Programa de Doctorado en Bioingeniería
Universidad Miguel Hernández de Elche

CXIP4 contiene un motivo ZCCHC, es esencial y participa en el *splicing* en *Arabidopsis thaliana*

Uri Israel Aceituno Valenzuela

Directora: María Rosa Ponce Molet
Codirectora: Raquel Sarmiento Mañús

Elche, 2025

CXIP4 contiene un motivo ZCCHC, es esencial y participa en el *splicing* en *Arabidopsis thaliana*

Trabajo realizado por el Ingeniero Uri Israel Aceituno Valenzuela, en la Unidad de Genética del Instituto de Bioingeniería de la Universidad Miguel Hernández de Elche, para optar al grado de Doctor.

Elche, 30 de enero de 2025

Esta Tesis Doctoral, titulada “CXIP4 contiene un motivo ZCCHC, es esencial y participa en el *splicing* en *Arabidopsis thaliana*”, se presenta bajo la modalidad de **tesis por compendio** de las siguientes **publicaciones**:

Aceituno-Valenzuela, U., Micol-Ponce, R., y Ponce, M.R. (2020). Genome-wide analysis of CCHC-type zinc finger (ZCCHC) proteins in yeast, Arabidopsis, and humans. *Cellular and Molecular Life Sciences* **77**, 3991-4014.

Aceituno-Valenzuela, U., Fontcuberta-Cervera, S., Micol-Ponce, R., Sarmiento-Mañús, R., Ruiz-Bayón, A., y Ponce, M.R. (2025). CAX-INTERACTING PROTEIN4 depletion causes early lethality and pre-mRNA missplicing in Arabidopsis. *Plant Physiology* **197**, kiae641.

MARÍA ROSA PONCE MOLET, Catedrática de Genética de la Universidad Miguel Hernández de Elche, y

RAQUEL SARMIENTO MAÑÚS, Doctora por la Universidad Miguel Hernández de Elche,

HACEMOS CONSTAR:

Que el presente trabajo ha sido realizado bajo nuestra dirección o codirección recoge fielmente la labor desarrollada por el Ingeniero Uri Israel Aceituno Valenzuela para optar al grado de Doctor. Las investigaciones reflejadas en esta memoria se han desarrollado íntegramente en la Unidad de Genética del Instituto de Bioingeniería de la UMH, según los términos y condiciones definidos en el Plan de Investigación del doctorando, y cumpliendo los objetivos inicialmente previstos de forma satisfactoria y lo establecido en el Código de Buenas Prácticas de la UMH.

Maria Rosa Ponce Molet	Raquel Sarmiento Mañús
Directora	Codirectora

Elche, 30 de enero de 2025

PIEDAD NIEVES DE AZA MOYA, Coordinadora del Programa de Doctorado en Bioingeniería de la Universidad Miguel Hernández de Elche por Resolución Rectoral 0169/17, de 1 de febrero de 2017:

HACE CONSTAR:

Que da su conformidad a la presentación de la Tesis Doctoral de Don Uri Israel Aceituno Valenzuela, titulada “CXIP4 contiene un motivo ZCCHC, es esencial y participa en el *splicing* en *Arabidopsis thaliana*”, que se ha desarrollado dentro del Programa de Doctorado en Bioingeniería bajo la dirección de la profesora María Rosa Ponce Molet y codirigida por la doctora Raquel Sarmiento Mañús.

Lo que firmo en Elche, a instancias del interesado y a los efectos oportunos, a treinta de enero de dos mil veinticinco.

Profesora PIEDAD NIEVES DE AZA MOYA
Coordinadora del Programa de Doctorado en Bioingeniería

A Paulina mi esposa, a mis hijos Lía y Vito,

A mis padres, hermanos y seres queridos.

ÍNDICE DE MATERIAS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

I.- PREFACIO

I.- PREFACIO

Siguiendo la normativa de la Universidad Miguel Hernández de Elche para la "Presentación de Tesis Doctorales por compendio de publicaciones", este documento se ha dividido en las partes siguientes:

I.- Este *Prefacio*.

II.- Un *Resumen* en español.

III.- Un *Summary* en inglés.

IV.- Una *Introducción*, en la que se presenta el tema de la Tesis y los antecedentes y objetivos del trabajo realizado.

V.- Un resumen de los *Materiales y métodos* de las publicaciones de la Tesis.

VI.- Un resumen de los *Resultados y discusión* de las publicaciones de la Tesis.

VII.- Un resumen de las *Conclusiones y perspectivas* del trabajo realizado.

VIII.- Una *Bibliografía* de los apartados IV-VII. Algunas de las referencias que incluye se repiten en las bibliografías de los artículos incluidos en esta memoria.

IX.- Un apartado de *Publicaciones,* que incluyen las dos siguientes, en las que se indica el factor de impacto [FI] del año correspondiente, o el más reciente disponible.

Aceituno-Valenzuela, U., Micol-Ponce, R., y Ponce, M.R. (2020). Genome-wide analysis of CCHC-type zinc finger (ZCCHC) proteins in yeast, Arabidopsis, and humans. *Cellular and Molecular Life Sciences* **77**, 3991-4014. [JCR IF: 9.262; D1; 30/295 en la categoría de Biochemistry & Molecular Biology].

Aceituno-Valenzuela, U., Fontcuberta-Cervera, S., Micol-Ponce, R., Sarmiento-Mañús, R., Ruiz-Bayón, A., y Ponce, M.R. (2025). CAX-INTERACTING PROTEIN4 depletion causes early lethality and pre-mRNA missplicing in Arabidopsis. *Plant Physiology* **197**, kiae641. [JCR FI: 6,6; D1; 18/265 en Plant Sciences].

Los "Supplementary Data Sets" de estos artículos no se han incluido en esta memoria por su gran longitud. Las correspondientes hojas de cálculo se remitirán a los miembros del tribunal en formato electrónico.

X.- Un apartado de *Agradecimientos*.

II.- RESUMEN

II.- RESUMEN

La NF-kappa B Activating Protein (NKAP) humana participa en varias facetas del metabolismo del ARN, como la ligación de exones durante el *splicing* y la lectura de la marca epitranscriptómica m^6A en los microARN (miARN) y los ARN mensajeros (ARNm); su presunta ortóloga en Arabidopsis es MORPHOLOGY OF ARGONAUTE1-52 SUPPRESSED (MAS2). El gen *MAS2* fue identificado en el laboratorio de María Rosa Ponce, en una búsqueda de supresores extragénicos del fenotipo morfológico de *argonaute1-52* (*ago1-52*), un alelo hipomorfo y viable de *AGO1*. La proteína AGO1 es una endorribonucleasa que juega un papel central en el silenciamiento postranscripcional mediado por miARN. El alelo *ago1-52* es portador de una mutación que daña uno de los 5′SS (sitio 5′ o donante del *splicing*) y causa la retención parcial del correspondiente intrón. Los alelos *mas2* identificados en el laboratorio de M.R. Ponce suprimen parcialmente el *splicing* aberrante del transcrito primario de *ago1-52*.

En el laboratorio de M.R. Ponce se realizó también un ensayo del doble híbrido de la levadura para identificar interactores de la proteína MAS2. La presunta interactora más representada en dicha búsqueda fue CAX-INTERACTING PROTEIN 4 (CXIP4), que contiene un dedo de zinc de 18 aminoácidos denominado motivo ZCCHC o “zinc knuckle”, cuya estructura central es $CX_2CX_4HX_4C$ (C es cisteína, H es histidina, y X, cualquier aminoácido), flanqueada por dos aminoácidos no conservados en cada extremo. Se conocen muchas proteínas con motivos ZCCHC, pero son muy pocas las que se han caracterizado.

En la parte inicial de esta Tesis se pretendió obtener una visión panorámica del conocimiento sobre las proteínas eucarióticas con motivos ZCCHC, que se ha plasmado en una revisión sistemática de estas proteínas presentes en *Saccharomyces cerevisiae*, *Arabidopsis thaliana* y *Homo sapiens*, en representación de los reinos de los hongos, las plantas y los metazoos, respectivamente. Hemos realizado búsquedas de proteínas que contienen la secuencia $CX_2CX_4HX_4C$ en bases de datos específicas de los tres organismos mencionados, utilizando los programas PatMatch para las de la levadura y Arabidopsis, y ScanProsite para las humanas. Hemos analizado la información estructural y funcional de todas estas proteínas disponible en UniProtKB, concluyendo que existen siete en la levadura, 34 en la especie humana y 69 en Arabidopsis. Aunque la mayoría de las proteínas que hemos considerado contenían un solo motivo ZCCHC, otras tenían más, hasta ocho. Los motivos ZCCHC estaban presentes en cualquier región de la proteína y aparecían solos o combinados con otros dominios o motivos, muchos de los cuales eran de unión a ARN. Además, encontramos que muchas proteínas que contienen motivos ZCCHC albergan regiones de baja

complejidad, principalmente ricas en prolina (P), leucina (L), glicina (G), arginina (R) o serina (S).

Muchas de las proteínas con motivos ZCCHC previamente estudiadas participan en diferentes facetas del metabolismo del ARNm, como la elongación de la transcripción, la poliadenilación, la traducción, el *splicing*, la exportación del ARNm maduro y su degradación, así como en la biogénesis de los ARN ribosómicos (ARNr) y los miARN, y en el silenciamiento génico postranscripcional mediado por estos últimos. Las proteínas con motivos ZCCHC son principalmente nucleares, aunque algunas tienen localización dual en el citoplasma y el núcleo, y muy pocas son exclusivamente citoplásmicas. Solo una de ellas se localiza en el cloroplasto de Arabidopsis, en donde participa en la regulación del *splicing* de los genes de este orgánulo que codifican ARNr.

Las proteínas presuntamente ortólogas que hemos estudiado conservan su localización subcelular y el número de sus motivos ZCCHC; en algunos casos, estos motivos determinan la localización subcelular de dichas proteínas o su translocación o su unión a otras proteínas o ácidos nucleicos. Construimos un árbol filogenético con los 198 motivos ZCCHC que identificamos en siete proteínas de la levadura, 69 de Arabidopsis y 34 humanas. Encontramos que los motivos ZCCHC de la mayoría de los presuntos ortólogos se agrupan en clados, lo que indica la existencia de una mayor conservación de los aminoácidos distintos a las tres C y la H características de este motivo. La conservación de estos aminoácidos sugiere su importancia para la actividad de la proteína ancestral de la que derivan dichos ortólogos y para ellos mismos.

Hemos estudiado en esta Tesis el gen *CXIP4* de Arabidopsis, que está anotado como un ortólogo lejano del *SREK1 Interacting Protein 1* (*SREK1IP1*) humano, al que se supone relacionado con el *splicing*. SREK1IP1 y CXIP4 solo presentan un motivo conocido: un ZCCHC, que está muy conservado y se agrupa en el árbol filogenético que hemos obtenido, lo que sugiere su trascendencia funcional. No se observan otras similitudes entre estas dos proteínas en el resto de su secuencia, aunque ambas presentan regiones de baja complejidad en su mitad carboxiterminal. Aunque existen presuntos ortólogos de *CXIP4* y *SREK1IP1* en muchas plantas y animales, no se ha descrito ningún alelo mutante.

Hemos caracterizado los mutantes insercionales *cxip4-1* y *cxip4-2*, que son portadores de un alelo nulo y otro hipomorfo de *CXIP4*, respectivamente. La letalidad postembrionaria temprana de *cxip4-1* indica que *CXIP4* es esencial en Arabidopsis. La viabilidad de *cxip4-2* nos ha permitido establecer la implicación de CXIP4 en varias rutas de desarrollo, ya que este mutante presenta un fenotipo muy pleiotrópico en todas las fases de su ciclo de vida. La pleiotropía del fenotipo morfológico de *cxip4-2* se corresponde con su perfil transcriptómico,

ya que una parte importante de sus genes desregulados codifican factores de transcripción cuya desregulación podría explicar diferentes rasgos de su fenotipo.

Hemos analizado el *splicing* global de las plantas mutantes *cxip4-2*, constatando un incremento del *splicing* alternativo, principalmente de los eventos de retención intrónica, aunque menor que los descritos en mutantes portadores de alelos de genes que codifican componentes del espliceosoma. Hemos confirmado algunos de los eventos de retención intrónica diferencial entre Col-0 y *cxip4-2*, utilizando para ello plantas del mutante presuntamente nulo *cxip4-1*, en el que las alteraciones del *splicing* son más evidentes, como esperábamos. Estos resultados sugieren que CXIP4 participa directa o indirectamente en el *splicing* de los pre-ARNm. El análisis del enriquecimiento en categorías funcionales de los genes que sufren *splicing* alternativo y se comportan diferencialmente en *cxip4-2* respecto a la estirpe silvestre sugiere que CXIP4 controla el *splicing* alternativo de grupos específicos de genes, entre ellos los implicados en la degradación de ARNm mediante desadenilación y la represión epigenética mediante metilación guiada por ARN no codificantes. La acumulación nuclear de ARN poliadenilados que hemos observado en las plantas *cxip4-2* es compatible con la implicación de CXIP4 en el *splicing* y refuerza la hipótesis de la conservación de las funciones de la SREK1IP1 humana y la CXIP4 de Arabidopsis.

III.- SUMMARY

III.- SUMMARY

Human NF-kappa B Activating Protein (NKAP) participates in diverse facets of RNA metabolism, such as exon ligation during splicing, and reading of the epitranscriptomic mark m^6A on microRNAs (miRNAs) and messenger RNAs (mRNAs). The putative orthologue of human NKAP in Arabidopsis is MORPHOLOGY OF ARGONAUTE1-52 SUPPRESSED (MAS2). The *MAS2* gene was identified in the laboratory of María Rosa Ponce, in a screen for extragenic suppressors of the morphological phenotype of *argonaute1-52* (*ago1-52*), a hypomorphic and viable allele of *AGO1*. The AGO1 protein is an endoribonuclease that plays a key role in miRNA-mediated post-transcriptional silencing. The *ago1-52* allele carries a mutation that damages one of its 5′SS (splicing acceptor), causing partial retention of the corresponding intron. The *mas2* alleles identified in the laboratory of M.R. Ponce partially suppress the aberrant splicing of the primary transcript of *ago1-52*.

A yeast two-hybrid assay was also performed in the laboratory of M.R. Ponce to identify interactors of the MAS2 protein. The most represented putative interactor found in such search was CAX-INTERACTING PROTEIN 4 (CXIP4), which contains a zinc finger called the ZCCHC motif or "zinc knuckle" with the structure $CX_2CX_4HX_4C$ (C is cysteine, H is histidine, and X is any amino acid). Many proteins with ZCCHC motifs are known, but very few have been characterized.

The initial part of this Thesis aimed to acquire a panoramic view of the state of the art on eukaryotic proteins with ZCCHC motifs, summarized in a systematic review of these proteins, covering *Saccharomyces cerevisiae*, *Arabidopsis thaliana*, and *Homo sapiens*, representing the fungal, plant, and metazoan kingdoms respectively. We conducted searches for proteins containing the $CX_2CX_4HX_4C$ sequence in databases specific of these organisms, using the programs Patmatch for yeast and Arabidopsis, and ScanProsite for humans. We analyzed the structural and functional information on these proteins available in UniProtKB, concluding that there are seven in yeast, 34 in humans, and 69 in Arabidopsis. While most of the proteins that we considered had a single ZCCHC motif, some had more, up to eight. These motifs were found in various regions of the protein, alone or combined with other domains or motifs, many of which are RNA-binding. Additionally, we found that many ZCCHC-containing proteins harbor regions of low complexity, mainly rich in proline (P), leucine (L), glycine (G), arginine (R), or serine (S), including RS domains found in proteins related to mRNA splicing.

Many previously studied ZCCHC motif-containing proteins participate in different facets of RNA metabolism, including transcription elongation, polyadenylation, translation, splicing, export and degradation of mRNA, as well as miRNA and ribosomal RNA (rRNA) biogenesis,

and miRNA-mediated post-transcriptional gene silencing. ZCCHC motif-containing proteins are primarily nuclear, although some have dual localization in the cytoplasm and nucleus, and very few are exclusively cytoplasmic. Only one of them is localized in the chloroplast of Arabidopsis, where it participates in splicing regulation of genes encoding rRNAs in this organelle.

Subcellular localization and the number of ZCCHC motifs are often conserved in orthologous proteins. In some cases, these motifs determine subcellular localization, translocation, or binding to other proteins or nucleic acids. We constructed a phylogenetic tree with the 198 ZCCHC motifs identified in seven yeast proteins, 69 from Arabidopsis, and 34 human. Our analysis revealed that the ZCCHC motifs of most putative orthologs group into clades, indicating higher conservation of amino acids beyond the characteristic three cysteines (C) and one histidine (H) in this motif. The conservation of these amino acids suggests their importance for the activity of the ancestral protein from which these orthologs are derived, as well as for their own function.

In this Thesis, we studied the Arabidopsis *CXIP4* gene, which was already annotated as a distant ortholog of *Human SREK1 Interacting Protein 1* (*SREK1IP1*), which is presumed to be related to splicing. SREK1IP1 and CXIP4 share only their single ZCCHC motif, which clearly cluster in our previously mentioned phylogenetic tree, suggesting its functional significance. No similarity was observed in the rest of the sequences of these two proteins, although both have regions of low complexity in their carboxy-terminal half. While putative orthologs of *CXIP4* and *SREK1IP1* exist in many plants and animals, no mutant alleles have been described.

We characterized the insertional mutants *cxip4-1* and *cxip4-2*, carrying a null and a hypomorphic allele of *CXIP4*, respectively. The early post-embryonic lethality of *cxip4-1* indicates that *CXIP4* is essential for Arabidopsis survival. The viability of *cxip4-2* plants allowed us to establish the involvement of *CXIP4* in various development pathways, as this mutant exhibits a highly pleiotropic phenotype throughout its life cycle. The pleiotropy of the morphological phenotype of *cxip4-2* corresponds to its transcriptomic profile, with a significant portion of its deregulated genes encoding transcription factors, and some of them may explain several traits of its phenotype.

We analyzed global splicing in *cxip4-2* mutant plants, finding increased alternative splicing, mainly intron retention events, although less pronounced than that described in mutants carrying alleles of genes encoding spliceosome components. We confirmed some such differential intron retention events using the likely null mutant *cxip4-1*, whose splicing alterations are more pronounced. These observation suggests that CXIP4 directly or indirectly

participates in pre-mRNA splicing but does not play a central role in this process. Functional category enrichment analysis of genes undergoing alternative splicing and behaving differently in *cxip4-2* compared to the wild type suggests that CXIP4 controls alternative splicing of specific gene categories. Our results also suggest that CXIP4 is related to epigenetic repression via non-coding RNA-guided methylation. The nuclear accumulation of polyadenylated RNA that we observed in *cxip4-2* plants is consistent with the involvement of CXIP4 in splicing and reinforces the hypothesis of functional conservation between human SREK1IP1 and Arabidopsis CXIP4.

.

IV.- INTRODUCCIÓN

IV.- INTRODUCCIÓN

IV.1.- Componentes del transcriptoma eucariótico

Los productos de la transcripción de los genes eucarióticos son de dos tipos: ARN mensajeros (ARNm) y ARN no codificantes, que codifican o no proteínas, respectivamente (revisado en Cech y Steitz, 2014; Ariel *et al.*, 2015). La mayor parte del transcriptoma de los eucariotas es no codificante (Bhat *et al.*, 2020). En la especie humana, por ejemplo, los ARN no codificantes representan el 90% del transcriptoma, y las regiones codificantes de los genes ocupan solo un 2% del genoma (Doolittle, 2013).

Dos tipos de ARN no codificantes participan en la traducción: los transferentes (ARNt), que descodifican las secuencias de los ARNm, y los ribosómicos (ARNr), que son componentes estructurales de los ribosomas (Bhogireddy *et al.*, 2021). Otros ARN no codificantes tienen funciones reguladoras y se clasifican según su longitud y mecanismo de biogénesis (Bhat *et al.*, 2020). Los ARN largos no codificantes (lncRNA), con más de 200 nucleótidos e implicados en la regulación transcripcional de la expresión génica, aunque muchos de ellos parecen no ser funcionales (revisado en Yang *et al.*, 2022). Los ARN pequeños no codificantes son muy heterogéneos; pertenecen a este grupo los microARN (miARN) y los pequeños ARN interferentes (siARN), que participan en la regulación postranscripcional y epitranscriptómica de la expresión génica, respectivamente. Los miARN son los ARN no codificantes mejor estudiados, por su gran importancia en la regulación postranscripcional de la expresión de muchos genes que codifican proteínas (revisado de Ariel *et al.*, 2015). Una prueba de su importancia es la reciente concesión del Premio Nobel en Fisiología o Medicina a sus descubridores Victor Ambros y Gary Ruvkum.

En los eucariotas, las endorribonucleasas de la familia ARGONAUTE (AGO) juegan un papel central en el silenciamiento génico mediado por los siARN y los miARN con los que se asocian, formando complejos que actúan a nivel transcripcional o postranscripcional, respectivamente (revisado en Hutvagner y Simard, 2008; O'Brien *et al.*, 2018). En *Arabidopsis thaliana* (en adelante, Arabidopsis) existen 10 proteínas AGO, siendo AGO1 la más importante en las rutas de silenciamiento génico postranscripcional mediadas por los miARN, que impiden la traducción de sus ARNm dianas (Hunter *et al.*, 2003; Baulcombe, 2004). Otras proteínas AGO se asocian con siARN endógenos o virales para controlar a nivel transcripcional la expresión génica (Carbonell y Carrington, 2015; Carbonell y Daròs, 2017). Los alelos nulos de *AGO1* son letales postembrionarios y los hipomorfos causan un fenotipo pleiotrópico (revisado en Albert *et al.*, 2015; Li *et al.*, 2022).

IV.2.- El metabolismo del ARN

Se denomina metabolismo del ARN al conjunto de los procesos que acontecen desde la síntesis de los transcritos primarios hasta la degradación de los ARN maduros, una vez cumplida su función. En muchas etapas del metabolismo de los ARN se modifican su secuencia primaria, estructura y localización subcelular (revisado en Matsui *et al.*, 2019).

IV.2.1.- Maduración de los pre-ARNm eucarióticos

Se denomina pre-ARNm a los transcritos primarios de los ARNm eucarióticos, que sufren una maduración compleja, fundamentalmente cotranscripcional, para generar un ARNm maduro, que es finalmente exportado al citoplasma, en donde es traducido. La maduración de un pre-ARNm comienza con la adición de una caperuza (cap): una guanosina metilada (m^7G) unida covalentemente a su extremo 5′ (revisado en Ramanathan *et al.*, 2016). Al extremo 3′ del pre-ARNm se agregan unas 200 adenosinas, la denominada cola poli(A). La caperuza y la cola poli(A) son necesarias para la estabilidad del ARNm, ya que lo protegen de las exorribonucleasas, y también contribuyen a su traducibilidad. La cola poli(A) interviene en la exportación de los ARNm del núcleo al citoplasma (revisado en Passmore y Coller, 2022).

IV.2.2.- El *splicing*

La etapa de la maduración de los pre-ARNm mejor estudiada es el *splicing*: la escisión de los intrones y la ligación de los exones adyacentes (revisado en Bentley, 2014), un proceso altamente regulado y complejo que lleva a cabo el espliceosoma, un macrocomplejo riboproteico dinámico, que se ensambla y desensambla formando distintos subcomplejos unidos a cada intrón durante su eliminación (revisado en Rappsilber *et al.*, 2002; Ru *et al.*, 2008; Hegele *et al.*, 2012). La presencia de intrones es común en los genes eucarióticos y posibilita su *splicing* alternativo, combinándose los intrones y exones de diversos modos, rindiendo distintas variantes de ARNm maduro, cuya traducción enriquece el proteoma, fundamentalmente en los animales y las plantas (revisado en Wilkinson *et al.*, 2020).

IV.2.2.1.- Secuencias reguladoras del *splicing*

En el *splicing* de los pre-ARNm ocurren dos reacciones de transesterificación consecutivas, cuya correcta ejecución depende del reconocimiento por el espliceosoma de cuatro regiones con secuencias conservadas. Dos de ellas se sitúan en los límites entre un intrón y sus exones flanqueantes: el 5′SS (sitio 5′ o donante del *splicing*) y el 3′SS (sitio 3′ o aceptor del *splicing*). Los 5′SS y 3′SS contienen invariablemente los dinucleótidos GU y AG en los extremos 5′ y 3′ de cada intrón, respectivamente. Las sustituciones o deleciones de

cualquiera de estos nucleótidos impiden o dificultan el *splicing* del intrón afectado. Los nucleótidos inmediatamente posteriores o anteriores a estos dinucleótidos también son parte de los 5′SS y 3′SS, pero su grado de conservación y requerimiento para un *splicing* correcto son menores (Brown *et al.*, 1996). Los intrones contienen otras dos regiones conservadas: el sitio de ramificación (branch point; BP), con una A invariante que es esencial para el *splicing*, y un tramo de una decena de pirimidinas contiguas que se encuentran aguas arriba del 3′SS, a una distancia variable de 15 a 50 nucleótidos (Tolstrup *et al.*, 1997) (Figura 1).

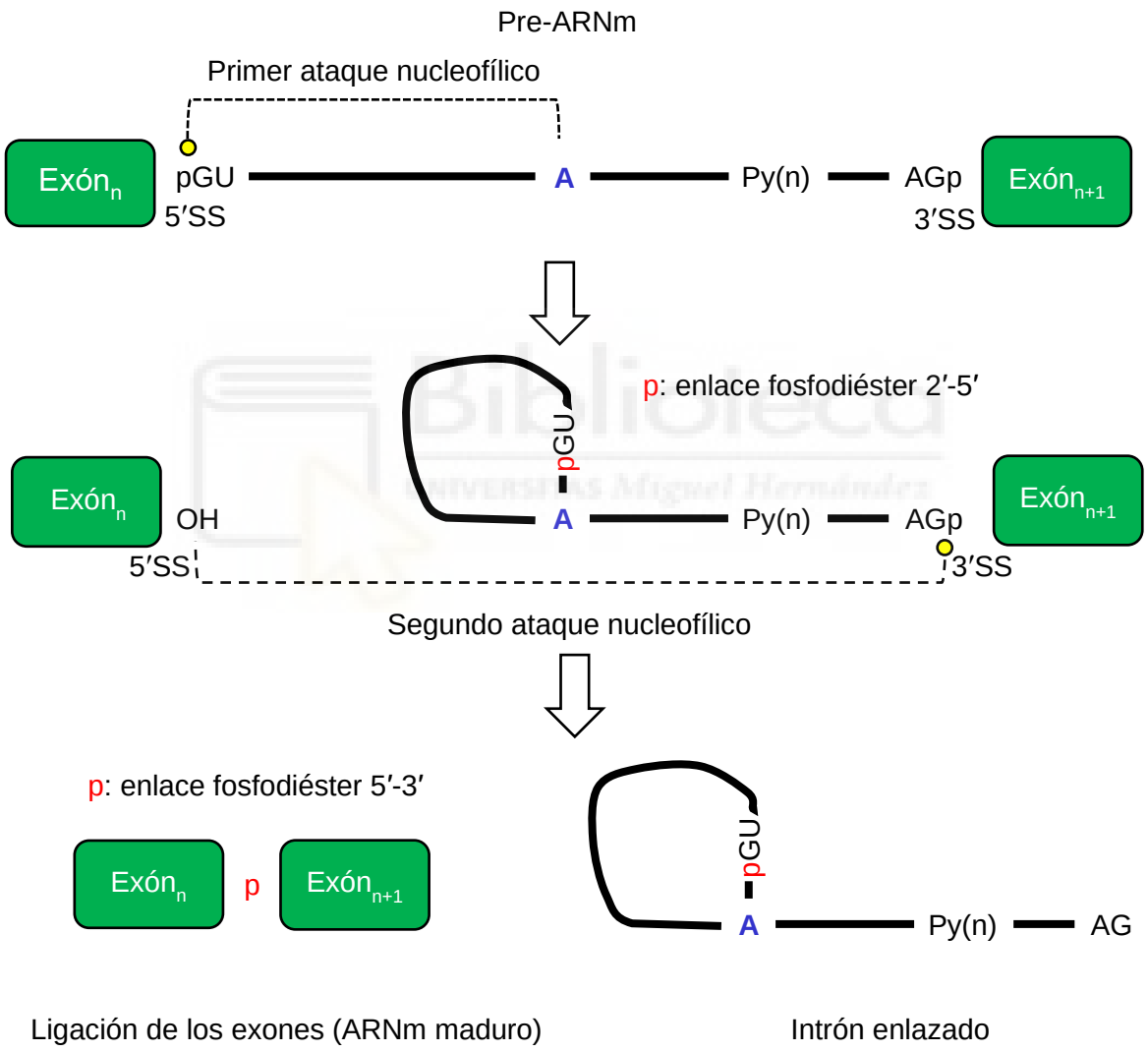


Figura 1.- Mecanismo del *splicing* de un intrón de tipo U2. Esquemas de las reacciones de transesterificación del *splicing* de un intrón de un pre-ARNm. Las líneas negras representan al intrón, y los rectángulos verdes, a sus dos exones adyacentes. Se destacan las secuencias conservadas de los 5′SS y 3′SS, el tramo de pirimidinas [Py(n)] y la adenina (A) del sitio de ramificación, en azul. Los ataques nucleofílicos se representan mediante líneas punteadas, y los grupos fosfato de los enlaces fosfodiéster que se generan, con letras rojas. Modificada a partir de Petrillo (2023).

Los intrones mencionados en el párrafo anterior son del tipo denominado U2, los mayoritarios en los genomas eucarióticos, que son procesados por el espliceosoma mayor. Los intrones minoritarios, denominados U12, difieren de los U2 en sus secuencias conservadas de los 5′SS y 3′SS, que son AU y AC, respectivamente, y son procesados por el espliceosoma menor (Marquez *et al.*, 2012). Arabidopsis solo cuenta con unos 300 intrones de tipo U12 (revisado en Turunen *et al.*, 2013; Ding *et al.*, 2022). En la mayoría de las publicaciones que abordan el *splicing* se hace referencia a los intrones U2 y al espliceosoma mayor, sin adjetivar, que es lo que se hará en lo sucesivo en esta memoria.

IV.2.2.2.- La maquinaria del *splicing*

En la primera reacción de transesterificación del *splicing*, el hidroxilo 2′ de la A conservada del sitio de ramificación ataca nucleofílicamente al grupo fosfato que une al exón anterior (el situado aguas arriba del intrón; $Exón_n$ en la Figura 1, en la página 10) con la G del 5′SS. Este corte endonucleolítico del 5′SS rinde dos productos, uno que contiene el exón anterior con su extremo 3′ libre, y otro con un lazo intrónico unido al extremo 5′ del exón siguiente (el que se encuentra aguas abajo del intrón; $Exón_{n+1}$ en la Figura 1) (revisado en Will y Lührmann, 2011; Shi, 2017; Wilkinson *et al.*, 2020). En la segunda transesterificación, un corte en el 3′SS libera al intrón, y se ligan los exones. Finalmente, el intrón es degradado y se reciclan sus ribonucleótidos (Brown *et al.*, 1996).

El espliceosoma mejor estudiado es el de *Saccharomyces cerevisiae*, a pesar de que, paradójicamente, el 90% de sus genes carece de intrones (Neuvéglise *et al.*, 2011). Son unas 170 las proteínas que forman parte del espliceosoma humano o se asocian con él en alguna etapa el *splicing* (revisado en Wahl *et al.*, 2009; Koncz *et al.*, 2012; Turunen *et al.*, 2013). Son unos 400 los genes de las plantas que codifican proteínas de las que se sabe o sospecha que participan en el *splicing* (Wang y Brendel, 2004; Koncz *et al.*, 2012).

Los principales subcomplejos del espliceosoma son las partículas ribonucleoproteicas pequeñas nucleares (snRNP) U1, U2, U4, U5 y U6, que están compuestas por un ARN pequeño nuclear (small nuclear RNA; snRNA), que es rico en uridinas y da nombre a la partícula. Existe un número variable de hasta 50 proteínas específicas de cada partícula U (revisado en Budak *et al.*, 2020).

Además de las proteínas que forman parte de las partículas U, otras 100 actúan como factores implicados en el *splicing* o en otros procesos relacionados con el metabolismo del ARNm, como la terminación de la transcripción, su exportación del núcleo al citoplasma y su control de calidad (Fabrizio *et al.*, 2009). En el *splicing* también intervienen helicasas de ARN del tipo DEAD-box, que constituyen una gran familia cuyos miembros también participan en

la regulación de la expresión génica y en otros aspectos del metabolismo de los ARN (revisado en Nidumukkala *et al.*, 2019; Xu *et al.*, 2023). Estas helicasas facilitan los reordenamientos y cambios estructurales que sufre el espliceosoma durante el *splicing* (revisado en De Bortoli *et al.*, 2021).

IV.2.2.3.- Secuencias reguladoras del *splicing* alternativo

El *splicing* alternativo es un mecanismo de regulación postranscripcional que se produce en la mayoría de los pre-ARNm eucarióticos que cuentan con más de un intrón (Nilsen y Graveley, 2010). Es consecuencia de la existencia de variaciones en las secuencias de los 5′SS o los 3′SS, o de la presencia de SS alternativos, que no son reconocidos y usados con igual eficacia por el espliceosoma (Petrillo, 2023). El *splicing* alternativo puede generar muchas variantes de ARNm maduros a partir de un único pre-ARNm, como consecuencia de la retención o eliminación parcial o total de intrones o exones. Se denomina exones constitutivos a los que se mantienen en todas las variantes de los ARNm tras el *splicing* alternativo, y alternativos a los que varían.

Además de los 5′SS y 3′SS, existen otras secuencias en los intrones y exones que actúan como elementos potenciadores o silenciadores en el *splicing* alternativo (Figura 2). Estas secuencias son los Exonic Splicing Enhancer (ESE) e ISE (Intronic Splicing Enhancer),

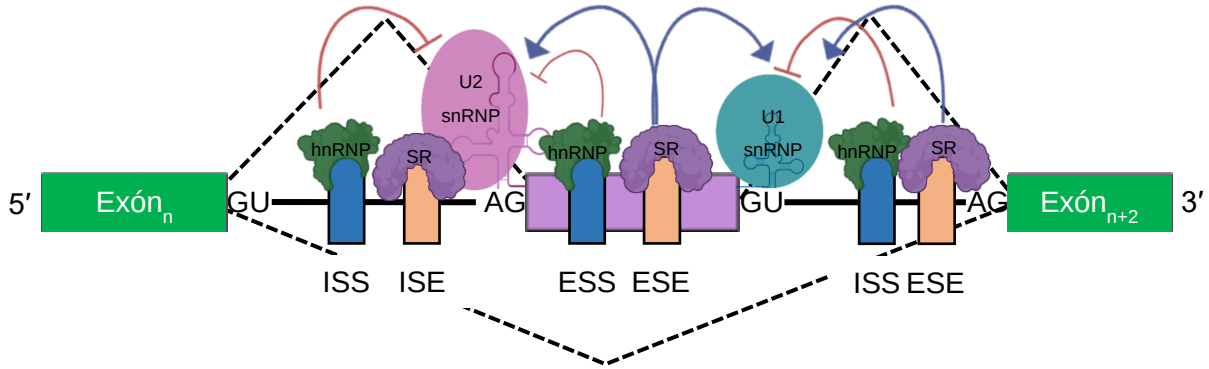


Figura 2.- Regulación del *splicing* alternativo de un pre-ARNm. Los exones e intrones se representan mediante rectángulos verdes (constitutivos) o violeta (alternativo) y líneas negras, respectivamente. Las señales potenciadoras del *splicing* (ISE y ESE) se muestran con rectángulos verticales azules, y las silenciadoras (ISS y ESS), con rectángulos verticales naranjas. Las proteínas SR unidas a las ISE y ESE incrementan la afinidad de las partículas U1 y U2 por los SS, favoreciendo la eliminación de los intrones (líneas con flechas rojas), mientras que las hnRNP juegan un papel antagónico, al unirse a las ESS e ISS (líneas con extremos romos). Las líneas punteadas señalan los posibles eventos de *splicing* alternativo que pueden ocurrir. Modificada a partir de Love *et al.* (2023).

que promueven el *splicing*, y los Exonic Splicing Silencer (ESS) e Intronic Splicing Silencer (ISS), que lo inhiben (revisado en Smith y Valcárcel, 2000). A estas secuencias se unen dos tipos de proteínas que regulan antagónicamente el *splicing* alternativo: las SR y hnRNP. Las proteínas SR son ricas en residuos de serina (S) y arginina (R) y se unen a los ESE e ISE, reclutando al espliceosoma a los SS; son necesarias para el reconocimiento y la selección de los SS (revisado de Zhang *et al.*, 2020). Las proteínas SR presentan uno o dos motivos de reconocimiento de ARN en su región aminoterminal, que median su unión a los pre-ARNm (revisado en Kufel *et al.*, 2022). Algunas proteínas SR contienen un motivo ZCCHC. Por su parte, las hnRNP se unen a los silenciadores del *splicing*, impidiendo la interacción del espliceosoma con los SS (Busch y Hertel, 2012; Jones *et al.*, 2022).

IV.2.2.4.- Tipos de eventos de *splicing* alternativo

Se asume generalmente que a la gran diversidad proteómica de los eucariotas contribuye notablemente el *splicing* alternativo, que explicaría al menos en parte el número relativamente bajo de genes de los genomas de los metazoos y las plantas. El *splicing* alternativo de los eucariotas pluricelulares ocurre de forma diferencial en el tiempo y en el espacio, generando distintas isoformas de proteínas que son específicas de tejido o etapa del desarrollo (revisado en Wilkinson *et al.*, 2020). La plasticidad de las plantas para adaptarse a diferentes condiciones ambientales también se ve favorecida por el *splicing* alternativo, tal como indican diversos análisis transcriptómicos en los que se han identificado ARNm alternativos específicos en respuesta a las condiciones lumínicas, de temperatura, disponibilidad de agua o concentraciones salinas y a patógenos (revisado en Dikaya *et al.*, 2021; Ling *et al.*, 2021; Kashkan *et al.*, 2022; Kufel *et al.*, 2022).

Pese a las numerosas evidencias de la existencia del *splicing* alternativo en distintos tejidos, etapas del desarrollo y condiciones ambientales, se desconoce su relevancia biológica, dada la escasez de análisis funcionales distintos de su análisis bioinformático. Las consecuencias del *splicing* alternativo, en los escasos casos en los que se ha analizado a nivel de sus consecuencias en la generación de isoformas de proteínas, ha revelado que estas últimas pueden actuar independiente o coordinadamente, jugando papeles específicos en distintos tejidos o localizaciones subcelulares. Además, las isoformas truncadas podrían competir con la canónica (la más abundante y completa), comportándose como lo haría el producto de un alelo mutante antimorfo (revisado en Kashkan *et al.*, 2022).

De los cinco tipos de eventos de *splicing* alternativo más frecuentes, la retención de intrones es el mayoritario en las plantas (Marquez *et al.*, 2012), y la eliminación de exones, en los animales (Chaudhary *et al.*, 2019) Según la base de datos Arabidopsis thaliana Reference

Transcript Dataset 2 (AtRTD2; https://ics.hutton.ac.uk/ atRTD/), el 60,4% de los genes multiexónicos de Arabidopsis sufren *splicing* alternativo, siendo de retención de intrones el 40% de los eventos. El reconocimiento y selección de un sitio 5′SS o 3′SS alternativo, presente en un exón, se da en un 8% y 16% de los eventos de *splicing* alternativo en Arabidopsis, respectivamente, siendo menores los casos de eliminación exónica. La exclusión mutua de exones ocurre raramente en las plantas (revisado en Kashkan *et al.*, 2022).

IV.2.2.5.- Análisis mutacional del *splicing*

El aislamiento y análisis de mutantes con defectos generalizados en el *splicing*, fundamentalmente en *Saccharomyces cerevisiae*, ha resultado especialmente útil para la identificación de genes que codifican componentes del espliceosoma y para la comprensión del mecanismo del *splicing* (revisado en Hossain *et al.*, 2016). Los alelos nulos de los genes que codifican componentes centrales del espliceosoma son letales, por lo que para su análisis genético se han obtenido alelos condicionales, fundamentalmente en *Saccharomyces cerevisiae*, o hipomorfos en otras especies (revisado de Plaschka *et al.*, 2019).

El análisis mutacional del *splicing* ha permitido constatar en Arabidopsis que las mutaciones hipomorfas en genes que codifican componentes del espliceosoma, o sus factores asociados, producen alteraciones pleiotrópicas del desarrollo vegetal, que pueden causar letalidad, dependiendo de su grado de insuficiencia de función y de la importancia del papel jugado por el producto del gen mutado. Por ejemplo, la homocigosis del alelo hipomorfo *prp8-6* del gen *PRE-MRNA PROCESSING FACTOR 8 (PRP8*) de Arabidopsis, que codifica un componente central del espliceosoma conservado en todos los eucariotas, provoca un retraso generalizado del desarrollo (Marquardt *et al.*, 2014). Otro ejemplo es el del mutante *sr45-1*, portador de un alelo del gen *SR45* de Arabidopsis, que codifica una proteína SR que participa en la regulación del *splicing* alternativo; este mutante muestra floración tardía, hojas apuntadas y un número variable de pétalos y estambres (Ali *et al.*, 2007). Las hojas apuntadas también son un rasgo del mutante *prp8-7*, en el que los eventos de retención intrónica afectan al 6,7% de los intrones (Sasaki *et al.*, 2015), a diferencia de *prp8-6,* en el que no se han constatado defectos globales en el *splicing* (Marquardt *et al.*, 2014).

La obtención de construcciones con genes testigo que incluyen intrones y SS con secuencias canónicas y mutadas en diferentes posiciones ha permitido identificar nuevos factores implicados el *splicing,* mediante la mutagénesis de individuos portadores de los transgenes y la evaluación de la actividad del gen testigo. En Arabidopsis, por ejemplo, el uso de construcciones de este tipo, que incorporan el gen de la proteína verde fluorescente (GFP),

ha permitido identificar nuevos alelos de genes que codifican componentes del espliceosoma y otros que no se habían relacionado previamente con el *splicing* (Kanno *et al.*, 2020).

Se han descrito mutaciones puntuales que alteran el *splicing* de los pre-ARNm de genes concretos, cuya función no está relacionada con el *splicing*. Algunas de estas mutaciones crean nuevos 5′SS y 3′SS o dañan los canónicos, alterando exclusivamente el patrón del *splicing* del gen mutado. Alrededor del 15% de las patologías humanas con base genética se deben a este tipo de mutaciones (Jiang y Chen, 2021). Además, se han identificado en células tumorales malignas mutaciones en los ESE y ESS de protooncogenes, que causan exclusión exónica y su sobreexpresión (Sterne-Weiler *et al.*, 2011; Mort *et al.*, 2014; Supek *et al.*, 2014).

La búsqueda de supresores extragénicos de los defectos del *splicing* de los pre-ARNm de genes concretos constituye otro abordaje fructífero; se han descubierto así nuevos genes implicados directa o indirectamente en el *splicing*, o nuevos alelos de genes previamente descritos en Arabidopsis (Kanno *et al.*, 2020).

IV.3.- Las proteínas con dedos de zinc

IV.3.1.- Estructura de los dedos de zinc

Los dedos de zinc son motivos estructurales de proteínas procarióticas y eucarióticas que incluyen residuos de cisteína (C), histidina (H) y, en menor medida, ácido aspártico (D), dispuestos en posiciones conservadas, que coordinan uno o más iones de zinc. Se conocen dedos de zinc que interaccionan con el ADN, el ARN, las proteínas, los lípidos y otras moléculas pequeñas. Estos motivos suelen agruparse de modos tales que dotan a la proteína a la que pertenecen de especificidad de unión a ácidos nucleicos o proteínas (revisado en Malgieri *et al.*, 2015).

Se han definido unos 30 tipos de dedos de zinc, dependiendo del número y distribución de sus C y H y de su estructura espacial. El dedo de zinc clásico es el C_2H_2 (o CCHH), el más abundante en todos los seres vivos. Su secuencia consenso es $X_2CX_{2,4}CX_{12}HX_{3,4,5}H$ y se une a ADN y ARN. El segundo tipo en abundancia es el C_3H (CCCH), con la secuencia consenso $CX_{4-17}CX_{4-6}CX_3H$, que forma parte, entre otras, de proteínas que regulan la estabilidad de los ARNm. El tercero es el C_2HC (CCHC), cuya secuencia consenso es $CX_2CX_4HX_4C$ y se encuentra en proteínas que participan en diversos procesos del metabolismo de los ARN (revisado en Armas y Calcaterra, 2012). Existen nueve tipos de dedos de zinc en las plantas: C_3H, C_4, C_6, C_8, C_2H_2, C_2HC, C_2H_5, C_3H_4 y C_3HC_4 (Takatsuji, 1998; Krishna *et al.*, 2003).

IV.3.2.- Las proteínas con motivos ZCCHC

Según se describe en la base de datos InterPro (http://www.ebi.ac.uk/interpro), el motivo ZCCHC (IPR025836) es una secuencia de 18 aminoácidos, 14 de los cuales son centrales, con la secuencia consenso $CX_2CX_4HX_4C$, y dos adicionales en cada extremo. En los motivos ZCCHC también están muy conservadas dos glicinas (G), situadas entre la segunda cisteína y la histidina. El motivo ZCCHC forma dos láminas beta conectadas por una hélice alfa, conformada por los cuatro aminoácidos que se encuentran entre la segunda C y la H, formándose así un dedo de zinc mucho más corto que el clásico de unión a ADN. El dedo de zinc del tipo ZCCHC recibe también el nombre de “zinc knuckle” o “Gag knuckle”, porque se descubrió en las proteínas Gag de retrovirus que están implicadas en el empaquetamiento del ARN viral (Armas y Calcaterra, 2012; Olson y Musier-Forsyth, 2019). A diferencia de otros tipos de dedos de zinc, los motivos ZCCHC se encuentran en una sola copia en muchas proteínas.

Cuando se inició esta Tesis se desconocía el número de genes eucarióticos que codifican proteínas con motivos ZCCHC, ya que no existía ningún registro unificado. Solo la base de datos HUGO Gene Nomenclature Committee (HGNC; http://www.genenames.org) reunía 25 proteínas humanas con motivos ZCCHC, que se denotaban con el término ZCCHC seguido de un número. La caracterización funcional de 16 de estas 25 proteínas humanas indicó su participación en el metabolismo de los ARNm y los ARN no codificantes (Figura 3, en la página 17; revisado en Armas y Calcaterra, 2012; Tweedie *et al.*, 2020).

IV.4.- Antecedentes y objetivos

IV.4.1.- Identificación de *MAS2*

En el laboratorio del profesor José Luis Micol se obtuvo una colección de mutantes foliares tras una mutagénesis con metanosulfonato de etilo (EMS) de la estirpe Landsberg *erecta* (L*er*) (Berná *et al.*, 1999). Algunos de ellos presentaban hojas recurvadas hacia el haz (hiponásticas), a diferencia de las silvestres, que son casi totalmente planas, y se denominaron *incurvata* (*icu*). La clonación posicional en el laboratorio de M.R. Ponce de cinco de los mutantes *icu* reveló que eran portadores de nuevos alelos hipomorfos o nulos de los genes *HASTY* (*HST*; Telfer y Poethig, 1998), *HYPONASTIC LEAVES1* (*HYL1*; Lu y Fedoroff, 2000), *ARGONAUTE1* (*AGO1*; Bohmert *et al.*, 1998) y *HUA ENHANCER1* (*HEN1*; Chen *et al.*, 2002). Todos estos genes codifican componentes de la maquinaria de silenciamiento génico postranscripcional mediada por miARN (Jover-Gil *et al.*, 2005; Jover-Gil *et al.*, 2012).

