 772 AtlincRNAs presentan expresión y conservación por sintenia y microhomología (Fig S22), 34 AtlincRNAs con conservación solo por sintenia (Fig S23) y 184 son especie específica.

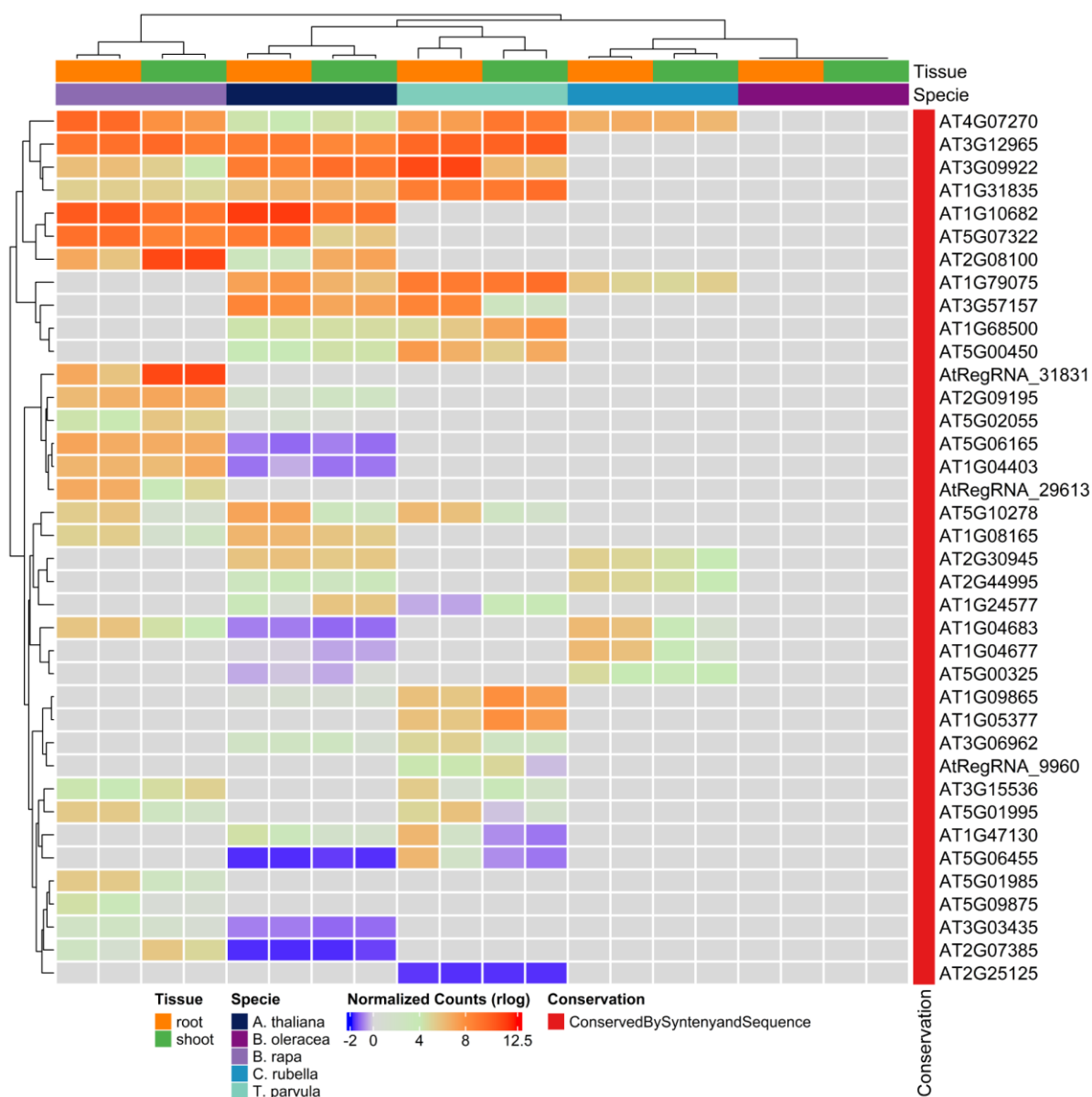


Figura 19. Patrones de expresión de los AtlincRNAs conservados por sintenia y microhomología en las diferentes especies. Los tipos de conservación identificados en cada lincRNA de *A. thaliana* se localizan en la parte derecha del heatmap. En la parte superior se encuentran las diferentes especies y órganos empleados para generar los transcriptomas

(raíces y parte aérea). AtlincRNAs: lincRNAs identificados en *A. thaliana*. Selección al azar de uno de los parálogos detectados en cada especie con respecto a cada gen de *A. thaliana*.

Analizando la expresión de los AtlincRNAs conservados por sintenia y microhomología a través de las especies, encontramos 38 genes de AtlincRNAs con expresión conservada en otras especies, excepto en *B. oleracea*, 29 de estos genes presentaban expresión y 9 de ellos no presentaron expresión en *A. thaliana* ([Fig 19](#)). La baja o nula expresión en los genes puede deberse a la selección al azar de uno de los parálogos solo para la representación gráfica (heatmap) y que este no tenga expresión, lo cual no descarta la posibilidad de contar con expresión en los otros parálogos detectados. Además, la localización de los lincRNAs a través de un alineamiento genómico no tiene una relación directa con la expresión del gen ([Fig S20](#)). Recordando que la expresión de los lincRNAs llega a depender de las condiciones, órganos y etapas del desarrollo en donde se localizan (Ariel et al., 2014a; Qin et al., 2017b; J. Wang et al., 2021). En el caso de los lincRNAs con expresión detectada en *A. thaliana*, encontramos 21 con expresión global en la plántula, lo cual se corrobora con la expresión de los lincRNAs *AT1G31835* (*BrRegRNA_27061*, *TpRegRNA_2375*) y *AT3G12965* (*BrRegRNA_15697*, *BoRegRNA_2886*, *TpRegRNA_6839*) en *B. rapa* y *T. parvula* ([Fig 19](#)); encontrando 6 lincRNAs con expresión conservada en las raíces a través de las especies, mientras que el resto de los lincRNAs expresados mostraban variaciones en la localización del lincRNA en las especies, por ejemplo el lincRNA *IPS1* en *A. thaliana* se encontraba expresado diferencialmente en la parte aérea, mientras que en *B. rapa* y en *T. parvula* se localiza con expresión global y raíces, respectivamente. Encontrando que este lincRNA puede expresarse tanto en raíces como en la parte aérea (Martin et al., 2000). Retomaremos a este lincRNA más tarde ([Sección 3.2.9](#)).

En el caso de los lincRNAs conservados solo por sintenia, localizamos 2 lincRNAs con expresión conservada el gen *AT4G30975* y *AT2G00580*. Encontramos los genes lincRNAs ortólogos del lincRNA *AT4G30975* en *B. oleracea* (*BoRegRNA_26102*) y en *C. rubella* (*CrRegRNA_16990*). En el caso del lincRNA *AT2G00580*, localizamos a este lincRNA en *B. rapa* (*BrRegRNA_25785*) y *B. oleracea* (*BoRegRNA_22995*) ([Fig 20](#)). En ambos casos sin expresión detectada en el ortólogo presente en *B. oleracea*. Para tener un indicio de sus posibles funciones, localizamos al lincRNA *AT2G00580* dentro del módulo 17 dentro de los resultados de co-expresión del capítulo 2, este módulo se encuentra relacionado con la regulación del RNA y la transcripción, lo cual cobra sentido con la retención de este lincRNA a través de las especies, a pesar de no localizar motivos en su secuencia. Por lo que, la expresión conservada de estos dos lincRNAs conservados por sintenia demuestra que a pesar de los bajos niveles en la conservación de secuencia dentro de sus regiones sinténicas, los lincRNAs pueden conservarse y tener expresión conservada a través en las especies, contando con funciones regulatorias importantes (L. Chen & Zhu, 2022; Palos et al., 2022).

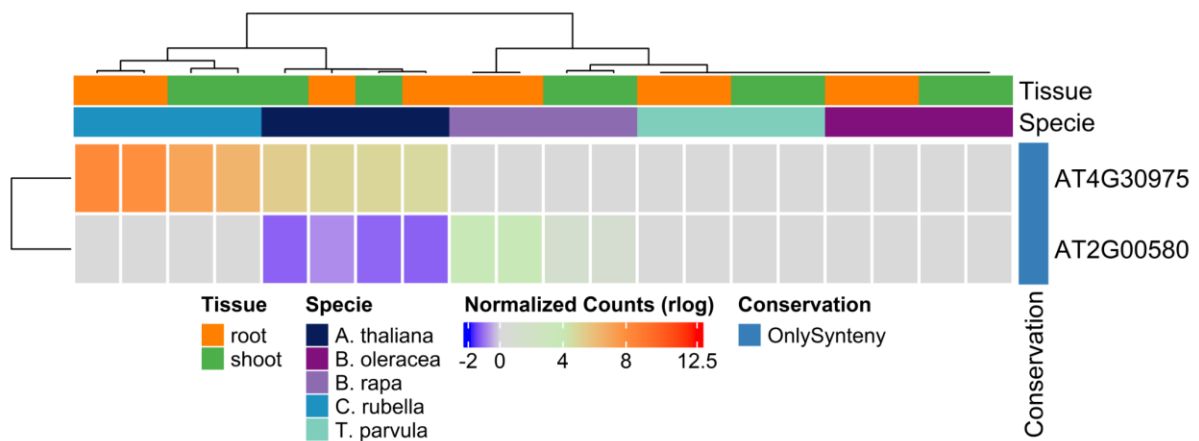


Figura 20. Patrones de expresión de los AtlincRNAs conservados solo por sintenia y en las diferentes especies. En la parte superior se encuentran las diferentes especies y órganos empleados para generar los transcriptomas (raíces y parte aérea). AtlincRNAs: lincRNAs identificados en *A. thaliana*.

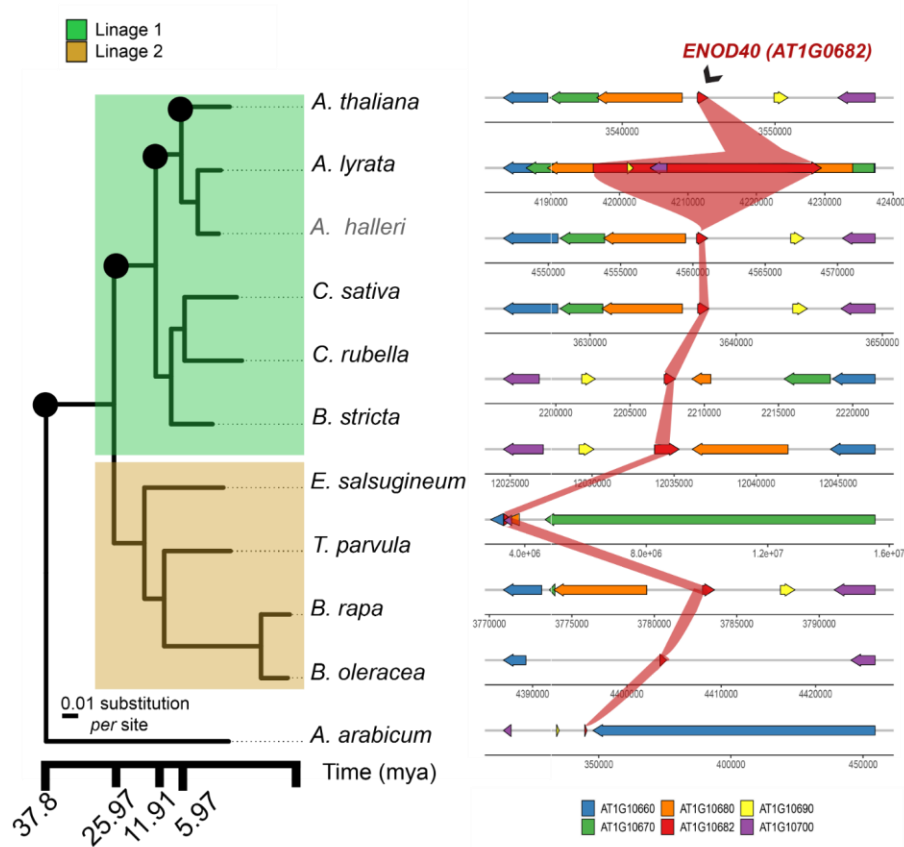
3.2.7 lincRNAs con expresión conservada en raíces

Dentro de los lincRNAs con expresión conservada en las raíces a través de las especies, encontramos a los lincRNAs *AT5G10278* (*sORF30*) en *B. rapa* (*BrRegRNA_31200*) y *T. parvula* (*TpRegRNA_16204*). Este lincRNAs se encuentra clasificado como *sORF*, dentro de esta categoría encontramos al lincRNA *ENOD40* (*AT1G10682*, *sORF1*) y *sORF5* (*AT3G57157*) los cuales cuentan con expresión conservada en las raíces a través de las especies (*ENOD40* en *B. rapa*, *BrRegRNA_14445*; *sORF5* en *T. parvula*, *Tp5g05420*). La expresión localizada de *ENOD40*, coincide con lo reportado previamente (Charon et al., 1997; Flemetakis et al., 2000). A pesar de su baja conservación a nivel de su secuencia (Flemetakis et al., 2000; Kouchi et al., 1999), hemos podido localizarlo conservado por sintenia y microhomología empleando el alineamiento genómico (Fig 21), además de encontrar su expresión conservada y localizada en raíces (Fig 22). Cabe destacar que este lincRNA no había podido ser localizado por sintenia en *A. thaliana* ni en *B. rapa* (Ganguly et al., 2021), siendo el primer reporte de localización por sintenia en este trabajo (Fig 21 y 22). Del mismo modo que en maíz (*Zea mays*), arroz (*Oryza sativa*), *Aeschynomene evenia*, el lincRNA *ENOD40* no provenia del gen *YSL7* (TRICHOME BIREFRINGENCE-LIKE 17), encontrando a ambos genes en cromosomas distintos, lo cual difiere con el lincRNA reportado en Medicago (Ganguly et al., 2021; Griesmann et al., 2018). A pesar de ser el primer lincRNA identificado en plantas, este lincRNA no se encuentra categorizado adecuadamente en Araport11 (C.-Y. Cheng et al., 2017), pero si se encuentra anotado en las bases de datos de NONCODE v6 (Y. Zhao et al., 2016), PlncDB v2 (J. Jin et al., 2020) y EvlncRNA v2 (B. Zhou et al., 2021). Adicionalmente, encontramos motivos conservados relacionados con la conservación de su estructura, lo cual coincide con lo anteriormente reportado (Gulyaev & Roussis, 2007). Los motivos 2 y 3 se encuentran localizados dentro del dominio 2 y la región 2, presentando complementación de bases entre ambos motivos para la formación del tallo en la estructura del dominio 2 (Fig 23), relacionando con la estructura conservada y predicha en *A. thaliana*, siendo el dominio 2 uno de los dominios más variantes en su secuencia (Gulyaev & Roussis, 2007). Además de relacionar a este dominio con sitios de unión a

metales (Mg II) esenciales para organizar la estructura secundaria del RNA (Podkowinski et al., 2009).

Adicionalmente, a partir de su localización genómica, logramos identificar al lincRNA *DONE40* (*MSTRG.1474* / *AtRegRNA_1474*), proveniente de la misma región del lincRNA *ENOD40* pero en sentido opuesto al mismo, identificado previamente en *Arachis hypogaea* (Ganguly et al., 2021). De acuerdo con la base de datos PlncRNADB (Bai et al., n.d.) toda la región comprendida por ambos transcritos presenta motivos de unión con proteínas (RNA binding protein, RBP), los cuales han sido previamente reportados en este lincRNA (Campalans et al., 2004) (Fig S19).

A



B

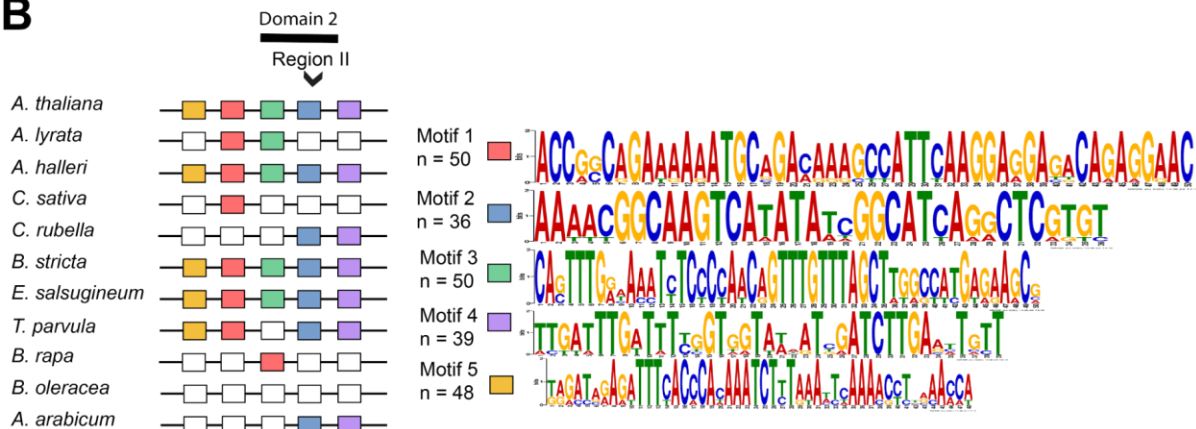


Figura 21. Conservación por sintenia y microhomología del lincRNA *ENOD40* identificados en *A. thaliana*. Tiempo de divergencia estimado de acuerdo con TimeTree.org.

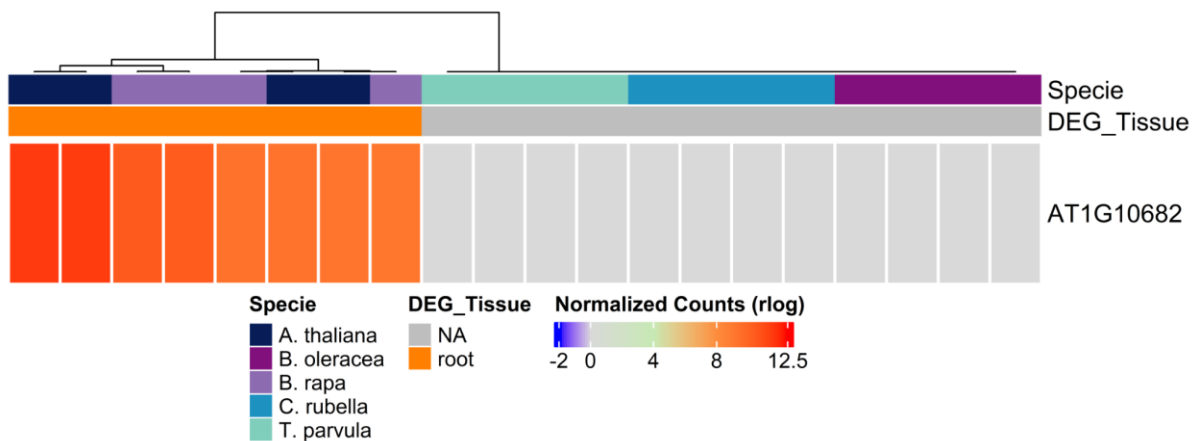


Figura 22. Expresión conservada y porcentaje de identidad en la secuencia del lincRNA *ENOD40*. En la parte superior encontramos la expresión localizada del lincRNA dividida en: global (azul), raíces (root, naranja) y parte aérea (shoot, verde), además, se muestran las 3 especies donde localizamos su expresión y conservación por sintenia y microhomología. El porcentaje de identidad se encuentra en un rango de 0 a 100 % con respecto al transcrito de *A. thaliana*. Los genes localizados en las otras son *BrRegRNA_3639* y *TpRegRNA_16718*, encontrados en *B. rapa* y *T. parvula*, respectivamente. Encontramos 4 réplicas por cada especie, debido a los 2 transcriptomas de cada órgano.

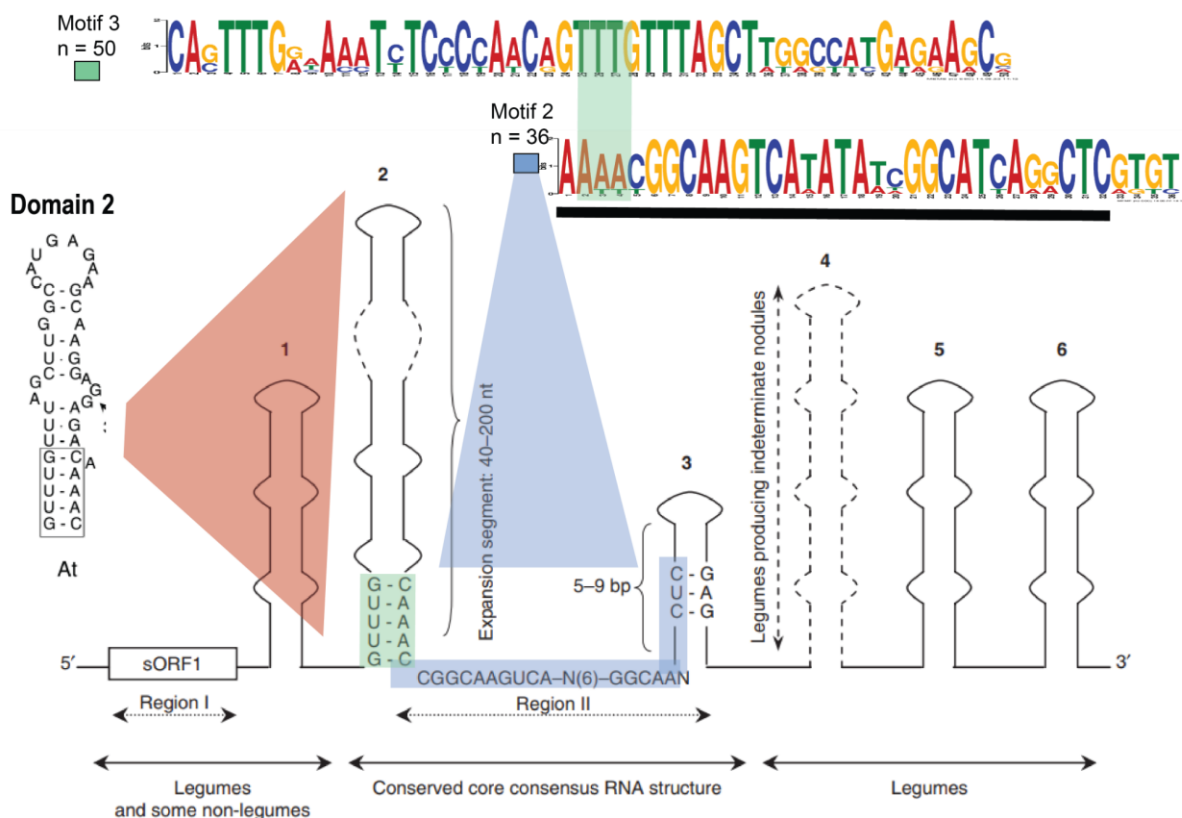


Figura 23. Estructura conservada del lincRNA *ENOD40* y los motivos conservados obtenidos a partir del análisis de microhomología. Los motivos 2 y 3 se encuentran localizados en color verde y azul, denotando la región complementaria entre ambos motivos para la formación del tallo en el dominio 2 del lincRNA. Del lado izquierdo se encuentra el loop predicho para la especie *A. thaliana*. Figura modificada de (Gulyaev & Roussis, 2007).

3.2.8 lincRNAs ortologos funcionales

Con el fin de analizar las posibles funciones biológicas relacionadas con los lincRNAs conservados por sintenia y microhomología, se compararon los resultados obtenidos con los lincRNAs con funciones anotadas tanto los reportados en EVlncRNAs V2.0 para la especie *A. thaliana* (B. Zhou et al., 2021), como los caracterizados funcionalmente. Siendo un total de 58 genes de lincRNAs con anotación funcional y expresados en *A. thaliana* (Tabla S5, Tabla 7, Fig 24). Dentro de este listado encontramos al lincRNAs relacionados con procesos biológicos como la germinación y desarrollo (*lncCOBRA1*, *lncCOBRA3*, *lncCOBRA5*, *MARS*) (Kramer et al., 2022; Roulé, Christ, et al., 2022), floración (*FLAIL*, *FLINC*, *LINC-AP2*) (Gao et al., 2016; Y. Jin et al., 2021; Severing et al., 2018), homeostasis de fosfato (*IPS1*, *At4*) (Franco-Zorrilla et al., 2007b; Shin et al., 2006a), y defensa contra patógenos (*LINC-AP2*, *ELENA1*) (Gao et al., 2016; Seo et al., 2017a). A pesar de que algunos se encuentren caracterizados funcionalmente en alguna condición, los hemos localizado expresados en plántulas con condiciones normales (Fig 24). Por lo que, posiblemente su expresión se relacione con funciones regulatorias en esta etapa del desarrollo, sin descartar su importancia biológica ante condiciones de estrés.

Tabla 7. AtlincRNAs con validez experimental y con conservación por sintenia y microhomología en las otras especies.

AtlincRNA geneID	Name	Especies con expresión	GeneID en otras especies
AT3G25795	TAS4	<i>B. rapa</i> , <i>B. oleracea</i>	<i>BrRegRNA_26595</i> , <i>BoRegRNA_31881</i>
<i>AtRegRNA_9454</i>	pri-miR858a	-	-
AT4G12735	npc60	-	-
AT1G67920	npc536	<i>B. rapa</i> , <i>B. oleracea</i> , <i>T. parvula</i>	<i>Bra004062</i> , <i>Bo7g050320</i> , <i>Tp5g22920</i>
<i>AtRegRNA_12326</i>	npc375	-	-
AT4G04692	npc156	-	-
AT5G00580	MARS/At5NC066580	-	-
AT3G05655	lncCOBRA5/At3NC040900	-	-
AT3G03435	lncCOBRA3/At3NC020890	<i>B. rapa</i>	<i>BrRegRNA_15395</i>
AT1G05913	lncCOBRA1/At1NC031460	-	-
AT4G09445	LINC-AP2/At4NC069370	-	-
AT3G09922	IPS1	<i>B. rapa</i> , <i>T. parvula</i>	<i>BrRegRNA_3639</i> , <i>TpRegRNA_16718</i>
AT2G18440	GUT15	-	-
AT1G08103	FLINC/At1NC078930	<i>B. oleracea</i>	<i>BoRegRNA_31476</i>

AT2G18735	FLAIL	-	-
AT1G10682	ENOD40	<i>B. rapa</i>	<i>BrRegRNA_14445</i>
<i>AtRegRNA_1474</i>	ENOD40	<i>T. parvula</i>	<i>TpRegRNA_816</i>
AT4G16355	ELENA1	-	-
AT2G18440	AtGUT15	-	-
AT1G08165	At4-3	<i>B. oleracea</i>	<i>BrRegRNA_24772</i>
AT2G08825	At4-2	<i>B. oleracea</i>	<i>BoRegRNA_12954</i>

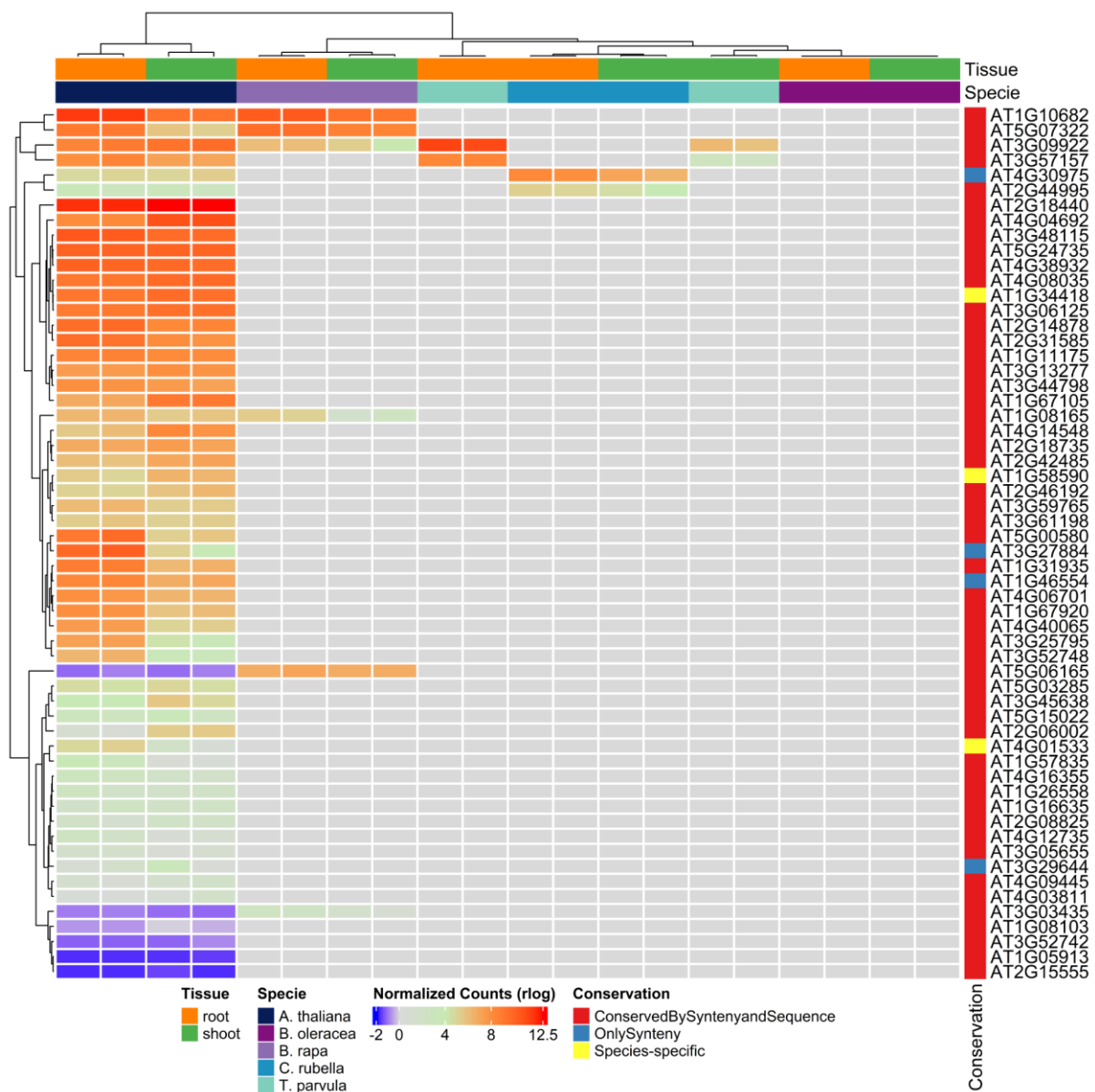


Figura 24. Patrones de expresión de los 58 genes At lincRNAs ortólogos con funciones anotadas. Los tipos de conservación identificados en cada lincRNA de *A. thaliana* se localizan en la parte derecha del heatmap. En la parte superior se encuentran las diferentes

especies y órganos empleados para generar los transcriptomas (raíces y parte aérea). AtlincRNAs: lincRNAs identificados en *A. thaliana*. Selección al azar de uno de los parálogos detectados en cada especie con respecto a cada gen de *A. thaliana*.

