UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escola Tècnica Superior d'Enginyeria Agronòmica i del Medi Natural

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Agronómica y del Medio Natural

Comparativa de herramientas para la predicción de ARNs no codificantes largos y análisis de su expresión diferencial en plantas de melón sometidas a frío

Trabajo Fin de Grado

Grado en Biotecnología

AUTOR/A: Núñez Salvador, Marta

Tutor/a: Forment Millet, José Javier

Cotutor/a externo: GOMEZ, GUSTAVO GERMAN

Director/a Experimental: CORELL SIERRA, JULIA

CURSO ACADÉMICO: 2021/2022

Título: Comparativa de herramientas para la predicción de ARNs no codificantes largos y análisis de su expresión diferencial en plantas de melón sometidas a frío.

Tittle: Comparison of tools for the prediction of long non-coding RNAs and analysis of their differential expression in melon plants subjected to cold.

Títol: Comparativa de ferramentes per a la predicció d'ARNs no codificants llargs y anàlisi de la seua expressió diferencial en plantes de meló sotmeses a fred.

Resumen

La respuesta de las plantas a las condiciones de estrés consiste en una compleja reprogramación de sus actividades transcripcionales, con el fin de reducir el impacto del estrés en su homeostasis fisiológica y celular. Entre los elementos reguladores de la respuesta a estrés destacan los ARNs no codificantes (ARNnc), que, pese a ser considerados inicialmente como material genético basura, en la actualidad se les atribuyen importantes papeles regulatorios de la expresión génica. Dentro de los ARNnc, aquellos con una longitud superior a 200 nucleótidos son clasificados como ARNs no codificantes largos (ARNncl) y, en comparación con animales, su estudio en plantas es aún incipiente. Por tanto, en el presente trabajo se ha realizado, en primer lugar, la evaluación de cinco programas de predicción de ARNncl (FEELnc, CPC2, CPAT, LncADeep, lncRNA_Mdeep) empleando cinco especies de plantas (*Arabidopsis thaliana, Cucumis sativus, Oryza sativa, Zea mays* y *Solanum lycopersicum*) en cada uno de ellos. Tanto CPC2 como FEELnc proporcionaron los mejores resultados, siendo CPC2 la herramienta más sensible y FEELnc la más precisa. En segundo lugar, se ha realizado un análisis de expresión diferencial de librerías de RNA-Seq de plantas de melón control y sometidas a frío. Para realizar este análisis se ha creado una anotación conjunta de transcritos codificantes y los ARNncl identificados por FEELnc. Con los resultados de este análisis se ha obtenido una lista de ARNncl y ARNs codificantes expresados diferencialmente sobre la que se ha realizado la búsqueda de potenciales ARNncl reguladores en *cis* de genes relacionados con la respuesta a estrés. Para ello se han localizado ambas clases de transcritos sobre el genoma y seleccionado aquellos que se encuentren a menos de 10 Kb. Entre los genes identificados, destacan los relacionados con el estrés oxidativo o la síntesis de calcio. En conclusión, este trabajo proporciona información relevante para el estudio de ARNncl en plantas, así como sobre el potencial papel regulatorio de ciertos ARNncl en la respuesta a estrés en melón.

Summary

The response of plants to stress conditions consists of a complex reprogramming of their transcriptional activities in order to reduce the impact of stress on their physiological and cellular homeostasis. Among the regulatory elements of the stress response, non-coding RNAs (ncRNAs) stand out, which, despite being initially considered as junk genetic material, are nowadays attributed important regulatory roles in gene expression. Within ncRNAs, those with a length of

more than 200 nucleotides are classified as long non-coding RNAs (lncRNAs) and, compared to animals, their study in plants is still incipient. Firstly, a benchmarking has been carried out in five lncRNA prediction programmes (FEELnc, CPC2, CPAT, LncADeep, lncRNA_Mdeep) using five plant species (*Arabidopsis thaliana, Cucumis sativus, Oryza sativa, Zea mays and Solanum lycopersicum*) in each of them. Both CPC2 and FEELnc provided the best results, with CPC2 being the most sensitive tool and FEELnc the most specific. Secondly, a differential expression analysis of RNA-Seq libraries from control and cold stress melon plants was performed. To carry out this analysis, was created a joint annotation of coding transcripts and ncRNAs identified by FEELnc. With the results of this analysis, a list of differentially expressed lncRNAs and coding RNAs was obtained, on which was carried out the search for potential cis-regulatory lncRNAs of genes related to the response to stress. Both classes of transcripts were located on the genome and selecting those that are less than 10 Kb away. Among the identify genes, those related to oxidative stress or calcium synthesis stand out. In conclusion, this work provides relevant information for the study of lncRNAs in plants, as well as on the potential regulatory role of certain lncRNAs in the response to stress in melon.

Palabras clave: ARNs no codificantes largos; expresión diferencial; evaluación comparativa; estrés; plantas; melón

Key words: long non-coding RNAs; differential expression, benchmarking; stress; plants; melon

Alumno/a: Marta Núñez Salvador

Localidad y fecha: Valencia, Julio 2020

Tutor académico: Prof. D**.** José Javier Forment Millet

Cotutor: D. Gustavo Gómez

Director experimental: Dña. Julia Corell Sierra

AGRADECIMIENTOS

En primer lugar, agradecer al Dr. Gustavo Gómez por darme la oportunidad de realizar este proyecto en su laboratorio. Al resto del grupo, Joan, Gabriela, Cristina, Antonio, gracias por ayudarme en todo lo que he necesitado y hacerme sentir tan cómoda y en especial, gracias a Julia y Pascual por escucharme siempre, apoyarme y no dejar que me desanime, pero, sobre todo, por hacerme ver que este sí puede ser mi camino.

Gracias a mi grupo de amigas de la universidad por todo el apoyo y por hacer que hasta de los peores momentos recuerde cosas buenas, no podría haber tenido más suerte con vosotras. Al resto de personas que me habéis acompañado durante estos años y habéis confiado en mí, deciros que gracias a vosotros el camino ha sido mucho más fácil. En especial, Andrea, gracias por levantarme el ánimo siempre que me venía abajo y por estar siempre dispuesta a echarme una mano, aunque tú también la necesitases, sin ti no hubiera sido lo mismo.

Por último, agradecer a toda mi familia por creer en mí y en especial a mis padres por estar siempre conmigo.

I. INDICE

II. ÍNDICE DE FIGURAS

III. ÍNDICE DE TABLAS

LISTA DE ABREVIATURAS

ADN	Ácido Desoxirribonucleico
APOLO	*lncRNA Auxin Regulated Promoter Loop*, Bucle del ARNncl Promotor regulado por la auxina
ARN	Ácido Ribonucleico
ARNm	ARN mensajero
ARNnc	ARN no codificante
ARNnci	ARN no codificante intronico
ARNncil	ARN no codificante intergénico largo
ARNncl	ARN no codificante largo
Ca^{+}	Ion de Calcio
COLDAIR /COOLAIR	*Cold assisted intronic noncoding RNAs*, ARNs no codificantes intrónicos asistido por frío
FN	Falsos negativos
FN_v	Falsos negativos validados
FP	Falsos positivos
GO	*Genic ontology,* Ontología génica
GTF	*Gene Transfer Format File*, Archivo de formato de transferencia de genes
HSVd	*Hop Stunt Viroid,* Viroide del Enanismo del Lúpulo
ID	Identificación
IPS1	*Inositol-3-phosphate synthase isozyme 1,* Inositol-3-fosfato sintasa isoenzima 1
Kb	Kilobases
LDMAR	*Long Day specific Male fertility associated RNA*, ARN específico para la fertilidad masculina
TAN	Transcritos anti-sentido naturales
ORF	*Open Reading Frame*, Marco de lectura abierto
PI	Punto isoeléctrico
ROS	*Reactive Oxigen Specie*, Especie Reactiva de Oxigeno
TB	TeraByte
TSV	*Tab Separated Values*, Valores Separados por Tabuladores
VN	Verdaderos negativos

VP	Verdaderos positivos
VPN	Valor Predictivo Negativo
VP_v	Verdaderos positivos validados
ARNnce	ARN no codificante exónico
Log2FC	Logaritmo en base 2 del *FoldChange*

1. INTRODUCCIÓN

Una de las principales características de las plantas, aunque no exclusiva ni excluyente, es que son organismos sésiles y en consecuencia sufren numerosos estreses ocasionados por el ambiente que les rodea, a los que deben hacer frente. A causa del cambio climático y el deterioro ambiental, los cultivos están sometidos cada vez a más condiciones adversas que limitan tanto su desarrollo como su productividad, ocasionando grandes pérdidas económicas y agrícolas. (Shahzad et al., 2021).

Gracias al desarrollo de nuevas técnicas de cultivo molecular e ingeniería genética es posible la introducción de tolerancia a estreses, pero para ello primero se debe entender el mecanismo molecular que tiene la respuesta en la planta.

1.1. INTERACCIÓN PLANTA – ESTRÉS ABIÓTICO

Cuando una planta se encuentra sometida a estrés es debido a que las condiciones del ambiente no son óptimas para su crecimiento, impidiendo la normal división y expansión de sus células. Frente a dichas condiciones, algunos de los cambios que se producen en la planta son respuestas no adaptativas, que reflejan el daño infligido. Sin embargo, la gran mayoría de los cambios son respuestas adaptativas, fisiológicas y celulares, tales como la activación de un gran número de mecanismos de regulación génica, encargados de reestablecer la homeostasis en el organismo y proporcionar resistencia (Zhang et al., 2022). Las moléculas implicadas en dicha respuesta adaptativa son un objetivo potencial para la mejora de los cultivos.

Tras la exposición de las plantas a ciertos estreses se pueden encontrar flujos iónicos alterados, desequilibrio fitohormonal, generación de especies reactivas de oxígeno (ROS), activación de quinasas y señalización mediada por hormonas, entre otros. (Mehrotra et al., 2020)

1.1.1. ESTRÉS POR BAJAS TEMPERATURAS

El estrés causado por bajas temperaturas (BT) se puede dividir en estrés BT leve (<20 °C) y congelante (<0 °C). Suelen experimentarlo las plantas que se cultivan en climas tropicales, como el melón (*Cucumis melo*), mientras que aquellas cultivadas en climas más templados, si son sometidas a temperaturas frías de forma progresiva, pueden aclimatarse y tolerar temperaturas incluso congelantes (Nurhasanah Ritonga & Chen, 2020).

Debido a que la expresión génica varía de forma específica según el tipo de estrés al que se encuentre sometida la planta, se puede intuir que estas pueden detectar y diferenciar entre estreses. Para entender como la planta responde a un estrés en concreto, en este caso las bajas temperaturas, es necesario conocer los mecanismos mediante los que es capaz de distinguir el frío de otros cambios ambientales (Mehrotra et al., 2020). Sin embargo, existe una gran dificultad para

identificar sensores celulares, y más aún, si se trata de sensores de señales físicas. Por lo que los pocos sensores identificados han sido resultado de estudios indirectos. La dificultad de la identificación recae en que un sensor puede ser una molécula única proveniente de un único gen; pero normalmente provienen de un conjunto de genes redundantes, provocando que la pérdida de función en un gen concreto no ocasione diferencias fenotípicas a pesar de sí estar implicado en la detección del estrés (Zhu, 2016).

Tras la detección del estrés por frio, ocurre una serie de cambios en el citoplasma celular, tales como la variación en la concentración de Ca^{+} y ROS, haciendo función de señalizador. Estos cambios celulares provocan la activación de mecanismos reguladores como ciertos factores de transcripción, responsables de la expresión de genes para combatir los daños sufridos por el frío. Además, en los últimos años se ha prestado especial atención al mecanismo regulador de otras moléculas denominadas ARN no codificante (ARNnc) (J. Wang et al., 2017).

1.1.2. EXPRESIÓN DE ARN NO CODIFICANTE

La identificación del ARN no codificante se ha dado a través de la tecnología de secuenciación masiva, que ha permitido identificar un gran número de ARNs no codificantes que se ubican en las regiones no anotadas del genoma (Nakaminami et al., 2012). Dentro de los ARNs no codificantes se puede encontrar una gran diversidad de mecanismos de acción, dianas y orígenes genómicos, que tienen una función clave en la regulación de la expresión génica (Liu et al., 2012.). A nivel funcional se catalogan como ARNnc estructurales y ARNnc reguladores, dentro de los cuales se dividen según su tamaño en ARNnc pequeños (inferiores a 200 nucleótidos) y ARNnc largos (mayores a 200 nucleótidos) (Ahmed et al., 2020) (Figura 1). Asimismo, dentro de los ARNs reguladores también se debe incluir el ARN circular (ARNcirc).

El ARNnc regulador, tal y como sugiere su nombre, es aquel que participa en la regulación de la expresión génica y en la modulación de la estabilidad y traducción del ARN. Uno de los procesos en los que se ve involucrado es en el empalme alternativo, el cual produce un aumento en la diversidad de proteínas, fenómeno que en plantas suele estar relacionado con condiciones de estrés. También participa en el control de la dinámica de degradación y estabilización del ARNm. Mediante los estudios llevados a cabo en *Arabidopsis thaliana* para el desarrollo de perfiles de expresión en una planta de referencia, se han podido explorar más profundamente las redes de señalización de estrés e identificar nuevos ARN no codificantes implicados en la respuesta a estrés, como los ARN intergénicos no codificantes largos COLDAIR y COOLAIR, inducidos por el frío (Nakaminami et al., 2012).

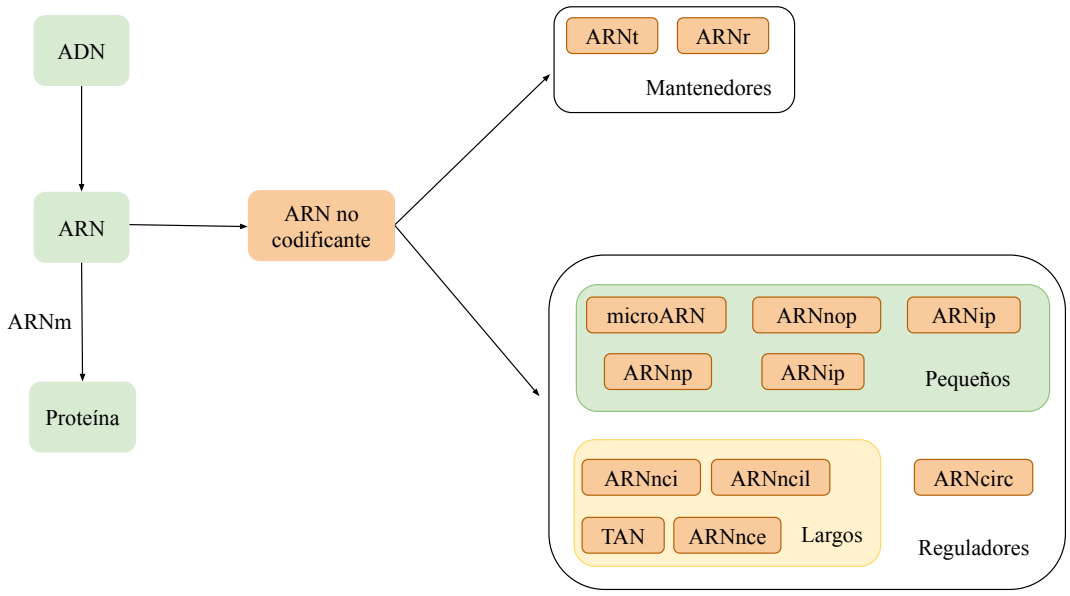


Figura 1. Clasificación de ARN no codificante. Dentro del ARNnc mantenedor se encuentre el ARN ribosómico (ARNr) y ARN de transferencia (ARNt). Por un lado, el ARNnc regulador pequeño puede ser micro-ARN, ARN nucleolar pequeño (ARNnop), ARN de interferencia pequeño (ARNip), ARN nuclear pequeño (ARNnp) y ARN con interacción piwi (ARNip). Por otro lado, el ARNnc regulador largo puede ser ARN no codificante intergénico largo (ARNncil), ARN no codificante intronico (ARNnci), ARN no codificante exónico (ARNnce) y transcritos antisentido naturales (NAT). También está el ARN circular como regulador. Adaptado de Ahmed et al., 2020

1.2. ARNS NO CODIFICANTES LARGOS

Gracias a los grandes avances en el estudio de la dinámica y propiedades del genoma, se conoce que a pesar de que la mayor parte de él se transcribe, solo una pequeña fracción tiene potencial codificante (Wilhelm et al., 2008). La gran cantidad de ARN no codificante, que hasta hace no demasiado tiempo era considerado como ADN basura, ha cobrado gran relevancia en la última década debido a su importante función reguladora, tal y como se ha comentado anteriormente. Se considera que la complejidad de un organismo no recae tanto en aquellos genes codificantes de proteínas si no en el repertorio de ARN no codificante (Quinn & Chang, 2015).