Uno de los mutantes aislados en dicha mutagénesis fue inicialmente denominado *icu9-2*. Era viable y relativamente fértil y presentaba un fenotipo morfológico fácilmente reconocible

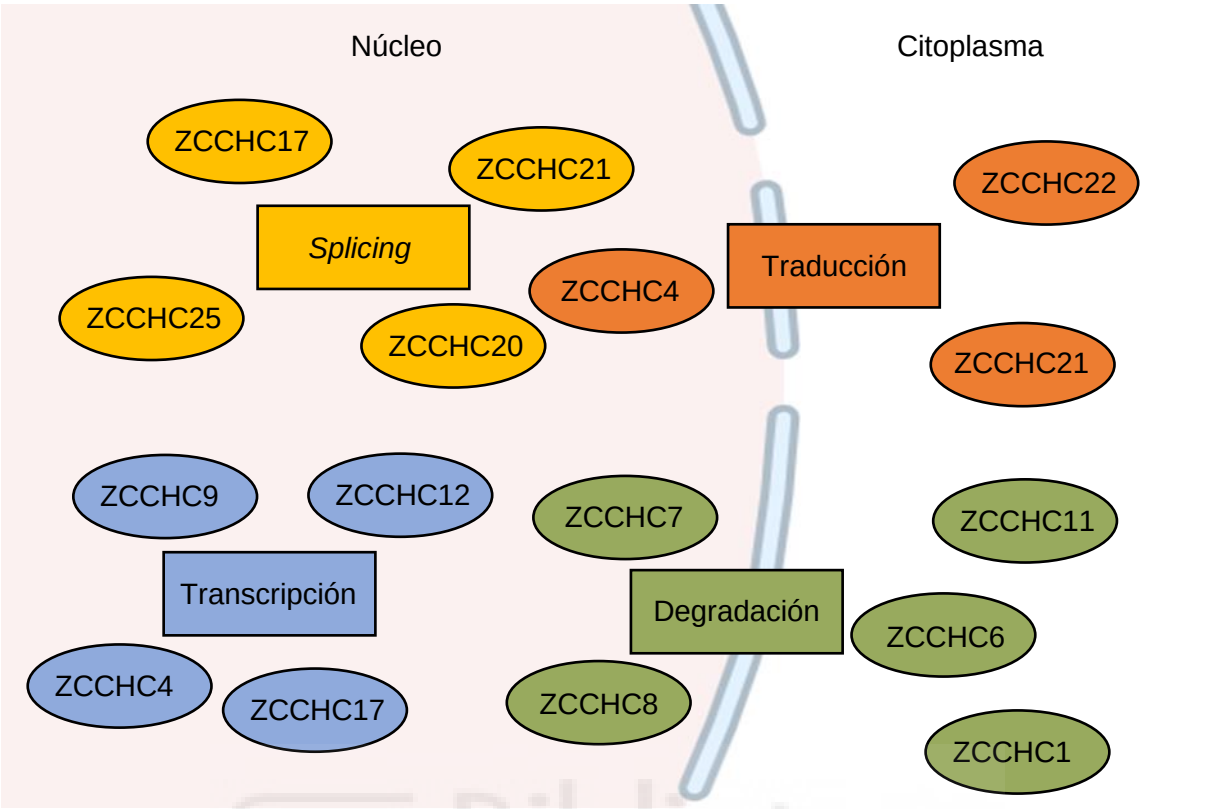


Figura 3.- Esquema de las localizaciones subcelulares de las proteínas humanas con motivos ZCCHC descritas en Wang *et al.* (2021) y sus funciones en el metabolismo del ARN.

a simple vista. Se estableció que era portador de un alelo del gen *AGO1*, razón por la que su denominación pasó a ser *ago1-52*. Resultó ser portador de una transición G → A en el penúltimo intrón de *AGO1*, que genera un nuevo 3′SS que el espliceosoma reconoce mejor y usa más frecuentemente que el genuino, que está inalterado. El *splicing* del pre-ARNm de *ago1-52* produce mayoritariamente un ARNm mutante que incorpora 10 nucleótidos de su penúltimo intrón, que desfasan su pauta de lectura y se traduce a una proteína truncada (Jover-Gil *et al.*, 2012; Sánchez-García *et al.*, 2015).

Para encontrar nuevos genes implicados en el metabolismo del ARN y relacionados con *AGO1*, se realizó en el laboratorio de M.R. Ponce una búsqueda de supresores extragénicos del fenotipo morfológico de *ago1-52*, tras una mutagénesis con EMS. Se identificaron así varios genes cuyas mutaciones dominantes suprimían la alteración del *splicing* de *ago1-52*, que se denominaron *MORPHOLOGY OF ARGONAUTE1-52 SUPPRESSED* (*MAS*).

La clonación posicional de *MAS2* reveló que era el ortólogo más plausible del gen que codifica la NF-kappa B Activating Protein (NKAP) humana (Sánchez-García *et al.*, 2015), una proteína multifuncional que ha sido implicada posteriormente en la ligación de exones (Fica *et*

al., 2019), en la selección de 3′SS alternativos con secuencias NAGNAG (Dybkov *et al.*, 2023) y en la lectura de las marcas epitranscriptómicas m^6A durante la maduración de los pri-miARN y pre-ARNm (Zhang *et al.*, 2019; Sun *et al.*, 2022b).

La clonación posicional de *MAS5* reveló que codificaba PRP8, el factor central del espliceosoma (apartado IV.2.2.5, en la página 14; Cabezas-Fuster *et al.*, 2023). Los alelos *mas2* y *mas5* identificados en el laboratorio de M.R. Ponce suprimían parcialmente los defectos del *splicing* de *ago1-52* de una manera singular, ya que propiciaban el uso del 3′SS genuino frente al alternativo (el creado por la mutación de la que es portador *ago1-52*). Estos alelos *mas2* y *mas5*, de hecho, incrementan la fidelidad del *splicing*.

IV.4.2.- Interacciones de MAS2 en ensayos del doble híbrido de la levadura

La caracterización molecular de MAS2 incluyó un ensayo del doble híbrido de la levadura, en el que se identificaron 13 presuntos interactores (Sánchez-García *et al.*, 2015), solo tres de los cuales habían sido caracterizados previamente a algún nivel (Tabla 1). Uno de ellos fue REGULATOR OF CBF GENE EXPRESSION 1 (RCF1), del que se sabía que codifica una helicasa de tipo DEAD-box conservada desde las levaduras hasta los mamíferos (Xu *et al.*, 2004), que estaba aparentemente implicada en el *splicing* de los pre-ARNm en

Tabla 1.- Interactores de MAS2 identificados en un ensayo del doble híbrido de la levadura

Interactor (código AGI[a])	Función descrita	Referencias
CXIP4 (AT2G28910)	Activador del antiportador CAX1	Cheng *et al.*, 2004
PPC2 (AT2G42600)	Metabolismo del C y N	Shi *et al.*, 2015
NF-YC10 (AT1G07980)	Regulador de la termotolerancia	Sato *et al.*, 2014
RCF1 (AT1G20920)	*Splicing* y biogénesis de los miARN	Guan *et al.*, 2013; Xu *et al.*, 2023
HIGLE (AT2G30350)	Biogénesis de los miARN	Verma *et al.*, 2022
RRP7 (AT5G38720)	Biogénesis del ribosoma	Micol-Ponce *et al.*, 2018
SMO4 (AT2G40430)	Biogénesis del ribosoma	Micol-Ponce *et al.*, 2020
RPS24B (AT5G28060)	Biogénesis del ribosoma	Cabezas-Fuster *et al.*, 2023
CWC25 (AT2G44200)	Presunto factor del *splicing*	--
SWAP (AT5G55100)	Presunto factor del *splicing*	--
NET1C (AT4G02710)	Desconocida	--
NPY6 (AT5G47800)	Desconocida	--
AT4G33690[b]	Desconocida	--

Todas las abreviaturas se definen en el texto de esta memoria, con las excepciones que siguen, que se han tomado literalmente del TAIR. [a]Arabidopsis Genome Initiative. [b]G patch domain protein. PPC2: PHOSPHOENOLPYRUVATE CARBOXYLASE 2. HIGLE: HYL1 INTERACTING GIY-YIG LIKE ENDONUCLEASE. CWC25: pre-mRNA splicing factor domain-containing protein. SWAP: Suppressor-of-White-APricot)/surp domain-containing protein. NET1C: NETWORKED 1C. NPY6: gene homologous to the NPY family based on deep phylogeny.

condiciones de estrés por frío (Guan *et al.*, 2013). Se ha confirmado recientemente que RCF1 participa en el *splicing* y que además está implicado en la biogénesis de los miARN (Xu *et al.*, 2023). Otro interactor fue NUCLEAR FACTOR Y, SUBUNIT C10 (NF-YC10), también denominado DNA POLYMERASE II SUBUNIT B3-1 (DPB3-1), que regula positivamente la termotolerancia (Sato *et al.*, 2014; Sato *et al.*, 2016). El tercer interactor al respecto del cual existía información previa era CAX-INTERACTING PROTEIN 4 (CXIP4), presuntamente implicado en el transporte de calcio (Cheng *et al.*, 2004), al que se dedica el apartado siguiente. Aunque se desconocía la función de los restantes, se dedujo de la búsqueda de homologías en las bases de datos que al menos dos podían estar relacionados con el *splicing,* y al menos tres, con la biogénesis del ribosoma (Sánchez-García *et al.*, 2015).

Se han caracterizado ocho de los 13 interactores indicados en la Tabla 1, tres de ellos en el laboratorio de M.R. Ponce, confirmándose su relación con la biogénesis del ribosoma: la RIBOSOMAL RNA PROCESSING 7 (RRP7), la NUCLEOLAR PROTEIN 53 (NOP53) y la RIBOSOMAL PROTEIN SMALL 24B (RPS24B). Estas tres proteínas participan en la maduración de los ARNr; RRP7 y RPS24B (así como su paráloga RPS24A), en la del ARNr 18S, y NOP53, en la del ARNr 5,8S (Micol-Ponce *et al.*, 2018; Micol-Ponce *et al.*, 2020; Cabezas-Fuster *et al.*, 2023). NOP53 fue denominado SMALL ORGAN 4 (SMO4) en Zhang *et al.* (2015), por los defectos en la proliferación celular causados por sus alelos mutantes, razón por la que se usó también esta denominación en Micol-Ponce *et al.* (2020).

IV.4.3.- CXIP4 es un interactor de MAS2 que contiene un motivo ZCCHC

En esta Tesis hemos estudiado la CAX-INTERACTING PROTEIN 4 (CXIP4) de Arabidopsis, cuya denominación se debe a su capacidad de activar un antiportador de H^+/Ca^{2+} en la vacuola de *Saccharomyces cerevisiae*, demostrada en un experimento de expresión heteróloga (Cheng *et al.*, 2004); CAX es la abreviatura de CALCIUM EXCHANGER 1. CXIP4 también fue identificado posteriormente en una búsqueda de reguladores de la respuesta a las proteínas mal plegadas (Unfolded Protein Response; UPR) de Arabidopsis, dado que las plantas transgénicas en las que se sobreexpresa *CXIP4* son menos sensibles que las silvestres al tratamiento con ditiotreitol y tunicamicina. Estos compuestos inducen la acumulación de proteínas mal plegadas que activan la UPR en el retículo endoplásmico (Hossain *et al.*, 2016).

CXIP4 fue el presunto interactor de MAS2 más representado en el ensayo del doble híbrido de la levadura comentado en el apartado anterior, ya que su ARNm fue identificado en 23 de los 55 clones que resultaron positivos (Sánchez-García *et al.*, 2015). Sin embargo, su presunta función anotada en las bases de datos se había deducido del análisis de su

expresión heteróloga en células de la levadura (Cheng *et al.*, 2004) y más tarde mediante su sobreexpresión en plantas transgénicas de Arabidopsis (Hossain *et al.*, 2016), abordajes experimentales que pueden generar artefactos. Tampoco se disponía de análisis funcionales de ninguno de sus ortólogos vegetales o de sus presuntos ortólogos animales, como SREK1 interacting protein 1 (SREK1IP1; SREK1 es la abreviatura de Splicing Regulatory Glutamic Acid And Lysine Rich Protein 1), cuyo dominio ZCCHC es muy similar al de CXIP4 (Figura 4).

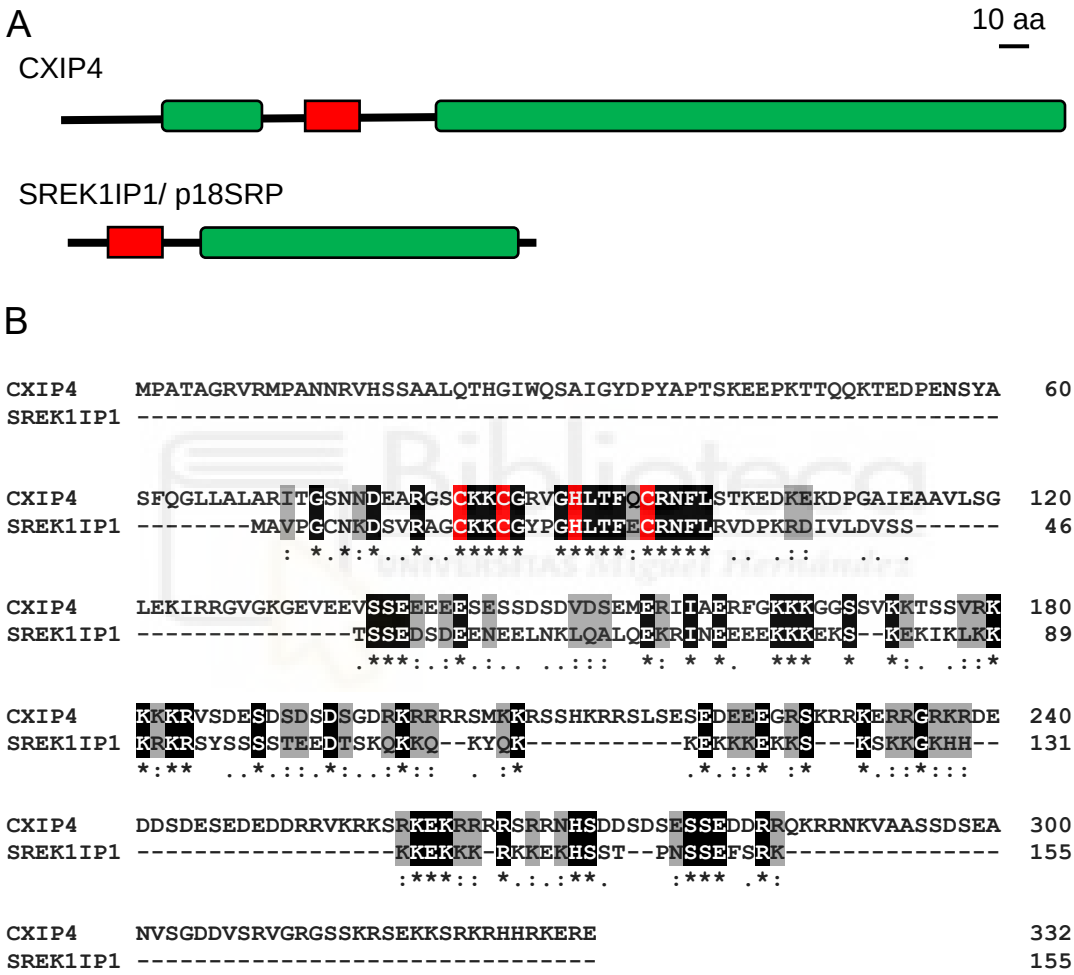


Figura 4.- Homología entre la CXIP4 de Arabidopsis y la SREK1IP1 humana. (A) Representación esquemática de estas proteínas, basada en la información de UniprotKB. Se representan en verde las regiones de baja complejidad, y en rojo, el motivo ZCCHC. (B) Alineamiento múltiple en el que se sombrean en rojo los aminoácidos conservados del motivo ZCCHC. El alineamiento se realizó con Clustal Omega (Sievers y Higgins, 2018) y se sombreó con Boxshade (https://junli.netlify.app/apps/boxshade/). Se indican los aminoácidos idénticos en negro, y los similares, en gris.

La SREK1IP1 humana fue identificada por primera vez en un ensayo del doble híbrido de la levadura, en el que interaccionó con SREK1. Esta última es una proteína SR de la que no se sabe mucho, salvo que interacciona con otras proteínas SR para la regulación del uso

de los 5′SS o 3′SS durante el *splicing* (Heese *et al.*, 2004). La SREK1IP1 humana y la CXIP4 de Arabidopsis podrían presentar funciones conservadas, ello a pesar de que su semejanza se limita casi exclusivamente al motivo ZCCHC y a las regiones de baja complejidad (Figura 4, en la página 20). Es particularmente digno de mención que en un ensayo de copurificación (Huttlin *et al.*, 2017) la NKAP humana interaccionó con SREK1IP1, que en la base de datos PANTHER está anotada como ortóloga lejana de la CXIP4 de Arabidopsis.

Aunque la CXIP4 está presente en todas las plantas, son muy pocas las publicaciones que la mencionan y ninguna en la que se hayan descrito alelos mutantes de los genes que la codifican. El gen *CXIP4* del arándano (*Vaccinium corymbosum*) se sobreexpresa tras un tratamiento con cadmio (Chen *et al.*, 2019), y el del trigo (*Triticum estivum*) se ha relacionado con la respuesta al tizón de la espiga, enfermedad causada por *Fusarium graminearum*, un hongo que genera grandes pérdidas a nivel mundial (Chen *et al.*, 2022). Estas observaciones sustentan las propuestas de que CXIP4 regula los niveles de cadmio y participa en la respuesta de defensa a *Fusarium graminearum*, respectivamente.

La localización subcelular de la CXIP4 de Arabidopsis se ha determinado en ensayos de expresión heteróloga con fusiones traduccionales del gen de la CXIP4 y los de la GFP o la CFP. Esta proteína se localiza en un patrón nuclear difuso en células BY-2 de *Nicotiana tabacum*, pero también en el citoplasma de la levadura, formando puntos discretos que no se corresponden con las mitocondrias (Cheng *et al.*, 2004). El transgén *35S::TaCXIP4:CFP* fue transferido a hojas de *Nicotiana benthamiana*, en las que se observó que se expresaba en el núcleo con un patrón punteado (Chen *et al.*, 2022).

IV.4.4.- Objetivos de esta Tesis

El objetivo principal de esta Tesis fue la caracterización funcional del gen *CXIP4,* mediante el análisis de los fenotipos morfológico y molecular causados por insuficiencia de función. Pretendíamos esclarecer su función en Arabidopsis, pero también contribuir a una mejor comprensión de la de *MAS2* y a la vez confirmar la relación funcional entre estos dos genes, así como su relación con el *splicing*.

Los análisis previos de la estructura de CXIP4 y su homología con otras proteínas revelaban la presencia de un motivo ZCCHC, su única región aparentemente conservada con su ortóloga animal más verosímil, la SREK1IP1 humana. Dado que la información disponible sobre los motivos ZCCHC era escasa, nos propusimos en primer lugar identificar y analizar las proteínas que los contenían, codificadas en los genomas humano, de Arabidopsis y *Saccharomyces cerevisiae*. Consideramos a estas tres especies representativas de los metazoos, las plantas y los hongos, respectivamente, y que en esa medida podían

proporcionarnos una visión panorámica de la filogenia de los motivos ZCCHC.

Nos propusimos en concreto: (1) Identificar las proteínas con motivos ZCCHC humanas, de Arabidopsis y de *Saccharomyces cerevisiae*, y realizar un análisis comparativo exhaustivo, (2) analizar los efectos fenotípicos de alelos insercionales de *CXIP4*, (3) establecer en qué procesos estaba implicado CXIP4, y (4) obtener el perfil transcriptómico del mutante hipomorfo *cxip4-2* y analizar sus defectos globales en el *splicing*.

V.- MATERIALES Y MÉTODOS

V.- MATERIALES Y MÉTODOS

Para la redacción de los apartados I a VII de esta memoria se han seguido las mismas pautas que en Tesis anteriores de los laboratorios de M.R. Ponce y J.L. Micol. En este apartado de Materiales y métodos se reproducen literalmente algunas frases procedentes de dichas Tesis. Se ha preferido usar los acrónimos castellanizados ADN y ARN —de uso común en los medios de comunicación españoles—, en lugar de los recomendados por la International Union of Pure and Applied Chemistry, DNA y RNA, para los ácidos desoxirribonucleico y ribonucleico, respectivamente. Esta elección no está basada en ningún argumento que se considere incontestable; ambas opciones son aceptadas por el *Diccionario de la Lengua Española* (vigesimotercera edición, 2014) de la Real Academia Española (RAE). Tal como recomienda la RAE en su *Ortografía de la lengua española* (2010), en esta memoria no se realiza el plural de las siglas añadiendo al final una s minúscula: se escribe “el ARN” y también “los ARN”.

La nomenclatura que se aplica en esta memoria a genes, mutaciones y fenotipos nuevos se atiene a las pautas propuestas para Arabidopsis por Meinke y Koornneef (1997). No hemos traducido al español muchos de los nombres de genes y proteínas que se mencionan en esta memoria; en estos casos solo hemos usado la cursiva para los genes. Los transgenes se denotan según lo establecido en las instrucciones a los autores de la revista *Plant Cell*. Los genotipos completos, como *cxip4-1/cxip4-2*, en los que los alelos de un gen en cromosomas homólogos se separan con una barra, se han utilizado únicamente cuando fue imprescindible. Salvo que se indique lo contrario, los individuos que se describen en este trabajo son homocigóticos para la mutación que se menciona en cada caso. Hemos utilizado en algunos casos un punto y coma como separador entre mutaciones no alélicas.

Las estirpes de Arabidopsis, su manipulación y las condiciones de cultivo usadas en esta Tesis se describen en la página 93. Hemos realizado análisis morfométricos, histoquímicos y de microscopía confocal de los mutantes a estudio (páginas 94 y 95). Hemos aislado ARN para su retrotranscripción seguida de RT-PCR y para su secuenciación masiva (páginas 93 y 94). Hemos construido fusiones transcripcionales y traduccionales con el gen de la proteína fluorescente verde (página 94). Hemos realizados alineamientos de secuencias aminoacídicas, por parejas y múltiples, así como otros análisis bioinformáticos y búsquedas de información *in silico* (página 95).

VI.- RESULTADOS Y DISCUSIÓN

VI.- RESULTADOS Y DISCUSIÓN

VI.1.- Identificación y análisis de proteínas con motivos ZCCHC en especies representativas de los hongos, los metazoos y las plantas

Cuando se inició esta tesis no se disponía de información anotada en bases de datos sobre el número de genes que codifican proteínas con motivos ZCCHC en ningún genoma, excepto el humano, en el que se habían descrito 25 de ellas (apartado IV.3.2, en la página 16). Tampoco existía ninguna revisión bibliográfica al respecto, salvo la de Armas y Calcaterra (2012), que al tratarse de un capítulo de libro no estaba indexada.

Nos propusimos inicialmente elaborar una revisión bibliográfica convencional sobre las proteínas con motivos ZCCHC de Arabidopsis. Se evidenció rápidamente la dificultad de encontrar dicha información usando como criterios de búsqueda “ZCCHC” o “zinc knuckle”, ya que se omitían en muchas de las publicaciones relevantes. También resultó evidente la utilidad de ampliar el espectro de nuestras búsquedas a los reinos animal, vegetal y de los hongos, usando bases de datos tanto específicas como generalistas y las herramientas bioinformáticas que conviniesen en cada caso. Nuestra revisión, en consecuencia, acabó siendo un análisis exhaustivo y sistemático de las proteínas con dedos de zinc del tipo CCHC (ZCCHC) en *Homo sapiens*, *Arabidopsis thaliana* y *Saccharomyces cerevisiae*, cuyas etapas se resumen en la Figura 5, en la página 25,y se detallan a continuación.

Realizamos nuestras primeras búsquedas en UniprotKB, usando los términos “CCHC”, “ZCCHC” o “zinc knuckle”, con los que obtuvimos registros redundantes, que incluían las mismas proteínas identificadas en distintas estirpes de la levadura, y muchos otros que solo contenían secuencias parciales de las proteínas potencialmente interesantes. También identificamos proteínas anotadas como portadoras de zinc knuckles que no eran motivos ZCCHC. Optamos entonces por una segunda estrategia, usando como criterio de búsqueda la secuencia $CX_2CX_4HX_4C$, y como herramientas, los programas PatMatch para los genomas de *Saccharomyces cerevisiae* y *Arabidopsis thaliana*, y ScanProsite para el humano (páginas 41 y 42). PatMatch se diseñó para la búsqueda de secuencias cortas, de tres a 30 nucleótidos o aminoácidos, inicialmente para *Saccharomyces cerevisiae* y más tarde para Arabidopsis (Chang *et al.*, 2003; Yan *et al.*, 2005); este programa permite identificar secuencias idénticas a la que se usa como criterio de búsqueda, pero también con desapareamientos (inserciones, deleciones o sustituciones) o degeneradas. ScanProsite (de Castro *et al.*, 2006) es un programa más limitado que PatMatch ya que solo permite identificar motivos de proteínas recogidos en la colección ScanProsite, o un presunto motivo definido por el usuario (https://prosite.expasy.org/scanprodite/).

Identificamos 11 genes en el genoma de la levadura, 135 en el de Arabidopsis y 68 en el humano (Supporting Datasets 1a, 1c y 1e de Aceituno-Valenzuela *et al.*, 2020), cuyas secuencias se tomaron de UniprotKB. Su análisis reveló que en muchas de las proteínas identificadas la secuencia consenso $CX_2CX_4HX_4C$ era parte de un motivo más largo, razón por la que las descartamos. También establecimos de este modo las posiciones y el número de los presuntos motivos ZCCHC en cada una de dichas proteínas.

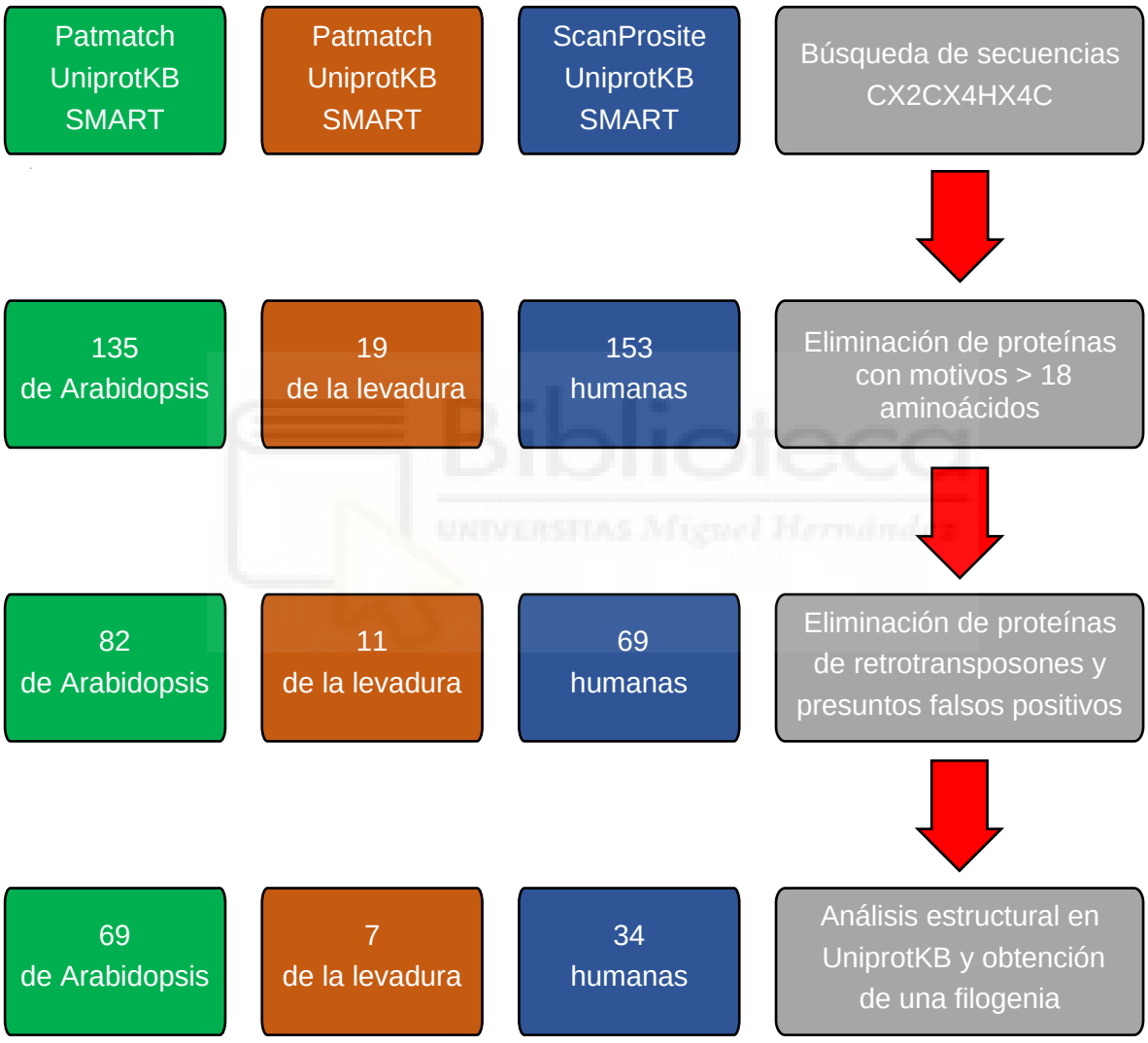


Figura 5.- Etapas en el análisis de proteínas con motivo ZCCHC humanas, de Arabidopsis y la levadura. Los rectángulos grises representan los filtros que se aplicaron para la identificación de los motivos ZCCHC. Los rectángulos verdes, marrones y azules indican los programas usados y las proteínas identificadas en las búsquedas realizadas.

Verificamos a continuación las secuencias de las proteínas predichas en las bases de datos PROSITE y Simple Modular Architecture Research Tool (SMART) o anotadas como

motivos ZCCHC en las bases de datos y descartamos 12 que no cumplían estos criterios: dos proteínas de la levadura, seis de Arabidopsis y cuatro humanas. Obtuvimos las anotaciones de las 119 proteínas seleccionadas y descartamos otras nueve de Arabidopsis que correspondían a retrotransposones. Seleccionamos finalmente 110 proteínas: siete de la levadura, 69 de Arabidopsis y 34 humanas (Supporting Datasets 1b, 1d y 1f de Aceituno-Valenzuela *et al.*, 2020). Identificamos de este modo 24 de las 25 proteínas ZCCHC humanas que estaban anotadas en la base de datos HUGO, excepto ZCCHC23, ya que su motivo ZCCHC resultó ser degenerado, con una asparagina (N) en lugar de la histidina (H) conservada, y que por tanto resultaba indetectable con ScanProsite al usar la secuencia canónica $CX_2CX_4HX_4C$ como criterio de búsqueda.

Localizamos los 198 motivos en las secuencias de las 110 proteínas y concluimos que la mayoría de ellas contienen un solo motivo ZCCHC, pero que algunas incluyen hasta ocho, que no siempre son idénticos. Completamos la información sobre estas 110 proteínas añadiendo la de los motivos, dominios y regiones de baja complejidad anotados en UniprotKB. Concluimos también que el motivo ZCHCC puede combinarse con otros dominios y motivos en una proteína dada o presentarse en solitario (páginas 43 a 45). También, que el 26% de las proteínas analizadas presentan regiones de baja complejidad, que probablemente confieren una mayor versatilidad en sus interacciones con ácidos nucleicos u otras proteínas, lo que les permitiría participar en procesos biológicos dispares (revisado en Coletta *et al.*, 2010). Estas observaciones sugieren que muchas proteínas con motivos ZCCHC son multifuncionales.

Establecimos también la localización subcelular de las 110 proteínas estudiadas, ateniéndonos a la información de las bases de datos UniprotKB, SGD de la levadura, TAIR de Arabidopsis y Human Protein Reference Database (HPRD). Concluimos que la mayoría de las que habían sido estudiadas a algún nivel eran nucleares, tal como cabía esperar de su implicación presunta o demostrada en el metabolismo del ARN.

VI.1.1.- Análisis de la semejanza entre las proteínas con motivos ZCCHC

Hemos establecido que el número de motivos ZCCHC de proteínas presuntamente ortólogas esta usualmente conservado (páginas 43 a 45). Hemos deducido una secuencia consenso genérica para los 198 motivos ZCCHC de las 110 proteínas que hemos estudiado, que coincide con la previamente determinada por Armas y Calcaterra (2012) y sugiere que la secuencia ancestral del motivo ZCCHC es $CX_2CX_3GHX_4C$, ya que la G es adyacente a la H en el 81% de las secuencias. Hemos construido una tabla de frecuencias de aminoácidos para las 18 posiciones del motivo ZCCHC (página 80), lo que nos ha permitido también proponer

la existencia de restricciones de ciertos aminoácidos en posiciones concretas del motivo. Por ejemplo, nuestras observaciones sugieren que una secuencia $CX_2CX_4HX_4C$ en la que el triptófano (W) ocupa una posición que no es la decimosegunda no es un motivo ZCCHC.

Los alineamientos múltiples de los motivos ZCCHC de cada una de las tres especies consideradas por separado también nos han permitido identificar una singularidad de los de Arabidopsis: un enriquecimiento del 42% en arginina (R) en la posición 14, que no se observa en los de las proteínas de la levadura o humanas (páginas 54 y 55).

No hemos podido obtener alineamientos múltiples de las secuencias completas de las proteínas de una misma especie, aunque lo hemos intentado a solicitud de uno de los revisores que juzgaron nuestro manuscrito. Suponemos que la razón es que el motivo ZCCHC, que es la única secuencia que dichas proteínas comparten, es demasiado corto, y su grado de conservación, muy bajo, ya que se limita a las tres C y la H, y en el 81% de ellos a una G adyacente a la H, tal como se mencionó anteriormente. Sin embargo, hemos obtenido un alineamiento múltiple de los 198 motivos ZCCHC de las 110 proteínas que hemos estudiado, y a partir de este, la correspondiente filogenia (página 56). Nos parece particularmente relevante que en dicha filogenia se agrupen en un clado los motivos ZCCHC de las proteínas con secuencias completas muy poco semejantes, que por tanto no deberían suponerse ortólogas. Este es el caso de los motivos ZCCHC de la SERK1IP1 humana y la CXIP4 de Arabidopsis, que hemos estudiado en esta Tesis (Aceituno-Valenzuela *et al.*, 2025); esta semejanza, evidenciada por nuestra filogenia, permite proponerlas como ortólogas lejanas, ya que difieren sustancialmente en el resto de sus secuencias (Figura 4, en la página 20).

VI.1.2.- Comparación de nuestros resultados con los de otros autores

Se han publicado otras dos búsquedas sistemáticas de proteínas con dominio ZCCHC posteriores a la nuestra. Una de ellas se realizó en el trigo almidonero silvestre (*Triticum dicoccoides*), en el que se identificaron 50 genes que codifican proteínas con motivos ZCCHC (Sun *et al.*, 2022a), de los que solo 22 tienen ortólogos en Arabidopsis. En la segunda se identificaron siete genes en *Ustilaginoidea virens*, un hongo patógeno del arroz (Chen *et al.*, 2021), un resultado que coincide con el nuestro para *Saccharomyces cerevisiae*.

En la especie humana no se ha realizado ninguna búsqueda sistemática, salvo la nuestra, en la que concluimos que las proteínas humanas con motivos ZCCHC son 31, un número superior a las 25 que se recogen en la base de datos HUGO y se mencionan en algunas revisiones bibliográficas (Wang *et al.*, 2021). No obstante, en un análisis reciente de proteínas con capacidad de unión a ARN demostrada experimentalmente, aunque carentes

de un motivo de unión a ARN conocido, se incluyeron otras cinco adicionales a las anotadas en HUGO y que nosotros identificamos previamente como proteínas con motivos ZCCHC, una de las cuales es SREK1IP1. Se ha propuesto recientemente que se incluya al motivo ZCCHC entre los de unión a ARN (Ray *et al.*, 2023).

VI.2.- Caracterización funcional del gen *CXIP4*

VI.2.1.- Análisis de mutantes insercionales del gen *CXIP4*

En esta Tesis también se ha abordado la caracterización funcional del gen *CXIP4* de Arabidopsis, utilizando sus dos alelos insercionales disponibles, a los que hemos denominado *cxip4-1* y *cxip4-2*. Aunque sus inserciones de ADN-T interrumpen el gen *CXIP4* en puntos muy próximos de su región codificante, sus efectos sobre el fenotipo morfológico y el molecular son diferentes (páginas 83 a 86 y 91). Hemos constatado que *cxip4-1* es recesivo y letal embrionario; no obstante, un 2% de las semillas *cxip4-1/cxip4-1* germinan y generan plántulas muy pequeñas, sin diferenciación de órganos visible y con estructuras compactas en forma de callo, que no son viables (páginas 83 y 84). Estas plántulas aberrantes se asemejan a los mutantes *pin-formed* (*pin*), cuyo transporte polar de la auxina está alterado (Bennett *et al.*, 1995; Křeček *et al.*, 2009), así como *atrtf2-2*, portador de un alelo mutante del gen que codifica la REPLICATION TERMINATION FACTOR 2 (RTF2) DOMAIN PROTEIN, que participa en la regulación del *splicing* (Sasaki *et al.*, 2015). Por su parte, el alelo *cxip4-2* es recesivo, viable en homocigosis y aparentemente de expresividad variable; su homocigosis causa un fenotipo morfológico altamente pleiotrópico. Las plantas *cxip4-2* manifiestan un retraso generalizado del desarrollo y hojas apuntadas con pocos tricomas y acumulación de antocianinas, lo que sugiere que están constitutivamente estresadas (páginas 83 a 85 y 101). Este fenotipo foliar se atenúa cuando las hojas se expanden, lo que sugiere que *CXIP4* es más importante para la proliferación que para la expansión celular.

Los fenotipos mutantes descritos en el párrafo anterior se deben a la pérdida de la función de *CXIP4*; lo hemos demostrado mediante la transferencia de un transgén portador del alelo silvestre de este gen a plantas *cxip4-1* y *cxip4-2* (páginas 83 a 86). Hemos concluido que *cxip4-1* produce una proteína más anómala que la de *cxip4-2*, que está truncada, lo que justificaría sus diferentes fenotipos y sugiere que *cxip4-1* es un alelo nulo, y *cxip4-2*, hipomorfo (páginas 84, 100 y 103). Nuestras observaciones sugieren que *CXIP4* es un gen esencial que participa en diversas etapas del desarrollo vegetal. La letalidad embrionaria causada por la pérdida de la función de *CXIP4* que hemos observado en el mutante *cxip4-1* (páginas 83 y 84) parece corresponderse con la descrita en cultivos de células humanas tras el silenciamiento del gen *SREK1IP1* mediante ARN interferentes (Heese *et al.*, 2004).

VI.2.2.- Análisis de los productos de la expresión de *CXIP4*

Hemos determinado el patrón de expresión espacial y temporal de *CXIP4*, construyendo dos transgenes, *CXIP4$_{proI}$:GUS* y *CXIP4$_{proII}$:GUS*. Hemos usado para ello dos segmentos de diferente longitud de la región intergénica situada aguas arriba de la unidad de transcripción de *CXIP4*, que suponíamos debía contener su promotor basal y secuencias reguladoras específicas. Hemos obtenido resultados similares de actividad del gen testigo *GUS* con ambos transgenes, detectando tinción GUS en la mayoría de los tejidos estudiados, que fue más intensa en los que se encontraban en proliferación activa. También detectamos tinción en embriones (página 104).

También hemos construido y transferido a plantas los transgenes *CXIP4$_{proI}$:CXIP4:GFP* y *CXIP4$_{proII}$:CXIP4:GFP*. Hemos detectado así la fluorescencia de la proteína de fusión CXIP4-GFP en el núcleo de las células de la raíz de las plantas transgénicas, en un patrón punteado (página 87), que debe corresponderse con el de la CXIP4 endógena, ya que la expresión del transgén rescata los fenotipos de los mutantes *cxip4-1* y *cxip4-2* (páginas 83 y 84). Suponemos que las acumulaciones discretas de fluorescencia que hemos encontrado corresponden a los así denominados *speckles* nucleares, tal como sugieren nuestros resultados obtenidos con el programa MULocDeep, que propone localizaciones suborganulares muy parecidas para CXIP4 y SREK1IP1 (página 107). MULocDeep es un programa de dominio público que permite predecir la localización de proteínas en 44 suborgánulos de los 10 principales compartimentos subcelulares, que han sido verificadas experimentalmente y están anotadas en UniprotKB (Jiang *et al.*, 2023).

Nuestros resultados coinciden con los de las proteínas CXIP4 caracterizadas en otras plantas, como el arándano (*Vaccinium corymbosum*; VcCXIP4) y el trigo harinero (*Triticum aestivum*; TaCXIP4), que también se localizan en el núcleo de las células del mesófilo de hojas del tabaco con un patrón punteado (Martins y Gerós, 2020; Chen *et al.*, 2022). La proteína humana SREK1IP1 es exclusivamente nuclear y manifiesta un patrón punteado (Huttlin *et al.*, 2017), sugiriendo una función conservada con la proteína humana SREK1IP1.

VI.2.3.- Análisis transcriptómico de *cxip4-2*

Hemos obtenido y comparado los transcriptomas de las plantas *cxip4-2* y Col-0 con el doble objetivo de identificar los genes desregulados y las alteraciones en su patrón de *splicing*. El escaso desarrollo de las plantas *cxip4-1/cxip4-1* y la baja frecuencia con la que aparecían en la descendencia de las plantas heterocigóticas *CXIP4/cxip4-1*, nos ha impedido utilizarlas también en los ensayos transcriptómicos.

Encontramos 1.094 genes sobreexpresados y 1.066 reprimidos, entre ellos *CXIP4* (páginas 89 y 109; Supplementary Data Set S1 de Aceituno-Valenzuela *et al.*, 2025). Hemos usado PANTHER para analizar separadamente el enriquecimiento en categorías funcionales de los genes sobreexpresados y reprimidos (Supplementary Data Set S1 de Aceituno-Valenzuela *et al.*, 2025). Hemos detectado entre los genes sobreexpresados el enriquecimiento de términos GO relacionados con la síntesis de antocianinas, así como la regulación y la biosíntesis de los flavonoides, lo que justificaría la acumulación de antocianinas en las plantas *cxip4-2* y nuestra suposición de que están constitutivamente estresadas. También hemos encontrado enriquecimiento, entre los genes sobreexpresados, de categorías relacionadas con la defensa frente a insectos y al daño mecánico.

Entre los genes reprimidos hemos detectado enriquecimiento en categorías relacionadas con la respuesta inmune frente a patógenos, lo que sugiere que CXIP4 participa en la regulación de mecanismos de defensa. La modulación de la respuesta inmune por CXIP4 concuerda con la participación de su ortóloga en el trigo en la respuesta de defensa contra *Fusarium graminearum* (Chen *et al.*, 2022). Hemos identificado entre los genes reprimidos a *GLABRA 1* (*GL1*) y entre los sobreexpresados a *PRODUCTION OF ANTHOCYANIN PIGMENT 1* (*PAP1*), que codifican activadores transcripcionales de genes implicados en el desarrollo de los tricomas y la producción de antocianinas, respectivamente (Larkin *et al.*, 1994; Teng *et al.*, 2005). La desregulación de estos dos genes, que hemos confirmado mediante RT-qPCR, podría explicar la baja densidad de tricomas y la acumulación de antocianinas observada en las hojas de *cxip4-2* (páginas 85 y 109).

Hemos analizado también las diferencias de los patrones de *splicing* de los pre-ARNm de las plantas *cxip4-2* y las silvestres. Hemos detectado 939 eventos de *splicing* alternativo significativamente diferentes de los de Col-0, siendo los más frecuentes los de retención intrónica, tal como se ha descrito para las plantas *prp8-7* (Sasaki *et al.*, 2015). Hemos comprobado mediante RT-PCR que la retención intrónica es mayor en el mutante nulo *cxip4-1* que en el hipomorfo *cxip4-2* (páginas 90 y 91; Supplementary Data Set S2 de Aceituno-Valenzuela *et al.*, 2025).

Hemos realizado también un análisis de enriquecimiento en categorías funcionales de genes que sufren un *splicing* alternativo diferente en *cxip4-2* y en el tipo silvestre. Hemos detectado así en *cxip4-2* enriquecimiento en genes implicados en la metilación del ADN guiada por ARN interferentes, el *splicing* y la remodelación de la cromatina (páginas 90 y 91; Supplementary Data Set S2 de Aceituno-Valenzuela *et al.*, 2025). Nos ha parecido muy relevante que la categoría de genes relacionados con la desadenilación de los ARNm, que es la primera etapa de la degradación de estos últimos, estuviera enriquecida entre los genes

cuyo *splicing* es diferente en las plantas *cxip4-2* respecto a las silvestres, lo que sugiere que CXIP4 contribuye al control de la degradación de los ARNm cuyo *splicing* ha sido aberrante.

VI.2.4.- Retención nuclear de ARN poliadenilados en *cxip4-2*

Las células eucarióticas han desarrollado mecanismos para prevenir la traducción de los pre-ARNm, entre ellos su retención nuclear hasta que su *splicing* se haya completado (Palazzo y Lee, 2018; Wegener y Müller-McNicoll, 2018; Rudzka *et al.*, 2022). Hemos detectado acumulación nuclear de ARN poliadenilados en las plantas *cxip4-2*, que no se observa en las silvestres (páginas 89 a 91) y es similar a la que manifiesta el mutante *prp8-7* (Cabezas-Fuster *et al.*, 2022). También es comparable a la retención nuclear observada en el mutante *sar1-4*, portador de un alelo hipomorfo del gen que codifica la nucleoporina NUP160, que juega un importante papel en la exportación de los ARNm del núcleo al citoplasma (Dong *et al.*, 2006; Parry *et al.*, 2006). Se ha observado en la levadura acumulación nuclear de ARN poliadenilados causada por mutaciones en genes que codifican componentes del espliceosoma o nucleoporinas (Paul y Montpetit, 2016).

Nuestros resultados sugieren que en las plantas *cxip4-2* se retienen en el núcleo celular moléculas de pre-ARNm incorrectamente procesadas por el espliceosoma, lo que a su vez permite proponer que CXIP4 participa en el *splicing* e interviene en la exportación de ARNm maduros del núcleo al citoplasma y/o en la degradación de los ARNm cuyo *splicing* ha sido aberrante. En ausencia de CXIP4, estos ARNm aberrantes no podrían ser exportados al citoplasma y se acumularían en el núcleo. Las interacciones anotadas en la base de datos HitPredict entre la SREK1IP1 humana y algunos factores del *splicing* y la exportación nuclear, principalmente la NUP153, son compatibles con nuestras hipótesis (páginas 115 y 116). Estas observaciones sugieren la conservación de las funciones de SREK1IP1 y CXIP4, en las que podría ser crucial el motivo ZCCHC, que es lo único que parecen compartir.

VII.- CONCLUSIONES Y PERSPECTIVAS

VII.- CONCLUSIONES Y PERSPECTIVAS

En esta Tesis hemos realizado una revisión sistemática de las proteínas que contienen motivos de dedos de zinc de tipo ZCCHC en *Saccharomyces cerevisiae*, *Arabidopsis thaliana* y *Homo sapiens*, a los que hemos considerado representativos de los hongos, las plantas y los metazoos, respectivamente. Hemos usado varias herramientas bioinformáticas y bases de datos para identificar las proteínas que contuviesen la secuencia $CX_2CX_4HX_4C$ y estuvieran anotadas o predichas. Hemos analizado las secuencias de los 198 presuntos motivos ZCCHC que hemos identificado en 110 proteínas y deducido tanto su secuencia consenso como sus restricciones de aminoácidos en ciertas posiciones del motivo. También hemos intentado inferir la trascendencia de la presencia de los motivos ZCCHC en base a su conservación entre proteínas consideradas ortólogas o que podrían serlo aunque solo comparten dicho motivo; este es el caso de la CXIP4 de Arabidopsis y la SRK1IP1 humana, y probablemente también el de otras proteínas de las familias CXIP4-like de las plantas y SREK1IP1-like de los animales.

Creemos haber identificado proteínas con motivos ZCCHC de una forma sistemática, sin sesgos, por lo que nuestros resultados permiten extraer conclusiones más fiables que algunas anotaciones en bases de datos, como PROSITE. De hecho, hemos establecido que el logo propuesto en PROSITE para la secuencia consenso del motivo ZCCHC estaba basado en el alineamiento de más de 500 proteínas, casi la mitad de las cuales corresponden a retrotransposones y otras muchas de mamíferos filogenéticamente muy próximos, mientras que las de las plantas estaban muy poco representadas.

En esta Tesis hemos intentado comprender mejor la función de la proteína MAS2 de Arabidopsis, mediante el análisis funcional de CXIP4, su interactor más representado en un ensayo del doble híbrido de la levadura previamente realizado. *MAS2* es un gen de Arabidopsis identificado y estudiado en Sánchez-García *et al.* (2015), un trabajo del laboratorio de M.R. Ponce, tras el aislamiento de sus alelos supresores de los defectos del *splicing* de *ago1-52* (página 16). En dicho trabajo se descubrió el mecanismo de supresión que relacionaba a MAS2 con la fidelidad del *splicing* y con el control de la expresión del ADNr 45S. La identificación de varios interactores de la proteína MAS2, cuyas funciones presuntas o demostradas los relacionaban con el *splicing* de los pre-ARNm o la biogénesis del ribosoma, fue clave para descubrir la implicación de MAS2 en estos dos procesos.