3.2.9 Conservación de *IPS1* y sus parálogos

Dentro de la lista de lincRNAs conservados y funcionales encontramos a los lincRNAs *IPS1*, sus parálogos *At4-3* y *At4-2*, los cuales presentan redundancia funcional en la regulación del fosfato en la planta (Franco-Zorrilla et al., 2007b; Shin et al., 2006a). Sin embargo, el contexto genómico de estos lincRNAs es diferente, prevaleciendo el motivo conservado y relacionado con su función en estos lincRNAs (Fig 25B). Motivo que se encuentra implicado con el secuestro del miR399, pero sin degradar del transcrito de *IPS1*, debido a la unión incompleta con el sitio de corte (Franco-Zorrilla et al., 2007b) (Fig 25B). Enfocándonos en el lincRNA *IPS1* encontramos 5 genes codificantes vecinos (*AT3G09920*, *AT3G09925*, *AT3G09930*, *AT3G09940* y *AT3G09950*) de este lincRNA a una distancia de 10 kb en *A. thaliana*, la gran mayoría del contexto genómica se encontraba conservado en las diferentes especies. Además, identificamos inversiones de los genes codificantes en *B. rapa*, *B. oleracea*, *Eutrema salsugieum* y *Boechera stricta* (Fig 25A). Por su parte, localizamos múltiples copias de este lincRNA en bloques sinténicos en *B. oleracea* (cromosomas C1, C2, C3, C5, C7y C9), *B. rapa* (A01, A02, A03, A05 y A10), *C. sativa* (CM002729, CM002743, CM002747), *E. salsugineum* (NW_006256829, NW_006256885) y *T. parvula* (CM001189, CM001192) (Tabla S6), relacionando estos sitios conservados con las duplicaciones a nivel del genoma y a la alta conservación por sintenia presente en estos genomas (Kagale, Koh, et al., 2014; Kagale, Robinson, et al., 2014; S. Liu et al., 2014). Pudiendo considerar la conservación de estas copias como un rasgo adquirido para la adaptación y supervivencia de estas especies ante los diversos cambios ambientales, lo cual ha sido previamente reportado en los genes codificantes de *C. sativa* (Song et al., 2020).

Posteriormente analizamos la región complementaria (CR) al miR399, siendo un motivo conservado e importante para su función como esponja de este miRNA. Las especies pertenecientes al Linaje I y II contenían este motivo conservado (Fig 25C). Por su parte, la especie *A. halleri* presentaba este motivo, a pesar de localizar solamente al gen codificante *AT3G09925* cercano a este lincRNA (Fig 25B). Lo cual nos permitió corroborar que el lincRNA identificado en esta especie era homólogo al gen *IPS1* de *A. thaliana*. Adicionalmente, encontramos diversas regiones de microhomología en las diferentes especies, siendo el motivo más preservado la secuencia CR al miR399 (Fig 25C), este motivo presenta un tamaño de apenas 23 nt y ha sido reportado conservado a través del reino vegetal (Rymarquis et al., 2008). Algo interesante por destacar, es el hecho de que el contexto genómico donde se encuentra este lincRNA en la especie *A. arabicum* se encuentra altamente conservado, sin embargo, su secuencia presenta una baja conservación, logrando identificar al menos dos regiones con microhomología con ayuda del alineamiento genómico (Fig 25C), aunque la CR al miR399 no se identificó (Fig 25B). Por lo cual, la combinación de ambos tipos de conservación, se muestra favorable para la localización de lincRNAs en plantas, siendo similar a lo que se reportó previamente para la identificación de los lincRNAs *roX1* y *roX2* en *Drosophila* (Quinn et al., 2016). Recordando que este lincRNA se encuentra ampliamente conservado en el reino vegetal (tomate, medicago, arroz, cebada, maíz) (Burleigh & Harrison, 1999; C. Y. Huang et al., 2011; C. Liu et al., 1997; Oono et al., 2013; Y. Zhu et al., 2022). La

localización de este lincRNA en *Selaginella moellendorffii* demuestra su existencia desde las plantas vasculares (Y. Zhu et al., 2022). Por lo que, la conservación por sintenia de este lincRNAs y los otros motivos localizados en su secuencia, se pueden relacionar a la conservación por estructura de este lincRNA, lo cual ha sido previamente reportado en la tesis de mi compañero Jose Antonio Corona Gomez ([Fig 26](#)).

Con relación a su expresión, el lincRNA *IPS1* se detectó expresado en *A. thaliana* (*AT3G09922*), *B. rapa* (*BrRegRNA_3639*) y *T. parvula* (*TpRegRNA_16718*) con una alta expresión en las 3 especies, aunque en el caso de *T. parvula* fue mayor su expresión a comparación del resto de las especies. Como se había mencionado previamente, la expresión de este lincRNA se puede determinar como global, al encontrarse presente en la parte aérea y en las raíces (Martin et al., 2000), en *A. thaliana* encontramos una expresión significativa en la parte aérea, mientras que en *T. parvula* fue en las raíces, y en el caso de *B. rapa* su expresión fue global, lo cual concuerda con lo anteriormente reportado (Martin et al., 2000). Al comparar las secuencias de los otros lincRNAs ortólogos con respecto a *A. thaliana*, encontramos un 32.10 % de conservación en *B. rapa* y un 32.47 % en *T. parvula* ([Fig 27](#)). Demostrando que a pesar de su baja conservación por secuencia, este lincRNA prevalece como una ventaja adaptativa en respuesta a los cambios ambientales (Y. Zhu et al., 2022).

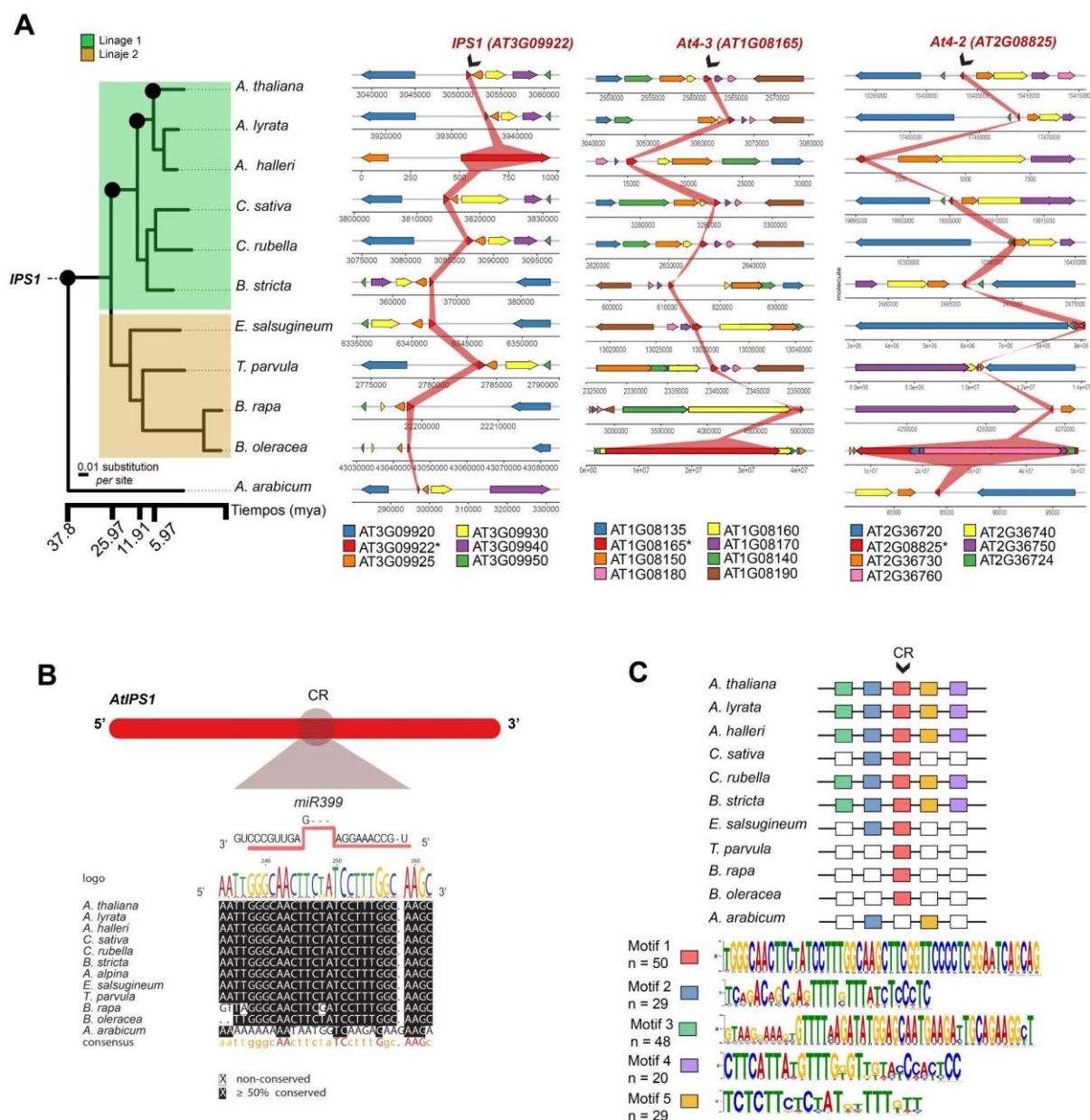


Figura 25. Conservación por sintenia y microhomología de los lincRNAs identificados en *A. thaliana*. (A) Bloque sinténico del lincRNA *IPS1* a través de la familia Brassicaceae. Distribución de los genes codificantes ortólogos y conservados por posición localizados a 10 kb (upstream y downstream) del lincRNA *IPS1* (rojo) en cada una de las especies. Las escalas de cada bloque sinténico fueron detectadas independientemente para cada especie. Gráfica obtenida con el paquete ggenes (v 0.4.1) en R. (B) Conservación de la región complementaria al miR-399 en las especies analizadas del lincRNA *IPS1*. La identificación de la región complementaria se realizó por medio del paquete msa (v1.28.0) en R, empleando Clustal Omega para la comparación. (C) Conservación de motivos localizados en el lincRNA *IPS1* en las diferentes especies. Se identificaron 5 motivos con un tamaño de 20 hasta 50 nt entre las secuencias analizadas, todos los motivos detectados eran significativos (con un p value como mínimo de 7.4×10^{-7}). CR = Complementary region (región complementaria). Tiempo de divergencia estimado de acuerdo con TimeTree.org.

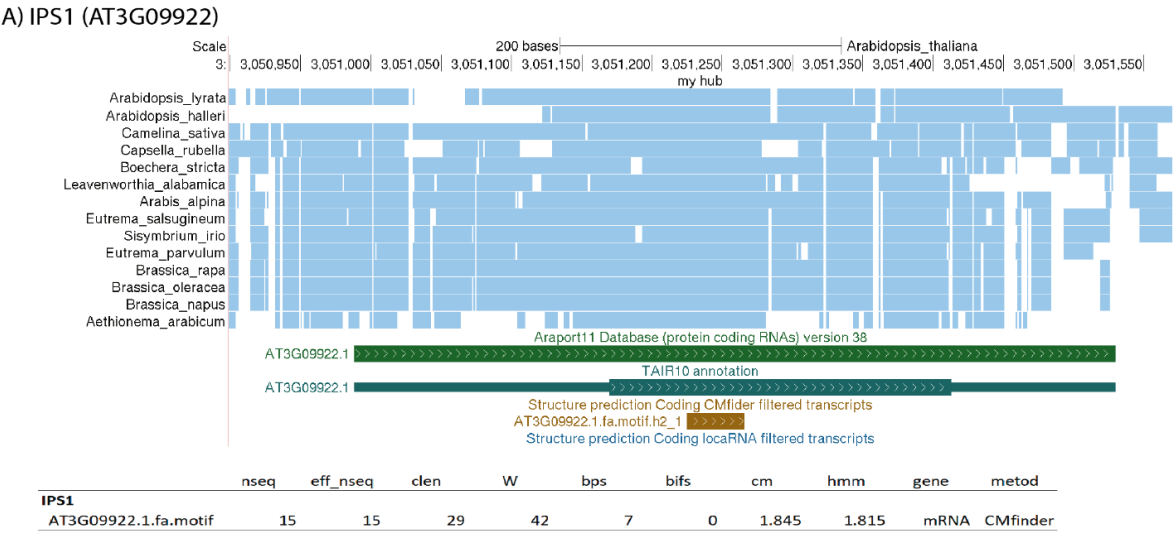


Figura 26. Conservación por estructura del lincRNA *IPS1*. Motivo estructural predicho por la herramienta CMfinder empleando el alineamiento genómico. Fragmento obtenido de la tesis de (Corona-Gomez, n.d.-a).

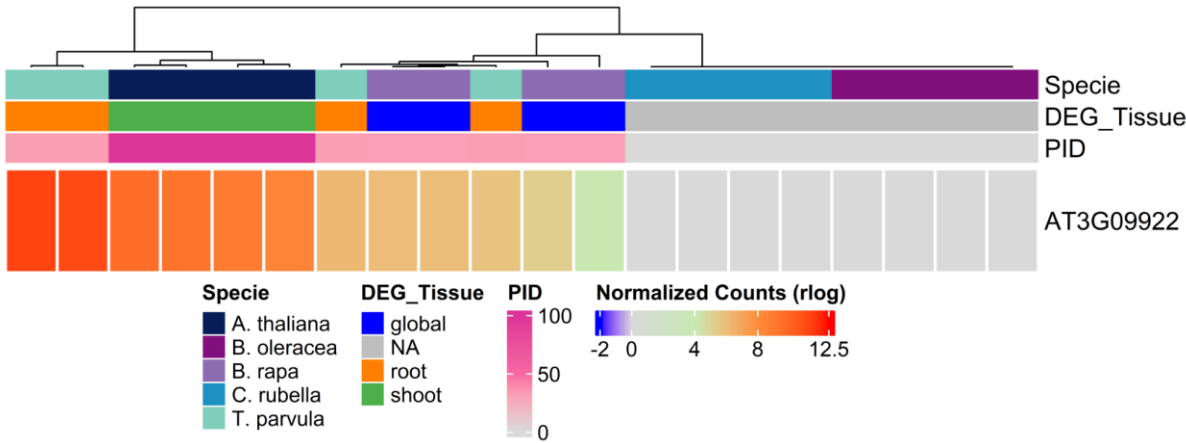


Figura 27. Expresión conservada y porcentaje de identidad en la secuencia del lincRNA *IPS1*. En la parte superior encontramos la expresión localizada del lincRNA dividida en: global (azul), raíces (root, naranja) y parte aérea (shoot, verde), además, se muestran las 3 especies donde localizamos su expresión y conservación por sintenia y microhomología. El porcentaje de identidad se encuentra en un rango de 0 a 100 % con respecto al transcrito de *A. thaliana*. Los genes localizados en las otras son *BrRegRNA_3639* y *TpRegRNA_16718*, encontrados en *B. rapa* y *T. parvula*, respectivamente. Encontramos 4 réplicas por cada especie, debido a los 2 transcriptomas de cada órgano.

3.2.10 Ganancia y pérdida de lincRNAs

La historia de una familia de genes permite conocer los cambios evolutivos relacionados con las variaciones fenotípicas entre las especies, permitiendo distinguir entre los genes ortólogos y los genes parálogos (Fitch, 1970; Ingram, 1961). Mientras que los genes ortólogos suelen mantener funciones similares, los genes parálogos pueden divergir en su función (Koonin, 2005). Se analizó la ganancia y pérdida de los lincRNAs empleando las matrices binarias

provenientes de los resultados de conservación por sintenia y microhomología ([Sección 3.1.4](#)). Encontrando 399 lincRNAs conservados en el nodo de las *Arabidopsis*, clasificándose como lincRNAs linaje específico. Conforme la distancia evolutiva aumenta se van detectando eventos de pérdida de los lincRNAs en la familia. Logrando obtener un total de 3601 lincRNAs conservados hasta el nodo ancestral compartido con *A. arabicum*. Aunque normalmente esta metodología es empleada para genes codificantes (D. Liu et al., 2016), el archivo de entrada es una matriz binaria, la cual puede ser alimentada de acuerdo a algún rasgo como el que empleamos de conservación ([Fig 28](#)). Por lo que, resulta interesante conocer a partir de estos resultados como se expresan los genes que encontramos y de qué manera se relacionan con los genes codificantes.

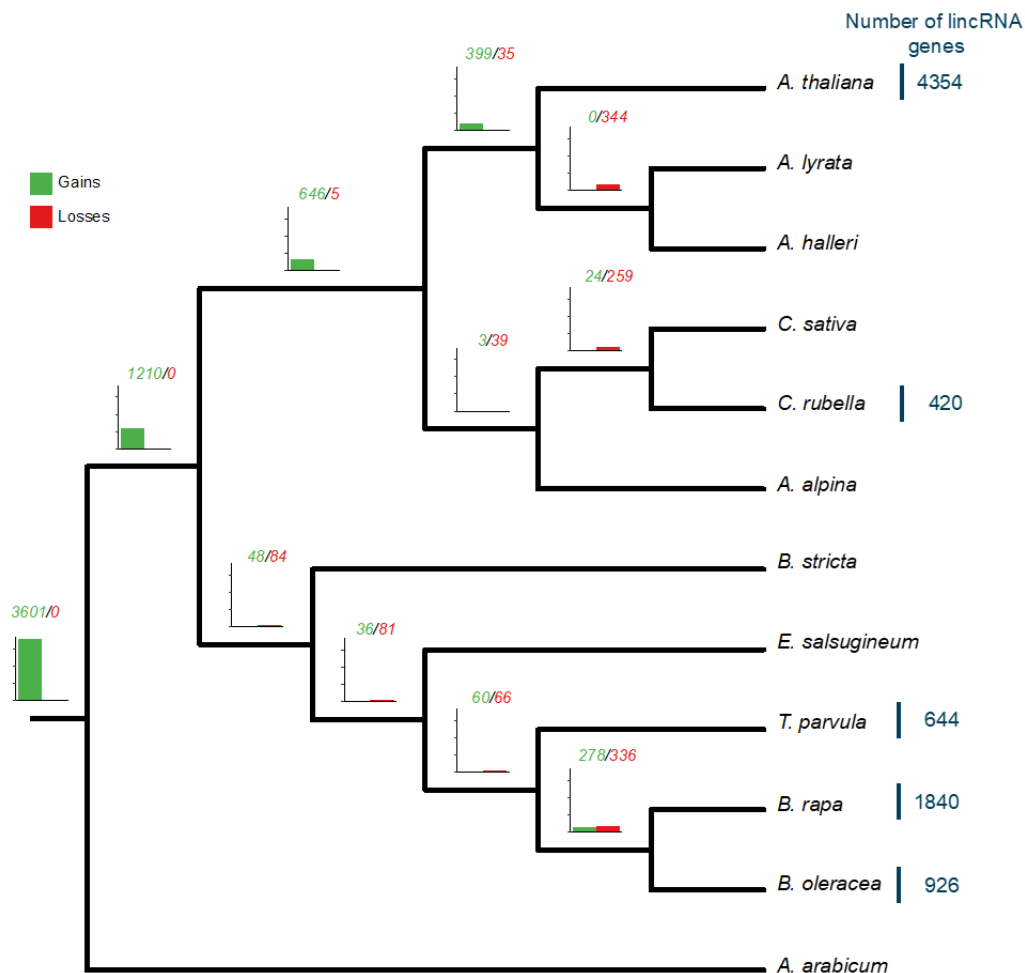


Figura 28. Ganancias y pérdidas de los lincRNA identificados en las diferentes especies. En verde se simbolizan las ganancias de los lincRNAs y en rojo las pérdidas en cada nodo. A la derecha encontramos el número de lincRNAs identificados en cada una de las especies. Tiempo de divergencia estimado de acuerdo con TimeTree.org.

3.3 Discusión

En este capítulo, se identificaron 1840 genes de lincRNAs *B. rapa*, 926 *B. oleracea*, 420 *C. rubella* y 644 *T. parvula* provenientes de la parte aérea y raíces de plántulas. En el caso de *A. thaliana* se empleó la notación generada en el capítulo 1, separando los lincRNAs

provenientes de los mismos transcriptomas (parte aérea y raíces), con el fin de poder comparar entre las especies. De los cuales más del 48 % de los lincRNAs se encuentran conservados por sintenia y más del 34.63 % presentan motivos conservados en su secuencia. Asimismo, más del 48 % se localizan en la especie basal *A. arabicum*, clasificándolos como ancestrales. Por su parte, una proporción de estos cuentan con expresión conservada a través de las especies, resaltando funciones relacionadas como el desarrollo, germinación, homeostasis del fósforo y modificaciones en el RNA.

3.3.1 Características de los lincRNAs

Los lincRNAs en general tienden a tener características inferiores a los genes codificantes, al comparar el tamaño de sus transcritos, el número de exones y su expresión, siendo reportado previamente en plantas (Deng et al., 2018; C.-Y. Wang et al., 2017), esponjas (Gaiti et al., 2015), mamíferos (Cabili et al., 2011) y animales (Hezroni et al., 2015). Sin embargo, los lincRNAs identificados en la especie *B. oleracea* son multiexónicos, lo cual ha sido previamente reportado en esta especie (X. Zhu et al., 2019) y siendo una característica predominante en los lincRNAs identificados en mamíferos (Hezroni et al., 2015), por su parte, en plantas como tomate, *Cleistogenes songorica*, durazno (*Prunus persica*), también se han reportado lincRNAs con transcritos multiexónicos (Yan et al., 2019; Z. Yang et al., 2019; H. Zhou et al., 2022). Además, sus transcritos presentaron un mayor tamaño a comparación de los BolincRNAs identificados en las otras especies y presentando una mayor expresión a comparación de los genes codificantes (Wilcoxon test, p value < 2.2e-16), lo cual difiere de lo previamente reportado en los lincRNAs. Lo cual puede sugerir implicaciones en el desarrollo por parte de estos lincRNAs en esta etapa analizada.

3.3.2 lincRNAs son importantes elementos regulatorios en plantas

Para analizar la conservación por sintenia y microhomología de los lincRNAs se analizaron los genes codificantes vecinos de cada lincRNA a una distancia de 10 Kb. A pesar de que *B. rapa* y *B. oleracea* tienen los genomas más grandes a comparación de las otras especies, encontramos más genes codificantes vecinos que los identificados en *T. parvula*, siendo una de las especies con el genoma más pequeño al igual que *A. thaliana*. Siendo *A. thaliana* la especie con más genes codificantes flanqueando a los lincRNAs que en las otras especies. Lo cual coincide con una alta conservación por sintenia por parte de estos genes en *A. thaliana*, *C. rubella*, *B. rapa* y *B. oleracea* (Kagale, Robinson, et al., 2014; S. Liu et al., 2014; Schranz et al., 2006; Slotte et al., 2013). Siendo una distancia de 10 Kb una buena opción para la localización de lincRNA aun en genomas grandes. En el caso de *T. parvula*, la baja localización de genes codificantes puede relacionarse con problemas en el archivo de anotación, por lo que, genes que no estén anotados aledaños a algún lincRNA no serán detectados en el análisis. Por otro lado, al emplear el alineamiento genómico como punto de partida en el análisis, identificamos inversiones de los bloques sinténicos detectados a través las especies, lo cual había sido previamente reportado para los genes ortólogos dentro de esta familia (Koch & Kiefer, 2005; Murat et al., 2015), pero no en lincRNAs. Permitiendo reconocer el impacto que tienen las modificaciones en el genoma y su relación con la conservación de genes importantes en la adaptación de las especies.

Siendo observado en los eventos de duplicación del genoma, en donde los nuevos genes obtenidos pueden proporcionar nuevas funciones génicas para la adaptación a diversas condiciones de crecimiento (Gehring et al., 2006; Moser, 2012; Serdyuk et al., 2022). Por lo que, a diferencia de uno de los trabajos más recientes en el análisis de conservación de lincRNAs en Brassicaceae, en donde solo multiplicaron la cantidad de lincRNAs identificadas por 3 para determinar el número de copias presentes en las especies (Palos et al., 2022). Sesgando el resultado al número de copias obtenidas en los genes codificantes (Kagale, Koh, et al., 2014; Kagale, Robinson, et al., 2014; S. Liu et al., 2014). A diferencia de este trabajo, nosotros podemos localizar las coordenadas de las múltiples copias (parálogos) de cada lincRNA a partir de los métodos empleados, identificando más de 2 copias de cada lincRNA en *A. thaliana* con respecto a *C. sativa*, *B. rapa* y *B. oleracea*. Lo cual difiere con la conservación mínima de 3 copias por cada gen codificante anotado en *A. thaliana* (Kagale, Koh, et al., 2014; Kagale, Robinson, et al., 2014; S. Liu et al., 2014). Las nuevas copias encontradas de los lincRNAs se relacionan con estos eventos de duplicación, por lo que, la conservación de estos lincRNAs puede conllevar a la adaptación a condiciones adversas de una especie a lo largo del tiempo, llegando a ser beneficiosos para la supervivencia y el crecimiento, como lo han llegado a observar en los genes codificantes (Song et al., 2020). Recordando que uno de los escenarios evolutivos importantes para el surgimiento de nuevos lincRNAs es la duplicación del genoma (Ulitsky & Bartel, 2013).

3.3.3 lincRNAs ortologos funcionales

Encontramos que más del 48 % (2090 genes) de los AtlincRNAs se encuentra conservado por sintenia y microhomología hasta la especie basal *A. arabicum*, siendo uno de los valores más altos de conservación reportados en plantas a comparación con los trabajos previos (L. Chen & Zhu, 2022; Mohammadin et al., 2015; Nelson et al., 2016; Palos et al., 2022). En las otras especies de referencia encontramos más del 49.52 % de los lincRNAs conservados como ancestrales. La conservación de los lincRNAs en bloques sinténicos puede ser relacionado en algunos casos con la regulación en *cis* que este puede llegar a tener con los genes codificantes vecinos, pero no siempre es el caso. Un ejemplo de un lincRNA con acción en *trans* es *IPS1*, encontrado ampliamente conservado en el reino vegetal (Burleigh & Harrison, 1999; C. Y. Huang et al., 2011; C. Liu et al., 1997; Oono et al., 2013; Y. Zhu et al., 2022), y con función conservada al actuar como esponja del miR399 para la homeostasis del fosfato (Burleigh & Harrison, 1999; Franco-Zorrilla et al., 2007b; C. Liu et al., 1997; Shin et al., 2006a). En este trabajo, lo identificamos conservado por sintenia y microhomología presente a lo largo de la filogenia.

Adicionalmente, identificamos 58 genes de lincRNAs con anotación funcional y expresión dentro de los lincRNAs conservados por sintenia y microhomología. Entre los que encontramos al lincRNA *IPS1*, *TAS4*, *npc60*, *npc536*, *npc375*, *npc156*, *FLAIL*, *FLINC*, *ENOD40*, entre otros. Algunos de los cuales no se habían reportado conservados por estos métodos o en algunas de las especies analizadas como es el caso del lincRNA *ENOD40*, el cual había sido previamente analizado en *A. thaliana* y *B. rapa*, pero sin ser localizado en estas por sintenia (Ganguly et al., 2021). Siendo importante destacar que dos de los motivos identificados concuerdan con la estructura conservada del dominio 2 en este lincRNA. Dominio que ha sido relacionado con sitios de unión a metales (Mg II), siendo esencial para organizar la estructura secundaria del RNA (Podkowinski et al., 2009). Permitiendo corroborar

la localización de lincRNAs ortólogos mediante la búsqueda de pequeños motivos conservados y prevalentes en la conservación por estructura.

Por su parte, encontramos lincRNAs conservados por sintenia y microhomología con expresión órgano específica en cada una de las especies, siendo más abundantes los lincRNAs con expresión en términos globales en esta etapa del desarrollo. De igual manera, identificamos lincRNAs conservados solo por sintenia expresados en algún órgano y con expresión global. Debido a esto podemos reforzar la idea sobre la importancia funcional en la regulación génica de los lincRNAs provenientes de regiones con muy baja conservación y con un bloque sinténico conservado, lo cual ha sido analizado previamente (L. Chen & Zhu, 2022; Palos et al., 2022). Siendo uno de los ejemplos más importantes el lincRNA *IPS1*, teniendo una conservación de secuencia de apenas el 32 %, presenta una alta expresión tanto en la parte aérea como en las raíces y encontrando conservado por ambos métodos a través de la familia Brassicaceae, hasta plantas vasculares como *S. moellendorffii* (Y. Zhu et al., 2022).

Identificamos lincRNAs provenientes de la parte aérea y de las raíces de 5 especies de la familia Brassicaceae. Analizamos la conservación por sintenia y microhomología de los lincRNAs provenientes de regiones intergénicas, encontrando más del 48 % de conservación de los lincRNAs a través de las especies, de los cuales una pequeña proporción se encontraban expresados en esta etapa del desarrollo analizada. Aunque los resultados obtenidos son solo de transcritos poliadenilados, excluyendo los que carecen de este rasgo. No obstante, encontramos lincRNAs diferencialmente expresados en las raíces y en la parte aérea, además de tener algunos casos con expresión conservada en las especies. Lo cual nos resalta de la importancia funcional de estos lincRNAs, así como de las implicaciones en su conservación a lo largo de la evolución, siendo importantes para la adaptación y supervivencia de las especies.

3.4 Conclusiones

Se identificaron los lincRNAs de 5 especies de la familia Brassicaceae, contando con 2490 lincRNAs en *A. thaliana*, 420 *C. rubella*, 644 *T. parvula*, 1840 *B. rapa* y 926 *B. oleracea* provenientes de la parte aérea y raíces.