A la hora de definir que es un ARN no codificante largo se debe tener en cuenta que no se les ha prestado tanta atención como a otros ARN no codificantes hasta hace pocos años. Por tanto, el conocimiento de sus funciones y características es limitado, especialmente en plantas, ya que en animales sí estan mejor caracterizados. De forma arbitraria, se considera que un ARN no codificante es largo cuando cuenta con más de 200 nucleótidos y contiene pautas de lectura abierta cortas (<100 nucleótidos) que dan lugar a péptidos muy pequeños (Tian et al., 2019). Debido a la generalidad de dicha definición, es muy probable que dentro de los ARNncl exista una gran variedad de funciones, careciendo hasta día de hoy de la identificación de un mecanismo de acción común.

Diversos informes muestran que los ARNncl pueden intervenir en una amplia variedad de procesos biológicos como la transcripción de genes, las modificaciones post-traduccionales, la traducción, la interferencia transcripcional, la modificación de proteínas y la metilación del ADN (Tian et al., 2019).

Cabe destacar que este grupo de ARNs no codificantes contiene bastantes características comunes con el ARNm, ya que suelen tener una cola poli A, CAP 5' y estar formados por distintos exones, por lo que puede tener lugar el empalme alternativo. A pesar de tener estas similitudes con los ARNm, los exones suelen ser menos abundantes y más largos y su nivel de expresión significativamente menor. En cuanto a su expresión, pueden ser transcritos por la polimerasa II, pero también por la polimerasa IV y V y a pesar de que los transcritos originados a partir de la polimerasa IV y V están menos descritos debido a su baja tasa de expresión, se conoce que no se produce la adición de cola poli A en el extremo 3'(Quinn & Chang, 2015).

Los ARNncl pueden encontrarse distribuidos a lo largo de todo el genoma, y dependiendo de su localización pueden clasificarse como ARNncl intrónico; cuando se encuentran en las regiones intrónicas de un gen, ARNncl exónico; cuando se encuentran en las regiones exónicas de un gen o ARNncl intergénico; cuando no se encuentran en la pauta de lectura de un gen. También se pueden clasificar según su dirección como sentido o anti-sentido (Figura 2). Los ARNncl en dirección anti-sentido también son conocidos como transcritos anti-sentido naturales (TAN) (Budak et al., 2020).

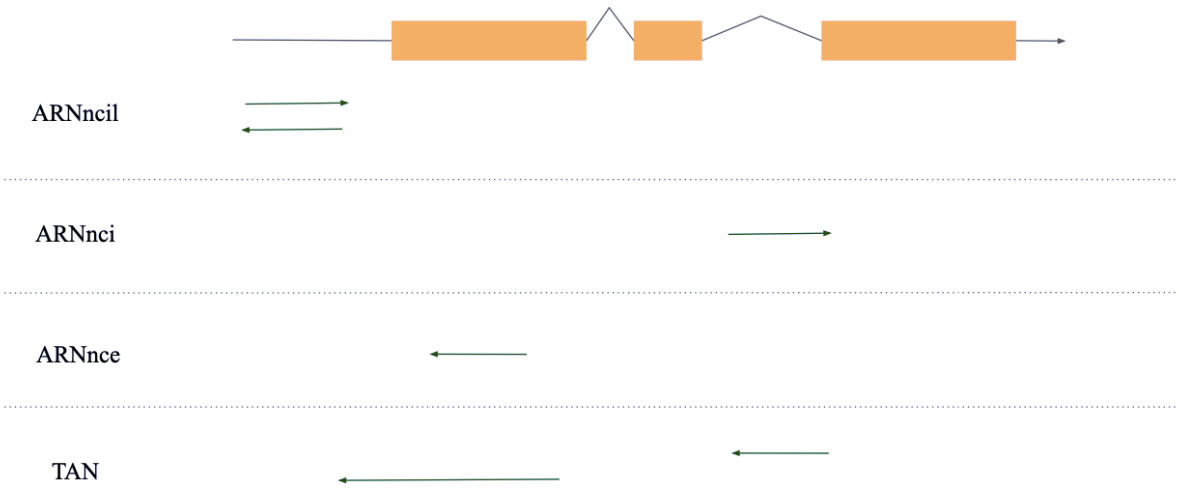


Figura 2. ARNncl según su dirección y localización en el genoma. Se clasifican en ARN no codificante intrónico (ARNnci), ARN no codificante intergénico largo (ARNncil), ARN no codificante exónico (ARNnce) y transcritos anti-sentido naturales (TAN), Adaptado de Budak et al., 2020.

Actualmente no existe una caracterización funcional de ningún ARNncl exónico en dirección sentido en plantas, pero sí del resto de clasificaciones. El grupo más estudiado dentro de los ARNncl son los intergénicos, y entre ellos se encuentran LDMAR, APOLO y IPS1. Este último actúa como señuelo

simulando regiones de una proteína objetivo. En cuanto a los ARNncl intrónicos, el siguiente grupo más estudiado, COOLAIR tiene función señalizadora en el proceso de floración. Por último, un ejemplo de un TAN es HvCesA6, el cual está implicado en la regulación de la síntesis de la pared celular (Budak et al., 2020)

Como consecuencia del desarrollo de bases de datos públicas en las que se acumula información genética de plantas se ha podido acelerar la identificación de un gran número de transcritos, y con la ayuda de herramientas bioinformáticas de predicción se ha podido identificar y separar según sus características los ARNncl del resto de transcritos. Es posible, incluso, el estudio funcional de dichos ARNncl y su mecanismo de acción. Una gran cantidad de herramientas se han desarrollado con dicha finalidad y se basan en el estudio de las posibles interacciones de ARNncl con otras moléculas como ARNm, ADN o microARN. Al estudiar la interacción del ARNncl con el microARN lo que se encuentra son posibles ARNs endógenos competitivos (ARNec), es decir, ARNncl que se encargan de "secuestrar" microARN evitando así el silenciamiento del gen diana (J. Wang et al., 2017).

A la hora de usar estas herramientas nunca se debe olvidar que los resultados son predicciones y siempre puede haber una tasa de error, por lo que para validar la función de un ARNncl será necesario hacerlo de forma experimental.

1.3. HERRAMIENTAS DE PREDICCIÓN DE ARNncl

Debido a la poca cantidad de ARNncl identificado en plantas, en comparación con otros tipos de ARN, resulta necesario el desarrollo de buenas herramientas de predicción que faciliten esta tarea. Con el uso de dichas herramientas se obtienen resultados *in silico* que pueden servir de guía en el diseño experimental (Pinkney et al., 2020).

Normalmente, la anotación de ARNncl se lleva a cabo con datos de secuenciación de ARN, por lo que el primer paso siempre será el ensamblaje de los transcritos. El ensamblaje puede llevarse a cabo alineando contra el genoma de referencia, en caso de contar con uno o mediante el ensamblaje *de novo*, en caso de no disponer de un genoma de referencia (Budak et al., 2020). Para que no haya problemas con los ARNnc pequeños siempre se hará un filtrado del ensamblaje para eliminar todos aquellos ARNs con menos de 200 nucleótidos.

Una vez se tienen los transcritos ensamblados y se ha hecho el filtrado ya pueden ser introducidos en una herramienta de predicción que se encargará de separar aquellos que considere ARNncl de aquellos que considere ARNm. Dichas herramientas presentan un gran desafío en su desarrollo debido al elevado número de características comunes que presentan los ARNncl y los ARNm. Además, cabe destacar que la adecuada predicción no dependerá solo del buen desarrollo de la herramienta sino también de la calidad de los datos de secuenciación. Los datos con una baja calidad tendrán mayores posibilidades de error (Budak et al., 2020).

1.3.1. PROPIEDADES DE LAS HERRAMIENTAS DE PREDICCIÓN

Los programas de predicción de ARNncl suelen tener en cuenta no solo la longitud de los transcritos si no también la del marco de lectura abierta, ya que tal y como se ha comentado anteriormente el marco de lectura abierta de los ARNncl se caracteriza por ser menor a 100 nucleótidos. También tienen en cuenta si estos marcos de lectura abierta cortos pueden codificar alguna proteína funcional midiendo el potencial de codificación.

El potencial de codificación puede calcularse teniendo en cuenta ciertas características a parte de la longitud e integridad de la ORF, como la frecuencia de nucleótidos, la cual puede medirse a partir del contenido de GC, el código de prueba de Ficket, la frecuencia de K-meros y la puntuación de hexámeros, o propiedades termodinámicas como la energía libre mínima del ARN y el punto isoeléctrico. Los distintos programas realizan dicho calculo y proceden a la posterior clasificación en función de los resultados, empleando algoritmos tanto de aprendizaje automático, como la máquina de soporte de vectores, bosque aleatorio o regresiones logísticas; como de aprendizaje profundo como las redes neuronales (Klapproth et al., 2021).

1.3.2. APRENDIZAJE AUTOMÁTICO Y APRENDIZAJE PROFUNDO

El aprendizaje automático está relacionado con la estadística computacional, la cual está ligada estrechamente con la elaboración de predicciones. Gracias al uso de algoritmos de aprendizaje automático la precisión de los cálculos del potencial de codificación es mucho más elevada y por tanto la predicción más precisa. Estos algoritmos cuentan con una primera fase de entrenamiento en la que aprenden, a partir de una serie de datos, a detectar patrones y tomar decisiones basadas en la información que recogen. Este entrenamiento les permite aplicar aquellos patrones detectados a nuevas series de datos y de esta forma clasificar en grupos (Mohammed et al., 2019).

El aprendizaje automático puede ser supervisado o no supervisado, durante el presente trabajo se trabajará con aprendizaje automático supervisado. Durante la fase de entrenamiento de dicho algoritmo se emplean datos previamente etiquetados como entrada, conociendo previamente las predicciones que debería generar (Ongsulee, 2018).

En los últimos años se han empezado a desarrollar algoritmos de aprendizaje profundo, llevando un paso más allá el aprendizaje automático. Estos algoritmos generan redes neuronales artificiales utilizando muchas capas sucesivas de procesamiento. Cada capa tiene como entrada la salida de la capa anterior y mediante una transformación no lineal se consigue la extracción de características para la clasificación (Ongsulee, 2018).

En referencia al presente trabajo, nos interesa las herramientas de predicción de ARNncl que usan algoritmos de aprendizaje automático o profundo supervisado y que no requieren un alineamiento con genoma de referencia. Dentro del aprendizaje automático se encuentra Coding Potencial

Calculator (CPC), una de las herramientas más antiguas y más citada y su versión más actualizada CPC2, la cual tiene un mayor rendimiento en la predicción de ARNncl (Kang et al., 2017) . También se encuentra entre las más citadas CPAT (L. Wang et al., 2013)y FEElnc (Wucher et al., 2017).

Debido a la novedad y complejidad de las herramientas que usan algoritmos de aprendizaje profundo, estas cuentan con muchas menos citaciones en comparación con el grupo anterior, pero entre las más citadas se encuentra LncADeep, que incorpora las características en una red neuronal profunda a partir de la que construye modelos de predicción (Yang et al., 2018) y lncRNA_Mdeep, la cual usa un marco multimodal (Fan et al., 2020).

1.3.3. EVALUACIÓN COMPARATIVA DE HERRAMIENTAS DE PREDICCIÓN

Debido a que siempre que se publica una nueva herramienta los responsables de su evaluación son los propios desarrolladores, puede caerse en la llamada “trampa de la autoevaluación” en la que debido a una evolución no sistémica pueden generarse sesgos en los resultados. Una buena solución para estas situaciones es la evaluación comparativa por parte de personas externas al desarrollo de la herramienta, la cual proporciona información de gran ayuda para otros investigadores (Mangul et al., 2019).

A la hora de realizar una evaluación comparativa se pueden utilizar datos simulados o datos experimentales, pero siempre teniendo en cuenta las limitaciones que presentan los datos simulados, ya que nunca van a mostrar totalmente la variabilidad que presentan los datos experimentales (Mangul et al., 2019). Además, para medir el rendimiento de las diferentes herramientas que se van a comparar es muy útil el cálculo de ciertos parámetros como la precisión, sensibilidad, especificidad o exactitud.

1.4. ESPECIES DE INTERÉS

En el presente trabajo se trabaja con *Cucumis melo*, de la familia de las cucurbitáceas, más comúnmente denominada melón, para el estudio de expresión diferencial de ARNlnc en condiciones de estrés por frío, ya que es un cultivo propio de climas secos y cálidos. Esto es debido a que es una especie de interés como modelo para la investigación de diversos procesos biológicos, tales como la tolerancia a estrés (Tian et al., 2019). El estudio de esta especie también resulta interesante debido a que es un importante cultivo hortícola en todo el mundo y su mejora frente a condiciones de estrés es de gran interés agronómico y económico.

Los estudios realizados en cucurbitáceas se centran principalmente en la resistencia a enfermedades y la calidad del fruto, sin embargo, a pesar de existir algunos estudios sobre la resistencia a estreses abióticos, aún queda mucho que profundizar en este campo, por ello en trabajos anteriores del grupo con melón, se ha estudiado la respuesta de la planta a diversas condiciones de estrés

simultáneamente, identificando 22 familias de micro-ARNs que responden a estreses dobles o triples (Villalba-Bermell et al., 2021).

Además, gracias a la existencia de bases de datos públicas como Melonomics (Ruggieri et al., 2018) o Cucurbitgenomics (Zheng et al., 2019), en las que se almacena la secuencia y anotación del genoma completo (Garcia-Mas et al., 2012), y al desarrollo de herramientas genéticas y moleculares, la búsqueda de nueva información genética resulta más sencilla y precisa.

Por otro lado, para realizar la comparativa de herramientas de predicción de ARNncl se han seleccionado 5 especies de interés agronómico y de las que se ha podido encontrar una amplia base de datos de ARNncl. Entre las especies seleccionadas está *Arabidopsis thaliana,* especie más comúnmente utilizada como organismo modelo del reino de las plantas en experimentación y en consecuencia la mejor y más ampliamente caracterizada de dicho reino. Por otro lado, también se han elegido dos de los cereales con mayor interés económico mundial, al ser muy comunes en la comida tradicional de muchas culturas, *Zea mays* (maíz) *y Oryza sativa* (arroz). Ambos cereales son especies monocotiledóneas, en contraposición a las otras tres que son dicotiledónea*s.* Por último, se seleccionó *Solanum lycopersicum,* al tratarse de uno de los cultivos hortícolas más estudiados *y Cucumis sativus* por pertenecer a la familia de las cucurbitáceas.

2. OBJETIVOS

El presente trabajo cuenta con dos objetivos principales. El primero de ellos se trata de realizar un estudio comparativo de cinco herramientas de predicción de ARNncl en cinco especies de plantas para medir su capacidad de diferenciar el ARNncl del ARNm.

En segundo lugar, se pretende llevar a la práctica el programa que mejores resultados presente en el estudio comparativo realizando un estudio de expresión diferencial de ARNncl en condiciones de estrés por frío.

3. MATERIALES Y MÉTODOS

3.1. ESTUDIO COMPARATIVO DE HERRAMIENTAS DE PREDICCIÓN

En primer lugar, se generan los ficheros que contienen el conjunto de datos prueba, los cuales se introducen como archivos de entrada de los cinco programas seleccionados. Con los archivos de salida se procede al análisis de resultados, generando una matriz de confusión que permite realizar el cálculo de siete métricas de evaluación. Dicho proceso se muestra de forma esquemática en la Figura 3.

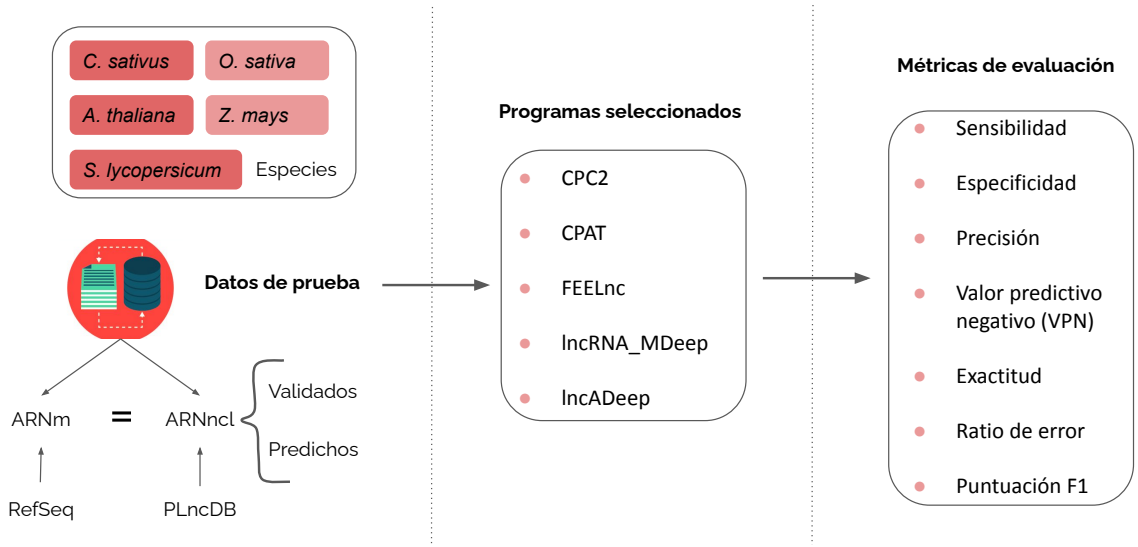


Figura 3. Flujo de trabajo para la realización del estudio comparativo de herramientas de predicción. Dentro de las especies seleccionadas, aquellas en color más oscuro se trata de dicotiledóneas, mientras que las de color más claro son monocotiledóneas.