Hemos realizado en esta Tesis el primer análisis funcional de alelos mutantes de pérdida de función de un gen de la familia PTHR31437, constituida por las proteínas CXIP4-like de las plantas y SREK1IP1-like de los animales. En concreto, dos alelos del gen *CXIP4*

de Arabidopsis: uno nulo, *cxip4-1*, y otro hipomorfo, *cxip4-2*. Los miembros de las familias CXIP4-like de las plantas y SREK1IP1-like de los metazoos se consideran ortólogos muy divergentes y apenas se habían estudiado. Hemos descubierto que el gen *CXIP4* de Arabidopsis es esencial y que su insuficiencia de función parcial retrasa el crecimiento y afecta a muchos aspectos del desarrollo vegetal. Nuestra demostración de que dichos efectos fenotípicos se atenúan a lo largo del crecimiento de las plantas *cxip4-2* y que el desarrollo embrionario de *cxip4-1* se interrumpe en fases tempranas de su desarrollo embrionario sugieren que CXIP4 se requiere en las etapas en las que la proliferación celular es más necesaria. Hemos encontrado también indicios de que CXIP4 participa en el control de la expresión génica mediante metilación del ADN mediada por ARN, que se intentarán confirmar más adelante.

VIII.- BIBLIOGRAFÍA DE LOS APARTADOS IV-VII

VIII.- BIBLIOGRAFÍA DE LOS APARTADOS IV-VII

Albert, I., Böhm, H., Albert, M., Feiler, C.E., Imkampe, J., Wallmeroth, N., Brancato, C., Raaymakers, T.M., Oome, S., Zhang, H., Krol, E., Grefen, C., Gust, A.A., Chai, J., Hedrich, R., Van den Ackerveken, G., y Nürnberger, T. (2015). An RLP23–SOBIR1–BAK1 complex mediates NLP-triggered immunity. *Nature Plants* **1,** 15140.

Ali, G.S., Palusa, S.G., Golovkin, M., Prasad, J., Manley, J.L., y Reddy, A.S.N. (2007). Regulation of plant developmental processes by a novel splicing factor. *PLOS ONE* **2,** e471.

Ariel, F., Romero-Barrios, N., Jégu, T., Benhamed, M., y Crespi, M. (2015). Battles and hijacks: noncoding transcription in plants. *Trends in Plant Science* **20,** 362-371.

Armas, P., y Calcaterra, N.B. (2012). Retroviral Zinc knuckles in eukaryotic cellular proteins, en *Zinc fingers: structure, properties, and applications*. Ciofani, R., y Makrlik, L., eds. Nova Biomedical Publishers.

Baulcombe, D. (2004). RNA silencing in plants. *Nature* **431,** 356-363.

Bennett, S.R.M., Alvarez, J., Bossinger, G., y Smyth, D.R. (1995). Morphogenesis in *pinoid* mutants of *Arabidopsis thaliana. Plant Journal* **8,** 505-520.

Bentley, D.L. (2014). Coupling mRNA processing with transcription in time and space. *Nature Reviews Genetics* **15,** 163-175.

Berná, G., Robles, P., y Micol, J.L. (1999). A mutational analysis of leaf morphogenesis in *Arabidopsis thaliana. Genetics* **152,** 729-742.

Bhat, A.A., Younes, S.N., Raza, S.S., Zarif, L., Nisar, S., Ahmed, I., Mir, R., Kumar, S., Sharawat, S.K., Hashem, S., Elfaki, I., Kulinski, M., Kuttikrishnan, S., Prabhu, K.S., Khan, A.Q., Yadav, S.K., El-Rifai, W., Zargar, M.A., Zayed, H., Haris, M., y Uddin, S. (2020). Role of non-coding RNA networks in leukemia progression, metastasis and drug resistance. *Molecular Cancer* **19,** 57.

Bhogireddy, S., Mangrauthia, S.K., Kudapa, H., Kumar, R., Pandey, A.K., Singh, S., Jain, A., Budak, H., y Varshney, R.K. (2021). Regulatory non-coding RNAs: a new frontier in regulation of plant biology. *Functional and integrative genomics* **21,** 313-330.

Bohmert, K., Camus, I., Bellini, C., Bouchez, D., Caboche, M., y Benning, C. (1998). *AGO1* defines a novel locus of *Arabidopsis* controlling leaf development. *EMBO Journal* **17,** 170-180.

Brown, J.W., Smith, P., y Simpson, C.G. (1996). Arabidopsis consensus intron sequences. *Plant Molecular Biology* **32,** 531-535.

Budak, H., Kaya, S.B., y Cagirici, H.B. (2020). Long non-coding RNA in plants in the era of reference sequences. *Frontiers in Plant Science* **11,** 276.

Busch, A., y Hertel, K.J. (2012). Evolution of SR protein and hnRNP splicing regulatory factors. *Wiley Interdisciplinary Reviews: RNA* **3,** 1-12.

Cabezas-Fuster, A., Micol-Ponce, R., Fontcuberta-Cervera, S., y Ponce, M.R. (2022). Missplicing suppressor alleles of Arabidopsis PRE-MRNA PROCESSING FACTOR 8 increase splicing fidelity by reducing the use of novel splice sites. *Nucleic Acids Research* **50,** 5513-5527.

Cabezas-Fuster, A., Micol-Ponce, R., Sarmiento-Mañús, R., y Ponce, M.R. (2023). Cross-kingdom conservation of Arabidopsis RPS24 function in 18S rRNA maturation. *bioRχiv* doi: 2023.2004.2021.537868.

Carbonell, A., y Carrington, J.C. (2015). Antiviral roles of plant ARGONAUTES. *Current Opinion in Plant Biology* **27,** 111-117.

Carbonell, A., y Daròs, J.A. (2017). Artificial microRNAs and synthetic trans-acting small interfering RNAs interfere with viroid infection. *Molecular Plant Pathology* **18,** 746-753.

Cech, Thomas R., y Steitz, Joan A. (2014). The noncoding RNA revolution-trashing old rules to forge new ones. *Cell* **157,** 77-94.

Chang, W.L., Lee, D.C., Leu, S., Huang, Y.M., Lu, M.C., y Ouyang, P. (2003). Molecular characterization of a novel nucleolar protein, pNO40. *Biochemical and Biophysical Research Communications* **307,** 569-577.

Chaudhary, S., Jabre, I., Reddy, A.S.N., Staiger, D., y Syed, N.H. (2019). Perspective on alternative splicing and proteome complexity in plants. *Trends in Plant Science* **24,** 496-506.

Chen, H., Su, Z., Tian, B., Hao, G., Trick, H.N., y Bai, G. (2022). TaHRC suppresses the calcium-mediated immune response and triggers wheat Fusarium head blight susceptibility. *Plant Physiol* **190,** 1566-1569.

Chen, S., Liu, Y., Deng, Y., Liu, Y., Dong, M., Tian, Y., Sun, H., y Li, Y. (2019). Cloning and functional analysis of the *VcCXIP4* and *VcYSL6* genes as Cd-regulating genes in blueberry. *Gene* **686,** 104-117.

Chen, X., Liu, J., Cheng, Y., y Jia, D. (2002). *HEN1* functions pleiotropically in *Arabidopsis* development and acts in C function in the flower. *Development* **129,** 1085-1094.

Chen, X., Pei, Z., Peng, L., Qin, Q., Duan, Y., Liu, H., Chen, X., Zheng, L., Luo, C., y Huang, J. (2021). Genome-wide identification and functional characterization of CCHC-type zinc finger genes in *Ustilaginoidea virens. Journal of Fungi* **7,** 947.

Cheng, N.H., Liu, J.Z., Nelson, R.S., y Hirschi, K.D. (2004). Characterization of CXIP4, a novel *Arabidopsis* protein that activates the H^+/Ca^{2+} antiporter, CAX1. *FEBS Letters* **559,** 99-106.

Coletta, A., Pinney, J.W., Solís, D.Y.W., Marsh, J., Pettifer, S.R., y Attwood, T.K. (2010). Low-complexity regions within protein sequences have position-dependent roles. *BMC Systems Biology* **4,** 43.

De Bortoli, F., Espinosa, S., y Zhao, R. (2021). DEAH-box RNA helicases in pre-mRNA splicing. *Trends in Biochemical Sciences* **46,** 225-238.

de Castro, E., Sigrist, C.J., Gattiker, A., Bulliard, V., Langendijk-Genevaux, P.S., Gasteiger, E., Bairoch, A., y Hulo, N. (2006). ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Research* **34,** 362-365.

Dikaya, V., El Arbi, N., Rojas-Murcia, N., Nardeli, S.M., Goretti, D., y Schmid, M. (2021). Insights into the role of alternative splicing in plant temperature response. *Journal of Experimental Botany* **72,** 7384-7403.

Ding, Z., Meng, Y.-R., Fan, Y.-J., y Xu, Y.-Z. (2022). Roles of minor spliceosome in intron recognition and the convergence with the better understood major spliceosome. *Wiley Interdisciplinary Reviews: RNA* **14,** e1761.

Dong, C.H., Hu, X., Tang, W., Zheng, X., Kim, Y.S., Lee, B.H., y Zhu, J.K. (2006). A putative *Arabidopsis* nucleoporin, AtNUP160, is critical for RNA export and required for plant tolerance to cold stress. *Molecular and Cellular Biology* **26,** 9533-9543.

Doolittle, W.F. (2013). Is junk DNA bunk? A critique of ENCODE. *Proceedings of the National Academy of Sciences USA* **110,** 5294-5300.

Dybkov, O., Preussner, M., El Ayoubi, L., Feng, V.Y., Harnisch, C., Merz, K., Leupold, P., Yudichev, P., Agafonov, D.E., Will, C.L., Girard, C., Dienemann, C., Urlaub, H., Kastner, B., Heyd, F., y Lührmann, R. (2023). Regulation of 3' splice site selection after step 1 of splicing by spliceosomal C* proteins. *Science Advances* **9,** eadf1785.

Fabrizio, P., Dannenberg, J., Dube, P., Kastner, B., Stark, H., Urlaub, H., y Luhrmann, R. (2009). The evolutionarily conserved core design of the catalytic activation step of the yeast spliceosome. *Molecular Cell* **36,** 593-608.

Fica, S.M., Oubridge, C., Wilkinson, M.E., Newman, A.J., y Nagai, K. (2019). A human postcatalytic spliceosome structure reveals essential roles of metazoan factors for exon ligation. *Science* **363,** 710-714.

Guan, Q., Wu, J., Zhang, Y., Jiang, C., Liu, R., Chai, C., y Zhu, J. (2013). A DEAD box RNA helicase is critical for pre-mRNA splicing, cold-responsive gene regulation, and cold tolerance in Arabidopsis. *Plant Cell* **25,** 342-356.

Heese, K., Fujita, M., Akatsu, H., Yamamoto, T., Kosaka, K., Nagai, Y., y Sawada, T. (2004). The splicing regulatory protein p18SRP is down-regulated in Alzheimer's disease brain. *Journal of Molecular Neuroscience* **24,** 269-276.

Hegele, A., Kamburov, A., Grossmann, A., Sourlis, C., Wowro, S., Weimann, M., Will, Cindy L., Pena,

V., Lührmann, R., y Stelzl, U. (2012). Dynamic protein-protein interaction wiring of the human spliceosome. *Molecular Cell* **45,** 567-580.

Hossain, M.A., Henríquez-Valencia, C., Gómez-Páez, M., Medina, J., Orellana, A., Vicente-Carbajosa, J., y Zouhar, J. (2016). Identification of novel components of the Unfolded Protein Response in Arabidopsis. *Frontiers in Plant Science* **7,** 650.

Hunter, C., Sun, H., y Poethig, R.S. (2003). The *Arabidopsis* heterochronic gene *ZIPPY* is an ARGONAUTE family member. *Current Biology* **13,** 1734-1739.

Huttlin, E.L., Bruckner, R.J., Paulo, J.A., Cannon, J.R., Ting, L., Baltier, K., Colby, G., Gebreab, F., Gygi, M.P., Parzen, H., Szpyt, J., Tam, S., Zarraga, G., Pontano-Vaites, L., Swarup, S., White, A.E., Schweppe, D.K., Rad, R., Erickson, B.K., Obar, R.A., Guruharsha, K.G., Li, K., Artavanis-Tsakonas, S., Gygi, S.P., y Harper, J.W. (2017). Architecture of the human interactome defines protein communities and disease networks. *Nature* **545,** 505-509.

Hutvagner, G., y Simard, M.J. (2008). Argonaute proteins: key players in RNA silencing. *Nature Reviews Molecular Cell Biology* **9,** 22-32.

Jiang, W., y Chen, L. (2021). Alternative splicing: Human disease and quantitative analysis from high-throughput sequencing. *Computational and Structural Biotechnology Journal* **19,** 183-195.

Jiang, Y., Jiang, L., Akhil, C.S., Wang, D., Zhang, Z., Zhang, W., y Xu, D. (2023). MULocDeep web service for protein localization prediction and visualization at subcellular and suborganellar levels. *Nucleic Acids Research* **51,** W343-W349.

Jones, A.N., Graß, C., Meininger, I., Geerlof, A., Klostermann, M., Zarnack, K., Krappmann, D., y Sattler, M. (2022). Modulation of pre-mRNA structure by hnRNP proteins regulates alternative splicing of MALT1. *Science Advances* **8,** eabp9153.

Jover-Gil, S., Candela, H., y Ponce, M.R. (2005). Plant microRNAs and development. *International Journal of Developmental Biology* **49,** 733-744.

Jover-Gil, S., Candela, H., Robles, P., Aguilera, V., Barrero, J.M., Micol, J.L., y Ponce, M.R. (2012). The microRNA pathway genes *AGO1*, *HEN1* and *HYL1* participate in leaf proximal-distal, venation and stomatal patterning in Arabidopsis. *Plant and Cell Physiology* **53,** 1322-1333.

Kanno, T., Venhuizen, P., Wu, M.T., Chiou, P., Chang, C.L., Kalyna, M., Matzke, A.J.M., y Matzke, M. (2020). A collection of pre-mRNA splicing mutants in *Arabidopsis thaliana*. *G3 Genes Genomes Genetics* **10,** 1983-1996.

Kashkan, I., Růžička, K., y Timofeyenko, K. (2022). How alternative splicing changes the properties of plant proteins. *Quantitative Plant Biology* **3,** e14.

Koncz, C., Dejong, F., Villacorta, N., Szakonyi, D., y Koncz, Z. (2012). The spliceosome-activating complex: molecular mechanisms underlying the function of a pleiotropic regulator. *Frontiers in Plant Science* **3,** 9.

Křeček, P., Skůpa, P., Libus, J., Naramoto, S., Tejos, R., Friml, J., y Zažímalová, E. (2009). The PIN-FORMED (PIN) protein family of auxin transporters. *Genome Biology* **10,** 249.

Krishna, S.S., Majumdar, I., y Grishin, N.V. (2003). Structural classification of zinc fingers: SURVEY AND SUMMARY. *Nucleic Acids Research* **31,** 532-550.

Kufel, J., Diachenko, N., y Golisz, A. (2022). Alternative splicing as a key player in the fine-tuning of the immunity response in Arabidopsis. *Molecular Plant Pathology* **23,** 1226-1238.

Larkin, J.C., Oppenheimer, D.G., Lloyd, A.M., Paparozzi, E.T., y Marks, M.D. (1994). Roles of the *GLABROUS1* and *TRANSPARENT TESTA GLABRA* genes in Arabidopsis trichome development. *Plant Cell* **6,** 1065-1076.

Li, Z., Li, W., Guo, M., Liu, S., Liu, L., Yu, Y., Mo, B., Chen, X., y Gao, L. (2022). Origin, evolution and diversification of plant ARGONAUTE proteins. *Plant Journal* **109,** 1086-1097.

Ling, Y., Mahfouz, M.M., y Zhou, S. (2021). Pre-mRNA alternative splicing as a modulator for heat stress response in plants. *Trends in Plant Science* **26,** 1153-1170.

Love, S.L., Emerson, J.D., Koide, K., y Hoskins, A.A. (2023). Pre-mRNA splicing-associated diseases and therapies. *RNA Biology* **20,** 525-538.

Lu, C., y Fedoroff, N. (2000). A mutation in the *Arabidopsis HYL1* gene encoding a dsRNA binding protein affects responses to abscisic acid, auxin, and cytokinin. *Plant Cell* **12,** 2351-2366.

Malgieri, G., Palmieri, M., Russo, L., Fattorusso, R., Pedone, P.V., y Isernia, C. (2015). The prokaryotic zinc-finger: structure, function and comparison with the eukaryotic counterpart. *The FEBS Journal* **282,** 4480-4496.

Marquardt, S., Raitskin, O., Wu, Z., Liu, F., Sun, Q., y Dean, C. (2014). Functional consequences of splicing of the antisense transcript *COOLAIR* on *FLC* transcription. *Molecular Cell* **54,** 156-165.

Marquez, Y., Brown, J.W., Simpson, C., Barta, A., y Kalyna, M. (2012). Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. *Genome Research* **22,** 1184-1195.

Martins, V., y Gerós, H. (2020). The grapevine CAX-interacting protein VvCXIP4 is exported from the nucleus to activate the tonoplast Ca^{2+}/H^{+} exchanger VvCAX3. *Planta* **252,** 35.

Matsui, A., Nakaminami, K., y Seki, M. (2019). Biological function of changes in RNA metabolism in plant adaptation to abiotic stress. *Plant and Cell Physiology* **60,** 1897-1905.

Meinke, D., y Koornneef, M. (1997). Community standards for *Arabidopsis* genetics. *Plant Journal* **12,** 247-253.

Micol-Ponce, R., Sarmiento-Mañús, R., Ruiz-Bayón, A., Montacié, C., Sáez-Vasquez, J., y Ponce, M.R. (2018). Arabidopsis RIBOSOMAL RNA PROCESSING7 is required for 18S rRNA maturation. *Plant Cell* **30,** 2855-2872.

Micol-Ponce, R., Sarmiento-Mañús, R., Fontcuberta-Cervera, S., Cabezas-Fuster, A., de Bures, A., Sáez-Vásquez, J., y Ponce, M.R. (2020). SMALL ORGAN4 is a ribosome biogenesis factor involved in 5.8S ribosomal RNA maturation. *Plant Physiology* **184,** 2022-2039.

Mort, M., Sterne-Weiler, T., Li, B., Ball, E.V., Cooper, D.N., Radivojac, P., Sanford, J.R., y Mooney, S.D. (2014). MutPred Splice: machine learning-based prediction of exonic variants that disrupt splicing. *Genome Biology* **15,** R19.

Neuvéglise, C., Marck, C., y Gaillardin, C. (2011). The intronome of budding yeasts. *Comptes Rendus Biologies* **334,** 662-670.

Nidumukkala, S., Tayi, L., Chittela, R.K., Vudem, D.R., y Khareedu, V.R. (2019). DEAD box helicases as promising molecular tools for engineering abiotic stress tolerance in plants. *Critical Reviews in Biotechnology* **39,** 395-407.

Nilsen, T.W., y Graveley, B.R. (2010). Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463,** 457-463.

O'Brien, J., Hayder, H., Zayed, Y., y Peng, C. (2018). Overview of microRNA biogenesis, mechanisms of actions, and circulation. *Frontiers in Endocrinology* **9,** 402.

Olson, E.D., y Musier-Forsyth, K. (2019). Retroviral Gag protein-RNA interactions: Implications for specific genomic RNA packaging and virion assembly. *Seminars in Cell and Developmental Biology* **86,** 129-139.

Palazzo, A.F., y Lee, E.S. (2018). Sequence determinants for nuclear retention and cytoplasmic export of mRNAs and lncRNAs. *Frontiers in Genetics* **9,** 440.

Parry, G., Ward, S., Cernac, A., Dharmasiri, S., y Estelle, M. (2006). The *Arabidopsis* SUPPRESSOR OF AUXIN RESISTANCE proteins are nucleoporins with an important role in hormone signaling and development. *Plant Cell* **18,** 1590-1603.

Passmore, L.A., y Coller, J. (2022). Roles of mRNA poly(A) tails in regulation of eukaryotic gene expression. *Nature Reviews Molecular Cell Biology* **23,** 93-106.

Paul, B., y Montpetit, B. (2016). Altered RNA processing and export lead to retention of mRNAs near transcription sites and nuclear pore complexes or within the nucleolus. *Molecular Biology of the Cell* **27,** 2742-2756.

Petrillo, E. (2023). Do not panic: An intron-centric guide to alternative splicing. *Plant Cell* **35,** 1752-1761.

Plaschka, C., Newman, A.J., y Nagai, K. (2019). Structural basis of nuclear pre-mRNA splicing: Lessons

from yeast. *Cold Spring Harbor Perspectives in Biology* **11,** a032391.

Ramanathan, A., Robb, G.B., y Chan, S.-H. (2016). mRNA capping: biological functions and applications. *Nucleic Acids Research* **44,** 7511-7526.

Rappsilber, J., Ryder, U., Lamond, A.I., y Mann, M. (2002). Large-scale proteomic analysis of the human spliceosome. *Genome Research* **12,** 1231-1245.

Ray, D., Laverty, K.U., Jolma, A., Nie, K., Samson, R., Pour, S.E., Tam, C.L., von Krosigk, N., Nabeel-Shah, S., Albu, M., Zheng, H., Perron, G., Lee, H., Najafabadi, H., Blencowe, B., Greenblatt, J., Morris, Q., y Hughes, T.R. (2023). RNA-binding proteins that lack canonical RNA-binding domains are rarely sequence-specific. *Scientific Reports* **13,** 5238.

Ru, Y., Wang, B.B., y Brendel, V. (2008). Spliceosomal proteins in plants, en *Nuclear pre-mRNA processing in plants*. Reddy, A.S.N., y Golovkin, M., eds. Springer.

Rudzka, M., Wroblewska-Ankiewicz, P., Majewska, K., Hyjek-Skladanowska, M., Golebiewski, M., Sikora, M., Smolinski, D.J., y Kolowerzo-Lubnau, A. (2022). Functional nuclear retention of pre-mRNA involving Cajal bodies during meiotic prophase in European larch (*Larix decidua*). *Plant Cell* **34,** 2404-2423.

Sánchez-García, A.B., Aguilera, V., Micol-Ponce, R., Jover-Gil, S., y Ponce, M.R. (2015). Arabidopsis *MAS2*, an essential gene that encodes a homolog of animal NF-kappa B activating protein, is involved in 45S ribosomal DNA silencing. *Plant Cell* **27,** 1999-2015.

Sasaki, T., Kanno, T., Liang, S.-C., Chen, P.-Y., Liao, W.-W., Lin, W.-D., Matzke, A.J.M., y Matzke, M. (2015). An Rtf2 domain-containing protein influences pre-mRNA splicing and is essential for embryonic development in *Arabidopsis thaliana*. *Genetics* **200,** 523-535.

Sato, H., Mizoi, J., Tanaka, H., Maruyama, K., Qin, F., Osakabe, Y., Morimoto, K., Ohori, T., Kusakabe, K., Nagata, M., Shinozaki, K., y Yamaguchi-Shinozaki, K. (2014). Arabidopsis DPB3-1, a DREB2A interactor, specifically enhances heat stress-induced gene expression by forming a heat stress-specific transcriptional complex with NF-Y subunits. *Plant Cell* **26,** 4954-4973.

Sato, H., Todaka, D., Kudo, M., Mizoi, J., Kidokoro, S., Zhao, Y., Shinozaki, K., y Yamaguchi-Shinozaki, K. (2016). The Arabidopsis transcriptional regulator DPB3-1 enhances heat stress tolerance without growth retardation in rice. *Plant Biotechnology Journal* **14,** 1756-1767.

Shi, J., Yi, K., Liu, Y., Xie, L., Zhou, Z., Chen, Y., Hu, Z., Zheng, T., Liu, R., Chen, Y., y Chen, J. (2015). Phosphoenolpyruvate carboxylase in Arabidopsis leaves plays a crucial role in carbon and nitrogen metabolism. *Plant Physiology* **167,** 671-681.

Shi, Y. (2017). Mechanistic insights into precursor messenger RNA splicing by the spliceosome. *Nature Reviews Molecular Cell Biology* **18,** 655-670.

Sievers, F., y Higgins, D.G. (2018). Clustal Omega for making accurate alignments of many protein sequences. *Protein Science* **27,** 135-145.

Smith, C.W., y Valcárcel, J. (2000). Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends in Biochemical Sciences* **25,** 381-388.

Sterne-Weiler, T., Howard, J., Mort, M., Cooper, D.N., y Sanford, J.R. (2011). Loss of exon identity is a common mechanism of human inherited disease. *Genome Research* **21,** 1563-1571.

Sun, A., Li, Y., He, Y., Zou, X., Chen, F., Ji, R., You, C., Yu, K., Li, Y., Xiao, W., y Guo, X. (2022a). Comprehensive genome-wide identification, characterization, and expression analysis of CCHC-type zinc finger gene family in wheat (*Triticum aestivum* L.). *Frontiers in Plant Science* **13,** 892105.

Sun, S., Gao, T., Pang, B., Su, X., Guo, C., Zhang, R., y Pang, Q. (2022b). RNA binding protein NKAP protects glioblastoma cells from ferroptosis by promoting *SLC7A11* mRNA splicing in an m6A-dependent manner. *Cell Death and Disease* **13,** 73.

Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T., y Lehner, B. (2014). Synonymous mutations frequently act as driver mutations in human cancers. *Cell* **156,** 1324-1335.

Takatsuji, H. (1998). Zinc-finger transcription factors in plants. *Cellular and Molecular Life Sciences* **54,** 582-596.

Telfer, A., y Poethig, R.S. (1998). *HASTY*: a gene that regulates the timing of shoot maturation in

Arabidopsis thaliana. *Development* **125,** 1889-1898.

Teng, S., Keurentjes, J., Bentsink, L., Koornneef, M., y Smeekens, S. (2005). Sucrose-specific induction of anthocyanin biosynthesis in Arabidopsis requires the MYB75/PAP1 gene. *Plant Physiology* **139,** 1840-1852.

Tolstrup, N., Rouzé, P., y Brunak, S. (1997). A branch point consensus from Arabidopsis found by non-circular analysis allows for better prediction of acceptor sites. *Nucleic Acids Research* **25,** 3159-3163.

Turunen, J.J., Niemelä, E.H., Verma, B., y Frilander, M.J. (2013). The significant other: splicing by the minor spliceosome. *Wiley Interdisciplinary Reviews: RNA* **4,** 61-76.

Tweedie, S., Braschi, B., Gray, K., Jones, T.E.M., Seal, Ruth L., Yates, B., y Bruford, E.A. (2020). Genenames.org: the HGNC and VGNC resources in 2021. *Nucleic Acids Research* **49,** D939-D946.

Verma, P., Kumari, P., Negi, S., Yadav, G., y Gaur, V. (2022). Holliday junction resolution by At-HIGLE: an SLX1 lineage endonuclease from Arabidopsis thaliana with a novel in-built regulatory mechanism. *Nucleic Acids Research* **50,** 4630-4646.

Wahl, M.C., Will, C.L., y Lührmann, R. (2009). The spliceosome: design principles of a dynamic RNP machine. *Cell* **136,** 701-718.

Wang, B.B., y Brendel, V. (2004). The ASRG database: identification and survey of Arabidopsis thaliana genes involved in pre-mRNA splicing. *Genome Biology* **5,** R102.

Wang, Y., Yu, Y., Pang, Y., Yu, H., Zhang, W., Zhao, X., y Yu, J. (2021). The distinct roles of zinc finger CCHC-type (ZCCHC) superfamily proteins in the regulation of RNA metabolism. *RNA Biology* **18,** 2107-2126.

Wegener, M., y Müller-McNicoll, M. (2018). Nuclear retention of mRNAs – quality control, gene regulation and human disease. *Seminars in Cell and Developmental Biology* **79,** 131-142.

Wilkinson, M.E., Charenton, C., y Nagai, K. (2020). RNA splicing by the spliceosome. *Annual Review of Biochemistry* **89,** 359-388.

Will, C.L., y Lührmann, R. (2011). Spliceosome structure and function. *Cold Spring Harbor Perspectives in Biology* **3,** a003707.

Xu, C., Zhang, Z., He, J., Bai, Y., Cui, J., Liu, L., Tang, J., Tang, G., Chen, X., y Mo, B. (2023). The DEAD-box helicase RCF1 plays roles in miRNA biogenesis and RNA splicing in Arabidopsis. *Plant Journal* **116,** 144-160.

Xu, Y.Z., Newnham, C.M., Kameoka, S., Huang, T., Konarska, M.M., y Query, C.C. (2004). Prp5 bridges U1 and U2 snRNPs and enables stable U2 snRNP association with intron RNA. *EMBO Journal* **23,** 376-385.

Yan, T., Yoo, D., Berardini, T.Z., Mueller, L.A., Weems, D.C., Weng, S., Cherry, J.M., y Rhee, S.Y. (2005). PatMatch: a program for finding patterns in peptide and nucleotide sequences. *Nucleic Acids Research* **33,** W262-W266.

Yang, J., Ariel, F., y Wang, D. (2022). Plant long non-coding RNAs: biologically relevant and mechanistically intriguing. *Journal of Experimental Botany* **74,** 2364-2373.

Zhang, D., Chen, M.-X., Zhu, F.-Y., Zhang, J., y Liu, Y.-G. (2020). Emerging functions of plant Serine/Arginine-rich (SR) proteins: Lessons from animals. *Critical Reviews in Plant Sciences* **39,** 173-194.

Zhang, J., Bai, R., Li, M., Ye, H., Wu, C., Wang, C., Li, S., Tan, L., Mai, D., Li, G., Pan, L., Zheng, Y., Su, J., Ye, Y., Fu, Z., Zheng, S., Zuo, Z., Liu, Z., Zhao, Q., Che, X., Xie, D., Jia, W., Zeng, M.-S., Tan, W., Chen, R., Xu, R.-H., Zheng, J., y Lin, D. (2019). Excessive miR-25-3p maturation via N6-methyladenosine stimulated by cigarette smoke promotes pancreatic cancer progression. *Nature Communications* **10,** 1858.

Zhang, X.R., Qin, Z., Zhang, X., y Hu, Y. (2015). Arabidopsis SMALL ORGAN 4, a homolog of yeast NOP53, regulates cell proliferation rate during organ growth. *Journal of Integrative Plant Biology* **57,** 810-818.

IX.- PUBLICACIONES

Cellular and Molecular Life Sciences (2020) 77:3991–4014
https://doi.org/10.1007/s00018-020-03518-7

Cellular and Molecular Life Sciences

REVIEW

Genome-wide analysis of CCHC-type zinc finger (ZCCHC) proteins in yeast, Arabidopsis, and humans

Uri Aceituno-Valenzuela[1] · Rosa Micol-Ponce[1] · María Rosa Ponce[1]

Received: 23 October 2019 / Revised: 6 March 2020 / Accepted: 30 March 2020 / Published online: 18 April 2020

Abstract

The diverse eukaryotic proteins that contain zinc fingers participate in many aspects of nucleic acid metabolism, from DNA transcription to RNA degradation, post-transcriptional gene silencing, and small RNA biogenesis. These proteins can be classified into at least 30 types based on structure. In this review, we focus on the CCHC-type zinc fingers (ZCCHC), which contain an 18-residue domain with the $CX_2CX_4HX_4C$ sequence, where C is cysteine, H is histidine, and X is any amino acid. This motif, also named the "zinc knuckle", is characteristic of the retroviral Group Antigen protein and occurs alone or with other motifs. Many proteins containing zinc knuckles have been identified in eukaryotes, but only a few have been studied. Here, we review the available information on ZCCHC-containing factors from three evolutionarily distant eukaryotes—*Saccharomyces cerevisiae*, *Arabidopsis thaliana*, and *Homo sapiens*—representing fungi, plants, and metazoans, respectively. We performed systematic searches for proteins containing the $CX_2CX_4HX_4C$ sequence in organism-specific and generalist databases. Next, we analyzed the structural and functional information for all such proteins stored in UniProtKB. Excluding retrotransposon-encoded proteins and proteins harboring uncertain ZCCHC motifs, we found seven ZCCHC-containing proteins in yeast, 69 in Arabidopsis, and 34 in humans. ZCCHC-containing proteins mainly localize to the nucleus, but some are nuclear and cytoplasmic, or exclusively cytoplasmic, and one localizes to the chloroplast. Most of these factors participate in RNA metabolism, including transcriptional elongation, polyadenylation, translation, pre-messenger RNA splicing, RNA export, RNA degradation, microRNA and ribosomal RNA biogenesis, and post-transcriptional gene silencing. Several human ZCCHC-containing factors are derived from neofunctionalized retrotransposons and act as proto-oncogenes in diverse neoplastic processes. The conservation of ZCCHCs in orthologs of these three phylogenetically distant eukaryotes suggests that these domains have biologically relevant functions that are not well known at present.

Keywords Zinc-knuckle · ZCCHC · RNA metabolism · Yeast · Human · Arabidopsis

Introduction

Zinc fingers are small zinc-ligating domains that make up the largest and most diverse superfamily of nucleic acid-binding proteins in eukaryotes. Zinc fingers stably fold via the coordination of one or more zinc ions (Zn^{2+}) through cysteine (C), histidine (H), and, less frequently, aspartate (D) residues. Zinc finger-containing proteins are classified depending on the sequence and structure of their zinc finger motifs, which contain diverse combinations of C and H [1]. The classical zinc finger is the C_2H_2 type, which consists of two C and two H residues, and is often described as $CX_{2-4}CX_{12}HX_{2-6}H$ (where X is any amino acid) [2]. The structures of zinc fingers resemble multiple finger-shaped protrusions, through which they bind to their DNA, RNA, or protein targets [3]. The first protein with zinc fingers described was the general transcription factor TFIIIA of *Xenopus laevis*, a DNA- and RNA-binding protein with nine C_2H_2 motifs [4]. About 2% of human proteins contain zinc fingers, which have been classified into 30 different types, as described in The Human Genome Organization Gene Nomenclature Committee database (HGNC; https://www.genenames.org/).

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00018-020-03518-7) contains supplementary material, which is available to authorized users.

✉ María Rosa Ponce
mrponce@umh.es

[1] Instituto de Bioingeniería, Universidad Miguel Hernández, Campus de Elche, 03202 Elche, Spain

As classified in the InterPro database (https://www.ebi.ac.uk/interpro; [5]), zinc knuckles are 18-residue, CCHC-type zinc fingers (ZCCHC) containing the $CX_2CX_4HX_4C$ consensus sequence (IPR025836). ZCCHC-containing factors have two short β-strands connected by a turn (the zinc knuckle) followed by a short helix or loop. The first ZCCHC was found in the murine leukemia virus (MLV) and Rous avian sarcoma virus (RSV) [6], and later in many other Group Antigen (Gag) proteins of retroviral nucleocapsids and eukaryotic retrotransposons [7]. Except for the spumaviruses, all retroviral nucleocapsid proteins have one or two ZCCHCs [8]. The two ZCCHCs of the *Human immunodeficiency virus* (HIV)-1 Gag protein are considered the ZCCHC prototype and are essential for viral genome packaging and infectivity [9]. Figure 1 shows a representation of the three-dimensional structure of a viral ZCCHC [10], obtained using the Visual Molecular Dynamics program [11].

ZCCHCs have not been found in bacteria, but a similar motif with residues of three conserved C and one H (CX_4CXHX_5C) is present in the C-terminal region of the essential *Escherichia coli* protein YjeQ [12]. YjeQ is a conserved bacterial GTPase that acts as a chaperone, facilitating the assembly of the 30S ribosomal subunit. YjeQ contributes to the folding of the decoding center for mRNA translation and binds the mature 30S subunit through its ZCCHC-like domain [13]. YjeQ depletion causes an accumulation of free 30S and 50S ribosomal subunits, and a reduction in 70S mature ribosomes [12]. YjeQ and its bacterial orthologs are considered promising targets for new antibiotics, since these proteins are required for *Staphylococcus aureus* to grow in experimental assays with mice (*Mus* sp.) [14, 15]. ZCCHCs have also been found in the DNA primase of the Archaea *Pyrococcus furiosus* [16].

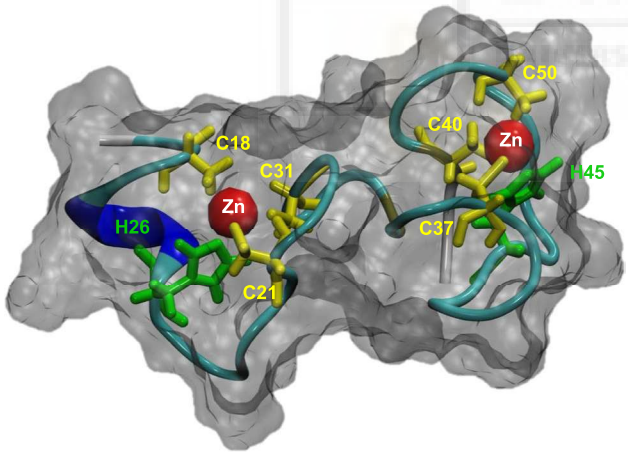


Fig. 1 Partial three-dimensional structure of the lentiviral EIAV nucleocapsid protein NCp11. The three-dimensional structure of a peptide of 37 amino acids (QTCYNCGKPGHLSSQCRAPKVCFKCKQPGHFSKQCRS) corresponding to residues 22–58 from the lentiviral EIAV nucleocapsid protein NCp11 (Protein Data Bank identification code for the partial structure: 2BL6) is shown. This peptide contains two ZCCHCs of the $CX_2CX_3GHX_4C$ type, which are underlined in the sequence above, and is complexed with zinc (red spheres). Only the relevant C and H residues of the ZCCHCs are highlighted. The structure shown in this figure was determined by two-dimensional hydrogen isotope nuclear magnetic resonance spectroscopy, as previously described [10]. We obtained this image using the Visual Molecular Dynamics molecular visualization program (https://www.ks.uiuc.edu/Research/vmd/) [11]

Here, we describe our systematic search and analysis of ZCCHC-containing factors from three evolutionarily distant eukaryotes, representing fungi, plants, and metazoa, respectively: *Saccharomyces cerevisiae* (hereafter, yeast), *Arabidopsis thaliana* (Arabidopsis), and *Homo sapiens* (humans). To our knowledge, no exhaustive compilation or comparative analysis of ZCCHC-containing factors in highly distant organisms has been published.

Genome-wide searches for ZCCHC-containing factors

To systematically and comprehensively identify ZCCHC-containing proteins, we carried out two different types of searches for each of the three species under study: the first used the search terms "knuckle" or "CCHC", and the second used the $CX_2CX_4HX_4C$ consensus sequence. We searched the UniProt Knowledgebase (UniProtKB; https://www.uniprot.org) using "knuckle" or "CCHC" and the name of the organism. Then, we repeated the search in organism-specific databases: the Saccharomyces Genome Database (https://www.yeastgenome.org), and the Arabidopsis Information Resource (TAIR; https://www.arabidopsis.org), and HGNC databases. The main problem that we found was that the results obtained from different searches were incoherent due to incomplete annotations and the presence of many duplicate, fragmented, or obsolete entries. This is likely because the UniProtKB database contains both reviewed (manually annotated) and unreviewed (computationally analyzed) records. For example, using "CCHC" and "*Saccharomyces cerevisiae*" as search terms on UniProtKB yielded 58 entries, many of which corresponded to the same protein from different yeast strains, and only eight of which were unique and harbored the consensus $CX_2CX_4HX_4C$ sequence. We then searched for the $CX_2CX_4HX_4C$ consensus sequence using the PatMatch pattern-matching program (https://www.yeastgenome.org/nph-patmatch), which can search for short sequences (< 30 amino acids), including ambiguous and degenerate patterns [17].

ZCCHC-containing factors encoded by the yeast genome

Using PatMatch, we obtained 19 unique sequence entries and 32 hits from 5,916 analyzed yeast sequences, several of which harbored more than one $CX_2CX_4HX_4C$ sequence. Eight of the 19 factors harboring a $CX_2CX_4HX_4C$ sequence were Gag proteins from retrotransposons. We next searched for structural information on the remaining 11 proteins (Supporting Dataset 1a) in the UniProtKB database, which compiles the information provided by three programs that predict protein domains: PROSITE (https://prosite.expasy.org), Pfam (https://pfam.xfam.org), and SMART (https://smart.embl.de). No domains or motifs were annotated for YOL029C or YER137C, two proteins of unknown function, and the $CX_2CX_4HX_4C$ sequences of other two proteins (DPOD/POL3 and HEL1) form part of two Cys-rich domains that are larger than that of the ZCCHC (Supporting Dataset 1a). The In Between RING finger (IBR) domains in HEL1 are ~ 50-residue C_6HC zinc finger domains ($CX_4CX_{14–30}CX_{1–4}CX_4\underline{CX_2CX_4HX_4C}$; the sequence found with PatMatch is underlined). IBRs have been described in members of some families of E3 ubiquitin-protein ligases, and form a Gag knuckle-like zinc-binding structure [18].

We concluded that the yeast genome encodes at least seven ZCCHC-containing factors, all with a known function; six of these are involved in RNA metabolism, as discussed later (Supporting Dataset 1b; for all yeast amino acid sequences, see Supplementary Fig. 1).

ZCCHC-containing factors encoded by the Arabidopsis genome

Equivalent results to those of yeast were obtained when we used the version of the PatMatch program found in the TAIR10 database (https://www.arabidopsis.org/cgi-bin/patmatch/nph-patmatch.pl), which includes 27,416 Arabidopsis proteins. We found 135 factors with a $CX_2CX_4HX_4C$ sequence; for 53 of these proteins, the $CX_2CX_4HX_4C$ sequence was annotated in UniProtKB to form part of domains or motifs that are not, and are larger than, a ZCCHC (Supporting Dataset 1c), mainly IBRs. Nine of the 82 remaining factors were Gag proteins of retrotransposons, and the other eight harbored a DUF4283 domain of unknown function (Supporting Dataset 1d), which is frequently associated with ZCCHC and with the reverse transcriptase domain, as described in the Pfam database. We classified these factors as putative retrotransposon-derived proteins.

Four other proteins (encoded by the AT3G16350, AT3G31950, AT3G50870, and AT4G08867 genes; Arabidopsis Genome Initiative gene identifiers [19] are used throughout this review) were not annotated as containing a ZCCHC in UniProtKB or Arabidopsis-specific databases, nor were predicted by PROSITE or SMART. In addition, proteins encoded by AT3G50870 and AT4G08867 belong to families without members harboring ZCCHCs, and we considered it very unlikely that their $CX_2CX_4HX_4C$ sequence was a true ZCCHC (Supporting Dataset 1c). However, AT3G16350 and AT3G31950 may encode genuine ZCCHC-containing factors. AT3G16350 encodes a MYB-type transcription factor with high similarity to KUODA1 (KUA1), sharing a very similar $CX_2CX_4HX_4C$ sequence that is annotated in UniProtKB as a ZCCHC. In the Aramemnon database (https://aramemnon.botanik.uni-koeln.de/), AT3G31950 is clustered with four proteins of unknown function (AT4G06479, AT4G06526, AT5G29070, and AT5G34870), the last three of which are annotated as zinc knuckle (CCHC-type) proteins. Therefore, we classified these four proteins as harboring uncertain ZCCHCs and did not include them in our subsequent analyses (Supporting Dataset 1c).

Most of the 61 remaining Arabidopsis factors with a ZCCHC are nuclear proteins involved in RNA metabolism (Supporting Dataset 1d; for all Arabidopsis amino acid sequences, see Supplementary Fig. 2). Many of these are predicted by homology, and only a few have been functionally studied.

ZCCHC-containing factors encoded by the human genome

There are PatMatch versions for yeast and Arabidopsis [17] searches, but none are integrated with human protein databases. Therefore, to search for human zinc knuckle proteins, we used the ScanProsite tool (https://prosite.expasy.org/scanprosite) with the $CX_2CX_4HX_4C$ sequence and "Homo sapiens" as a filter for the taxonomy option. As a result, we found 245 hits in 153 sequences, including isoforms corresponding to 69 different genes. When we analyzed these proteins in UniProtKB, we found a predicted ZCCHC in 34 factors (Supporting Dataset 1e). The $CX_2CX_4HX_4C$ sequence was placed in a Cys-rich region in four proteins, and in the remaining 31, it was a part of a larger domain, in most cases an IBR (Supporting Dataset 1e).

The 34 human ZCCHC-containing factors that we found in UniProtKB included the 24 proteins that are annotated in the HGNC database as ZCCHC1–ZCCHC25 (https://www.genenames.org/data/genegroup/#!/group/74), with the exception of ZCCHC23, because it harbors a degenerate ZCCHC in which the H residue is substituted by asparagine (N). Most human ZCCHC-containing factors of known function are involved in messenger RNA (mRNA), microRNA (miRNA), and ribosomal RNA (rRNA) metabolism (Supporting Dataset 1f; for all human amino acid sequences, see Supplementary Fig. 3).

Most ZCCHC-containing proteins are involved in RNA metabolism

We then analyzed each of the proteins that we considered to be genuine ZCCHC-containing factors (Supporting Datasets 1b, 1d, and 1f), looking for their orthologs in the other two species under study, and using databases and the Basic Local Alignment Search Tool for proteins (BLASTP) [20]. Then, we gathered information on each protein in general and specific databases, as well as in publications. We present the proteins that have been studied to some extent in three figures: yeast ZCCHC-containing factors, with their human and Arabidopsis orthologs, in Fig. 2; the remaining human factors and their Arabidopsis orthologs in Fig. 3; and the remaining Arabidopsis factors in Fig. 4. We did not include in these figures putative orthologs not harboring ZCCHCs. We added the structural information that we found in UniProtKB: motifs, domains, and low-complexity regions (Figs. 2, 3, 4).

We found that the structures and functions of ZCCHC-containing factors were very diverse, harboring a wide variety of domains and motifs in addition to the ZCCHC, and participating in many different processes. As described in UniProtKB and publications, most of the ZCCHC-containing factors participate in RNA metabolism, including transcriptional elongation, polyadenylation, translation, pre-mRNA splicing, RNA export, RNA degradation, miRNA and rRNA biogenesis, and post-transcriptional gene silencing. However, when we tried to group them into functional categories, the results were barely informative, because only a few proteins had been assigned a Gene Ontology (GO) or KEGG term (results of the GO analysis are included in Supporting Datasets 1b, 1d, and 1f). In this section, we summarize the most relevant information on the proteins that play a role in RNA metabolism, mostly from studies on yeast and humans.

Yeast Msl5 is involved in the assembly of early spliceosome complexes

Pre-mRNA splicing occurs in two consecutive transesterification reactions, executed by the spliceosome to remove introns and join exons, which co-transcriptionally contribute to mRNA maturation. More than 150 *trans*-acting factors and five small nuclear uridine-rich ribonucleoproteins (snRNPs; U1, U2, U4, U5, and U6) participate in splicing together with other splicing factors, such as serine (S)/arginine (R)-rich (SR) proteins—which contain an Arg-Ser-rich region (the RS domain)—and heterogeneous nuclear ribonucleoproteins (hnRNPs). In addition, several intronic sequences play essential roles in splicing, including the 5′ and 3′ splice sites and the branch point (reviewed in [21–24]).