En el análisis de conservación por sintenia y microhomología, se identificaron 1932 genes de lincRNAs conservados por sintenia y microhomología en *A. thaliana*, 1425 lincRNAs en *B. rapa*, 864 *B. oleracea*, 279 *C. rubella* y 337 *T. parvula* dentro de la familia Brassicaceae (11 especies). El resto de los lincRNAs se clasificaron en solo sintenia y especie específica en todas las especies. Contando con más del 48 % de los lincRNAs conservados por ambos métodos (sintenia y microhomología) y localizados hasta la especie basal *A. arabicum*.

La metodología empleada nos permite la localización por coordenadas de los lincRNAs duplicados en los diferentes genomas. Relacionando la conservación de estos lincRNAs con funciones importantes para la adaptación, brindando un panorama sobre la evolución de estos lincRNAs a través de las especies.

Con relación a su expresión, la mayoría de los lincRNAs presentaban patrones de expresión global en las plántulas, siendo muy pocos de ellos órgano específicos con expresión conservada entre las especies. Aunque podemos resaltar algunos ejemplos como los

lincRNAs *IPS1* y *ENOD40*. Lo cual refuerza la idea de que los lincRNAs cuentan con funciones importantes en el desarrollo de las plántulas y también actúan de forma específica en los distintos órganos.

Nuestros resultados proporcionan un aporte al análisis de los lincRNAs, así como a los mecanismos responsables de su conservación por sintenia y microhomología, teniendo la oportunidad de visualizar los cambios de organización que pueden sufrir las plantas a lo largo del tiempo, pero identificando contextos genómicos conservados a pesar de estos cambios en su genoma.

Conclusiones generales

En esta tesis exploramos las posibles funciones biológicas de los lincRNAs en *A. thaliana*, los cuales pueden llegar a diferir dependiendo de la etapa del desarrollo y órgano en el que se localicen. Recordando la importancia de los lincRNAs como reguladores genéticos en las especies, sin embargo aún es muy poca la información referente a su conservación y función, logramos identificar algunos de los lincRNAs con validez experimental como son: *COOLAIR*, *COLD AIR*, *ELENA1*, y *HID1*, dentro de los módulos de co-expresión.

Dentro de la misma familia encontramos especies con relevancia económica como son *B. rapa*, *B. oleracea* y *C. sativa*, las cuales cuentan con duplicaciones en su genoma, sin embargo, cuentan con una alta conservación por sintenia en los genes codificantes analizados. Por lo que, resultó interesante encontrar más del 48 % de los lincRNAs conservados dentro de bloques sinténicos y una parte de estos con múltiples copias de cada gen anotado en *A. thaliana* con respecto a estas tres especies con demanda económica. Considerando que estos nuevos genes conservados y derivados de la duplicación del genoma pueden llegar a ser responsables de las ventajas adaptativas o económicas de estas especies. Encontrando ejemplos como *IPS1* para la homeostasis en carencia de fosfato y su amplia conservación en el reino vegetal, lo cual llega a considerarlo uno de los lincRNAs más importantes que facilitaron la evolución de las plantas terrestres actuales ante condiciones adversas.

Perspectivas

1. Enriquecer el análisis de co-expresión alimentando al programa con más transcriptomas con otras condiciones, como la inducción de hormonas. Además de poder escalar el análisis a otras especies de la misma familia, como *B. rapa* o *B. oleracea*.
2. Corroborar y localizar la expresión *in vivo* de los lncRNAs conservados por sintenia y microhomología. Además, analizar su expresión a lo largo del desarrollo en plantas bajo condiciones normales.
3. Esperamos que los lncRNAs conservados por sintenia y microhomología en todas las especies, puedan ser retomados para analizar sus posibles funciones por medio de los módulos de co-expresión y pudiendo enfocar el análisis en otras especies, además de *A. thaliana*.
4. Analizar las funciones y mecanismos biológicos de los lncRNAs conservados solo por sintenia mediante la generación de mutantes (sobrexpresión y represión), no solo analizarlo a nivel de plántula, si no llegar a escalar hasta su floración y obtención de frutos.
5. Analizar la regulación en *cis* y *trans* de los lncRNAs conservados (sintenia y microhomología), así como sus posibles implicaciones biológicas.
6. A partir de las observaciones de las mutantes, se puede analizar la recuperación funcional de los lncRNAs empleando especies distantes evolutivamente (*A. thaliana* con *B. oleracea*).
7. Analizar la expresión y posibles funciones de las múltiples copias pertenecientes a cada copia única de *A. thaliana* localizadas en *C. sativa*, *B. rapa* o en *B. oleracea* con el fin de conocer los efectos de la domesticación y adaptación que han sufrido a lo largo del tiempo.

Referencias bibliográficas

- Ajmera, I., Shi, J., Giri, J., Wu, P., Stekel, D. J., Lu, C., & Hodgman, T. C. (2018). Regulatory feedback response mechanisms to phosphate starvation in rice. *NPJ Systems Biology and Applications*, 4, 4. <https://doi.org/10.1038/s41540-017-0041-0>
- Alekseyenko, A. A., Ellison, C. E., Gorchakov, A. A., Zhou, Q., Kaiser, V. B., Toda, N., Walton, Z., Peng, S., Park, P. J., Bachtrog, D., & Kuroda, M. I. (2013). Conservation and de novo acquisition of dosage compensation on newly evolved sex chromosomes in *Drosophila*. *Genes & Development*, 27(8), 853–858. <https://doi.org/10.1101/gad.215426.113>
- Alexa, A., Rahnenführer, J., & Lengauer, T. (2006). Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, 22(13), 1600–1607. <https://doi.org/10.1093/bioinformatics/btl140>
- Allen, E., Xie, Z., Gustafson, A. M., & Carrington, J. C. (2005). microRNA-directed phasing during trans-acting siRNA biogenesis in plants. *Cell*, 121(2), 207–221. <https://doi.org/10.1016/j.cell.2005.04.004>
- Allen, T. A., Von Kaenel, S., Goodrich, J. A., & Kugel, J. F. (2004). The SINE-encoded mouse B2 RNA represses mRNA transcription in response to heat shock. *Nature Structural & Molecular Biology*, 11(9), 816–821. <https://doi.org/10.1038/nsmb813>
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. In *Journal of Molecular Biology* (Vol. 215, Issue 3, pp. 403–410). [https://doi.org/10.1016/s0022-2836\(05\)80360-2](https://doi.org/10.1016/s0022-2836(05)80360-2)
- Andrew, S. (2010). *FastQC: A quality control tool for high throughput sequence data*. Babraham Bioinformatics. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
- Andrews, S. J., & Rothnagel, J. A. (2014). Emerging evidence for functional peptides encoded by short open reading frames. *Nature Reviews. Genetics*, 15(3), 193–204. <https://doi.org/10.1038/nrg3520>
- Ariel, F., Jegu, T., Latrasse, D., Romero-Barrios, N., Christ, A., Benhamed, M., & Crespi, M. (2014a). Noncoding transcription by alternative RNA polymerases dynamically regulates an auxin-driven chromatin loop. *Molecular Cell*, 55(3), 383–396. <https://doi.org/10.1016/j.molcel.2014.06.011>
- Ariel, F., Jegu, T., Latrasse, D., Romero-Barrios, N., Christ, A., Benhamed, M., & Crespi, M. (2014b). Noncoding transcription by alternative RNA polymerases dynamically regulates an auxin-driven chromatin loop. *Molecular Cell*, 55(3), 383–396. <https://doi.org/10.1016/j.molcel.2014.06.011>
- Ariel, F., Lucero, L., Christ, A., Mammarella, M. F., Jegu, T., Veluchamy, A., Mariappan, K., Latrasse, D., Blein, T., Liu, C., Benhamed, M., & Crespi, M. (2020). R-Loop Mediated trans Action of the APOLO Long Noncoding RNA. *Molecular Cell*, 77(5), 1055–1065.e4. <https://doi.org/10.1016/j.molcel.2019.12.015>
- Ariel, F., Romero-Barrios, N., Jégú, T., Benhamed, M., & Crespi, M. (2015). Battles and hijacks: noncoding transcription in plants. *Trends in Plant Science*, 20(6), 362–371. <https://doi.org/10.1016/j.tplants.2015.03.003>
- Bailey, T. L., Johnson, J., Grant, C. E., & Noble, W. S. (2015). The MEME Suite. In *Nucleic Acids Research* (Vol. 43, Issue W1, pp. W39–W49). <https://doi.org/10.1093/nar/gkv416>
- Bai, Y., Dai, X., Ye, T., Zhang, P., Yan, X., Gong, X., Liang, S., & Chen, M. (n.d.). PlncRNADB: A Repository of Plant lncRNAs and lncRNA-RBP Protein Interactions.

Current Bioinformatics, 14(7), 621–627. Retrieved October 31, 2022, from <http://www.eurekaselect.com/article/96272>

Bancroft, I., Morgan, C., Fraser, F., Higgins, J., Wells, R., Clissold, L., Baker, D., Long, Y., Meng, J., Wang, X., Liu, S., & Trick, M. (2011). Dissecting the genome of the polyploid crop oilseed rape by transcriptome sequencing. In *Nature Biotechnology* (Vol. 29, Issue 8, pp. 762–766). <https://doi.org/10.1038/nbt.1926>

Bardou, F., Ariel, F., Simpson, C. G., Romero-Barrios, N., Laporte, P., Balzergue, S., Brown, J. W. S., & Crespi, M. (2014). Long noncoding RNA modulates alternative splicing regulators in Arabidopsis. *Developmental Cell*, 30(2), 166–176. <https://doi.org/10.1016/j.devcel.2014.06.017>

Bartolomei, M. S., Zemel, S., & Tilghman, S. M. (1991). Parental imprinting of the mouse H19 gene. In *Nature* (Vol. 351, Issue 6322, pp. 153–155). <https://doi.org/10.1038/351153a0>

Bazin, J., & Bailey-Serres, J. (2015). Emerging roles of long non-coding RNA in root developmental plasticity and regulation of phosphate homeostasis. *Frontiers in Plant Science*, 6, 400. <https://doi.org/10.3389/fpls.2015.00400>

Beilstein, M. A., Al-Shehbaz, I. A., Mathews, S., & Kellogg, E. A. (2008). Brassicaceae phylogeny inferred from phytochrome A and ndhF sequence data: tribes and trichomes revisited. *American Journal of Botany*, 95(10), 1307–1327. <https://doi.org/10.3732/ajb.0800065>

Ben Amor, B., Wirth, S., Merchan, F., Laporte, P., d'Aubenton-Carafa, Y., Hirsch, J., Maizel, A., Mallory, A., Lucas, A., Deragon, J. M., Vaucheret, H., Thermes, C., & Crespi, M. (2009). Novel long non-protein coding RNAs involved in Arabidopsis differentiation and stress responses. *Genome Research*, 19(1), 57–69. <https://doi.org/10.1101/gr.080275.108>

Berardini, T. Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E., & Huala, E. (2015). The arabidopsis information resource: Making and mining the “gold standard” annotated reference plant genome. In *genesis* (Vol. 53, Issue 8, pp. 474–485). <https://doi.org/10.1002/dvg.22877>

Bertani, S., Sauer, S., Bolotin, E., & Sauer, F. (2011). The noncoding RNA Mistral activates Hoxa6 and Hoxa7 expression and stem cell differentiation by recruiting MLL1 to chromatin. *Molecular Cell*, 43(6), 1040–1046. <https://doi.org/10.1016/j.molcel.2011.08.019>

Bhatia, G., Sharma, S., Upadhyay, S. K., & Singh, K. (2019). Long Non-coding RNAs Coordinate Developmental Transitions and Other Key Biological Processes in Grapevine. *Scientific Reports*, 9(1), 3552. <https://doi.org/10.1038/s41598-019-38989-7>

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. In *Bioinformatics* (Vol. 30, Issue 15, pp. 2114–2120). <https://doi.org/10.1093/bioinformatics/btu170>

Bolser, D., Staines, D. M., Pritchard, E., & Kersey, P. (2016). Ensembl Plants: Integrating Tools for Visualizing, Mining, and Analyzing Plant Genomics Data. *Methods in Molecular Biology*, 1374, 115–140. https://doi.org/10.1007/978-1-4939-3167-5_6

Boyes, D. C. (2001). Growth Stage-Based Phenotypic Analysis of Arabidopsis: A Model for High Throughput Functional Genomics in Plants. In *THE PLANT CELL ONLINE* (Vol. 13, Issue 7, pp. 1499–1510). <https://doi.org/10.1105/tpc.13.7.1499>

Brannan, C. I., Dees, E. C., Ingram, R. S., & Tilghman, S. M. (1990). The product of the H19 gene may function as an RNA. In *Molecular and Cellular Biology* (Vol. 10, Issue 1,

pp. 28–36). <https://doi.org/10.1128/mcb.10.1.28>

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5), 525–527. <https://doi.org/10.1038/nbt.3519>

Brown, J. B., Boley, N., Eisman, R., May, G. E., Stoiber, M. H., Duff, M. O., Booth, B. W., Wen, J., Park, S., Suzuki, A. M., Wan, K. H., Yu, C., Zhang, D., Carlson, J. W., Cherbas, L., Eads, B. D., Miller, D., Mockaitis, K., Roberts, J., ... Celniker, S. E. (2014). Diversity and dynamics of the *Drosophila* transcriptome. *Nature*, 512(7515), 393–399. <https://doi.org/10.1038/nature12962>

Buenrostro, J. D., Araya, C. L., Chircus, L. M., Layton, C. J., Chang, H. Y., Snyder, M. P., & Greenleaf, W. J. (2014). Quantitative analysis of RNA-protein interactions on a massively parallel array reveals biophysical and evolutionary landscapes. *Nature Biotechnology*, 32(6), 562–568. <https://doi.org/10.1038/nbt.2880>

Burleigh, S. H., & Harrison, M. J. (1999). The Down-Regulation of *Mt4*-Like Genes by Phosphate Fertilization Occurs Systemically and Involves Phosphate Translocation to the Shoots1. In *Plant Physiology* (Vol. 119, Issue 1, pp. 241–248). <https://doi.org/10.1104/pp.119.1.241>

Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., & Rinn, J. L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & Development*, 25(18), 1915–1927. <https://doi.org/10.1101/gad.17446611>

Calixto, C. P. G., Tzioutziou, N. A., James, A. B., Hornyik, C., Guo, W., Zhang, R., Nimmo, H. G., & Brown, J. W. S. (2019a). Cold-Dependent Expression and Alternative Splicing of Arabidopsis Long Non-coding RNAs. *Frontiers in Plant Science*, 10, 235. <https://doi.org/10.3389/fpls.2019.00235>

Calixto, C. P. G., Tzioutziou, N. A., James, A. B., Hornyik, C., Guo, W., Zhang, R., Nimmo, H. G., & Brown, J. W. S. (2019b). Cold-Dependent Expression and Alternative Splicing of Arabidopsis Long Non-coding RNAs. *Frontiers in Plant Science*, 10, 235. <https://doi.org/10.3389/fpls.2019.00235>

Campalans, A., Kondorosi, A., & Crespi, M. (2004). Enod40, a Short Open Reading Frame-Containing mRNA, Induces Cytoplasmic Localization of a Nuclear RNA Binding Protein in *Medicago truncatula*. *The Plant Cell*, 16(4), 1047–1059. <https://doi.org/10.1105/tpc.019406>

Cesana, M., Cacchiarelli, D., Legnini, I., Santini, T., Sthandier, O., Chinappi, M., Tramontano, A., & Bozzoni, I. (2011). A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell*, 147(2), 358–369. <https://doi.org/10.1016/j.cell.2011.09.028>

Charon, C., Johansson, C., Kondorosi, E., Kondorosi, A., & Crespi, M. (1997). enod40 induces dedifferentiation and division of root cortical cells in legumes. *Proceedings of the National Academy of Sciences of the United States of America*, 94(16), 8901–8906. <https://doi.org/10.1073/pnas.94.16.8901>

Cheng, C.-Y., Krishnakumar, V., Chan, A. P., Thibaud-Nissen, F., Schobel, S., & Town, C. D. (2017). Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *The Plant Journal: For Cell and Molecular Biology*, 89(4), 789–804. <https://doi.org/10.1111/tpj.13415>

Cheng, F., Wu, J., Cai, C., Fu, L., Liang, J., Borm, T., Zhuang, M., Zhang, Y., Zhang, F., Bonnema, G., & Wang, X. (2016). Genome resequencing and comparative variome

analysis in a *Brassica rapa* and *Brassica oleracea* collection. In *Scientific Data* (Vol. 3, Issue 1). <https://doi.org/10.1038/sdata.2016.119>

Chen, H., & Boutros, P. C. (2011). VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics*, 12, 35. <https://doi.org/10.1186/1471-2105-12-35>

Chen, L., & Zhu, Q.-H. (2022). The evolutionary landscape and expression pattern of plant lincRNAs. In *RNA Biology* (Vol. 19, Issue 1, pp. 1190–1207). <https://doi.org/10.1080/15476286.2022.2144609>

Chen, M.-J. M., Chen, L.-K., Lai, Y.-S., Lin, Y.-Y., Wu, D.-C., Tung, Y.-A., Liu, K.-Y., Shih, H.-T., Chen, Y.-J., Lin, Y.-L., Ma, L.-T., Huang, J.-L., Wu, P.-C., Hong, M.-Y., Chu, F.-H., Wu, J.-T., Li, W.-H., & Chen, C.-Y. (2016). Integrating RNA-seq and ChIP-seq data to characterize long non-coding RNAs in *Drosophila melanogaster*. *BMC Genomics*, 17, 220. <https://doi.org/10.1186/s12864-016-2457-0>

Chodroff, R. A., Goodstadt, L., Sirey, T. M., Oliver, P. L., Davies, K. E., Green, E. D., Molnár, Z., & Ponting, C. P. (2010). Long noncoding RNA genes: conservation of sequence and brain expression among diverse amniotes. *Genome Biology*, 11(7), R72. <https://doi.org/10.1186/gb-2010-11-7-r72>

Choi, S.-W., Kim, H.-W., & Nam, J.-W. (2019). The small peptide world in long noncoding RNAs. *Briefings in Bioinformatics*, 20(5), 1853–1864. <https://doi.org/10.1093/bib/bby055>

Compaan, B., Ruttink, T., Albrecht, C., Meeley, R., Bisseling, T., & Franssen, H. (2003). Identification and characterization of a *Zea mays* line carrying a transposon-tagged ENOD40. *Biochimica et Biophysica Acta*, 1629(1-3), 84–91. <https://doi.org/10.1016/j.bbaexp.2003.08.004>

Conway, J. R., Lex, A., & Gehlenborg, N. (2017). UpSetR: an R package for the visualization of intersecting sets and their properties. In *Bioinformatics* (Vol. 33, Issue 18, pp. 2938–2940). <https://doi.org/10.1093/bioinformatics/btx364>

Corona-Gomez, J. A. (n.d.-a). *Conservación de ARNs largos no codificantes en Brasicáceas*.

Corona-Gomez, J. A. (n.d.-b). *Splicing conservation signals in plant long non-coding RNAs*. <https://doi.org/10.26226/morressier.5ebd45acffea6f735881b089>

Corona-Gomez, J. A., Coss-Navarrete, E. L., Garcia-Lopez, I. J., Pérez-Patiño, J. A., & Fernandez-Valverde, S. L. (n.d.). *Transcriptome-guided annotation and functional classification of long non-coding RNAs in Arabidopsis thaliana*. <https://doi.org/10.1101/2022.04.18.488676>

Corona-Gomez, J. A., Garcia-Lopez, I. J., Stadler, P. F., & Fernandez-Valverde, S. L. (2020). Splicing conservation signals in plant long noncoding RNAs. *RNA*, 26(7), 784–793. <https://doi.org/10.1261/rna.074393.119>

Crespi, M. D., Jurkevitch, E., Poiret, M., d'Aubenton-Carafa, Y., Petrovics, G., Kondorosi, E., & Kondorosi, A. (1994). enod40, a gene expressed during nodule organogenesis, codes for a non-translatable RNA involved in plant growth. *The EMBO Journal*, 13(21), 5099–5112. <https://doi.org/10.1002/j.1460-2075.1994.tb06839.x>

Csorba, T., Questa, J. I., Sun, Q., & Dean, C. (2014). Antisense COOLAIR mediates the coordinated switching of chromatin states at FLC during vernalization. *Proceedings of the National Academy of Sciences of the United States of America*, 111(45), 16160–16165. <https://doi.org/10.1073/pnas.1419030111>

Cui, J., Luan, Y., Jiang, N., Bao, H., & Meng, J. (2017). Comparative transcriptome

analysis between resistant and susceptible tomato allows the identification of lncRNA16397 conferring resistance to *Phytophthora infestans* by co-expressing glutaredoxin. In *The Plant Journal* (Vol. 89, Issue 3, pp. 577–589).
<https://doi.org/10.1111/tpj.13408>

Darbellay, F., & Necșulea, A. (2020). Comparative Transcriptomics Analyses across Species, Organs, and Developmental Stages Reveal Functionally Constrained lncRNAs. *Molecular Biology and Evolution*, 37(1), 240–259.
<https://doi.org/10.1093/molbev/msz212>

Deng, P., Liu, S., Nie, X., Weining, S., & Wu, L. (2018). Conservation analysis of long non-coding RNAs in plants. *Science China. Life Sciences*, 61(2), 190–198.
<https://doi.org/10.1007/s11427-017-9174-9>

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D. G., Lagarde, J., Veeravalli, L., Ruan, X., Ruan, Y., Lassmann, T., Carninci, P., Brown, J. B., Lipovich, L., Gonzalez, J. M., ... Guigó, R. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Research*, 22(9), 1775–1789.
<https://doi.org/10.1101/gr.132159.111>

Di, C., Yuan, J., Wu, Y., Li, J., Lin, H., Hu, L., Zhang, T., Qi, Y., Gerstein, M. B., Guo, Y., & Lu, Z. J. (2014). Characterization of stress-responsive lncRNAs in *Arabidopsis thaliana* by integrating expression, epigenetic and structural features. *The Plant Journal: For Cell and Molecular Biology*, 80(5), 848–861. <https://doi.org/10.1111/tpj.12679>

Di Marsico, M., Di Marsico, M., Gallart, A. P., Sanseverino, W., & Cigliano, R. A. (2022). GreenNC 2.0: a comprehensive database of plant long non-coding RNAs. In *Nucleic Acids Research* (Vol. 50, Issue D1, pp. D1442–D1447).
<https://doi.org/10.1093/nar/gkab1014>

Dinger, M. E., Pang, K. C., Mercer, T. R., & Mattick, J. S. (2008). Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Computational Biology*, 4(11), e1000176. <https://doi.org/10.1371/journal.pcbi.1000176>

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1), 15–21. <https://doi.org/10.1093/bioinformatics/bts635>

Du, Q., Wang, K., Zou, C., Xu, C., & Li, W.-X. (2018). The -miR399 Regulatory Module Is Important for Low Phosphate Tolerance in Maize. *Plant Physiology*, 177(4), 1743–1753. <https://doi.org/10.1104/pp.18.00034>

Edgar, R., Domrachev, M., & Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1), 207–210. <https://doi.org/10.1093/nar/30.1.207>

Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, 20(1), 238. <https://doi.org/10.1186/s13059-019-1832-y>

Engreitz, J. M., Haines, J. E., Perez, E. M., Munson, G., Chen, J., Kane, M., McDonel, P. E., Guttman, M., & Lander, E. S. (2016). Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature*, 539(7629), 452–455.
<https://doi.org/10.1038/nature20149>

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19), 3047–3048. <https://doi.org/10.1093/bioinformatics/btw354>

Fedak, H., Palusinska, M., Krzyczmonik, K., Brzezniak, L., Yatusevich, R., Pietras, Z., Kaczanowski, S., & Swiezewski, S. (2016). Control of seed dormancy in *Arabidopsis* by a cis-acting noncoding antisense transcript. *Proceedings of the National Academy of Sciences of the United States of America*, 113(48), E7846–E7855. <https://doi.org/10.1073/pnas.1608827113>

Feisenstein, J. (1989). *PHYLP: Phylogeny Inference Package Version 3.2 Manual*. <https://books.google.com/books/about/PHYLP.html?hl=&id=6G8fzQEACAAJ>

Fernandez-Pozo, N., Metz, T., Chandler, J. O., Gramzow, L., Mérai, Z., Maumus, F., Mittelsten Scheid, O., Theißen, G., Schranz, M. E., Leubner-Metzger, G., & Rensing, S. A. (2021). *Aethionema arabicum* genome annotation using PacBio full-length transcripts provides a valuable resource for seed dormancy and Brassicaceae evolution research. *The Plant Journal: For Cell and Molecular Biology*, 106(1), 275–293. <https://doi.org/10.1111/tpj.15161>

Finn, R. D., Bateman, A., Clements, J., Coghill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J., & Punta, M. (2014). Pfam: the protein families database. In *Nucleic Acids Research* (Vol. 42, Issue D1, pp. D222–D230). <https://doi.org/10.1093/nar/gkt1223>

Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, 39(Web Server issue), W29–W37. <https://doi.org/10.1093/nar/gkr367>

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Systematic Zoology*, 19(2), 99–113. <https://www.ncbi.nlm.nih.gov/pubmed/5449325>

Fitch, W. M. (2000). Homology a personal view on some of the problems. *Trends in Genetics: TIG*, 16(5), 227–231. [https://doi.org/10.1016/s0168-9525\(00\)02005-9](https://doi.org/10.1016/s0168-9525(00)02005-9)

Flemetakis, E., Kavroulakis, N., Quaadvlieg, N. E., Spaink, H. P., Dimou, M., Roussis, A., & Katinakis, P. (2000). *Lotus japonicus* contains two distinct ENOD40 genes that are expressed in symbiotic, nonsymbiotic, and embryonic tissues. *Molecular Plant-Microbe Interactions: MPMI*, 13(9), 987–994. <https://doi.org/10.1094/MPMI.2000.13.9.987>

Franco-Zorrilla, J. M., Valli, A., Todesco, M., Mateos, I., Puga, M. I., Rubio-Somoza, I., Leyva, A., Weigel, D., García, J. A., & Paz-Ares, J. (2007a). Target mimicry provides a new mechanism for regulation of microRNA activity. *Nature Genetics*, 39(8), 1033–1037. <https://doi.org/10.1038/ng2079>

Franco-Zorrilla, J. M., Valli, A., Todesco, M., Mateos, I., Puga, M. I., Rubio-Somoza, I., Leyva, A., Weigel, D., García, J. A., & Paz-Ares, J. (2007b). Target mimicry provides a new mechanism for regulation of microRNA activity. *Nature Genetics*, 39(8), 1033–1037. <https://doi.org/10.1038/ng2079>

Gabaldón, T., & Koonin, E. V. (2013). Functional and evolutionary implications of gene orthology. *Nature Reviews. Genetics*, 14(5), 360–366. <https://doi.org/10.1038/nrg3456>

Gaiti, F., Fernandez-Valverde, S. L., Nakanishi, N., Calcino, A. D., Yanai, I., Tanurdzic, M., & Degnan, B. M. (2015). Dynamic and Widespread lncRNA Expression in a Sponge and the Origin of Animal Complexity. In *Molecular Biology and Evolution* (Vol. 32, Issue 9, pp. 2367–2382). <https://doi.org/10.1093/molbev/msv117>

Ganguly, P., Roy, D., Das, T., Kundu, A., Cartieaux, F., Ghosh, Z., & DasGupta, M. (2021). The Natural Antisense Transcript Derived from the lncRNA Locus Interacts with SET Domain Protein ASHR3 During Inception of Symbiosis in. *Molecular Plant-Microbe Interactions: MPMI*, 34(9), 1057–1070. <https://doi.org/10.1094/MPMI-12-20-0357-R>

Gao, R., Liu, P., Irwanto, N., Loh, D. R., & Wong, S.-M. (2016). Upregulation of LINC-

AP2 is negatively correlated with AP2 gene expression with Turnip crinkle virus infection in *Arabidopsis thaliana*. *Plant Cell Reports*, 35(11), 2257–2267.
<https://doi.org/10.1007/s00299-016-2032-9>

Gehring, A., Friedt, W., Lühs, W., & Snowdon, R. J. (2006). Genetic mapping of agronomic traits in false flax (*Camelina sativa* subsp. *sativa*). In *Genome* (Vol. 49, Issue 12, pp. 1555–1563). <https://doi.org/10.1139/g06-117>

Gelaw, T. A., & Sanan-Mishra, N. (2021). Non-Coding RNAs in Response to Drought Stress. *International Journal of Molecular Sciences*, 22(22), 12519.
<https://doi.org/10.3390/ijms222212519>

Gene Frequently Asked Questions. (2018). In *Gene Help [Internet]*. National Center for Biotechnology Information (US). <https://www.ncbi.nlm.nih.gov/books/NBK3840/>

Gibbs, D. J., Voß, U., Harding, S. A., Fannon, J., Moody, L. A., Yamada, E., Swarup, K., Nibau, C., Bassel, G. W., Choudhary, A., Lavenus, J., Bradshaw, S. J., Stekel, D. J., Bennett, M. J., & Coates, J. C. (2014). AtMYB93 is a novel negative regulator of lateral root development in *Arabidopsis*. *The New Phytologist*, 203(4), 1194–1207.
<https://doi.org/10.1111/nph.12879>

Golicz, A. A., Bhalla, P. L., & Singh, M. B. (2018). lncRNAs in Plant and Animal Sexual Reproduction. *Trends in Plant Science*, 23(3), 195–205.
<https://doi.org/10.1016/j.tplants.2017.12.009>

Golicz, A. A., Singh, M. B., & Bhalla, P. L. (2018). The Long Intergenic Noncoding RNA (lincRNA) Landscape of the Soybean Genome. In *Plant Physiology* (Vol. 176, Issue 3, pp. 2133–2147). <https://doi.org/10.1104/pp.17.01657>

Gong, C., & Maquat, L. E. (2011). lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. *Nature*, 470(7333), 284–288.
<https://doi.org/10.1038/nature09701>

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., & Rokhsar, D. S. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research*, 40(Database issue), D1178–D1186. <https://doi.org/10.1093/nar/gkr944>

Gosai, S. J., Foley, S. W., Wang, D., Silverman, I. M., Selamoglu, N., Nelson, A. D. L., Beilstein, M. A., Daldal, F., Deal, R. B., & Gregory, B. D. (2015). Global analysis of the RNA-protein interaction and RNA secondary structure landscapes of the *Arabidopsis* nucleus. *Molecular Cell*, 57(2), 376–388. <https://doi.org/10.1016/j.molcel.2014.12.004>

Griesmann, M., Chang, Y., Liu, X., Song, Y., Haberer, G., Crook, M. B., Billault-Penneteau, B., Lauressergues, D., Keller, J., Imanishi, L., Roswanjaya, Y. P., Kohlen, W., Pujic, P., Battenberg, K., Alloisio, N., Liang, Y., Hilhorst, H., Salgado, M. G., Hocher, V., ... Cheng, S. (2018). Phylogenomics reveals multiple losses of nitrogen-fixing root nodule symbiosis. *Science*, 361(6398). <https://doi.org/10.1126/science.aat1743>

Griffith, M., Timonin, M., Wong, A. C. E., Gray, G. R., Akhter, S. R., Saldanha, M., Rogers, M. A., Weretilnyk, E. A., & Moffatt, B. (2007). *Thellungiella*: an *Arabidopsis*-related model plant adapted to cold temperatures. *Plant, Cell & Environment*, 30(5), 529–538. <https://doi.org/10.1111/j.1365-3040.2007.01653.x>

Grote, P., & Herrmann, B. G. (2015). Long noncoding RNAs in organogenesis: making the difference. *Trends in Genetics: TIG*, 31(6), 329–335.
<https://doi.org/10.1016/j.tig.2015.02.002>

Gulyaev, A. P., & Roussis, A. (2007). Identification of conserved secondary structures and expansion segments in enod40 RNAs reveals new enod40 homologues in plants.