3.1.1. PROGRAMAS

Se han seleccionado 5 programas para realizar el estudio comparativo de la capacidad de predicción de ARNncl. Algunos criterios seguidos a la hora de seleccionar los programas fueron: contar con un número razonable de citas, estar testados en plantas (o sin especificidad hacia ninguna especie), no necesitar el alineamiento con un genoma de referencia y tener la posibilidad de instalarse y funcionar en la línea de comandos de Linux. Se seleccionaron tanto herramientas que emplean algoritmos de aprendizaje automático como de aprendizaje profundo. En la Tabla 1 se muestran en resumen los programas seleccionados y las características que tiene en cuenta cada uno a la hora de predecir, además de donde se pueden encontrar.

Tabla 1. Listado de programas seleccionados. Se expone el año de publicación, el algoritmo y las características que usa cada programa para la identificación y predicción de ARNncl, así como donde se pueden encontrar.

Software	Año	Algoritmo	Características	URL
CPC2	2017	Máquina de soporte de vectores (MSV)	Código de prueba de Ficket, longitud del ORF, integridad del ORF y predicción del punto isoeléctrico	http://cpc2.gao-lab.org/download.php
CPAT (v.3.0.0)	2013	Regresión logística (RL)	Longitud del ORF, cobertura del ORF (longitud ORF/longitud transcrito), código prueba de Ficket y frecuencia de hexámeros.	https://sourceforge.net/projects/rna-cpat/files/?source=navbar
FEELnc	2017	Bosque aleatorio (BA)	Frecuencia de K-meros, longitud e integridad del ORF	https://github.com/tderrien/ FEELnc.
lncADeep	2018	Red de creencia profunda (RCP)	Cobertura y longitud del ORF, el perfil de densidad de la entropía del ORF, la frecuencia de hexámeros, longitud de la región codificante, código prueba de Ficket, el contenido de CG y la longitud de la región no codificante de un ARNm	http://cqb.pku.edu.cn/ZhuLab/lncadeep/
ARNncl_Mdeep	2020	Dos modelos de redes neuronales profundas (RNP) y un modelo de red neuronal de convolución (RNC)	Cobertura y longitud del ORF, código prueba de Ficket, frecuencia de hexámeros y frecuencia de K-meros	https://github.com/NWPU903PR/lncRNA_Mdeep

3.1.2. DATOS DE PRUEBA

Con la finalidad de que el estudio comparativo de los programas sea más fiable se probaron 5 especies diferentes con cada programa (*Arabidopsis thaliana, Cucumis sativus, Oryza sativa, Solanum lycopersicum y Zea mays*). Para ello se genera un fichero de datos de prueba para cada una de las especies que debe contener ARNm y ARNncl. Las secuencias de ARNm se obtienen de RefSeq (Pruitt et al., 2007), donde se puede encontrar secuencias completas y bien anotadas, mientras que las de ARNncl se obtienen de PLncDB V2.0 (Jin et al., 2021), una base de datos de ARNncl específica de plantas.

En PLncDB la mayoría ARNncl provienen del análisis computacional de muestras de RNA-Seq, por tanto, se trata de ARNncl no validados experimentalmente. Lo más adecuado sería realizar el estudio comparativo con transcritos validados, pero este es un problema intrínseco de los programas de predicción de ARNncl en plantas ya que el número de ARNncl validados en dicho reino es muy escaso. De hecho, los conjuntos de datos con los que se entrena a los programas están formados mayoritariamente por ARNncl putativos. Sin embargo, todas las especies seleccionadas cuenta con algunos ARNncl validados, que se incluyen también en los ficheros de datos de prueba y de esta forma se tiene una idea del sesgo que ocasiona el uso de ARNncl putativos para el entrenamiento de los programas.

En resumen, los ficheros de prueba generados contienen dos clases de transcritos, los ARNm y los ARNncl (validados y putativos). La proporción entre ambas clases debe ser balanceada, es decir, debe haber la misma cantidad de transcritos codificantes y no codificantes. Debido a que el factor limitante son los ARNncl, se cogerán tantos ARNm como ARNncl tenga cada especie (Tabla 2).

Tabla 2. Número de transcritos que se obtienen de cada clase para cada especie. Las tres primeras columnas corresponden al número de ARNm (incluyendo isoformas), ARNncl putativos y de ARNncl validados y la última columna corresponde al número de transcritos que contienen los ficheros que serán usados para probar los programas, todos con la misma cantidad de ARNm que de ARNncl.

Especie	ARNm	ARNncl (putativo)	ARNncl (validado)	Fichero final
A. thaliana	48151	13455	144	27198
O. sativa	46343	11525	40	23130
Z. mays	64045	32371	26	64794
S. lycopersicum	40265	8722	19	17482
C. sativus	33003	8745	8	17506

Para facilitar el análisis de los resultados, la cabecera de cada uno de los transcritos constará del nombre de la especie a la que pertenece, el tipo de transcrito (ARNm, ARNncl o ARNncl-validado) y un numero distintivo. Un ejemplo de un transcrito codificante de *O. sativa* sería: 'O.sativa_ARNm_1'. La finalidad de estandarizar la cabecera de los transcritos es poder trabajar con bucles, evitando tener que repetir el mismo proceso para cada especie.

3.1.3. ANÁLISIS DE RESULTADOS

Cada programa proporciona un fichero de salida diferente (Tabla 3), por ello es necesario ordenar y sintetizar estos ficheros para quedarse solo con aquella información que sea necesaria.

Tabla 3. Comandos usados y columnas del fichero de salida de cada programa. En la siguiente tabla aparecen las diferentes columnas que contiene el fichero de salida de cada uno de los programas separadas por ';'. En negrita se encuentran las dos que se deben seleccionar para analizar los resultados.

Programa	Comando	Columnas
CPAT	`cpat.py -x <tabla con frecuencia de hexámeros> -d <modelo de regresión logística> --antisense -g <fichero de entrada> -o <fichero de salida>`	**ID de la secuencia**; ID; ARNm; Cadena del ORF; Pauta del ORF; Comienzo del ORF; Final del ORF; ORF; Fickett; Hexamero ; **Probabilidad codificante**
CPC2	`CPC2.py -i <fichero de entrada> -o <fichero de salida> --ORF`	**ID**; Longitude del transcrito; Longitude del péptido; Puntuación Fickett; pI; Integridad del ORF; Comienzo del ORF ; Probabilidad codificante; **Etiqueta**
FEELnc	`FEELnc_codpot.pl --outdir="directorio de salida" -o <fichero de salida> -i <fichero de entrada> -a <fichero FASTA con secuencias de ARNm de la especie> --mode=shuffle --spethres=0.99,0.99`	**ID**; Puntuaciónkmer_1mer; Puntuaciónkmer_2mer; Puntuaciónkmer_3mer; Puntuaciónkmer_6mer; Puntuaciónkmer_9mer; Puntuaciónkmer_12mer; Cobertura del ORF; Tamaño ARN; potencial codificante; **Etiqueta**
LncADeep	`python LncADeep.py -MODE lncRNA -f <fichero de entrada> -o <fichero de salida>`	**ID**; Número de clasificadores que consideran el transcrito como codificante (a partir de 10 se considera codificante); **Etiqueta**; 21 columnas que corresponden

		a 21 clasificadores que miden el potencial codificante.
lncRNA_Mdeep	`python3 lncRNA_Mdeep.py -i <fichero de entrada> -o <fichero de salida>`	**ID**; **Etiqueta**; Potencial codificante

Se obtiene un fichero de dos columnas; una con la cabecera de cada transcrito y la otra con la predicción del programa. En cuanto a la columna correspondiente a la predicción, algunos programas como Feelnc, CPC2, LncADeep y lncRNA_Mdeep cuentan con una columna que indica directamente si el transcrito es codificante o no codificante, columna a la que se denomina 'etiqueta'. Mientras, otros como CPAT cuentan con una columna que debe procesarse para saber si es codificante o no. Se consideran como no codificantes en el fichero de salida de CPAT, aquellos transcritos que presentan una puntuación de más de 0.364 en la columna de 'Probabilidad codificante', ya que según se comprueba en L. Wang et al., 2013 mediante una curva ROC, este es el umbral donde se consigue la mayor especificidad y sensibilidad.

El fichero con dos columnas (ID y etiqueta) es el que se usa para generar la matriz de confusión (Tabla 4), herramienta que nos permite visualizar de forma practica el rendimiento de un programa de predicción. La matriz de confusión se construye a partir de filas, que corresponde a la clasificación original, y columnas, que corresponde a las predicciones realizadas por el programa.

En este caso se trabaja con una clasificación binaria, considerando positivo a los ARNncl y negativo a las ARNm, además, se deben tener en cuenta los ARNncl validados, por lo que se añade una fila más. Con esta tercera fila es posible calcular ciertas medidas de rendimiento con los ARNncl validados y compararlas con las medidas de los putativos (Tabla 4).

Tabla 4. Matriz de confusión

		Predicción	
		Positivos predichos	Negativos predichos
Clasificación original	Positivos	VP	FN
	Negativos	FP	VN
	Positivos validados	VPv	FNv

Al generar la matriz se obtienen cuatro opciones:

- Verdaderos positivos: ARNncl que el programa predice como no codificante
- Verdaderos negativos: ARNm que el programa predice como codificante
- Falsos positivos: ARNm que el programa predice como no codificante
- Falsos negativos: ARNncl que el programa predice como codificante

Más las dos opciones extra de los ARNncl validados:

- Verdaderos positivos validados: ARNncl validado que el programa predice como no codificante
- Falsos negativos validados: ARNncl validado que el programa predice como codificante

3.1.4. MÉTRICAS DE EVALUACIÓN

A partir de la matriz de confusión de cada uno de los programas en las cinco especies se calculan siete indicadores que permiten evaluar la herramienta de predicción que mejores resultados presente. Los indicadores calculados se encuentran en la Tabla 5.

Tabla 5. Métricas de evaluación y sus ecuaciones. Adaptado de (Hossin & Sulaiman, 2020).

Indicador	Ecuación	Descripción
Sensibilidad	$\frac{\mathrm{VP}}{\mathrm{VP}+\mathrm{FN}}$	Indicador del potencial de una herramienta de predicción para clasificar de forma correcta los casos positivos.
Especificidad	$\frac{\mathrm{VN}}{\mathrm{VN}+\mathrm{FP}}$	Indicador del potencial de una herramienta de predicción para clasificar de forma correcta los casos negativos
Valor predictivo negativo (VPN)	$\frac{\mathrm{VN}}{\mathrm{VN}+\mathrm{FN}}$	Medida de dispersión, porcentaje de verdaderos negativos dentro de los negativos predichos.
Precisión	$\frac{\mathrm{VP}}{\mathrm{VP}+\mathrm{FP}}$	Medida de dispersión, porcentaje de verdaderos positivos dentro de los positivos predichos.
Exactitud	$\frac{\mathrm{VP}+\mathrm{VN}}{\mathrm{total}}$	Proporción de aciertos, tanto positivos como negativos.
Ratio de error	$\frac{\mathrm{FP}+\mathrm{FN}}{\mathrm{total}}$	Proporción de errores, tanto positivos como negativos
Puntuación F1	$\frac{2*\mathrm{VP}}{2*\mathrm{VP}+\mathrm{FP}+\mathrm{FN}}$	Resumen de la precisión y sensibilidad

3.2. ANÁLISIS BIOINFORMÁTICO PARA EL ESTUDIO DE LA EXPRESIÓN DIFERENCIAL DE GENES Y ARNncl

3.2.1. DATOS DE RNA-SEQ

El material vegetal usado en este trabajo se cultivó y trató tal y como se describe en Sanz-Carbonell et al. 2019, pero en este caso se llevaron a cabo solo dos estreses; por frío y por infección HSVd y se realizaron dos controles, uno con las plantas antes de la inoculación y otro de plantas sin tratamiento. Se realizaron tres replicas para cada una de las condiciones y cada muestra se generó a partir de tres plantas.

La extracción de ADN y ARN se realizó como se describe en Márquez-Molins et al., 2022 y todas las librerías fueron construidas por Novogene Europe (Cambridge, Reino Unido) según sus procedimientos estándar.

Finalmente se obtuvieron doce librerías de RNA-Seq *paired-end* direccionales de 150 pb de *Cucumis melo*. Todas ellas son usadas para la generación de un nuevo transcriptoma que incluya ARNncl. Para el análisis de expresión diferencial, sin embargo, solo se usarán las tres replicas correspondientes al estrés por frío y al no tratamiento.

Las librerías se encuentran en formato FASTQ, en el cual se almacenan tanto las secuencias nucleótidos como sus métricas de calidad. Antes de utilizar estos datos brutos es necesario un control de calidad de las secuencias, el cual se realizó con FastQC (v 0.9.11) (Andrews, S., 2010).

3.2.2. GENERACIÓN DE UN TRANSCRIPTOMA QUE CONTENGA ARNncl

Es necesaria la generación de un nuevo transcriptoma previo al análisis de expresión diferencial, dado que el fichero de anotación de referencia no contiene ARNncl y por tanto no podría realizarse la cuantificación de estos.

Para facilitar el seguimiento del proceso de generación de un transcriptoma que contenga ARNncl se muestra, de forma esquemática en la Figura 4, la metodología seguida, los ficheros necesarios y los programas usados.

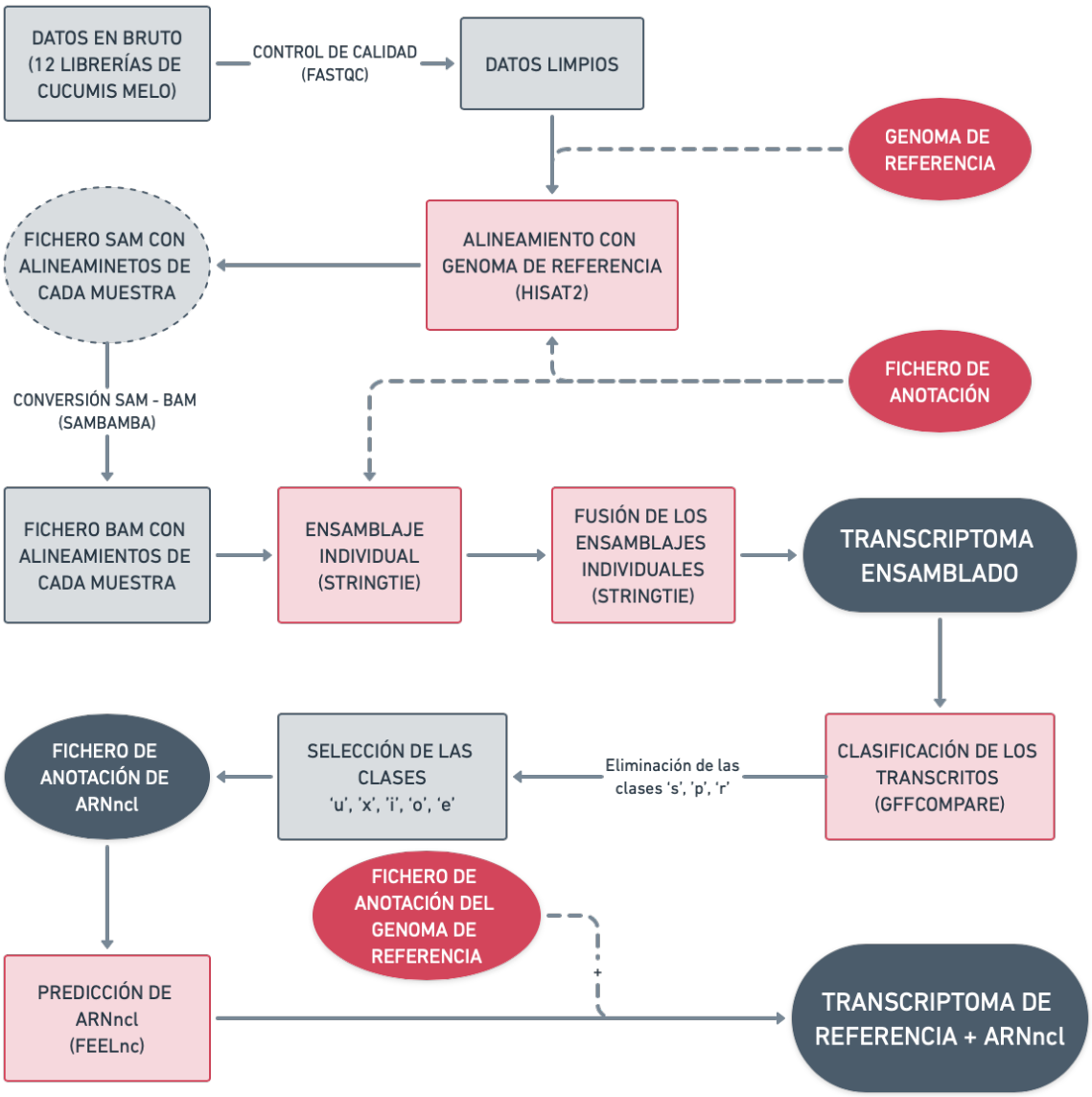


Figura 4. Flujo de trabajo para la generación de un transcriptoma que contenga ARNncl. Se parte de doce librerías de RNA-Seq que se alinean contra el genoma de referencia, se ensamblan de forma individual y después se fusionan dando como resultado un transcriptoma ensamblado. A partir de aquí se seleccionan los potenciales ARNncl a partir de los códigos de clases y se introducen en un programa de predicción (FEELnc). Aquellos transcritos que pasen el filtro de la predicción son considerados como ARNncl y se añaden al transcriptoma de referencia ya existente.