Organism	Protein	Structure (motifs, domains and low-complexity regions)	Function or process
Yeast	Msl5/Bbp/Sf1	476	Assembly of early spliceosome complexes
Human	SF1/ZCCHC25	638	Assembly of early spliceosome complexes
Arabidopsis	SF1	804	pre-mRNA alternative splicing
Yeast	Slu7	382	Component of the spliceosome; splicing fidelity
Human	SLU7	586	Component of the spliceosome; splicing fidelity
Arabidopsis	AT3G45950, SMP1*, SMP2	535	Leaf development
Yeast	Mpe1	441	pre-mRNA cleavage and polyadenylation
Human	RBBP6	1792	pre-mRNA cleavage and polyadenylation
Arabidopsis	PQT3*, PQT3L	826	Regulation of oxidative stress tolerance
Yeast	Air1*, Air2	360	RNA processing, degradation and surveillance
Human	ZCCHC7	543	RNA processing, degradation and surveillance
Yeast	Gis2	153	Regulation of translation
Human	CNBP/ZCCHC22	177	Regulation of translation
Yeast	Bik1	440	Dynamics of microtubule cytoskeleton
Human	CLIP-170	1438	Dynamics of microtubule cytoskeleton

ZCCHC KH RRM DWNN RING CAP-Gly Prp18-interacting p53-interacting
RB1-interacting RS Pro-rich Lys-rich Gly-rich Arg-rich Arg/Gly-rich Ser-rich

Fig. 2 Schematic representation of the structures of yeast ZCCHC-containing factors, with indication of their putative Arabidopsis and human orthologs. Schemes have been drawn to scale. Motifs, domains, and low-complexity regions are represented by symbols, and were obtained from the UniProtKB database and publications. The Arabidopsis Genome Initiative [19] gene code is shown for Arabidopsis proteins. In cases of several paralogs, the asterisk indicates which protein is represented

Organism	Protein	Structure (motifs, domains and low-complexity regions)	Function or process
Human	ZCRB1/ZCCHC19	217	Component of the U12 spliceosome
Arabidopsis	U11/U12-31K	261	Component of the U12 spliceosome
Human	SRSF7/ZCCHC20	237	Nucleocytoplasmic mRNA export
Arabidopsis	RSZ22	200	Splicing factor (putative)
Human	pNO40/ZCCHC17	241	Splicing and ribosome biogenesis
Human	LIN28A/ZCCHC1	209	Suppressor of miRNA biogenesis
Arabidopsis	CSP2	203	Unknown
Human	LIN28B	250	Suppressor of miRNA biogenesis
Human	TUT4/ZCCHC11	1645	Suppressor of miRNA biogenesis
Human	TUT7/ZCCHC6	1495	Suppressor of miRNA biogenesis
Human	CPSF4/CPSF30	269	pre-mRNA cleavage and polyadenylation
Human	ZCCHC8	707	RNA surveillance
Arabidopsis	AT1G67210*, AT5G38600	403	RNA surveillance (putative)
Human	RBM4A/ZCCHC21	364	Alternative splicing; translational repression
Human	RBM4B/ZCCHC15	359	Unknown
Human	XRN2	950	RNA surveillance
Arabidopsis	XRN2*, XRN3, XRN4	1012	RNA surveillance
Human	ZCCHC4	513	rRNA biogenesis
Human	ZCCHC9	271	rRNA biogenesis
Arabidopsis	AT5G52380	268	Unknown
Human	DDX41	622	Innate immune response
Arabidopsis	RH35*, RH43	591	Unknown
Human	ZCCHC3	403	Innate immune response
Human	ZCCHC10	192	Tumor suppressor
Human	SREK1IP1	155	Unknown

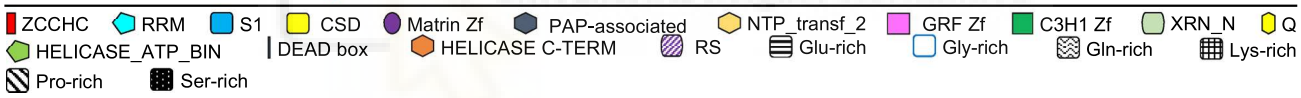


Fig. 3 Schematic representation of the structure of human ZCCHC-containing factors with indication of their putative Arabidopsis orthologs. Other details are as described in the legend of Fig. 2

Yeast Mud synthetic-lethal 5 (Msl5), also termed Branchpoint binding protein (Bbp) and Splicing factor 1 (Sf1), specifically interacts with the seven-nucleotide branch point sequences (UACUAAC) that are found close to the 3′ splice sites of pre-mRNA introns [25], and participates in the assembly of early spliceosome complexes [26]. The human ortholog of Msl5, termed SF1 or ZCCHC25 (Fig. 2), is encoded by an essential gene and binds to the degenerate sequence CURAY (where R and Y are purine and pyrimidine, respectively) [25]. Yeast Msl5 and human SF1 interact with Mutant U1 Die 2 (Mud2) and its human ortholog U2 Auxiliary Factor 65 kDa (U2AF65) splicing factor, respectively, facilitating branch point recognition [27]. SF1 harbors an extended K homology (KH) domain, a ZCCHC, and a Pro-rich C-terminal half. Deletion assays have shown that the KH domain, but not the ZCCHC or the Pro-rich region, is required for RNA binding and pre-spliceosome assembly [28]. Although similar in structure to human SF1, Msl5 includes a second ZCCHC (Fig. 2), whose specific function has not been determined, although it does not seem to be involved in RNA binding [29]. *MSL5* is also an essential gene in yeast. Conditional *msl5* mutants show aberrant splicing, mainly of introns with weak consensus splice sequences, and high levels of pre-mRNAs in the cytoplasm, suggesting that Msl5 is required for the nuclear retention of unspliced and misspliced pre-mRNAs [30].

The protein encoded by AT5G51300 is considered to be the Arabidopsis homolog of yeast Msl5 and human SF1 [31], and is annotated at TAIR and UniProtKB as ARABIDOPSIS SF1 HOMOLOG (AtSF1). The protein participates in alternative splicing of pre-mRNAs and its structure is very similar to those of its yeast and human orthologs, containing a KH domain, two ZCCHCs, an RNA recognition motif

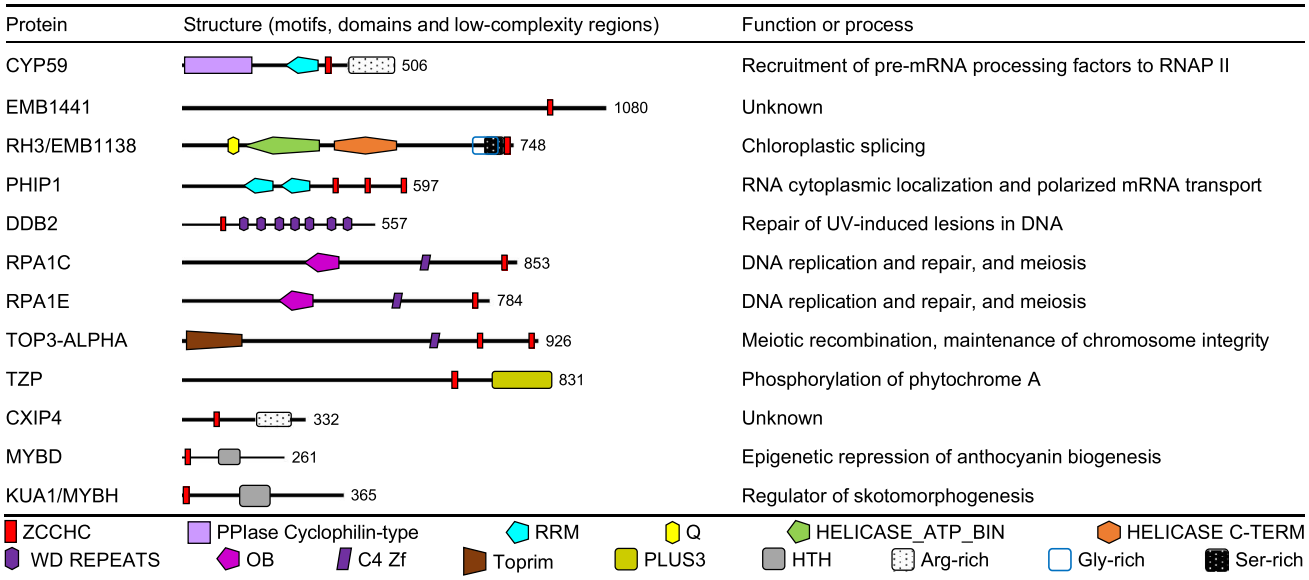


Fig. 4 Schematic representation of the structure of Arabidopsis ZCCHC-containing factors. Other details are as described in the legend of Fig. 2

(RRM), and a Pro-rich C-terminal region (Fig. 2). Unlike its *SF1* and *MLS5* orthologs, AT5G51300 seems to be a single-copy gene (according to our BLASTP search), which is not essential, and its loss of function causes dwarfism, early flowering, and hypersensitivity to abscisic acid [31].

Yeast Slu7 and its human ortholog are required for splicing fidelity

The yeast Pre-mRNA processing factor 18 (Prp18) stabilizes the interaction between adjacent exons and the U5 snRNA during the second step of pre-mRNA splicing [32]. Prp18 and Synergistic lethal with U5 snRNA protein 7 (Slu7) cooperate to promote the second step of splicing, and Slu7 is required to recruit Prp18 to the spliceosome [33, 34]. Yeast *SLU7* is essential and encodes a protein with a ZCCHC in its N-terminal half, followed by a Prp18-interacting region (Fig. 2). Slu7 is required for splicing fidelity, since it participates in the selection of the 3′ splice site [35].

Human SLU7 also seems to be involved in splicing fidelity, and its depletion alters the use of the splicing acceptor sequence (AG), so that the spliceosome indiscriminately binds to other AG dinucleotides [36]. The ZCCHC of human SLU7 is also located in its N-terminus, within a nuclear localization signal. This signal, together with the ZCCHC, is required to retain human SLU7 in the nucleus, but not for its import from the cytoplasm [37].

Arabidopsis has three putative SLU7 orthologs encoded by AT3G45950, AT1G65660, and AT4G37120 (Fig. 2). Alignment of these three proteins with other eukaryotic SLU7 factors revealed the presence of a conserved ZCCHC [37, 38]. AT3G45950 is annotated in TAIR and UniProtKB as a "Pre-mRNA splicing Prp18-interacting factor"; AT1G65660 and AT4G37120 are annotated in UniProtKB as "Pre-mRNA-splicing factor SLU7-A" (SLU7-A) and SLU7-B, respectively, and in TAIR as SWELLMAP1 (SMP1) and SMP2, respectively, because both were identified in a screen for leaf developmental abnormalities. Loss of function of *SMP* genes causes a decrease in plant size, narrow and pointed leaves, and defects in leaf venation pattern. *SMP1* and *SMP2* are functionally redundant during development [38]. The relationship among these three Arabidopsis proteins and pre-mRNA splicing has been inferred only from their similarity to SLU7 factors, and the role of their ZCCHC has not been determined.

Human ZCRB1/ZCCH19 is a component of the minor spliceosome

The U2-type spliceosome is present in all eukaryotes and processes most introns. Many eukaryotes also contain a minor spliceosome that excises the rare U12-type introns, which represent only 0.5% of the introns in a given organism and have different donor- and branch-site consensus sequences. The U12-type spliceosome has been lost independently in distantly related species, including yeast [39]. Most of the genes containing U12-type introns contain only one of these introns, and also have several U2-type introns. Splicing is slower in U12-type than in U2-type introns, possibly allowing control of the cellular levels of mature mRNAs because inefficient splicing can cause retention of the U12 intron and mRNA decay [40].

Human zinc finger CCHC-type and RNA-binding domain-containing 1 (ZCRB1), also named ZCCHC19 and

U11/U12-31K, is one of the components of the U11/U12 snRNPs of the U12 spliceosome, connecting both ends of the intron being processed [41]. The essential Arabidopsis gene AT3G10400 is the ortholog of human *ZCRB1*, encoding a U11/U12-31K protein, which acts as an RNA chaperone for proper splicing of U12-type introns [42]. The human and Arabidopsis U11/U12-31K proteins have 62% sequence similarity [43] and a similar structure, with an RRM domain followed by a ZCCHC (Fig. 3). Moreover, they both exclusively localize to the nucleus, as would be expected from their role in pre-mRNA splicing [41, 42]. In addition, the C-terminal half of the Arabidopsis AT3G10400 protein harbors a low-complexity region that is rich in glutamic acid (E); this region is absent from human ZCRB1. There is no yeast U11/U12-31K ortholog, since *Saccharomyces cerevisiae* lacks the entire U12 spliceosome.

Human SRSF7/ZCCHC20 is an SR factor involved in RNA export and decay

Members of the large and conserved SR family of RNA-binding proteins function in eukaryotic pre-mRNA metabolism and mRNA decay, export, and translation. SR proteins share a C-terminal domain, the RS domain (a low-complexity region rich in alternating S and R that in SR proteins is longer than 50 amino acids), and an RRM in their N-terminal region [44]. The RS domain confers flexibility to the SR proteins; the conformation of these proteins varies depending on the proteins or RNA molecules with which they interact [45, 46].

The best-studied role of SR proteins is in the regulation of constitutive and alternative splicing of pre-mRNAs. In constitutive splicing, SR proteins bind pre-mRNA exonic splicing enhancers through their RRM domain and recruit the spliceosome to nearby splice sites through their RS domain [47, 48]. SR factors antagonize hnRNPs, which repress alternative splicing and bind exonic and intronic splicing silencers [49].

Most SR proteins are nucleoplasmic, although some move from the nucleus to the cytoplasm, suggesting a role in mRNA export, translation control, and/or other cytoplasmic processes [46]. In the nucleus, SR proteins are usually found in multiple aggregates, termed speckles, and are frequently used as markers of these nuclear compartments. Human serine/arginine-rich splicing factor 7 (SRSF7), initially termed 9G8 and then ZCCHC20, contains an RRM domain followed by a single ZCCHC and an RS domain at its C-terminal half (Fig. 3). ZCCHC20 and SRp20 (also termed serine/arginine-rich splicing factor 3; SRSF3) are two SR proteins that show a highly dynamic behavior in their intracellular localization, shuttling between the nucleus and cytoplasm [50]. Both ZCCHC20 and SRp20 (which does not harbor a ZCCHC) bind RNA, suggesting, together with their dynamic localization, that they have a role in RNA transport. ZCCHC20 and SRp20 specifically bind to the intron-less mRNA transport element of the histone H2A, to promote its export to the cytoplasm [51]. The process of mRNA export requires the binding of any mRNA to adapter proteins that interact with receptors. For example, human mRNAs bound to ZCCHC20 and SRp20 selectively interact with nuclear RNA export factor 1 (NXF1), which is considered the predominant receptor for mRNA export from the nucleus to the cytoplasm [51].

The Arabidopsis genome encodes two subfamilies of SR factors with ZCCHCs

As with many other gene families, the number of SR family members is larger in plants (up to 22 in rice [*Oryza sativa*]) than in yeast (2) or humans (9). Only one of the human SR proteins, ZCCHC20, contains a ZCCHC [50]. The Arabidopsis genome encodes 19 SR proteins, belonging to six subfamilies, termed SR, RS, SC, SCL, RSZ, and RS2Z [52]. SR45 may represent an additional independent subfamily [53]. Only the members of the RSZ (RSZ21 [encoded by AT1G23860], RSZ22 [AT4G31580], and RSZ22A [AT2G24590]) and RS2Z (RS2Z32 [AT3G53500] and RSZ33 [AT2G37340]) subfamilies have one and two ZCCHCs, respectively [52, 53]. The ZCCHCs of human ZCCHC20 and its putative Arabidopsis ortholog, RSZ22, are involved in RNA recognition specificity [54, 55]. The structure of RSZ22 is very similar to that of human ZCCHC20 (Fig. 3); however, the function of Arabidopsis RSZ and RS2Z factors is not well known. Arabidopsis RSZ and RS2Z factors co-localize and accumulate in speckles that are distinct from those of other members of the SR family and splicing factors [53]. Some ZCCHC-containing proteins interact with SR factors of the SCL subfamily in yeast two-hybrid (Y2H) and co-immunoprecipitation assays [53, 55].

The localization of Arabidopsis RSZ22 is dynamic, as it moves between the nucleus and cytoplasm, and its ZCCHC is not required for its nuclear localization, but is required for it to interact with other splicing factors [56]. RSZ22 has been found in nuclear speckles and within the nucleolus, and mutations affecting its RRM or ZCCHC domains prevent it from translocating to the nucleolus.

A role for Arabidopsis RSZ33 in splicing has also been proposed based on its interactions with other putative splicing factors (RSZ22 and RSZ21) and the alterations in splicing some mRNAs, including its own, which cause its overexpression [57]. RSZ22 and RSZ21 interact with U1-70K, a component of U1 snRNP [58]. In a directed Y2H assay, RSZ33 interacted with CYCLIN-DEPENDENT PROTEIN KINASE G1 (CDKG1), which regulates the splicing of *CALLOSE SYNTHASE 5* (*CALS5*) pre-mRNA, encoding the

main synthase involved in the synthesis of the pollen callose wall [59]. As a consequence, loss of *CALS5* and *CDKG1* causes defects in pollen wall formation and male fertility. Based on these results, it has been proposed that splicing of *CALS5* pre-mRNA is facilitated by its indirect binding to CDKG1 through RSZ33 [59].

pNO40, a human ZCCHC-containing factor, plays a double role in splicing and ribosome biogenesis

The human nucleolar protein of 40 kDa (pNO40), also named ZCCHC17, is a multifunctional protein that participates in ribosome biogenesis and splicing. pNO40 has been found in the nucleolus and cytoplasm as part of the 60S ribosomal subunit [60, 61]. pNO40 has recently been identified as repressing transcription of the 47S rDNA, which encodes the 28S, 18S, and 5.8S rRNAs, by inhibiting the binding of the basal transcription factor upstream binding factor (UBF), to the 35S rDNA promotor [62]. pNO40 interacts with SR proteins to recruit them from nuclear speckles into the nucleolus, where they are retained. Sequestration of SR factors into the nucleolus perturbs RNA metabolism, including mRNA export, as shown in cells overexpressing pNO40, which display nuclear retention of poly(A)$^+$ RNAs and alterations of pre-mRNA splicing [63]. In addition to its ZCCHC located in the middle of the protein, human pNO40 harbors an S1 RNA-binding domain in its N-terminal half, and an extended basic domain (Lys-rich) in its C-terminal half (Fig. 3). No pNO40 ortholog has been found in yeast [61], and we did not identify any similar Arabidopsis proteins in databases or in our BLASTP analysis.

Metazoan LIN28 suppresses the biogenesis of the let-7 microRNA

Genetic screens have identified key genes controlling developmental timing in the nematode *Caenorhabditis elegans*, including *lin-4* and *let-7*, the first two genes identified in eukaryotes as encoding miRNAs (reviewed in [64]). miRNAs are small (about 22 nt), non-coding RNAs that post-transcriptionally repress gene expression. miRNAs are encoded by endogenous genes that have been found in plant, animal, and protist genomes, and inhibit the translation of their complementary mRNA targets. Mature, functional miRNAs are generated from longer precursors that form stem-loop structures; these stem-loops are processed by Dicer, a member of the RNase III family of endoribonucleases. In metazoans, transcription of the genes encoding miRNAs, the *MIR* genes, yields a primary precursor (pri-miRNA) that is processed in the nucleus by Drosha, another type III RNase, generating a shorter precursor (pre-miRNA) that is exported to the cytoplasm, where it is then processed by Dicer to form a duplex, one strand of which is the mature miRNA (reviewed in [65]).

The *lin-28* gene of *C. elegans* encodes an RNA-binding factor; the *lin-28* mRNA is targeted by the *lin-4* miRNA [31]. The LIN28 protein is highly conserved at the structural and functional levels in a variety of animals, from worms to humans. In invertebrates, *lin28* is a single-copy gene, whereas all vertebrates have two paralogs, *LIN28A* and *LIN28B* [66]. Human LIN28A (also termed ZCCHC1) and LIN28B block *let-7* maturation at different steps: LIN28A moves from the nucleus to the cytoplasm, where it blocks *pre-let-7* processing by Dicer; and LIN28B sequesters *pri-let-7* in the nucleolus, away from Drosha [67]. Metazoan LIN28 proteins are RNA-binding factors with two ZCCHCs and a nucleic acid-interacting domain, called the cold-shock domain (CSD; Fig. 3 and Supporting Dataset 1f) [68]. The CSD is about 70 amino acids long and includes two consensus RNA-binding domains. CSDs are found in many different proteins from a wide range of eubacteria and eukaryotes; these proteins bind RNA and single-stranded DNA, acting as RNA-chaperones [69]. CSD-containing proteins are structurally and functionally diverse, participating in multiple steps of RNA metabolism [70]. Human LIN28A and LIN28B are also termed COLD-SHOCK DOMAIN DNA-BINDING PROTEIN 1 (CSDD1) and CSDD2, respectively.

Human LIN28A binds to *pre-let-7* and circumvents maturation of this miRNA by recruiting terminal uridylyltransferase 4 (TUT4), which catalyzes *pre-let-7* 3′ uridylation and induces its degradation, thereby preventing *pre-let-7* processing by Dicer [64, 71]. TUT4, also termed ZCCHC11, and its paralog TUT7 (ZCCHC6), which plays a minor role in the oligouridylation of *pre-let-7*, are members of the non-canonical Poly(A) Polymerases (PAPs), belonging to the DNA polymerase β superfamily. TUT4 and TUT7 harbor three ZCCHCs. In addition to the ZCCHC, TUT4 and TUT7 also harbor a Matrin C_2H_2-type zinc finger, a nucleotidyltransferase (NTP_transf) domain, two Cid1 family PAP-associated domains, and Pro- and Gln-rich regions in TUT4, but a Glu-rich region in TUT7 (Fig. 3 and Supporting Dataset 1f). A specific interaction between the two ZCCHCs of mouse LIN28A and *pre-let-7* is necessary and sufficient to induce its oligouridylation by TUT4 [64]. Mouse LIN28A has been found in the periphery of the endoplasmic reticulum and suppresses endoplasmic reticulum-associated translation of mRNAs that form small hairpins containing AAGNNG, AAGNG, or UGUG sequences, which are the binding targets for LIN28A [72].

Our analyses on PROSITE and PatMatch did not identify any RNA uridylyltransferase containing a ZCCHC in Arabidopsis. However, Arabidopsis HEN1 SUPPRESSOR1 (HESO1, encoded by AT2G39740) and UTP:RNA URIDYLYLTRANSFERASE 1 (URT1, encoded by AT2G45620) have been shown to function redundantly in

miRNA 3′ uridylation, a process in which HESO1 plays the predominant role [73]. URT1 could be the functional ortholog of TUT7 or TUT4 because, in addition to their similar uridylation activity, it displayed the highest similarity to both factors in our BLASTP analysis: 35% and 33%, respectively. The reciprocal BLASTP, using the AT2G45620 protein to search in human databases, identified TUT7 and TUT4 as the most likely homologs. In addition, the structures of TUT4, TUT7, and URT1 are similar, sharing the NT (PF01909) and Cid1 family PAP (PF03828) domains, although URT1 lacks a ZCCHC.

The Arabidopsis genome encodes several Gly-rich proteins similar to LIN28

The similarity between GLYCINE-RICH PROTEIN2 (GRP2; encoded by AT4G38680), also named COLD SHOCK DOMAIN PROTEIN2 (CSDP2) or COLD SHOCK PROTEIN2 (CSP2), and human LIN28 has been previously described [66, 74]. CSP2 and its paralog CSP1 belong to the plant GRP superfamily of proteins, whose five families (I–V) have repeats of glycines with the G_nX structure. Family IV is composed of RNA-binding GRPs (RBGs), which have one or two RNA-binding domains in their N-terminus and a 30–70% G content in the C-terminus. Four RBG subclasses have been defined based on their structure (IVa–d). Members of the IVa, b, and d subclasses contain a single RRM, and members of the IVc subclass contain a CSD in their N-terminus [75, 76]. The Arabidopsis genome encodes 19 IVa–d RGBs, seven of which belong to the IVa subclass, three to IVb, four to IVc, and five to IVd [77]. In addition to an RNA-binding domain (either RRM or CSD), members of the IVb subclass, which are termed RNA-BINDING GLYCINE-RICH PROTEIN B1 (RBGB1) to RBGB3, and members of the IVc subclass, which are termed CSP1 to CSP4, have one (RBGB) or several (seven in CSP1 and CSP3, and two in CSP2) ZCCHCs in the Gly-rich region of their C-terminus (Fig. 3 and Supplementary Fig. 2) [76].

CSP1 and CSP2 associate with polyribosomes via mRNA, display mRNA-chaperone activity, and localize to the nucleolus and the cytoplasm [78, 79]. Nucleolar localization of CSP2 is dependent on its C-terminal GR/ZCCHC region. CSP2 and its closest paralog, CSP4/GRP2B (encoded by AT2G21060), which are partially redundant, negatively regulate seed germination and tolerance to freezing and salt stress [80]. CSP3 (encoded by AT2G17870) is required for abiotic stress tolerance, because its loss of function increases sensitivity to freezing, drought, and salt stress; its overexpression increases tolerance to these stresses. However, the molecular function of CSP3 related to stress or RNA metabolism is yet to be elucidated [81].

CSP3 seems to be a versatile protein whose interactions and subcellular localization suggest a role in ribosome biogenesis and RNA processing and stability. CSP3 has been found in the nucleolus, nucleoplasm, nuclear speckles, and cytoplasm. In the nucleolus and nucleoplasm, CSP3 interacts with NUCLEOLIN1 (NUC1), ribosome biogenesis factors, and 60S ribosomal subunit proteins. Arabidopsis NUC1 regulates the transcription of the 45S rDNA and the processing of the 45S pre-rRNA, the primary precursor of the 25S, 18S, and 5.8S rRNAs [82, 83]. CSP3 also interacts with several poly(A)-binding proteins in nuclear speckles, and with the DECAPPING5 PROTEIN5 (DCP5) protein in the cytoplasm [81, 84].

Little information is available on the Arabidopsis RBGB subfamily, which is composed of three members, RBGB1–RBGB3, and exhibits RNA-chaperone activity when expressed in *Escherichia coli*. Transcription of *RBGB2* (also termed *RZ-1A*) is induced by cold treatment but not by other types of abiotic stress or exogenous abscisic acid treatment, and its overexpression confers tolerance to freezing in Arabidopsis and to cold exposure in *E. coli* [85]. RBGB1 (RZ-1B) and RBGB3 (RZ-1C) localize in nuclear speckles. Simultaneous loss of function of *RZ-1B* and *RZ-1C* perturbs the splicing of many genes. Several aspects of the development of *rz-1b rz-1c* double mutants are delayed, including germination and flowering time, and their leaves are serrated. RZ-1B and RZ-1C interact in Y2H assays with some SR proteins of different subfamilies, such as RSZ21 and RSZ22, all of which contain ZCCHCs. In addition, RZ-1B and RZ-1C interact with themselves and with each other through their C-terminal domain, which is required for their localization in nuclear speckles [86].

ZCCHC-containing factors involved in pre-mRNA polyadenylation

In eukaryotic pre-mRNA maturation, which yields mature mRNAs competent for nuclear export and translation, 3′-end cleavage and polyadenylation are key steps. In mammals, about 20 factors constitute the core poly(A) machinery, which includes PAP and the conserved cleavage and polyadenylation specificity factor (CPSF) and cleavage stimulation factor (CstF) complexes, which are required for pre-mRNA cleavage, but not for polyadenylation. PAP is recruited by CPSF and catalyzes the addition of 200–250 adenosines to the 3′ end of the cleaved pre-mRNA in a template-independent fashion [87].

Yeast Mutant PCF11 extragenic suppressor 1 (Mpe1) is a component of the CPSF complex. Mpe1 is required for pre-mRNA cleavage and for polyadenylation at the *bona fide* site, as shown by examination of *mpe1* mutants [88]. Mpe1 is an E3 ligase with a RING finger domain at its C-terminus, a variant of a ubiquitin-like (UBL) domain in the

N-terminus, termed Domain With No Name (DWNN), and a ZCCHC in its central region (Fig. 2). Mutational analysis of these domains revealed that they are required for pre-mRNA polyadenylation, because the DWNN mediates protein–protein interactions within the CPSF complex, and the RING finger and the ZCCHCs mediate RNA binding [89].

The human homolog of yeast Mpe1 is the Retinoblastoma (Rb) binding protein 6 (RBBP6), also termed p53-associated cellular protein-testes derived (PACT). RBBP6 contains the same three domains as Mpe1, i.e., DWNN, ZCCHC, and RING finger. In addition, RBBP6 is four times larger than yeast Mpe1, and harbors an RS domain, followed by p53- and Rb-binding regions and a Lys-rich region in the C-terminus (Fig. 2). RBBP6 binds the p53 and Rb tumor suppressor factors, and thus interferes with the binding of p53 to DNA [90, 91]. Knockdown of *Rbbp6* in mice enhances the accumulation of p53 by reducing its polyubiquitination and causes embryo lethality. Consequently, loss of RBBP6 in mice increases p53-dependent gene transcription, revealing that RBBP6 is a negative regulator of p53 [92]. Knockdown of *RBBP6* in human cells with a small interfering RNA (siRNA) dramatically reduces the efficiency of 3′ pre-mRNA cleavage, causing a decrease in the abundance of mRNAs and an increase in the use of distal poly(A) sites. The decrease in mRNA levels is particularly evident for transcripts containing AU-rich elements (AREs) in their 3′-UTRs, which are naturally unstable transcripts targeted by the exosome. The DWNN is essential for binding to CstF, and, therefore, also essential for 3′ pre-mRNA cleavage, as shown by 3′ cleavage and polyadenylation assays in cells after knockdown of *RBBP6* with and without transgenes expressing different regions of the RBBP6 protein [93].

The Arabidopsis genes AT4G17410 and AT5G47430 encode putative orthologs of Mpe1 and RBBP6 (Fig. 2), both of which are annotated in the Aramemnon database as “Mpe1-like components of the CPSF complex” and in the TAIR database as “encoding a CCHC-type zinc finger protein with a DWNN domain”. We found AT4G17410 and AT5G47430 to be the only Arabidopsis genes annotated as encoding a protein with a DWNN. AT4G17410 has been named PARAQUAT TOLERANCE 3 (PQT3) and is a negative regulator of oxidative stress tolerance [94]. Its plant, fungal, and protist orthologs lack the p53- and Rb-binding and SR domains [95]. However, PQT3 and its paralog AT5G47430 (PQT3-like; PQT3L), which also contain the DWNN, ZCCHC, and RING finger domains, have a Gly- and an Arg-rich region in their C-terminal half (Fig. 2). The function of PQT3 and PQTL3L in pre-mRNA cleavage and polyadenylation is not known.

The human CPSF4 (also termed CPSF30) is also a component of the CPSF complex, acting as a co-factor of the cleavage and polyadenylation machinery. CPSF4 and its animal orthologs harbor a ZCCHC in their C-terminus (Fig. 3), which is absent from fungi and plants [96, 97]. In vitro RNA-binding assays revealed that its ZCCHC enhances poly(U)-binding ability. In addition, the ZCCHC alone produces the same enhancement of poly(U) binding as the other five zinc fingers together, which are C3H1-type ($CX_8CX_5CX_3H$) zinc fingers, and are present in all CPSF30/CPSF4 orthologs, including the Yeast 30 kDa homolog 1 (YTH1) [96].

Air proteins are key factors in RNA degradation by the exosome

The yeast Arginine methyltransferase-interacting RING finger protein 1 (Air1) was identified in a screen based on the Y2H assay for structures that interact with Histone methyltransferase 1 (Hmt1), the major arginine methyltransferase in yeast. Many RNA-binding proteins undergo arginine methylation by Hmt1, including histones and hnRNPs [98]. Arginine methylation by yeast Hmt1 and its human counterpart occurs in the Arg/Gly-rich C-terminal domain, termed the RGG box, of hnRNPs [99]. RGG/RG domains are encoded by all eukaryotic genomes and are the second most abundant among human RNA-binding domains [100]. RGG domains are intrinsically disordered, and mediate degenerate specificity in RNA binding. RGGs are found as single RNA-binding domains or combined with more structured domains, such as the RRM- or KH-type domains [101].

The eukaryotic exosome is a multiprotein complex with 3′ to 5′ exonucleolytic activity that processes and/or degrades normal and aberrant RNA species, and the by-products of their maturation, in both the nucleus and the cytoplasm. To be catalytically active and to recruit its RNA substrates, the exosome is assisted by cofactors. These cofactors include protein complexes that act as activators, RNA helicases, RNA-binding proteins that act as adaptors, including ZCCHC proteins, and other proteins that do not fit in any of these categories. In yeast, the Trf4/5-Air1/2-Mtr4 polyadenylation (TRAMP) and Superkiller (Ski) complexes are the best-characterized exosome cofactors, which recruit the exosome to its RNA targets in the nucleus and cytoplasm, respectively. The TRAMP and Ski complexes are required for RNA surveillance. TRAMP is also required for maturation of snRNAs and rRNAs, while the Ski complex also participates in the turnover of functional mRNAs [102].

The yeast TRAMP complex is composed of the non-canonical Trf4/5 PAP, RNA helicase Mtr4 (mRNA transport 4), and RNA-binding proteins Air1 and Air2 [103]. The number of ZCCHCs in Air1 and Air2 is not clear. These two proteins have four $CX_2CX_4HX_4C$ sequences (Supporting Dataset 1b), but they have been described as harboring five ZCCHCs [104, 105]. However, their second ZCCHC is not canonical, since there is an additional residue between the second C and the H ($CX_2CX_5HX_4C$). In agreement with

the four $CX_2CX_4HX_4C$ sequences present in both proteins, four ZCCHCs are described for Air1 in UniProtKB, but only three are annotated for Air2; the $CX_2CX_5HX_4C$ sequence is not considered a ZCCHC in either case.

Air1 and Air2 bind Trf4 and Trf5 through the ZCCHCs to recognize RNA substrates of TRAMP. The Trf4 subunit of TRAMP adds a short poly(A) tail to cryptic unstable transcripts (CUTs). It can also polyadenylate many different improperly processed or aberrant non-coding RNAs (ncRNAs) that participate in splicing [small nuclear RNAs (snRNAs)] or in ribosome biogenesis and translation [transfer RNAs (tRNAs), small nucleolar RNAs (snoRNAs), and rRNAs]. Polyadenylation of these TRAMP targets stimulates their degradation by the nuclear exosome. Mutating *AIR1* and *AIR2*, which encode proteins with a substitution in the second C of the third and fourth ZCCHC, revealed that the second C in these motifs is critical for Air function in vivo [104].

Co-immunoprecipitation analysis coupled with mass spectrometry allowed for the discovery of two different MTR4-containing human complexes, one of which is restricted to the nucleoplasm and the second to the nucleolus. The nuclear exosome-targeting (NEXT) complex is composed of MTR4, RNA Binding Motif Protein 7 (RBM7), and ZCCHC8, which possesses a single ZCCHC that can potentially interact with RNA. Based on its dual interaction with the exosome and the spliceosome, an additional role has been proposed for the NEXT complex in the recruitment of the exosome for decay and/or processing of intron-encoded snoRNAs [106].

ZCCHC7 is the putative human ortholog of yeast Air1 and Air2, and co-purified with the nucleolar fraction of MTR4 and the exonuclease RRP6 (also known as EXOSC10). RRP6 is a catalytic subunit of the RNA exosome that is highly conserved in eukaryotes and has been found in the nucleolus and the nucleoplasm. RRP6 is required for proper maturation of ncRNA substrates, such as rRNAs, snRNAs, snoRNAs, and CUTs (reviewed in [107]). Three RRP6-like proteins (RRP6L1 to RRP6L3) have been found in Arabidopsis: RRP6L1 in the nucleus and in the nucleolar vacuole, RRP6L2 mainly in the nucleolus, and RRP6L3 in the cytoplasm, but its functional relationship with the exosome has not been determined [108].

Loss of function of ZCCHC7 causes the accumulation of by-products excised from the 5′-ETS of 47S pre-rRNA processing [109]. ZCCHC7 is a nucleolar protein, consistent with its role in rRNA biogenesis. ZCCHC7 has four ZCCHCs (Fig. 2), which mediate the interaction with PAP-associated domain-containing 5 (PAPD5) and PAPD7, the human orthologs of yeast TRF4-2 and TRF4-1, respectively, suggesting the existence of a human nucleolar TRAMP-like complex [110]. Using BLASTP, we found several Arabidopsis ZCCHC-containing proteins of unknown function with limited sequence similarity to human ZCCHC7, no one of which was a credible ZCCHC7 homolog.

AT1G67210 and AT5G38600 are annotated in the TAIR10 database as encoding proline-rich spliceosome-associated family protein/zinc knuckle (CCHC-type) family proteins, and are proposed to be ZCCHC8 homologs [111, 112] with structures similar to that of human ZCCHC8, harboring a single ZCCHC. They have also been identified in co-immunoprecipitation assays using HUA ENHANCER 2 (HEN2) as bait [112]. Arabidopsis HEN2 and MTR4 are the two Arabidopsis co-orthologs of the yeast Mtr4 helicase; HEN2 is nucleoplasmic and MTR4 is mainly nucleolar, as shown by their hybrid proteins with GFP [111, 112]. Both of these helicases associate with core components of the nuclear exosome, but most proteins that co-precipitate with MTR4-GFP are ribosomal processing factors that do not co-purify with HEN2-GFP. Plants lacking HEN2 activity accumulate many types of polyadenylated nuclear RNA species, including unspliced pre-mRNAs or misspliced mRNAs, excised introns, and by-products of the processing of ncRNAs. Based on these results, the existence of two specialized RNA helicases, acting as cofactors of the exosome, has been proposed in Arabidopsis: MTR4 assists in the nucleolus to the exosome in the maturation of rRNAs, and HEN2 in the degradation of many different RNA nuclear substrates. The functions of AT1G67210 and AT5G38600 have not yet been determined, but both may be part of two different NEXT-like complexes, together with HEN2 and RBM7, and participate in the degradation and/or processing of nuclear exosome RNA substrates [112].

The yeast Gis2 and human CNBP and RBM4 proteins promote cap-independent translation

Expansion of CCTG repeats in the first intron of the human gene encoding the CCHC-type zinc finger nucleic acid-binding protein (CNBP), also termed Zinc finger 9 (ZNF9) and ZCCHC22, causes the autosomal dominant form of Myotonic dystrophy type 2 [113]. Human CNBP was the first cellular eukaryotic ZCCHC-containing protein described, and was found to bind to the conserved Sterol Regulatory Element present in promoters of genes encoding enzymes of the cholesterol biosynthetic pathway [114]. CNBP was later identified as an RNA-binding protein in *Xenopus laevis* by its ability to bind the 5′-UTR of mRNAs encoding ribosomal proteins, which suggests a role in their translational control [115, 116]. CNBP orthologs, including yeast Glucose inhibition of gluconeogenic growth suppressor 2 protein (Gis2), harbor seven ZCCHCs (Fig. 2), and interact with ribosomal proteins and the translating ribosome. Human CNBP and yeast Gis2 are located in stress granules [117], which

are cytoplasmic compartments containing translationally repressed mRNAs. Stress granules form when translation initiation is impaired and function in translational repression and/or mRNA decay [118]. Human CNBP and yeast Gis2 are also considered translational activators because they promote cap-independent translation through interactions with the 5′-terminal oligopyrimidine (5′ TOP) tract of mRNAs and the translating ribosome [119, 120]. Many eukaryotic 5′ TOP mRNAs are transcribed from genes that encode factors involved in translation, including ribosomal proteins and ribosomal biogenesis factors [121]. Human CNBP enhances global transcription and translation by unfolding DNA and RNA G-quadruplex (G4) structures, located in the promoters of genes or 5′-UTRs of mRNAs, respectively [122, 123]. Although CNBP orthologs have been identified in fungi and metazoans, such as *Drosophila melanogaster* [124], no CNBP or Gis2 orthologs have been described in plants. Our BLASTP searches of the yeast Gis2 and human CNBP sequence against the TAIR10 protein database yielded CSP1 and CSP3 as the most similar factors in the Arabidopsis proteome.

RNA-binding domain protein 4A (RBM4A), also named ZCCHC21, and RBM4B (ZCCH15) are the human orthologs of *Drosophila melanogaster* LARK (low-complexity amyloid-like reversible kinked segments); LARK and the two human RBM4 factors harbor two RRM domains and a ZCCHC (Fig. 3). The RRMs and C-terminus of LARK are required for its splicing function. Deletion analysis of LARK suggests that its ZCCHC and second RRM domain act in concert in translational regulation [125]. RBM4A is a dynamic protein that shuttles between the nucleus and cytoplasm, and is involved in controlling alternative pre-mRNA splicing and translational repression. In the cytoplasm, RBM4A suppresses cap-dependent translation in unstressed conditions, but under stress, it promotes translation mediated by IRES (internal ribosome entry sites) of stress-responsive genes. In the nucleus, RBM4B is localized in nuclear speckles and the nucleolus, where it might play a role in rRNA biogenesis, since it interacts with the 40 kDa subunit of RNA polymerase I (RNAP I) in Y2H assays (reviewed in [126]). Similar to human CNBP, metazoan LARK/RBM4 proteins have been found to be DNA G4-binding proteins [127]. There is no *Drosophila melanogaster* LARK homolog described in Arabidopsis, and our searches for Arabidopsis homologs of the yeast LARK protein and of its human orthologs did not yield similar factors.

XRNs contribute to RNA turnover and post-transcriptional gene silencing

Eukaryotic 5′→3′ exoribonucleases (XRNs) contribute to RNA turnover and maintenance of RNA homeostasis, which involves degradation of a wide range of normal and defective mRNAs. XRNs also participate in the maturation of many species of ncRNAs, such as rRNAs, tRNAs, and snoRNAs. XRNs have been found in the nucleoplasm, the nucleolus, and the cytoplasm. XRNs were first identified in yeast, which has two of these enzymes: Xrn1 (also named Pacman) and Xrn2 (also named ribonucleic acid trafficking, Rat1), which act in the cytoplasm and nucleus, respectively. Yeast Xrn1 has been found predominantly in the cytoplasm, but also in the nucleus, and it has been implicated in the degradation of decapped or cleaved mRNAs, and with the processing of ncRNAs, such as snoRNAs or tRNAs [128, 129]. More recently, Xrn1 has been shown to act in the nucleus as an activator of the transcription of genes encoding unstable transcripts, many of which encode proteins involved in ribosome biogenesis and mRNA translation [130]. Xrn2/Rat1 is required for the termination of transcription by RNAP I and II [131, 132], and for the maturation of snoRNAs and rRNAs [133, 134].

The human and Arabidopsis proteomes lack an ortholog of yeast Xrn1, but they contain one and three homologs of yeast Xrn2/Rat1, respectively, all of which have a ZCCHC, unlike yeast XRNs. In addition to the ZCCHC, all XRNs, including yeast Xrn2/Rat1, harbor a 5′→3′ exonuclease domain in their N-terminus (Fig. 3). Human XRN2 is involved in the termination of transcription by RNA polymerase II, and it degrades nascent RNA downstream of the 3′ cleavage site [135]. Human XRN2 is also involved in the suppression of replication stress and maintenance of genomic stability [136]. The three Xrn2/Rat1 orthologs of Arabidopsis are XRN2 (encoded by AT5G42540), XRN3 (AT1G75660), and XRN4 (AT1G54490). Arabidopsis XRN2 and XRN3 are nucleolar and nucleoplasmic factors, and XRN4 is cytoplasmic [137]. XRN2 and XRN3 act in 45S pre-rRNA processing [138], and as endogenous RNA silencing suppressors, and are required for the degradation of excised miRNA loops produced during miRNA maturation. XRN4 localizes in the processing bodies with the decapping enzymes DCP1 and DCP2, all of which are involved in mRNA decay. Plants without XRN4 activity over-accumulate 3′ fragments of RNAs, including targets of miRNAs [139]. XRN4 is a suppressor of post-transcriptional gene silencing, because it degrades decapped mRNAs, preventing them from becoming templates for RNA-dependent RNA polymerases, thus producing small interfering RNAs that would enter into the post-transcriptional gene silencing pathway [140].

Human ZCCHC4 and ZCCHC9 are involved in rRNA biogenesis

Human ZCCHC4 is a nucleolar and cytoplasmic RNA m^6A methyltransferase involved in the methylation of the 28S

rRNA in the cytoplasm. The 28S rRNA is a component of the 60S ribosomal subunit, and the loss of ZCCHC4 reduces 60S ribosomal subunit levels and global translation, although it has no effect on mature 28S and 18S rRNA production [141]. ZCCHC4 harbors a ZCCHC and a zinc finger GRF-type, containing three conserved glycine (G), R, and phenylalanine (F) residues in the center of the domain (Fig. 3). ZCCHC4 does not have orthologs in yeast, and even though it has been claimed to be conserved in multicellular organisms, ZCCHC4 orthologs have only been analyzed in animals [141]. We did not find any Arabidopsis protein with enough similarity to be considered a credible ortholog of ZCCHC4.

ZCCHC9 is a nucleoplasmic and nucleolar protein [142]. Like Air1, Air2, and ZCCHC7 (but not ZCCHC8), ZCCHC9 harbors four ZCCHCs (Fig. 3), and its similarity to these proteins has been pointed out; however, ZCCHC9 is considered a protein without a yeast homolog [143]. ZCCHC9 is assumed to be a ribosome biogenesis factor because its knockdown causes high over-accumulation of 21S pre-rRNA, a precursor of 18S rRNA [144]. AT5G52380 encodes the Arabidopsis protein most similar to human ZCCHC9, according to the Aramemnon database and our own BLASTP analysis, but it has not been studied. The reciprocal BLASTP analysis, searching with the AT5G52380 protein in human databases, also shows that ZCCHC9 is the human protein most similar to Arabidopsis AT5G52380. This Arabidopsis protein harbors five $CX_2CX_4HX_4C$ sequences (Supporting Dataset 1d) recognized by the SMART program as ZCCHCs, but only three of them were predicted with PROSITE (Fig. 3).

Other Arabidopsis-specific ZCCHC-containing factors that are also involved in RNA metabolism

Arabidopsis CYCLOPHILIN59 (CYP59; encoded by AT1G53720) and its orthologs, Silencer of Germline 7 (SIG-7) of *Caenorhabditis elegans* and RRM-containing cyclophilin regulating transcription (Rct1) of *Schizosaccharomyces pombe*, participate in regulating the recruitment of pre-mRNA processing factors to RNAP II. They interact with the C-terminal domain (CTD) of the large subunit of RNAP II, which is composed of tandem repeats of the conserved heptapeptide YSPTSPS, and whose S residues are dynamically phosphorylated and dephosphorylated. The phosphorylation status of the CTD determines its ability to recruit diverse factors, including splicing factors, to RNAP II. Similar to the histone code of protein modifications that affect chromatin structure and gene expression, the existence of a CTD code that affects the recruitment of factors to RNAP II has been postulated. The CTD must be unphosphorylated to initiate transcription, but phosphorylation of S5 and S2 is required for transcriptional elongation, and phosphorylation of S2 is required for termination [145]. CYP59 and its orthologs act as peptidyl isomerases for the *trans*-to-*cis* isomerization of P residues present in the CTD heptapeptide repeats. Arabidopsis CYP59 is an RNA-binding protein that interacts with SR factors in Y2H and pull-down assays, as well as with the CTD of RNAP II, and localizes in a punctuate pattern next to splicing speckles where most nuclear SR factors are found, as well as in the nucleolus. In plants overexpressing CYP59, phosphorylation of the RNAP II CTD is reduced [146]. Arabidopsis CYP59 and its plant orthologs harbor a ZCCHC between a RRM domain and a low-complexity region in its C-terminal region (Fig. 4). ZCCHCs are absent in their yeast and animal orthologs.

Two Arabidopsis ZCCHC-containing factors, encoded by AT5G49930 and AT5G26742, were isolated because of the effects of their loss of function, which cause embryonic lethality, so they are termed EMBRYO DEFECTIVE 1441 (EMB1441) and EMB1138, respectively [147]. AT5G49930 is annotated in the Aramemnon database as "putative (yeast RQC2)-like component of ribosome-associated quality control complex", and as the ortholog of the human Nuclear Export Mediator Factor (NEMF) in HomoloGene. NEMF is a component of the ribosome quality-control complex that facilitates the recognition and ubiquitination of stalled 60S subunits [148]. As described in HomoloGene, all NEMF orthologs in animals, plants, and yeast share two domains of unknown function, DUF814 and DUF3441, but only plant orthologs (Arabidopsis and rice) harbor a ZCCHC.

EMB1138 (AT5G26742) or RH3 and its maize ortholog are DEAD-box RNA helicases with RNA-chaperone activity that participate in chloroplastic splicing and could also be involved in the assembly of 50S ribosomal subunits in the chloroplast [149, 150]. RH3/EMB1138 is described in the HomoloGene database as the ortholog of human DExD-box helicase 50 (DDX50), which localizes in nucleoli and nuclear speckles and participates in antiviral responses [151], but only plant orthologs seem to harbor a ZCCHC (Fig. 4). RH3 is the only chloroplastic ZCCHC-containing protein that we found.

AT3G55340 encodes PHRAGMOPLASTIN-INTERACTING PROTEIN 1 (PHIP1), a plant-specific protein that seems to be required for RNA cytoplasmic localization and polarized mRNA transport, which involves the movement of ribonucleoprotein complexes along the microtubular cytoskeleton [152]. PHIP1 contains three ZCCHCs and two RRM domains (Fig. 4).