Nucleic Acids Research, 35(9), 3144–3152. <https://doi.org/10.1093/nar/gkm173>

Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., Feldser, D., Huarte, M., Zuk, O., Carey, B. W., Cassady, J. P., Cabili, M. N., Jaenisch, R., Mikkelsen, T. S., Jacks, T., Hacohen, N., Bernstein, B. E., Kellis, M., Regev, A., Rinn, J. L., & Lander, E. S. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. In *Nature* (Vol. 458, Issue 7235, pp. 223–227). <https://doi.org/10.1038/nature07672>

Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M. J., Gnirke, A., Nusbaum, C., Rinn, J. L., Lander, E. S., & Regev, A. (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnology*, 28(5), 503–510. <https://doi.org/10.1038/nbt.1633>

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., MacManes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C. N., ... Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, 8(8), 1494–1512. <https://doi.org/10.1038/nprot.2013.084>

Han, G., Cheng, C., Zheng, Y., Wang, X., Xu, Y., Wang, W., Zhu, S., & Cheng, B. (2019). Identification of Long Non-Coding RNAs and the Regulatory Network Responsive to Arbuscular Mycorrhizal Fungi Colonization in Maize Roots. *International Journal of Molecular Sciences*, 20(18). <https://doi.org/10.3390/ijms20184491>

Hedges, S. B., Dudley, J., & Kumar, S. (2006). TimeTree: a public knowledge-base of divergence times among organisms. *Bioinformatics*, 22(23), 2971–2972. <https://doi.org/10.1093/bioinformatics/btl505>

He, H., Zhou, Y.-F., Yang, Y.-W., Zhang, Z., Lei, M.-Q., Feng, Y.-Z., Zhang, Y.-C., Chen, Y.-Q., Lian, J.-P., & Yu, Y. (2021). Genome-Wide Analysis Identified a Set of Conserved lncRNAs Associated with Domestication-Related Traits in Rice. *International Journal of Molecular Sciences*, 22(9). <https://doi.org/10.3390/ijms22094742>

Henriques, R., Wang, H., Liu, J., Boix, M., Huang, L.-F., & Chua, N.-H. (2017a). The antiphasic regulatory module comprising CDF5 and its antisense RNA FLORE links the circadian clock to photoperiodic flowering. *The New Phytologist*, 216(3), 854–867. <https://doi.org/10.1111/nph.14703>

Henriques, R., Wang, H., Liu, J., Boix, M., Huang, L.-F., & Chua, N.-H. (2017b). The antiphasic regulatory module comprising CDF5 and its antisense RNA FLORE links the circadian clock to photoperiodic flowering. *The New Phytologist*, 216(3), 854–867. <https://doi.org/10.1111/nph.14703>

Heo, J. B., & Sung, S. (2011). Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. *Science*, 331(6013), 76–79. <https://doi.org/10.1126/science.1197349>

Hezroni, H., Koppstein, D., Schwartz, M. G., Avrutin, A., Bartel, D. P., & Ulitsky, I. (2015). Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Reports*, 11(7), 1110–1122. <https://doi.org/10.1016/j.celrep.2015.04.023>

Hickey, G., Paten, B., Earl, D., Zerbino, D., & Haussler, D. (2013). HAL: a hierarchical format for storing and analyzing multiple genome alignments. In *Bioinformatics* (Vol. 29, Issue 10, pp. 1341–1342). <https://doi.org/10.1093/bioinformatics/btt128>

Hirsch, J., Lefort, V., Vankersschaver, M., Boualem, A., Lucas, A., Thermes, C., d'Aubenton-Carafa, Y., & Crespi, M. (2006). Characterization of 43 non-protein-coding mRNA genes in Arabidopsis, including the MIR162a-derived transcripts. *Plant Physiology*, 140(4), 1192–1204. <https://doi.org/10.1104/pp.105.073817>

Hon, C.-C., Ramilowski, J. A., Harshbarger, J., Bertin, N., Rackham, O. J. L., Gough, J., Denisenko, E., Schmeier, S., Poulsen, T. M., Severin, J., Lizio, M., Kawaji, H., Kasukawa, T., Itoh, M., Burroughs, A. M., Noma, S., Djebali, S., Alam, T., Medvedeva, Y. A., ... Forrest, A. R. R. (2017). An atlas of human long non-coding RNAs with accurate 5' ends. *Nature*, 543(7644), 199–204. <https://doi.org/10.1038/nature21374>

Housman, G., & Ulitsky, I. (2016). Methods for distinguishing between protein-coding and long noncoding RNAs and the elusive biological purpose of translation of long noncoding RNAs. *Biochimica et Biophysica Acta*, 1859(1), 31–40. <https://doi.org/10.1016/j.bbagrm.2015.07.017>

Hsu, P. Y., Calviello, L., Wu, H.-Y. L., Li, F.-W., Rothfels, C. J., Ohler, U., & Benfey, P. N. (2016). Super-resolution ribosome profiling reveals unannotated translation events in. *Proceedings of the National Academy of Sciences of the United States of America*, 113(45), E7126–E7135. <https://doi.org/10.1073/pnas.1614788113>

Huang, C. Y., Shirley, N., Genc, Y., Shi, B., & Langridge, P. (2011). Phosphate utilization efficiency correlates with expression of low-affinity phosphate transporters and noncoding RNA, IPS1, in barley. *Plant Physiology*, 156(3), 1217–1229. <https://doi.org/10.1104/pp.111.178459>

Huang, X., Zhang, H., Wang, Q., Guo, R., Wei, L., Song, H., Kuang, W., Liao, J., Huang, Y., & Wang, Z. (2021). Genome-wide identification and characterization of long non-coding RNAs involved in flag leaf senescence of rice. *Plant Molecular Biology*, 105(6), 655–684. <https://doi.org/10.1007/s11103-021-01121-3>

Hu, W., Alvarez-Dominguez, J. R., & Lodish, H. F. (2012). Regulation of mammalian cell differentiation by long non-coding RNAs. In *EMBO reports* (Vol. 13, Issue 11, pp. 971–983). <https://doi.org/10.1038/embor.2012.145>

Ilik, I. A., Quinn, J. J., Georgiev, P., Tavares-Cadete, F., Maticzka, D., Toscano, S., Wan, Y., Spitale, R. C., Luscombe, N., Backofen, R., Chang, H. Y., & Akhtar, A. (2013). Tandem stem-loops in roX RNAs act together to mediate X chromosome dosage compensation in Drosophila. *Molecular Cell*, 51(2), 156–173. <https://doi.org/10.1016/j.molcel.2013.07.001>

Ingram, V. M. (1961). Gene Evolution and the Hæmoglobins. In *Nature* (Vol. 189, Issue 4766, pp. 704–708). <https://doi.org/10.1038/189704a0>

Iyer, M. K., Niknafs, Y. S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T. R., Prensner, J. R., Evans, J. R., Zhao, S., Poliakov, A., Cao, X., Dhanasekaran, S. M., Wu, Y.-M., Robinson, D. R., Beer, D. G., Feng, F. Y., Iyer, H. K., & Chinnaiyan, A. M. (2015). The landscape of long noncoding RNAs in the human transcriptome. *Nature Genetics*, 47(3), 199–208. <https://doi.org/10.1038/ng.3192>

Jiao, Y., Wickett, N. J., Ayyampalayam, S., Chanderbali, A. S., Landherr, L., Ralph, P. E., Tomsho, L. P., Hu, Y., Liang, H., Soltis, P. S., Soltis, D. E., Clifton, S. W., Schlarbaum, S. E., Schuster, S. C., Ma, H., Leebens-Mack, J., & dePamphilis, C. W. (2011). Ancestral polyploidy in seed plants and angiosperms. *Nature*, 473(7345), 97–100. <https://doi.org/10.1038/nature09916>

Jin, J., Liu, J., Wang, H., Wong, L., & Chua, N.-H. (2013). PLncDB: plant long non-coding RNA database. *Bioinformatics*, 29(8), 1068–1071.

<https://doi.org/10.1093/bioinformatics/btt107>

Jin, J., Lu, P., Xu, Y., Li, Z., Yu, S., Liu, J., Wang, H., Chua, N.-H., & Cao, P. (2020). PLncDB V2.0: a comprehensive encyclopedia of plant long noncoding RNAs. *Nucleic Acids Research*, 49(D1), D1489–D1495. <https://doi.org/10.1093/nar/gkaa910>

Jin, Y., Ivanov, M., Dittrich, A. N., Nelson, A. D. L., & Marquardt, S. (2021). A trans-acting long non-coding RNA represses flowering in Arabidopsis. In *bioRxiv* (p. 2021.11.15.468639). <https://doi.org/10.1101/2021.11.15.468639>

Ji, Z., Song, R., Regev, A., & Struhl, K. (2015). Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *eLife*, 4, e08890. <https://doi.org/10.7554/eLife.08890>

Kagale, S., Koh, C., Nixon, J., Bollina, V., Clarke, W. E., Tuteja, R., Spillane, C., Robinson, S. J., Links, M. G., Clarke, C., Higgins, E. E., Huebert, T., Sharpe, A. G., & Parkin, I. A. P. (2014). The emerging biofuel crop *Camelina sativa* retains a highly undifferentiated hexaploid genome structure. *Nature Communications*, 5, 3706. <https://doi.org/10.1038/ncomms4706>

Kagale, S., Robinson, S. J., Nixon, J., Xiao, R., Huebert, T., Condie, J., Kessler, D., Clarke, W. E., Edger, P. P., Links, M. G., Sharpe, A. G., & Parkin, I. A. P. (2014). Polyploid evolution of the Brassicaceae during the Cenozoic era. *The Plant Cell*, 26(7), 2777–2791. <https://doi.org/10.1105/tpc.114.126391>

Kapranov, P., Cheng, J., Dike, S., Nix, D. A., Dutttagupta, R., Willingham, A. T., Stadler, P. F., Hertel, J., Hackermüller, J., Hofacker, I. L., Bell, I., Cheung, E., Drenkow, J., Dumais, E., Patel, S., Helt, G., Ganesh, M., Ghosh, S., Piccolboni, A., ... Gingeras, T. R. (2007). RNA Maps Reveal New RNA Classes and a Possible Function for Pervasive Transcription. In *Science* (Vol. 316, Issue 5830, pp. 1484–1488). <https://doi.org/10.1126/science.1138341>

Karreth, F. A., Tay, Y., Perna, D., Ala, U., Tan, S. M., Rust, A. G., DeNicola, G., Webster, K. A., Weiss, D., Perez-Mancera, P. A., Krauthammer, M., Halaban, R., Provero, P., Adams, D. J., Tuveson, D. A., & Pandolfi, P. P. (2011). In Vivo Identification of Tumor-Suppressive PTEN ceRNAs in an Oncogenic BRAF-Induced Mouse Model of Melanoma. In *Cell* (Vol. 147, Issue 4, p. 948). <https://doi.org/10.1016/j.cell.2011.10.032>

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC. *Genome Research*, 12(6), 996–1006. <https://doi.org/10.1101/gr.229102>

Kern, C., Wang, Y., Chitwood, J., Korf, I., Delany, M., Cheng, H., Medrano, J. F., Van Eenennaam, A. L., Ernst, C., Ross, P., & Zhou, H. (2018). Genome-wide identification of tissue-specific long non-coding RNA in three farm animal species. *BMC Genomics*, 19(1), 684. <https://doi.org/10.1186/s12864-018-5037-7>

Kim, D.-H., Xi, Y., & Sung, S. (2017). Modular function of long noncoding RNA, COLDAIR, in the vernalization response. *PLoS Genetics*, 13(7), e1006939. <https://doi.org/10.1371/journal.pgen.1006939>

Kindgren, P., Ard, R., Ivanov, M., & Marquardt, S. (2018). Transcriptional read-through of the long non-coding RNA SVALKa governs plant cold acclimation. In *Nature Communications* (Vol. 9, Issue 1). <https://doi.org/10.1038/s41467-018-07010-6>

Kino, T., Hurt, D. E., Ichijo, T., Nader, N., & Chrousos, G. P. (2010). Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor. *Science Signaling*, 3(107), ra8. <https://doi.org/10.1126/scisignal.2000568>

Koch, M. A., & Kiefer, M. (2005). Genome evolution among cruciferous plants: a lecture

from the comparison of the genetic maps of three diploid species--*Capsella rubella*, *Arabidopsis lyrata* subsp. *petraea*, and *A. thaliana*. In *American Journal of Botany* (Vol. 92, Issue 4, pp. 761–767). <https://doi.org/10.3732/ajb.92.4.761>

Kohnen, M. V., Schmid-Siebert, E., Trevisan, M., Petrolati, L. A., Sénéchal, F., Müller-Moulé, P., Maloof, J., Xenarios, I., & Fankhauser, C. (2016). Neighbor Detection Induces Organ-Specific Transcriptomes, Revealing Patterns Underlying Hypocotyl-Specific Growth. *The Plant Cell*, 28(12), 2889–2904. <https://doi.org/10.1105/tpc.16.00463>

Kong, L., Zhang, Y., Ye, Z.-Q., Liu, X.-Q., Zhao, S.-Q., Wei, L., & Gao, G. (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Research*, 35(Web Server issue), W345–W349. <https://doi.org/10.1093/nar/gkm391>

Koonin, E. V. (2005). Orthologs, Paralogs, and Evolutionary Genomics. In *Annual Review of Genetics* (Vol. 39, Issue 1, pp. 309–338). <https://doi.org/10.1146/annurev.genet.39.073003.114725>

Kouchi, H., & Hata, S. (1993). Isolation and characterization of novel nodulin cDNAs representing genes expressed at early stages of soybean nodule development. *Molecular & General Genetics: MGG*, 238(1-2), 106–119. <https://doi.org/10.1007/BF00279537>

Kouchi, H., Takane, K., So, R. B., Ladha, J. K., & Reddy, P. M. (1999). Rice ENOD40: isolation and expression analysis in rice and transgenic soybean root nodules. *The Plant Journal: For Cell and Molecular Biology*, 18(2), 121–129. <https://doi.org/10.1046/j.1365-313x.1999.00432.x>

Kramer, M. C., Kim, H. J., Palos, K. R., Garcia, B. A., Lyons, E., Beilstein, M. A., Nelson, A. D. L., & Gregory, B. D. (2022). A Conserved Long Intergenic Non-coding RNA Containing snoRNA Sequences, , Affects Arabidopsis Germination and Development. *Frontiers in Plant Science*, 13, 906603. <https://doi.org/10.3389/fpls.2022.906603>

Kumar, S., Stecher, G., Suleski, M., & Hedges, S. B. (2017). TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Molecular Biology and Evolution*, 34(7), 1812–1819. <https://doi.org/10.1093/molbev/msx116>

Kumpf, R., Thorstensen, T., Rahman, M. A., Heyman, J., Nenseth, H. Z., Lammens, T., Herrmann, U., Swarup, R., Veiseth, S. V., Emberland, G., Bennett, M. J., De Veylder, L., & Aalen, R. B. (2014). The ASH1-RELATED3 SET-domain protein controls cell division competence of the meristem and the quiescent center of the Arabidopsis primary root. *Plant Physiology*, 166(2), 632–643. <https://doi.org/10.1104/pp.114.244798>

Kung, J. T. Y., Colognori, D., & Lee, J. T. (2013). Long noncoding RNAs: past, present, and future. *Genetics*, 193(3), 651–669. <https://doi.org/10.1534/genetics.112.146704>

Ladomery, M., & Pisignano, G. (2021). *Post-transcriptional Regulation through Long Non-coding RNAs (lncRNAs)*. MDPI. https://books.google.com/books/about/Post_transcriptional_Regulation_through.html?hl=&id=qUY-EAAQBAJ

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. In *BMC Bioinformatics* (Vol. 9, Issue 1). <https://doi.org/10.1186/1471-2105-9-559>

Lavin, M., Herendeen, P. S., & Wojciechowski, M. F. (2005). Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the tertiary.

Systematic Biology, 54(4), 575–594. <https://doi.org/10.1080/10635150590947131>

Lee, S.-Y., Hwang, E. Y., Seok, H.-Y., Tarte, V. N., Jeong, M. S., Jang, S. B., & Moon, Y.-H. (2015). Arabidopsis AtERF71/HRE2 functions as transcriptional activator via cis-acting GCC box or DRE/CRT element and is involved in root development through regulation of root cell expansion. *Plant Cell Reports*, 34(2), 223–231. <https://doi.org/10.1007/s00299-014-1701-9>

Leighton, P. A., Ingram, R. S., Eggenschwiler, J., Efstratiadis, A., & Tilghman, S. M. (1995). Disruption of imprinting caused by deletion of the H19 gene region in mice. In *Nature* (Vol. 375, Issue 6526, pp. 34–39). <https://doi.org/10.1038/375034a0>

Li, L., Eichten, S. R., Shimizu, R., Petsch, K., Yeh, C.-T., Wu, W., Chetoor, A. M., Givan, S. A., Cole, R. A., Fowler, J. E., Evans, M. M. S., Scanlon, M. J., Yu, J., Schnable, P. S., Timmermans, M. C. P., Springer, N. M., & Muehlbauer, G. J. (2014). Genome-wide discovery and characterization of maize long non-coding RNAs. *Genome Biology*, 15(2), R40. <https://doi.org/10.1186/gb-2014-15-2-r40>

Liscum, E., & Reed, J. W. (2002). Genetics of Aux/IAA and ARF action in plant growth and development. *Plant Molecular Biology*, 49(3-4), 387–400. <https://www.ncbi.nlm.nih.gov/pubmed/12036262>

Little, D. Y., Rao, H., Oliva, S., Daniel-Vedele, F., Krapp, A., & Malamy, J. E. (2005). The putative high-affinity nitrate transporter NRT2.1 represses lateral root initiation in response to nutritional cues. *Proceedings of the National Academy of Sciences of the United States of America*, 102(38), 13693–13698. <https://doi.org/10.1073/pnas.0504219102>

Liu, C., Muchhal, U. S., & Raghothama, K. G. (1997). Differential expression of TPS11, a phosphate starvation-induced gene in tomato. *Plant Molecular Biology*, 33(5), 867–874. <https://doi.org/10.1023/a:1005729309569>

Liu, D., Yu, X., Wang, S., Dai, E., Jiang, L., Wang, J., Yang, Q., Yang, F., Zhou, S., & Jiang, W. (2016). The gain and loss of long noncoding RNA associated-competing endogenous RNAs in prostate cancer. *Oncotarget*, 7(35), 57228. <https://doi.org/10.18632/oncotarget.11128>

Liu, F., Shi, T., Qi, L., Su, X., Wang, D., Dong, J., & Huang, Z. Y. (2019). lncRNA profile of *Apis mellifera* and its possible role in behavioural transition from nurses to foragers. *BMC Genomics*, 20(1), 393. <https://doi.org/10.1186/s12864-019-5664-7>

Liu, G., Liu, F., Wang, Y., & Liu, X. (2022). A novel long noncoding RNA CIL1 enhances cold stress tolerance in Arabidopsis. In *Plant Science* (Vol. 323, p. 111370). <https://doi.org/10.1016/j.plantsci.2022.111370>

Liu, J., Jung, C., Xu, J., Wang, H., Deng, S., Bernad, L., Arenas-Huertero, C., & Chua, N.-H. (2012). Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in Arabidopsis. *The Plant Cell*, 24(11), 4333–4345. <https://doi.org/10.1105/tpc.112.102855>

Liu, S., Liu, Y., Yang, X., Tong, C., Edwards, D., Parkin, I. A. P., Zhao, M., Ma, J., Yu, J., Huang, S., Wang, X., Wang, J., Lu, K., Fang, Z., Bancroft, I., Yang, T.-J., Hu, Q., Wang, X., Yue, Z., ... Paterson, A. H. (2014). The Brassica oleracea genome reveals the asymmetrical evolution of polyploid genomes. *Nature Communications*, 5, 3930. <https://doi.org/10.1038/ncomms4930>

Liu, S., Sun, Z., & Xu, M. (2018). Identification and characterization of long non-coding RNAs involved in the formation and development of poplar adventitious roots. In *Industrial Crops and Products* (Vol. 118, pp. 334–346).

<https://doi.org/10.1016/j.indcrop.2018.03.071>
 LncRNA/circRNA–miRNA–mRNA networks regulate the development of root and shoot meristems of *Populus*. (2019). *Industrial Crops and Products*, 133, 333–347.
<https://doi.org/10.1016/j.indcrop.2019.03.048>
 Love, M. I., Huber, W., & Anders, S. (2014a). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), 550.
<https://doi.org/10.1186/s13059-014-0550-8>
 Love, M. I., Huber, W., & Anders, S. (2014b). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), 550.
<https://doi.org/10.1186/s13059-014-0550-8>
 Ma, L., Bajic, V. B., & Zhang, Z. (2013). On the classification of long non-coding RNAs. *RNA Biology*, 10(6), 925–933. <https://doi.org/10.4161/rna.24604>
 Martin, A. C., del Pozo, J. C., Iglesias, J., Rubio, V., Solano, R., de la Pena, A., Leyva, A., & Paz-Ares, J. (2000). Influence of cytokinins on the expression of phosphate starvation responsive genes in *Arabidopsis*. In *The Plant Journal* (Vol. 24, Issue 5, pp. 559–567). <https://doi.org/10.1046/j.1365-313x.2000.00893.x>
 Matsumoto, A., Pasut, A., Matsumoto, M., Yamashita, R., Fung, J., Monteleone, E., Saghatelian, A., Nakayama, K. I., Clohessy, J. G., & Pandolfi, P. P. (2017). mTORC1 and muscle regeneration are regulated by the LINC00961-encoded SPAR polypeptide. *Nature*, 541(7636), 228–232. <https://doi.org/10.1038/nature21034>
 Mattick, J. S. (2011). The central role of RNA in human development and cognition. *FEBS Letters*, 585(11), 1600–1616. <https://doi.org/10.1016/j.febslet.2011.05.001>
 Mattick, J. S., & Rinn, J. L. (2015). Discovery and annotation of long noncoding RNAs. In *Nature Structural & Molecular Biology* (Vol. 22, Issue 1, pp. 5–7). <https://doi.org/10.1038/nsmb.2942>
 Matzke, M. A., & Mosher, R. A. (2014). RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. *Nature Reviews. Genetics*, 15(6), 394–408. <https://doi.org/10.1038/nrg3683>
 Mercer, T. R., Dinger, M. E., Sunkin, S. M., Mehler, M. F., & Mattick, J. S. (2008). Specific expression of long noncoding RNAs in the mouse brain. In *Proceedings of the National Academy of Sciences* (Vol. 105, Issue 2, pp. 716–721). <https://doi.org/10.1073/pnas.0706729105>
 Mercer, T. R., & Mattick, J. S. (2013). Structure and function of long noncoding RNAs in epigenetic regulation. *Nature Structural & Molecular Biology*, 20(3), 300–307. <https://doi.org/10.1038/nsmb.2480>
 Mohammadin, S., Edger, P. P., Pires, J. C., & Schranz, M. E. (2015). Positionally-conserved but sequence-diverged: identification of long non-coding RNAs in the Brassicaceae and Cleomaceae. *BMC Plant Biology*, 15, 217. <https://doi.org/10.1186/s12870-015-0603-5>
 Moison, M., Pacheco, J. M., Lucero, L., Fonouni-Farde, C., Rodríguez-Melo, J., Mansilla, N., Christ, A., Bazin, J., Benhamed, M., Ibañez, F., Crespi, M., Estevez, J. M., & Ariel, F. (2021). The lncRNA APOLO interacts with the transcription factor WRKY42 to trigger root hair cell expansion in response to cold. *Molecular Plant*, 14(6), 937–948. <https://doi.org/10.1016/j.molp.2021.03.008>
 Moriwaki, T., Miyazawa, Y., Kobayashi, A., Uchida, M., Watanabe, C., Fujii, N., & Takahashi, H. (2011). Hormonal regulation of lateral root development in *Arabidopsis* modulated by MIZ1 and requirement of GNOM activity for MIZ1 function. *Plant*

Physiology, 157(3), 1209–1220. <https://doi.org/10.1104/pp.111.186270>

Moser, B. R. (2012). Biodiesel from alternative oilseed feedstocks: camelina and field pennycress. In *Biofuels* (Vol. 3, Issue 2, pp. 193–209). <https://doi.org/10.4155/bfs.12.6>

Murat, F., Louis, A., Maumus, F., Armero, A., Cooke, R., Quesneville, H., Roest Crollius, H., & Salse, J. (2015). Understanding Brassicaceae evolution through ancestral genome reconstruction. *Genome Biology*, 16, 262. <https://doi.org/10.1186/s13059-015-0814-y>

Nawrocki, E. P., Burge, S. W., Bateman, A., Daub, J., Eberhardt, R. Y., Eddy, S. R., Floden, E. W., Gardner, P. P., Jones, T. A., Tate, J., & Finn, R. D. (2015). Rfam 12.0: updates to the RNA families database. In *Nucleic Acids Research* (Vol. 43, Issue D1, pp. D130–D137). <https://doi.org/10.1093/nar/gku1063>

Nawrocki, E. P., & Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, 29(22), 2933–2935. <https://doi.org/10.1093/bioinformatics/btt509>

Necsulea, A., Soumillon, M., Warnefors, M., Liechti, A., Daish, T., Zeller, U., Baker, J. C., Grützner, F., & Kaessmann, H. (2014). The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature*, 505(7485), 635–640. <https://doi.org/10.1038/nature12943>

Nelson, A. D. L., Forsythe, E. S., Devisetty, U. K., Clausen, D. S., Haug-Batzell, A. K., Meldrum, A. M. R., Frank, M. R., Lyons, E., & Beilstein, M. A. (2016). A Genomic Analysis of Factors Driving lincRNA Diversification: Lessons from Plants. *G3 Genes|Genomes|Genetics*, 6(9), 2881–2891. <https://doi.org/10.1534/g3.116.030338>

Ng, S.-Y., Johnson, R., & Stanton, L. W. (2012). Human long non-coding RNAs promote pluripotency and neuronal differentiation by association with chromatin modifiers and transcription factors. *The EMBO Journal*, 31(3), 522–533. <https://doi.org/10.1038/emboj.2011.459>

Nordin, M., Bergman, D., Halje, M., Engström, W., & Ward, A. (2014). Epigenetic regulation of the Igf2/H19 gene cluster. *Cell Proliferation*, 47(3), 189–199. <https://doi.org/10.1111/cpr.12106>

Nystrom, S. L., & McKay, D. J. (2021). Memes: A motif analysis environment in R using tools from the MEME Suite. In *PLOS Computational Biology* (Vol. 17, Issue 9, p. e1008991). <https://doi.org/10.1371/journal.pcbi.1008991>

Oh, D.-H., Dassanayake, M., Haas, J. S., Kropornika, A., Wright, C., d'Urzo, M. P., Hong, H., Ali, S., Hernandez, A., Lambert, G. M., Inan, G., Galbraith, D. W., Bressan, R. A., Yun, D.-J., Zhu, J.-K., Cheeseman, J. M., & Bohnert, H. J. (2010). Genome structures and halophyte-specific gene expression of the extremophile *Thellungiella parvula* in comparison with *Thellungiella salsuginea* (*Thellungiella halophila*) and *Arabidopsis*. *Plant Physiology*, 154(3), 1040–1052. <https://doi.org/10.1104/pp.110.163923>

Oono, Y., Kawahara, Y., Yazawa, T., Kanamori, H., Kuramata, M., Yamagata, H., Hosokawa, S., Minami, H., Ishikawa, S., Wu, J., Antonio, B., Handa, H., Itoh, T., & Matsumoto, T. (2013). Diversity in the complexity of phosphate starvation transcriptomes among rice cultivars based on RNA-Seq profiles. *Plant Molecular Biology*, 83(6), 523–537. <https://doi.org/10.1007/s11103-013-0106-4>

Pachnis, V., Belayew, A., & Tilghman, S. M. (1984). Locus unlinked to alpha-fetoprotein under the control of the murine raf and Rif genes. *Proceedings of the National Academy of Sciences of the United States of America*, 81(17), 5523–5527. <https://doi.org/10.1073/pnas.81.17.5523>

Palos, K., Nelson Dittrich, A. C., Yu, L., Yang, Brock, J. R., Railey, C. E., Wu, H.-Y. L., Sokolowska, E., Skirycz, A., Hsu, P. Y., Gregory, B. D., Lyons, E., Beilstein, M. A., & Nelson, A. D. L. (2022). Identification and functional annotation of long intergenic non-coding RNAs in Brassicaceae. *The Plant Cell*, 34(9), 3233–3260. <https://doi.org/10.1093/plcell/koac166>

Park, S.-W., Kang, Y. I., Sypula, J. G., Choi, J., Oh, H., & Park, Y. (2007). An evolutionarily conserved domain of roX2 RNA is sufficient for induction of H4-Lys16 acetylation on the Drosophila X chromosome. *Genetics*, 177(3), 1429–1437. <https://doi.org/10.1534/genetics.107.071001>

Paule, M. R. (2000). SURVEY AND SUMMARY Transcription by RNA polymerases I and III. In *Nucleic Acids Research* (Vol. 28, Issue 6, pp. 1283–1298). <https://doi.org/10.1093/nar/28.6.1283>