3.2.2.1.ALINEAMIENTO DE LAS MUESTRAS CON EL GENOMA DE REFERENCIA

El genoma de referencia usado y su fichero de anotación corresponden a la versión v 4.0 obtenida de Melonomics. El fichero de anotación es un fichero GFF3 y se convierte a GTF con gffread (v 0.12.7) (Pertea & Pertea, 2020):

```
gffread <fichero de anotación de referencia GFF3> -T -o <fichero de anotación de referencia GTF>
```

El alineamiento de las muestras contra el genoma de referencia se lleva a cabo con hisat2 (v 2.1.1) (Kim et al. 2015), ya que las muestras provienen de un RNA-Seq. Al tratarse de transcritos maduros el programa usado para alinear debe tener en cuenta la pérdida de los intrones y los saltos que esto supone en el genoma. Para facilitar la tarea de alineamiento es necesario, por un lado, realizar un indexado del genoma de referencia con hisat2-build y por otro, guardar los sitios de empalme con el fichero hisat2_extract_splice_sites.py (Tabla 6).

Tabla 6. Comandos usados en hisat2

Indexado	`hisat2-build <FASTA del genoma de referencia> <directorio de salida>`
Sitios de empalme	`Python hisat2/hisat2_extract_splice_sites.py <fichero de anotación de referencia> > <fichero de salida>`
Hisat2	`hisat2 -x <directorio del índice del genoma de referencia> -1 <lista de ficheros que contienen "_1"> -2 <lista de ficheros que contienen "_2"> -S <nombre de las muestras> --max-intronlen 10000 --dta -p 20 --rna-strandness RF --known-splicesite-infile <fichero con los sitios de empalme> --novel-splicesite-outfile <fichero de salida de los sitios de empalme> --summary-file <fichero de salida>`

Los resultados del alineamiento se obtienen en formato SAM que serán convertidos a BAM y ordenados con el programa sambamba (v 0.8.0) (Tarasov et al., 2015). Los comandos usados son:

```
sambamba view -t 20 -f bam -F "not unmapped" -S -o <fichero SAM>
sambamba sort -t 20 --tmpdir="directorio de salida" -o <fichero de salida>
```

3.2.2.2.ENSAMBLAJE GUIADO POR FICHERO DE ANOTACIÓN:

Se realiza el ensamblaje de cada una de las muestras con el programa Stringtie (v 2.2.0) (Pertea et al., 2015) utilizando los siguientes comandos:

```
stringtie -m 100 -p 10 -G <fichero de anotación> -o <fichero de salida GTF> --rf -A <fichero de salida con las abundancias de los genes> <fichero de entrada BAM>
```

Una vez se tiene el ensamblaje de cada muestra es necesario combinar todas las muestras en un solo fichero, de nuevo se usa la herramienta Stringtie:

```
stringtie --merge -o <fichero de salida (transcriptoma ensamblado)> -g 250 -F 0.5 -v -G <fichero de anotación de referencia> <archivo con las rutas del GTF generado de cada muestra>
```

Se añade el argumento -g 250 ya que de este modo aquellas lecturas que se encuentren a menos de 250 nucleótidos se fusionan en un mismo paquete de procesamiento, maximizando el número de potenciales ARNncls con 2 exones. Se debe tener en cuenta que al fusionar las lecturas que se encuentren a 250 nucleótidos se pueden perder algunos genes con ubicaciones cercanas, en los que uno de ellos tenga una función y el otro se trate de una proteína desconocida.

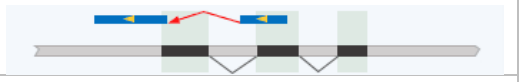
3.2.2.3.FILTRADO Y CLASIFICACIÓN

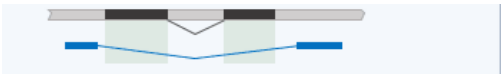
Con ayuda del programa gffcompare (v 0.12.6) (Pertea & Pertea, 2020)se compara el fichero de anotación del genoma de referencia con el transcriptoma ensamblado:

```
gffcompare -V <GTF del transcriptoma ensamblado> -o <fichero de salida> -r <fichero de anotación el genoma de referencia>
```

Se usa el argumento -r para crear una columna llamada código de clase, que clasifica a los transcritos según su localización, tal como aparece en la Tabla 7.

Tabla 7. Código de clases usado por gffcompare. En negro se muestran los exones y transcritos de referencia y en azul los transcritos que deben clasificarse. Adaptado de Pertea & Pertea, 2020

=	Coincide exacta de la cadena de intrones. 	o	Otros solapamientos de la misma cadena con los exones de referencia. 
c	Contenidos en la referencia (compatible con el intrón). 	s	Coincidencia de intrones en la cadena opuesta (como un error de mapeo). 
k	Contención de la referencia (contención inversa). 	x	Solapamiento exónico en la cadena opuesta (como ‘o’ o ‘e’ pero en la cadena opuesta). 
m	Intrones retenidos, todos los intrones coinciden o se mantienen. 	i	Totalmente contenido en un intrón de referencia. 

n	Intrones retenidos, no todos los intrones coinciden. 	y	Contiene una referencia dentro de sus intrones. 
j	Multi-exónico con al menos una coincidencia de unión. 	p	Posible escurrimiento de la polimerasa. 
e	Transfragmentación de un solo exón parcialmente cubriendo un intrón, posible fragmento de pre-ARNm. 	r	Repeticiones.
		u	No se ubica dentro de nada conocido

A partir del código de clases se realiza, por un lado, una selección negativa con un programa de Python que puede consultarse en el Anexo 2, Código 1, dicho programa se encarga de eliminar las clases que puedan ser debidas a errores (r, s y p).

Por otro lado, de selecciona positivamente las clases que sean potenciales ARNncl (u, x, i, o, e) con otro programa de Python que puede encontrarse en el Anexo 2, Código 2. Además, cada uno de los códigos de clases se asocia a un tipo de ARNncl; "u": ARN no codificante intergénico largo (ARNncil), "x": Transcritos anti-sentido naturales (TAN), "i": ARN no codificante intrónico (ARNnci) y "o" y "e": ARN no codificante exónico (ARNnce).

Con ayuda de otro programa de Python se genera una tabla en la que aparece el número de exones y la longitud de cada transcrito, permitiendo seleccionar los transcritos que tienen más de un exón y más de 200 nucleótidos. Gracias a estos filtros se consigue un fichero de anotación que solo contiene potenciales ARNncl.

3.2.2.4.PREDICCIÓN ARNNCLS CON FEELNC

Para conseguir un nivel más de seguridad de los potenciales ARNncl se usa una herramienta de predicción, FEELnc (Wucher et al., 2017). El programa FEELnc necesita un fichero FASTA como entrada, el cual se genera a partir del genoma de referencia y el fichero de anotación del transcriptoma ensamblado con la herramienta gffread (v 0.12.7):

```
gffread -w <fichero de salida FASTA> -W -F -g <genoma de referencia> <fichero GTF del transcriptoma ensamblado>
```

Una vez se tiene el fichero FASTA con los posibles ARNncl se lleva a cabo la predicción:

```
FEELnc_codpot.pl --outdir="directorio de salida" -o <fichero de salida> -i <fichero de entrada> -a <fichero FASTA con secuencias de ARNm de la especie> --mode=shuffle --spethres=0.99,0.99
```

El fichero FASTA con secuencias de ARNm de *Cucumis melo* aquí especificado se genera también con gffread:

```
gffread -w <fichero de salida FASTA> -W -F -g <genoma de referencia> <fichero de anotación de referencia>
```

El fichero de salida de FEELnc es a un fichero de texto con once columnas (Tabla 3) de las cuales solo nos interesa la primera (ID) y la última (etiqueta). Con ayuda de dichas columnas se genera un fichero con los IDs de los transcritos que han sido predichos como ARNncl. A partir de los identificadores se obtiene un fichero FASTA usando seqtk (v 1.3) (Lh3/Seqtk, 2018);

```
seqtk subseq <fichero FASTA de entrada de FEELnc> <fichero con las identificaciones de los ARNncl predichos> > <fichero de salida>
```

y un fichero GTF usando las siguientes funciones de bash;

```
> <fichero de salida>

while read id; do

        grep "transcript_id \""$id"\"" <fichero GTF con los potenciales ARNncl antes de la predicción >> <fichero de salida>

done < <fichero con las identificaciones de ARNncl predichos>
```

Por último, el fichero de anotación generado con los ARNncl predichos se combina con el fichero de anotación del genoma de referencia, obteniendo un nuevo transcriptoma que sí contiene ARNncl y por tanto permite su cuantificación.

3.2.3. CUANTIFICACIÓN Y ANÁLISIS DE EXPRESIÓN DIFERENCIAL

Para llevar a cabo el estudio de expresión diferencial es necesaria la cuantificación de los transcritos, para ello se usa el programa Salmon (v 1.6.0) (Patro et al., 2017) (Tabla 8)

Tabla 8. Comandos para la cuantificación de transcritos con Salmon. Para cuantificar con Salmon es necesario el indexado del transcriptoma

Indexado	`salmon index -t <fichero FASTA con genes y ARNncl> -I <directorio de salida>`

Salmon	`salmon quant -i <transcritos indexados> -l A -1 <fichero lecturas directas> " -2 <fichero lecturas reversas> -g <genoma de referencia> -p 11 --validateMapping -o $sample --gcBias`

Los archivos de salida con los conteos se importan a R (v 4.1.0) con el paquete tximport (1.24.0) (Soneson et al., 2016), tal y como se recomienda en (Love et al., 2014).

Una vez importados los recuentos se usa el paquete DESeq2 (v 1.34.0) (Love et al., 2014) el cual corrige el tamaño de la librería, por lo que los valores transformados o normalizados no deben usarse como entrada. Además, este paquete es el que proporciona la información de qué transcritos están diferencialmente expresados mediante una tabla que contiene el número de lecturas normalizado, el *log2(foldchange)*, el p-valor y el p-valor ajustado (Tabla S3). Para considerar como significativa la expresión diferencial de los transcritos se necesita un p-valor ajustado < 0,05.

3.2.4. BÚSQUEDA DE ELEMENTOS REGULADORES EN *CIS*

A partir del fichero de salida de DEseq2 se genera una lista solo con los transcritos (ARNncl y ARNm) expresados diferencialmente para realizar una búsqueda de potenciales ARNncl reguladores en *cis* que puedan intervenir en la respuesta a estrés.

Se usa la herramienta bedtools (v 2.30.0) (Quinlan & Hall, 2010) para la búsqueda de ARNncl diferencialmente expresados que se encuentren a una distancia de 10kb de algún ARNm también diferencialmente expresado. Para usar dicha herramienta es necesario generar un fichero GTF de ARNncl y otro de ARNm que contenga solo los transcritos diferencialmente expresados, utilizando el transcriptoma que contiene los ARNncl y la lista de ARNncl y ARNm expresados diferencialmente. Dichos ficheros se convierten a formato BED con convert2bed (v 2.4.40) (Neph et al., 2012):

```
convert2bed -i gtf < <fichero de entrada> --attribute-key=transcript_id | awk '$8 == "transcript" {print $0}' > <fichero de salida>
```

A continuación, con los siguientes comandos se creó un archivo BED de ARNncl con rangos de 10kb aguas abajo/aguas arriba:

```
bedtools slop -i <fichero de entrada> -g <fichero con el tamaño de los cromosomas> -b 10000 > <fichero de salida>
```

Se intersectó el archivo BED de ARNncl con rangos y el archivo BED de Genes:

```
bedtools intersect -a <fichero de entrada> -b <fichero BED con los genes diferencialmente expresados> -s -wao -nonamecheck > <fichero de salida>
```

El fichero con el tamaño de los cromosomas se genera con SAMtools (v 0.1.19) (Danecek et al., 2021): `samtools faidx <genoma de referencia>`

Mediante el uso de términos GO obtenidos de la base de datos *Cucurbitgenomics* y usando la función *merge* de R se genera una tabla con los genes diferencialmente expresados, que cuentan con ARNncl también diferencialmente expresados a menos de 10kb aguas arriba/aguas abajo, y su correspondiente término GO.

En la Figura 5 se presenta de forma resumida la metodología seguida para realizar el estudio de expresión diferencial.

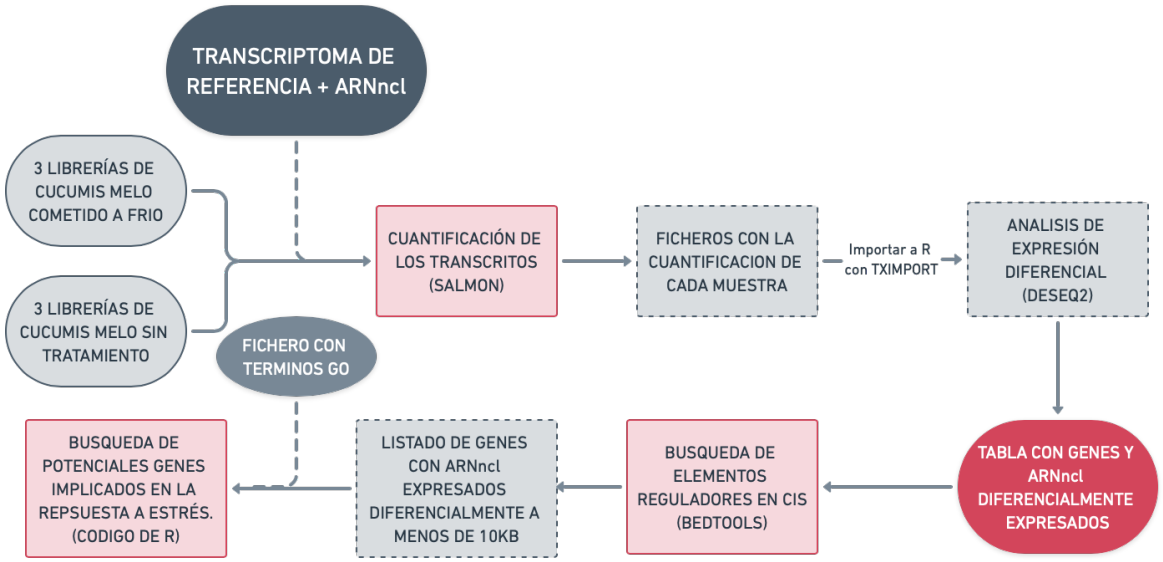


Figura 5. Flujo de trabajo seguido para el análisis de expresión diferencial y la búsqueda de ARNncl reguladores en *cis*. A partir de seis librerías de un ARNseq y con ayuda del nuevo transcriptoma que contiene ARNncl, se realiza la cuantificación tanto de genes como de ARNncl y su posterior análisis de expresión diferencial. Por último, se realiza la búsqueda de ARNncl reguladores de genes que intervienen en la respuesta a estrés en *cis.*

3.3. OBTENCIÓN DE LAS FIGURAS

Para el tratamiento de los datos, tanto en el análisis de expresión diferencial como en el estudio comparativo de herramientas de predicción se utilizó R (v 4.1.0). Las representaciones se obtuvieron usando las librerías de R ggplot2 (v 3.3.6) y ggpubr (v 0.4.0).