ZCCHC-containing proteins that are not associated with RNA metabolism

Yeast Bilateral karyogamy defect 1 (Bik1) and its human ortholog, CAP-Gly domain-containing linker protein 1 (CLIP1, better known as CLIP-170), play critical roles in the regulation of microtubule polymerization, stabilization, and dynamics. Both factors harbor a ZCCHC in their C-terminal regions [153] and are cytoplasmic (Fig. 2 and Supporting Dataset 1b and 1f). Our results indicate that Bik1 is the only yeast ZCCHC-containing factor with a known function that is not involved in RNA metabolism. The ZCCHC described for Bik1 is not recognized by the ScanProsite or SMART programs (Supporting Dataset 1b), but it has been found to be involved in protein–protein interactions [154]. Two ZCCHCs have been described in human CLIP-170 (Fig. 2 and Supporting Dataset 1f). The second ZCCHC of CLIP-170, which is absent from Bik1, contains an additional amino acid between the second C and the H ($CX_2C\underline{X}_5HX_4C$). This atypical domain is functional, as shown by site-directed mutagenesis [155]. We have not found credible Bik1 or CLIP-170 orthologs using BLASTP searches, but several Arabidopsis factors have been proposed, based on amino acid composition instead of sequence identity [156].

We found two human ZCCHC-containing factors involved in innate immune responses: DDX41 and ZCCHC3. DDX41 is an intracellular sensor of pathogenic double-stranded DNA that activates the immune response [157]. Arabidopsis RH35 (encoded by AT5G51280) and its paralog RH43 (AT4G33370) are described in HomoloGene as orthologs of human DDX41; RH35 seems to be nuclear, and RH43 cytoplasmic, but they are not well studied. These four orthologs harbor a ZCCHC in their C-terminal regions, as described in UniProtKB and HomoloGene databases (Fig. 3 and Supporting Dataset 1d and 1f). Human ZCCHC3 is considered a co-sensor for cytosolic viral RNA and DNA [158, 159]. No other domains or motifs that three ZCCHCs located in its C-terminal region are described for ZCCHC3 in UniProtKB (Fig. 3).

Six of the human ZCCHC-containing proteins are encoded by long terminal repeat (LTR) retrotransposon-derived genes, belonging to two different families: Mammalian retrotransposon-derived (Mart) and the Paraneoplastic Ma Antigen (PNMA) (Supporting Dataset 1f). Members of these families are classified as domesticated Gag proteins, encoded by neofunctionalized retrotransposons whose genes are present in mammalian genomes and seem to have been domesticated during evolution [160]. The human genome encodes 11 *Mart* genes with deletions affecting the regions encoding reverse transcriptase integrase activity, but only three (PEG10 [Paternally expressed gene 10 protein], ZCCHC5, and ZCCHC16) harbor a ZCCHC [161]. The human PNMA family has 19 members, three of which harbor a ZCCHC in their C-terminal region with highly similar sequences: the paraneoplastic antigen Ma3 (PNMA3), PNMA7A (ZCCHC12), and PNMA7B (ZCCHC18). These PNMA proteins have been associated with Paraneoplastic Neuronal Disorder (PND), a non-metastatic complication of cancer; they seem to play an oncogenic role [162]. ZCCHC10 has recently also been related to cancer, with a suppressor effect instead of an oncogenic effect. ZCCHC10 binds and hence stabilizes p53 by inhibiting its interaction with murine double minute 2 (MDM2), an E3 ubiquitin-protein ligase that ubiquitinates p53 for proteasomal degradation [163].

Some other Arabidopsis factors with ZCCHCs have not been associated with RNA metabolism, such as DNA DAMAGE-BINDING PROTEIN 2 (DDB2; AT5G58760), which is involved, together with ARGONAUTE1 (AGO1), in the repair of UV-induced lesions. Arabidopsis AGO1 is the main effector of post-transcriptional gene silencing mediated by miRNAs, other endogenous siRNAs, and viral siRNAs [164]. DDB2 and AGO1 form a chromatin-bound complex, and photoproduct-derived siRNAs facilitate the recognition of UV-damaged DNA for repair [165]. REPLICATION PROTEIN A 1C (RPA1C) and RPA1E play a redundant role in DNA replication, repair, and meiosis (crossing over and repair); all plant RPA1C orthologs harbor at least one ZCCHC that is also present in all RPA1E orthologs in the Brassicaceae [166]. The ZCCHC of TOP3-alpha (TOP3α) is required for meiotic recombination and chromosome integrity, preventing extra crossovers [167]. TANDEM ZINC KNUCKLE PROTEIN (TZP; AT5G43630) is a nuclear protein that was identified as a quantitative trait locus negatively controlling morning-specific growth in Arabidopsis [168]. TZP is required for nuclear phosphorylation of phytochrome A (phyA), the primary photoreceptor for far-red light. TZP interacts with phyA through its Plus3 domain [169]. We found five genes encoding transcription factors of the MYB family (AT1G70000, AT3G16350, AT5G47390, AT5G56840, and AT5G61620), one of them with an uncertain ZCCHC, and only two of which have already been studied. MYB-LIKE DOMAIN (MYBD; AT1G70000) is involved in the epigenetic repression of anthocyanin biogenesis [170]. KUA1 (encoded by AT5G47390), also named MYB HYPOCOTYL ELONGATION-RELATED (MYBH), was first identified as a regulator of skotomorphogenesis [171], and later as a circadian clock-regulated gene that promotes leaf cell expansion, controlling the homeostasis of reactive oxygen species [172].

Another Arabidopsis factor that has been studied to some extent is CAX-INTERACTING PROTEIN 4 (CXIP4; AT2G28910), a nuclear and cytoplasmic protein that was initially identified, because its expression in yeast activates

the H^+/Ca^{2+} antiporter CAX1 [173]. However, CXIP4 was later found to be an interactor of MORPHOLOGY OF ARGONAUTE1-52 SUPPRESSED 2 (MAS2) in a screen based on the Y2H assay. MAS2 is the presumed ortholog of human NF-kappa-B-activating protein, and seems to be a key player in the regulation of rRNA synthesis that also interacts with splicing and ribosomal biogenesis factors, so CXIP4 may be involved in RNA metabolism [174].

ZCCHC sequences are conserved among orthologs

A consensus sequence has been proposed for ZCCHCs, based on the alignment of retroviral and cellular ZCCHCs: X_2-C-ϕ-ψ-C-G-ψ-X-G-H-ω-(A/S)-(+)-(−)-C-P-X, where ϕ is an aromatic amino acid, ψ is a polar or charged residue able to form hydrogen bonds, ω is a hydrophobic residue, and (+) and (−) are positively and negatively charged amino acids [175]. To obtain a consensus sequence, we considered factors found in the three organisms studied here (110 factors, excluding the retrotransposon-related factors and proteins with uncertain ZCCHCs) and aligned their 198 putative ZCCHCs in two ways, either all together or grouped by species (20 from yeast, 121 from Arabidopsis, and 57 human). We then generated logos from the multiple alignments obtained, using the WebLogo program (https://weblogo.threeplusone.com/create.cgi), resulting in similar findings for the three organisms (Fig. 5) as those previously described by other authors [175] and the corresponding PROSITE profile. The PROSITE database is composed of entries describing protein domains and profiles to identify them [176–178]. PROSITE profiles provide the corresponding alignment and logo for any group of proteins that share a given domain. However, the analysis of ZCCHCs (https://prosite.expasy.org/PDOC50158) may be biased, since it includes ZCCHCs from 260 retrotransposons and 285 cellular proteins. These sequences seem to have been chosen irrespectively of their phylogenetic distances, with mammalian factors being over-represented, including orthologs of very closely related species, and those of plants being under-represented.

The results of our multiple alignments showed that in 160 (81%) sequences, G was at the 10th position, suggesting that the $CX_2CX_3GHX_4C$ is the ancestral sequence of ZCCHCs (Fig. 5 and Supplementary Figs. 1–4). In fact, in the InterPro database, ZCCHCs have two different accession numbers, IPR025836 for the $CX_2CX_4HX_4C$ sequence, and IPR025829 for $CX_2CX_3GHX_4C$. We also found G in the 7th position in 140 (71%) of the sequences, and R at the 14th position in 51 of the 121 sequences of Arabidopsis (42%), an enrichment that was not found for yeast or human ZCCHCs (Supplemental Fig. 4), suggesting a lineage-specific evolution of the ZCCHCs in plants.

Tryptophan (W) displayed the most structural constraints, because it only occurred once at the 2nd, 4th, and 17th positions, but was found nineteen times at the 12th position. Hence, sequences in which any position is represented only once could be false positives or ZCCHCs with peculiar features. For example, we only found a ZCCHC-containing yeast factor with a W at its 2nd position: Bik1, the only ZCCHC-containing yeast factor that is not nuclear or involved in RNA metabolism. Its putative human ortholog, CLIP-170, harbors a tyrosine (Y), which is also an aromatic amino acid, instead of a W at the 2nd position of the ZCCHC. Therefore, the presence of these aromatic amino acids at the second position of the ZCCHCs could be necessary for them to interact with proteins instead with RNA. The AT1G75560 and AT4G00980 Arabidopsis proteins have W in the 4th and 17th positions, respectively, and both are annotated as "Zinc knuckle (CCHC-type) family protein" but have not been studied. PROSITE and SMART predicted the $CX_2CX_4HX_4C$ sequences present in these factors to be ZCCHCs (Supporting Dataset 1d). No human protein contained a W other than that of the 12th position of the ZCCHC.

To ascertain if there is conservation in the sequences of the ZCCHCs among functionally related factors, we constructed phylogenetic trees with the 198 ZCCHCs that we considered genuine (20 of yeast, 121 of Arabidopsis, and 57 of human ZCCHC proteins). Then, we focused our analysis on factors harboring a single ZCCHC, because the ZCCHCs of a specific protein that harbor more than one, grouped with the ZCCHCs of different proteins (Fig. 6), making it difficult to draw conclusions. We observed that most paralogs were grouped in clades, and that several clades included orthologs or co-orthologs. This was the case for human DDX41 and Arabidopsis RH34 and RH35; the yeast Mpe1, human RBBP6, and Arabidopsis PQT3 and PQT3L; the Slu7 orthologs of the three organisms; yeast Bik1 and human CLIP-170, each of which clustered with two Arabidopsis unknown proteins; and human and Arabidopsis XRN factors and U11/U12-31K spliceosomal proteins. However, we did not observe clusters among ZCCHCs from other proteins involved in the same general process (e.g., splicing, translation, and polyadenylation), except one with human SF1 and pNO40, which participate directly (SF1) or indirectly (pNO40) in pre-mRNA splicing. We also found an unexpected cluster formed by human SRSF7 and CPSF4 factors. SRSF7 is an SR factor involved in RNA export, but CPSF4 is a co-factor of the cleavage and polyadenylation machinery. However, the structures of the members of these two pairs of proteins are very different, with different domains and ZCCHC positions (Figs. 2, 3).

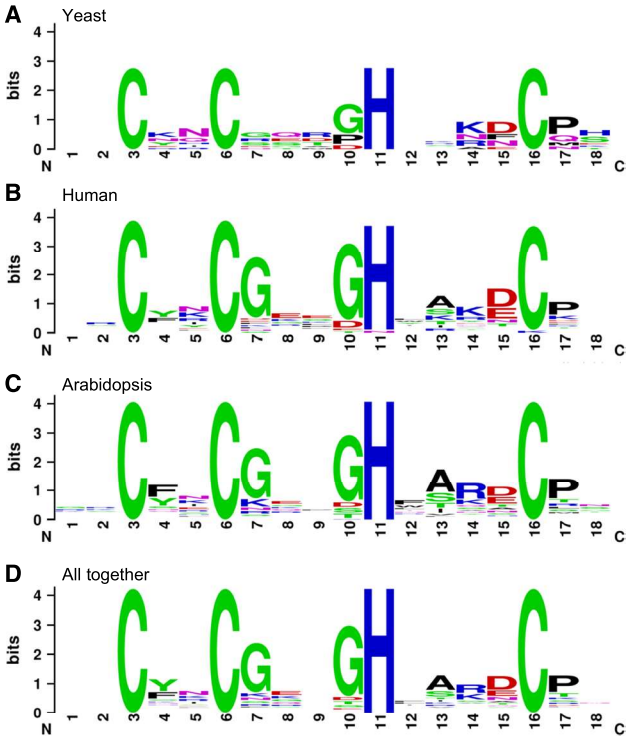


Fig. 5 Logos representing normalized amino acid frequencies obtained from the multiple alignment of the 18 amino acids that constitute the ZCCHCs of yeast, human, and Arabidopsis proteins. Sequences were taken grouped by species (**a–c**) or all together (**d**) to generate the corresponding logos from multiple alignments, using the WebLogo version 3 software (https://weblogo.threeplusone.com/create.cgi). Uncertain and retrotransposon-related factors were not taken into account, and the resulting 198 sequences were aligned, all together or grouped by species

In addition, we found the ZCCHC of AT2G15180, an unknown protein annotated as "zinc knuckle (CCHC-type) family protein", clearly grouped with the two human RBM4 factors. In our BLASTP searches, however, we did not find any Arabidopsis protein similar to RBM4 factors, as previously mentioned. RBM4 factors harbor two RRM domains (Fig. 3), but none is annotated for AT2G15180, suggesting that is not their Arabidopsis functional ortholog. We also found a cluster formed by the plant-specific CXIP4 and human SREK1-interacting protein 1 (SREK1IP1), also known as p18 splicing regulatory protein (P18SRP), whose function is unknown, although it is known to interact with the SR splicing regulatory protein SRrp86 [179]. Only the ZCCHC is predicted in both proteins, in what seem to be similar regions, which may not be sufficient to consider them structural orthologs (Figs. 3, 4). The finding of similar $CX_2CX_4HX_4C$ sequences in proteins without any other structural homology may have not biological relevance.

Conclusions and perspectives

In this work, we compiled information available in public databases and the literature on ZCCHC-containing factors in yeast, Arabidopsis, and humans, which we considered representative of the Fungi, Plantae, and Animalia kingdoms, respectively. No representative of the Protista kingdom was considered because of the scarcity of functional genomics information available. We used PatMatch (for yeast and Arabidopsis) and ScanProsite (for human) to search for proteins containing the $CX_2CX_4HX_4C$ sequence, whose primary structures were later analyzed on UniProtKB, a database that identifies protein domains, motifs and low-complexity regions by compiling the information obtained from diverse programs, such as PROSITE, and SMART.

We found a wide array of combinations of ZCCHCs within proteins; in most cases, each protein had only a single copy, but there could be as many as eight. Moreover, they were present in any region of the protein, and could be alone or combined with other domains or motifs (Figs. 2, 3, 4). In addition, we found that many ZCCHC-containing proteins harbor low-complexity regions, which are mainly rich in P, L, G, R, or S, including RS domains that we found in proteins related to pre-mRNA splicing. Proteins with low-complexity regions are likely to have more interactors than those lacking these regions [180]; therefore, low-complexity regions could cooperate with ZCCHCs for RNA or protein binding, or have their own biological functions. Constructs expressing human CNBP recombinant proteins lacking the RG-rich region (Arg/Gly-rich) located between their first and second ZCCHC (Fig. 2) are required for high-affinity RNA binding. The proline-rich segment of ZCCHC8 (Fig. 2) is the site of interaction with the RRM domain of RBM7 in the NEXT complex of the exosome [106].

Our results obtained from PROSITE and SMART did not always coincide. For example, PROSITE predicted only three ZCCHCs in yeast Air2, but SMART predicted five, and we found four $CX_2CX_4HX_4C$ sequences within this protein (Supplementary Fig. 1 and Supporting Dataset 1a, 1b). Air2 is also considered by some studies to harbor five ZCCHCs, but the sequence of the second one is $CX_2CX_5HX_4C$ instead of $CX_2CX_4HX_4C$, and its deletion does not impair its function [104]. A similar case is that of Retinitis pigmentosa 9 (RP9), which is annotated as a ZCCHC-containing factor in UniProtKB, but also harbors a CCHC sequence with a different spacing ($CX_2CX_4HX_6C$). We also found in the literature that some variants of the canonical ZCCHC seem to behave as true zinc knuckles, but they are not annotated in UniProtKB, nor have they been predicted with any other program available. This is the case with the human RecQ4 helicase, which harbors an atypical ZCCHC in which the second C is substituted with an Asparagine (CNHC), while

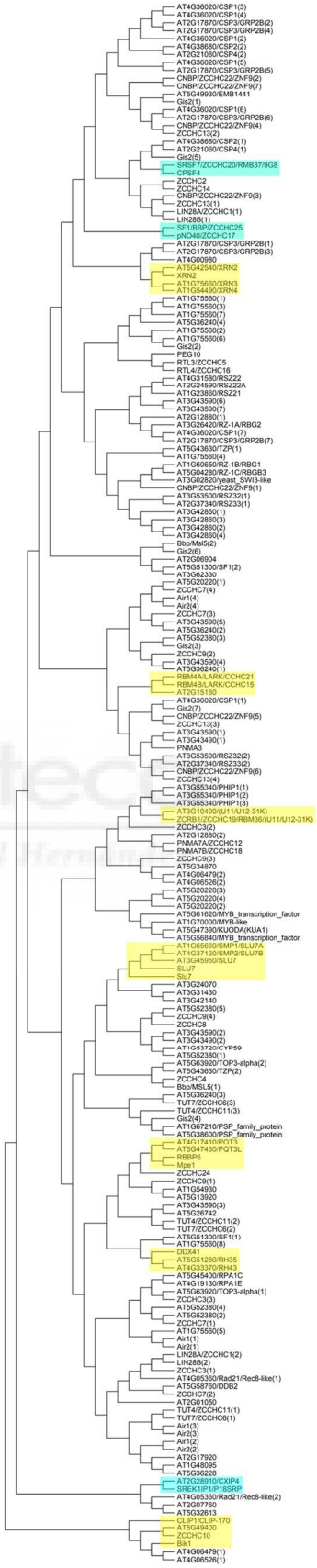


Fig. 6 Phylogeny of the ZCCHCs of yeast, Arabidopsis, and human proteins. The unrooted tree was obtained with the MEGA X software (https://www.megasoftware.net/) from the alignment of the 18 amino acids of the 198 ZCCHCs found in yeast, human, and Arabidopsis proteins. The Neighbor-Joining algorithm was used, with bootstrap values from 1000 replicates. The position of a given ZCCHC in proteins with more than one of these motifs is indicated in brackets (1, 2, and 3 indicate first, second, and third, respectively, within the protein sequence, from its N- to its C-terminus). Clades highlighted in yellow or blue include ZCCHCs from proteins with a single ZCCHC that are known to be orthologs, or without a known functional relationship, respectively

its animal (*Drosophila*, mouse, rat, and Xenopus) homologs harbor a canonical ZCCHC. Structural and functional analyses have shown that human and *Xenopus* RecQ4 bind nucleic acids in vitro, mainly forked and single-strand RNA substrates, and that the substitution of N by C, reconstituting a canonical ZCCHC, impairs binding to DNA substrates [181].

We found that the presence and number of ZCCHCs are usually conserved in orthologous proteins (Figs. 2, 3), suggesting a role that, in some cases, is involved in subcellular localization, translocation, or binding to other proteins or nucleic acids. More remarkable yet is the finding of similarly clustered ZCCHCs in evolutionarily distant orthologs (Fig. 6), which suggests selective pressure acting within highly conserved ortholog groups, probably because these specific sequences play an irreplaceable role in the process in which their proteins are involved. Phylogenetic analysis of ZCCHC sequences should allow discovery and/or discrimination of true functional paralogs and orthologs. Particularly interesting will be the genomes containing large multigene families, like those of plants.

Most ZCCHC-containing proteins are nuclear, a minority are nuclear and/or cytoplasmic, and very few are only cytoplasmic. Unexpectedly, we found an Arabidopsis ZCCHC-containing factor, the DEAD-box RNA helicase RH3, to be chloroplastic. This Arabidopsis protein and its maize co-orthologs, ZmRH3A and ZmRH3B, are involved in chloroplastic splicing and chloroplastic ribosome biogenesis [149], and database searches showed that they harbor a ZCCHC, as does their rice ortholog, encoded by Os03g61220. Most chloroplast proteins are encoded by nuclear genes, synthetized in the cytoplasm and imported into the chloroplast through their N-terminal transit peptide, which targets them to the chloroplast, where they are cleaved off. The ZCCHC at the end of the Gly-Ser-rich C-terminal region of the Arabidopsis RH3 protein (Fig. 4) likely forms part of the mature chloroplastic protein and is probably functional.

We also found six human neofunctionalized genes derived from the retrotransposon Gag protein, belonging to two different families (Supporting Dataset 1f). The proteins encoded by these genes interact with other proteins instead of RNA. The discovery of their interacting proteins allowed researchers to

infer their functions [160]. Some of them seem to function in diverse tumor-related processes, promoting cell division, inhibiting apoptosis, or hindering the degradation of key proteins [182]. We found that many Arabidopsis proteins annotated as "Zinc knuckle (CCHC-type) family protein" are very similar to Ta11-like non-LTR retrotransposons, sharing the DUF4283 domain, and they could derive from retrotransposons and encode new functions, as in humans. Indeed, one of these genes, AT5G36228, is considered to be a transposable element locus whose polymorphisms in natural Arabidopsis populations have been proposed to play a role in adaptive evolution to changing environments [183].

We found that ZCCHC-containing factors are involved in most steps of RNA metabolism and affect mRNAs and ncRNAs, such as rRNAs or miRNAs. On the other hand, we found very few ZCCHC-containing factors involved in processes not related to mRNA metabolism. It is possible that other ZCCHC-containing factors participate in the other processes. The vast majority of ZCCHC-containing factors discussed here are only annotated as unknown proteins; this is especially true in the case of Arabidopsis, in which 28 of the 69 proteins identified are annotated as a "zinc knuckle (CCHC-type) family protein" or "nucleic acid binding/zinc ion binding protein" (Supporting Dataset 1d). The discovery of the DNAs, RNAs, and proteins that interact with ZCCHC-containing factors, and the search for these factors in diverse eukaryotes, along with deletion analyses of ZCCHCs, will provide a better understanding of their functions and a fruitful avenue of research to provide insight into RNA metabolism across eukaryotes.

Therefore, the identification of true ZCCHC-containing factors does not seem to be straightforward, and the number of these proteins may differ from what we found. The number probably would have been higher if we had considered in this review functional ZCCHCs with atypical (retaining the ability to bind zinc, as described in UniProtKB) or degenerate (not retaining) domains. However, it would have been lower if we had taken into account not only the primary conserved sequence of the putative ZCCHCs, but also their three-dimensional structures.

Acknowledgements We apologize to those researchers whose work could not be cited due to space constraints. We thank J. L. Micol for his critical reading, comments, and suggestions on this paper.

Funding This work was supported by the Ministerio de Ciencia e Innovación of Spain (Grant number BIO2017-89728-R MCI/AEI/FEDER EU) and the Generalitat Valenciana (Grant number PROMETEO/2019/117), both to M.R.P.; U.A.-V. held a predoctoral fellowship (GRISOLIAP/2016/134) from the Generalitat Valenciana.

References

1. Matthews JM, Sunde M (2002) Zinc fingers-folds for many occasions. IUBMB Life 54(6):351–355. https://doi.org/10.1080/15216540216035
2. Iuchi S (2001) Three classes of C_2H_2 zinc finger proteins. Cell Mol Life Sci 58(4):625–635. https://doi.org/10.1007/PL00000885
3. Laity JH, Lee BM, Wright PE (2001) Zinc finger proteins: new insights into structural and functional diversity. Curr Opin Struct Biol 11(1):39–46. https://doi.org/10.1016/S0959-440X(00)00167-6
4. Miller J, McLachlan AD, Klug A (1985) Repetitive zinc-binding domains in the protein transcription factor IIIA from *Xenopus* oocytes. EMBO J 4(6):1609–1614. https://doi.org/10.1002/j.1460-2075.1985.tb03825.x
5. Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, Chang HY, Dosztanyi Z, El-Gebali S, Fraser M, Gough J, Haft D, Holliday GL, Huang H, Huang X, Letunic I, Lopez R, Lu S, Marchler-Bauer A, Mi H, Mistry J, Natale DA, Necci M, Nuka G, Orengo CA, Park Y, Pesseat S, Piovesan D, Potter SC, Rawlings ND, Redaschi N, Richardson L, Rivoire C, Sangrador-Vegas A, Sigrist C, Sillitoe I, Smithers B, Squizzato S, Sutton G, Thanki N, Thomas PD, Tosatto SC, Wu CH, Xenarios I, Yeh LS, Young SY, Mitchell AL (2017) InterPro in 2017-beyond protein family and domain annotations. Nucleic Acids Res 45(D1):D190–D199. https://doi.org/10.1093/nar/gkw1107
6. Henderson LE, Copeland TD, Sowder RC, Smythers GW, Oroszlan S (1981) Primary structure of the low molecular weight nucleic acid-binding proteins of murine leukemia viruses. J Biol Chem 256(16):8400–8406
7. Green LM, Berg JM (1989) A retroviral Cys-Xaa_2-Cys-Xaa_4-His-Xaa_4-Cys peptide binds metal ions: spectroscopic studies and a proposed three-dimensional structure. Proc Natl Acad Sci USA 86(11):4047–4051. https://doi.org/10.1073/pnas.86.11.4047
8. Klein DJ, Johnson PE, Zollars ES, De Guzman RN, Summers MF (2000) The NMR structure of the nucleocapsid protein from the mouse mammary tumor virus reveals unusual folding of the C-terminal zinc knuckle. Biochemistry 39(7):1604–1612. https://doi.org/10.1021/bi9922493
9. De Guzman RN, Wu ZR, Stalling CC, Pappalardo L, Borer PN, Summers MF (1998) Structure of the HIV-1 nucleocapsid protein bound to the SL3 psi-RNA recognition element. Science 279(5349):384–388. https://doi.org/10.1126/science.279.5349.384
10. Amodeo P, Castiglione Morelli MA, Ostuni A, Battistuzzi G, Bavoso A (2006) Structural features in EIAV NCp11: a lentivirus nucleocapsid protein with a short linker. Biochemistry 45(17):5517–5526. https://doi.org/10.1021/bi0524924
11. Humphrey W, Dalke A, Schulten K (1996) VMD: visual molecular dynamics. J Mol Graph 14(1):33–38. https://doi.org/10.1016/0263-7855(96)00018-5
12. Daigle DM, Rossi L, Berghuis AM, Aravind L, Koonin EV, Brown ED (2002) YjeQ, an essential, conserved, uncharacterized protein from *Escherichia coli,* is an unusual GTPase with circularly permuted G-motifs and marked burst kinetics. Biochemistry 41(37):11109–11117. https://doi.org/10.1021/bi020355q
13. Jeganathan A, Razi A, Thurlow B, Ortega J (2015) The C-terminal helix in the YjeQ zinc-finger domain catalyzes the release of RbfA during 30S ribosome subunit assembly. RNA 21(6):1203–1216. https://doi.org/10.1261/rna.049171.114
14. Campbell TL, Henderson J, Heinrichs DE, Brown ED (2006) The *yjeQ* gene is required for virulence of *Staphylococcus*

aureus. Infect Immun 74(8):4918–4921. https://doi.org/10.1128/IAI.00258-06
15. Comartin DJ, Brown ED (2006) Non-ribosomal factors in ribosome subunit assembly are emerging targets for new antibacterial drugs. Curr Opin Pharm 6(5):453–458. https://doi.org/10.1016/j.coph.2006.05.005
16. Augustin MA, Huber R, Kaiser JT (2001) Crystal structure of a DNA-dependent RNA polymerase (DNA primase). Nat Struct Biol 8(1):57–61. https://doi.org/10.1038/83060
17. Yan T, Yoo D, Berardini TZ, Mueller LA, Weems DC, Weng S, Cherry JM, Rhee SY (2005) PatMatch: a program for finding patterns in peptide and nucleotide sequences. Nucleic Acids Res 33(Web Server issue):262–266. https://doi.org/10.1093/nar/gki368
18. Beasley SA, Hristova VA, Shaw GS (2007) Structure of the Parkin in-between-ring domain provides insights for E3-ligase dysfunction in autosomal recessive Parkinson's disease. Proc Natl Acad Sci USA 104(9):3095–3100. https://doi.org/10.1073/pnas.0610548104
19. Fusaro AF, Bocca SN, Ramos RL, Barroco RM, Magioli C, Jorge VC, Coutinho TC, Rangel-Lima CM, De Rycke R, Inze D, Engler G, Sachetto-Martins G (2007) AtGRP2, a cold-induced nucleo-cytoplasmic RNA-binding protein, has a role in flower and seed development. Planta 225(6):1339–1351. https://doi.org/10.1007/s00425-006-0444-4
20. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402. https://doi.org/10.1093/nar/25.17.3389
21. Hoskins AA, Moore MJ (2012) The spliceosome: a flexible, reversible macromolecular machine. Trends Biochem Sci 37(5):179–188. https://doi.org/10.1016/j.tibs.2012.02.009
22. Meyer K, Koester T, Staiger D (2015) Pre-mRNA splicing in plants: in vivo functions of RNA-binding proteins Implicated in the splicing process. Biomolecules 5(3):1717–1740. https://doi.org/10.3390/biom5031717
23. Shi Y (2017) Mechanistic insights into precursor messenger RNA splicing by the spliceosome. Nat Rev Mol Cell Biol 11:655–670. https://doi.org/10.1038/nrm.2017.86
24. Shi Y (2017) The spliceosome: a protein-directed metalloribozyme. J Mol Biol 429(17):2640–2653. https://doi.org/10.1016/j.jmb.2017.07.010
25. Berglund JA, Chua K, Abovich N, Reed R, Rosbash M (1997) The splicing factor BBP interacts specifically with the pre-mRNA branchpoint sequence UACUAAC. Cell 89(5):781–787. https://doi.org/10.1016/S0092-8674(00)80261-5
26. Kramer A (1992) Purification of splicing factor SF1, a heat-stable protein that functions in the assembly of a presplicing complex. Mol Cell Biol 12(10):4545–4552. https://doi.org/10.1128/mcb.12.10.4545
27. Berglund JA, Abovich N, Rosbash M (1998) A cooperative interaction between U2AF65 and mBBP/SF1 facilitates branchpoint region recognition. Genes Dev 12(6):858–867. https://doi.org/10.1101/gad.12.6.858
28. Rain JC, Rafi Z, Rhani Z, Legrain P, Kramer A (1998) Conservation of functional domains involved in RNA binding and protein-protein interactions in human and *Saccharomyces cerevisiae* pre-mRNA splicing factor SF1. RNA 4(5):551–565. https://doi.org/10.1017/s1355838298980335
29. Garrey SM, Voelker R, Berglund JA (2006) An extended RNA binding site for the yeast branch point-binding protein and the role of its zinc knuckle domains in RNA binding. J Biol Chem 281(37):27443–27453. https://doi.org/10.1074/jbc.M603137200
30. Rutz B, Seraphin B (2000) A dual role for BBP/ScSF1 in nuclear pre-mRNA retention and splicing. EMBO J 19(8):1873–1886. https://doi.org/10.1093/emboj/19.8.1873
31. Jang YH, Park HY, Lee KC, Thu MP, Kim SK, Suh MC, Kang H, Kim JK (2014) A homolog of splicing factor SF1 is essential for development and is involved in the alternative splicing of pre-mRNA in *Arabidopsis thaliana*. Plant J 78(4):591–603. https://doi.org/10.1111/tpj.12491
32. Crotti LB, Bacikova D, Horowitz DS (2007) The Prp18 protein stabilizes the interaction of both exons with the U5 snRNA during the second step of pre-mRNA splicing. Genes Dev 21(10):1204–1216. https://doi.org/10.1101/gad.1538207
33. Jones MH, Frank DN, Guthrie C (1995) Characterization and functional ordering of Slu7p and Prp17p during the second step of pre-mRNA splicing in yeast. Proc Natl Acad Sci USA 92(21):9687–9691. https://doi.org/10.1073/pnas.92.21.9687
34. James SA, Turner W, Schwer B (2002) How Slu7 and Prp18 cooperate in the second step of yeast pre-mRNA splicing. RNA 8(8):1068–1077. https://doi.org/10.1017/s1355838202022033
35. Frank D, Guthrie C (1992) An essential splicing factor, SLU7, mediates 3′ splice site choice in yeast. Genes Dev 6(11):2112–2124. https://doi.org/10.1101/gad.6.11.2112
36. Chua K, Reed R (1999) The RNA splicing factor hSlu7 is required for correct 3′ splice-site choice. Nature 402(6758):207–210. https://doi.org/10.1038/46086
37. Shomron N, Reznik M, Ast G (2004) Splicing factor hSlu7 contains a unique functional domain required to retain the protein within the nucleus. Mol Biol Cell 15(8):3782–3795. https://doi.org/10.1091/mbc.E04-02-0152
38. Clay NK, Nelson T (2005) The recessive epigenetic *swellmap* mutation affects the expression of two step II splicing factors required for the transcription of the cell proliferation gene *STRUWWELPETER* and for the timing of cell cycle arrest in the Arabidopsis leaf. Plant Cell 17(7):1994–2008. https://doi.org/10.1105/tpc.105.032771
39. Lin CF, Mount SM, Jarmolowski A, Makalowski W (2010) Evolutionary dynamics of U12-type spliceosomal introns. BMC Evol Biol 10:47. https://doi.org/10.1186/1471-2148-10-47
40. Niemela EH, Frilander MJ (2014) Regulation of gene expression through inefficient splicing of U12-type introns. RNA Biol 11(11):1325–1329. https://doi.org/10.1080/15476286.2014.996454
41. Wang H, Gao MX, Li L, Wang B, Hori N, Sato K (2007) Isolation, expression, and characterization of the human *ZCRB1* gene mapped to 12q12. Genomics 89(1):59–69. https://doi.org/10.1016/j.ygeno.2006.07.009
42. Kim WY, Jung HJ, Kwak KJ, Kim MK, Oh SH, Han YS, Kang H (2010) The Arabidopsis U12-type spliceosomal protein U11/U12-31K is involved in U12 intron splicing via RNA chaperone activity and affects plant development. Plant Cell 22(12):3951–3962. https://doi.org/10.1105/tpc.110.079103
43. Will CL, Schneider C, Hossbach M, Urlaub H, Rauhut R, Elbashir S, Tuschl T, Luhrmann R (2004) The human 18S U11/U12 snRNP contains a set of novel proteins not found in the U2-dependent spliceosome. RNA 10(6):929–941. https://doi.org/10.1261/rna.7320604
44. Manley JL, Krainer AR (2010) A rational nomenclature for serine/arginine-rich protein splicing factors (SR proteins). Genes Dev 24(11):1073–1074. https://doi.org/10.1101/gad.1934910
45. Haynes C, Iakoucheva LM (2006) Serine/arginine-rich splicing factors belong to a class of intrinsically disordered proteins. Nucleic Acids Res 34(1):305–312. https://doi.org/10.1093/nar/gkj424
46. Howard JM, Sanford JR (2015) The RNAissance family: SR proteins as multifaceted regulators of gene expression. Wiley Interdiscip Rev RNA 6(1):93–110. https://doi.org/10.1002/wrna.1260

47. Long JC, Caceres JF (2009) The SR protein family of splicing factors: master regulators of gene expression. Biochem J 417(1):15–27. https://doi.org/10.1042/BJ20081501
48. Graveley BR (2000) Sorting out the complexity of SR protein functions. RNA 6(9):1197–1211. https://doi.org/10.1017/S1355838200000960
49. Maniatis T, Tasic B (2002) Alternative pre-mRNA splicing and proteome expansion in metazoans. Nature 418(6894):236–243. https://doi.org/10.1038/418236a
50. Busch A, Hertel KJ (2012) Evolution of SR protein and hnRNP splicing regulatory factors. Wiley Interdiscip Rev RNA 3(1):1–12. https://doi.org/10.1002/wrna.100
51. Huang Y, Steitz JA (2001) Splicing factors SRp20 and 9G8 promote the nucleocytoplasmic export of mRNA. Mol Cell 7(4):899–905. https://doi.org/10.1016/S1097-2765(01)00233-7
52. Rauch HB, Patrick TL, Klusman KM, Battistuzzi FU, Mei W, Brendel VP, Lal SK (2014) Discovery and expression analysis of alternative splicing events conserved among plant SR proteins. Mol Biol Evol 31(3):605–613. https://doi.org/10.1093/molbev/mst238
53. Lorkovic ZJ, Hilscher J, Barta A (2008) Co-localisation studies of Arabidopsis SR splicing factors reveal different types of speckles in plant cell nuclei. Exp Cell Res 314(17):3175–3186. https://doi.org/10.1016/j.yexcr.2008.06.020
54. Cavaloc Y, Bourgeois CF, Kister L, Stevenin J (1999) The splicing factors 9G8 and SRp20 transactivate splicing through different and specific enhancers. RNA 5(3):468–483. https://doi.org/10.1017/s1355838299981967
55. Lopato S, Gattoni R, Fabini G, Stevenin J, Barta A (1999) A novel family of plant splicing factors with a Zn knuckle motif: examination of RNA binding and splicing activities. Plant Mol Biol 39(4):761–773. https://doi.org/10.1023/a:1006129615846
56. Rausin G, Tillemans V, Stankovic N, Hanikenne M, Motte P (2010) Dynamic nucleocytoplasmic shuttling of an Arabidopsis SR splicing factor: role of the RNA-binding domains. Plant Physiol 153(1):273–284. https://doi.org/10.1104/pp.110.154740
57. Lopato S, Forstner C, Kalyna M, Hilscher J, Langhammer U, Indrapichate K, Lorkovic ZJ, Barta A (2002) Network of interactions of a novel plant-specific Arg/Ser-rich protein, atRSZ33, with atSC35-like splicing factors. J Biol Chem 277(42):39989–39998. https://doi.org/10.1074/jbc.M206455200
58. Golovkin M, Reddy AS (1998) The plant U1 small nuclear ribonucleoprotein particle 70K protein interacts with two novel serine/arginine-rich proteins. Plant Cell 10(10):1637–1648. https://doi.org/10.1105/tpc.10.10.1637
59. Huang XY, Niu J, Sun MX, Zhu J, Gao JF, Yang J, Zhou Q, Yang ZN (2013) CYCLIN-DEPENDENT KINASE G1 is associated with the spliceosome to regulate *CALLOSE SYNTHASE5* splicing and pollen wall formation in Arabidopsis. Plant Cell 25(2):637–648. https://doi.org/10.1105/tpc.112.107896
60. Chang WL, Lee DC, Leu S, Huang YM, Lu MC, Ouyang P (2003) Molecular characterization of a novel nucleolar protein, pNO40. Biochem Biophys Res Commun 307(3):569–577. https://doi.org/10.1016/s0006-291x(03)01208-7
61. Gueydan C, Wauquier C, De Mees C, Huez G, Kruys V (2002) Identification of ribosomal proteins specific to higher eukaryotic organisms. J Biol Chem 277(47):45034–45040. https://doi.org/10.1074/jbc.M208551200
62. Lin YM, Chu PH, Ouyang P (2019) Ectopically expressed pNO40 suppresses ribosomal RNA synthesis by inhibiting UBF-dependent transcription activation. Biochem Biophys Res Commun. https://doi.org/10.1016/j.bbrc.2019.06.057
63. Lin YM, Chu PH, Li YZ, Ouyang P (2017) Ribosomal protein pNO40 mediates nucleolar sequestration of SR family splicing factors and its overexpression impairs mRNA metabolism. Cell Signal 32:12–23. https://doi.org/10.1016/j.cellsig.2017.01.010
64. Moss EG (2007) Heterochronic genes and the nature of developmental time. Curr Biol 17(11):425–434. https://doi.org/10.1016/j.cub.2007.03.043
65. Ha M, Kim VN (2014) Regulation of microRNA biogenesis. Nat Rev Mol Cell Biol 15(8):509–524. https://doi.org/10.1038/nrm3838
66. Moss EG, Tang L (2003) Conservation of the heterochronic regulator Lin-28, its developmental expression and microRNA complementary sites. Dev Biol 258(2):432–442. https://doi.org/10.1016/s0012-1606(03)00126-x
67. Piskounova E, Polytarchou C, Thornton JE, LaPierre RJ, Pothoulakis C, Hagan JP, Iliopoulos D, Gregory RI (2011) Lin28A and Lin28B inhibit let-7 microRNA biogenesis by distinct mechanisms. Cell 147(5):1066–1079. https://doi.org/10.1016/j.cell.2011.10.039
68. Tsialikas J, Romer-Seibert J (2015) LIN28: roles and regulation in development and beyond. Development 142(14):2397–2404. https://doi.org/10.1242/dev.117580
69. Graumann PL, Marahiel MA (1998) A superfamily of proteins that contain the cold-shock domain. Trends Biochem Sci 23(8):286–290. https://doi.org/10.1016/S0968-0004(98)01255-9
70. Mihailovich M, Militti C, Gabaldon T, Gebauer F (2010) Eukaryotic cold shock domain proteins: highly versatile regulators of gene expression. BioEssays 32(2):109–118. https://doi.org/10.1002/bies.200900122
71. Heo I, Joo C, Kim YK, Ha M, Yoon MJ, Cho J, Yeom KH, Han J, Kim VN (2009) TUT4 in concert with Lin28 suppresses microRNA biogenesis through pre-microRNA uridylation. Cell 138(4):696–708. https://doi.org/10.1016/j.cell.2009.08.002
72. Cho J, Chang H, Kwon SC, Kim B, Kim Y, Choe J, Ha M, Kim YK, Kim VN (2012) LIN28A is a suppressor of ER-associated translation in embryonic stem cells. Cell 151(4):765–777. https://doi.org/10.1016/j.cell.2012.10.019
73. Wang X, Zhang S, Dou Y, Zhang C, Chen X, Yu B, Ren G (2015) Synergistic and independent actions of multiple terminal nucleotidyl transferases in the 3′ tailing of small RNAs in Arabidopsis. PLoS Genet 11(4):e1005091. https://doi.org/10.1371/journal.pgen.1005091
74. Kingsley PD, Palis J (1994) GRP2 proteins contain both CCHC zinc fingers and a cold shock domain. Plant Cell 6(11):1522–1523. https://doi.org/10.1105/tpc.6.11.1522
75. Sachetto-Martins G, Franco LO, de Oliveira DE (2000) Plant glycine-rich proteins: a family or just proteins with a common motif? Biochim Biophys Acta 1492(1):1–14. https://doi.org/10.1016/s0167-4781(00)00064-6
76. Mangeon A, Junqueira RM, Sachetto-Martins G (2010) Functional diversity of the plant glycine-rich proteins superfamily. Plant Signal Behav 5(2):99–104. https://doi.org/10.4161/psb.5.2.10336
77. Krishnamurthy P, Kim JA, Jeong MJ, Kang CH, Lee SI (2015) Defining the RNA-binding glycine-rich (RBG) gene superfamily: new insights into nomenclature, phylogeny, and evolutionary trends obtained by genome-wide comparative analysis of Arabidopsis, Chinese cabbage, rice and maize genomes. Mol Genet Genom 290(6):2279–2295. https://doi.org/10.1007/s00438-015-1080-0
78. Juntawong P, Sorenson R, Bailey-Serres J (2013) Cold shock protein 1 chaperones mRNAs during translation in *Arabidopsis thaliana*. Plant J 74(6):1016–1028. https://doi.org/10.1111/tpj.12187
79. Sasaki K, Kim MH, Imai R (2007) Arabidopsis COLD SHOCK DOMAIN PROTEIN2 is a RNA chaperone that is regulated by cold and developmental signals. Biochem Biophys Res Commun 364(3):633–638. https://doi.org/10.1016/j.bbrc.2007.10.059