Paytuví Gallart, A., Hermoso Pulido, A., Anzar Martínez de Lagrán, I., Sanseverino, W., & Aiese Cigliano, R. (2016). GREENC: a Wiki-based database of plant lncRNAs. *Nucleic Acids Research*, 44(D1), D1161–D1166. <https://doi.org/10.1093/nar/gkv1215>

Pertea, G., & Pertea, M. (2020). GFF Utilities: GffRead and GffCompare. In *F1000Research* (Vol. 9, p. 304). <https://doi.org/10.12688/f1000research.23297.2>

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, 33(3), 290–295. <https://doi.org/10.1038/nbt.3122>

Petersen, T. N., Brunak, S., von Heijne, G., & Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. In *Nature Methods* (Vol. 8, Issue 10, pp. 785–786). <https://doi.org/10.1038/nmeth.1701>

Petryszak, R., Keays, M., Amy Tang, Y., Fonseca, N. A., Barrera, E., Burdett, T., Füllgrabe, A., Fuentes, A. M.-P., Jupp, S., Koskinen, S., Mannion, O., Huerta, L., Megy, K., Snow, C., Williams, E., Barzine, M., Hastings, E., Weisser, H., Wright, J., ... Brazma, A. (2016). Expression Atlas update—an integrated database of gene and protein expression in humans, animals and plants. In *Nucleic Acids Research* (Vol. 44, Issue D1, pp. D746–D752). <https://doi.org/10.1093/nar/gkv1045>

Plants of the world online. (n.d.). Plants of the World Online. Retrieved November 22, 2022, from <https://powo.science.kew.org/>

Podkowinski, J., Zmienko, A., Florek, B., Wojciechowski, P., Rybarczyk, A., Wrzesinski, J., Ciesiolka, J., Blazewicz, J., Kondorosi, A., Crespi, M., & Legocki, A. (2009). Translational and structural analysis of the shortest legume ENOD40 gene in *Lupinus luteus*. *Acta Biochimica Polonica*, 56(1), 89–102. <https://www.ncbi.nlm.nih.gov/pubmed/19287803>

Ponting, C. P., Oliver, P. L., & Reik, W. (2009). Evolution and Functions of Long Noncoding RNAs. In *Cell* (Vol. 136, Issue 4, pp. 629–641). <https://doi.org/10.1016/j.cell.2009.02.006>

Pruitt, K. D. (2004). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. In *Nucleic Acids Research* (Vol. 33, Issue Database issue, pp. D501–D504). <https://doi.org/10.1093/nar/gki025>

Qin, T., Zhao, H., Cui, P., Albeshier, N., & Xiong, L. (2017a). A Nucleus-Localized Long Non-Coding RNA Enhances Drought and Salt Stress Tolerance. *Plant Physiology*, 175(3), 1321–1336. <https://doi.org/10.1104/pp.17.00574>

Qin, T., Zhao, H., Cui, P., Albeshier, N., & Xiong, L. (2017b). A Nucleus-Localized Long Non-Coding RNA Enhances Drought and Salt Stress Tolerance. *Plant Physiology*,

175(3), 1321–1336. <https://doi.org/10.1104/pp.17.00574>

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6), 841–842. <https://doi.org/10.1093/bioinformatics/btq033>

Quinn, J. J., Zhang, Q. C., Georgiev, P., Ilik, I. A., Akhtar, A., & Chang, H. Y. (2016). Rapid evolutionary turnover underlies conserved lncRNA-genome interactions. *Genes & Development*, 30(2), 191–207. <https://doi.org/10.1101/gad.272187.115>

Ramírez-Colmenero, A., Oktaba, K., & Fernandez-Valverde, S. L. (2020). Evolution of Genome-Organizing Long Non-coding RNAs in Metazoans. In *Frontiers in Genetics* (Vol. 11). <https://doi.org/10.3389/fgene.2020.589697>

Ran, X., Zhao, F., Wang, Y., Liu, J., Zhuang, Y., Ye, L., Qi, M., Cheng, J., & Zhang, Y. (2020). Plant Regulomics: a data-driven interface for retrieving upstream regulators from plant multi-omics data. *The Plant Journal: For Cell and Molecular Biology*, 101(1), 237–248. <https://doi.org/10.1111/tpj.14526>

Rice, P., Longden, I., & Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends in Genetics: TIG*, 16(6), 276–277. [https://doi.org/10.1016/s0168-9525\(00\)02024-2](https://doi.org/10.1016/s0168-9525(00)02024-2)

Rinn, J. L., Kertesz, M., Wang, J. K., Squazzo, S. L., Xu, X., Bruggmann, S. A., Goodnough, L. H., Helms, J. A., Farnham, P. J., Segal, E., & Chang, H. Y. (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell*, 129(7), 1311–1323. <https://doi.org/10.1016/j.cell.2007.05.022>

Rohrig, H., Schmidt, J., Miklashevichs, E., Schell, J., & John, M. (2002). Soybean ENOD40 encodes two peptides that bind to sucrose synthase. *Proceedings of the National Academy of Sciences of the United States of America*, 99(4), 1915–1920. <https://doi.org/10.1073/pnas.022664799>

Rosa, S., Duncan, S., & Dean, C. (2016). Mutually exclusive sense-antisense transcription at FLC facilitates environmentally induced gene repression. *Nature Communications*, 7, 13031. <https://doi.org/10.1038/ncomms13031>

Ross, C. J., & Ulitsky, I. (2022). Discovering functional motifs in long noncoding RNAs. *Wiley Interdisciplinary Reviews. RNA*, 13(4), e1708. <https://doi.org/10.1002/wrna.1708>

Roulé, T., Ariel, F., Hartmann, C., Hussain, N., Benhamed, M., Gutierrez-Marcos, J., Crespi, M., & Blein, T. (2022). The lncRNA MARS modulates the epigenetic reprogramming of the maternal cluster in response to ABA. *Mol Plant*. <https://doi.org/10.1101/2020.08.10.236562>

Roulé, T., Christ, A., Hussain, N., Huang, Y., Hartmann, C., Benhamed, M., Gutierrez-Marcos, J., Ariel, F., Crespi, M., & Blein, T. (2022). The lncRNA MARS modulates the epigenetic reprogramming of the maternal cluster in response to ABA. *Molecular Plant*, 15(5), 840–856. <https://doi.org/10.1016/j.molp.2022.02.007>

Rymarquis, L. A., Kastenmayer, J. P., Hüttenhofer, A. G., & Green, P. J. (2008). Diamonds in the rough: mRNA-like non-coding RNAs. In *Trends in Plant Science* (Vol. 13, Issue 7, pp. 329–334). <https://doi.org/10.1016/j.tplants.2008.02.009>

Salmena, L., Poliseno, L., Tay, Y., Kats, L., & Pandolfi, P. P. (2011). A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell*, 146(3), 353–358. <https://doi.org/10.1016/j.cell.2011.07.014>

Sánchez-Calderón, L., López-Bucio, J., Chacón-López, A., Gutiérrez-Ortega, A., Hernández-Abreu, E., & Herrera-Estrella, L. (2006). Characterization of low phosphorus insensitive mutants reveals a crosstalk between low phosphorus-induced determinate

root development and the activation of genes involved in the adaptation of Arabidopsis to phosphorus deficiency. *Plant Physiology*, 140(3), 879–889.
<https://doi.org/10.1104/pp.105.073825>

Sande, K. van de, van de Sande, K., Pawlowski, K., Czaja, I., Wieneke, U., Schell, J., Schmidt, J., Walden, R., Matvienko, M., Wellink, J., van Kammen, A., Franssen, H., & Bisseling, T. (1996). Modification of Phytohormone Response by a Peptide Encoded by **ENOD40** of Legumes and a Nonlegume. In *Science* (Vol. 273, Issue 5273, pp. 370–373). <https://doi.org/10.1126/science.273.5273.370>

Sang, S., Chen, W., Zhang, D., Zhang, X., Yang, W., & Liu, C. (2021). Data integration and evolutionary analysis of long non-coding RNAs in 25 flowering plants. *BMC Genomics*, 22(Suppl 3), 739. <https://doi.org/10.1186/s12864-021-08047-6>

Sarropoulos, I., Marin, R., Cardoso-Moreira, M., & Kaessmann, H. (2019). Developmental dynamics of lncRNAs across mammalian organs and species. *Nature*, 571(7766), 510–514. <https://doi.org/10.1038/s41586-019-1341-x>

Satbhai, S. B., Setzer, C., Freynschlag, F., Slovak, R., Kerdaffrec, E., & Busch, W. (2017). Natural allelic variation of FRO2 modulates Arabidopsis root growth under iron deficiency. *Nature Communications*, 8, 15603. <https://doi.org/10.1038/ncomms15603>

Schranz, M. E., Lysak, M. A., & Mitchell-Olds, T. (2006). The ABC's of comparative genomics in the Brassicaceae: building blocks of crucifer genomes. *Trends in Plant Science*, 11(11), 535–542. <https://doi.org/10.1016/j.tplants.2006.09.002>

Seo, J. S., Sun, H.-X., Park, B. S., Huang, C.-H., Yeh, S.-D., Jung, C., & Chua, N.-H. (2017a). ELF18-INDUCED LONG-NONCODING RNA Associates with Mediator to Enhance Expression of Innate Immune Response Genes in Arabidopsis. *The Plant Cell*, 29(5), 1024–1038. <https://doi.org/10.1105/tpc.16.00886>

Seo, J. S., Sun, H.-X., Park, B. S., Huang, C.-H., Yeh, S.-D., Jung, C., & Chua, N.-H. (2017b). ELF18-INDUCED LONG-NONCODING RNA Associates with Mediator to Enhance Expression of Innate Immune Response Genes in Arabidopsis. *The Plant Cell*, 29(5), 1024–1038. <https://doi.org/10.1105/tpc.16.00886>

Serdyuk, O., Trubina, V., & Gorlova, L. (2022). Comparative evaluation of diseases affection of winter and spring false flax (*Camelina sativa* (L.) Crantz) in the Krasnodar region. In *IOP Conference Series: Earth and Environmental Science* (Vol. 949, Issue 1, p. 012105). <https://doi.org/10.1088/1755-1315/949/1/012105>

Severing, E., Faino, L., Jamge, S., Busscher, M., Kuijer-Zhang, Y., Bellinazzo, F., Busscher-Lange, J., Fernández, V., Angenent, G. C., Immink, R. G. H., & Pajoro, A. (2018). Arabidopsis thaliana ambient temperature responsive lncRNAs. *BMC Plant Biology*, 18(1), 145. <https://doi.org/10.1186/s12870-018-1362-x>

Shen, E., Zhu, X., Hua, S., Chen, H., Ye, C., Zhou, L., Liu, Q., Zhu, Q.-H., Fan, L., & Chen, X. (2018). Genome-wide identification of oil biosynthesis-related long non-coding RNAs in allopolyploid Brassica napus. *BMC Genomics*, 19(1), 745. <https://doi.org/10.1186/s12864-018-5117-8>

Shin, H., Shin, H.-S., Chen, R., & Harrison, M. J. (2006a). Loss of At4 function impacts phosphate distribution between the roots and the shoots during phosphate starvation. *The Plant Journal: For Cell and Molecular Biology*, 45(5), 712–726. <https://doi.org/10.1111/j.1365-313X.2005.02629.x>

Shin, H., Shin, H.-S., Chen, R., & Harrison, M. J. (2006b). Loss of At4 function impacts phosphate distribution between the roots and the shoots during phosphate starvation. *The Plant Journal: For Cell and Molecular Biology*, 45(5), 712–726.

<https://doi.org/10.1111/j.1365-313X.2005.02629.x>

Shi, Y., Chen, J., & Hou, X. (2020). Similarities and Differences of Photosynthesis Establishment Related mRNAs and Novel lncRNAs in Early Seedlings (Coleoptile/Cotyledon vs. True Leaf) of Rice and. *Frontiers in Genetics*, *11*, 565006. <https://doi.org/10.3389/fgene.2020.565006>

Sievers, F., & Higgins, D. G. (2018). Clustal Omega for making accurate alignments of many protein sequences. *Protein Science: A Publication of the Protein Society*, *27*(1), 135–145. <https://doi.org/10.1002/pro.3290>

Slotte, T., Hazzouri, K. M., Ågren, J. A., Koenig, D., Maumus, F., Guo, Y.-L., Steige, K., Platts, A. E., Escobar, J. S., Newman, L. K., Wang, W., Mandáková, T., Vello, E., Smith, L. M., Henz, S. R., Steffen, J., Takuno, S., Brandvain, Y., Coop, G., ... Wright, S. I. (2013). The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. *Nature Genetics*, *45*(7), 831–835. <https://doi.org/10.1038/ng.2669>

Soneson, C., Love, M. I., & Robinson, M. D. (2015a). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, *4*, 1521. <https://doi.org/10.12688/f1000research.7563.2>

Soneson, C., Love, M. I., & Robinson, M. D. (2015b). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, *4*, 1521. <https://doi.org/10.12688/f1000research.7563.2>

Song, Y., Cui, H., Shi, Y., Xue, J., Ji, C., Zhang, C., Yuan, L., & Li, R. (2020). Genome-wide identification and functional characterization of the *Camelina sativa* WRKY gene family in response to abiotic stress. *BMC Genomics*, *21*(1), 786. <https://doi.org/10.1186/s12864-020-07189-3>

Stauffer, E., & Maizel, A. (2014). Post-transcriptional regulation in root development. *Wiley Interdisciplinary Reviews. RNA*, *5*(5), 679–696. <https://doi.org/10.1002/wrna.1239>

Stoesser, G. (2004). NCBI (National Center for Biotechnology Information). In *Dictionary of Bioinformatics and Computational Biology*. <https://doi.org/10.1002/0471650129.dob0477>

Sun, H.-X., & Chua, N.-H. (2019). Bioinformatics Approaches to Studying Plant Long Noncoding RNAs (lncRNAs): Identification and Functional Interpretation of lncRNAs from RNA-Seq Data Sets. *Methods in Molecular Biology*, *1933*, 197–205. https://doi.org/10.1007/978-1-4939-9045-0_11

Sun, M., & Kraus, W. L. (2015). From discovery to function: the expanding roles of long noncoding RNAs in physiology and disease. *Endocrine Reviews*, *36*(1), 25–64. <https://doi.org/10.1210/er.2014-1034>

Sun, Z., Huang, K., Han, Z., Wang, P., & Fang, Y. (2020a). Genome-wide identification of *Arabidopsis* long noncoding RNAs in response to the blue light. *Scientific Reports*, *10*(1), 1–10. <https://doi.org/10.1038/s41598-020-63187-1>

Sun, Z., Huang, K., Han, Z., Wang, P., & Fang, Y. (2020b). Genome-wide identification of *Arabidopsis* long noncoding RNAs in response to the blue light. *Scientific Reports*, *10*(1), 6229. <https://doi.org/10.1038/s41598-020-63187-1>

Supek, F., Bošnjak, M., Škunca, N., & Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *PloS One*, *6*(7), e21800. <https://doi.org/10.1371/journal.pone.0021800>

Szcześniak, M. W., Bryzghalov, O., Ciomborowska-Basheer, J., & Makołowska, I. (2019). CANTATAdb 2.0: Expanding the Collection of Plant Long Noncoding RNAs. In *Methods in Molecular Biology* (pp. 415–429). <https://doi.org/10.1007/978-1-4939-9045-0>

- Szcześniak, M. W., Rosikiewicz, W., & Makałowska, I. (2016). CANTATadb: A Collection of Plant Long Non-Coding RNAs. *Plant & Cell Physiology*, 57(1), e8. <https://doi.org/10.1093/pcp/pcv201>
- Tang, H., & Lyons, E. (2012). Unleashing the genome of brassica rapa. *Frontiers in Plant Science*, 3, 172. <https://doi.org/10.3389/fpls.2012.00172>
- Tan, X., Li, S., Hu, L., & Zhang, C. (2020). Genome-wide analysis of long non-coding RNAs (lncRNAs) in two contrasting rapeseed (*Brassica napus* L.) genotypes subjected to drought stress and re-watering. *BMC Plant Biology*, 20(1), 81. <https://doi.org/10.1186/s12870-020-2286-9>
- Tripathi, V., Ellis, J. D., Shen, Z., Song, D. Y., Pan, Q., Watt, A. T., Freier, S. M., Bennett, C. F., Sharma, A., Bubulya, P. A., Blencowe, B. J., Prasanth, S. G., & Prasanth, K. V. (2010). The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. *Molecular Cell*, 39(6), 925–938. <https://doi.org/10.1016/j.molcel.2010.08.011>
- Tsai, M.-C., Manor, O., Wan, Y., Mosammaparast, N., Wang, J. K., Lan, F., Shi, Y., Segal, E., & Chang, H. Y. (2010). Long noncoding RNA as modular scaffold of histone modification complexes. *Science*, 329(5992), 689–693. <https://doi.org/10.1126/science.1192002>
- Ulitsky, I. (2016). Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. In *Nature Reviews Genetics* (Vol. 17, Issue 10, pp. 601–614). <https://doi.org/10.1038/nrg.2016.85>
- Ulitsky, I., & Bartel, D. P. (2013). lincRNAs: genomics, evolution, and mechanisms. *Cell*, 154(1), 26–46. <https://doi.org/10.1016/j.cell.2013.06.020>
- Ulitsky, I., Shkumatava, A., Jan, C. H., Sive, H., & Bartel, D. P. (2011). Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell*, 147(7), 1537–1550. <https://doi.org/10.1016/j.cell.2011.11.055>
- UniProt Consortium. (2021). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1), D480–D489. <https://doi.org/10.1093/nar/gkaa1100>
- Vakirlis, N., Carvunis, A.-R., & McLysaght, A. (2020). Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. In *eLife* (Vol. 9). <https://doi.org/10.7554/elife.53500>
- Van Bel, M., Diels, T., Vancaester, E., Kreft, L., Botzki, A., Van de Peer, Y., Coppens, F., & Vandepoele, K. (2017). PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. *Nucleic Acids Research*, 46(D1), D1190–D1196. <https://doi.org/10.1093/nar/gkx1002>
- Wang, A., Hu, J., Gao, C., Chen, G., Wang, B., Lin, C., Song, L., Ding, Y., & Zhou, G. (2019). Genome-wide analysis of long non-coding RNAs unveils the regulatory roles in the heat tolerance of Chinese cabbage (*Brassica rapa* ssp. *chinensis*). *Scientific Reports*, 9(1), 5002. <https://doi.org/10.1038/s41598-019-41428-2>
- Wang, C.-Y., Liu, S.-R., Zhang, X.-Y., Ma, Y.-J., Hu, C.-G., & Zhang, J.-Z. (2017). Genome-wide screening and characterization of long non-coding RNAs involved in flowering development of trifoliate orange (*Poncirus trifoliata* L. Raf.). *Scientific Reports*, 7, 43226. <https://doi.org/10.1038/srep43226>
- Wang, J., Chen, Q., Wu, W., Chen, Y., Zhou, Y., Guo, G., & Chen, M. (2021). Genome-wide analysis of long non-coding RNAs responsive to multiple nutrient stresses in *Arabidopsis thaliana*. *Functional & Integrative Genomics*, 21(1), 17–30.

<https://doi.org/10.1007/s10142-020-00758-5>

Wang, K. C., Yang, Y. W., Liu, B., Sanyal, A., Corces-Zimmerman, R., Chen, Y., Lajoie, B. R., Protacio, A., Flynn, R. A., Gupta, R. A., Wysocka, J., Lei, M., Dekker, J., Helms, J. A., & Chang, H. Y. (2011). A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature*, 472(7341), 120–124.

<https://doi.org/10.1038/nature09819>

Wang, L., Park, H. J., Dasari, S., Wang, S., Kocher, J.-P., & Li, W. (2013). CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Research*, 41(6), e74. <https://doi.org/10.1093/nar/gkt006>

Wang, P., Yu, X., Zhu, Z., Zhai, Y., Zhao, Q., Meng, Y., Li, J., Lou, Q., & Chen, J. (2020). Global Profiling of lncRNAs Expression Responsive to Allopolyploidization in Cucumis. In *Genes* (Vol. 11, Issue 12, p. 1500). <https://doi.org/10.3390/genes11121500>

Wang, T., Song, H., Wei, Y., Li, P., Hu, N., Liu, J., Zhang, B., & Peng, R. (2020). High throughput deep sequencing elucidates the important role of lncRNAs in Foxtail millet response to herbicides. *Genomics*, 112(6), 4463–4473.

<https://doi.org/10.1016/j.ygeno.2020.07.045>

Wang, T.-Z., Liu, M., Zhao, M.-G., Chen, R., & Zhang, W.-H. (2015). Identification and characterization of long non-coding RNAs involved in osmotic and salt stress in *Medicago truncatula* using genome-wide high-throughput sequencing. *BMC Plant Biology*, 15, 131. <https://doi.org/10.1186/s12870-015-0530-5>

Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., Liu, S., Bai, Y., Mun, J.-H., Bancroft, I., Cheng, F., Huang, S., Li, X., Hua, W., Wang, J., Wang, X., Freeling, M., Pires, J. C., Paterson, A. H., Chalhoub, B., ... Brassica rapa Genome Sequencing Project Consortium. (2011). The genome of the mesopolyploid crop species *Brassica rapa*. *Nature Genetics*, 43(10), 1035–1039. <https://doi.org/10.1038/ng.919>

Wang, Y., Fan, X., Lin, F., He, G., Terzaghi, W., Zhu, D., & Deng, X. W. (2014a). Arabidopsis noncoding RNA mediates control of photomorphogenesis by red light. *Proceedings of the National Academy of Sciences of the United States of America*, 111(28), 10359–10364. <https://doi.org/10.1073/pnas.1409457111>

Wang, Y., Fan, X., Lin, F., He, G., Terzaghi, W., Zhu, D., & Deng, X. W. (2014b). Arabidopsis noncoding RNA mediates control of photomorphogenesis by red light. *Proceedings of the National Academy of Sciences of the United States of America*, 111(28), 10359–10364. <https://doi.org/10.1073/pnas.1409457111>

Wunderlich, M., Groß-Hardt, R., & Schöffl, F. (2014). Heat shock factor HSFB2a involved in gametophyte development of *Arabidopsis thaliana* and its expression is controlled by a heat-inducible long non-coding antisense RNA. In *Plant Molecular Biology* (Vol. 85, Issue 6, pp. 541–550). <https://doi.org/10.1007/s11103-014-0202-0>

Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., Bar-Even, A., Horn-Saban, S., Safran, M., Domany, E., Lancet, D., & Shmueli, O. (2005). Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics*, 21(5), 650–659.

<https://doi.org/10.1093/bioinformatics/bti042>

Yang, W. C., Katinakis, P., Hendriks, P., Smolders, A., de Vries, F., Spee, J., van Kammen, A., Bisseling, T., & Franssen, H. (1993). Characterization of GmENOD40, a gene showing novel patterns of cell-specific expression during soybean nodule development. *The Plant Journal: For Cell and Molecular Biology*, 3(4), 573–585.

<https://doi.org/10.1046/j.1365-313x.1993.03040573.x>

Yang, Z., Yang, C., Wang, Z., Yang, Z., Chen, D., & Wu, Y. (2019). LncRNA expression profile and ceRNA analysis in tomato during flowering. *PloS One*, 14(1), e0210650. <https://doi.org/10.1371/journal.pone.0210650>

Yan, Q., Wu, F., Yan, Z., Li, J., Ma, T., Zhang, Y., Zhao, Y., Wang, Y., & Zhang, J. (2019). Differential co-expression networks of long non-coding RNAs and mRNAs in *Cleistogenes songorica* under water stress and during recovery. *BMC Plant Biology*, 19(1), 23. <https://doi.org/10.1186/s12870-018-1626-5>

Yi, X., Zhang, Z., Ling, Y., Xu, W., & Su, Z. (2015). PNRD: a plant non-coding RNA database. *Nucleic Acids Research*, 43(Database issue), D982–D989. <https://doi.org/10.1093/nar/gku1162>

Yoon, J.-H., Abdelmohsen, K., Srikantan, S., Yang, X., Martindale, J. L., De, S., Huarte, M., Zhan, M., Becker, K. G., & Gorospe, M. (2012). LincRNA-p21 suppresses target mRNA translation. *Molecular Cell*, 47(4), 648–655. <https://doi.org/10.1016/j.molcel.2012.06.027>

Yuan, J., Li, J., Yang, Y., Tan, C., Zhu, Y., Hu, L., Qi, Y., & Lu, Z. J. (2018). Stress-responsive regulation of long non-coding RNA polyadenylation in *Oryza sativa*. *The Plant Journal: For Cell and Molecular Biology*, 93(5), 814–827. <https://doi.org/10.1111/tpj.13804>

Yuan, J., Zhang, Y., Dong, J., Sun, Y., Lim, B. L., Liu, D., & Lu, Z. J. (2016). Systematic characterization of novel lncRNAs responding to phosphate starvation in *Arabidopsis thaliana*. *BMC Genomics*, 17, 655. <https://doi.org/10.1186/s12864-016-2929-2>

Yuan, W., Luo, X., Li, Z., Yang, W., Wang, Y., Liu, R., Du, J., & He, Y. (2016). A cis cold memory element and a trans epigenome reader mediate Polycomb silencing of FLC by vernalization in *Arabidopsis*. In *Nature Genetics* (Vol. 48, Issue 12, pp. 1527–1534). <https://doi.org/10.1038/ng.3712>

Yu, X., Yang, J., Li, X., Liu, X., Sun, C., Wu, F., & He, Y. (2013). Global analysis of cis-natural antisense transcripts and their heat-responsive nat-siRNAs in *Brassica rapa*. *BMC Plant Biology*, 13, 208. <https://doi.org/10.1186/1471-2229-13-208>

Zhang, J., Li, J., Saeed, S., Batchelor, W. D., Alariqi, M., Meng, Q., Zhu, F., Zou, J., Xu, Z., Si, H., Wang, Q., Zhang, X., Zhu, H., Jin, S., & Yuan, D. (2022). Identification and Functional Analysis of lncRNA by CRISPR/Cas9 During the Cotton Response to Sap-Sucking Insect Infestation. *Frontiers in Plant Science*, 13, 784511. <https://doi.org/10.3389/fpls.2022.784511>

Zhang, Q., Zhao, C., Li, M., Sun, W., Liu, Y., Xia, H., Sun, M., Li, A., Li, C., Zhao, S., Hou, L., Picimbon, J.-F., Wang, X., & Zhao, Y. (2013). Genome-wide identification of *Thellungiella salsuginea* microRNAs with putative roles in the salt stress response. *BMC Plant Biology*, 13, 180. <https://doi.org/10.1186/1471-2229-13-180>

Zhang, W., Han, Z., Guo, Q., Liu, Y., Zheng, Y., Wu, F., & Jin, W. (2014). Identification of maize long non-coding RNAs responsive to drought stress. *PloS One*, 9(6), e98958. <https://doi.org/10.1371/journal.pone.0098958>

Zhang, Y.-C., Liao, J.-Y., Li, Z.-Y., Yu, Y., Zhang, J.-P., Li, Q.-F., Qu, L.-H., Shu, W.-S., & Chen, Y.-Q. (2014). Genome-wide screening and functional analysis identify a large number of long noncoding RNAs involved in the sexual reproduction of rice. *Genome Biology*, 15(12), 512. <https://doi.org/10.1186/s13059-014-0512-1>

Zhao, T., & Schranz, M. E. (2019). Network-based microsynteny analysis identifies major differences and genomic outliers in mammalian and angiosperm genomes. *Proceedings of the National Academy of Sciences of the United States of America*,

116(6), 2165–2174. <https://doi.org/10.1073/pnas.1801757116>

Zhao, X., Gan, L., Yan, C., Li, C., Sun, Q., Wang, J., Yuan, C., Zhang, H., Shan, S., & Liu, J. N. (2019). Genome-Wide Identification and Characterization of Long Non-Coding RNAs in Peanut. *Genes*, 10(7). <https://doi.org/10.3390/genes10070536>

Zhao, X., Li, J., Lian, B., Gu, H., Li, Y., & Qi, Y. (2018). Global identification of Arabidopsis lncRNAs reveals the regulation of MAF4 by a natural antisense RNA. *Nature Communications*, 9(1), 5056. <https://doi.org/10.1038/s41467-018-07500-7>

Zhao, Y., Li, H., Fang, S., Kang, Y., Wu, W., Hao, Y., Li, Z., Bu, D., Sun, N., Zhang, M. Q., & Chen, R. (2016). NONCODE 2016: an informative and valuable data source of long non-coding RNAs. *Nucleic Acids Research*, 44(D1), D203–D208. <https://doi.org/10.1093/nar/gkv1252>

Zhou, B., Ji, B., Liu, K., Hu, G., Wang, F., Chen, Q., Yu, R., Huang, P., Ren, J., Guo, C., Zhao, H., Zhang, H., Zhao, D., Li, Z., Zeng, Q., Yu, J., Bian, Y., Cao, Z., Xu, S., ... Wang, J. (2021). EVLncRNAs 2.0: an updated database of manually curated functional long non-coding RNAs validated by low-throughput experiments. *Nucleic Acids Research*, 49(D1), D86–D91. <https://doi.org/10.1093/nar/gkaa1076>

Zhou, H., Ren, F., Wang, X., Qiu, K., Sheng, Y., Xie, Q., Shi, P., Zhang, J., & Pan, H. (2022). Genome-wide identification and characterization of long noncoding RNAs during peach (*Prunus persica*) fruit development and ripening. *Scientific Reports*, 12(1), 11044. <https://doi.org/10.1038/s41598-022-15330-3>

Zhou, Y., Cho, W. K., Byun, H.-S., Chavan, V., Kil, E.-J., Lee, S., & Hong, S.-W. (2019). Genome-wide identification of long non-coding RNAs in tomato plants irradiated by neutrons followed by infection with. *PeerJ*, 7, e6286. <https://doi.org/10.7717/peerj.6286>

Zhu, M., Zhang, M., Xing, L., Li, W., Jiang, H., Wang, L., & Xu, M. (2017). Transcriptomic Analysis of Long Non-Coding RNAs and Coding Genes Uncovers a Complex Regulatory Network That Is Involved in Maize Seed Development. *Genes*, 8(10). <https://doi.org/10.3390/genes8100274>

Zhu, X., Tai, X., Ren, Y., Chen, J., & Bo, T. (2019). Genome-Wide Analysis of Coding and Long Non-Coding RNAs Involved in Cuticular Wax Biosynthesis in Cabbage (*Brassica oleracea* L. var. capitata). In *International Journal of Molecular Sciences* (Vol. 20, Issue 11, p. 2820). <https://doi.org/10.3390/ijms20112820>

Zhu, Y., Chen, L., Hong, X., Shi, H., & Li, X. (2022). Revealing the novel complexity of plant long non-coding RNA by strand-specific and whole transcriptome sequencing for evolutionarily representative plant species. *BMC Genomics*, 23(4), 1–16. <https://doi.org/10.1186/s12864-022-08602-9>

Zhu, Y., Chen, L., Zhang, C., Hao, P., Jing, X., & Li, X. (2017). Global transcriptome analysis reveals extensive gene remodeling, alternative splicing and differential transcription profiles in non-seed vascular plant *Selaginella moellendorffii*. In *BMC Genomics* (Vol. 18, Issue S1). <https://doi.org/10.1186/s12864-016-3266-1>

Zou, C., Wang, Q., Lu, C., Yang, W., Zhang, Y., Cheng, H., Feng, X., Prosper, M. A., & Song, G. (2016). Transcriptome analysis reveals long noncoding RNAs involved in fiber development in cotton (*Gossypium arboreum*). *Science China. Life Sciences*, 59(2), 164–171. <https://doi.org/10.1007/s11427-016-5000-2>

Datos suplementarios

- **Capítulo 2: Anotación y clasificación funcional de lncRNAs en *A. thaliana***

Dataset S1. Nombre y posición de los genes de lncRNAs en formato BED12, así como la longitud de su transcrito, CPAT score, localización genómica, clasificación y el nivel de confianza (LC o SM) (link: <https://figshare.com/s/5d19c00590cbd7fe73a1>).