3.4. COMPUTADOR UTILIZADO

Los programas usados en el estudio comparativo de herramientas de predicción, al igual que aquellos usados en el análisis de expresión diferencial, han sido ejecutados en el Centro de Procesamiento de Datos (CPD) del Instituto de Biología Integrativa de Sistemas (I2SysBio). El computador contiene 9 nodos de cálculo con 40 *cores* cada uno, 256 TB de disco y 1,5 TB de memoria. Para la gestión y planificación de tareas el *cluster* hace uso de SLURM (*Simple Linux Utility for Resources Management*)(Jette et al., 2002).

4. RESULTADOS Y DISCUSIÓN

4.1. ESTUDIO COMPARATIVO DE HERRAMIENTAS DE PREDICCIÓN

Los resultados obtenidos se presentan de forma resumida en la Tabla 9 y de forma más extendida en la Tabla S2 del Anexo 1. Además, en la Tabla S1 pueden consultarse los datos utilizados para el cálculo de las métricas.

Tabla 9. Medias de las métricas de evaluación calculadas para las cinco especies con que se ha probado cada programa. En negrita se encuentran señalados los mejores resultados para cada una de las métricas.

Programa	Exactitud	Ratio error	Sensibilidad	Sensibilidad Validados	Especificidad	Precisión	VPN	Puntuación F1
FEELnc	**97,38**	**2,62**	91,50	89,69	**100,00**	**99,99**	96,27	95,39
CPAT	86,61	13,39	75,45	78,15	97,79	97,14	80,29	84,69
CPC2	96,12	3,88	**98,12**	**96,26**	94,11	94,35	**98,06**	**96,19**
LncADeep	92,24	7,76	88,36	94,83	96,11	95,77	89,42	91,84
lncRNA_Mdeep	91,82	8,18	86,88	86,39	96,75	96,39	88,25	91,31

A simple vista, teniendo en cuenta la Tabla 9, se observa que FEELnc y CPC2 son las herramientas que mejores resultados presentan para todas las métricas. FEELnc presenta una elevada precisión y una especificidad de prácticamente el 100%. Además, tiene el mejor resultado de exactitud y ratio de error, por lo que el número de aciertos, tanto positivos como negativos, es el más elevado entre los programas de predicción. Mientras, CPC2 destaca por la sensibilidad, tanto de ARNncl validados como putativos, y el valor predictivo negativo (VPN).

Cuando un programa presenta una mayor sensibilidad indica una mayor capacidad de detectar positivos, del mismo modo ocurre con la especificidad y los resultados negativos. En la Figura 6A se observa que la tendencia general de los programas es tener una mayor especificidad y una menor sensibilidad, a excepción de CPC2 que presenta una mayor sensibilidad. Existe un equilibrio entre ambas métricas que depende del punto de corte elegido para considerar un caso como positivo, tal y como se comenta en Chu (1999). CPAT, a pesar de tener el valor de sensibilidad más bajo, presenta uno de los valores de especificidad más altos, siendo una herramienta que claramente no cuenta con dicho equilibrio entre ambas métricas.

A pesar de que estas dos métricas nos ofrecen información muy valiosa, también es necesario fijarse en otras como la precisión o el valor predictivo negativo. Por un lado, la precisión indica, dentro de los positivos predichos, la proporción de verdaderos positivos. Es decir, la proporción de ARNncl

predichos que son realmente ARNncl. Por otro lado, el VPN indica la proporción de ARNm predichos (negativos) que son realmente ARNm (verdaderos negativos).

Las tendencias de la precisión y el VPN se presentan en la Figura 6B, en la cual se observa que los resultados son muy similares a los de la Figura 6A. Aquellos programas con mayor especificidad también tienen una mayor precisión, pero un menor VPN, al contrario que en los que tienen mejores valores de sensibilidad. Con esto se puede ver como aquellos programas con mayor potencial de detección de positivos, también presentan una mayor proporción de falsos positivos y menor de falsos negativos, mientras que aquellos con mayor potencial de detección de negativos, presentan mayor proporción de falsos negativos pero menor de falsos positivos.

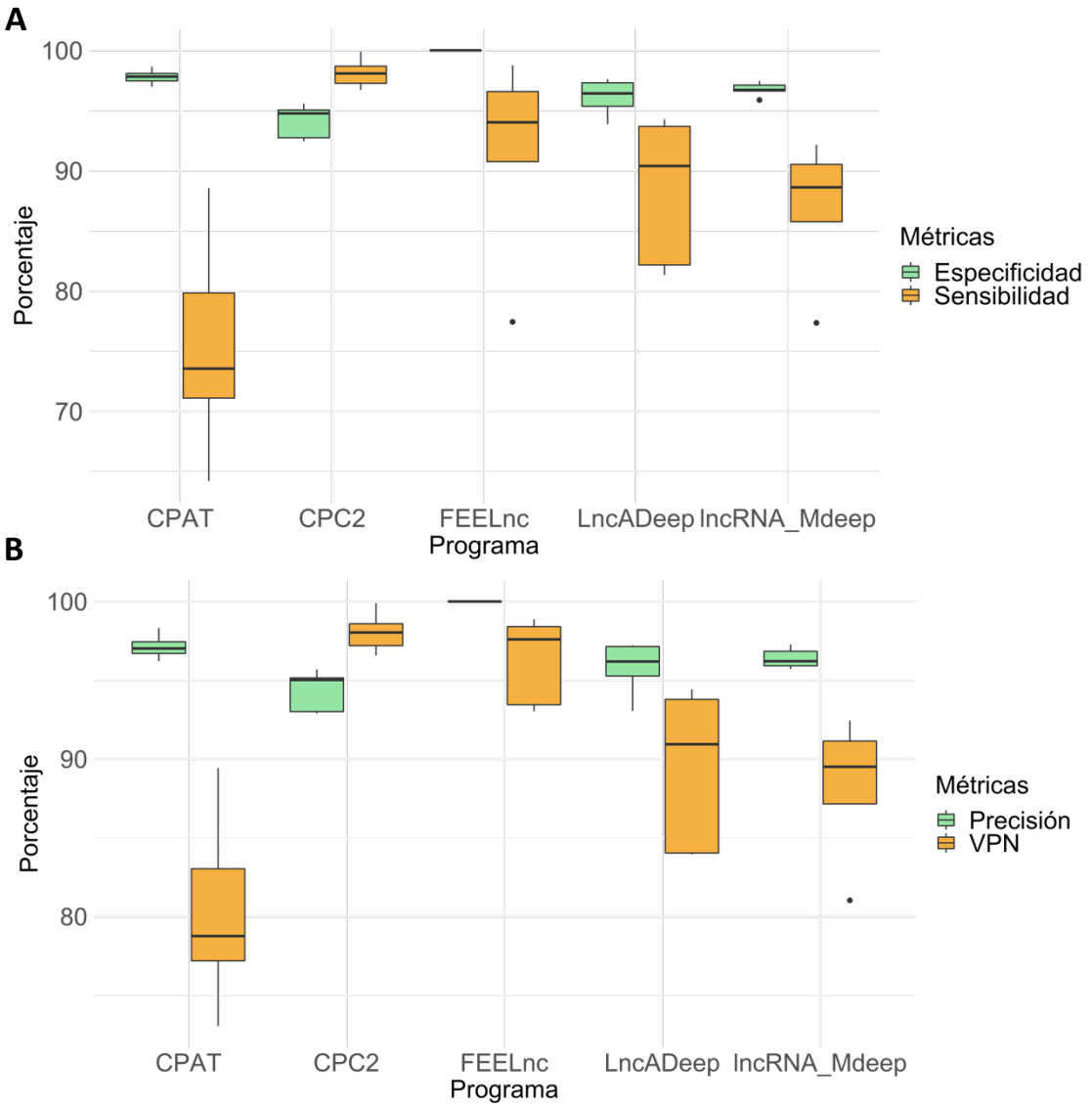


Figura 6. Diagramas de cajas. A) Especificidad (verde) y sensibilidad (naranja) juntas para cada uno de los programas. B) Precisión (verde) y VPN (naranja) para cada uno de los programas.

En cuanto a la exactitud y el ratio de error, medidas que muestran la proporción de aciertos y falsos positivos y negtivos de las herramientas de predicción, tanto CPC2 como FEELnc presentan muy buenos valores respecto al resto de herramientas, aunque FEELnc ligeramente superiores.

Por último, la puntuación F1, la cual fusiona las métricas de sensibilidad y precisión, permite medir el equilibrio entre la cantidad de positivos predichos y la calidad de los mismos. En este caso los valores entre FEELnc y CPC2 también son muy similares, siendo CPC2 un poco mayor. Esta medida resulta muy interesante siempre que se busque un equilibrio entre la sensibilidad y la precisión.

Tanto FEELnc como CPC2 presentan resultados muy buenos y para decidir cuál de los dos usar será necesario considerar la finalidad de su uso. Siempre que se busque una mayor identificación de ARNncl a pesar de no tener tanta fiabilidad, se seleccionará CPC2, mientras si lo que se busca es una mayor calidad de la predicción a pesar de que la cantidad sea menor, se seleccionará FEELnc. Si se busca un equilibrio entre la calidad y cantidad de predichos tanto positivos como negativos ambos programas serán adecuados, aunque posiblemente CPC2 sea un poco superior, ya que como se presenta en la Figura 7 es de las herramientas que mejores resultados tiene usando ARNncl validados.

En el presente trabajo hay un mayor interés en la precisión que en la sensibilidad, es decir, se antepone la fiabilidad de los ARNncl predichos a la cantidad, por lo que el programa seleccionado es FEELnc.

4.1.1. SESGO OCASIONADO POR EL USO DE ARNncl PUTATIVOS PARA ENTRENAR LAS HERRAMIENTAS

Tal y como se comentó en la metodología, la mayoría de los programas específicos de plantas o inespecíficos de especie han sido entrenados por ficheros de datos con ARNncl putativos, no validados. Debido a que los resultados de cada programa dependen de la integridad de los transcritos con los que se entrena y pocos de ellos se encuentran validados, es posible que ciertas características asociadas a los ARNncl no se tengan en cuenta y que por efecto de confusión se clasifiquen como ARNm. Un ejemplo ilustrativo es el comentado en Klapproth et al., 2021; aquellos programas que usen la cola poliA como característica y hayan tenido unos archivos de prueba en los que ningún ARNncl contenga cola poliA clasificarán de forma errónea todos los transcritos con cola poliA como codificantes. El sesgo que ocasiona esta situación no ha sido ampliamente investigado hasta la fecha, pero en la Figura 7 se representa la sensibilidad de cada uno de los programas con ARNncl putativos frente a la sensibilidad de los pocos ARNncl que existen.

La tendencia general que se observa en la Figura 7 es que la sensibilidad de los programas es mayor con datos putativos, tal y como se esperaba. Sin embargo, LncADeep presenta mejores resultados y

una menor dispersión con datos validados y CPAT aunque presenta una dispersión muy elevada en los resultados, parece que podría tener mejores resultados también con datos validados.

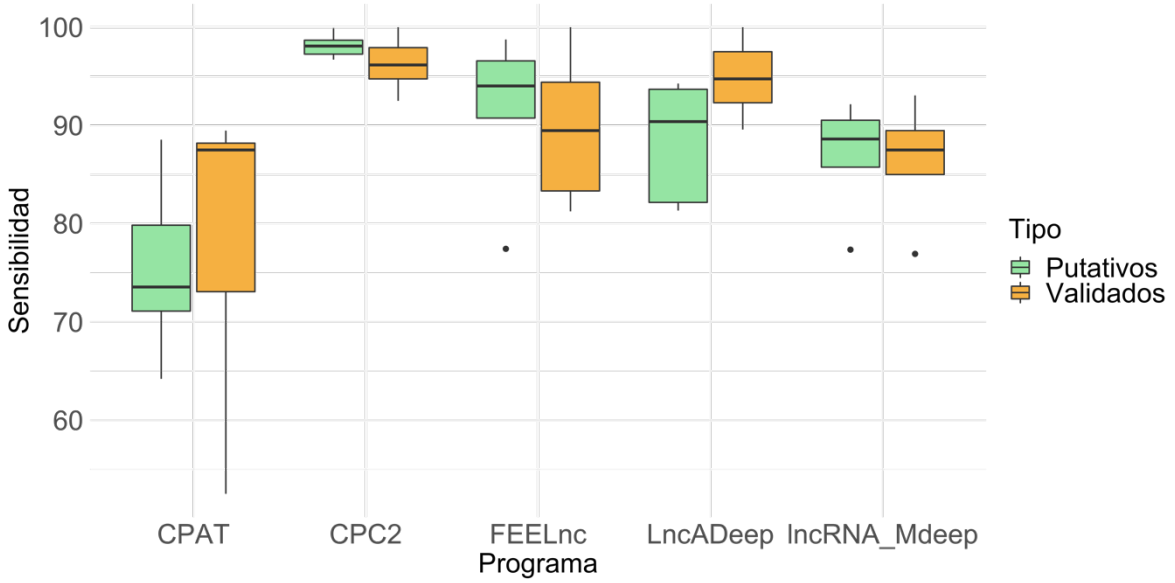


Figura 7. Diagrama de cajas de la sensibilidad de cada uno de los programas, tanto usando ARNncl putativos (verde) como ARNncl validados (naranja).

No se debe olvidar que estos datos simplemente son orientativos ya que la cantidad de ARNncl validados es significativamente menor a la de ARNncl putativos. A pesar de que en humanos y en algunos mamíferos el estudio de ARNncl está muy desarrollado, en plantas aún queda mucho trabajo que hacer. Resulta necesaria la validación experimental de un mayor número de ARNncl en plantas para poder entrenar los programas con datos validados y mejorar tanto la precisión como la sensibilidad en la predicción. De este modo se pueden diseñar experimentos más eficaces y dirigidos, además de ahorrar en material y tiempo.

4.1.2. DIFICULTADES POR LA NO ESTANDARIZACIÓN DE LOS PROGRAMAS

En el presente trabajo se han probado solo cinco herramientas de predicción de ARNncl, sin embargo, con una búsqueda rápida se pueden encontrar más de treinta. Cuando existe una variedad tan amplia de herramientas con una misma finalidad resulta necesaria la publicación de estudios comparativos o de bases de datos que faciliten la búsqueda de la mejor herramienta según el análisis que se pretenda realizar. Es cierto que los ARNncl aún son un área novedosa y por ello no tienen un volumen muy elevado de estudios comparativos o de bases de datos. Algunos ejemplos de estudios comparativos que se han publicado hasta la fecha serían Klapproth et al., 2021; Pinkney et al., 2020 y Signal et al., 2016 , en todos ellos queda reflejada la reducida información que hay disponible para trabajar con ARNncl en plantas respecto a animales. Por otro lado, sería interesante la generación de una base de datos que contenga de forma clasificada y ordenada los métodos de

análisis de ARNncl disponibles, tal y como se puede encontrar para los micro-ARNs en tools4miRs (Lukasik et al., 2016).

Otra iniciativa muy valiosa es la publicación de medidas estandarizadas para la evaluación de herramientas de aprendizaje automático tal y como se hace en Walsh et al., (2021). Pero además de una correcta evaluación, las herramientas necesitan un continuo mantenimiento y cabe destacar que algunas de las herramientas que se pueden encontrar disponibles no están siendo mantenidas, provocando dificultades en su instalación o incluso el no funcionamiento de estas.

4.2. ANÁLISIS DE EXPRESIÓN DIFERENCIAL

El análisis de expresión diferencial de las plantas de melón resultó en 17870 transcritos codificantes y 1256 ARNncl que cambian su expresión como consecuencia del estrés por frío. Dentro de los transcritos codificantes; 6339 estan sobreexpresados y 6223 reprimidos mientras que, de los ARNncl, 213 estan sobreexpresados y 319 reprimidos (Figura 8).