80. Sasaki K, Kim MH, Imai R (2013) Arabidopsis COLD SHOCK DOMAIN PROTEIN 2 is a negative regulator of cold acclimation. New Phytol 198(1):95–102. https://doi.org/10.1111/nph.12118
81. Kim MH, Sato S, Sasaki K, Saburi W, Matsui H, Imai R (2013) COLD SHOCK DOMAIN PROTEIN 3 is involved in salt and drought stress tolerance in Arabidopsis. FEBS Open Biol 3:438–442. https://doi.org/10.1016/j.fob.2013.10.003
82. Pontvianne F, Matia I, Douet J, Tourmente S, Medina FJ, Echeverria M, Saez-Vasquez J (2007) Characterization of *AtNUC-L1* reveals a central role of nucleolin in nucleolus organization and silencing of *AtNUC-L2* gene in Arabidopsis. Mol Biol Cell 18(2):369–379. https://doi.org/10.1091/mbc.E06-08-0751
83. Petricka JJ, Nelson TM (2007) Arabidopsis nucleolin affects plant development and patterning. Plant Physiol 144(1):173–186. https://doi.org/10.1104/pp.106.093575
84. Kim MH, Sonoda Y, Sasaki K, Kaminaka H, Imai R (2013) Interactome analysis reveals versatile functions of Arabidopsis COLD SHOCK DOMAIN PROTEIN 3 in RNA processing within the nucleus and cytoplasm. Cell Stress Chaperones 18(4):517–525. https://doi.org/10.1007/s12192-012-0398-3
85. Kim YO, Kim JS, Kang H (2005) Cold-inducible zinc finger-containing glycine-rich RNA-binding protein contributes to the enhancement of freezing tolerance in *Arabidopsis thaliana*. Plant J 42(6):890–900. https://doi.org/10.1111/j.1365-313X.2005.02420.x
86. Wu Z, Zhu D, Lin X, Miao J, Gu L, Deng X, Yang Q, Sun K, Zhu D, Cao X, Tsuge T, Dean C, Aoyama T, Gu H, Qu LJ (2016) RNA binding proteins RZ-1B and RZ-1C play critical roles in regulating pre-mRNA splicing and gene expression during development in Arabidopsis. Plant Cell 28(1):55–73. https://doi.org/10.1105/tpc.15.00949
87. Neve J, Patel R, Wang Z, Louey A, Furger AM (2017) Cleavage and polyadenylation: ending the message expands gene regulation. RNA Biol 14(7):865–890. https://doi.org/10.1080/15476286.2017.1306171
88. Vo LT, Minet M, Schmitter JM, Lacroute F, Wyers F (2001) Mpe1, a zinc knuckle protein, is an essential component of yeast cleavage and polyadenylation factor required for the cleavage and polyadenylation of mRNA. Mol Cell Biol 21(24):8346–8356. https://doi.org/10.1128/MCB.21.24.8346-8356.2001
89. Lee SD, Moore CL (2014) Efficient mRNA polyadenylation requires a ubiquitin-like domain, a zinc knuckle, and a RING finger domain, all contained in the Mpe1 protein. Mol Cell Biol 34(21):3955–3967. https://doi.org/10.1128/MCB.00077-14
90. Sakai Y, Saijo M, Coelho K, Kishino T, Niikawa N, Taya Y (1995) cDNA sequence and chromosomal localization of a novel human protein, RBQ-1 (RBBP6), that binds to the retinoblastoma gene product. Genomics 30(1):98–101. https://doi.org/10.1006/geno.1995.0017
91. Simons A, Melamed-Bessudo C, Wolkowicz R, Sperling J, Sperling R, Eisenbach L, Rotter V (1997) PACT: cloning and characterization of a cellular p53 binding protein that interacts with Rb. Oncogene 14(2):145–155. https://doi.org/10.1038/sj.onc.1200825
92. Li L, Deng B, Xing G, Teng Y, Tian C, Cheng X, Yin X, Yang J, Gao X, Zhu Y, Sun Q, Zhang L, Yang X, He F (2007) PACT is a negative regulator of p53 and essential for cell growth and embryonic development. Proc Natl Acad Sci USA 104(19):7951–7956. https://doi.org/10.1073/pnas.0701916104
93. Di Giammartino DC, Li W, Ogami K, Yashinskie JJ, Hoque M, Tian B, Manley JL (2014) RBBP6 isoforms regulate the human polyadenylation machinery and modulate expression of mRNAs with AU-rich 3' UTRs. Genes Dev 28(20):2248–2260. https://doi.org/10.1101/gad.245787.114
94. Luo C, Cai XT, Du J, Zhao TL, Wang PF, Zhao PX, Liu R, Xie Q, Cao XF, Xiang CB (2016) PARAQUAT TOLERANCE3 Is an E3 Ligase that switches off activated oxidative response by targeting histone-modifying PROTEIN METHYLTRANSFERASE4b. PLoS Genet 12(9):e1006332. https://doi.org/10.1371/journal.pgen.1006332
95. Pugh DJ, Ab E, Faro A, Lutya PT, Hoffmann E, Rees DJ (2006) DWNN, a novel ubiquitin-like domain, implicates RBBP6 in mRNA processing and ubiquitin-like pathways. BMC Struct Biol 6:1. https://doi.org/10.1186/1472-6807-6-1
96. Barabino SM, Hubner W, Jenny A, Minvielle-Sebastia L, Keller W (1997) The 30-kD subunit of mammalian cleavage and polyadenylation specificity factor and its yeast homolog are RNA-binding zinc finger proteins. Genes Dev 11(13):1703–1716. https://doi.org/10.1101/gad.11.13.1703
97. Chakrabarti M, Hunt AG (2015) CPSF30 at the interface of alternative polyadenylation and cellular signaling in plants. Biomolecules 5(2):1151–1168. https://doi.org/10.3390/biom5021151
98. Liu Q, Dreyfuss G (1995) In vivo and in vitro arginine methylation of RNA-binding proteins. Mol Cell Biol 15(5):2800–2808. https://doi.org/10.1128/mcb.15.5.2800
99. Siebel CW, Guthrie C (1996) The essential yeast RNA binding protein Np13p is methylated. Proc Natl Acad Sci USA 93(24):13641–13646. https://doi.org/10.1073/pnas.93.24.13641
100. Jarvelin AI, Noerenberg M, Davis I, Castello A (2016) The new (dis)order in RNA regulation. Cell Commun Signal 14:9. https://doi.org/10.1186/s12964-016-0132-3
101. Ozdilek BA, Thompson VF, Ahmed NS, White CI, Batey RT, Schwartz JC (2017) Intrinsically disordered RGG/RG domains mediate degenerate specificity in RNA binding. Nucleic Acids Res 45(13):7984–7996. https://doi.org/10.1093/nar/gkx460
102. Zinder JC, Lima CD (2017) Targeting RNA for processing or destruction by the eukaryotic RNA exosome and its cofactors. Genes Dev 31(2):88–100. https://doi.org/10.1101/gad.294769.116
103. Houseley J, Tollervey D (2008) The nuclear RNA surveillance machinery: the link between ncRNAs and genome structure in budding yeast? Biochim Biophys Acta 1779(4):239–246. https://doi.org/10.1016/j.bbagrm.2007.12.008
104. Fasken MB, Leung SW, Banerjee A, Kodani MO, Chavez R, Bowman EA, Purohit MK, Rubinson ME, Rubinson EH, Corbett AH (2011) Air1 zinc knuckles 4 and 5 and a conserved IWRXY motif are critical for the function and integrity of the Trf4/5-Air1/2-Mtr4 polyadenylation (TRAMP) RNA quality control complex. J Biol Chem 286(43):37429–37445. https://doi.org/10.1074/jbc.M111.271494
105. Vanácová S, Wolf J, Martin G, Blank D, Dettwiler S, Friedlein A, Langen H, Keith G, Keller W (2005) A new yeast poly(A) polymerase complex involved in RNA quality control. PLoS Biol 3(6):e189. https://doi.org/10.1371/journal.pbio.0030189
106. Falk S, Finogenova K, Melko M, Benda C, Lykke-Andersen S, Jensen TH, Conti E (2016) Structure of the RBM7-ZCCHC8 core of the NEXT complex reveals connections to splicing factors. Nat Commun 7:13573. https://doi.org/10.1038/ncomms13573
107. Fox MJ, Mosley AL (2016) Rrp 6: Integrated roles in nuclear RNA metabolism and transcription termination. Wiley Interdiscip Rev RNA 7(1):91–104. https://doi.org/10.1002/wrna.1317
108. Lange H, Holec S, Cognat V, Pieuchot L, Le Ret M, Canaday J, Gagliardi D (2008) Degradation of a polyadenylated rRNA maturation by-product involves one of the three RRP6-like proteins in *Arabidopsis thaliana*. Mol Cell Biol 28(9):3038–3044. https://doi.org/10.1128/MCB.02064-07
109. Sudo H, Nozaki A, Uno H, Ishida Y, Nagahama M (2016) Interaction properties of human TRAMP-like proteins and their role

in pre-rRNA 5'ETS turnover. FEBS Lett 590(17):2963–2972. https://doi.org/10.1002/1873-3468.12314
110. Lubas M, Christensen MS, Kristiansen MS, Domanski M, Falkenby LG, Lykke-Andersen S, Andersen JS, Dziembowski A, Jensen TH (2011) Interaction profiling identifies the human nuclear exosome targeting complex. Mol Cell 43(4):624–637. https://doi.org/10.1016/j.molcel.2011.06.028
111. Lange H, Sement FM, Gagliardi D (2011) MTR4, a putative RNA helicase and exosome co-factor, is required for proper rRNA biogenesis and development in *Arabidopsis thaliana*. Plant J 68(1):51–63. https://doi.org/10.1111/j.1365-313X.2011.04675.x
112. Lange H, Zuber H, Sement FM, Chicher J, Kuhn L, Hammann P, Brunaud V, Berard C, Bouteiller N, Balzergue S, Aubourg S, Martin-Magniette ML, Vaucheret H, Gagliardi D (2014) The RNA helicases AtMTR4 and HEN2 target specific subsets of nuclear transcripts for degradation by the nuclear exosome in *Arabidopsis thaliana*. PLoS Genet 10(8):e1004564. https://doi.org/10.1371/journal.pgen.1004564
113. Liquori CL, Ricker K, Moseley ML, Jacobsen JF, Kress W, Naylor SL, Day JW, Ranum LP (2001) Myotonic dystrophy type 2 caused by a CCTG expansion in intron 1 of *ZNF9*. Science 293(5531):864–867. https://doi.org/10.1126/science.1062125
114. Rajavashisth TB, Taylor AK, Andalibi A, Svenson KL, Lusis AJ (1989) Identification of a zinc finger protein that binds to the sterol regulatory element. Science 245(4918):640–643. https://doi.org/10.1126/science.2562787
115. Pellizzoni L, Lotti F, Maras B, Pierandrei-Amaldi P (1997) Cellular nucleic acid binding protein binds a conserved region of the 5′ UTR of *Xenopus laevis* ribosomal protein mRNAs. J Mol Biol 267(2):264–275. https://doi.org/10.1006/jmbi.1996.0888
116. Pellizzoni L, Lotti F, Rutjes SA, Pierandrei-Amaldi P (1998) Involvement of the *Xenopus laevis* Ro60 autoantigen in the alternative interaction of La and CNBP proteins with the 5′ UTR of L4 ribosomal protein mRNA. J Mol Biol 281(4):593–608. https://doi.org/10.1006/jmbi.1998.1961
117. Rojas M, Farr GW, Fernandez CF, Lauden L, McCormack JC, Wolin SL (2012) Yeast Gis2 and its human ortholog CNBP are novel components of stress-induced RNP granules. PLoS One 7(12):e52824. https://doi.org/10.1371/journal.pone.0052824
118. Buchan JR, Parker R (2009) Eukaryotic stress granules: the ins and outs of translation. Mol Cell 36(6):932–941. https://doi.org/10.1016/j.molcel.2009.11.020
119. Huichalaf C, Schoser B, Schneider-Gold C, Jin B, Sarkar P, Timchenko L (2009) Reduction of the rate of protein translation in patients with myotonic dystrophy 2. J Neurosci 29(28):9042–9049. https://doi.org/10.1523/JNEUROSCI.1983-09.2009
120. Sammons MA, Samir P, Link AJ (2011) *Saccharomyces cerevisiae* Gis2 interacts with the translation machinery and is orthogonal to myotonic dystrophy type 2 protein ZNF9. Biochem Biophys Res Commun 406(1):13–19. https://doi.org/10.1016/j.bbrc.2011.01.086
121. Iadevaia V, Caldarola S, Tino E, Amaldi F, Loreni F (2008) All translation elongation factors and the e, f, and h subunits of translation initiation factor 3 are encoded by 5′-terminal oligopyrimidine (TOP) mRNAs. RNA 14(9):1730–1736. https://doi.org/10.1261/rna.1037108
122. David AP, Pipier A, Pascutti F, Binolfi A, Weiner AMJ, Challier E, Heckel S, Calsou P, Gomez D, Calcaterra NB, Armas P (2019) CNBP controls transcription by unfolding DNA G-quadruplex structures. Nucleic Acids Res 47(15):7901–7913. https://doi.org/10.1093/nar/gkz527
123. Benhalevy D, Gupta SK, Danan CH, Ghosal S, Sun HW, Kazemier HG, Paeschke K, Hafner M, Juranek SA (2017) The human CCHC-type zinc finger nucleic acid-binding protein binds G-Rich elements in target mRNA coding sequences and promotes translation. Cell Rep 18(12):2979–2990. https://doi.org/10.1016/j.celrep.2017.02.080
124. Antonucci L, D'Amico D, Di Magno L, Coni S, Di Marcotullio L, Cardinali B, Gulino A, Ciapponi L, Canettieri G (2014) CNBP regulates wing development in *Drosophila melanogaster* by promoting IRES-dependent translation of *dMyc*. Cell Cycle 13(3):434–439. https://doi.org/10.4161/cc.27268
125. McNeil GP, Schroeder AJ, Roberts MA, Jackson FR (2001) Genetic analysis of functional domains within the Drosophila LARK RNA-binding protein. Genetics 159(1):229–240
126. Markus MA, Morris BJ (2009) RBM4: a multifunctional RNA-binding protein. Int J Biochem Cell Biol 41(4):740–743. https://doi.org/10.1016/j.biocel.2008.05.027
127. Niu K, Xiang L, Jin Y, Peng Y, Wu F, Tang W, Zhang X, Deng H, Xiang H, Li S, Wang J, Song Q, Feng Q (2019) Identification of LARK as a novel and conserved G-quadruplex binding protein in invertebrates and vertebrates. Nucleic Acids Res 47(14):7306–7320. https://doi.org/10.1093/nar/gkz484
128. Parker R (2012) RNA degradation in *Saccharomyces cerevisae*. Genetics 191(3):671–702. https://doi.org/10.1534/genetics.111.137265
129. Parker R, Song H (2004) The enzymes and control of eukaryotic mRNA turnover. Nat Struct Mol Biol 11(2):121–127. https://doi.org/10.1038/nsmb724
130. Medina DA, Jordan-Pla A, Millan-Zambrano G, Chavez S, Choder M, Perez-Ortin JE (2014) Cytoplasmic 5'–3' exonuclease Xrn1p is also a genome-wide transcription factor in yeast. Front Genet 5:1. https://doi.org/10.3389/fgene.2014.00001
131. Luo W, Johnson AW, Bentley DL (2006) The role of Rat1 in coupling mRNA 3'-end processing to transcription termination: implications for a unified allosteric-torpedo model. Genes Dev 20(8):954–965. https://doi.org/10.1101/gad.1409106
132. El Hage A, Koper M, Kufel J, Tollervey D (2008) Efficient termination of transcription by RNA polymerase I requires the 5′ exonuclease Rat1 in yeast. Genes Dev 22(8):1069–1081. https://doi.org/10.1101/gad.463708
133. Petfalski E, Dandekar T, Henry Y, Tollervey D (1998) Processing of the precursors to small nucleolar RNAs and rRNAs requires common components. Mol Cell Biol 18(3):1181–1189. https://doi.org/10.1128/mcb.18.3.1181
134. Geerlings TH, Vos JC, Raue HA (2000) The final step in the formation of 25S rRNA in *Saccharomyces cerevisiae* is performed by 5′ → 3′exonucleases. RNA 6(12):1698–1703. https://doi.org/10.1017/s1355838200001540
135. West S, Gromak N, Proudfoot NJ (2004) Human 5′ → 3′ exonuclease Xrn2 promotes transcription termination at co-transcriptional cleavage sites. Nature 432(7016):522–525. https://doi.org/10.1038/nature03035
136. Morales JC, Richard P, Patidar PL, Motea EA, Dang TT, Manley JL, Boothman DA (2016) XRN2 links transcription termination to DNA damage and replication stress. PLoS Genet 12(7):e1006107. https://doi.org/10.1371/journal.pgen.1006107
137. Kastenmayer JP, Green PJ (2000) Novel features of the XRN-family in Arabidopsis: evidence that AtXRN4, one of several orthologs of nuclear Xrn2p/Rat1p, functions in the cytoplasm. Proc Natl Acad Sci USA 97(25):13985–13990. https://doi.org/10.1073/pnas.97.25.13985
138. Zakrzewska-Placzek M, Souret FF, Sobczyk GJ, Green PJ, Kufel J (2010) *Arabidopsis thaliana* XRN2 is required for primary cleavage in the pre-ribosomal RNA. Nucleic Acids Res 38(13):4487–4502. https://doi.org/10.1093/nar/gkq172
139. Souret FF, Kastenmayer JP, Green PJ (2004) AtXRN4 degrades mRNA in *Arabidopsis* and its substrates include selected miRNA targets. Mol Cell 15(2):173–183. https://doi.org/10.1016/j.molcel.2004.06.006

140. Gazzani S, Lawrenson T, Woodward C, Headon D, Sablowski R (2004) A link between mRNA turnover and RNA interference in Arabidopsis. Science 306(5698):1046–1048. https://doi.org/10.1126/science.1101092
141. Ma H, Wang X, Cai J, Dai Q, Natchiar SK, Lv R, Chen K, Lu Z, Chen H, Shi YG, Lan F, Fan J, Klaholz BP, Pan T, Shi Y, He C (2019) N(6-)Methyladenosine methyltransferase ZCCHC4 mediates ribosomal RNA methylation. Nat Chem Biol 15(1):88–94. https://doi.org/10.1038/s41589-018-0184-3
142. Zhou A, Zhou J, Yang L, Liu M, Li H, Xu S, Han M, Zhang J (2008) A nuclear localized protein ZCCHC9 is expressed in cerebral cortex and suppresses the MAPK signal pathway. J Genet Genom 35(8):467–472. https://doi.org/10.1016/s1673-8527(08)60064-8
143. Sanudo M, Jacko M, Rammelt C, Vanacova S, Stefl R (2011) 1H, 13C, and 15N chemical shift assignments of ZCCHC9. Biomol NMR Assign 5(1):19–21. https://doi.org/10.1007/s12104-010-9257-2
144. Tafforeau L, Zorbas C, Langhendries JL, Mullineux ST, Stamatopoulou V, Mullier R, Wacheul L, Lafontaine DL (2013) The complexity of human ribosome biogenesis revealed by systematic nucleolar screening of Pre-rRNA processing factors. Mol Cell 51(4):539–551. https://doi.org/10.1016/j.molcel.2013.08.011
145. Hsin JP, Manley JL (2012) The RNA polymerase II CTD coordinates transcription and RNA processing. Genes Dev 26(19):2119–2137. https://doi.org/10.1101/gad.200303.112
146. Gullerova M, Barta A, Lorkovic ZJ (2006) AtCyp59 is a multidomain cyclophilin from *Arabidopsis thaliana* that interacts with SR proteins and the C-terminal domain of the RNA polymerase II. RNA 12(4):631–643. https://doi.org/10.1261/rna.2226106
147. Tzafrir I, Pena-Muralla R, Dickerman A, Berg M, Rogers R, Hutchens S, Sweeney TC, McElver J, Aux G, Patton D, Meinke D (2004) Identification of genes required for embryo development in Arabidopsis. Plant Physiol 135(3):1206–1220. https://doi.org/10.1104/pp.104.045179
148. Shao S, Brown A, Santhanam B, Hegde RS (2015) Structure and assembly pathway of the ribosome quality control complex. Mol Cell 57(3):433–444. https://doi.org/10.1016/j.molcel.2014.12.015
149. Asakura Y, Galarneau E, Watkins KP, Barkan A, van Wijk KJ (2012) Chloroplast RH3 DEAD box RNA helicases in maize and Arabidopsis function in splicing of specific group II introns and affect chloroplast ribosome biogenesis. Plant Physiol 159(3):961–974. https://doi.org/10.1104/pp.112.197525
150. Gu L, Xu T, Lee K, Lee KH, Kang H (2014) A chloroplast-localized DEAD-box RNA helicaseAtRH3 is essential for intron splicing and plays an important role in the growth and stress response in *Arabidopsis thaliana*. Plant Physiol Biochem 82:309–318. https://doi.org/10.1016/j.plaphy.2014.07.006
151. Han P, Ye W, Lv X, Ma H, Weng D, Dong Y, Cheng L, Chen H, Zhang L, Xu Z, Lei Y, Zhang F (2017) DDX50 inhibits the replication of dengue virus 2 by upregulating IFN-beta production. Arch Virol 162(6):1487–1494. https://doi.org/10.1007/s00705-017-3250-3
152. Ma L, Xie B, Hong Z, Verma DP, Zhang Z (2008) A novel RNA-binding protein associated with cell plate formation. Plant Physiol 148(1):223–234. https://doi.org/10.1104/pp.108.120527
153. Miller RK, D'Silva S, Moore JK, Goodson HV (2006) The CLIP-170 orthologue Bik1p and positioning the mitotic spindle in yeast. Curr Top Dev Biol 76:49–87. https://doi.org/10.1016/S0070-2153(06)76002-1
154. Blake-Hodek KA, Cassimeris L, Huffaker TC (2010) Regulation of microtubule dynamics by Bim1 and Bik1, the budding yeast members of the EB1 and CLIP-170 families of plus-end tracking proteins. Mol Biol Cell 21(12):2013–2023. https://doi.org/10.1091/mbc.E10-02-0083
155. Weisbrich A, Honnappa S, Jaussi R, Okhrimenko O, Frey D, Jelesarov I, Akhmanova A, Steinmetz MO (2007) Structure-function relationship of CAP-Gly domains. Nat Struct Mol Biol 14(10):959–967. https://doi.org/10.1038/nsmb1291
156. Gardiner J, Overall R, Marc J (2011) Putative Arabidopsis homologues of metazoan coiled-coil cytoskeletal proteins. Cell Biol Int 35(8):767–774. https://doi.org/10.1042/CBI20100719
157. Jiang Y, Zhu Y, Liu ZJ, Ouyang S (2017) The emerging roles of the DDX41 protein in immunity and diseases. Protein Cell 8(2):83–89. https://doi.org/10.1007/s13238-016-0303-4
158. Lian H, Zang R, Wei J, Ye W, Hu MM, Chen YD, Zhang XN, Guo Y, Lei CQ, Yang Q, Luo WW, Li S, Shu HB (2018) The zinc-finger protein ZCCHC3 binds RNA and facilitates viral RNA sensing and activation of the RIG-I-like receptors. Immunity 49(3):438–448. https://doi.org/10.1016/j.immuni.2018.08.014
159. Lian H, Wei J, Zang R, Ye W, Yang Q, Zhang XN, Chen YD, Fu YZ, Hu MM, Lei CQ, Luo WW, Li S, Shu HB (2018) ZCCHC3 is a co-sensor of cGAS for dsDNA recognition in innate immune response. Nat Commun 9(1):3349. https://doi.org/10.1038/s41467-018-05559-w
160. Campillos M, Doerks T, Shah PK, Bork P (2006) Computational characterization of multiple Gag-like human proteins. Trends Genet 22(11):585–589. https://doi.org/10.1016/j.tig.2006.09.006
161. Brandt J, Schrauth S, Veith AM, Froschauer A, Haneke T, Schultheis C, Gessler M, Leimeister C, Volff JN (2005) Transposable elements as a source of genetic innovation: expression and evolution of a family of retrotransposon-derived neogenes in mammals. Gene 345(1):101–111. https://doi.org/10.1016/j.gene.2004.11.022
162. Pang SW, Lahiri C, Poh CL, Tan KO (2018) PNMA family: Protein interaction network and cell signalling pathways implicated in cancer and apoptosis. Cell Signal 45:54–62. https://doi.org/10.1016/j.cellsig.2018.01.022
163. Ning Y, Hui N, Qing B, Zhuo Y, Sun W, Du Y, Liu S, Liu K, Zhou J (2019) ZCCHC10 suppresses lung cancer progression and cisplatin resistance by attenuating MDM2-mediated p53 ubiquitination and degradation. Cell Death Dis 10(6):414. https://doi.org/10.1038/s41419-019-1635-9
164. Mallory A, Vaucheret H (2010) Form, function, and regulation of ARGONAUTE proteins. Plant Cell 22(12):3879–3889. https://doi.org/10.1105/tpc.110.080671
165. Schalk C, Cognat V, Graindorge S, Vincent T, Voinnet O, Molinier J (2017) Small RNA-mediated repair of UV-induced DNA lesions by the DNA DAMAGE-BINDING PROTEIN 2 and ARGONAUTE 1. Proc Natl Acad Sci USA 114(14):E2965–E2974. https://doi.org/10.1073/pnas.1618834114
166. Aklilu BB, Culligan KM (2016) Molecular evolution and functional diversification of replication protein A1 in plants. Front Plant Sci 7:33. https://doi.org/10.3389/fpls.2016.00033
167. Séguéla-Arnaud M, Choinard S, Larchevêque C, Girard C, Froger N, Crismani W, Mercier R (2017) RMI1 and TOP3alpha limit meiotic CO formation through their C-terminal domains. Nucleic Acids Res 45(4):1860–1871. https://doi.org/10.1093/nar/gkw1210
168. Loudet O, Michael TP, Burger BT, Le Mette C, Mockler TC, Weigel D, Chory J (2008) A zinc knuckle protein that negatively controls morning-specific growth in *Arabidopsis thaliana*. Proc Natl Acad Sci USA 105(44):17193–17198. https://doi.org/10.1073/pnas.0807264105
169. Zhang S, Li C, Zhou Y, Wang X, Li H, Feng Z, Chen H, Qin G, Jin D, Terzaghi W, Gu H, Qu LJ, Kang D, Deng XW, Li J (2018) TANDEM ZINC-FINGER/PLUS3 is a key component

of phytochrome A signaling. Plant Cell 30(4):835–852. https://doi.org/10.1105/tpc.17.00677
170. Nguyen NH, Jeong CY, Kang GH, Yoo SD, Hong SW, Lee H (2015) MYBD employed by HY5 increases anthocyanin accumulation via repression of *MYBL2* in Arabidopsis. Plant J 84(6):1192–1205. https://doi.org/10.1111/tpj.13077
171. Kwon Y, Kim JH, Nguyen HN, Jikumaru Y, Kamiya Y, Hong SW, Lee H (2013) A novel Arabidopsis MYB-like transcription factor, MYBH, regulates hypocotyl elongation by enhancing auxin accumulation. J Exp Bot 64(12):3911–3922. https://doi.org/10.1093/jxb/ert223
172. Lu D, Wang T, Persson S, Mueller-Roeber B, Schippers JH (2014) Transcriptional control of ROS homeostasis by KUODA1 regulates cell expansion during leaf development. Nat Commun 5:3767. https://doi.org/10.1038/ncomms4767
173. Cheng NH, Liu JZ, Nelson RS, Hirschi KD (2004) Characterization of CXIP4, a novel Arabidopsis protein that activates the H+/Ca2+ antiporter, CAX1. FEBS Lett 559(1–3):99–106. https://doi.org/10.1016/S0014-5793(04)00036-5
174. Sánchez-García AB, Aguilera V, Micol-Ponce R, Jover-Gil S, Ponce MR (2015) Arabidopsis *MAS2*, an essential gene that encodes a homolog of animal NF-kappa B Activating Protein, is involved in 45S ribosomal DNA silencing. Plant Cell 27(7):1999–2015. https://doi.org/10.1105/tpc.15.00135
175. Armas P, Calcaterra NB (2012) Retroviral Zinc knuckles in eukaryotic cellular proteins. In: Ciofani R, Makrlik L (eds) Zinc fingers: structure, properties, and applications. Nova Biomedical Publishers, Hauppauge, pp 51–80
176. de Castro E, Sigrist CJ, Gattiker A, Bulliard V, Langendijk-Genevaux PS, Gasteiger E, Bairoch A, Hulo N (2006) ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Res 34(Web Server issue):W362–W365. https://doi.org/10.1093/nar/gkl124
177. Sigrist CJ, De Castro E, Langendijk-Genevaux PS, Le Saux V, Bairoch A, Hulo N (2005) ProRule: a new database containing functional and structural information on PROSITE profiles. Bioinformatics 21(21):4060–4066. https://doi.org/10.1093/bioinformatics/bti614
178. Sigrist CJ, Cerutti L, Hulo N, Gattiker A, Falquet L, Pagni M, Bairoch A, Bucher P (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. Brief Bioinform 3(3):265–274. https://doi.org/10.1093/bib/3.3.265
179. Heese K, Fujita M, Akatsu H, Yamamoto T, Kosaka K, Nagai Y, Sawada T (2004) The splicing regulatory protein p18SRP is down-regulated in Alzheimer's disease brain. J Mol Neurosci 24(2):269–276. https://doi.org/10.1385/JMN:24:2:269
180. Coletta A, Pinney JW, Solis DY, Marsh J, Pettifer SR, Attwood TK (2010) Low-complexity regions within protein sequences have position-dependent roles. BMC Syst Biol 4:43. https://doi.org/10.1186/1752-0509-4-43
181. Marino F, Mojumdar A, Zucchelli C, Bhardwaj A, Buratti E, Vindigni A, Musco G, Onesti S (2016) Structural and biochemical characterization of an RNA/DNA binding motif in the N-terminal domain of RecQ4 helicases. Sci Rep 6:21501. https://doi.org/10.1038/srep21501
182. Xie T, Pan S, Zheng H, Luo Z, Tembo KM, Jamal M, Yu Z, Yu Y, Xia J, Yin Q, Wang M, Yuan W, Zhang Q, Xiong J (2018) *PEG10* as an oncogene: expression regulatory mechanisms and role in tumor progression. Cancer Cell Int 18:112. https://doi.org/10.1186/s12935-018-0610-3
183. Li ZW, Hou XH, Chen JF, Xu YC, Wu Q, Gonzalez J, Guo YL (2018) Transposable Elements contribute to the adaptation of *Arabidopsis thaliana*. Genome Biol Evol 10(8):2140–2150. https://doi.org/10.1093/gbe/evy171

Genome-wide analysis of CCHC-type zinc finger (ZCCHC) proteins in yeast, Arabidopsis, and humans

Supplementary Figures

Uri Aceituno-Valenzuela, Rosa Micol-Ponce, and María Rosa Ponce

Instituto de Bioingeniería, Universidad Miguel Hernández, Campus de Elche, 03202 Elche, Spain

Corresponding author:
María Rosa Ponce
email: mrponce@umh.es

Air1; 1 $CX_2CX_4HX_4C$ and 3 $CX_2CX_3GHX_4C$

MSTLLSEVESIDTLPYVKDTTPTGSDSSSFNKLLAPSIEDVDANPEELRTLRGQGRYFGITDYDSNGAIMEAEPKCNNCSQRGHLKRNC
PHVICTYCGFMDDHYSQHCPKAIICTNCNANGHYKSQCPHKWKKVFCTLCNSKRHSRERCPSIWRSYLLKTKDTNQGDFDFQTVFCYNC
GNAGHFGDDCAEGRSSRVPNTDGSAFCGDNLATKFKQHYFNQLKDYKREASQRQHFDNEHEFNLLDYEYNDDAYDLPGSRTYRDKMKWK
GKVQSTRNRNSSNNRYESGNNRKKKSPFSAQNYKVTKNKRVQTHPLDFPRSSQNNRTNDYSSQFSYNRDDFPKGPKNKRGRSSSNKSQR
NGRY

Air2; 1 $CX_2CX_4HX_4C$ and 3 $CX_2CX_3GHX_4C$

MEKNTAPFVVDTAPTTPPDKLVAPSIEEVNSNPNELRALRGQGRYFGVSDDDKDAIKEAAPKCNNCSQRGHLKKDCPHIICSYCGATDD
HYSRHCPKAIQCSKCDEVGHYRSQCPHKWKKVQCTLCKSKKHSKERCPSIWRAYILVDDNEKAKPKVLPFHTIYCYNCGGKGHFGDDCK
EKRSSRVPNEDGSAFTGSNLSVELKQEYYRHMNRNSDENEDYQFSESIYDEDPLPRPSHKRHSQNDHSHSGRNKRRASNFHPPPYQKSN
VIQPTIRGETLSLNNNISKNSRYQNTKVNVSSISENMYGSRYNPSTYVDNNSISNSSNYRNYNSYQPYRSGTLGKRR

Msl5/Bbp/Sf1; 1 $CX_2CX_4HX_4C$ and 1 $CX_2CX_3GHX_4C$

MSFRRINSRYFENRKGSSMEEKKAKVPPNVNLSLWRKNTVESDVHRFNSLPSKISGALTREQIYSYQVMFRIQEITIKLRTNDFVPPSR
KNRSPSPPPVYDAQGKRTNTREQRYRKKLEDERIKLVEIALKTIPYFVPPDDYKRPTKFQDKYYIPVDQYPDVNFVGLLLGPRGRTLRK
LQEDSNCKIAIRGRGSVKEGKNASDLPPGAMNFEDPLHCLIIADSEDKIQKGIKVCQNIVIKAVTSPEGQNDLKRGQLRELAELNGTLR
EDNRPCPICGLKDHKRYDCPNRKIPNIQGIVCKICGQTGHFSRDCNSSSQRMSRFDRNATVNNSAPIQSNDVHYNSNTHPIQAPKRSRY
DNNSTEPPLKFPASSRYAPSPSPPASHISRQAQNVTPTPPPGLTSSSFSSGVPGIAPPPLQSPPESEQPKFSLPPPPGMTTVQSSIAPP
PGLSGPPGFSNNMGNDINKPTPPGLQGPPGL

Bik1; 1 $CX_2CX_3GHX_4C$

MDRYQRKIGCFIQIPNLGRGQLKYVGPVDTKAGMFAGVDLLANIGKNDGSFMGKKYFQTEYPQSGLFIQLQKVASLIEKASISQTSRRT
TMEPLSIPKNRSIVRLTNQFSPMDDPKSPTPMRSFRITSRHSGNQQSMDQEASDHHQQQEFGYDNREDRMEVDSILSSDRKANHNTTSD
WKPDNGHMNDLNSSEVTIELREAQLTIEKLQRKQLHYKRLLDDQRMVLEEVQPTFDRYEATIQEREKEIDHLKQQLELERRQQAKQKQF
FDAENEQLLAVVSQLHEEIKENEERNLSHNQPTGANEDVELLKKQLEQLRNIEDQFELHKTKWAKEREQLKMHNDSLSKEYQNLSKELF
LTKPQDSSSEEVASLTKKLEEANEKIKQLEQAQAQTAVESLPIFDPPAPVDTTAGRQQWCEHCDTMGHNTAECPHHNPDNQQFF

Gis2; 1 $CX_2CX_4HX_4C$ and 6 $CX_2CX_3GHX_4C$

MSQKACYVCGKIGHLAEDCDSERLCYNCNKPGHVQTDCTMPRTVEFKQCYNCGETGHVRSECTVQRCFNCNQTGHISRECPEPKKTSRF
SKVSCYKCGGPNHMAKDCMKEDGISGLKCYTCGQAGHMSRDCQNDRLCYNCNETGHISKDCPKA

Mpe1; 1 $CX_2CX_4HX_4C$

MSSTIFYRFKSQRNTSRILFDGTGLTVFDLKREIIQENKLGDGTDFQLKIYNPDTEEEYDDDAFVIPRSTSVIVKRSPAIKSFSVHSRL
KGNVGAAALGNATRYVTGRPRVLQKRQHTATTTANVSGTTEEERIASMFATQENQWEQTQEEMSAATPVFFKSQTNKNSAQENEGPPPP
GYMCYRCGGRDHWIKNCPTNSDPNFEGKRIRRTTGIPKKFLKSIEIDPETMTPEEMAQRKIMITDEGKFVVQVEDKQSWEDYQRKRENR
QIDGDETIWRKGHFKDLPDDLKCPLTGGLLRQPVKTSKCCNIDFSKEALENALVESDFVCPNCETRDILLDSLVPDQDKEKEVETFLKK
QEELHGSSKDGNQPETKKMKLMDPTGTAGLNNNTSLPTSVNNGGTPVPPVPLPFGIPPFPMFPMPFMPPTATITNPHQADASPKK

Slu7; 1 $CX_2CX_3GHX_4C$

MNNNSRNNENRSTINRNKRQLQQAKEKNENIHIPRYIRNQPWYYKDTPKEQEGKKPGNDDTSTAEGGEKSDYLVHHRQKAKGGALDIDN
NSEPKIGMGIKDEFKLIRPQKMSVRDSHSLSFCRNCGEAGHKEKDCMEKPRKMQKLVPDLNSQKNNGTVLVRATDDDWDSRKDRWYGYS
GKEYNELISKWERDKRNKIKGKDKSQTDETLWDTDEEIELMKLELYKDSVGSLKKDDADNSQLYRTSTRLREDKAAYLNDINSTESNYD
PKSRLYKTETLGAVDEKSKMFRRHLTGEGLKLNELNQFARSHAKEMGIRDEIEDKEKVQHVLVANPTKYEYLKKKREQEETKQPKIVSI
GDLEARKVDGTKQSEEQRNHLKDLYG

YOL029C; 1 $CX_2CX_4HX_4C$

MKPVTCCNQKNNIMPSLVPVCCSEKKIESDAKKSISKCCGDKEIYDSENRPITKEDGSWIPGSCKQCRSDPHSRNFCQSLSNKCSSSSF
SSNSALSPDLNEQQTDVNYNSIKLPEICSCKNAQMNAASDAKRYLPISYTYQKIRQHMQKNKSIQEQLNPEDSTSISSALENIASGLHV
RGQKVELQSIKDALHKMDKNVLE

Supplementary Figure 1. Yeast ZCCHC domain-containing proteins. The number of $CX_2CX_4HX_4C$ or $CX_2CX_3GHX_4C$ sequences in each protein is shown; these sequences have been highlighted. A slash separates alternative names given by different authors to the same protein.

AT5G51280 RH35; 1 $CX_2CX_3GHX_4C$

MESIMEEADSYIEYVSVAERRAIAAQKILQRKGKASELEEEADKEKLAEAKPSLLVQATQLKRDVPEVSATEQIILQEKEMMEHLSDKK
TLMSVRELAKGITYTEPLLTGWKPPLHIRKMSSKQRDLIRKQWHIIVNGDDIPPPIKNFKDMKFPRPVLDTLKEKGIVQPTPIQVQGLP
VILAGRDMIGIAFTGSGKTLVFVLPMIMIALQEEMMMPIAAGEGPIGLIVCPSRELARQTYEVVEQFVAPLVEAGYPPLRSLLCIGGID
MRSQLEVVKRGVHIVVATPGRLKDMLAKKKMSLDACRYLTLDEADRLVDLGFEDDIREVFDHFKSQRQTLLFSATMPTKIQIFARSALV
KPVTVNVGRAGAANLDVIQEVEYVKQEAKIVYLLECLQKTSPPVLIFCENKADVDDIHEYLLLKGVEAVAIHGGKDQEDREYAISSFKA
GKKDVLVATDVASKGLDFPDIQHVINYDMPAEIENYVHRIGRTGRCGKTGIATTFINKNQSETTLLDLKHLLQEAKQRIPPVLAELNDP
MEEAETIANASGVKG**CAYCGGLGHRIRDC**PKLEHQKSVAISNSRKDYFGSGGYRGEI

AT5G26742 RH3; 1 $CX_2CX_3GHX_4C$

MASTVGVPSLYQVPHLEISKPNSKKRSNCLSLSLDKPFFTPLSLVRRTRRIHSSSLLVPSAVATPNSVLSEEAFKSLGLSDHDEYDLDG
DNNNVEADDGEELAISKLSLPQRLEESLEKRGITHLFPIQRAVLVPALQGRDIIARAKTGTGKTLAFGIPIIKRLTEEAGDYTAFRRSG
RLPKFLVLAPTRELAKQVEKEIKESAPYLSTVCVYGGVSYTIQQSALTRGVDVVVGTPGRIIDLIEGRSLKLGEVEYLVLDEADQMLAV
GFEEAVESILENLPTKRQSMLFSATMPTWVKKLARKYLDNPLNIDLVGDQDEKLAEGIKLYAIATTSTSKRTILSDLITVYAKGGKTIV
FTQTKRDADEVSLALSNSIATEALHGDISQHQRERTLNAFRQGKFTVLVATDVASRGLDIPNVDLVIHYELPNDPETFVHRSGRTGRAG
KEGSAILMHTSSQKRTVRSLERDVGCHFEFISPPTVGDLLESSADQVVATLNGVHPDSIKFFSATAQKLYEEKGTDALAAALAHLSGFS
QPPSSRSLLSHEKGWVTLQLIRDPTNARGFLSARSVTGFLSDLYRTAADEVGKIFLIADDRIQGAVFDLPEEIAKELLEKDVPEGNSLS
MITKLPPLQDDGPSSDNYGRFSSRDRMPRGGGGSRGSRGGRGGSSRGRDSWGGDDDRGSRRSSGGGSSWSRGGSSSRGSSDDWLIGGRS
SSSSRAPSRERSFGGS**CFICGKSGHRATDC**PDKRGF

AT4G33370 RH43; 1 $CX_2CX_3GHX_4C$

MEVDDGYVEYVPVEERLAQMKRKVVEEPGKGMMEHLSDKKKLMSVGELARGITYTEPLSTWWKPPLHVRKMSTKQMDLIRKQWHITVNG
EDIPPPIKNFMDMKFPSPLLRMLKDKGIMHPTPIQVQGLPVVLSGRDMIGIAFTGSGKTLVFVLPMIILALQEEIMMPIAAGEGPIALV
ICPSRELAKQTYDVVEQFVASLVEDGYPRLRSLLCIGGVDMRSQLDVVKKGVHIVVATPGRLKDILAKKKMSLDACRLLTLDEADRLVD
LGFEDDIRHVFDHFKSQRQTLLFSATMPAKIQIFATSALVKPVTVNVGRAGAANLDVIQEVEYVKQEAKIVYLLECLQKTTPPVLIFCE
NKADVDDIHEYLLLKGVEAVAIHGGKDQEDRDYAISLFKAGKKDVLVATDVASKGLDFPDIQHVINYDMPGEIENYVHRIGRTGRCGKT
GIATTFINKNQSEITLLDLKHLLQEAKQRIPPVLAELNGPMEETETIANASGVKG**CAYCGGLGHRILQC**PKFEHQKSVAISSSRKDHFG
SDGYRGEV

AT4G36020 CSP1; 7 $CX_2CX_3GHX_4C$

MASEDQSAARSTGKVNWFNASKGYGFITPDDGSVELFVHQSSIVSEGYRSLTVGDAVEFAITQGSDGKTKAVNVTAPGGGSLKKENNSR
GNGARRGGGGSG**CYNCGELGHISKDC**GIGGGGGGGERRSRGGEG**CYNCGDTGHFARDC**TSAGNGDQRGATKGGNDG**CYTCGDVGHVARD**
CTQKSVGNGDQRGAVKGGNDG**CYTCGDVGHFARDC**TQKVAAGNVRSGGGGSGT**CYSCGGVGHIARDC**ATKRQPSRG**CYQCGGSGHLARD**
CDQRGSGGGGNDNA**CYKCGKEGHFAREC**SSVA

AT4G38680 CSP2/GRP2; 2 $CX_2CX_3GHX_4C$

MSGDNGGGERRKGSVKWFDTQKGFGFITPDDGGDDLFVHQSSIRSEGFRSLAAEEAVEFEVEIDNNNRPKAIDVSGPDGAPVQGNSGGG
SSGGRGGFGGGRGGGRGSGGGYGGGGGGYGGRGGGGRGGSD**CYKCGEPGHMARDC**SEGGGGYGGGGGGYGGGGGYGGGGGGYGGGGRGG
GGGGGS**CYSCGESGHFARDC**TSGGR

AT2G17870 CSP3/GRP2B; 7 $CX_2CX_3GHX_4C$

MAMEDQSAARSIGKVSWFSDGKGYGFITPDDGGEELFVHQSSIVSDGFRSLTLGESVEYEIALGSDGKTKAIEVTAPGGGSLNKKENSS
RGSGGN**CFNCGEVGHMAKDC**DGGSGGKSFGGGGGRRSGGEGE**CYMCGDVGHFARDC**RQSGGGNSGGGGGGGRP**CYSCGEVGHLAKDC**RG
GSGGNRYGGGGGRGSGGDG**CYMCGGVGHFARDC**RQNGGGNVGGGGST**CYTCGGVGHIAKVC**TSKIPSGGGGGGRA**CYECGGTGHLARDC**
DRRGSGSSGGGGGSNK**CFICGKEGHFAREC**TSVA

AT2G21060 CSP4/GRP2b; 2 $CX_2CX_3GHX_4C$

MSGGGDVNMSGGDRRKGTVKWFDTQKGFGFITPSDGGDDLFVHQSSIRSEGFRSLAAEESVEFDVEVDNSGRPKAIEVSGPDGAPVQGN
SGGGGSSGGRGGFGGGGGRGGGRGGGSYGGGYGGRGSGGRGGGGGDNS**CFKCGEPGHMAREC**SQGGGGYSGGGGGGRYGSGGGGGGGGG
GLS**CYSCGESGHFARDC**TSGGAR

AT5G45400 RPA1C; 1 $CX_2CX_4HX_4C$

MAVSLTEGVVMKMLNGEVTSETDMMPVLQVTELKLIQSKLHQNQESSNRYKFLLSDGTDLAAGMLNTSLNSLVNQGTIQLGSVIRLTHY
ICNLIQTRRIVVIMQLEVIVEKCNIIGNPKEPGHSSINPQRGGVNTQSNGGSEQQQARRSDVNGGRYGVSANSPQPQVVHNSSDAGRYC
VSANSPQPQVVHSSSDAGRYGVSANSPQRQVVHNSPDAGRYGQPQVSQRYGTGSGYPETSPSTRPYVSSNAGYGGSRQDQPRAPTATTA
YSRPVQSAYQPQQPPMYVNRGPVARNEAPPRINPIAALNPYQGRWTIKVRVTSKADLRRFNNPRGEGKLFSFDLLDADGGEIRVTCFND
AVDQFFDKIVVGNVYLISRGNLKPAQKNFNHLPNDYEIHLDSASTIQPCEDDGTIPRYHFHFRNIGDIENMENNSTTDVIGIVSSISPT
VAIMRKNLTEVQKRSLQLKDMSGRSVEVTMWGNFCNAEGQKLQNLCDSGVFPVLALKAGRIGEFNGKQVSTIGASQFFIEPDFPEAREL
RQWYEREGRNAHFTSISREFSGVGRQEVRKVIAQIKDEKLGTSEKPDWITVCATISFMKVENFCYTACPIMNGDRPCSKKVTNNGDGTW
RCEKCDKCVDECDYRYILQIQLQDHTDLTWATAFQEAGEEIMGMSAKDLYYVKYENQDEEKFEDIIRSVAFTKYIFKLKIKEETYSDEQ
RVKATVVKAEKLNYSSNTRFMLEAIDKLKIGDANSLPIKAESSNYRSDAFNSGVGTSGTRDTASVDARREFGLPAANQVGQYGNQYSSD
ARSLGGFTS**CNVCRSNSHVSANC**PTLMSEPQGQYMGGTNAGGGMPRQHVGSY

AT4G19130 RPA1E; 1 $CX_2CX_3GHX_4C$

MEVSLTAGAIGKIMNGEVTTEADMIPVLQVTDLKQIMAQQDPTRERFRMVLSDGTYLHQGMLGTDLNNLVKEGTLQPGSIVRLTRFVGD

VIKGRRIVIVPQLEVLKQISDIIGHPVPGGKHNDQRGADSGIKFNTTEQQGSGIRQVNNIEPGRSNAAISPQVGGTGSSVPASTTPSTR
AYSNPSSGNGVTRQDYARDPPTSYPHQPQPPPPMYANRGPVARNEAPPKIIPVNALSPYSGRWTIKARVTNKAALKQYSNPRGEGKVFN
FDLLDADGGEIRVTCFNAVADQFYDQIVVGNLYLISRGSLRPAQKNFNHLRNDYEIMLDNASTIKQCYEEDAAIPRHQFHFRTIGDIES
MENNCIVDVIGIVSSISPTVTITRKNGTATPKRSLQLKDMSGRSVEVTMWGDFCNAEGQRLQSLCDSGVFPVLAVKAGRISEFNGKTVS
TIGSSQLFIDPDFVEAEKLKNWFEREGKSVPCISLSREFSGSGKVDVRKTISQIKDEKLGTSEKPDWITVSATILYLKFDNFCYTACPI
MNGDRPCSKKVTDNGDGTWRCEKCDKSVDECDYRYILQLQIQDHTDLTCVTAFQEAGEEIMGISAKDLYYVKNEHKDEEKFEDIIRKVA
FTKYNFKLKVKEETFSDEQRVKATVVKVDKLNYSADTRTMLGAMDKLRTRDANSLPINPEGSDYNADVVNTGIGSSGTRDPSSVQRRDF
GLHAHQSGQSGNHYSGGGATTSCNVCGNSGHVSAKCPGATKPQEQGQYMGGSYRGTTGSYGGGLPRQHVGSY

AT5G63920 TOP3A; 1 $CX_2CX_4HX_4C$ and 1 $CX_2CX_3GHX_4C$

MSRRGGGPVTVLNVAEKPSVAKSVAGILSRGTFRTREGRSRYNKIFEFDYAINGQPCRMLMTSVIGHLMELEFADRYRKWHSCDPADLY
QAPVMKHVPEDKKDIKKTLEEEARKSDWLVLWLDCDREGENIAFEVVDVCRAVKHNLFIRRAHFSALIDRDIHEAVQNLRDPNQLFAEA
VDARQEIDLRIGASFTRFQTMLLRDRFAIDSTGEERSRVISYGPCQFPTLGFIVERYWEIQAHEPEEFWTINCSHQSEEGLATFNWMRG
HLFDYASAVILYEMCVEEPTATVMNVPHPRERFKYPPYPLNTIELEKRASRYFRLSSEHTMKVAEELYQAGFISYPRTETDSFSSRTDL
RAMVEEQTRHPAWGSYAQRLLEPEGGLWRNPANGGHDDKAHPPIHPTKFSSGESNWSRDHLNVYELVVRHYLACVSQPAVAAETTVEID
IAGERFSASGRAILAKNYLEVYRFESWGGSVIPVYEKGQQFIPTTLTLDAAVTRPPPLLCEADLLSCMDKAGIGTDATMHDHIKKLLDR
GYATKDANTRFSPTNLGEALVMGYDDMGYELWKPNLRALMEHDMNEVSVGRKTKAEVLETCLQQMKACFLDARVKKSKLLEAMTIFFER
SNNTDESESQTAGEVVRRCNLCNESDMALRKNRDGNFMVGCMNYPQCRNAVWLPGPTLEASVTTNVCQSCGPGPVYKILFKFRQIGIPP
GFDVNHLGCVGGCDDILKQLIDICGTGSRSQARRTPGTAPSNNIQGSNTRQSNVCIHCQQRGHASTNCPSRVPASRNSRPTATNPRNDE
STVSCNTCGSQCVLRTANTEANRGRQFFSCPTQGCSFFAWEDSINNSSGNATTGSNSGGSGRRGSRGRGRGGRGGQSSGGRRGSGTSFV
SATGEPVSGIRCFSCGDPSHFANACPNRNNSNGNYF

AT5G47390 KUA1; 1 $CX_2CX_3GHX_4C$

MTRRCSHCNHNGHNSRTCPNRGVKLFGVRLTEGSIRKSASMGNLSHYTGSGSGGHGTGSNTPGSPGDVPDHVAGDGYASEDFVAGSSSS
RERKKGTPWTEEEHRMFLLGLQKLGKGDWRGISRNYVTTRTPTQVASHAQKYFIRQSNVSRRKRRSSLFDMVPDEVGDIPMDLQEPEED
NIPVETEMQGADSIHQTLAPSSLHAPSILEIEECESMDSTNSTTGEPTATAAAASSSSRLEETTQLQSQLQPQPQLPGSFPILYPTYFS
PYYPFPFPIWPAGYVPEPPKKEETHEILRPTAVHSKAPINVDELLGMSKLSLAESNKHGESDQSLSLKLGGGSSSRQSAFHPNPSSDSS
DIKSVIHAL

AT5G61620 MYB_related-type transcription factor; 1 $CX_2CX_3GHX_4C$

MVKETVTVAKTCSHCGHNGHNARTCLNGVNKASVKLFGVNISSDPIRPPEVTALRKSLSLGNLDALLANDESNGSGDPIAAVDDTGYHS
DGQIHSKKGKTAHEKKKGKPWTEEEHRNFLIGLNKLGKGDWRGIAKSFVSTRTPTQVASHAQKYFIRLNVNDKRKRRASLFDISLEDQK
EKERNSQDASTKTPPKQPITGIQQPVVQGHTQTEISNRFQNLSMEYMPIYQPIPPYYNFPPIMYHPNYPMYYANPQVPVRFVHPSGIPV
PRHIPIGLPLSQPSEASNMTNKDGLDLHIGLPPQATGASDLTGHGVIHVK