Dataset S2. lncRNAs redundantes (o con más de una clasificación) en formato BED13, (BED12 + columna de clasificación) (link: <https://figshare.com/s/859d12a6da94ba62c9fa>).

Dataset S3. Tabla de la metadata e IDs de todos los transcriptomas empleados en este proyecto (link: <https://figshare.com/s/70c23951cda438d0fe35>).

Dataset S4. Tabla de los valores de *Tau* obtenidos para los lncRNAs órgano-específicos (link: <https://figshare.com/s/5c47f80a3b2929ac45e4>).

Dataset S5. Tabla de los lncRNAs y genes codificantes localizados en cada módulo con WGCNA (link: <https://figshare.com/s/0ca9c981b318c96f5f7c>).

- **Capítulo 3: Conservación de lincRNAs por sintenia y microhomología**

Especie	Anotación de lncRNAs (BED12)	lncRNAs con los tipos de conservación (csv) *	lncRNAs y genes codificantes conservados (sintenia y microhomología) y expresados (tsv) (rlog)
<i>A. thaliana</i> (224 transcriptomas)	Dataset S1	-	-
<i>A. thaliana</i> (4 transcriptomas)	-	At THESIS lncRNAs annotated bed.csv (Dataset S10)	At expression plot.tsv (Dataset S15)
<i>C. rubella</i> (4 transcriptomas)	Cr lncRNAs clean2 RegRNAIDs. bed (Dataset S6)	Cr THESIS lncRNAs annotated bed.csv (Dataset S11)	Cr expression plot.tsv (Dataset S16)
<i>T. parvula</i> (4 transcriptomas)	Tp lncRNAs clean2 RegRNAIDs. bed (Dataset S7)	Tp THESIS lncRNAs annotated bed.csv (Dataset S12)	Tp expression plot.tsv (Dataset S17)
<i>B. rapa</i> (4 transcriptomas)	Br lncRNAs clean2 RegRNAIDs. bed (Dataset S8)	Br THESIS lncRNAs annotated bed.csv (Dataset S13)	Br expression plot.tsv (Dataset S18)
<i>B. oleracea</i> (4 transcriptomas)	Bo lncRNAs clean2 RegRNAIDs. bed (Dataset S9)	Bo THESIS lncRNAs annotated bed.csv (Dataset S14)	Bo expression plot.tsv (Dataset S19)

* Los 3 tipos de conservación son: "OnlySynteny", "Species-specific" y "ConservedBySyntenyandSequence".

Tablas suplementarias

Tabla S1. Información sobre los transcriptomas generados de la parte aérea y raíces de las especies analizadas.

Especie	órganos	Muestras	BioProject	BioSample	SRA*	% GC	Total de reads	Reads alineados
Arabidopsis thaliana	Parte aérea	At_shoot R1	PRJNA765039	SAMN21582179	SRR16093081	46	36,075,409	31,423,073
	Parte aérea	At_shoot R2		SAMN21542145	SRR16093080	46	49,835,624	43,879,678
	Raíces	At_rootR1		SAMN21542144	SRR16093079	45	40,506,616	35,421,322
	Raíces	At_rootR2		SAMN21542146	SRR16093078	45	40,231,170	35,455,572
Brassicarapa	Parte aérea	Br_shoot R1	PRJNA877244	SAMN30689713	SRR21485487	49	42,619,781	36,226,788
	Parte aérea	Br_shoot R2		SAMN30689714	SRR21486522	47	42,858,023	35,172,385
	Raíces	Br_rootR1		SAMN30689715	SRR21487071	46	40,093,722	34,894,215
	Raíces	Br_rootR2		SAMN30689716	SRR21487072	46	41,619,143	34,732,317
Brassica oleracea	Parte aérea	Bo_shoot R1	PRJNA877256	SAMN30689795	SRR21491682	46	29,022,629	3,087,072
	Parte	Bo_shoot		SAMN30689	SRR214916	47	34,544,44	3,916,134

	aérea	R2		796	81		2	
	Raíces	Bo_rootR1		SAMN30689797	SRR21491680	46	42,661,491	5,125,640
	Raíces	Bo_rootR2		SAMN30689798	SRR21491679	46	24,173,645	2,951,559
Capsella rubella	Parte aérea	Cr_shootR1	PRJNA877255	SAMN30691554	SRR21498893	46	38,871,033	5,415,288
	Parte aérea	Cr_shootR2		SAMN30691555	SRR21498892	46	32,361,544	3,857,282
	Raíces	Cr_rootR1		SAMN30691556	SRR21498891	46	31,716,337	3,900,646
	Raíces	Cr_rootR2		SAMN30691557	SRR21498890	45	27,839,797	3,531,978
Thellungiella parvula	Parte aérea	Tp_shootR1	PRJNA877258	SAMN30691650	SRR21504237	46	38,007,402	31,824,549
	Parte aérea	Tp_shootR2		SAMN30691651	SRR21504236	46	43,067,272	36,256,501
	Raíces	Tp_rootR1		SAMN30691652	SRR21504235	47	48,068,798	38,370,303
	Raíces	Tp_rootR2		SAMN30691653	SRR21504234	47	40,111,744	33,349,668

*Los datos son *Paired-end* (Forward and reverse data), R = Réplica biológica, At = Arabidopsis thaliana, Br = Brassica rapa, Bo = Brassica oleracea, Cr = Capsella rubella, Tp = Thellungiella parvula.

Tabla S2. lncRNAs con funciones conocidas con valores altos de *Tau* en un solo 6rgano.

Gene ID	Tissue/organ	<i>Tau</i>	lncRNA
AT2G34655	petiole	0.657	<i>APOLO</i>
AT5G03545	root	0.583	<i>AT4</i>
AT5G03545	plant callus	0.572	<i>AT4</i>
AT4G06195	hypocotyl	0.81	<i>AT4G06195</i>
AT4G06195	cotyledon	0.788	<i>AT4G06195</i>
AT4G06195	whole plant	0.543	<i>AT4G06195</i>
AT4G08415	embryo	1	<i>AT4G08415</i>
AT5G01155	SAM	0.739	<i>AT5G01155</i>
AT5G01155	flower	0.699	<i>AT5G01155</i>
AT5G01155	siliqua	0.559	<i>AT5G01155</i>
AT5G01675	seed	0.902	<i>COOLAIR</i>
AT5G01675	embryo	0.707	<i>COOLAIR</i>
AT1G21529	plant callus	0.975	<i>DRIR</i>
AT4G16355	petal	0.946	<i>ELENA1</i>
AT1G08103	SAM	0.913	<i>FLINC</i>
AT2G35747	SAM	1	<i>HID1</i>
AT3G09922	plant callus	0.874	<i>IPS1</i>
AT3G09922	root	0.836	<i>IPS1</i>
AT3G09922	cotyledon	0.609	<i>IPS1</i>
AT5G00580	seedling	0.957	<i>MARS</i>
AT5G00580	root	0.868	<i>MARS</i>
AT2G27400	hypocotyl	0.596	<i>TAS1a</i>
AT2G27400	cotyledon	0.566	<i>TAS1a</i>
AT2G27400	SAM	0.542	<i>TAS1a</i>
AT2G27400	root	0.54	<i>TAS1a</i>

Tabla S3. Genomas empleados en este trabajo.

Especies	Ensembl ID	Recursos
<i>Arabidopsis thaliana</i>	GCA_000001735.1 TAIR10	Ensembl Plantas/Araport11
<i>Arabidopsis lyrata</i>	GCA_000004255.1	Ensembl Plants
<i>Arabidopsis halleri</i>	GCA_900078215.1	Phytozome V13
<i>Aethionema arabicum</i>	GCA_000411095.1	NCBI
<i>Arabis alpina</i>	GCA_000733195.1	NCBI
<i>Bochera stricta</i>	GCA_002079875.1	NCBI
<i>Brassica rapa</i>	GCA_000309985.1	Ensembl Plants
<i>Brassica oleracea</i>	GCA_000695525.1	Ensembl Plants
<i>Camelina sativa</i>	GCA_000633955.1	NCBI
<i>Capsella rubella</i>	GCA_000375325.1	Phytozome v13
<i>Thellungiella parvula</i>	GCA_000218505.1	NCBI
<i>Eutrema salsugieum</i>	GCA_000478725.1	NCBI

Tabla S4. Archivos de anotación de cada especie analizada.

Especies	Recursos
<i>Arabidopsis thaliana</i>	Ensembl Plants (TAIR10/Araport11)
<i>Brassica rapa</i>	Ensembl Plants
<i>Brassica oleracea</i>	Ensembl Plants
<i>Capsella rubella</i>	Phytozome v13
<i>Thellungiella parvula</i>	Thellungiella.org (<i>Schrenkiella parvula</i> / <i>Eutrema parvulum</i> , version 2)

Tabla S5. Archivos empleados para la anotación de *T. parvula*.

Files	Recursos
Genome 1 (G1)	NCBI (<i>Schrenkiella-parvula_eutrema-parvulum_v01.genome.fa</i>)
Genome 2 (G2)	Thellungiella.org (<i>Thellungiella_parvula_v2.fa</i>)
Annotation file	Thellungiella.org (<i>Schrenkiella parvula</i> / <i>Eutrema parvulum</i> , version 2)

(G2)	
Annotation file created (G1)	Thellungiella parvula v2.gtf

Tabla S6. Archivos empleados para OrthoFinder.

Species	coding genes	Peptides file	Source	Genome version
Arabidopsis thaliana	27,628	Arabidopsis_thaliana.TAIR10.pep.all.fa.gz	Ensembl Plants	TAIR10/Arabidopsis11
Arabidopsis lyrata	32,667	Arabidopsis_lyrata.v.1.0.pep.all.fa.gz	Ensembl Plants	1.0
Arabidopsis halleri	32,158	Arabidopsis_halleri.Ahal2.2.pep.all.fa.gz	Ensembl Plants	2.2
Camelina sativa	89,275	Camelina_sativa.Cs.pep.all.fa.gz	Ensembl Plants	1
Capsella rubella	27,682	Crubella_474_v1.1.protein_primaryTranscriptOnly.fa.gz	Phytozome v13	1.1
Arabis alpina	21,609	Arabis_alpina.A_alpina_V4.pep.all.fa.gz	Phytozome v13	4
Boechera stricta	27,416	Bstricta_278_v1.2.cds_primaryTranscriptOnly.fa.gz	Ensembl Plants	1.2
Brassica oleracea	59,220	Brassica_oleracea.BOL.pep.all.fa.gz	Ensembl Plants	1
Brassica rapa	41,018	Brassica_rapa.Brapa_1.0.pep.all.fa.gz	Ensembl Plants	1.0
Thellungiella parvula	26,847	Sparvula_574_v2.2.transcript_primaryTranscriptOnly.fa.gz	Phytozome v13	2.2
Eutrema salsugineum	26,847	Eutrema_salsugineum.Eut_salg1_0.pep.all.fa.gz	Ensembl Plants	1.0
Aethionema arabicum	22,230	aeth.pep	SynNet	2.5

Aethionema arabicum lo obtuve de <https://github.com/zhaotao1987/SynNet-Pipeline>

Tabla S7. Porcentaje de ortogrupos conservados de genes codificantes.

	Especies de referencia										
Especies*	Ae	Ah	Al	Ah	Bs	Bo	Br	Cs	Cr	Tp	Es
Ah	93.04	100	89.41	91.67	91.33	76.48	82.98	77.76	91.64	92.19	92.42
Al	92.99	91.67	100	93.93	92.53	77.15	83.07	78.91	92.86	92.38	93.45
At	95.24	89.04	88.99	100	94.66	77.78	84.41	78.79	94.58	94.48	95.49
Bo	93.60	85.42	84.06	89.44	90.95	100	93.86	78.56	90.15	93.25	93.42
Br	93.50	84.90	82.91	88.93	90.20	85.98	100	76.16	89.80	93.01	92.91
Cs	95.92	90.81	89.89	94.73	95.34	82.14	86.92	100	96.40	95.23	96.19
Cr	94.53	88.08	87.06	93.60	94.48	77.58	84.35	79.34	100	94.11	94.94
Es	93.67	84.58	83.43	89.97	90.98	76.54	83.10	75.38	90.39	92.76	100
Tp	93.55	83.93	82.04	88.56	89.82	76	82.76	74.23	89.14	100	92.28
Ae	100	72	70.20	75.88	77.04	64.85	70.72	63.56	76.11	79.52	79.21
Bs	94.64	86.83	85.82	92.66	100	77.42	83.81	77.61	93.46	93.80	94.52

Abreviación de los nombres de las especies: Ah = *Arabidopsis halleri*, Al = *Arabidopsis lyrata*, At = *Arabidopsis thaliana*, Bo = *Brassica oleracea*, Br = *Brassica rapa*, Cs = *Camelina sativa*, Cr = *Capsella rubella*, Es = *Eutrema salsugineum*, Tp = *Thellungiella parvula*, Ae = *Aethionema arabicum*, Bs = *Boechera stricta*.

Tabla S8. Distribución de los genes codificantes ortólogos y lincRNAs ortólogos (conservados por sintenia y microhomología) con expresión en cada órgano y especie.

Especie	Tipo	Expresión localizada	Genes expresados	Total	Proporción (%)
<i>A. thaliana</i>	lincRNA	global	700	783	89.4
<i>A. thaliana</i>	lincRNA	Raíces	40	783	5.11
<i>A. thaliana</i>	lincRNA	Parte aérea	43	783	5.49
<i>A. thaliana</i>	Genes codificantes	global	10876	23293	46.69
<i>A. thaliana</i>	Genes codificantes	Raíces	6249	23293	26.83
<i>A. thaliana</i>	Genes codificantes	Parte aérea	6168	23293	26.48
<i>B. oleracea</i>	lincRNA	global	410	826	49.64
<i>B. oleracea</i>	lincRNA	Raíces	250	826	30.27
<i>B. oleracea</i>	lincRNA	Parte aérea	166	826	20.1
<i>B. oleracea</i>	Genes codificantes	global	15643	29370	53.26
<i>B. oleracea</i>	Genes codificantes	Raíces	7280	29370	24.79
<i>B. oleracea</i>	Genes codificantes	Parte aérea	6447	29370	21.95
<i>B. rapa</i>	lincRNA	global	659	997	66.1
<i>B. rapa</i>	lincRNA	Raíces	173	997	17.35
<i>B. rapa</i>	lincRNA	Parte aérea	165	997	16.55
<i>B. rapa</i>	Genes codificantes	global	16533	28622	57.76
<i>B. rapa</i>	Genes codificantes	Raíces	6144	28622	21.47
<i>B. rapa</i>	Genes codificantes	Parte aérea	5945	28622	20.77
<i>C. rubella</i>	lincRNA	global	157	267	58.8
<i>C. rubella</i>	lincRNA	Raíces	65	267	24.34
<i>C. rubella</i>	lincRNA	Parte aérea	45	267	16.85
<i>C. rubella</i>	Genes codificantes	global	10025	21887	45.8
<i>C. rubella</i>	Genes codificantes	Raíces	6182	21887	28.25

<i>C. rubella</i>	Genes codificantes	Parte aérea	5680	21887	25.95
<i>T. parvula</i>	lincRNA	global	293	494	59.31
<i>T. parvula</i>	lincRNA	Raíces	124	494	25.1
<i>T. parvula</i>	lincRNA	Parte aérea	77	494	15.59
<i>T. parvula</i>	Genes codificantes	global	11779	19395	60.73
<i>T. parvula</i>	Genes codificantes	Raíces	4050	19395	20.88
<i>T. parvula</i>	Genes codificantes	Parte aérea	3566	19395	18.39

Tabla S9. AtlincRNAs conservados (por sintenia y microhomología) y anotados como funcionales en EvlncRNA v2.0.

EvLncRNA ID	AtlincRNA geneID	AtlincRNA	Especies con expresión	GeneID en otras especies	Expresión localizada en <i>A. thaliana</i>	Expresión localizada en las otras especies
N/A	AT3G25795	TAS4	<i>B. rapa</i> , <i>B. oleracea</i>	<i>BrRegRNA_26595</i> , <i>BoRegRNA_31881</i>	raíces	raíces
N/A	<i>AtRegRNA_9454</i>	pri-miR858a	-	-	global	-
EL2785	AT4G12735	npc60	-	-	global	-
EL2784	AT1G67920	npc536	<i>B. rapa</i> , <i>B. oleracea</i> , <i>T. parvula</i>	<i>Bra004062</i> , <i>Bo7g050320</i> , <i>Tp5g22920</i>	raíces	global
EL2779	<i>AtRegRNA_12326</i>	npc375	-	-	global	-
EL2774	AT4G04692	npc156	-	-	parte aérea	-
N/A	AT5G00580	MARS/At5NC066580	-	-	raíces	-
N/A	AT3G05655	lncCOBRA5/At3NC040900	-	-	global	-
N/A	AT3G03435	lncCOBRA3/At3NC020890	<i>B. rapa</i>	<i>BrRegRNA_15395</i>	global	-
N/A	AT1G05913	lncCOBRA1/At1NC031460	-	-	global	-
EL1794	AT4G09445	LINC-AP2/At4NC	-	-	global	-

		069370				
EL1331	AT3G09922	IPS1	<i>B. rapa</i> , <i>T. parvula</i>	<i>BrRegRNA_3639</i> , <i>TpRegRNA_16718</i>	parte aérea	global(Br) y raíces (Tp)
EL1181	AT2G18440	GUT15	-	-	parte aérea	-
EL1004	AT1G08103	FLINC/At1NC078930	<i>B. oleracea</i>	<i>BoRegRNA_31476</i>	global	global
N/A	AT2G18735	FLAIL	-	-	global	-
EL0872	AT1G10682	enod40	<i>B. rapa</i>	<i>BrRegRNA_14445</i>	raíces	raíces
EL0872	<i>AtRegRNA_1474</i>	enod40	<i>T. parvula</i>	<i>TpRegRNA_816</i>	raíces	raíces
EL0864	AT4G16355	ELENA1	-	-	global	-
EL0450	AT2G18440	AtGUT15	-	-	parte aérea	-
EL0437	AT5G24735	AT5G24735	-	-	global	-
EL0435	AT5G15022	AT5G15022	<i>B. oleracea</i>	<i>BoRegRNA_3830</i>	global	raíces
EL0434	AT5G07322	AT5G07322	<i>B. rapa</i>	<i>BrRegRNA_6752</i>	raíces	raíces
EL0433	AT5G06165	AT5G06165	<i>B. rapa</i>	<i>BrRegRNA_6707</i>	global	global
EL0433	<i>AtRegRNA_33960</i>	AT5G06165	-	-	global	-
EL0432	AT5G03285	AT5G03285	-	-	global	-
EL0428	<i>AtRegRNA_30196</i>	At4NC047210	-	-	global	-
EL0427	AT4G40065	AT4G40065	-	-	raíces	-
EL0426	AT4G38932	AT4G38932	-	-	global	-

EL0424	AT4G30975	AT4G30975	-	-	global	-
EL0421	AT4G14548	AT4G14548	-	-	parte aérea	-
EL0419	AT4G08035	AT4G08035	-	-	global	-
EL0418	AT4G06701	AT4G06701	-	-	raíces	-
EL0417	AT4G03811	AT4G03811	-	-	global	-
EL0413	AT1G08165	At4-3	<i>B. oleracea</i>	<i>BrRegRNA_24772</i>	global	raíces
EL0412	AT2G08825	At4-2	<i>B. oleracea</i>	<i>BoRegRNA_12954</i>	global	parte aérea
EL0407	AT3G61198	AT3G61198	-	-	global	-
EL0405	AT3G59765	AT3G59765	-	-	global	-
EL0404	AT3G57157	AT3G57157	<i>B. rapa, T. parvula</i>	<i>Bra007307, Tp5g05420</i>	raíces	raíces
EL0402	AT3G52748	AT3G52748	-	-	raíces	-
EL0401	AT3G52742	AT3G52742	-	-	global	-
EL0399	AT3G48115	AT3G48115	-	-	raíces	-
EL0398	AT3G45638	AT3G45638	-	-	global	-
EL0397	AT3G44798	AT3G44798	-	-	global	-
EL0396	AT3G29644	AT3G29644	-	-	global	-

		4				
EL0394	AT3G27884	AT3G27884	-	-	raíces	-
EL0393	AT3G13277	AT3G13277	-	-	global	-
EL0391	AT3G06125	AT3G06125	-	-	global	-
EL0386	AT2G46192	AT2G46192	-	-	global	-
EL0383	AT2G44995	AT2G44995	<i>C. rubella</i>	<i>CrRegRNA_20009</i>	global	global
EL0380	AT2G42485	AT2G42485	-	-	parte aérea	-
EL0374	AT2G31585	AT2G31585	-	-	raíces	-
EL0372	AT2G18735	AT2G18735	-	-	global	-
EL0371	AT2G15555	AT2G15555	-	-	global	-
EL0369	AT2G14878	AT2G14878	-	-	raíces	-
EL0367	AT2G06002	AT2G06002	-	-	parte aérea	-
EL0358	AT1G67105	AT1G67105	-	-	parte aérea	-
EL0355	AT1G57835	AT1G57835	-	-	raíces	-
EL0351	AT1G31935	AT1G31935	<i>B. rapa</i>	<i>Bra033864</i>	raíces	global

EL0350	<i>AT1G26558</i>	AT1G2655 8	-	-	global	-
EL0346	<i>AT1G16635</i>	AT1G1663 5	-	-	global	-
EL0340	<i>AT1G11175</i>	AT1G1117 5		-	global	-
EL0339	<i>AT1G10682</i>	AT1G1068 2	<i>B. rapa</i>	<i>BrRegRNA_14445</i>	raíces	raíces

Tabla S10. Ubicación de las copias localizadas de AtIPS1 a través de las especies.

Especie	Cromosoma	Inicio	Fin	strand
Camelina_sativa_CM002747.1	AT3G09920	4115421	4121703	-
Camelina_sativa_CM002747.1	AT3G09922	4128308	4128943	+
Camelina_sativa_CM002747.1	AT3G09925	4129186	4130230	-
Camelina_sativa_CM002747.1	AT3G09930	4130669	4134585	+
Camelina_sativa_CM002747.1	AT3G09940	4135217	4138085	-
Camelina_sativa_CM002747.1	AT3G09950	4138846	4139497	-
Camelina_sativa_CM002743.1	AT3G09920	3983414	3989546	-
Camelina_sativa_CM002743.1	AT3G09922	3995943	3996556	+
Camelina_sativa_CM002743.1	AT3G09925	3996675	3997735	-
Camelina_sativa_CM002743.1	AT3G09930	3998179	4000585	+
Camelina_sativa_CM002743.1	AT3G09940	4001180	4011149	-
Camelina_sativa_CM002743.1	AT3G09950	4011953	4012572	-
Camelina_sativa_CM002729.1	AT3G09920	3801137	3807557	-
Camelina_sativa_CM002729.1	AT3G09922	3814252	3815148	+
Camelina_sativa_CM002729.1	AT3G09925	3815390	3823816	-
Camelina_sativa_CM002729.1	AT3G09930	3817110	3826513	+
Camelina_sativa_CM002729.1	AT3G09940	3819691	3829839	-
Camelina_sativa_CM002729.1	AT3G09950	3830615	3831209	-
Eutrema_salsugineum_NW_006256829.1	AT3G09922	854137	854751	+
Eutrema_salsugineum_NW_006256829.1	AT3G09930	870736	878353	+
Eutrema_salsugineum_NW_006256885.1	AT3G09920	6349174	6852836	-
Eutrema_salsugineum_NW_006256885.1	AT3G09922	6341544	6342162	+
Eutrema_salsugineum_NW_006256885.1	AT3G09925	6339982	6340938	-
Eutrema_salsugineum_NW_006256885.1	AT3G09930	6336327	6338862	+
Eutrema_salsugineum_NW_006256885.1	AT3G09950	6335408	6336012	-
Thellungiella_parvula_CM001192.1	AT3G09922	1740916	17409793	+

		7		
Thellungiella_parvula_CM001192.1	AT3G09930	1738262 6	17395630	+
Thellungiella_parvula_CM001189.1	AT3G09920	2323033	2777855	-
Thellungiella_parvula_CM001189.1	AT3G09922	2783570	2784135	+
Thellungiella_parvula_CM001189.1	AT3G09925	2784135	2785026	-
Thellungiella_parvula_CM001189.1	AT3G09930	2785761	2788309	+
Thellungiella_parvula_CM001189.1	AT3G09950	2788743	2789286	-
Brassica_rapa_A05	AT3G09920	2221171 0	22216825	+
Brassica_rapa_A05	AT3G09922	2219802 4	22199048	-
Brassica_rapa_A05	AT3G09925	2219643 0	22197737	+
Brassica_rapa_A05	AT3G09930	2219463 1	22195146	-
Brassica_rapa_A05	AT3G09950	2219207 2	22192643	+
Brassica_rapa_A10	AT3G09922	1707399 7	17074607	-
Brassica_rapa_A03	AT3G09920	1562028 5	15620644	+
Brassica_rapa_A03	AT3G09922	499257	15727668	-
Brassica_rapa_A03	AT3G09925	1572555 5	15726692	+
Brassica_rapa_A02	AT3G09922	1853317	1853925	+
Brassica_rapa_A02	AT3G09930	1865050	1870047	+
Brassica_rapa_A01	AT3G09920	2780406 3	27804440	-
Brassica_rapa_A01	AT3G09922	2751038	27510870	-

		6		
Brassica_oleracea_C9	AT3G09922	53871127	53874708	-
Brassica_oleracea_C7	11147597	11148683	AT3G09922	+
Brassica_oleracea_C7	11149246	11149313	AT3G09925	-
Brassica_oleracea_C5	AT3G09920	43077748	43082646	+
Brassica_oleracea_C5	AT3G09922	43043852	43044628	-
Brassica_oleracea_C5	AT3G09925	43041001	43042277	+
Brassica_oleracea_C5	AT3G09930	43034197	43034741	-
Brassica_oleracea_C5	AT3G09950	43031310	43031881	+
Brassica_oleracea_C3	AT3G09920	23556100	23556458	+
Brassica_oleracea_C3	AT3G09922	642587	23697765	+
Brassica_oleracea_C3	AT3G09925	23695626	37502978	+
Brassica_oleracea_C3	AT3G09930	665494	676379	+
Brassica_oleracea_C2	AT3G09922	697736	698360	+
Brassica_oleracea_C2	AT3G09930	707076	716082	+
Brassica_oleracea_C1	AT3G09920	41480700	41481121	-
Brassica_oleracea_C1	AT3G09922	40979044	40979531	-

Figuras suplementarias

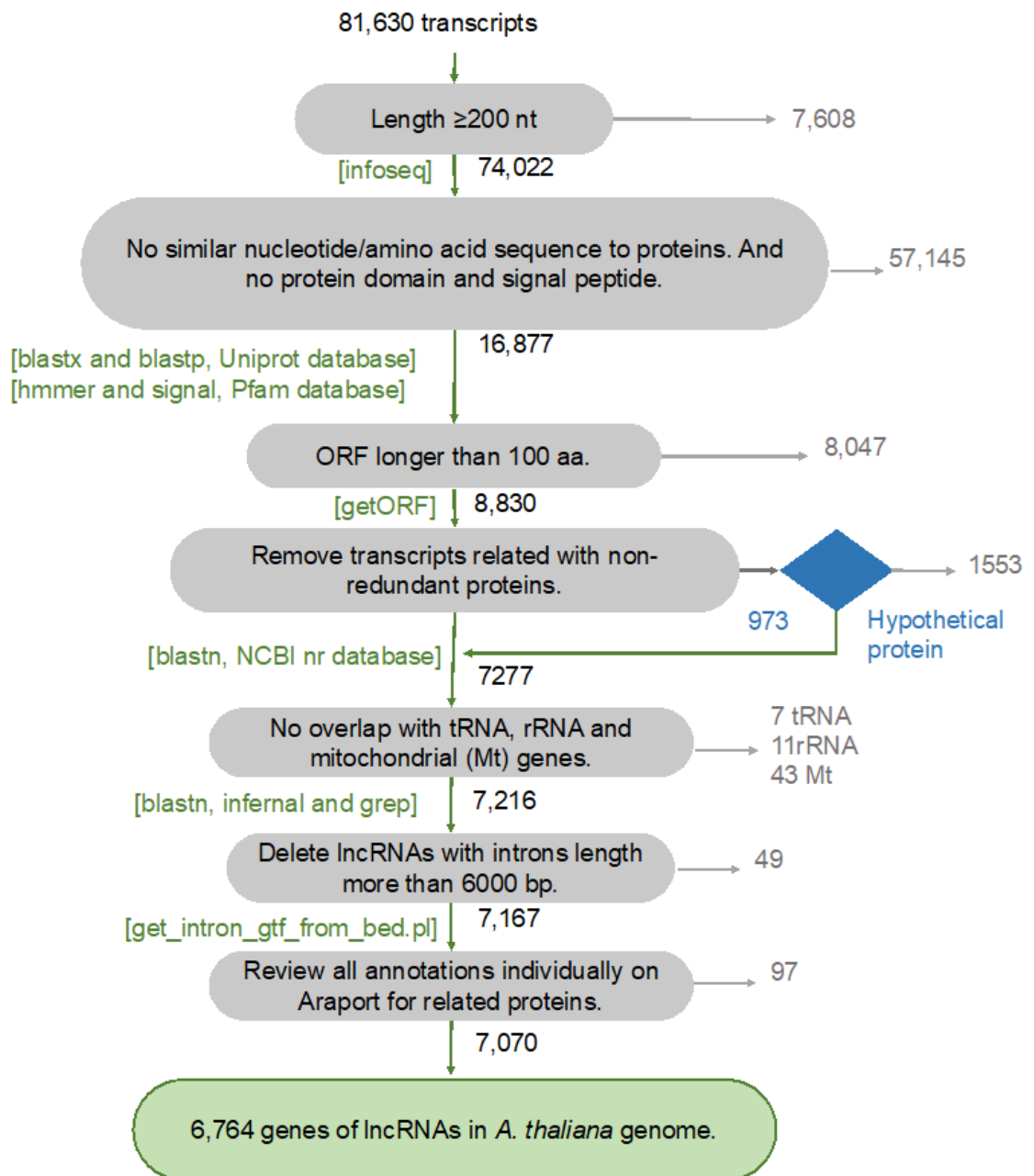


Figura S1. Diagrama de los filtros empleados para la identificación de lncRNAs.