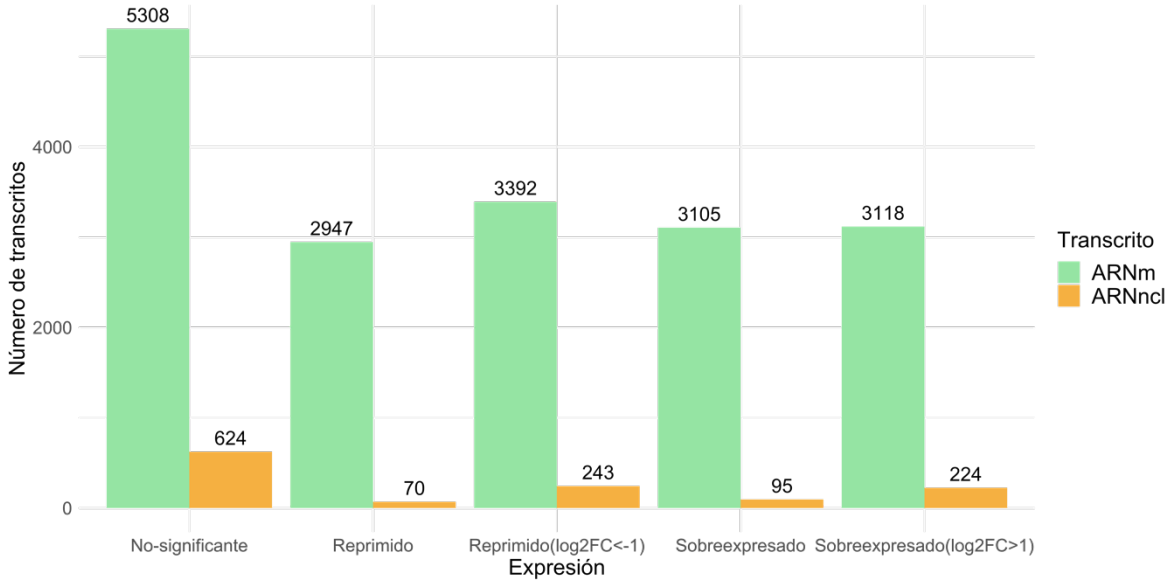


Figura 8. Diagrama de barras con el número de ARNm y ARNncl clasificados según su expresión. Los transcritos reprimidos que cuentan con un log2FC < -1 han reducido su expresión, como mínimo, a la mitad. Del mismo modo, los sobreexpresados que cuentan con un log2FC > 1, han duplicado, como mínimo, su expresión.

Dentro de los ARNncl, los más abundantes son los ARNncil, con 364 transcritos expresados diferencialmente, seguidos de los TAN, con 197 (Figura 9). Ambos tipos corresponden a los ARNncl más estudiados y caracterizados. Todos estos datos se han obtenido a partir del fichero de salida de DEseq2, en la Tabla S3 del Anexo 1 se puede observar un pequeño ejemplo de la información que contiene dicho fichero.

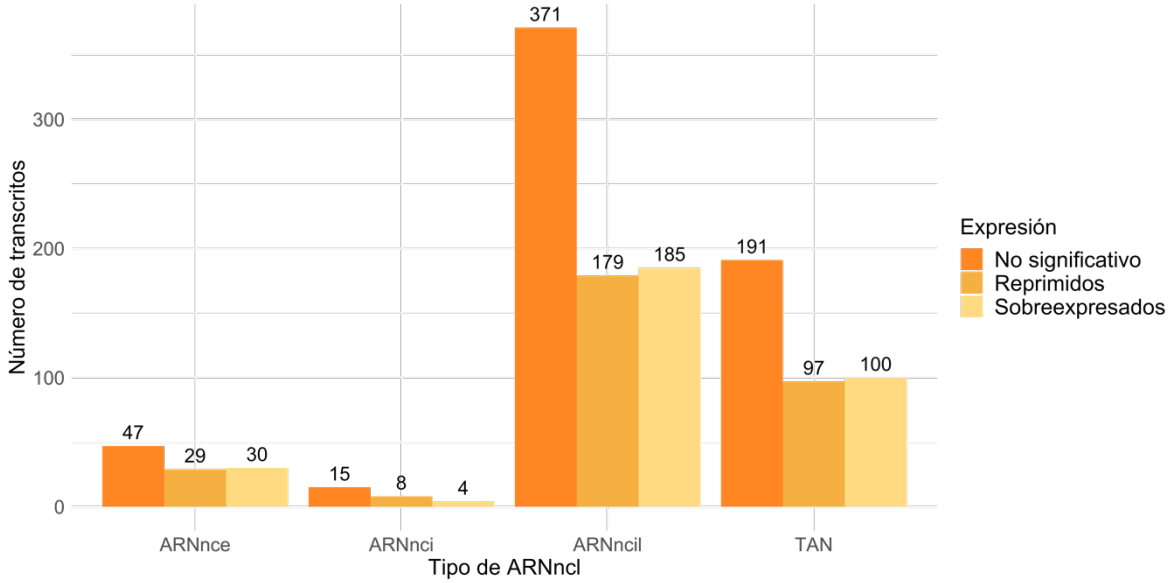


Figura 9. Número de transcritos de cada tipo de ARNncl. Cada tipo de transcrito se encuentra clasificado según si está sobreexpresado, reprimido o si su expresión diferencial no es significativa ARN no codificante intrónico (ARNnci), ARN no codificante intergénico largo (ARNncil), ARN no codificante exónico (ARNnce) y transcritos anti-sentido naturales (TAN).

En la Figura 10 (A y B) se observa la existencia de una gran simetría entre la sobreexpresión y la represión de los transcritos, tanto en ARNncl como en ARNm, lo que sugiere que durante la respuesta a estrés ocurren tanto mecanismos de activación como de represión.

Además, en la Figura 10A también se puede visualizar la media de expresión de cada transcrito; en la zona con una media de expresión baja (zona izquierda) no se encuentra un gran número de transcritos significativos, mientras que en la zona central de la gráfica se encuentran la mayoría de ellos, tanto ARNm como ARNncl. Resulta interesante recordar que la expresión de ARNncl suele ser menor en cantidad a la de ARNm, lo cual concuerda con que aquellos transcritos con una media de expresión más elevada (zona derecha) sean mayormente ARNm.

En cuanto a la Figura 10B, permite visualizar los valores de p-valor de cada transcrito y por tanto la significancia estadística de la expresión diferencial de cada uno de ellos.

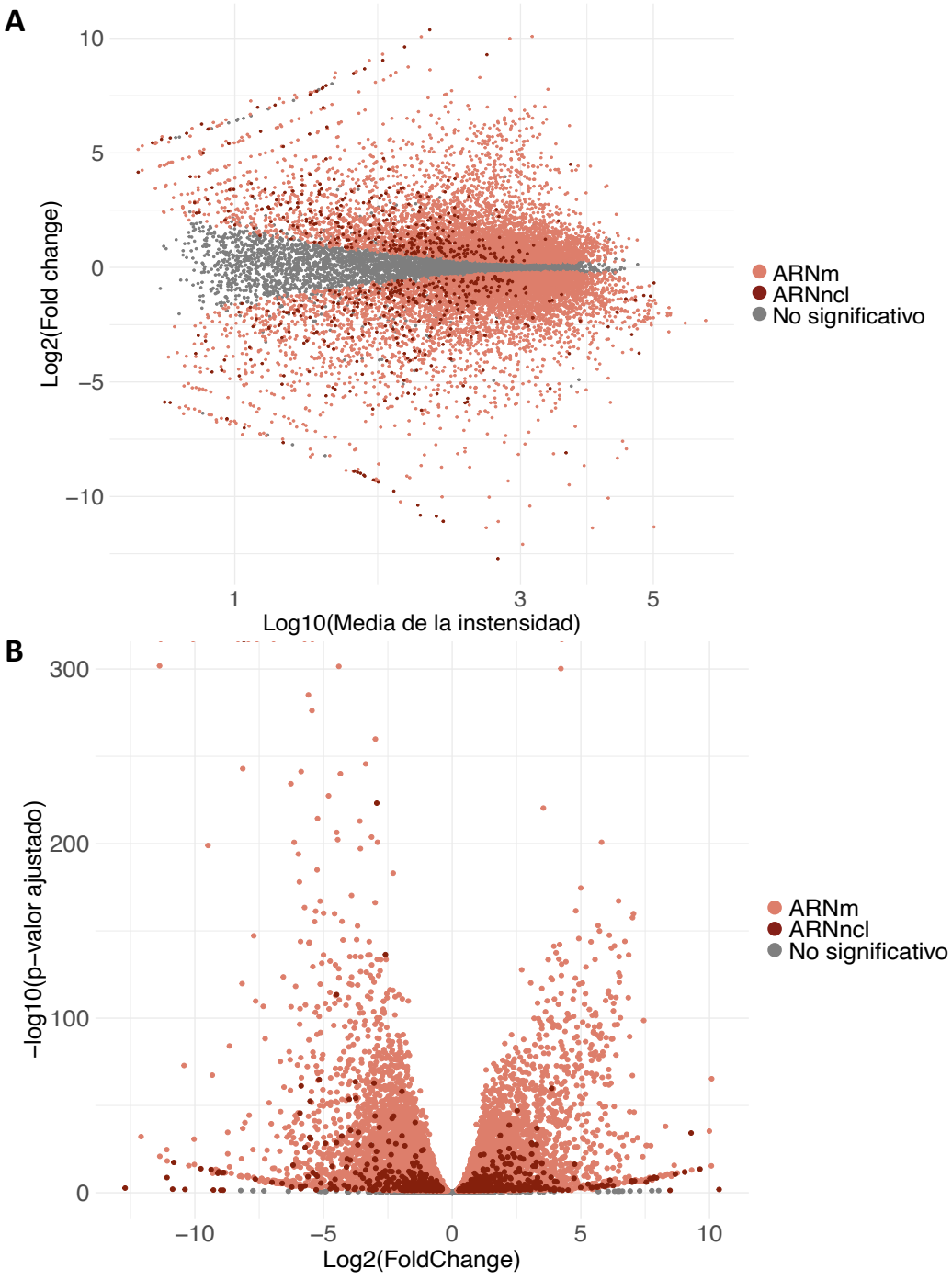


Figura 10. A) ***MA plot:*** Presenta, en base logarítmica, la expresión diferencial entre la condición de estrés y la condición control, frente a la media de expresión de cada uno de los transcritos en las diferentes muestras. **B)** ***Volcano plot****:* Presenta, en base logarítmica, la significancia estadística (p-valor ajustado) de cada transcrito frente a la expresión diferencial entre la condición de estrés y el control. Ambos gráficos contienen ARNm y ARNncl, si estos tienen un p-valor significativo aparecen como puntos de color rosa los ARNm y rojo los ARNncl, mientras que si no es significativo son puntos grises. Los datos utilizados para la obtención de estas gráficas pueden consultarse en la Tabla S3 del Anexo 1.

4.2.1. POTENCIALES ARNncl REGULADORES EN *CIS*

La búsqueda de ARNncl que pudiesen tener un efecto en *cis* ha resultado en 331 parejas de ARNncl y ARNm que se localizan a menos de 10 kb aguas arriba/abajo y que estan expresados diferencialmente. Dichas parejas se muestran en un fichero como el presentado en la Tabla S4 del Anexo 1, al que se le ha añadido los p-valor ajustado de cada miembro de la pareja. Gracias a los valores del log2FC se ha podido observar que las parejas de transcritos muestran una correlación positiva de su expresión diferencial (Figura 11), por lo que cuando un miembro de la pareja aumenta su expresión, el otro también lo hace y del mismo modo ocurre con la represión. Cabe destacar que a pesar de que la relación lineal no es muy intensa (cuenta con una R de Pearson baja), sí es significativa y no debida al azar ya que presenta un valor p igual a 1,2e-09. Dicha correlación también se ha observado en trabajos como Engreitz et al. (2016), donde se ve como los *loci* de ARNncl influyen en la expresión de los genes vecinos.

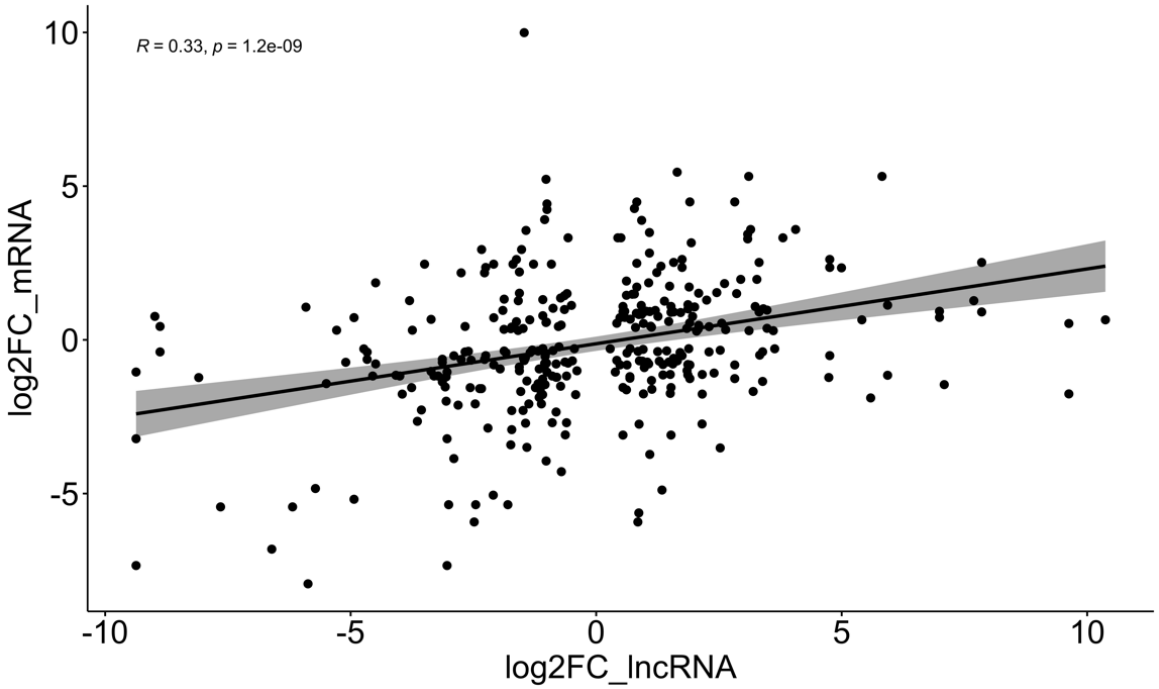


Figura 11. Correlación de Pearson entre los log2FC de las parejas de ARNncl y ARNm que se encuentran en posición *cis.* En la parte superior de la gráfica se observa la significancia estadística de la correlación (p = 1,2e-09) y el grado de correlación (R= 0,33).

De forma paralela, se han encontrado cuatro términos GO relacionados con la respuesta a estrés de aquellos genes que están potencialmente regulados en *cis* por un ARNncl, dos de ellos relacionados con la respuesta a estrés oxidativo y dos con la respuesta a estrés (Tabla 10). En estos cuatro genes, además, se puede observar claramente la correlación positiva en su expresión con los respectivos ARNncl de la que se ha hablado anteriormente. También se han encontrado algunos genes

relacionados con la síntesis de calcio, que al igual que el estrés oxidativo, pueden intervenir en la activación de la respuesta a estrés por frío, tal y como se comentó en la introducción

Tabla 10. Identificadores de genes expresados diferencialmente con términos GO relacionados con la respuesta a estrés, junto al identificador del ARNncl expresado diferencialmente que se encuentra a menos de 10 KB y los respectivos log2FC. Esta Tabla ha sido generada a partir de la información que contienen los ficheros que se ejemplifican en la Tabla S4 y S5 del Anexo 1

Término GO	ARNm	ARNncl	Log2FC ARNm	Log2FC ARNncl
Respuesta a estrés	MELO3C002099.2	MSTRG.21873.1	-1.274014	-1.089241
Respuesta a estrés	MELO3C015266.2	MSTRG.2293.1	-6.807791	-6.607133
Respuesta a estrés oxidativo	MELO3C014348.2	MSTRG.8271.1	0.9675370	3.479854
Respuesta a estrés oxidativo	MELO3C025286.2	MSTRG.3236.3	-0.8655323	-1.055966

Sería interesante, en investigaciones futuras, la validación experimental y el estudio *in vivo* de estos ARNncl para comprobar si realmente tienen una función reguladora en *cis* y por tanto un papel importante en la respuesta a estrés de la planta de melón.

5. CONCLUSIÓN

Se puede concluir que:

- A pesar de la novedad y la falta de bases de datos curadas o herramientas específicas de ARNncl en plantas, las herramientas de predicción disponibles han funcionado con muy buenos resultados en las cinco especies de plantas utilizadas.
- De acuerdo con los parámetros analizados, FEELnc es el programa más preciso y específico, además de tener el mejor resultado de exactitud y ratio de error. Por otra parte, CPC2 destaca como la herramienta más sensible.
- El estrés por frío en melón ocasiona una respuesta transcripcional amplia, produciéndose tanto procesos de sobreexpresión como de represión, en la misma magnitud e indistintamente entre ARNncl y ARNm.
- Los resultados muestran una correlación positiva entre la expresión de los ARNncl y los ARNm que se encuentran a menos de 10 kb, apuntando a una relación de co-regulación entre ambos. Buena muestra de esto son los cuatro genes relacionados con la respuesta a estrés que se han mostrado y sus respectivos ARNncl.