AT1G70000 putative MYB_related-type transcription factor; 1 $CX_2CX_3GHX_4C$

MSRSCSQCGNNGHNSRTCPTDITTTGDNNDKGGGEKAIMLFGVRVTEASSSCFRKSVSMNNLSQFDQTPDPNPTDDGGYASDDVVHASG
RNRERKRGTPWTEEEHRLFLTGLHKVGKGDWRGISRNFVKTRTPTQVASHAQKYFLRRTNQNRRRRRSSLFDITPDSFIGSSKEENQLQ
TPLELIRPVPIPIPIPPSRKMADLNLNKKKTPATTEMFPLSLNLQRPSSSTSSSSNEQKARGSRASSGFEAMSSNGDSIMGVA

AT5G56840 putative MYB_related-type transcription factor; 1 $CX_2CX_3GHX_4C$

MGRRCSHCGNVGHNSRTCSSYQTRVVTSSSPPPPPPPSILAAAIKKSFSMDCLPACSSSSSSFAGYLSDGLAHKTPDRKKGVPWTAEEH
RTFLIGLEKLGKGDWRGISRNFVVTKSPTQVASHAQKYFLRQTTTLHHKRRRTSLFDMVSAGNVEENSTTKRICNDHIGSSSKVVWKQG
LLNPRLGYPDPKVSVSGSGNSGGLDLELKLASIQSPESNIRPISVT

AT2G28910 CXIP4; 1 $CX_2CX_3GHX_4C$

MPATAGRVRMPANNRVHSSAALQTHGIWQSAIGYDPYAPTSKEEPKTTQQKTEDPENSYASFQGLLALARITGSNNDEARGSCKKCGRV
GHLTFQCRNFLSTKEDKEKDPGAIEAAVLSGLEKIRRGVGKGEVEEVSSEEEEESESSDSDVDSEMERIIAERFGKKKGGSSVKKTSSV
RKKKKRVSDESDSDSDSGDRKRRRRSMKKRSSHKRRSLSESEDEEEGRSKRRKERRGRKRDEDDSDESEDEDDRRVKRKSRKEKRRRRS
RRNHSDDSDSESSEDDRRQKRRNKVAASSDSEANVSGDDVSRVGRGSSKRSEKKSRKRHHRKERE

AT5G58760 DDB2; 1 $CX_2CX_3GHX_4C$

MSSTRSRRKRDPEIVIARDTDSELSSSEEEEEEEDNYPFSESEEEDEAVKNGGKIELEKNKAKGKAPITVKLIKKVCKVCKQPGHEAGF
KGATYIDCPMKPCFLCKMPGHTTMSCPHRVVTDHGILPTSHRNTKNPIDFVFKRQLQPRIPPIKPKYVIPDQVHCAVIRYHSRRVTCLE
FHPTKNNILLSGDKKGQIGVWDFGKVYEKNVYGNIHSVQVNNMRFSPTNDDMVYSASSDGTIGYTDLETGTSSTLLNLNPDGWQGANSW
KMLYGMDINSEKGVVLAADNFGFLHMIDHRTNNSTGEPILIHKQGSKVCGLDCNPVQPELLLSCGNDHFARIWDMRKLQPKASLHDLAH
KRVVNSAYFSPSSGTKILTTCQDNRIRIWDSIFGNLDLPSREIVHSNDFNRHLTPFKAEWDPKDTSESLIVIGRYISENYNGTALHPID
FIDASNGQLVAEVMDPNITTITPVNKLHPRDDVLASGSSRSLFIWRPQDNTEMVEEKKDKKIIICYGDSKKKGKKQKRGSDDEDDEDDI
FSSKGKNIKVNKYQAKTTKKTKT

AT5G38600 proline-rich spliceosome-associated (PSP) family protein; 1 $CX_2CX_4HX_4C$

METEDVLDIPASSNFGSEVKKNSLESGNGSPEANSLVGNDENVKGNLDLDLTEENLRIVGGQESGEILTEQVSDVFNASVESVAVDEKL
GIQKETLVHSTTLDVSSKAGVKRPRTSYDEQQPTVHVTYKHLTRASKQKLESLLQKWSEWEAENTSLAQDQEQLFESGEETCFPAIRVG
LQKTSSVSFWIDNQTGHKPLEDFVLVESSTTPLYDRKFAIGLNSADGSRNVEGGLEIIDDDPPRCFNCGGYSHSLRECPRPFDRSAVNS
ARKLQKSKRNQNSSGPRLPSRYYQKTQTGKYDGLKPGTLDAETRQLLNLGELDPPPWLNRMREIGYPPGYLAPEDDHLSGITIFGEEVE

TREEIESEDGEILEKANHPPEPQMKKTVEFPGINAPFPENADEWLWEAAPSHRNSSRSGRWQQQKTSRGHDYRDDGPLGVEPSSYPPRY
GSRYDYGYGSNEYGSRSRSPGIDRSLSERSKRDYSSYDADFRERDRDRDRNRDRDRDRDWDRDRDRCDRDRDRDDRDWSYRLSSRR

AT5G43630 TZP; 1 $CX_2CX_3GHX_4C$ and 1 $CX_2CX_4HX_4C$

MGDGDEQSKELGGVSSSSRRCSSGTAGAANAEARMKFAAVDAITELVWSPSNGLSLRCADISFTGKAKLLSPNFFDIGLTNMAIHSNST
SIEDQEDHVDVELRNRDQVNQAMIGGSVEDMKPEMVEDKVETNDDIKNEEAGCSKRSSDSPKAMEGETRDLLVNEQLRMESAGSQEEGD
KAHNRVDRLESMDENNLATLAVVACEGKGDYLPEGEAGPSGSYRRREKAKGKEKALSDENFGGDGEDEDEESFGSVESCNSAGLLSRGK
KRPGFEEQLIFGSKRLKTLNQECLGSTSKLKQDSSFMNWISNMTKGIWKGNEEDNSPFVALTTTSNANGHGQVNAIVDQQQLSPCCVKE
NSGCRNTGFQSFFQSIYCPKKQSQDVVDMDFPNDVNAAPLQELPWIPEHCDISKGDDLSSSGNEIGPVAEPNISSGKVVFNQTSKTQSS
ENKREDKEPNISLMSLSKSKPNEEPKTCGEADGKVSPCLTNRNSGLKSLWISRFSSKGSFPQKKASETAKEANASASDAAKTRDSRKML
ADKNVIRPSISSVDGPDKPDTVLPIVSSMRIESSEAMASLFARRLEAMKSIMPSGSLAENAEEEQRDLI**CFYCGKKGHCLRDC**LEVTDT
ELRDLVQNISVRNGREEASSL**CIRCFQLSHWAATC**PNAPLYGSGAEGRAMKNALASTSGMKLPISGFTDVPRAVFDAVQVLRLSRTDVL
KWINTKKSVSGLEGFFLRLRLGKWEEGLGGTGYYVARIDGDTEGQSSRRHSEKSLISVKVKGVTCLVESQFISNQDFLEEELKAWWQSA
GKSARTSGYDGIPSAEELSRKIQQRKMLGF

AT5G51300 SF1; 2 $CX_2CX_3GHX_4C$

MESVEMNNPNSQTLDQPPPSSNGDTAPLALDHMNPQNSESVALNGSSTPIPDTNGSSAKPELLRPLLSENGVSKTLSGNDKDQSGGEEE
TTSRRKRRSRWDPPPSESINNPSAEGGTDSGTGTRKRKSRWADDEPRTQIQLPDFMKDFTGGIEFDPEIQALNSRLLEISRMLQSGMPL
DDRPEGQRSPSPEPVYDNMGIRINTREYRARERLNRERQEIIAQIIKKNPAFKPPADYRPPKLHKKLFIPMKEFPGYNFIGLIIGPRGN
TQKRMERETGAKIVIRGKGSVKEGRHQQKKDLKYDPSENEDLHVLVEAETQEALEAAAGMVEKLLQPVDEVLNEHKRQQLRELATLNGT
IRDEEF**CRLCGEPGHRQYAC**PSRTNTFKSDVL**CKICGDGGHPTIDC**PVKGTTGKKMDDEYQNFLAELGGTVPESSLKQSATLALGPGSS
GSNPPWANNAGNGASAHPGLGSTPTKPPSKEYDETNLYIGFLPPMLEDDGLINLFSSFGEIVMAKVIKDRVTGLSKGYGFVKYADVQMA
NTAVQAMNGYRFEGRTLAVRIAGKSPPPIAPPGPPAPQPPTQGYPPSNQPPGAYPSQQYATGGYSTAPVPWGPPVPSYSPYALPPPPPG
SYHPVHGQHMPPYGMQYPPPPPHVTQAPPPGTTQNPSSSEPQQSFPPGVQADSGAATSSIPPNVYGSSVTAMPGQPPYMSYPSYYNAVP
PPTPPAPASSTDHSQNMGNMPWANNPSVSTPDHSQGLVNAPWAPNPPMPPTVGYSQSMGNVPWAPKPPVQPPAENPSSVGESEYEKFMA
EMK

AT1G23860 RSZ21; 1 $CX_2CX_3GHX_4C$

MTRVYVGNLDPRVTERELEDEFKAFGVLRNVWVARRPPGYAFLEFDDERDALDAISALDRKNGWRVELSHKDKGGRGGGGGRRGGIEDS
K**CYECGELGHFAREC**RRGRGSVRRRSPSPRRRRSPDYGYARRSISPRGRRSPPRRRSVTPPRRGRSYSRSPPYRGSRRDSPRRRDSPYG
RRSPYANGV

AT4G31580 RSZ22; 1 $CX_2CX_3GHX_4C$

MSRVYVGNLDPRVTERELEDEFRAFGVVRSVWVARRPPGYAFLDFEDPRDARDAIRALDGKNGWRVEQSHNRGERGGGGRGGDRGGGGG
GRGGRGGSDLK**CYECGETGHFAREC**RNRGGTGRRRSKSRSRTPPRYRRSPSYGRRSYSPRARSPPPPRRRSPSPPPARGRSYSRSPPPY
RAREEVPYANGNGLKERRRSRS

AT2G24590 RSZ22A; 1 $CX_2CX_3GHX_4C$

MSRVYVGNLDPRVTERELEDEFRSFGVIRSVWVARRPPGYAFLDFEDSRDARDAIREVDGKNGWRVEQSHNRGGGGGRGGGRGGGDGGR
GRGGSDLK**CYECGESGHFAREC**RSRGGSGGRRRSRSRSRSPPRYRKSPTYGGRRSYSPRARSPPPPRRRSPSPRGRNYSRSPPPYRARD
EVPYANGNGLKDVRRSRS

AT3G53500 RSZ32; 2 $CX_2CX_3GHX_4C$

MPRYDDRYGNTRLYVGRLSSRTRTRDLERLFSRYGRVRDVDMKRDYAFVEFSDPRDADDARYYLDGRDFDGSRITVEASRGAPRGSRDN
GSRGPPPGSGR**CFNCGVDGHWARDC**TAGDWKNK**CYRCGERGHIERNC**KNSPSPKKARQGGSYSRSPVKSRSPRRRRSPSRSRSYSRGRS
YSRSRSPVRREKSVEDRSRSPKAMERSVSPKGRDQSLSPDRKVIDASPKRGSDYDGSPKENGNGRNSASPIVGGGESPVGLNGQDRSPI
DDEAELSRPSPKGSESP

AT2G37340 RSZ33; 2 $CX_2CX_3GHX_4C$

MPRYDDRYGNTRLYVGRLSSRTRTRDLERLFSRYGRVRDVDMKRDYAFVEFGDPRDADDARHYLDGRDFDGSRITVEFSRGAPRGSRDF
DSRGPPPGAGR**CFNCGVDGHWARDC**TAGDWKNK**CYRCGERGHIERNC**KNSPKKLRRSGSYSRSPVRSRSPRRRRSPSRSLSRSRSYSRS
RSPVRRRERSVEERSRSPKRMDDSLSPRARDRSPVLDDEGSPKIIDGSPPPSPKLQKEVGSDRDGGSPQDNGRNSVVSPVVGAGGDSSK
EDRSPVDDDYEPNRTSPRGSESP

AT1G60650 RZ-1B/RBGB1; 1 $CX_2CX_3GHX_4C$

MKDRENDGNLESRIFVGGLSWDVTERQLESTFDRYGKITECQIMVGRDTGRPRGFGFITFTDRRGADDAIKHMHGRELGNKVISVNKAE
PKVGGEDVDQLKKGGGYSSRGKGTEDE**CFKCRRPGHWARDC**PSTGDDRERFRVPLAMRSRIGDIDGHRDRYGDRDLERERERERREFDRY
MDGRRDRDGGRYSYRDRFDSGDKYEPRDHYPFERYAPPGDRFVSDRYGMPEHHLENEYRGRERSYDRDRYARDTSDRYGDMGPIRDEGR
PYRSRPGPYDRPSRPGGRPSSYERW

AT5G04280 RZ-1C/RBGB3; 1 $CX_2CX_3GHX_4C$

MAAKEGSRIFVGGLSPEVTDRDLERAFSRFGDILDCQIMLERDTGRSRGFGFITFADRRAMDESIREMHGRDFGDRVISVNRAEPKLGR
DDGESHGSRGGRDSGYSIAGKGSFGGGGGGGGRVGEDE**CFKCGRVGHWARDC**PSAGGGRGGPVGGFSSRASAYGGSDGRVDRYADRDRY
VDRERYIDDRYDGAARYGARDRFDSREAYIPRDRYASDRYAAPADRFAGGDRYSRGSDRYPPGSYDKARSFERDIAPSAGSDRYGGGRA
GGPIRGGGEEGRGFRSRAGAPYERPSRSGGGGAYPSSSTFDRY

AT3G10400 U11/U12-31K; 1 $CX_2CX_3GHX_4C$

MKKRKIHHSDDEEDDTFYYRYSSVAAPPPSNPKHQPSSSAKSSAPGGGSGGLAPSKSTLYVSNLDFSLTNSDIHTLFSTFGKVARVTVL
KDRHTRQSRGVAFVLYVSREDAAKAARSMDAKILNGRKLTVSIAADNGRASEFIKKRVYKDKSR**CYECGDEGHLSYEC**PKNQLGPRERP
PPPKKRGRRKEEEGEAEEISWSAAAPSLAVAEEEFEEENWASVVDNEAGERLRKREAEEEEERRMKRKEKKVSYFSDESDDED

AT3G55340 PHIP1; 2 $CX_2CX_3GHX_4C$

MVLSNKKLKQRIRQDLAESLSVSVSETNPQSQSLKLLLDSSSHKPRLSKREKRRNCETFAREDDEIRENEVGNGGSSEKTDTKIKKKRK
RDDAVEVDELEGDEGTKEEQKPQKKKNKKKKKRKVNKTPKKAEEGNVEEKVKVEEIEVNTDNKEEDGVVPNKLYVGGIPYQSTEDEIR
SYFRSCGVIIKVDCKMRPEDGAFSGIAFITFDTEDGAKRALAFDRAAMGDRYLTIQQYVKTTTPSIPRRKTSSGFAPEMVDGYNRVYIG
NLAWDTTERDIRKLFSDCVINSVRLGKNKETGEFKGYAHVDFKDSVSVAIALKLDQQVICGRPVKICCALKDRPATDHTPGETNNAGSY
NMEDTYAAADPVPALAGRSEVDDGNYFATTVSSSKVKRRV**CYECGEKGHLSTAC**PIKLQKADDQANSKLGQETVDGRPAMQSYGLPKNS
GDSYYMNETYASTNETYNGGYSASAVGTGKVKRRN**CYECGEKGHLSTAC**PIKLQNTSHTNSTLDHQTVEAGPTQVTSYSLQKKTRDTEN
NGGSFMDESYATVPISIDVTNGANDASLTSAVSTGKIKKRNCYECGEKGHLSSACPNKLQKQG

AT1G53720 CYP59; 1 $CX_2CX_4HX_4C$

MSVLIVTSLGDIVIDLHSDKCPLTCKNFLKLCKIKYYNGCLFHTVQKDFTAQTGDPTGTGAGGDSIYKFLYGEQARFYKDEIHLDLKHS
KTGTVAMASGGENLNASQFYFTLRDDLDYLDGKHTVFGQIAEGFDTLTRINEAYVDPKNRPYKNIRIKHTHILDDPFDDPPQLAEMMPD
ASPEGKPKEEVKDDVRLEDDWVPMDEELGAQELEEVIREKAAHSSAVVLESIGDIPEAEVKPPDNVLFVCKLNPVTEDEDLHTIFSRFG
TVVSADVIRDFKTGDSLCYAFIEFENKESCEQAYFKMDNALIDDRRIHVDFSQSVSKLWSQFRQKDSQKGKGNG**CFKCGSTDHIAKDC**V
GGPSSKFIVKDQNRQHGGGEGYEMVFEGDVHETPKHNSHERERSEKIQRRSPHGNGEGKRQHRDERDDGRRQHDREDARELERKHRERK
ERESREDEDRRRRRRREESRDKESRRERDEDDHRSHRDYKERRRERDDRHGREARHERRDR

AT5G49930 EMB1441; 1 $CX_2CX_3GHX_4C$

MVKVRMNTADVAAEVKCLKRLIGMRCSNVYDISPKTYMFKLLNSSGITESGESEKVLLLMESGVRLHTTAYVRDKSNTPSGFTLKLRKH
IRTRRLEDVRQLGYDRIIVFQFGLGANAHYVILELYAQGNIILTDSEYMIMTLLRSHRDDNKGFAIMSRHRYPIEICRVFERTTVSKLQ
ESLTAFVLKDHDAKQIEPKEQNGGKKGGKSNDSTGAKQYTLKNILGDALGYGPQLSEHIILDAGLVPTTKLSEDKKLDDNEIQLLVQAV
IVFEDWLEDIINGQKVPEGYILMQKQILANDTTSESGGVKKMYDEFCSILLNQFKSRVYEKFETFDAALDEFYSKIESQRSEQQQKAKE
DSASLKLNKIRQDQENRVQILKKEVNHCVNMAELIEYNLEDVDAAILAVRVALAKGMGWDDLARMVKEEKKLGNPVAGVIDRLYLEKNC
MTLLLCNNLDEMDDDEKTVPVEKVEVDLSLSAHGNARRWYEMKKKQETKQEKTVSAHEKAFRAAEKKTRHQLSQEKVVATISHMRKVHW
FEKFNWFISSENYLVISGRDAQQNEMIVKRYMSKGDLYVHAELHGASSTVIKNHKPEQNVPPLTLNQAGCFTVCHSQAWDSKIVTSAWW
VYPHQVTKTAPTGEYLTVGSFMIRGKKNFLPPHPLIMGFGLLFRLDESSLGAHLNERRVRGEEEGMNDVVMETHAPDEHSDTESENEAV
NEVVSASGEVDLQESSTALSQDTSSLDMSSSGITEENVASATSQLEDLLDRTLGLGAATVAGKKDTIETSKDDMEEKMKQEEKNAVVRD
KPYMSKAERRKLKMGQSGNTAADGNTGQEKQQRKEKDVSSLSQATKSIPDNKPAGEKVSRGQRGKLKKMKEKYADQDEDERKIRMALLA
SSGKPQKTDVESQNAKTAVTEVKKPSEETDDAVKI**CYRCKKVGHLARDC**HGKETSDMDKVVMEEDDIHEVGDEEKEKLIDVDYLTGNPL
PTDILLYAVPVCGPYNALQSYKYRVKAIPGSMKKGKAAKTAMNLFTHMSEASVREKELMKACTDPELMAALVGNVKITAAGLTQLKQKQ
KKGKKSGKQQHS

AT5G47430 PQT3L; 1 $CX_2CX_3GHX_4C$

MAIYYKFKSARDYDTIAMDGPFISVGILKDKIFETKHLGTGKDLDIVVSNAQTNEEYLDEAMLIPKNTSVLIRRVPGRPRITVITTQEP
RIQNKVEDVQAETTNFPVADPSAAEFPEDEYDEFGTDLYSIPDTQDAQHIIPRPHLATADDKVDEESKIQALIDTPALDWQQRQGQDTF
GAGRGYGRGMPGRMNGRGFGMERKTPPPGYV**CHRCNIPGHFIQHC**PTNGDPNYDVKRVKPPTGIPKSMLMATPDGSYSLPSGAVAVLKP
NEDAFEKEMEGLPSTTRSVGELPPELKCPLCKEVMKDAALTSKCCYKSFCDKCIRDHIISKSMCVCGRSDVLADDLLPNKTLRDTINRI
LEAGNDSTENVGSVGHIPDLESARCPPPKALSPTTSVASKGEKKPVLSNNNDASTLKAPMEVAEITSAPRASAEVNVEKPVDACESTQG
SVIVKEATVSKLNTQAPKEEMQQQVAAGEPGKKKKKKPRVPGNDMQWNPVPDLAGPDYMMQMGPGPQYFNGMQPGFNGVQPGFNGVQPG
FNGFHPGFNGFGGPFPGAMPPFMGYGLNPMDMGFGGGMNMMHPDPFMAQGFGFPNIPPPHRDLAEMGNRMNLQRAMMGRDEAEARNAEM
LRKRENERRPEGGKMFRDGENSRMMMNNGTSASASSINPNKSRQAPPPPIHDYDRRRRPEKRLSPEHPPTRKNISPSRDSKRKSERYPD
ERDRQRDRERSRHQDVDREHDRTRDRRDEDRSRDHRHHRGETERSQHHHRKRSEPPSSEPPVPATKAEIENNLKSSVFARISFPEEETS
SGKRRKVPSSSSTSVTDPSASASAAAAVGTSVHRHSSRKEIEVADYESSDEDRHFKRKPSRYARSPPVVVSDVSEDKLRYSKRGKGERS
RA

AT4G17410 PQT3; 1 $CX_2CX_4HX_4C$

MAIYYKFKSARDYDTISMDGPFITVGLLKEKIYETKHLGSGKDLDIVISNAQTNEEYLDEAMLIPKNTSVLIRRVPGRPRIRIITREEP
RVEDKVENVQADMNNVITADASPVEDEFDEFGNDLYSIPDAPAVHSNNLCHDSAPADDEETKLKALIDTPALDWHQQGADSFGPGRGYG
RGMAGRMGGRGFGMERTTPPPGYV**CHRCNVSGHFIQHC**STNGNPNFDVKRVKPPTGIPKSMLMATPNGSYSLPSGAVAVLKPNEDAFEK
EMEGLTSTTRSVGEFPPELKCPLCKEVMRDAALASKCCLKSYCDKCIRDHIIAKSMCVCGATHVLADDLLPNKTLRDTINRILESGNSS
AENAGSMCQVQDMESVRCPPPKALSPTTSAASGGEKKPAPSNNNETSTLKPSIEIAEITSAWASAEIVKVEKPVDASANIQGSSNGKEA
AVSQLNTQPPKEEMPQQVASGEQGKRKKKKPRMSGTDLAGPDYMMPMGPGPGNQYFNGFQPGFNGVQHGFNGVQPGFNGFHHGFNGFPG
PFPGAMPPFVGYGFGGVIHPDPFAAQGFGFPNIPPPYRDLAEMGNRMNLQHPIMGREEFEAKKTEMKRKRENEIRRSEGGNVVRDSEKS
RIMNNSAVTSSPVKPKSRQGPPPPISSDYDRRRRSDRSSPERQSSRRFTSPPRSSSRKSERDRHHDLDSEHDRRRDRPRETDRKHRKRS
EKSSSDPTVEIDDNNKSNVFTRISFPEESSGKQRKTSKSSPAPPESSVAPVSSGRRHHSRREREMVEYDSSDDEDRHFKRKPSRYKRSP
SVAPSDAGDEHFRHSKRSKGERARA

AT3G45950 Pre-mRNA splicing Prp18-interacting factor; 1 $CX_2CX_4HX_4C$

MATASVAFKSRKDHRKQKELEEARKAGLAPAEVDEGGKEINLHIPKYLTIPPLYAKSEKPSLKHQKNWKTKPVSTTSYYDRGAKTYQAE
KYRKGA**CQNCGAMTHDVKTC**MERPRKVGAKYTDKNIAPDEKIESLEFDYDGKRDRWNGYDPSSYCHVRDRHEAKENAREKYLNEQQLIA
KLEEKNIDGEEEDLRVDEAKIDESMQVDFAKVKKRVRTTDGGSKGTVRNLRIREDPAKYLLNLDVNSAYYDPKSRSMREDPLPYTDPNE

KFCLRDNQYRNSGQAIEFKQQNMYSCEAFDKGQDIHMQAAPSQAELCYKRVKIAKEKLNSQRKDAIIAKYGDAAAKDDIPMELLLGQSK
LIKTSQANGIKLVPNVFVNLCFLVFVFIS

AT1G65660 SMP1/SLU7A; 1 $CX_2CX_4HX_4C$

MATASVAFKSREDHRKQIELEEARKAGLAPAEVDEDGKEINPHIPQYMSSAPWYLNSEKPSLKHQRKWKSDPNYTKSWYDRGAKIFQAE
KYRKGACQNCGAMTHTAKACMDRPRKIGAKYTNMNIAPDEKIESFELDYDGKRDRWNGYDPSTYHRVIDLYEAKEDARKKYLKEQQLKK
LEEKNNNEKGDDANSDGEEDEDDLRVDEAKVDESRQMDFAKVEKRVRTTGGGSTGTVRNLRIREDTAKYLLNLDVNSAHYDPKTRSMRE
DPLPDADPNDKFYLGDNQYRNSGQALEFKQLNIHSWEAFDKGQDMHMQAAPSQAELLYKSFQVAKEKLKSQTKDTIMDKYGNAATEDEI
PMELLLGQSERQVEYDRAGRIIKGQEVILPKSKYEEDVHANNHTSVWGSYWKDHQWGYKCCQQIIRNSYCTGSAGIEAAEAALDLMKAN
IARKEATEESPKKVEEKRMASWGTDIPEDLELNEEALANALKKEDLSRREEKDERKRKYNVKYNNDVTPEEMEAYRMKRVHHEDPMKDF
L

AT4G37120 SMP2/SLU7B; 1 $CX_2CX_4HX_4C$

MATASVAFKSREDHRKKLELEEARKAGLAPAEVDEDGKEINPHIPEYMSKAPWYLKSEQPSLKHQKNWKIEPEPKKIWYDRGKKIYQAE
QYRKGACINCGAMTHSSKACMDRPRKIGAKYTNMNIAADEKIESFELDYDGKRDRWNGYDTSTYRHVVDRYDAKEEARKKYLKEQQLKK
LEEKNNNENGDDATSDGEEDLDDLRVDEAKVDESRQMDFAKVEKRVRTTGGGSTGTVRNLRIREDTAKYLLNLDVNSAHYDPKTRSMRE
DPLPDADPNEKFYLGDNQYRNSGQALEFKQINIHSCEAFDKGHDMHMQAAPSQAELLYKNFKVAKEKLKTQTKDTIMEKYGNAATEGEI
PMELLLGQSERQIEYDRAGRIMKGQEVIIPKSKYEEDVHANNHTSVWGSWWKDHQWGYKCCQQTIRNSYCTGSAGIEAAEASIDLMKAN
IARKEASKESPKKVEEKKMATWGTDIPEDLELNEEALANALKKEDLSRREEKDERKRKYNVNYTNDVTSEEMEAYRMKRVHHEDPMRNF
PG

AT5G42540 XRN2; 1 $CX_2CX_3GHX_4C$

MGVPSFYRWLIQRYPLTIQEVIEEEPLEVNGGGVTIPIDSSKPNPNGYEYDNLYLDMNGIIHPCFHPEDKPSPTTFTEVFQCMFDYIDR
LFVMVRPRKLLFMAIDGVAPRAKMNQQRARRFRAAKDAAEAAAEEEQLREEFEREGKKLPPKVDSQVFDSNVITPGTEFMATLSFALRY
YIHVRLNSDPGWKNIKVILSDANVPGEGEHKIMSYIRCNKNHPGYNPNTHHCLYGLDADLIMLSLATHEIHFSILREVVFFPGEEGKCF
LCGQEGHRAADCEGKIKRKTGEMLDNTEADVVVKKPYEFVNIWILREYLEHDMQIPGAKKNLDRLIDDFIFICFFVGNDFLPHMPTLEI
REGAIELLMSVYKNKFRSAKKYLTDSSKLNLRNVERFIKAVGMYENQIFQKRAQVQQRQSERFRRDKARDKARDNARDNAQASRQFSGK
LVQLDSLDEVSDSLHSSPSRKYLRLSLDDNIGVANVETENSLKAEELDNEEDLKFKLKKLLRDKGDGFRSGNGEQDKVKLNKVGWRERY
YEEKFAAKSVEEMEQIRRDVVLKYTEGLCWIMHYYYHGVCSWNWFYPYHYAPFASDLKGLEKLDIKFELGSPFKPFNQLLAVLPSASAH
ALPECYRSLMTNPDSPIADFYPADFEIDMNGKRYSWQGISKLPFVEEKRLLEAAAQVEKSLTNEEIRRNSALFDMLFVVASHPLGELIR
SLNSRTNNLSNEERATIIEKIDPGLSDGMNGYIASCGGDSQPSCFCSTVEGMEDVLTNQVICAIYKLPEDIRGSEITHQIPRLAIPKKT
ISLVDLKSGGLLWHEDGDKRRAPPKVIKIKRYNPEGSISGGRLGKASHRLVLQTINAQPDYMNINSEPALCPNTVFQNERVPKKIPTFK
DNGIQWISPPPSQITPKKMNSPQRQKAWKKDETPQSREKSKKLKSSLKVNPLKMKKTKSPQREFTREKKKENITPQRKLTKAQRQVKHI
RMMEEAKMIKQRKKEKYLRKKAKYAQGAPPKTA

AT1G75660 XRN3; 1 $CX_2CX_3GHX_4C$

MGVPSFYRWLAEKYPLLVADVIEEEPVEIEGIKIPVDTSKPNPNNLEYDNLYLDMNGIIHPCFHPEDRPSPTTFEEVFQCMFDYIDRLF
VMVRPRKLLYMAIDGVAPRAKMNQQRSRRFRSAKDASDAAAEEERLREEFEREGRRLPPKVDSQVFDSNVITPGTEFMGVLSIALQYYV
HLRLNHDVGWKNIKVILSDANVPGEGEHKIMSYIRLQRNLPGFDPNTRHCLYGLDADLIMLGLATHEVHFSILREVVYTPGQQERCFLC
GQMGHFASNCEGKPKKRAGESDEKGDGNDFVKKPYQFLHIWVLREYLELEMRIPNPPFEIDLERIVDDFIFICFFVGNDFLPHMPTLEI
REGAINLLMAVYKKEFRSFDGYLTDGCKPNLKRVEQFIQAVGSFEDKIFQKRAMQHQRQAERVKRDKAGKATKRMDDEAPTVQPDLVPV
ARFSGSRLASAPTPSPFQSNDGRSAPHQKVRRLSPGSSVGAAIVDVENSLESDERENKEELKTKLKELIREKSDAFNSDTTEEDKVKLG
QPGWRERYYEEKFSVVTPEEMERVRKDVVLKYTEGLCWVMHYYMEGVCSWQWFYPYHYAPFASDLKDLGEMDIKFELGTPFKPFNQLLG
VFPAASSHALPERYRTLMTDPNSPIIDFYPTDFEVDMNGKRFSWQGIAKLPFIDERRLLEAVSEVEFTLTDEEKRRNSRMCDMLFIATS
HRLAELVFSLDNHCRQLSARERVDFKVKIKPKLSDGMNGYLTPCSGETHPPVFRSPMEGMEDILTNQVICCIYRLPDAHEHITRPPPGV
IFPKKTVDIGDLKPPPALWHEDNGRRPMHNNHGMHNNHGMHNNQGRQNPPGSVSGRHLGNAAHRLVSNSLQMGTDRYQTPTDVPAPGYG
YNPPQYVPPIPYQHGGYMAPPGAQGYAQPAPYQNRGGYQPRGPSGRFPSEPYQSQSREGQHASRGGGYSGNHQNQHQQQQWHGQGGSEQ
NNPRGYNGQHHHQQGGDHDRRGRGRGSHHHHDQGGNPRHRY

AT1G54490 XRN4; 1 $CX_2CX_4HX_4C$

MGVPAFYRWLADRYPKSISDVVEEEPTDGGRGDLIPVDITRPNPNGFEFDNLYLDMNGIIHPCFHPEGKPAPATYDDVFKSMFEYIDHL
FTLVRPRKILYLAIDGVAPRAKMNQQRSRRFRAAKDAAEAEAEEERLRKDFEMEGQILSAKEKAETCDSNVITPGTPFMAILSVALQYY
IQSRLNHNPGWRYVKVILSDSNVPGEGEHKIMSYIRLQRNLPGFDPNTRHCLYGLDADLIMLSLATHEVHFSILREVITYPGQQEKCFV
CGQTGHFASDCPGKSGSNNAAADIPIHKKKYQFLNIWVLREYLQYELAIPDPPFMINFERIIDDFVFLCFFVGNDFLPHMPTLEIREGA
INLLMHVYRKEFTAMGGYLTDSGEVLLDRVEHFIQAVAVNEDKIFQKRTRIKQSMDNNEEMKQRSRRDPSEVPPEPIDDKIKLGEPGYK
ERYYAEKFSTTNPEETEQIKQDMVLKYVEGLCWVCRYYYQGVCSWQWFYPYHYAPFASDLKNLPDLEITFFIGEPFKPFDQLMGTLPAA
SSNALPGEYRKLMTDPSSPILKFYPADFELDMNGKRFAWQGIAKLPFIEEKLLLAATRKLEETLTVEEQQRNSVMLDLLYVHPAHPLGQ
RILQYYHFYQHMPPHECLPWMIDPNSSQGMNGFLWFSERNGFQTRVDSPVNGLPCIEQNRALNVTYLCPAKHSHISEPPRGAIIPDKIL
TSVDIKPFPPLWHEDNSNRRRQARDRPQVVGAIAGPSLGEAAHRLIKNTLNMKSSTGAASGLIDPNGYYRNVPGNYSYGGVNRPRAPGP
SPYRKAYDDDSSYYYGKYNNSTQGTFNNGPRYPYPSNGSQDYNRNYNSKIVAEQHNRGGLGAGMSGLSIEDNGRSKQLYSSYTEAANAN
LNPLPSPPTQWIGTQPGGNFVGGYYRDGVGYSETNGKSVKKVIYQAKTQPSHRGANL

AT1G54930 protein of unknown function; 1 $CX_2CX_4HX_4C$

MSSGDCFHCHQPGHWAKNCPLKTTTKPTAAAAPSPPDIHCPCNAGPCNTVTSKTEKNPNRRFYTCPSCGYFKWCDQGLGDCGFFKWEDG
ESLLHETELDSDGNVKRNRLGVVVELELNPASSSESNTLGNRIVDTLPVNPVTVGKESIPVFAGFNDQGSVSNSIVPSFDLITLYDDAV
RLETKEPVLPSVAPKHLDTQVETLCGNALEAVESSSQNTSDLVLNANNKPEHIHQRAATSGETEASYSGSSSMMVLIEQYKSEKLYLES

ISMKHVEALTAYTGSYKQLESLRDRAHSLKKQLLEVEKQVKLCEAETSEFAASVQEVSGEMAKSQKKMVEIAGKVAKEVRVDKQRD

AT1G67210 proline-rich spliceosome-associated (PSP) family protein; 1 $CX_2CX_4HX_4C$

MAASSGSGLEAEEGEISIDMEEDMDLTEDDFRNVSGQFSGQASIVEVGDAVDVRVETVKVDVSSKSGVKRARTISLEQQPSVHVTYKHL
TRDSKQKLESLLQQWSEWEAEQNSLSEDQEQVLEAGDETYFPALRVGLQKTSSVSFWFDYQTGHSSSKKSVPVESSTTPLYNRGFTIGL
DSGSNNVEGGLEIIDDPPRCFNCGAYSHSIRECPRPFDRSAVSNARRQHKRKRNQTPGSRLPSRYYQSLQRGKYDGLKPGSLDAETRKL
LGLKELDPPPWLNRMREIGYPPGYFAVEEDDDDHSRITIFGEEETKEEEEVKTEEGEILEKASPQEPRKIMTVGFPGINAPIPENADSW
LWEQRNSNTGHTNYHNHLRPQYEMGPLGIQLSSSFPPMHGIRYDHRFGL

AT1G75560 protein of unknown function; 8 $CX_2CX_3GHX_4C$

MSSMSRSRSRSRSRSRDRFRSRSPRDRRMRSERVSYHDAPSRREREPRRAFSQGNLCNNCKRPGHFARDCSNVSVCNNCGLPGHIAAEC
TAESRCWNCREPGHVASNCSNEGICHSCGKSGHRARDCSNSDSRAGDLRLCNNCFKQGHLAADCTNDKACKNCRTSGHIARDCRNDPVC
NICSISGHVARHCPKGDSNYSDRGSRVRDGGMQRGGLSRMSRDREGVSAMIICHNCGGRGHRAYECPSARVADRGFRRY

AT2G06904 protein of unknown function; 1 $CX_2CX_3GHX_4C$

MPLKRSRGRESVEEEGRSLDLPSRRYDLRERVRGIAAMPERIIREDVQIRDGMMIDSDCEDDFVITGVRTVEQKLASEGHEIVMVHDYV
SVSSKSSASPSFALSLYSPISYESDPKEDQDPPISPYVPEAVDHSPDSPFALLWEGTSPILRFEADMRADAIPTFDGGFTSGIKILFCF
ICGEDGHYPQDCQFYTQYIPWPLLVSTVESMSIMRLPVC

AT2G12880 putative CCHC-type zinc finger protein; 2 $CX_2CX_3GHX_4C$

MAPKFTLNQERDNDRERTRASYNNDRRRNDYDPRACYKCGKLGHFARSCHVVTQPTTAYITCYFCSEEGHRSNGCPNKRTDQVNPKGHC
YWCGNQDHRFNLIIWRSRCLFRLMKPKYAF

AT2G15180 putative CCHC-type zinc finger protein; 1 $CX_2CX_3GHX_4C$

METIIYSSSDEEDSYGCSQDSYYNETRFVESWHDDGYSSSSDFEEEPDGGAPEPEPPDRYSSHATPPSYSKPWISWNPIPKTNFYPTFY
PTRKVIYKLFVHNGSKTYGPEFFLFSGLGYLQWESNMNYYFEFHSTAQEDKLSIALGQLKGSALWWWDQDEYNRWYERRAPIRTWERLK
WNMCAKYSPQSLSPAHVVQQKQKPTFLPQMQATTQGKCTFQTKHVELTCYRCKQEGHIAKICPTRETTTKVGLEQQELLKAKEKQEIVS
SPKGKNEQAETCQTDLNNSMNVITHLSSAKSIAKVSGTKENITDQGEASTMEKVFTEESKNQGGPTLDEITVNKDESVNDTIQIKGEPS
DAQQFPKTQCNLLNPYTLQWTNLTYLCVGKKVLRTKPLEEGGDDVRMGADVPADALVDRLLAGFFDGLSDGLPEVLHVSNQLVRTKESY
PFLPFCASQSHIWKLGDPLRHPEPFTLNP

AT3G02820 putative SWI3-like component of replication fork protection complex; 1 $CX_2CX_3GHX_4C$

MESAPTGCFKCGRPGHWSRDCPSSAPVAGNNSVSSSSAPSQIPNNEFQRSSSKSGTSIAPAPKVTKTRVQRPKLTPELLLSEDGLGYVL
RYFPKSFKYRGRGKEVSDLGNLIRLYSEWHTHLLPYYSFDHFVHKVQQVASTKRVKNCINELRERVASGVDPNKLYEKQEENTVPSDDQ
DMDQPSHDEENIPSKSVDADTNADAFEDSMLNEIFDNASKLPSDEQNMDKSSELTEEQRARMEANRLKAMEKAQNISEEQRVRMEANRL
KALERAKARLQPNQD

AT3G42860 putative CCHC-type zinc finger protein; 4 $CX_2CX_3GHX_4C$

MKKITIPVESLDEEDDFLLQLAAIEAEAAAKRPRVSSIPEGPYMAALKGSKSDQWQQSPLNPASKSRSVAVTTGGFQRSDGGGGVAGEQ
DFPEKSCPCGVGICLILTSNTPKNPGRKFYKCPNREENGGCGFFQWCDAVQSSGTSTTTSNSYGNGNDTKFPDHQCPCGAGLCRVLTAK
TGENVGRQFYRCPVFEGSCGFFKWCNDNVVSSPTSYSVTKNSNFGDSDTRGYQNAKTGTPCYKCGKEGHWARDCTVQSDTGPVKSTSAA
GDCFKCGKPGHWSRDCTAQSGNPKYEPGQMKSSSSSGECYKCGKQGHWSRDCTGQSSNQQFQSGQAKSTSSTGDCYKCGKAGHWSRDCT
SPAQTTNTPGKRQRQY

AT3G43490 putative CCHC-type zinc finger protein; 2 $CX_2CX_4HX_4C$

MPPKRKDINFLFDDGEDTVAIIDDEANEDLRLKILEKAFSRRNVDNKLDSDLSFDPGVVSTVMVNGGKSEEVKNSKSNKKMKRNKLEAA
NEIVTHCVERQDEDNMVEDVVRGEEEDGETTSNSVMTKLLRGARYFDPLDAGWVTCYSCGEKDHITVSCPTLTNCRKSCFICASLEHGA
RQCTKVWDIDANLGIHQDKTQRFKGKLCGSGDDDEVTDLMLNPQHRGLENMIQGLFLARIRDYTKPKWFDSTNYLELARTMI

AT3G43590 putative CCHC-type zinc finger protein; 6 $CX_2CX_3GHX_4C$ and 1 $CX_2CX_4HX_4C$

MPRQNKDEKFVFDDGEDEDREDPVAITEVDNGEEEDDDEANEDLSLKILEKALSRRDVGNKLDSDLSSDSGVVSTVMVNGVKSKVKKSE
SSKKMKRNKLEADHEIPIVWNDQDEEKVVEEIVKGEGEDDEVERSDEPKTEETASNLVLKKLLRGARYFDPPDAGWVSCYSCGEQGHTS
FNCPTPTKRRKPCFICGSLEHGAKQCSKGHDCYICKKTGHRAKDCPDKYKNGSKGAVCLRCGDFGHDMILCKYEYSKEDLKDVQCYICK
SFGHLCCVEPGNSLSWAVSCYRCGQLGHSGLACGRHYEESNENDSATPERLFNSREASECYRCGEEGHFARECPNSSSISTSHGRESQT
LCYRCNGSGHFARECPNSSQVSKRDRETSTTSHKSRKKNKENSEHDSTPHESNGKTKKKKKKKTHKEEQPQTSPRKRKHRGGWITEEPE
EESFQRGKMRRPKSPITPSGYNRSPSTHIGHNYRSPKFNSGGHYPGSQSSRHHSGPSPSRWQPSHQHHHHHQHLHHHHQNHSYEPAPPR
HGRANRYSEFAGNYERW

AT3G62330 putative CCHC-type zinc finger protein; 1 $CX_2CX_3GHX_4C$

MESTRSDPELDDDFSEIYKEYTGPASAVTNNNIQDKDKPVKQRSEERCDEEEEQLPDPNSVPTDFTSREAKVWEAKSKATERNWKKRKE
EEMICKICGESGHFTQGCPSTLGANRKSQEFFERVPARDNNVRVLFTEKVMESIERETSCKIKLDEKFIIVSGKDRLILRKGVDAVHKV
KEDGEMKSSSVSHRSRSRSPRRTSVGPSRARNSEPQRQQLPSHGSSSFPERSGRQDKFVDNRFREETRVRENQRNVPRGSPQAYGSDRA
RSRSTHSKSPGRPRYSGWDKPYDRQKPEVSGYRSERWDQERMGGSSDIQVSHQFERPPFPQTLEELELEYTRDALELEKKRDKEEDEEN
NKHRETIRELRESYMKKLAGLRGMNAKQWDDFLQLDAQRRQQQARQQNSGLSYGNYRQFPPYAEFDDGYSSNPPPYGGNNMPMDSKGRY
PNHGDNYSSRHQDNNYGGFQRQRREEYGKAYNRY

AT4G00980 protein of unknown function; 1 $CX_2CX_3GHX_4C$

MENSLLDSGETMEIVATQKIEETVKSILSESDMDQMTEFKLRLDASAKLGIDLSGTNHKKLVRDVLEVFLLSTPGEALVPETVAPAKNE
TVSVAAASVGGEDERFICKLSEKQNATVQRYRGQPFLSIGSQEHGKAFRGAHLSTNQWSVIKKNFAAIEDGIKQCQSKLKSEAARNGDT
SEAVDKDSSHGFSVIKISRFDGKSYLYWASQMELFLKQLKLTYVLSEPCPSIGSSQGPETNPREITRADATGKKWLRDDYLCYTHLMNS
LSDHLYRRYSQKFKHAKELWDELKWVYQCDESKSKRSQVRKYIEFRMVEERPILEQVQVFNKIADSIVSAGMFLDEAFHVSTIISKFPP
SWRGFCTRLMEEEYLPVWMLMERVKAEEELLRNGAKGVTYRPATGSSQMERTPSLGTTHRGSQSVGWKRKEPERDERVIIVCDNCGRKG
HLAKHCWGSKSDERASGKSNRINSSVAAPVESETQATTNNDRG

AT4G05360 putative Rad21/Rec8-like protein; 1 $CX_2CX_4HX_4C$ and 1 $CX_2CX_3GHX_4C$

MGKIKDQKGKEHDHSQCRRYVLQSYLLRTEHISPNSTRNINMITQDDSLCLLSISIYTLNSSHASLWLFTRSEEESADFGESEDPSKTV
RKRKKLSSSKLGFRRLDNQSKKDQLFNEPLFTGFSNVLLSVFKKDCVAKSYFAAPKEPASVPVSSPTREAETEINSASPVTQSTVPDST
NQESTVQRSSGQQTEHFQGVALKFVLPGLNILEYMLSPPPRSSPFRTNGFTIHPETWETGSYRTQPSTSNNTEELHFLEEGVNTPVRSP
VTQDSGGFSGRTRALAQHLKERYSGYLSLNKILEGKTRKIAARMFYETLGEVPEIVQKIIQEVRQSLKSRLMEILRLPFKKRFSWNERN
RALHEFKVDGNKEVLYKSSKGKRCFECKGFRHMCSECANLMKEKEKKFIMSDSEIDSDDGEELKNLVAFTTFESSIASASASGPTSASA
TGSTSASATGPATGSDNDQSDDDDLSISDEEFAENYKALYEHCVKVVEENSVLTKEKLKLEAKVVKTLKFAAEKEEEASQLEETQKNLR
MLNNGTKKLGHILSIGKTDKCGLGFKGNPSKSDPVFVYGGKITSASGTVKETATVAEIASDTRTDSRTDTETTSGTRKVQKLQSEPRRV
FRPVCHHCGVVGHIRPRCFRLLREKNRLMNAYDVRFHGPKCYHYGVQGHIKRNCFRFIRECSHEGLRRNKVWVRKDDFHGSGGENDDVF
GDVVLELELMSSQHSQKGGDLKM

AT5G20220 putative CCHC-type zinc finger protein; 4 $CX_2CX_3GHX_4C$

MRKKKTRFQFLLNPNSKFVFFPRAISSSSSSSSDNNNDGSVSSSRQNNRQMGYDPSEELFGVDFKPRFISGDSREPRSWFGPNGQYIRE
LPCPTCRGRGYTSCSNCGIERSRLDCPQCKGKGIMTCLRCLGDCVIWEESIDERPWEKARSSSPFRVKEDDEVDNLEIKFSKRRKSKRI
YQSPTPEVGQKISRSLKSLNAKTGLFSKRMKIIHRDPVLHAQRVAAIKKAKGTPAARKHASESMKAFFSNPVNREQRSLSMKGTKFYCK
NCGQEGHRRHYCPELGTNADRKFRCRGCGGKGHNRRTCPKSKSIVTKGISTRYHKCGICGERGHNSRTCRKPTGVNPSCSGENSGEDGV
GKITYACGFCKKMGHNVRTCPSKQVSDSDSCLEQEGS