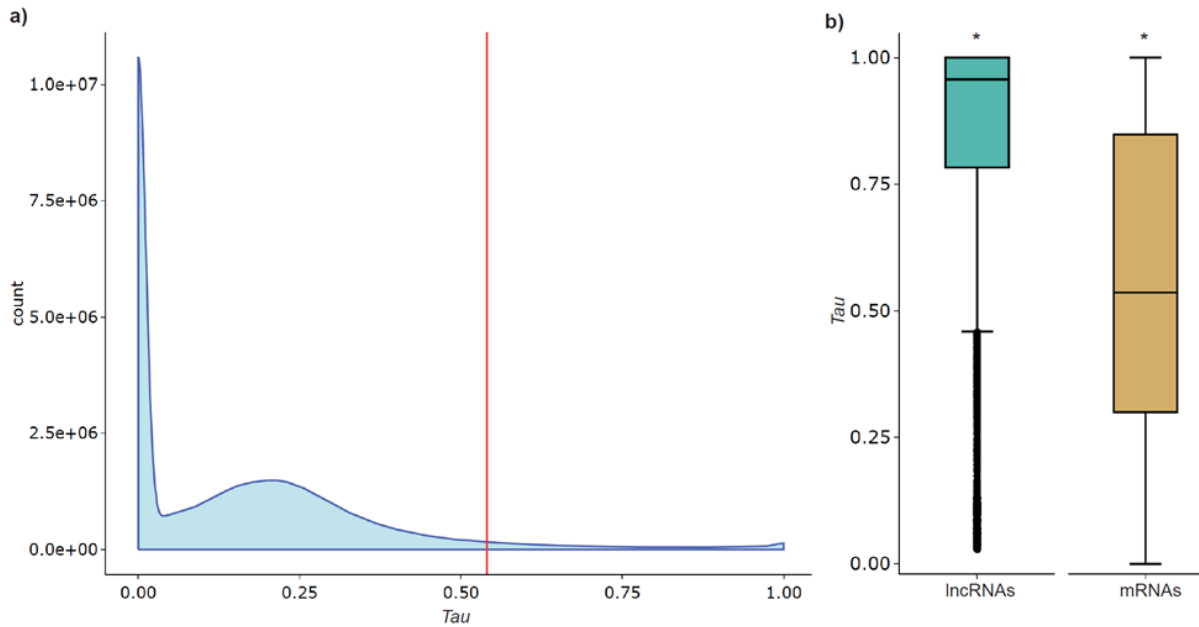


Figure S2. Distribution of values of τ . a) Cut-off point of τ (0.54) values. b) Boxplot with top values of τ in lncRNAs and mRNAs. lncRNAs have significantly higher values than mRNAs (Wilcoxon test $p = 2.2 \times 10^{-16}$).

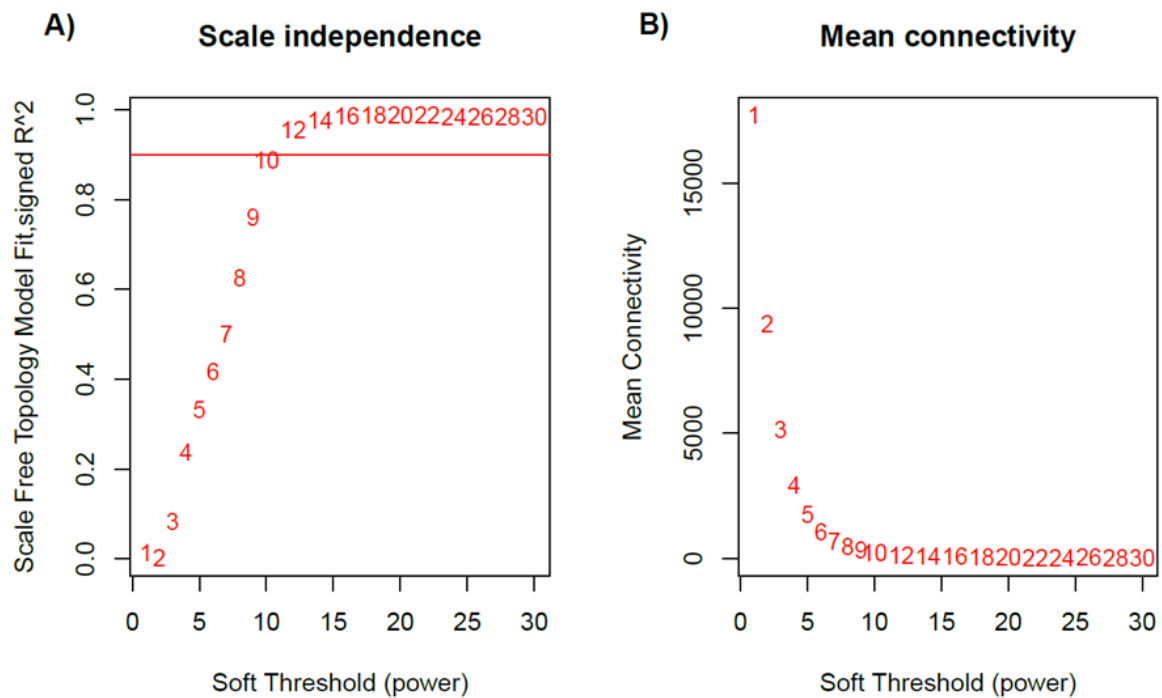


Figure S3. Network topology analysis. Various soft-thresholding powers (x-axis) are shown for the scale free topology fit index (A) and mean connectivity (B) (y axis). Based on these analyses we selected a power β soft threshold of 12. The red line in panel A corresponds to using an R^2 cut-off of $h=0.9$.

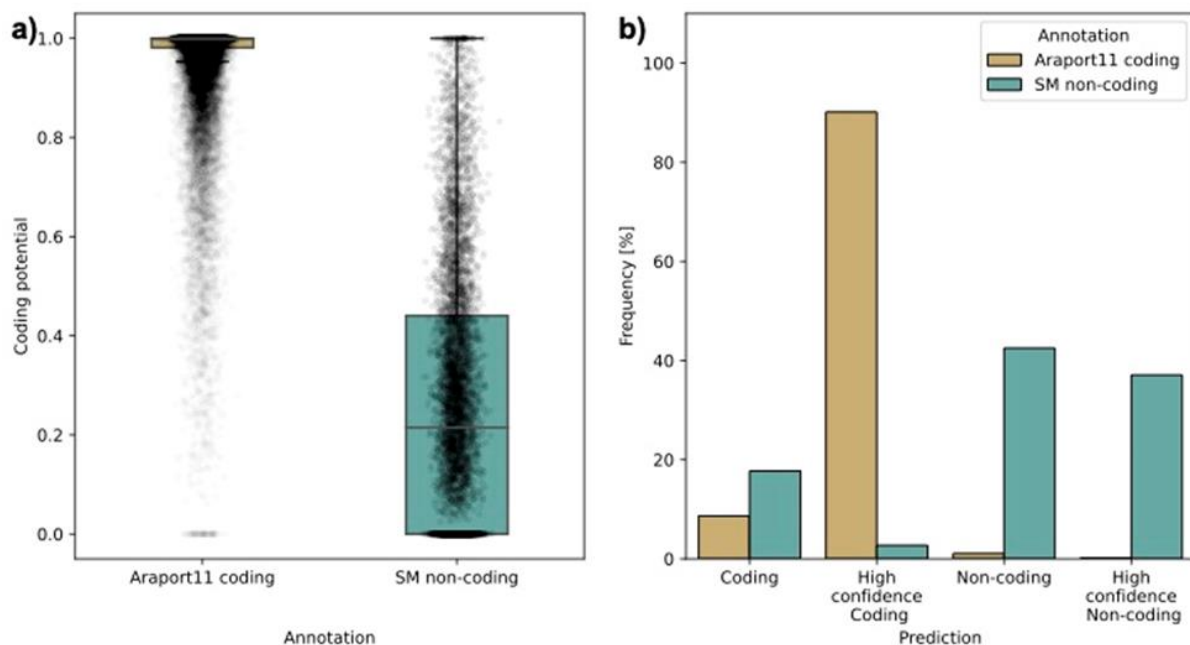


Figure S4. Coding potential assessment of Araport11 coding genes (ochre) and SM non-coding (aqua) a) Boxplot with CPAT coding potential values for both gene sets. LncRNAs have significantly lower coding potential values than mRNAs (Wilcoxon test $p < 10 \times e^{-15}$). b) Percentage of both classes classified by CPAT as coding (coding potential > 0.5), non-coding (coding potential ≤ 0.5), high confidence coding (coding potential ≥ 0.9) and high confidence non-coding (coding potential ≤ 0.1).

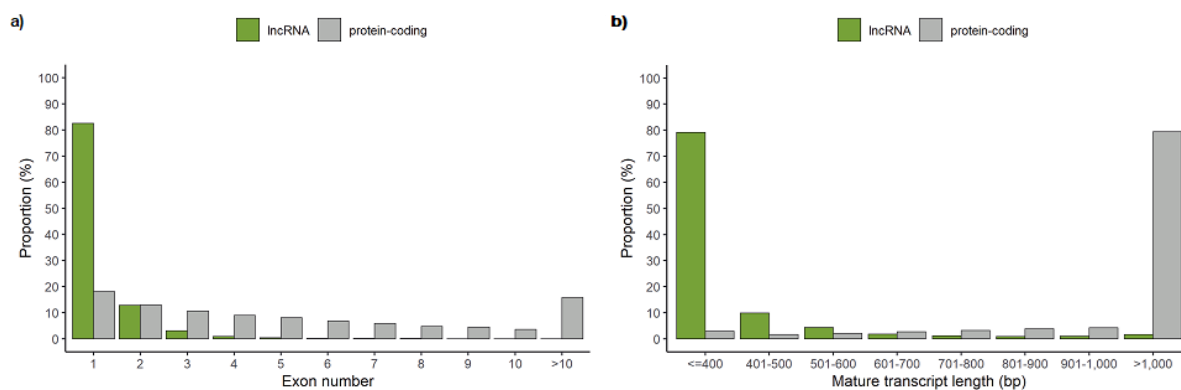


Figure S5. a) Proportion of the number of exons contained in the lncRNAs and coding genes. b) Proportion of the average size of the lncRNAs compared to the coding genes.

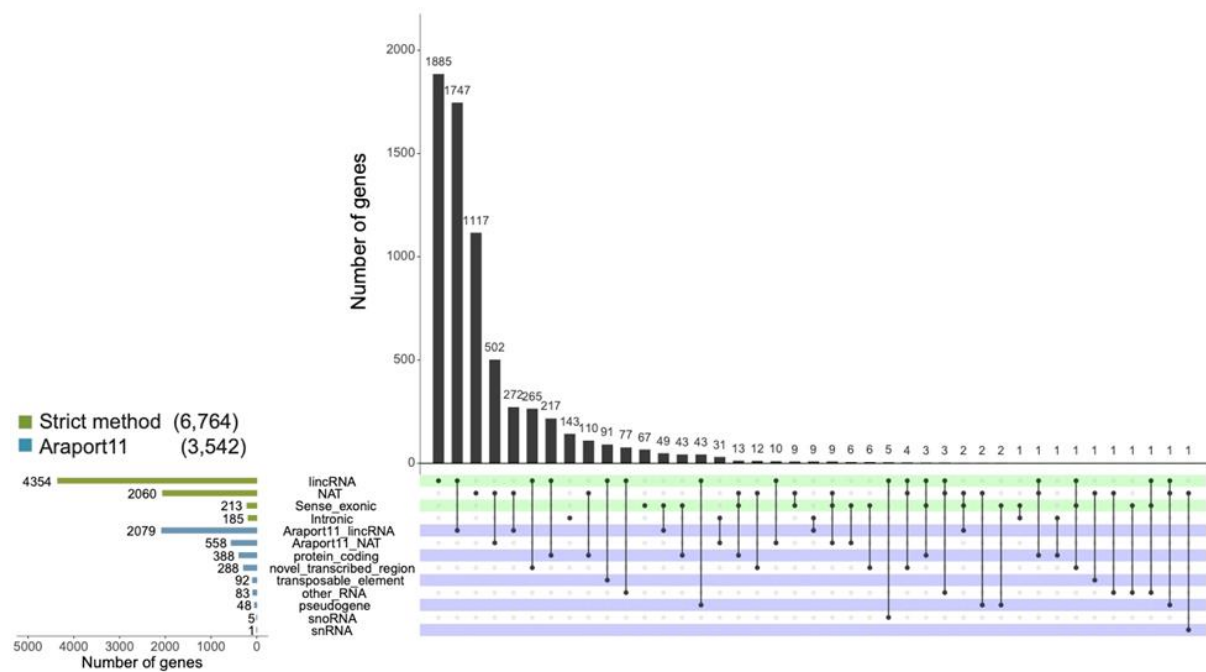


Figure S6. Comparison of each of the categories of lncRNAs present in Araport11 against the categories of the newly generated annotation.

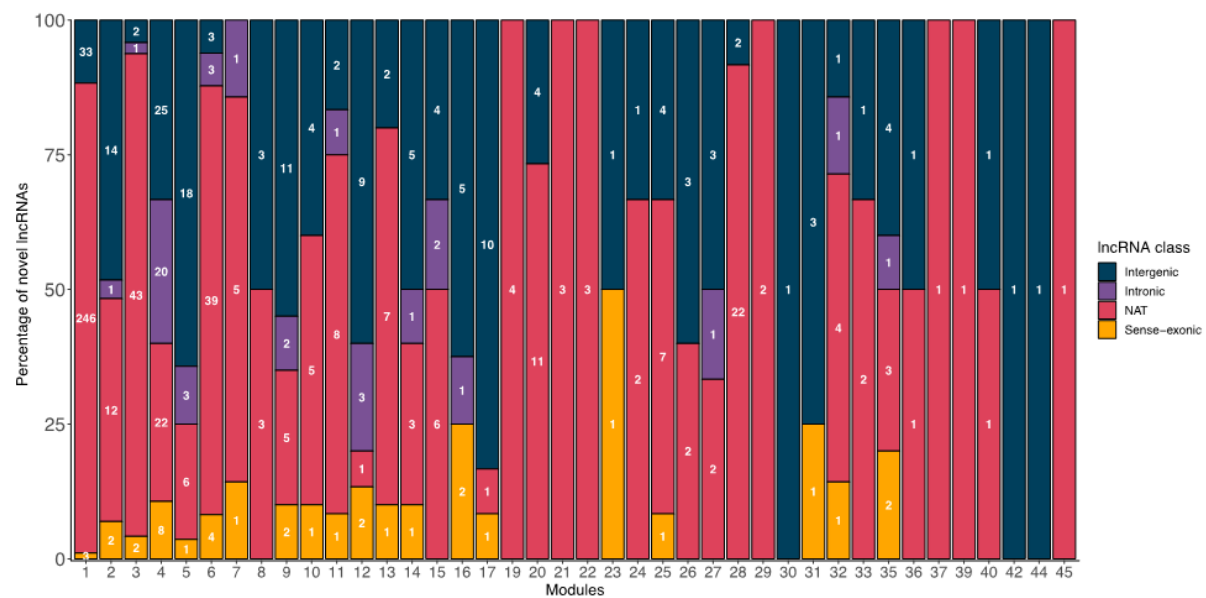


Figure S7. Novel lncRNAs per module. Proportion of novel lncRNAs (without annotation in Araport11) in each co-expression module and classified by biotype (Intergenic (lincRNA) - blue, Intronic - purple, NAT - pink, Sense-exonic - yellow).

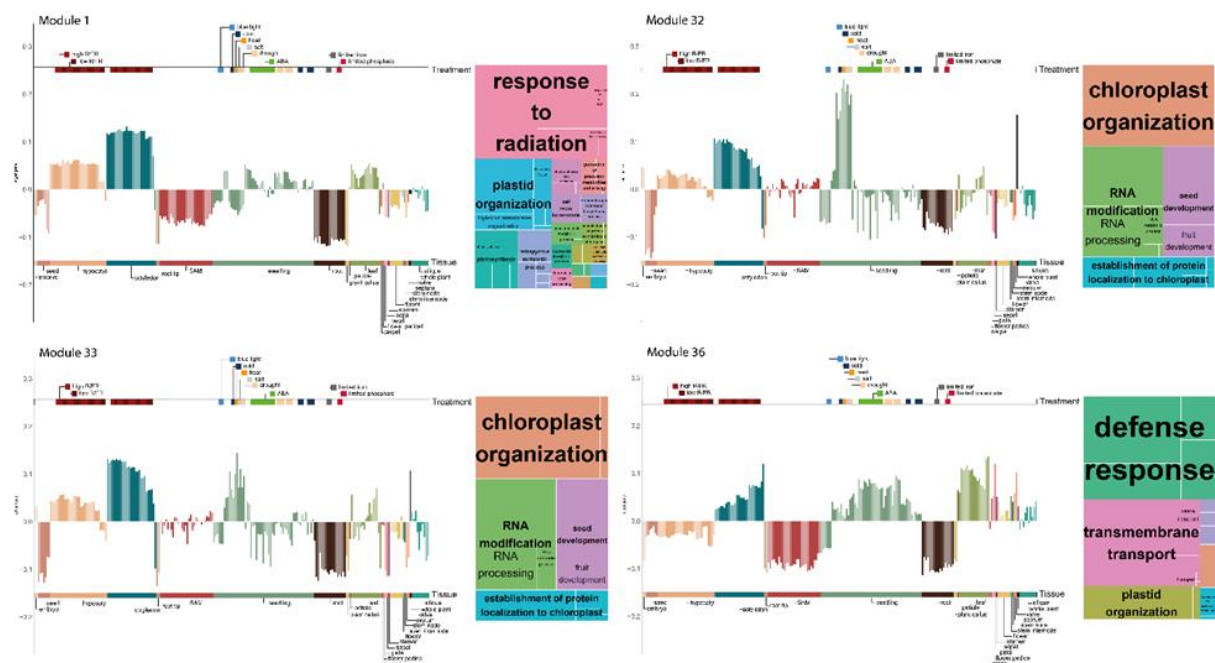


Figure S8. Eigengene expression per module for modules 1, 32, 33 and 36, chloroplast organization and photosynthesis functional category (4 modules with 409 lncRNAs).

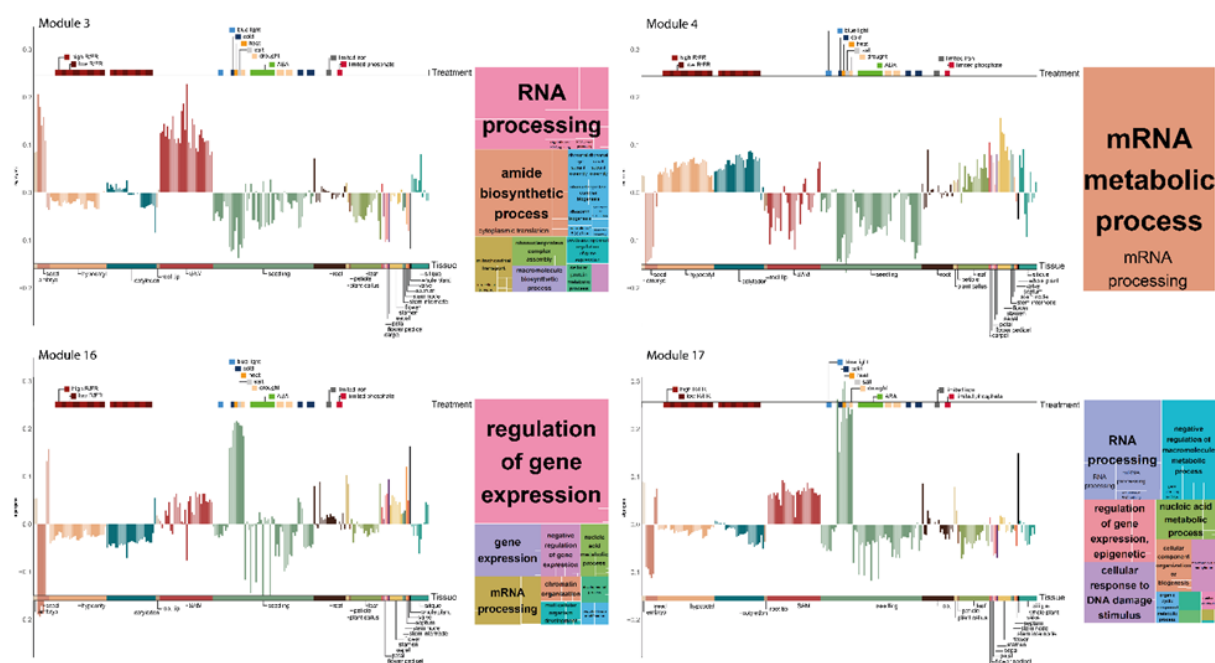


Figure S9. Eigengenes expression per module for modules 3, 4, 16 and 17; RNA regulation and transcription functional category (4 modules with 375 lncRNAs).

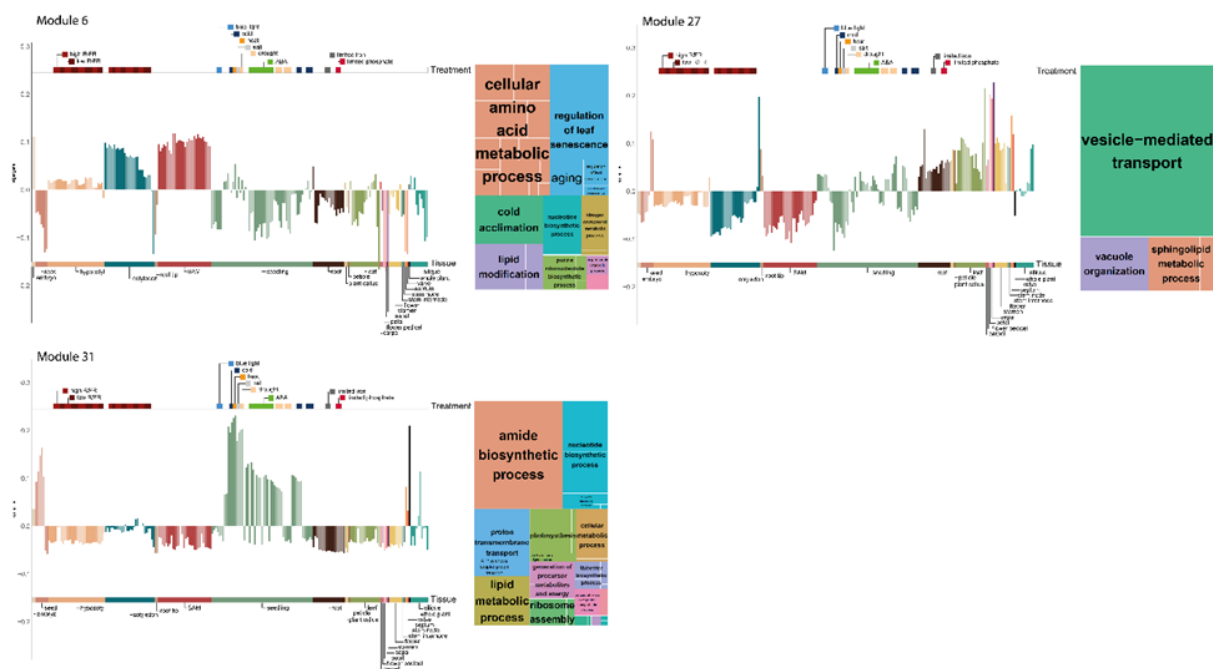


Figure S13. Eigengene expression per module for modules 6, 27 and 31; lipids and membranes functional category (3 modules with 97 lncRNAs).

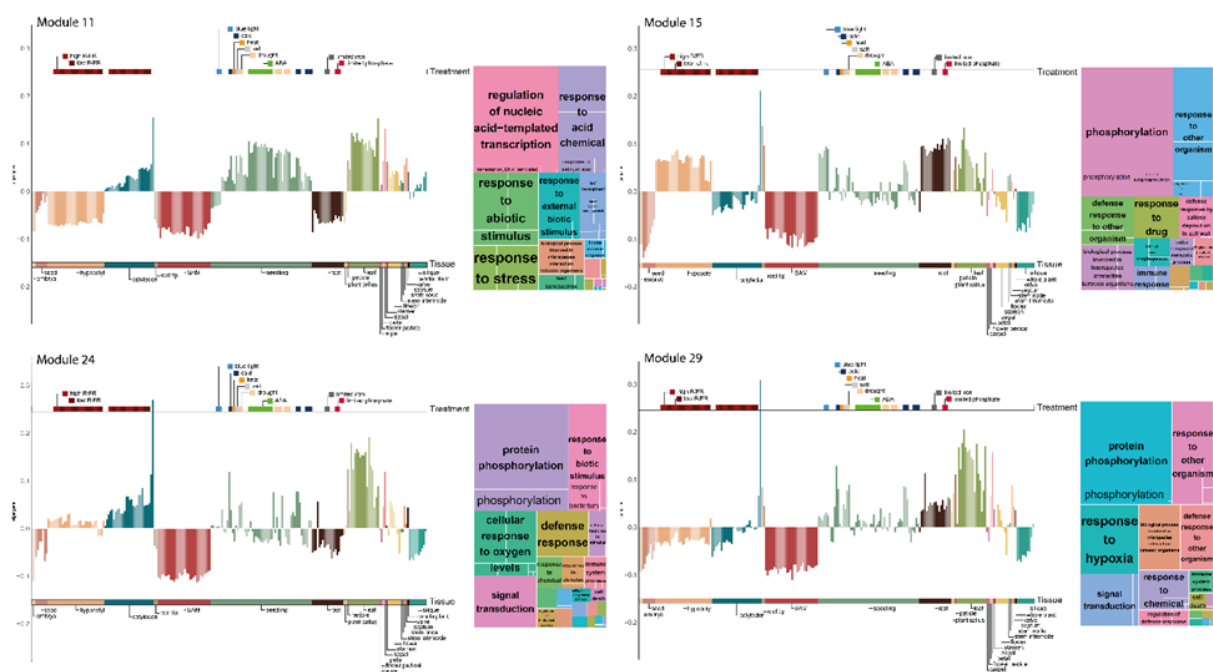


Figure S14. Eigengene expression per module for modules 11, 15, 24 and 29; response to pathogens functional category (4 modules with 72 lncRNAs).

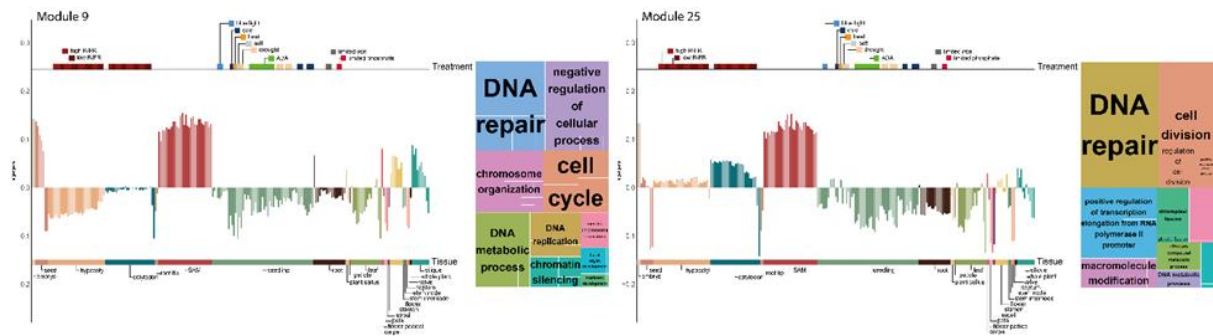


Figure S15. Eigengene expression per module for modules 9 and 25; DNA repair functional category (2 modules with 61 lncRNAs)

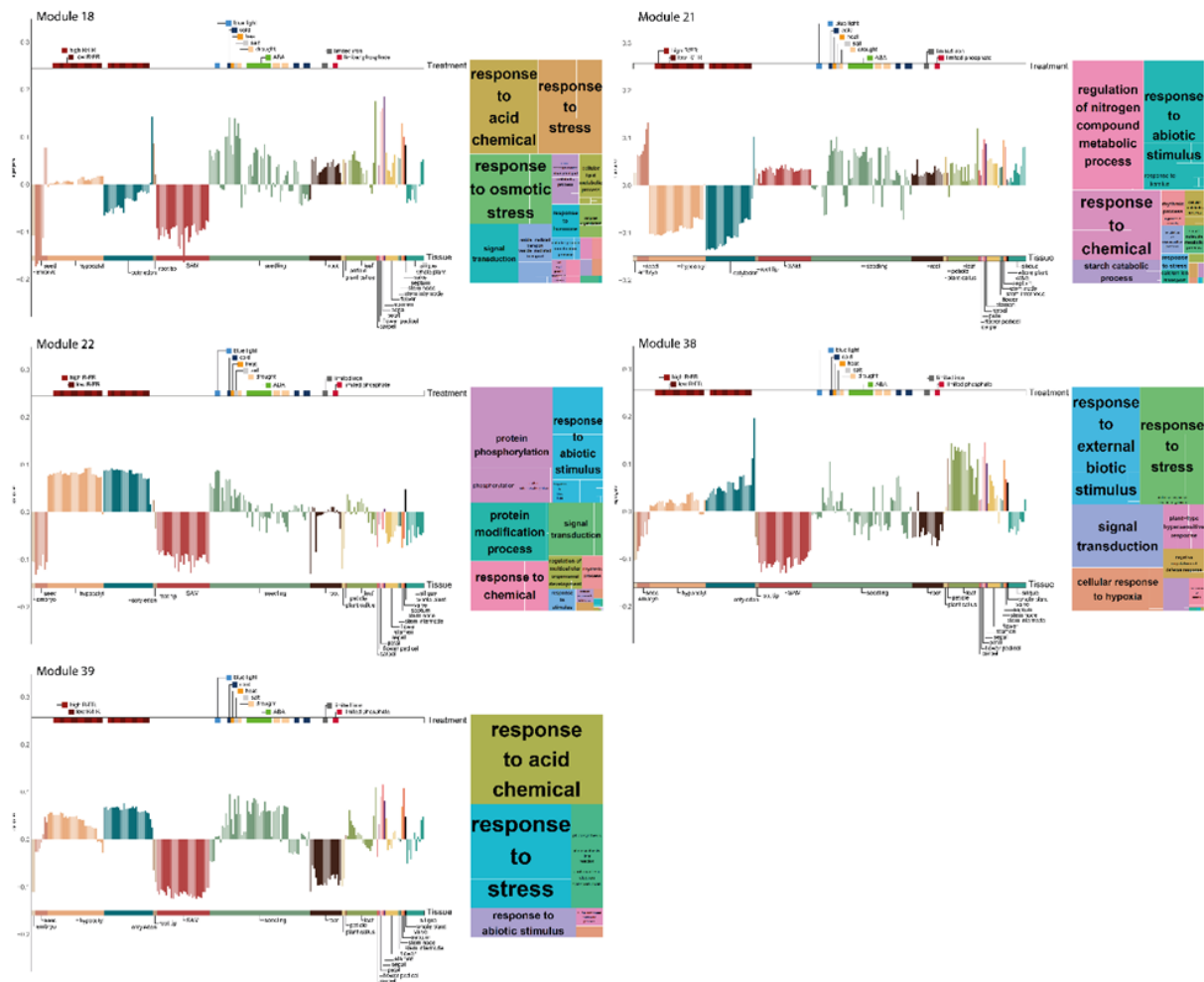


Figure S16. Eigengene expression per module for modules 18, 21, 22, 38 and 39; response to stress functional category (5 modules with 17 lncRNAs).