6. BIBLIOGRAFÍA

Ahmed, W., Xia, Y., Li, R., Bai, G., Siddique, K. H. M., & Guo, P. (2020). Non-coding RNAs: Functional roles in the regulation of stress response in Brassica crops. *Genomics*, *112*(2), 1419–1424. https://doi.org/10.1016/J.YGENO.2019.08.011

Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data. Retrieved June 4, 2022, from http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Budak, H., Kaya, S. B., & Cagirici, H. B. (2020). Long Non-coding RNA in Plants in the Era of Reference Sequences. *Frontiers in Plant Science*, *11*, 276. https://doi.org/10.3389/FPLS.2020.00276/BIBTEX

Chu, K. (1999). An introduction to sensitivity, specificity, predictive values and likelihood ratios. *Emergency Medicine*, *11*(3), 175–181. https://doi.org/10.1046/J.1442-2026.1999.00041.X

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2). https://doi.org/10.1093/GIGASCIENCE/GIAB008

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 2016 539:7629*, *539*(7629), 452–455. https://doi.org/10.1038/nature20149

Fan, X. N., Zhang, S. W., Zhang, S. Y., & Ni, J. J. (2020). Lncrna_mdeep: An alignment-free predictor for distinguishing long non-coding rnas from protein-coding transcripts by multimodal deep learning. *International Journal of Molecular Sciences*, *21*(15), 1–11. https://doi.org/10.3390/ijms21155222

Garcia-Mas, J., Benjak, A., Sanseverino, W., Bourgeois, M., Mir, G., Gonźalez, V. M., Heńaff, E., Cam̂ara, F., Cozzuto, L., Lowy, E., Alioto, T., Capella-Gutieŕrez, S., Blancae, J., Cañizares, J., Ziarsolo, P., Gonzalez-Ibeas, D., Rodriǵuez-Moreno, L., Droege, M., Du, L., … Puigdomeǹech, P. (2012). The genome of melon (Cucumis melo L.). *Proceedings of the National Academy of Sciences of the United States of America*, *109*(29), 11872–11877. https://doi.org/10.1073/PNAS.1205415109/SUPPL_FILE/SD06.XLS

GitHub - lh3/seqtk: Toolkit for processing sequences in FASTA/Q formats. (2018). Retrieved June 15, 2022, from https://github.com/lh3/seqtk

Hossin, M., & Sulaiman. (2020). A review on evaluation metrics for data classification evaluations. *IJDKP) International Journal of Data Mining & Knowledge Management Process (IJDKP)*, *5*(2). https://doi.org/10.5121/ijdkp.2015.5201

Jette, M. A., Jette, M. A., Yoo, A. B., & Grondona, M. (2002). SLURM: Simple Linux Utility for Resource Management. *In lecture notes in computer science: proceedings of job scheduling strategies for parallel processing (JSSPP) 2003*, *2862*, 44--60. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.10.6834

Jin, J., Lu, P., Xu, Y., Li, Z., Yu, S., Liu, J., Wang, H., Chua, N. H., & Cao, P. (2021). 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

Kang, Y. J., Yang, D. C., Kong, L., Hou, M., Meng, Y. Q., Wei, L., & Gao, G. (2017). CPC2: A fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Research*, *45*(W1), W12–W16. https://doi.org/10.1093/nar/gkx428

Klapproth, C., Sen, R., Stadler, P. F., Findeiß, S., & Fallmann, J. (2021). Common Features in lncRNA Annotation and Classification: A Survey. *Non-Coding RNA*, *7*(4), 77. https://doi.org/10.3390/ncrna7040077

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 1–21. https://doi.org/10.1186/S13059-014-0550-8/FIGURES/9

Mangul, S., Martin, L. S., Hill, B. L., Ka-Mei Lam, A., Distler, M. G., Zelikovsky, A., Eskin, E., & Flint, J. (2019). Systematic benchmarking of omics computational tools. Nat Commun, 27;10(1):1393. https://doi.org/10.1038/s41467-019-09406-4

Márquez-Molins, J., Villalba-Bermell, P., Corell-Sierra, J., Pallás, V., Gómez, G., & Gustavo Gomez, C. G. (2022). Multiomic analisys reveals that viroid infection induces a temporal reprograming of plant-defence mechanisms at multiple regulatory levels. *BioRxiv*, 2022.01.06.475203. https://doi.org/10.1101/2022.01.06.475203

Mehrotra, S., Verma, S., Kumar, S., Kumari, S., & Mishra, B. N. (2020). Transcriptional regulation and signalling of cold stress response in plants: An overview of current understanding. *Environmental and Experimental Botany*, *180*, 104243. https://doi.org/10.1016/J.ENVEXPBOT.2020.104243

Mohammed, A. R., Mohammed, S. A., & Shirmohammadi, S. (2019). Machine Learning and Deep Learning Based Traffic Classification and Prediction in Software Defined Networking. *2019 IEEE International Symposium on Measurements and Networking, M and N 2019 - Proceedings*. https://doi.org/10.1109/IWMN.2019.8805044

Nakaminami, K., Matsui, A., Shinozaki, K., & Seki, M. (2012). RNA regulation in plant abiotic stress responses. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, *1819*(2), 149–153. https://doi.org/10.1016/J.BBAGRM.2011.07.015

Neph, S., Kuehn, M. S., Reynolds, A. P., Haugen, E., Thurman, R. E., Johnson, A. K., Rynes, E., Maurano, M. T., Vierstra, J., Thomas, S., Sandstrom, R., Humbert, R., & Stamatoyannopoulos, J. A. (2012). BEDOPS: high-performance genomic feature operations. *Bioinformatics*, *28*(14), 1919–1920. https://doi.org/10.1093/BIOINFORMATICS/BTS277

Nurhasanah Ritonga, F., & Chen, S. (2020). Physiological and Molecular Mechanism Involved in Cold Stress Tolerance in Plants. *Plants 2020, Vol. 9, Page 560*, *9*(5), 560. https://doi.org/10.3390/PLANTS9050560

Ongsulee, P. (2018). Artificial intelligence, machine learning and deep learning. *International Conference on ICT and Knowledge Engineering*, 1–6. https://doi.org/10.1109/ICTKE.2017.8259629

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods 2017 14:4*, *14*(4), 417–419. https://doi.org/10.1038/nmeth.4197

Pertea, M., & Pertea, G. (2020). GFF Utilities: GffRead and GffCompare. *F1000Research*, *9*. https://doi.org/10.12688/F1000RESEARCH.23297.1

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., Salzberg, S. L., & Biotechnol, N. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads HHS Public Access Author manuscript. *Nat Biotechnol*, *33*(3), 290–295. https://doi.org/10.1038/nbt.3122

Pinkney, H. R., Wright, B. M., & Diermeier, S. D. (2020). The lncrna toolkit: Databases and in silico tools for lncrna analysis. In *Non-coding RNA* (Vol. 6, Issue 4, pp. 1–25). MDPI AG. https://doi.org/10.3390/ncrna6040049

Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, *35*(suppl_1), D61–D65. https://doi.org/10.1093/NAR/GKL842

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., & Chang, H. Y. (2016). Unique features of long non-coding RNA biogenesis and function. *Nature Reviews Genetics*, *17*(1), 47-62. https://doi.org/10.1038/nrg.2015.10

Ruggieri, V., Alexiou, K. G., Morata, J., Argyris, J., Pujol, M., Yano, R., Nonaka, S., Ezura, H., Latrasse, D., Boualem, A., Benhamed, M., Bendahmane, A., Cigliano, R. A., Sanseverino, W., Puigdomènech, P., Casacuberta, J. M., & Garcia-Mas, J. (2018). An improved assembly and annotation of the melon (Cucumis melo L.) reference genome. *Scientific Reports 2018 8:1*, *8*(1), 1–9. https://doi.org/10.1038/s41598-018-26416-2

Sanz-Carbonell, A., Marques, M. C., Bustamante, A., Fares, M. A., Rodrigo, G., & Gomez, G. (2019). Inferring the regulatory network of the miRNA-mediated response to biotic and abiotic stress in melon. *BMC Plant Biology*, *19*(1), 1–17. https://doi.org/10.1186/S12870-019-1679-0/FIGURES/9

Shahzad, A., Ullah, S., Afzal, &, Dar, A., Fahad Sardar, M., Mehmood, T., Tufail, M. A., Shakoor, A., & Haris, M. (2021). Nexus on climate change: agriculture and possible solution to cope future climate change stresses. *Environmental Science and Pollution Research*, *28*, 14211–14232. https://doi.org/10.1007/s11356-021-12649-8/Published

Soneson, C., Love, M. I., Robinson, M. D., & Floor, S. N. (2016). Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research 2015 4:1521*, *4*, 1521. https://doi.org/10.12688/f1000research.7563.1

Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J., & Prins, P. (2015). Sambamba: fast processing of NGS alignment formats. *Bioinformatics*, *31*(12), 2032–2034. https://doi.org/10.1093/BIOINFORMATICS/BTV098

Tian, Y., Bai, S., Dang, Z., Hao, J., Zhang, J., & Hasi, A. (2019). Genome-wide identification and characterization of long non-coding RNAs involved in fruit ripening and the climacteric in Cucumis melo. *BMC Plant Biology*, *19*(1). https://doi.org/10.1186/s12870-019-1942-4

Villalba-Bermell, P., Marquez-Molins, J., Marques, M. C., Hernandez-Azurdia, A. G., Corell-Sierra, J., Picó, B., Monforte, A. J., Elena, S. F., & Gomez, G. G. (2021). Combined Stress Conditions in Melon Induce Non-additive Effects in the Core miRNA Regulatory Network. *Frontiers in Plant Science*, *12*, 2522. https://doi.org/10.3389/FPLS.2021.769093/BIBTEX

Walsh, I., Fishman, D., Garcia-Gasulla, D., Titma, T., Pollastri, G., Harrow, J., Psomopoulos, F. E., & Tosatto, S. C. E. (2021). *DOME: recommendations for supervised machine learning validation in biology*. Nat Methods, 18(10):1122-1127. https://doi.org/10.1038/s41592-021-01205-4.

Wang, J., Meng, X., Dobrovolskaya, O. B., Orlov, Y. L., & Chen, M. (2017). Non-coding RNAs and Their Roles in Stress Response in Plants Wang J et al / miRNA and lncRNA in Plant Stress Response. In *Genomics, Proteomics and Bioinformatics* (Vol. 15, Issue 5, pp. 301–312). Beijing Genomics Institute. https://doi.org/10.1016/j.gpb.2017.01.007

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). https://doi.org/10.1093/nar/gkt006

Wilhelm, B. T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., Penkett, C. J., Rogers, J., & Bähler, J. (2008). Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, *453*(7199), 1239–1243. https://doi.org/10.1038/nature07002

Wucher, V., Legeai, F., Hédan, B., Rizk, G., Lagoutte, L., Leeb, T., Jagannathan, V., Cadieu, E., David, A., Lohi, H., Cirera, S., Fredholm, M., Botherel, N., Leegwater, P. A. J., le Béguec, C., Fieten, H., Johnson, J., Alföldi, J., André, C., … Derrien, T. (2017). FEELnc: A tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Research*, *45*(8). https://doi.org/10.1093/nar/gkw1306

Yang, C., Yang, L., Zhou, M., Xie, H., Zhang, C., Wang, M. D., & Zhu, H. (2018). LncADeep: An ab initio lncRNA identification and functional annotation tool based on deep learning. *Bioinformatics*, *34*(22), 3825–3834. https://doi.org/10.1093/bioinformatics/bty428

Zhang, H., Zhu, J., Gong, Z., & Zhu, J. K. (2022). Abiotic stress responses in plants. In *Nature Reviews Genetics* (Vol. 23, Issue 2, pp. 104–119). Nature Research. https://doi.org/10.1038/s41576-021-00413-0

Zheng, Y., Wu, S., Bai, Y., Sun, H., Jiao, C., Guo, S., Zhao, K., Blanca, J., Zhang, Z., Huang, S., Xu, Y., Weng, Y., Mazourek, M., Reddy, U. K., Ando, K., McCreight, J. D., Schaffer, A. A., Burger, J., Tadmor, Y., … Fei, Z. (2019). Cucurbit Genomics Database (CuGenDB): a central portal for comparative and functional genomics of cucurbit crops. *Nucleic Acids Research*, *47*(D1), D1128–D1136. https://doi.org/10.1093/NAR/GKY944

Zhu, J. K. (2016). Abiotic Stress Signaling and Responses in Plants. *Cell*, *167*(2), 313–324. https://doi.org/10.1016/J.CELL.2016.08.029

ANEXO 1: MATERIAL SUPLEMENTARIO

Tabla S1. Valores de VP, VN, FP, FN, VPv y FNv usados para calcular las métricas de cada uno de los programas para cada una de las especies.

Especie	Programa	VP	VN	FP	FN	VP_v	FN_v
A.thaliana	FEELnc	6163	13488	1	219	118	7
O.sativa	FEELnc	4462	11527	0	284	25	0
Z.mays	FEELnc	8285	32161	2	2414	13	3
S.lycopersicum	FEELnc	7889	8709	0	100	17	2
C.sativus	FEELnc	5984	8697	0	610	5	1
A.thaliana	CPAT	12041	13188	411	1558	127	17
O.sativa	CPAT	9234	11407	158	2331	21	19
Z.mays	CPAT	20800	31424	820	11597	19	7
S.lycopersicum	CPAT	6215	8548	191	2526	17	2
C.sativus	CPAT	6438	8583	169	2315	7	1
A.thaliana	CPC2	13419	12572	1027	180	141	3
O.sativa	CPC2	11554	10959	606	11	37	3
Z.mays	CPC2	31328	30043	2354	1069	25	1
S.lycopersicum	CPC2	8573	8354	387	168	18	1
C.sativus	CPC2	8513	8319	434	240	8	0
A.thaliana	LncADeep	12292	13112	487	1307	129	15
O.sativa	LncADeep	10834	11027	538	731	39	1
Z.mays	LncADeep	26619	30409	1988	5778	24	2
S.lycopersicum	LncADeep	8239	8505	236	502	18	1
C.sativus	LncADeep	7119	8543	210	1634	8	0
A.thaliana	lncRNA_Mdeep	12532	13038	561	1067	134	10
O.sativa	lncRNA_Mdeep	10469	11271	294	1096	34	6
Z.mays	lncRNA_Mdeep	25060	31330	1067	7337	20	6
S.lycopersicum	lncRNA_Mdeep	7746	8488	253	995	17	2
C.sativus	lncRNA_Mdeep	7506	8457	296	1247	7	1

Tabla S2. Métricas de evaluación de cada uno de los programas para cada especie.

Especie	Programa	Exactitud	Ratio error	Sensibilidad	Sensibilidad validados	Especificidad	Precisión	VPN	PuntuaciónF1
A.thaliana	FEELnc	98,89	1,11	96,57	94,40	99,99	99,98	98,40	98,25
O.sativa	FEELnc	98,25	1,75	94,02	100,00	100,00	100,00	97,60	96,92
Z.mays	FEELnc	94,36	5,64	77,44	81,25	99,99	99,98	93,02	87,27
S.lycopersicum	FEELnc	99,40	0,60	98,75	89,47	100,00	100,00	98,86	99,37
C.sativus	FEELnc	96,01	3,99	90,75	83,33	100,00	100,00	93,45	95,15
A.thaliana	CPAT	92,76	7,24	88,54	88,19	96,98	96,70	89,43	92,44
O.sativa	CPAT	89,24	10,76	79,84	52,50	98,63	98,32	83,03	88,12
Z.mays	CPAT	80,79	19,21	64,20	73,08	97,46	96,21	73,04	77,01
S.lycopersicum	CPAT	84,46	15,54	71,10	89,47	97,81	97,02	77,19	82,06
C.sativus	CPAT	85,81	14,19	73,55	87,50	98,07	97,44	78,76	83,83
A.thaliana	CPC2	95,56	4,44	98,68	97,92	92,45	92,89	98,59	95,70
O.sativa	CPC2	97,33	2,67	99,90	92,50	94,76	95,02	99,90	97,40
Z.mays	CPC2	94,72	5,28	96,70	96,15	92,73	93,01	96,56	94,82
S.lycopersicum	CPC2	96,83	3,17	98,08	94,74	95,57	95,68	98,03	96,86
C.sativus	CPC2	96,15	3,85	97,26	100,00	95,04	95,15	97,20	96,19
A.thaliana	LncADeep	93,40	6,60	90,39	89,58	96,42	96,19	90,94	93,20
O.sativa	LncADeep	94,51	5,49	93,68	97,50	95,35	95,27	93,78	94,47
Z.mays	LncADeep	88,01	11,99	82,17	92,31	93,86	93,05	84,03	87,27
S.lycopersicum	LncADeep	95,78	4,22	94,26	94,74	97,30	97,22	94,43	95,71
C.sativus	LncADeep	89,47	10,53	81,33	100,00	97,60	97,13	83,94	88,53
A.thaliana	lncRNA_Mdeep	94,01	5,99	92,15	93,06	95,87	95,72	92,44	93,90
O.sativa	lncRNA_Mdeep	93,99	6,01	90,52	85,00	97,46	97,27	91,14	93,77
Z.mays	lncRNA_Mdeep	87,03	12,97	77,35	76,92	96,71	95,92	81,03	85,64
S.lycopersicum	lncRNA_Mdeep	92,86	7,14	88,62	89,47	97,11	96,84	89,51	92,54
C.sativus	lncRNA_Mdeep	91,19	8,81	85,75	87,50	96,62	96,21	87,15	90,68

Tabla S3. Ejemplo de las 20 primeras líneas del fichero de salida de Deseq2

ID_transcript	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj	COLD-S1_S1	OLD-S2_S2	COLD-S3_S3	NT-S11_S8	NT-S1_S5	NT-S2_S7	ID_Gene	log10baseMean	Type	ID
maker-chr01-exonerate_est2genome-gene-0,0-mRNA-1	2,44E+13	4,85E-01	8,06E-02	6,03E+14	1,69E+05	5,48E+05	2,81E+14	2,89E+14	2,83E+14	2,08E+13	2,04E+14	1,98E+14	maker-chr01	3,39E+14	up-regulated	mRNA
maker-chr01-exonerate_est2genome-gene-0,1-mRNA-1	1,33E+13	2,44E-01	1,07E-01	2,28E+14	0,02288176	0,03409788	1,47E+14	1,34E+14	1,53E+13	1,14E+14	1,27E+13	1,25E+14	maker-chr01	3,13E+13	up-regulated	mRNA
maker-chr01-exonerate_est2genome-gene-0,2-mRNA-1	3,34E+14	-1,03E+14	1,35E-01	-7,64E+14	2,21E+00	1,01E+00	2,22E+14	2,13E+14	2,25E+14	4,34E+14	4,64E+14	4,46E+14	maker-chr01	2,53E+14	down-regulated (<=-1)	mRNA
maker-chr01-exonerate_est2genome-gene-0,3-mRNA-1	1,02E+14	-1,43E+13	1,22E-01	-1,18E+14	4,66E-18	5,37E-17	6,19E+14	4,72E+14	5,69E+14	1,56E+14	1,47E+14	1,45E+14	maker-chr01	4,01E+13	down-regulated (<=-1)	mRNA
maker-chr01-exonerate_est2genome-gene-0,4-mRNA-1	6,48E+14	1,77E+13	1,40E-01	1,27E+14	9,64E-23	1,33E-21	9,81E+14	1,12E+14	9,04E+13	2,77E+14	3,15E+14	2,90E+14	maker-chr01	2,81E+14	up-regulated (>=1)	mRNA
maker-chr01-exonerate_est2genome-gene-0,5-mRNA-1	1,50E+14	-1,91E+12	1,42E-01	-1,35E+13	1,55E-27	2,62E-26	6,92E+14	4,98E+14	6,93E+14	2,36E+14	2,40E+14	2,33E+14	maker-chr01	4,18E+14	down-regulated (<=-1)	mRNA
maker-chr01-exonerate_est2genome-gene-0,6-mRNA-1	2,97E+14	-5,36E-01	1,71E-01	-3,14E+14	0,00167204	0,00294235	2,64E+14	2,02E+14	2,64E+14	3,56E+14	3,30E+13	3,64E+14	maker-chr01	2,47E+13	down-regulated	mRNA
maker-chr01-exonerate_est2genome-gene-0,7-mRNA-1	4,95E+14	4,82E+14	5,86E-01	8,23E+14	1,92E-02	9,96E-03	7,32E+14	1,19E+14	9,45E+14	2,45E+13	3,63E+14	3,98E+14	maker-chr01	1,70E+14	up-regulated (>=1)	mRNA
maker-chr01-exonerate_est2genome-gene-0,8-mRNA-1	1,49E+14	-9,18E-01	2,06E-01	-4,46E+14	8,23E+08	1,90E+09	1,07E+14	1,05E+14	9,62E+14	2,26E+14	1,91E+14	1,68E+14	maker-chr01	2,18E+14	down-regulated	mRNA
maker-chr01-exonerate_est2genome-gene-0,9-mRNA-1	1,78E+14	6,15E-01	1,75E-01	3,51E+14	0,00044197	0,00084041	2,29E+14	2,26E+14	1,92E+14	1,07E+14	1,46E+14	1,70E+14	maker-chr01	3,25E+14	up-regulated	mRNA
maker-chr01-exonerate_est2genome-gene-1,0-mRNA-1	5,32E+14	-3,16E-01	1,64E-01	-1,92E+14	0,05434383	0,07597723	5,31E+14	4,34E+14	4,58E+14	6,87E+14	5,84E+13	5,00E+14	maker-chr01	2,73E+14	non-significant	mRNA
maker-chr01-exonerate_est2genome-gene-1,1-mRNA-1	6,78E+14	-3,51E-01	1,46E-01	-2,41E+14	0,01593242	0,02437683	6,00E+14	5,05E+14	6,88E+14	7,48E+14	7,55E+14	7,76E+14	maker-chr01	2,83E+14	down-regulated	mRNA
maker-chr01-exonerate_est2genome-gene-1,11-mRNA-1	1,37E+14	-1,58E+14	1,80E-01	-8,77E+14	1,86E-04	1,09E-03	7,83E+14	4,88E+14	7,88E+14	2,14E+14	1,95E+14	2,05E+14	maker-chr01	4,14E+14	down-regulated (<=-1)	mRNA
maker-chr01-exonerate_est2genome-gene-1,12-mRNA-1	9,14E+14	-4,82E-01	9,68E-02	-4,98E+14	6,44E+07	1,67E+08	7,60E+14	7,83E+14	7,44E+14	1,05E+14	1,08E+13	1,07E+14	maker-chr01	2,96E+13	down-regulated	mRNA
maker-chr01-exonerate_est2genome-gene-1,13-mRNA-1	4,73E+14	-4,40E+14	1,18E-01	-3,74E+14	2,15E-292	3,42E-288	4,33E+14	4,00E+14	4,48E+14	9,68E+14	8,00E+14	9,39E+14	maker-chr01	3,67E+14	down-regulated (<=-1)	mRNA
maker-chr01-exonerate_est2genome-gene-1,2-mRNA-1	1,59E+14	7,10E-01	1,99E-01	3,56E+14	0,00036955	0,00070971	1,87E+14	2,07E+14	1,98E+14	8,18E+14	1,40E+14	1,41E+14	maker-chr01	3,20E+14	up-regulated	mRNA
maker-chr01-exonerate_est2genome-gene-1,3-mRNA-1	4,74E+14	9,41E-01	1,85E-01	5,09E+14	3,57E+07	9,50E+07	6,56E+14	6,63E+14	5,49E+14	2,81E+14	4,07E+14	2,86E+14	maker-chr01	2,68E+14	up-regulated	mRNA
maker-chr01-exonerate_est2genome-gene-1,4-mRNA-1	5,68E+13	-1,91E+14	1,40E-01	-1,36E+14	2,52E-28	4,37E-27	2,29E+14	2,67E+14	2,17E+14	9,50E+13	8,34E+14	9,12E+14	maker-chr01	2,76E+14	down-regulated (<=-1)	mRNA
maker-chr01-exonerate_est2genome-gene-1,5-mRNA-1	1,35E+14	-6,48E-01	2,38E-01	-2,73E+13	0,00640316	0,01038633	8,49E+14	1,05E+13	1,27E+13	1,37E+14	1,71E+14	1,86E+14	maker-chr01	2,13E+14	down-regulated	mRNA
maker-chr01-exonerate_est2genome-gene-1,7-mRNA-1	4,25E+14	4,64E+14	6,43E-01	7,22E+14	5,07E+01	2,11E+01	6,17E+13	1,21E+14	6,26E+13	1,18E+14	3,53E+14	4,95E+14	maker-chr01	1,64E+14	up-regulated (>=1)	mRNA

Tabla S4: Ejemplo de las 20 primeras filas de las parejas de ARNm y ARNncl localizados a menos de 10 kb aguas arriba/abajo y sus respectivos log2FC.

ID_transcript	ID_lncRNA	log2FC_lncRNA	log2FC_mRNA
maker-chr01-exonerate_est2genome-gene-118.3-m...	MSTRG.768.1	-2.2501079	-0.5127701
maker-chr01-exonerate_est2genome-gene-118.3-m...	MSTRG.774.2	4.7564580	-0.5127701
maker-chr01-exonerate_est2genome-gene-118.3-m...	MSTRG.774.1	1.7492346	-0.5127701
maker-chr01-exonerate_est2genome-gene-118.4-m...	MSTRG.768.1	-2.2501079	2.3555790
maker-chr01-exonerate_est2genome-gene-118.4-m...	MSTRG.774.1	1.7492346	2.3555790
maker-chr01-exonerate_est2genome-gene-118.4-m...	MSTRG.774.2	4.7564580	2.3555790
maker-chr01-exonerate_est2genome-gene-118.5-m...	MSTRG.774.1	1.7492346	2.6160581
maker-chr01-exonerate_est2genome-gene-118.5-m...	MSTRG.774.2	4.7564580	2.6160581
maker-chr01-exonerate_est2genome-gene-149.41-...	MSTRG.898.1	1.9040826	0.5526302
maker-chr01-exonerate_est2genome-gene-197.22-...	MSTRG.1157.4	1.7203557	0.8879218
maker-chr01-exonerate_est2genome-gene-278.34-...	MSTRG.1477.42	1.2530796	-0.3809477
maker-chr01-exonerate_est2genome-gene-302.3-m...	MSTRG.1608.2	-8.8823056	-0.3924978
maker-chr01-exonerate_est2genome-gene-302.3-m...	MSTRG.1608.5	-2.6696582	-0.3924978
maker-chr01-exonerate_est2genome-gene-302.4-m...	MSTRG.1608.2	-8.8823056	0.4356979
maker-chr01-exonerate_est2genome-gene-302.4-m...	MSTRG.1608.5	-2.6696582	0.4356979
maker-chr01-exonerate_est2genome-gene-324.6-m...	MSTRG.1826.1	-1.4116113	-1.3380051
maker-chr01-exonerate_est2genome-gene-324.7-m...	MSTRG.1826.1	-1.4116113	-3.4961570
maker-chr01-exonerate_est2genome-gene-335.1-m...	MSTRG.1952.3	1.0921980	0.9200227
maker-chr01-exonerate_est2genome-gene-335.10-...	MSTRG.1950.4	2.0871976	0.4714210
maker-chr01-exonerate_est2genome-gene-335.11-...	MSTRG.1950.4	2.0871976	1.5167210

Tabla S5. Ejemplo de las 20 primeras filas del fichero que contiene las parejas de ARNm y ARNncl localizados a menos de 10 kb aguas arriba/abajo y los correspondientes términos GO de los ARNm.

Name_transcript	GO_term	Function	Ubication	ID_transcript	ID_lncRNA
MELO3C001983.2	GO:0016020	cellular_component	membrane	maker-chr12-exonerate_est2genome-gene-256.13-...	MSTRG.21985.4
MELO3C002014.2	GO:0005484	molecular_function	SNAP receptor activity	maker-chr12-exonerate_est2genome-gene-254.7-m...	MSTRG.21955.1
MELO3C002014.2	GO:0015031	biological_process	protein transport	maker-chr12-exonerate_est2genome-gene-254.7-m...	MSTRG.21955.1
MELO3C002014.2	GO:0006810	biological_process	transport	maker-chr12-exonerate_est2genome-gene-254.7-m...	MSTRG.21955.1
MELO3C002014.2	GO:0007049	biological_process	cell cycle	maker-chr12-exonerate_est2genome-gene-254.7-m...	MSTRG.21955.1
MELO3C002014.2	GO:0051301	biological_process	cell division	maker-chr12-exonerate_est2genome-gene-254.7-m...	MSTRG.21955.1
MELO3C002014.2	GO:0005886	cellular_component	plasma membrane	maker-chr12-exonerate_est2genome-gene-254.7-m...	MSTRG.21955.1
MELO3C002014.2	GO:0051707	biological_process	response to other organism	maker-chr12-exonerate_est2genome-gene-254.7-m...	MSTRG.21955.1
MELO3C002014.2	GO:0016021	cellular_component	integral component of membrane	maker-chr12-exonerate_est2genome-gene-254.7-m...	MSTRG.21955.1
MELO3C002014.2	GO:0009737	biological_process	response to abscisic acid	maker-chr12-exonerate_est2genome-gene-254.7-m...	MSTRG.21955.1
MELO3C002014.2	GO:0016192	biological_process	vesicle-mediated transport	maker-chr12-exonerate_est2genome-gene-254.7-m...	MSTRG.21955.1
MELO3C002014.2	GO:0016020	cellular_component	membrane	maker-chr12-exonerate_est2genome-gene-254.7-m...	MSTRG.21955.1
MELO3C002014.2	GO:0061025	biological_process	membrane fusion	maker-chr12-exonerate_est2genome-gene-254.7-m...	MSTRG.21955.1
MELO3C002014.2	GO:0005515	molecular_function	protein binding	maker-chr12-exonerate_est2genome-gene-254.7-m...	MSTRG.21955.1
MELO3C002014.2	GO:0009507	cellular_component	chloroplast	maker-chr12-exonerate_est2genome-gene-254.7-m...	MSTRG.21955.1
MELO3C002014.2	GO:0009612	biological_process	response to mechanical stimulus	maker-chr12-exonerate_est2genome-gene-254.7-m...	MSTRG.21955.1
MELO3C002014.2	GO:0009504	cellular_component	cell plate	maker-chr12-exonerate_est2genome-gene-254.7-m...	MSTRG.21955.1
MELO3C002014.2	GO:0000911	biological_process	cytokinesis by cell plate formation	maker-chr12-exonerate_est2genome-gene-254.7-m...	MSTRG.21955.1
MELO3C002015.2	GO:0005515	molecular_function	protein binding	maker-chr12-exonerate_est2genome-gene-254.6-m...	MSTRG.21955.1
MELO3C002044.2	GO:0003677	molecular_function	DNA binding	maker-chr12-exonerate_est2genome-gene-252.15-...	MSTRG.21923.5

ANEXO 2: CÓDIGOS SUPLEMENTARIOS

Código 1. Programa de Python usado para eliminar los transcritos con código de clases correspondientes a los errores (p, r, s).

```
#!/usr/bin/env python3
# -*- coding: utf-8 -*-

"""
This script is aim to remove transcripts with class codes not interesting or incorrectly assembled (p,r,s).
"""
## MODULES

import sys

## PIPELINE

path_GTF_initial = sys.argv[1]
path_GTF_filtered = sys.argv[2]

GTF_initial = open(path_GTF_initial, "r+")
GTF_filtered = open(path_GTF_filtered, "w")

for line in GTF_initial:
   line=line.rstrip().split("\t")

   if line[2] == "transcript":
      if ('class_code "p"' not in line[-1]) and \
               ('class_code "r"' not in line[-1]) and \
               ('class_code "s"' not in line[-1]):
         var="+"
         GTF_filtered.write("%s\n"%("\t".join(line)))
      else:
         var="-"

   elif line[2]=="exon":
      if var=="+":
         GTF_filtered.write("%s\n"%("\t".join(line)))
```

Código 2. Programa de Python usado para seleccionar los transcritos con código de clases correspondientes a los ARNncl (u, x, i, o, e).

```
#!/usr/bin/env python3
# -*- coding: utf-8 -*-

"""
This script is aim to select the class codes: "x" (antisense), "i" (intronic),"u" (intergenic) and "o"/"e" (exon sense) as novel transcripts.
"""

## MODULES

import sys

## PIPELINE

path_GTF_initial = sys.argv[1]
path_GTF_filtered = sys.argv[2]

GTF_initial = open(path_GTF_initial, "r+")
GTF_filtered = open(path_GTF_filtered, "w")

for line in GTF_initial:
   line=line.rstrip().split("\t")

   if line[2] == "transcript":
      if ('class_code "x"' in line[-1]) or \
               ('class_code "u"' in line[-1]) or \
               ('class_code "i"' in line[-1]) or \
               ('class_code "o"' in line[-1]) or \
               ('class_code "e"' in line[-1]):
         var="+"
         GTF_filtered.write("%s\n"%("\t".join(line)))
      else:
         var="-"

   elif line[2]=="exon":
      if var=="+":
         GTF_filtered.write("%s\n"%("\t".join(line)))
```