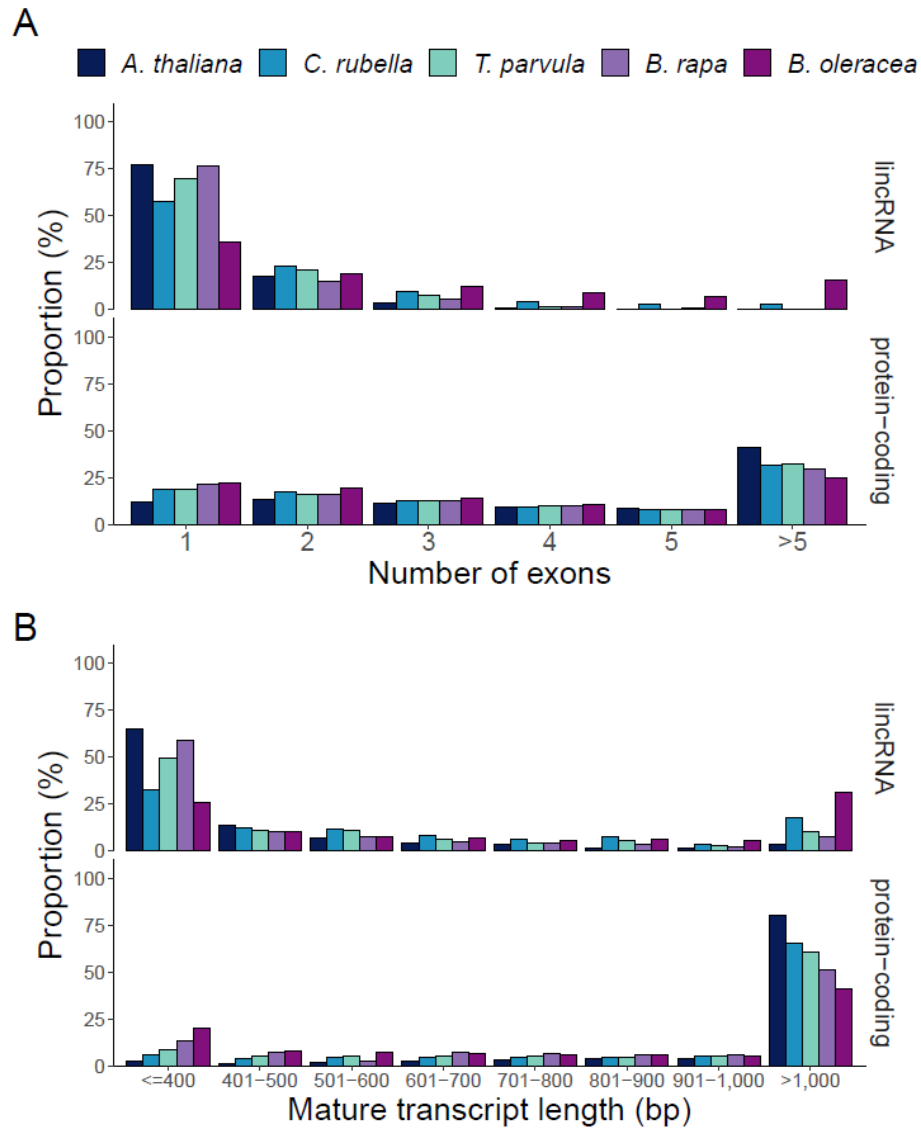


Figura S17. Distribución del número de exones y el tamaño del transcrito maduro en los lincRNAs identificados en cada una de las especies analizadas. (A) Número de exones y (B) tamaño del transcrito maduro de los lincRNAs identificados a partir de los transcriptomas generados en plántulas (parte aérea y raíces, ~ 8 días después de la germinación).

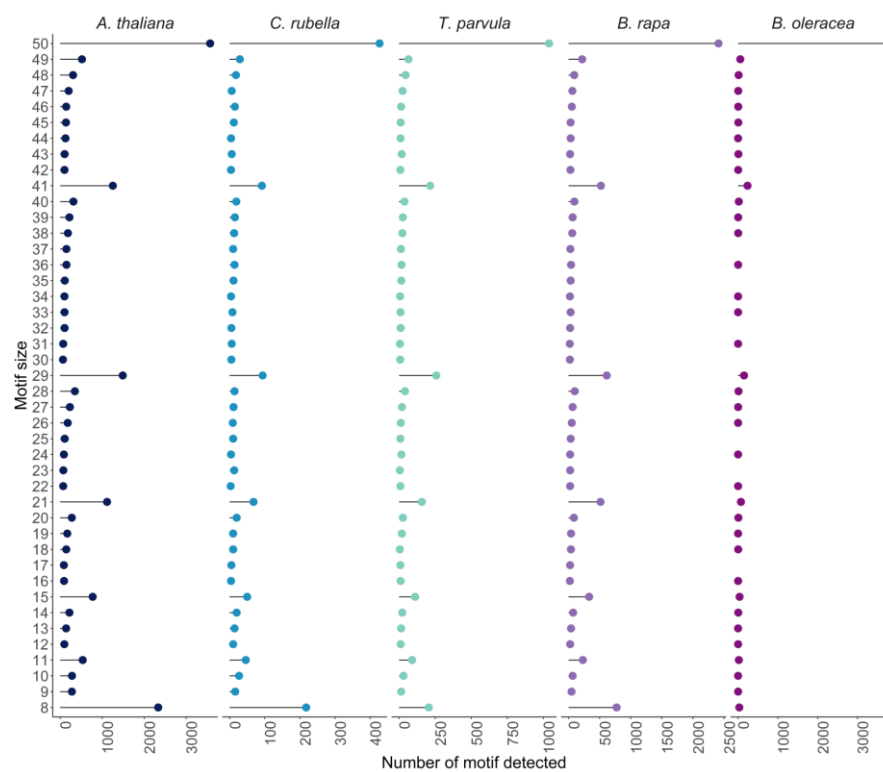


Figura S18. Tamaño de los motivos encontrados en el análisis de microhomología.

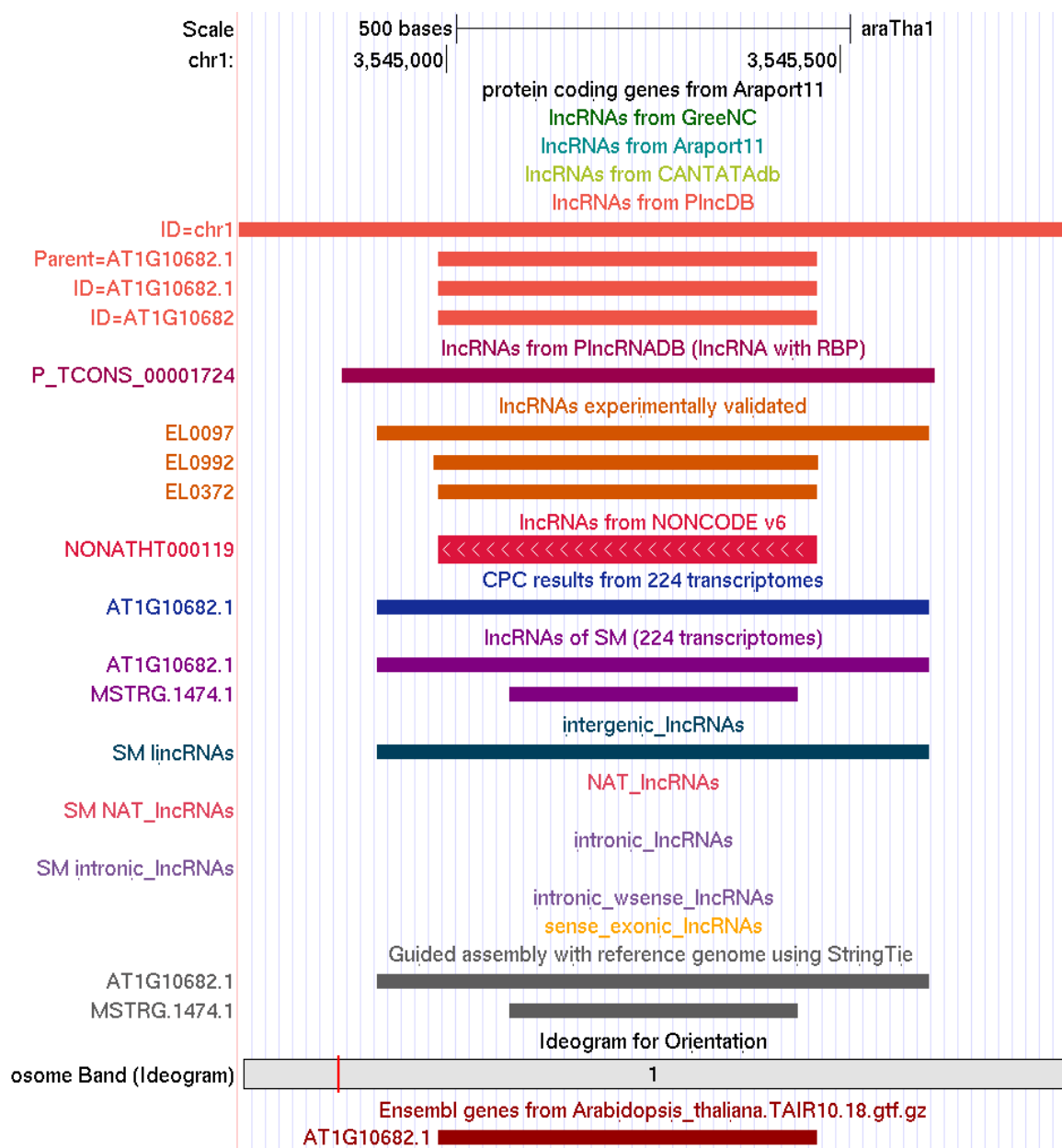


Figura S19. Visualización del lincRNA *ENOD40* en UCSC.

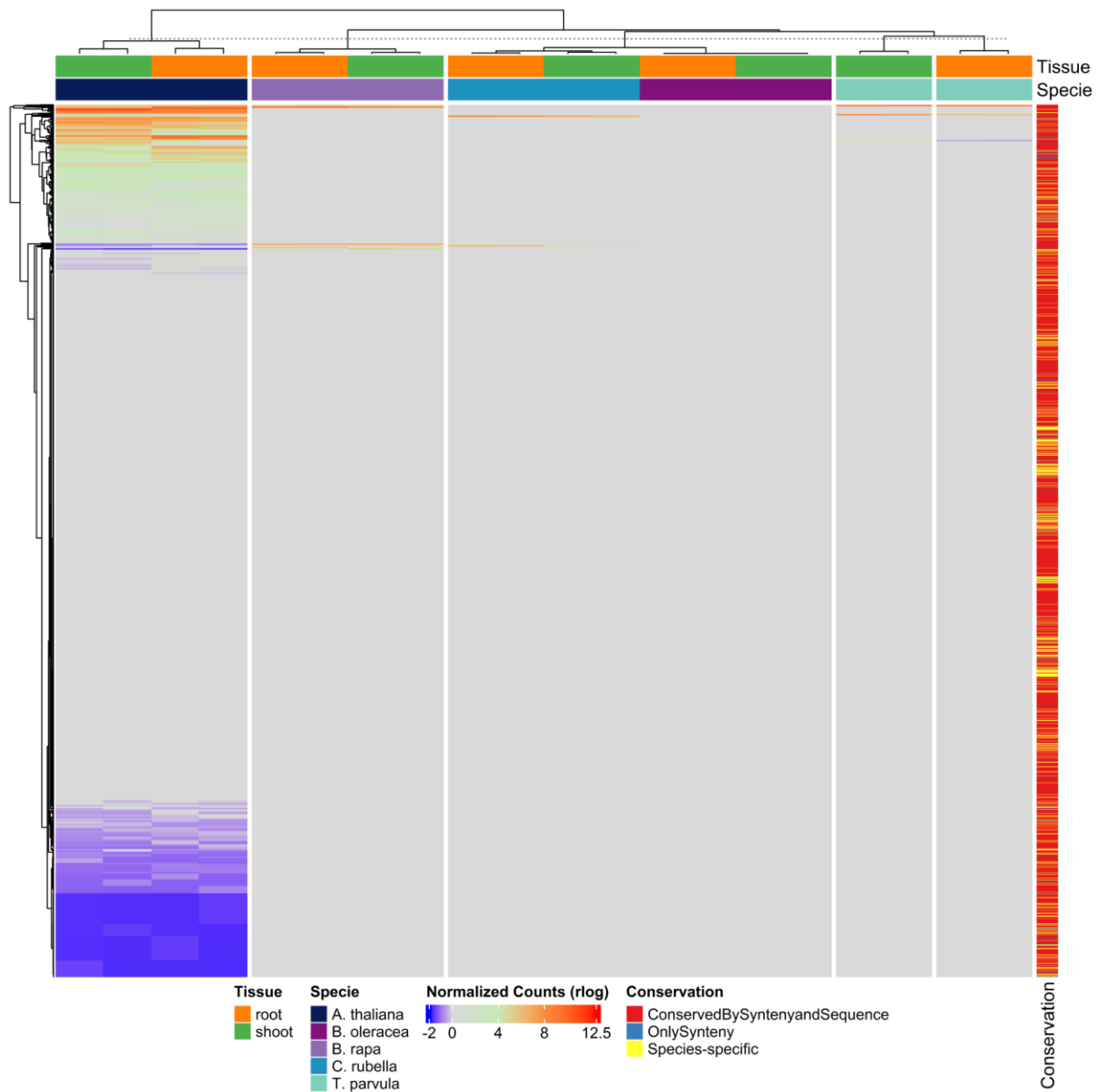


Figura S20. Patrones de expresión de los 2490 genes de lincRNAs identificados en *A. thaliana* en las otras especies de referencia. Los tipos de conservación identificados en cada lincRNA de *A. thaliana* se localizan en la parte derecha del heatmap. En la parte superior se encuentran las diferentes especies y órganos empleados para generar los transcriptomas (raíces y parte aérea).

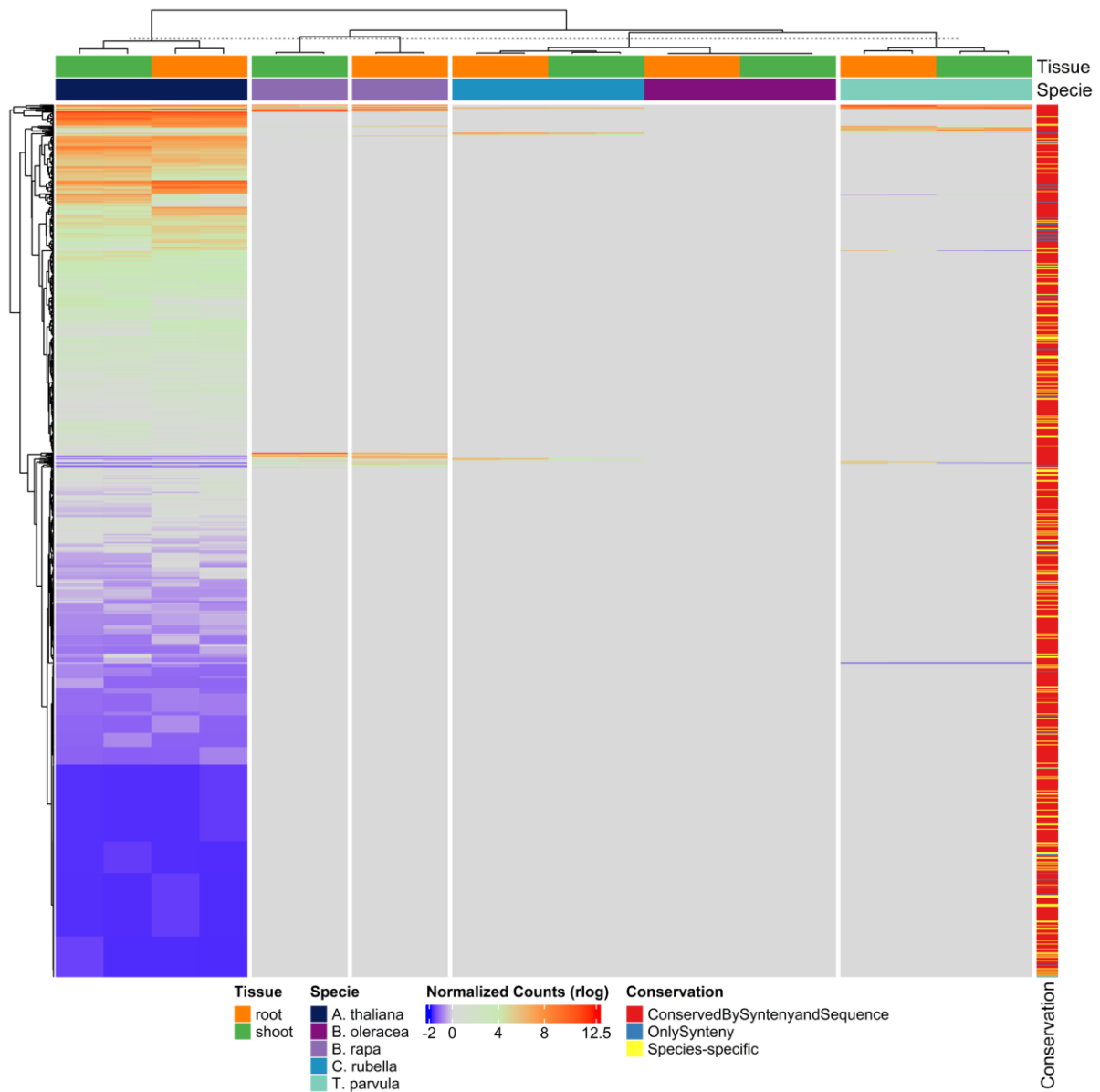


Figura S21. Patrones de expresión de los lincRNAs (990 genes) expresados en *A. thaliana* en las otras especies de referencia. De los 2490 genes, encontramos 990 genes con expresión en *A. thaliana*. Los tipos de conservación identificados en cada lincRNA de *A. thaliana* se localizan en la parte derecha del heatmap. En la parte superior se encuentran las diferentes especies y órganos empleados para generar los transcriptomas (raíces y parte aérea).

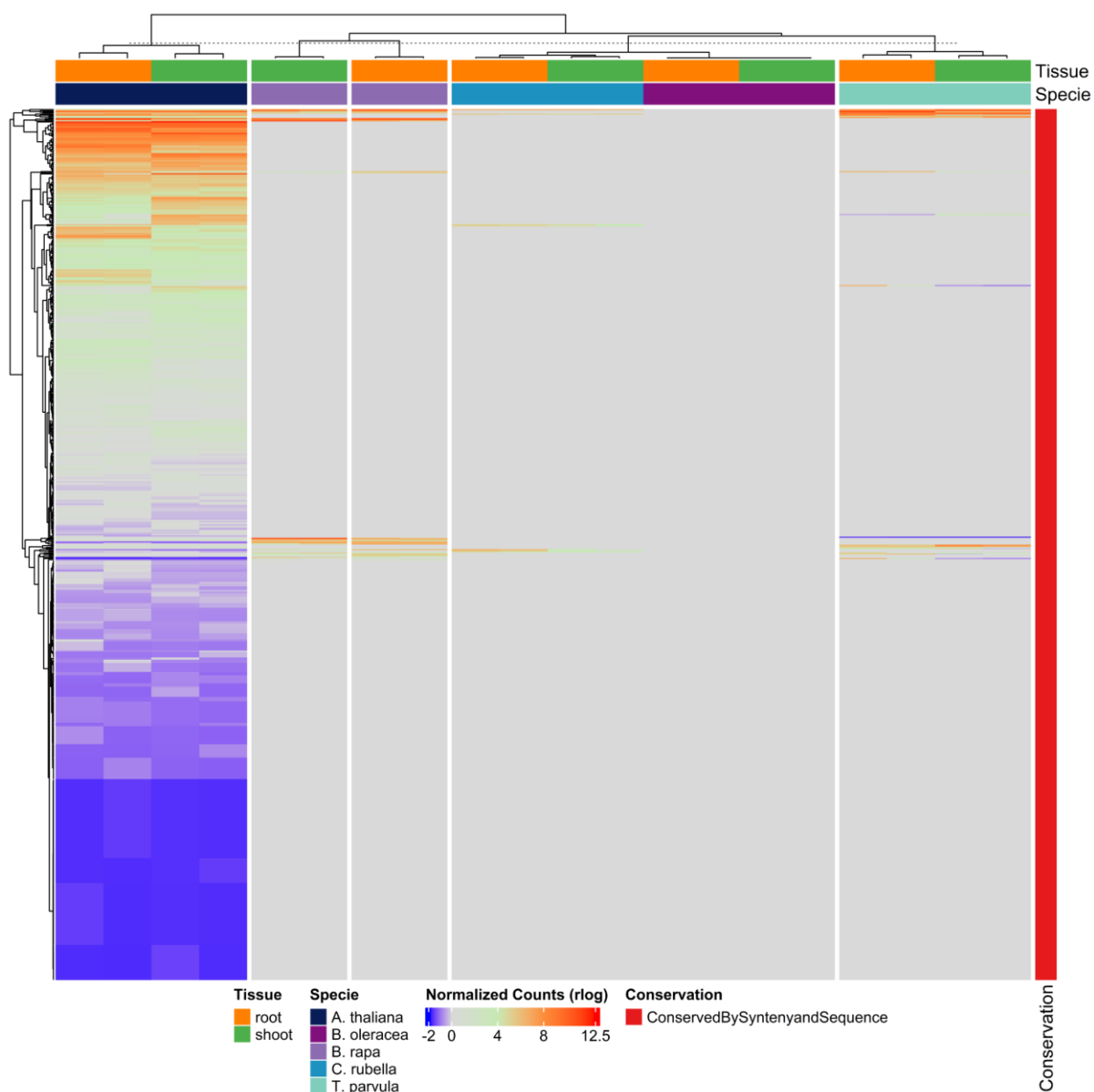


Figura S22. Patrones de expresión de los lincRNAs conservados por sintenia y microhomología (772 genes) en *A. thaliana*. Los tipos de conservación identificados en cada lincRNA de *A. thaliana* se localizan en la parte derecha del heatmap. En la parte superior se encuentran las diferentes especies y órganos empleados para generar los transcriptomas (raíces y parte aérea).

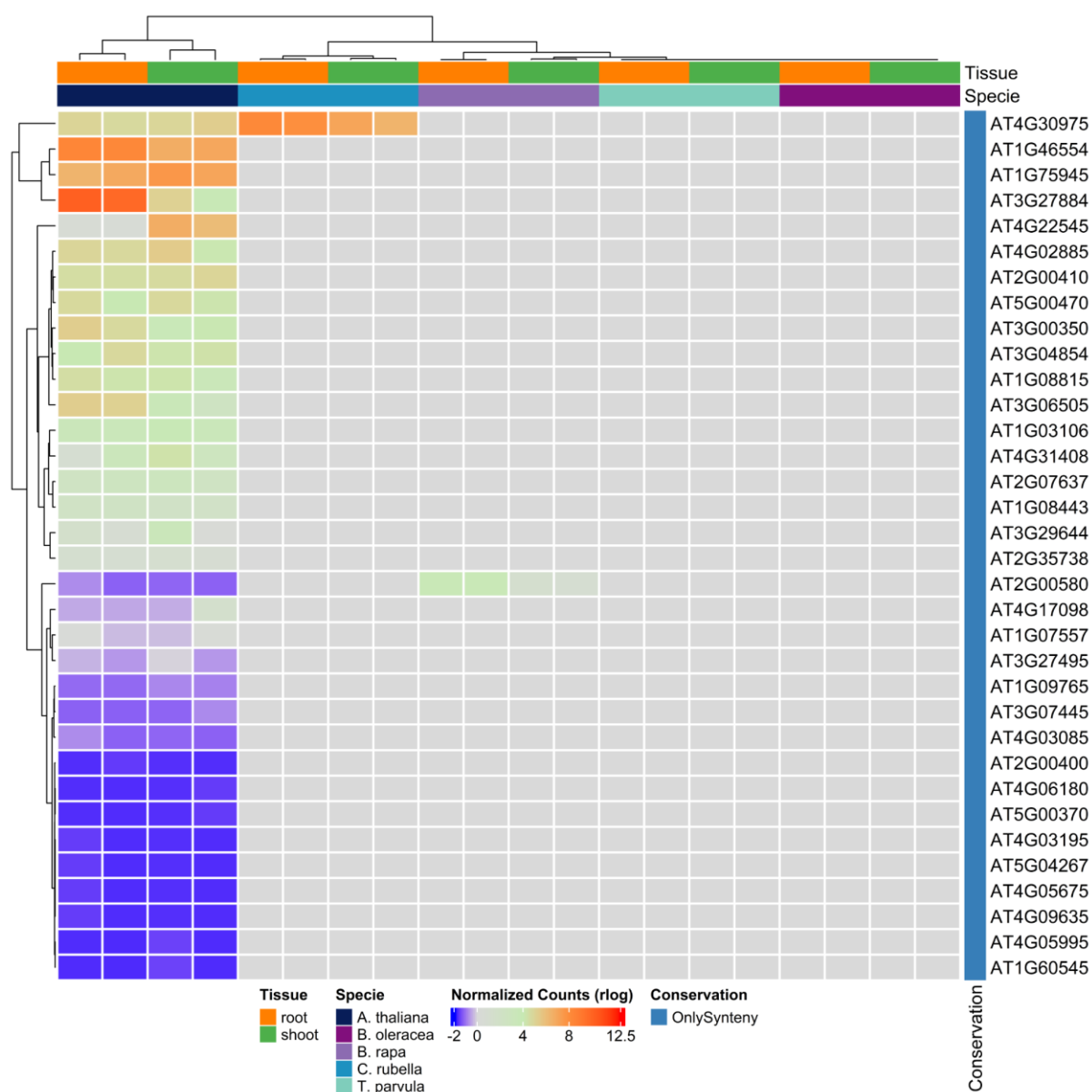


Figura S23. Patrones de expresión de los lincRNAs conservados solo por sintenia (34 genes) en *A. thaliana*. Los tipos de conservación identificados en cada lincRNA de *A. thaliana* se localizan en la parte derecha del heatmap. En la parte superior se encuentran las diferentes especies y órganos empleados para generar los transcriptomas (raíces y parte aérea).

Como parte final de esta tesis me gustaría recalcar que todos los resultados producidos en este trabajo de tesis, tienen como finalidad aportar y apoyar en el entendimiento de los lncRNAs y su evolución en las plantas. Por lo que, con los resultados de esta tesis se ha generado uno de los artículos contemplados, mientras que el otro permanece en proceso. A continuación enlisto el plan de publicación y el contenido de cada uno de los artículos:

- **Artículo 1:** Coss-Navarrete, Evelia L.*; Corona-Gómez, José Antonio*; García-López, Irving J; Pérez-Patiño, Jaime Alejandro; Fernández-Valverde, Selene L, 2022. *Co-autores. **“Transcriptome-guided annotation and functional classification of long non-coding RNAs in *Arabidopsis thaliana*”** publicado en [Scientific Reports](#).
 - En este artículo, generamos un nuevo archivo de anotación de lncRNAs, los clasificamos de acuerdo a su contexto genómico e inferimos sus posibles funciones biológicas mediante el empleo de redes de co-expresión.
- **Artículo 2:** Coss-Navarrete, Evelia L.*Corona-Gomez José Antonio, Gillmor Stewart, Hurst Laurence, Oktaba Katarzyna, y Fernandez-Valverde Selene L. Título propuesto: **“Syntenic and microhomology conservation of plant lincRNAs”**.
 - Este artículo tiene como objetivo analizar la conservación de los lincRNAs por sintenia y microhomología, empleando nueva anotación del artículo 1 para *A. thaliana* y generando nuevas anotaciones para 4 especies más de la misma familia. Con el fin de entender la expresión conservada de los lincRNAs en la parte aérea y raíces a través de las especies.
 - Actualmente se encuentra en escritura, por lo que, considerando que los análisis contemplados han sido finalizados, esperamos poder terminar este artículo y someterlo el año que viene.