AT4G06479 protein of unknown function; 1 $CX_2CX_4HX_4C$ and 1 $CX_2CX_3GHX_4C$

MAKDSESTESDISIESDPSEYLESTEDANNHDGPLSSEEANHVEWMITEDGELFAVPTENAIEQQTNDSLNNMLSQFYVPSTDKEGSFG
GPNLGTIGGHGLELEEADGAFGAMNEYGADDEYDMLMLMGNNVEDENGAIENEIENELEHVSLSVINISLDSSHAILMIPNSSKADSCW
EDEMFKGDEPSEDEENGRWDKMLLEQELNHESLMNHLRLDEERSKGKLVDKAGENEKWKGKRAVEMNMESEKAKDQRIVGVNLHPVGIG
YGINSKVKPGIQATARKSVGKKLATPAKRPCEICSHTDHPTEECLYPPQTIPYTDDYAKCYYCEGMGHMSMYCLYIAPNAGEGSLRGVG
PLMTTTTEEKCLNE

AT4G06526 protein of unknown function; 1 $CX_2CX_4HX_4C$ and 1 $CX_2CX_3GHX_4C$

MSSTKGSNIVSPLHGLNVPNLENFVLPSSIDDENVDNHDGMLSSEEANHVEWMITEDGELYAVPTENATEEQACDSLNNILSRFYVPST
DEEGSFSVPNMGTIGEHGLELEEVNGDMFDVPLIQANGAFGAMNKYGDDDEYDMLMLMGNDEEDESGVIENEIENELEQAIDKLIAYKA
DSCWEGEMFKGNEPSEDEGNGRLDEEKSKGKLVDKAEENERWKGKRAAEMGMESEKAKDQRIVGVNPLPLGMEYGINTKVKPRIQATAR
KSVGKKLATPAKRPCEICSHTDHPTKECLYPPHTMPYMDDCAKCSCCGGVGHMSMYCPYVAPNASEGSSRGVRPLMTTVAEEKCLNEGM
PSFLWVRRFKTFVKMMGEFIEVGNDQSQLTAYAQTLPKLQRAKMSMNGHEWAKMGKVGRAAAKFALPAAELTWVGHLTRGALRLNRSLW
GGGVTPVATCRQALLSGWLYEKFHLVEVGPETPVACGHDICSDMWQQAGIVDCGLGDSPET

AT1G48095 protein of unknown function; 1 $CX_2CX_4HX_4C$

MSDFENIVLPSSCDSQVTAWPVLELDFNQDTSTYVAFIRVKIRLVFTDRLRFFRRVRFESREGAMIGFEYEKLRRICTNCCRITHHVNN
CPFLNAPVIHDDEVEVLDVPVWEEGAASNTITDQTHMTSSDSSDISSASLISQPLPPASLASHEPHLVAGSEASRLVNPRPIPQNHFSS
SSSDFKGKGKAKMEIGECSKRKKDKQVVDDALRNVRQCRKDQGIRIMILQINLDIELASFVRTAGAVASHKMMKLLTPFSTVVLRRSKY
SILDDLEGYITSGVYTYKVTISTKLKICFSLYFHLF

AT5G13920 protein of unknown function; contains GRF-type zinc-binding domain; 1 $CX_2CX_3GHX_4C$

MLNGTMQTGDCFRCRQAGHWINDCPLKSYTDDPPPAIQCPCGGGFCEIKVANTRENPGRKFYKCPTAQNCCFFKWCDKVTDEDIKFRPA
FTIPICSCGAGPCRRVKDVSGRAYLICCIKKGFGACGFFKWEDVEMIPSCDVMDEIDFWVEADQILSDVESSLQARGGVIPEIANQMAS
EKECQASVSGIEDDCVVECQDSVSAEDDSTLENLDSISMSVSDVHSTALNQGISLFDPVTSEPEEPWKKTQNGDQPTNSALSKLSVDEA
MSDLIRDTVSSGSVVIHGRTNHEQPEIDGAEWSFPCLQNLIDQYNSEKLQLESISGKHVQMLSEFMASYRRLRLLHEKTSHLRKTLLET
EKEMVCCEAETLKFGASCREVAGEMAESQKRMQETADKLGKEVEVFKQNEFVGLKRRRT

AT5G34870 protein of unknown function; 1 $CX_2CX_3GHX_4C$

MSSSKGSNTVSLLHSVNVPNLENSAPSSSTDDESLFVAMNDQFMAEDSDSTESDTSIESDPSDVWKLFAVPTENATEEQVNDSLNNMLT
QFYVPSTDEEGSFGVPNLGTIGEHGLKLEELNADLFDVPVIHANGAFGAMNEYGANDEYEMLMLMGNNKEDENEAIENELEHAIDQLIA
ECAQDFGLIGEKGGGNGNDGGDSSSDSFESLTLEPKSPSVINISSDSSQPIVMMLNPSKPESCWEIEMFKGDEPSEEEGRGRWDPMLLE
QELNHEYLMNRLKLDEERSKEKMAQRIVGVNLHPLRIGYGINSKVKPRIRATARKSVRRKLPTPAKRPWEICGYIDHPTEQCLHPPQAM
PYMVDCAKCYSCGGVGHVSMYCPYLAPNAGEGSSRDALVFMGKGETSHERRHEGDSKTFVKMSEEFVGDRMRKSCLNAWPIDRMCPEST
KPAKGKDGQEWARMGKDGKSWACRGRNFPCRVQNRHVWTPHLRCLAAQVTNLGAKVGLVLEPKLAESWSRS

AT5G36240 putative CCHC-type zinc finger protein; 4 $CX_2CX_3GX_4C$

MIKLPNPISRVCDPLIMPRHNDDEAEVCLRCGGFGHDMTLCKYEYSHEDLKNIKCYVCNSLGHLCCIEPGHTQSWTVSCYRCGQLGHTG
LACGRHYDDSVSPSCFICGREGHFEHQCHNSFSVCFPEDSSEDECQGPDSSSVRFQENTREEEEGHFEHQCPDSSSVCFQEISREEGFI
SLNSSSKSTSKGRETRRLCYECKGKGHIARDCPNSSQVVSISLSTFCALSSSVLSFLSCYYSRINFLCFFQDKYGI

AT5G49400 zinc knuckle (CCHC-type) family protein; 1 $CX_2CX_3GHX_4C$

MSNSKDEKSQDAADRIKAATLTAAKGLSRTQAERAAAAAARNVNAYGQKEEGPSRWQEKRELKRQMYLMSTEKAVRLGERKDKTMSASA
VGSSASAASQCQKCFQAGHWTYECKNERVYISRPSRTQQLKNPKLRTKPSVDDLDGSDDDDEEERPDATNGKAEVEKRSKKSKRKHRSK
SDSESDSEASVFETDSDGSSGESSSEYSSSSDSEDERRRRRKAKKSKKKQKQRKERRRRYSSSSSESSESESASDSDSDEDRSRRKKKS
KRHSNKRR

AT5G52380 CCHC-type zinc finger protein; 3 $CX_2CX_3GHX_4C$ and 2 $CX_2CX_4HX_4C$

MVNQRRRLAQKRYKEANPELFPKAEPTPPKDPNKKKKKKSLFKKKKPGSSTDRPQRTGSSTRHPLRVPGMKPGEGCFICHSKTHIAKLC
PEKSEWERNKICLQCRRRGHSLKNCPEKNNESSEKKLCYNCGDTGHSLSHCPYPMEDGGTKFASCFICKGQGHISKNCPENKHGIYPMG
GCCKVCGSVAHLVKDCPDKFNQESAQPKKTSRFDATPRGKLTKFSGDDLEDDFTEEPKSSKKINTSDDSAQNSVEVKKKKQGPKIVNFV
G

AT2G01050 protein of unknown function; 1 $CX_2CX_3GHX_4C$

MLDVGEKGRPPGDPPDKLESWATKVKGSAGGGILKPEDVIDDEFVRERVGLEFPDGEDEEPVITIGEEVLEAMNGLWKKCMIVKVLGSQ
IPISVLNRKLRELWKPSGVMTVMDLPRQFFMIRFELEEEYMAALTGGPWRVLGNYLLVQDWSSRFDPLRDDIVTTPVWVRLSNIPYNYY
HRCLLMEIARGLGRPLKVDMNTINFDKGRFARVCIEVNLAKPLKGTVLINGDRYFVAYEGLSKICSSCGIYGHLVHSCPRNVVVKVSAG
AETVTDRAVVPVGMEGDDGFTVVQRTARRPAAPVQKMVFAVGASGGRSKQRLRELPKNQGVDLANRFGGLDGNGDLPDLREVAITEGPN
KENEYHGRNVGKVMGVPLVKEARGSTQMEKGKGGSKGGFKWKRNGGMKALEPIGPKQKHGAANKPARGLIFGPTKDANSVPVGEDLLSN
GKRLRVEQRDVGRPGGVYSSAMGSHAHEASFDLDSSSTLSQRFQREDLMSEIAVVSHEGSEVGNSSEGMA

AT5G36228 protein of unknown function; 1 $CX_2CX_4HX_4C$

MSDELWNAIQHMDLGREEPELYIPYHAYVGALASNRLSLLGRILNPQTQSVERAILELPYQWGLGTQVHGRILDDRCFQVRFRSEIDLL
NGLRRAPWVFNEWFIALQRWEDFPTEDFLTFIDVWVHIRGIPLPYVSERTVEIIASTLGEVVAMDFNEETTSQITFIRVKVRMDFTEPL
RFFRRVRFASRERAMIGFEYEKLQRVCTNCCRVNHQVSHCPYVVHQEEMDNEPDVLVSPERYDDEDSLNQEDHGRHSQSSVISSFSSLT
PISLNAPPVVNWNDNMIGNIPHRFPSTSVSSSHTVSDGYLAASEWRPKDQVSYEVGESSKRKKGKQVLEVPERSIRQRRMGSGIRFYPV
NGENP

AT2G07760 protein of unknown function; 1 $CX_2CX_3GHX_4C$

MNSVSVLTQGKETEVGPKVVVVDSSNVNSVPDPRWPYLTRWTQNSPSPSPSVISPSISLVKPLTIGSSPKIARSSSLDFVNLTSSLDST
LPVTVEPPIEHVVSITSLSYSAEVSLPTSALPPIVSATLDESIAPVKLVSPTLKGAWAKQLKFTSSSASQQDVGGAPMQSHLEATKEDN
VRFPWAAKMDPAARNLYRATSPEFMEDGIPKRGLWHVDDCLMFVAPWSTVNTFDLPEISTIPVWVTLKNIPNRLYSILGISHIASGLGA
PMATYKPRLDPSLMSEANILVEVELSKAFPPRIAAVDKKGNISMVNVEYAWIPAKCGKCGQLGHKASRCMKPHLAHEKVTEIVSEEIIT
PAIVSLASATNLVSPITLQTKTPIDVPITNSKIQIDTVFDIEAGPIQDKDNCTGVADCFTNVEAEVVTEETCTVESVFEIAGTKDKFSR
LGSSFPDGDSLHSDEDSIVSVESDSGSELMELMTPSGQRLLRERPVKPSIKAKEIQASSTTRGRGNRGRGNRGRGNRGRGNRGRG

AT3G24070 protein of unknown function; 1 $CX_2CX_4HX_4C$

MWLKIRGIPIQYLCDGTVREIASSMGEVMEVELDDGMVDLSSVRARVNVCVDTRLCFKKVARFDSGEVKIVSFRYEDIGMSKARFKFCF
NCGDMNHLARNCLIPWVDVPDPYERSLSPPPHESNSDGSAEGNCGDGGTSSGTLELLELVDAVQDFPVGDGEQLELAGVVQDLPVGDGE
QQQQGLDGVEEANATQVDSEMVGISSEGSKRKFDAVEQGDENPEKRLRGITDATEGGSTDATQGGSTDATEEGLGVSLKPLQGE

AT3G31430 protein of unknown function; 1 $CX_2CX_4HX_4C$

MATDKAICRRLLYYLFVTPLVTVYREFAGKVVCFFTAHSLVVISTVESTIHRDATSSASGNVLLKFRFTIPESTNLISLKPPNHFPLRF
YQPLSQMADNLRRAVQDINLGVDDIPFALPEDIVNHAVAENRFILFGRPVMPRRQNLRSIVASMPRIWGQSGLVHGRIMEGRQFHFIFT
LEESLETVLRRGPWAFNDWMILLQRWEPQIPLFPFIPFWVQIRGIPFQFLNRGVVEHIGRALGQVLDTDFNVEVVARMDFARVLLHWDI
THPLRFQRHFQFTAGVNTLLRFRYERLRGFCEVCGMLTHDFGACLIQNGGEEQADDDDDDEEHPQTYHN

AT3G42140 protein of unknown function; 1 $CX_2CX_4HX_4C$

MADNLQRQMQEIALGVQEEAINLLIELCDEALRKPNLASLQNRCPKNVDEEVVGRILEIHKIEFLFQSEESMFSILRRGPWSFNDWMCV
IQRWTKLHSDAEFKRIPFWIQIRGIPLRFLTARIITSIGERMGLFLETNLGRDVSVLKFQYEKLKNFCTTCGMLSHDASECPTSGNQGP
HADDDDDDDDNDANEDHPDVPGDNPDQVHGPMNQQHPEESTEDEHQNAPKRGKLKPLLQRLTMEVPWCVATCARPMPQKILNTAWLKDR
EGNHKF

AT5G32613 protein of unknown function; 1 $CX_2CX_3GHX_4C$

MPKKKKSVRPSWPPIPSKFFRVLAVSSAPGSSVSSVPDSVLSCVGADASSTLVEMNRVRPSSPKVLGSMLELCQDIVPLTSFDLVEVLD
KSVKLPEARISSEIAVPVSENSSIGTVELEQGCEITSATAPELEISSSVPVDTLINSSFDRITNPSPTSSKGPLWMTPSTFLEDGTPMV
VAPASVLLKTAEMSLWEHYGSSVLEDGGVDFHSFSKYSPMGESPLEPQELQTAQTWAILKNVPPQLYSLEGISVIASGIGEPLHTEKSR
LGPVNIGRTKVKVVTNLGTPLPDSIVVRDVQGNTARVAVTYPRPPPKCLNCGRYGHLLSRCSKPLMKKLPFKKDLPSGSKEVQILVISL
PTSQEAQRGIMLESSIEDQKTTTTQAKSKCRRSRSKKRSASLPSASIGPLEIQKGVKEGKSDRLAAVAKPKWIVKADVKRPGTASQPTL
SSPTIIDASCEL

AT2G17920 Nucleic acid binding/zinc ion binding protein; 1 $CX_2CX_4HX_4C$

MADELWDEIQNLELRQEGPSLFIPNEAYIMVAGRNRLSIIARPLNPRVQNLQAIITALPRAWGLTAHVHGRIIDDTYVQFLFQSEMDLL
SVQRREPWLFNNWFVASQRWQPAPALNFVTTIDLWVQMRGIPFLYVSEETALEIAQEIGAIISLDFHDTTSTQIAYIRVRVRVGITDSL
RFFQRITFESGESALIRFQYERLRRICSNCFRFTHNRNYCPYRPRVQILDRERAALHDSVLRSSNNSQSQMTESSFPAPLTPPPRVPAP
PLNHQELAAATPYFPSKNCSFPALHSSYYSKWYWQTSSFP

AT3G21000 Gag-Pol-related retrotransposon family protein; 1 $CX_2CX_4HX_4C$

MATANIRDGVVSDQFDYEIWAPITKSTLIEQGLWDVVVNGVPQDPSKNPELAATIQPEELSKWRDFVVKDAKALQILQSSLTDSVFRKT
LSASSAKDVWDLLRKGNEQATIRRLEQVTIRRLEKQLEDLKMVDKESGSSYLDKALEILERLGRAKLEKSDYEICKNVFTTLSGSFDGL
DSMLEELIDVHKMTSKSLVEYFYYRVHESSTEEAIFGLLKDLRLKSKSEKW**CGLCYKNNHNQEDC**KFRIHTDKEEKEDEIVVDYRLETV
PNLGAKTYDDDIWIIHKMAPINMTPYVKYFTTLDRTFKATVGTVDGTVLLVEGKGDVKIRMKEGKKKTIRNVIFVPGLNRNVLSFGKMV
SKRYSISTGMQGECIVCDRGENKLGDAMWMTDETEMALRLKVIEGKLTY

AT5G18636 Ta11-like non-LTR retrotransposon; 1 $CX_2CX_4HX_4C$

MADELWDEIQNLELGQEDPALFIPHEAYVMVEASNNLSLIARPLNPRVQNLNSVVVALPRSWGLTTQVHGRVLDATYVQFVFANEIDLM
MVQRREPWLFNNWFVAATRWQVAPAHNLVTTIDLWVQIRGIPLPYVSEETVLEIAQDLGEIISLDFHEATSPQIAFIRVRVRFGITDRL
RFFQRIIFDSGETATIRFQYERLRRL**CSSCFRFTHNRAYC**PYRQRSLSIARERALFCDSVQRSSMNSQSQMTESSFPVPMTPPPRVDPP
PMNHSEFVAAYPHLATATNVNYRFTGDSSTSRQDLSSGSNNILPRRTTHFTDHRRCFEAGQSSRQNENREPRRPTERMLPPSHFDHVQR
SGGILKPPKKR

AT5G25200 Ta11-like non-LTR retrotransposon; 1 $CX_2CX_4HX_4C$

MADELWDEIQNLELGQEDPALFIPHEAYVMVEASNNLSLIARPLNPRVQNLNSVVVALPRSWGLTTQVHGRVLDATYVQFLFANEIDLM
MVQRREPWLFNNWFVAATRWQVAPAHNLVTTIDLWVQIRGIPLPYVSEETVLEIAQDLGEIISLDFHEATSPQIAFIRVRVRFGITDRL
RFFQRIIFDSGETATIRFQYERLRRL**CSSCFRFTHNRAYC**PYRQRPLSIARERALFCDSVQRSSMNSQSQMTESSFPVPMTPPPRVDPP
PMNHSEFVAAYPHLATATNVNYRFTGDSSTSRQDLSSGSNNILPRRTTHFTDHRRCFETGQSSRQNENREPRRPTERMLPPSHFDHVQR
AGGILKPPKKR

AT2G13450 Ta11-like non-LTR retrotransposon; 1 $CX_2CX_4HX_4C$

MADELWDELQHLELGREDPALFIPHEAYAIVESRNRLSLIARPLNPRSQNLHAVISALPRAWGLTNRVHGRVLNDTFVQFIFQSEIDLL
SVLRREPWLYNNWFVTAQRWEVNLTFHLLTSIELWVQMRGIPLLYVCEETALEIAHELGEIITLDFHDSTTTQIAYIRVRIRFGITDRL
RFFLRIIFDSGETALISFQYERLRRI**CSSCFRMTHHRNSC**LYRQIESLHRVTNTTAQRNVREEVFMRDENLRSSMNSQSQMSESSFPTP
IDPPPRIPHPPLNPDELVAAYYPHTRATSLPNFAGPLPQVPLRRNVDERDSNIQPFSGPAFAAHSPRLVEVGESSRQGENTQNVHTVEK
GDSSKRKNMGGPKFKDDARKSNEDEHMNGGILKPPKKR

AT2G16676 Ta11-like non-LTR retrotransposon; 1 $CX_2CX_4HX_4C$

MRGIPLLYVCEATVTEIALGLGQIISLDFHDATTTQIAFIRVRIRFEITDRIRFFQRITFDSGETALIRFQYERLRRI**CSSCFRLTHHR
NYC**PYRQPEPRSIIRGPTNNLTRSRREGVCTRDEYHRSSLNSQSQMSENAFPAPIEPPPRVAAPPLNPDEFRAAYFPEGRAGSLPNIGT
LNLNPPSRRQEASRNSNVQPFTGPAFGANVPRVVEVGECS

AT2G41590 Ta11-like non-LTR retrotransposon; 1 $CX_2CX_4HX_4C$

MSDELWNEIQNLELGQEDPALFIPHEAYVMVEATNRLSMIARPLNPRVQNLNSVVVALPRTWGLTNQVHGRILDATYVQFLFQNEIDLM
MVQRKEPWLFNNWFVAATRWEVAPAHNFVTTIDLWVQIRGIPLPYVSEETVMEIAQDLGEVLMLDYHDTTSIQIAYIRVRVRFGITDRL
RFFQRIVFDSGETATIRFQYERLRRI**CSSCFRFTHNRAYC**PYRPRPLSIARERALFRDSVHRSSMNSQSQMTDSSFPIPQTPPPRISHP
PLNHDEFVAAYPHLDSGRNDHIRCEGESSNFRQDLSSASNSITPRREPQYLTDRRHFEPGQSSRRHDVRDLRRGSERIGNLNQQNYVQR
SGGILKPPKKR

AT3G47920 Ta11-like non-LTR retrotransposon;1 $CX_2CX_4HX_4C$

MLDATYVQFLFQNEVDLLSVQRRELWLFNNWFVANHRWEPAPVLNFVTTIDLWVQMRGIPLLYVCEETALEIAHEIGEIITLDFHDATM
TQIAYIRVRVRIGITDRLRFFQRITFDSGETALIRFQYERLRRI**CSSCFRVTHHRNYC**PYRPRLPNYGRERAVFHDERLRSSMNSQSQM
TESSFPAPVLPPPRIVTPPLNHGEFLAAHPNFPAPREGLNHQGRGTYYQGLCQGQQVSTDSNITPSVGTALSTGSRRVFEVGQSSRGVE
TRETRKRQEEKGTHDEQDKAHMKGGILNPPKKR

AT4G02000 Ta11-like non-LTR retrotransposon; 1 $CX_2CX_4HX_4C$

MTSLTIDLLSVLRREPWLYNNWFVTTHRWEVNLTFHLLTSIELWVQMRGIPLLYVCEETALEIAHELGKILTLDFHDSTTTQIAYIRVR
IRFGITDRLRFFQRIIFDFGEAALISFQYERLRRI**CSSCFRMTHHRNSC**PYRQIEPLHRVTNSTAQRNVREEVFMRDENLRSSMNSQSQ
MSESSFPTPIDPPPRIPHPPLNPDELVAAYYPHTRATSLPNFAGPLPQVPLRKNVDERDSNIQPFSGPAFAAHSPRLVEVGESSRQGEN
TQNVHTVEKGDSSKRKNMGGPRFKDDARKSNEDEHMNGGILKPPKKR

AT2G02103 Ta11-like non-LTR retrotransposon; 1 $CX_2CX_4HX_4C$

MADELWDEIQNLELGQEDPALFIPHEAYVMVEASNNLSLIARPLNPRVQNLNSVVVALPRSWGLTTQVHGRILDATYVQFLFANEIDLL
MVQRREPWLFNNWFVAATRWQVAPAHNLVTTIDLWVQIRGIPLPYVSEETVMEIAHDLGEIISLDFHEATSPQIAFIRVRVRFGITDRL
RFFQRVIFDSGETATIRFQYERLRRL**CSSCFRFTHNRAYC**PYRQRPLSIARERALFCDSVQRSSMNSQSQMTESSFPVPLTPPPRVDPP
PVNHAEFTAAYPHLATATNENHRDVGESSTSRQDMSSGSNNFLARRTTHFTDHRRCFEGGQASRQNENREPRRPTERMRSPTHFDHVQR
SGGILKPPKKTLNHSIVFCHFLGKIAGLEASQLQDEEACLVGYGFWL

Supplementary Figure 2. Arabidopsis ZCCHC domain-containing proteins. The number of $CX_2CX_4HX_4C$ or $CX_2CX_3GHX_4C$ sequences in each protein is shown; these sequences have been highlighted. The Arabidopsis Genome Initiative (AGI) gene identifier (ATNGNNNNN) is shown, together with the TAIR10 annotation or protein name. A slash separates alternative names given by different authors to the same protein.

AIR1/ZCCHC7; 4 $CX_2CX_3GHX_4C$

MMFGGYETIEAYEDDLYRDESSSELSVDSEVEFQLYSQIHYAQDLDDVIREEEHEEKNSGNSESSSSKPNQKKLIVLSDSEVIQLSDGS
EVITLSDEDSIYRCKGKNVRVQAQENAHGLSSSLQSNELVDKKCKSDIEKPKSEERSGVIREVMIIEVSSSEEEESTISEGDNVESWML
LGCEVDDKDDDILLNLVGCENSVTEGEDGINWSISDKDIEAQIANNRTPGRWTQRYYSANKNIICRNCDKRGHLSKNCPLPRKVRRCFL
CSRRGHLLYSCPAPLCEYCPVPKMLDHSCLFRHSWDKQCDRCHMLGHYTDACTEIWRQYHLTTKPGPPKKPKTPSRPSALAYCYHCAQK
GHYGHECPEREVYDPSPVSPFICYYDDKYEIQEREKRLKQKIKVLKKNGVIPEPSKLPYIKAANENPHHDIRKGRASWKSNRWPQENKE
TQKEMKNKNRNWEKHRKADRHREVDEDFPRGPKTYSSPGSFKTQKPSKPFHRSSHYHTSREDKSPKEGKRGKQKKKERCWEDDDNDNLF
LIKQRKKKS

AIR2/ZCCHC8; 1 $CX_2CX_4HX_4C$

MAAEVYFGDLELFEPFDHPEESIPKPVHTRFKDDDGDEEDENGVGDAELRERLRQCEETIEQLRAENQELKRKLNILTRPSGILVNDTK
LDGPILQILFMNNAISKQYHQEIEEFVSNLVKRFEEQQKNDVEKTSFNLLPQPSSIVLEEDHKVEESCAIKNNKEAFSVVGSVLYFTNF
CLDKLGQPLLNENPQLSEGWEIPKYHQVFSHIVSLEGQEIQVKAKRPKPHCFNCGSEEHQMKDCPMPRNAARISEKRKEYMDACGEANN
QNFQQRYHAEEVEERFGRFKPGVISEELQDALGVTDKSLPPFIYRMRQLGYPPGWLKEAELENSGLALYDGKDGTDGETEVGEIQQNKS
VTYDLSKLVNYPGFNISTPRGIPDEWRIFGSIPMQACQQKDVFANYLTSNFQAPGVKSGNKRSSSHSSPGSPKKQKNESNSAGSPADME
LDSDMEVPHGSQSSESFQFQPPLPPDTPPLPRGTPPPVFTPPLPKGTPPLTPSDSPQTRTASGAVDEDALTLEELEEQQRRIWAALEQA
ESVNSDSDVPVDTPLTGNSVASSPCPNELDLPVPEGKTSEKQTLDEPEVPEIFTKKSEAGHASSPDSEVTSLCQKEKAELAPVNTEGAL
LDNGSVVPNCDISNGGSQKLFPADTSPSTATKIHSPIPDMSKFATGITPFEFENMAESTGMYLRIRSLLKNSPRNQQKNKKASE

CLIP1/CLIP-170; 1 $CX_2CX_3GHX_4C$

MSMLKPSGLKAPTKILKPGSTALKTPTAVVAPVEKTISSEKASSTPSSETQEEFVDDFRVGERVWVNGNKPGFIQFLGETQFAPGQWAG
IVLDEPIGKNDGSVAGVRYFQCEPLKGIFTRPSKLTRKVQAEDEANGLQTTPASRATSPLCTSTASMVSSSPSTPSNIPQKPSQPAAKE
PSATPPISNLTKTASESISNLSEAGSIKKGERELKIGDRVLVGGTKAGVVRFLGETDFAKGEWCGVELDEPLGKNDGAVAGTRYFQCQP
KYGLFAPVHKVTKIGFPSTTPAKAKANAVRRVMATTSASLKRSPSASSLSSMSSVASSVSSRPSRTGLLTETSSRYARKISGTTALQEA
LKEKQQHIEQLLAERDLERAEVAKATSHVGEIEQELALARDGHDQHVLELEAKMDQLRTMVEAADREKVELLNQLEEEKRKVEDLQFRV
EEESITKGDLEQKSQISEDPENTQTKLEHARIKELEQSLLFEKTKADKLQRELEDTRVATVSEKSRIMELEKDLALRVQEVAELRRRLE
SNKPAGDVDMSLSLLQEISSLQEKLEVTRTDHQREITSLKEHFGAREETHQKEIKALYTATEKLSKENESLKSKLEHANKENSDVIALW
KSKLETAIASHQQAMEELKVSFSKGLGTETAEFAELKTQIEKMRLDYQHEIENLQNQQDSERAAHAKEMEALRAKLMKVIKEKENSLEA
IRSKLDKAEDQHLVEMEDTLNKLQEAEIKVKELEVLQAKCNEQTKVIDNFTSQLKATEEKLLDLDALRKASSEGKSEMKKLRQQLEAAE
KQIKHLEIEKNAESSKASSITRELQGRELKLTNLQENLSEVSQVKETLEKELQILKEKFAEASEEAVSVQRSMQETVNKLHQKEEQFNM
LSSDLEKLRENLADMEAKFREKDEREEQLIKAKEKLENDIAEIMKMSGDNSSQLTKMNDELRLKERDVEELQLKLTKANENASFLQKSI
EDMTVKAEQSQQEAAKKHEEEKKELERKLSDLEKKMETSHNQCQELKARYERATSETKTKHEEILQNLQKTLLDTEDKLKGAREENSGL
LQELEELRKQADKAKAAQTAEDAMQIMEQMTKEKTETLASLEDTKQTNAKLQNELDTLKENNLKNVEELNKSKELLTVENQKMEEFRKE
IETLKQAAAQKSQQLSALQEENVKLAEELGRSRDEVTSHQKLEEERSVLNNQLLEMKKRESKFIKDADEEKASLQKSISITSALLTEKD
AELEKLRNEVTVLRGENASAKSLHSVVQTLESDKVKLELKVKNLELQLKENKRQLSSSSGNTDTQADEDERAQESQIDFLNSVIVDLQR
KNQDLKMKVEMMSEAALNGNGDDLNNYDSDDQEKQSKKKPRLFCDICDCFDLHDTEDCPTQAQMSEDPPHSTHHGSRGEERPYCEICEM
FGHWATNCNDDETF

CNBP/ZCCHC22/ZNF9; 7 $CX_2CX_3GHX_4C$

MSSNECFKCGRSGHWARECPTGGGRGRGMRSRGRGGFTSDRGFQFVSSSLPDICYRCGESGHLAKDCDLQEDACYNCGRGGHIAKDCKE
PKREREQCCYNCGKPGHLARDCDHADEQKCYSCGEFGHIQKDCTKVKCYRCGETGHVAINCSKTSEVNCYRCGESGHLARECTIEATA

CPSF4; 1 $CX_2CX_3GHX_4C$

MQEIIASVDHIKFDLEIAVEQQLGAQPLPFPGMDKSGAAVCEFFLKAACGKGGMCPFRHISGEKTVVCKHWLRGLCKKGDQCEFLHEYD
MTKMPECYFYSKFGECSNKECPFLHIDPESKIKDCPWYDRGFCKHGPLCRHRHTRRVICVNYLVGFCPEGPSCKFMHPRFELPMGTTEQ
PPLPQQTQPPAKQSNNPPLQRSSSLIQLTSQNSSPNQQRTPQVIGVMQSQNSSAGNRGPRPLEQVTCYKCGEKGHYANRCTKGHLAFLS
GQ

DDX41; 1 $CX_2CX_3GHX_4C$

MEESEPERKRARTDEVPAGGSRSEAEDEDDEDYVPYVPLRQRRQLLLQKLLQRRRKGAAEEEQQDSGSEPRGDEDDIPLGPQSNVSLLD
QHQHLKEKAEARKESAKEKQLKEEEKILESVAEGRALMSVKEMAKGITYDDPIKTSWTPPRYVLSMSEERHERVRKKYHILVEGDGIPP
PIKSFKEMKFPAAILRGLKKKGIHHPTPIQIQGIPTILSGRDMIGIAFTGSGKTLVFTLPVIMFCLEQEKRLPFSKREGPYGLIICPSR
ELARQTHGILEYYCRLLQEDSSPLLRCALCIGGMSVKEQMETIRHGVHMMVATPGRLMDLLQKKMVSLDICRYLALDEADRMIDMGFEG
DIRTIFSYFKGQRQTLLFSATMPKKIQNFAKSALVKPVTINVGRAGAASLDVIQEVEYVKEEAKMVYLLECLQKTPPPVLIFAEKKADV
DAIHEYLLLKGVEAVAIHGGKDQEERTKAIEAFREGKKDVLVATDVASKGLDFPAIQHVINYDMPEEIENYVHRIGRTGRSGNTGIATT
FINKACDESVLMDLKALLLEAKQKVPPVLQVLHCGDESMLDIGGERGCAFCGGLGHRITDCPKLEAMQTKQVSNIGRKDYLAHSSMDF

LIN28A/ZCCHC1/CSDD1; 2 $CX_2CX_4HX_4C$

MGSVSNQQFAGGCAKAAEEAPEEAPEDAARAADEPQLLHGAGICKWFNVRMGFGFLSMTARAGVALDPPVDVFVHQSKLHMEGFRSLKE
GEAVEFTFKKSAKGLESIRVTGPGGVFCIGSERRPKGKSMQKRRSKGDRCYNCGGLDHHAKECKLPPQPKKCHFCQSISHMVASCPLKA
QQGPSAQGKPTYFREEEEEIHSPTLLPEAQN

LIN28B/CSDD2; 2 $CX_2CX_4HX_4C$

MAEGGASKGGGEEPGKLPEPAEEESQVLRGTGHCKWFNVRMGFGFISMINREGSPLDIPVDVFVHQSKLFMEGFRSLKEGEPVEFTFKK
SSKGLESIRVTGPGGSPCLGSERRPKGKTLQKRKPKGDRCYNCGGLDHHAKECSLPPQPKKCHYCQSIMHMVANCPHKNVAQPPASSQG
RQEAESQPCTSTLPREVGGGHGCTSPPFPQEARAEISERSGRSPQEASSTKSSIAPEEQSKKGPSVQKRKKT

pNO40/ZCCHC17; 1 $CX_2CX_3GHX_4C$

MNSGRPETMENLPALYTIFQGEVAMVTDYGAFIKIPGCRKQGLVHRTHMSSCRVDKPSEIVDVGDKVWVKLIGREMKNDRIKVSLSMKV
VNQGTGKDLDPNNVIIEQEERRRRSFQDYTGQKITLEAVLNTTCKKCGCKGHFAKDCFMQPGGTKYSLIPDEEEEKEEAKSAEFEKPDP
TRNPSRKRKKEKKKKKHRDRKSSDSDSSDSESDTGKRARHTSKDSKAAKKKKKKKKHKKKHKE

RBBP6; 1 $CX_2CX_3GHX_4C$

MSCVHYKFSSKLNYDTVTFDGLHISLCDLKKQIMGREKLKAADCDLQITNAQTKEEYTDDNALIPKNSSVIVRRIPIGGVKSTSKTYVI
SRTEPAMATTKAIDDSSASISLAQLTKTANLAEANASEEDKIKAMMSQSGHEYDPINYMKKPLGPPPPSYTCFRCGKPGHYIKNCPTNG
DKNFESGPRIKKSTGIPRSFMMEVKDPNMKGAMLTNTGKYAIPTIDAEAYAIGKKEKPPFLPEEPSSSSEEDDPIPDELLCLICKDIMT
DAVVIPCCGNSYCDECIRTALLESDEHTCPTCHQNDVSPDALIANKFLRQAVNNFKNETGYTKRLRKQLPPPPPPIPPPRPLIQRNLQP
LMRSPISRQQDPLMIPVTSSSTHPAPSISSLTSNQSSLAPPVSGNPSSAPAPVPDITATVSISVHSEKSDGPFRDSDNKILPAAALASE
HSKGTSSIAITALMEEKGYQVPVLGTPSLLGQSLLHGQLIPTTGPVRINTARPGGGRPGWEHSNKLGYLVSPPQQIRRGERSCYRSINR
GRHHSERSQRTQGPSLPATPVFVPVPPPPLYPPPPHTLPLPPGVPPPQFSPQFPPGQPPPAGYSVPPPGFPPAPANLSTPWVSSGVQTA
HSNTIPTTQAPPLSREEFYREQRRLKEEEKKKSKLDEFTNDFAKELMEYKKIQKERRRSFSRSKSPYSGSSYSRSSYTYSKSRSGSTRS
RSYSRSFSRSHSRSYSRSPPYPRRGRGKSRNYRSRSRSHGYHRSRSRSPPYRRYHSRSRSPQAFRGQSPNKRNVPQGETEREYFNRYRE
VPPPYDMKAYYGRSVDFRDPFEKERYREWERKYREWYEKYYKGYAAGAQPRPSANRENFSPERFLPLNIRNSPFTRGRREDYVGGQSHR
SRNIGSNYPEKLSARDGHNQKDNTKSKEKESENAPGDGKGNKHKKHRKRRKGEESEGFLNPELLETSRKSREPTGVEENKTDSLFVLPS
RDDATPVRDEPMDAESITFKSVSEKDKRERDKPKAKGDKTKRKNDGSAVSKKENIVKPAKGPQEKVDGERERSPRSEPPIKKAKEETPK
TDNTKSSSSSQKDEKITGTPRKAHSKSAKEHQETKPVKEEKVKKDYSKDVKSEKLTTKEEKAKKPNEKNKPLDNKGEKRKRKTEEKGVD
KDFESSSMKISKLEVTEIVKPSPKRKMEPDTEKMDRTPEKDKISLSAPAKKIKLNRETGKKIGSTENISNTKEPSEKLESTSSKVKQEK
VKGKVRRKVTGTEGSSSTLVDYTSTSSTGGSPVRKSEEKTDTKRTVIKTMEEYNNDNTAPAEDVIIMIQVPQSKWDKDDFESEEEDVKS
TQPISSVGKPASVIKNVSTKPSNIVKYPEKESEPSEKIQKFTKDVSHEIIQHEVKSSKNSASSEKGKTKDRDYSVLEKENPEKRKNSTQ
PEKESNLDRLNEQGNFKSLSQSSKEARTSDKHDSTRASSNKDFTPNRDKKTDYDTREYSSSKRRDEKNELTRRKDSPSRNKDSASGQKN
KPREERDLPKKGTGDSKKSNSSPSRDRKPHDHKATYDTKRPNEETKSVDKNPCKDREKHVLEARNNKESSGNKLLYILNPPETQVEKEQ
ITGQIDKSTVKPKPQLSHSSRLSSDLTRETDEAAFEPDYNESDSESNVSVKEEESSGNISKDLKDKIVEKAKESLDTAAVVQVGISRNQ
SHSSPSVSPSRSHSPSGSQTRSHSSSASSAESQDSKKKKKKKEKKKHKKHKKHKKHKKHAGTEVELEKSQKHKHKKKKSKKNKDKEKEK
EKDDQKVKSVTV

RBM4B/LARK/ZCCHC15; 1 $CX_2CX_3GHX_4C$

MVKLFIGNLPREATEQEIRSLFEQYGKVLECDIIKNYGFVHIEDKTAAEDAIRNLHHYKLHGVNINVEASKNKSKASTKLHVGNISPTC
TNQELRAKFEEYGPVIECDIVKDYAFVHMERAEDAVEAIRGLDNTEFQGKRMHVQLSTSRLRTAPGMGDQSGCYRCGKEGHWSKECPVD
RTGRVADFTEQYNEQYGAVRTPYTMGYGESMYYNDAYGALDYYKRYRVRSYEAVAAAAAASAYNYAEQTMSHLPQVQSTTVTSHLNSTS
VDPYDRHLLPNSGAAATSAAMAAAAATTSSYYGRDRSPLRRAAAMLPTVGEGYGYGPESELSQASAATRNSLYDMARYEREQYVDRARY
SAF

RBM4A/LARK/ZCCHC21; 1 $CX_2CX_3GHX_4C$

MVKLFIGNLPREATEQEIRSLFEQYGKVLECDIIKNYGFVHIEDKTAAEDAIRNLHHYKLHGVNINVEASKNKSKTSTKLHVGNISPTC
TNKELRAKFEEYGPVIECDIVKDYAFVHMERAEDAVEAIRGLDNTEFQGKRMHVQLSTSRLRTAPGMGDQSGCYRCGKEGHWSKECPID
RSGRVADLTEQYNEQYGAVRTPYTMSYGDSLYYNNAYGALDAYYKRCRAARSYEAVAAAAAASVYNYAEQTLSQLPQVQNTAMASHLTST
SLDPYDRHLLPTSGAAATAAAAAAAAAAVTAASTSYYGRDRSPLRRATAPVPTVGEGYGYGHESELSQASAAARNSLYDMARYEREQYA
DRARYSAF

ZCCHC4; 1 $CX_2CX_4HX_4C$

MAASRNGFEAVEAEGSAGCRGSSGMEVVLPLDPAVPAPLCPHGPTLLFVKVTQGKEETRRFYACSACRDRKDCNFFQWEDEKLSGARLA
AREAHNRRCQPPLSRTQCVERYLKFIELPLTQRKFCQTCQQLLLPDDWGQHSEHQVLGNVSITQLRRPSQLLYPLENKKTNAQYLFADR
SCQFLVDLLSALGFRRVLCVGTPRLHELIKLTASGDKKSNIKSLLLDIDFRYSQFYMEDSFCHYNMFNHHFFDGKTALEVCRAFLQEDK
GEGIIMVTDPPFGGLVEPLAITFKKLIAMWKEGQSQDDSHKELPIFWIFPYFFESRICQFFPSFQMLDYQVDYDNHALYKHGKTGRKQS
PVRIFTNIPPNKIILPTEEGYRFCSPCQRYVSLENQHCELCNSCTSKDGRKWNHCFLCKKCVKPSWIHCSICNHCAVPDHSCEGPKHGC
FICGELDHKRSTCPNIATSKRANKAVRKQKQRKSNKMKMETTKGQSMNHTSATRRKKRRERAHQYLGS

SF1/BBP/ZCCHC25; 1 $CX_2CX_3GHX_4C$

MATGANATPLDFPSKKRKRSRWNQDTMEQKTVIPGMPTVIPPGLTREQERAYIVQLQIEDLTRKLRTGDLGIPPNPEDRSPSPEPIYNS
EGKRLNTREFRTRKKLEEERHNLITEMVALNPDFKPPADYKPPATRVSDKVMIPQDEYPEINFVGLLIGPRGNTLKNIEKECNAKIMIR
GKGSVKEGKVGRKDGQMLPGEDEPLHALVTANTMENVKKAVEQIRNILKQGIETPEDQNDLRKMQLRELARLNGTLREDDNRILRPWQS
SETRSITNTTVCTKCGGAGHIASDCKFQRPGDPQSAQDKARMDKEYLSLMAELGEAPVPASVGSTSGPATTPLASAPRPAAPANNPPPP
SLMSTTQSRPPWMNSGPSESRPYHGMHGGGPGGPGGGPHSFPHPLPSLTGGHGGHPMQHNPNGPPPPWMQPPPPPMNQGPHPPGHHGPP
PMGKSVPGKYACGLWGLSPASRKRYDAATTYGHDAAAAAASQWAAPTPSLWSSSPMATTAAAASATPSAQQQYGFQYPLAMAAKIPPRG
GDGPSHESEDFPRPLVTLPGRQPQQRPWWTGWFGKAA

SLU7; 1 $CX_2CX_4HX_4C$

MSATVVDAVNAAPLSGSKEMSLEEPKKMTREDWRKKKELEEQRKLGNAPAEVDEEGKDINPHIPQYISSVPWYIDPSKRPTLKHQRPQP
EKQKQFSSSGEWYKRGVKENSIITKYRKGACENCGAMTHKKKDCFERPRRVGAKFTGTNIAPDEHVQPQLMFDYDGKRDRWNGYNPEEH
MKIVEEYAKVDLAKRTLKAQKLQEELASGKLVEQANSPKHQWGEEEPNSQMEKDHNSEDEDEDKYADDIDMPGQNFDSKRRITVRNLRI
REDIAKYLRNLDPNSAYYDPKTRAMRENPYANAGKNPDEVSYAGDNFVRYTGDTISMAQTQLFAWEAYDKGSEVHLQADPTKLELLYKS
FKVKKEDFKEQQKESILEKYGGQEHLDAPPAELLLAQTEDYVEYSRHGTVIKGQERAVACSKYEEDVKIHNHTHIWGSYWKEGRWGYKC

CHSFFKYSYCTGEAGKEIVNSEECIINEITGEESVKKPQTLMELHQEKLKEEKKKKKKKKKHRKSSSDSDDEEKKHEKLKKALNAEEA
RLLHVKETMQIDERKRPYNSMYETREPTEEEMEAYRMKRQRPDDPMASFLGQ

SREK1IP1/P18SRP; 1 $CX_2CX_3GHX_4C$

MAVPGCNKDSVRAGCKKCGYPGHLTFECRNFLRVDPKRDIVLDVSSTSSEDSDEENEELNKLQALQEKRINEEEEKKKEKSKEKIKLKK
KRKRSYSSSSTEEDTSKQKKQKYQKKEKKKEKKSKSKKGKHHKKEKKKRKKEKHSSTPNSSEFSRK

SRSF7/ZCCHC20/RMB37/9G8; 1 $CX_2CX_3GHX_4C$

SRYGRYGGETKVYVGNLGTGAGKGELERAFSYYGPLRTVWIARNPPGFAFVEFEDPRDAEDAVRGLDGKVICGSRVRVELSTGMPRRSR
FDRPPARRPFDPNDRCYECGEKGHYAYDCHRYSRRRRSRSRSRSHSRSRGRRYSRSRSRSRGRRSRSASPRRSRSISLRRSRSASLRRS
RSGSIKGSRYFQSPSRSRSRSRSISRPRSSRSKSRSPSPKRSRSPSGSPRRSASPERMD

TUT4/ ZCCHC11; 3 $CX_2CX_3GHX_4C$

MEESKTLKSENHEPKKNVICEESKAVQVIGNQTLKARNDKSVKEIENSSPNRNSSKKNKQNDICIEKTEVKSCKVNAANLPGPKDLGLV
LRDQSHCKAKKFPNSPVKAEKATISQAKSEKATSLQAKAEKSPKSPNSVKAEKASSYQMKSEKVPSSPAEAEKGPSLLLKDMRQKTELQ
QIGKKIPSSFTSVDKVNIEAVGGEKCALQNSPRSQKQQTCTDNTGDSDDSASGIEDVSDDLSKMKNDESNKENSSEMDYLENATVIDES
ALTPEQRLGLKQAEERLERDHIFRLEKRSPEYTNCRYLCKLCLIHIENIQGAHKHIKEKRHKKNILEKQEESELRSLPPPSPAHLAALS
VAVIELAKEHGITDDDLRVRQEIVEEMSKVITTFLPECSLRLYGSSLTRFALKSSDVNIDIKFPPKMNHPDLLIKVLGILKKNVLYVDV
ESDFHAKVPVVVCRDRKSGLLCRVSAGNDMACLTTDLLTALGKIEPVFIPLVLAFRYWAKLCYIDSQTDGGIPSYCFALMVMFFLQQRK
PPLLPCLLGSWIEGFDPKRMDDFQLKGIVEEKFVKWECNSSSATEKNSIAEENKAKADQPKDDTKKTETDNQSNAMKEKHGKSPLALET
PNRVSLGQLWLELLKFYTLDFALEEYVICVRIQDILTRENKNWPKRRIAIEDPFSVKRNVARSLNSQLVYEYVVERFRAAYRYFACPQT
KGGNKSTVDFKKREKGKISNKKPVKSNNMATNGCILLGETTEKINAEREQPVQCDEMDCTSQRCIIDNNNLLVNELDFADHGQDSSSLS
TSKSSEIEPKLDKKQDDLAPSETCLKKELSQCNCIDLSKSPDPDKSTGTDCRSNLETESSHQSVCTDTSATSCNCKATEDASDLNDDDN
LPTQELYYVFDKFILTSGKPPTIVCSICKKDGHSKNDCPEDFRKIDLKPLPPMTNRFREILDLVCKRCFDELSPPCSEQHNREQILIGL
EKFIQKEYDEKARLCLFGSSKNGFGFRDSDLDICMTLEGHENAEKLNCKEIIENLAKILKRHPGLRNILPITTAKVPIVKFEHRRSGLE
GDISLYNTLAQHNTRMLATYAAIDPRVQYLGYTMKVFAKRCDIGDASRGSLSSYAYILMVLYFLQQRKPPVIPVLQEIFDGKQIPQRMV
DGWNAFFFDKTEELKKRLPSLGKNTESLGELWLGLLRFYTEEFDFKEYVISIRQKKLLTTFEKQWTSKCIAIEDPFDLNHNLGAGVSRK
MTNFIMKAFINGRKLFGTPFYPLIGREAEYFFDSRVLTDGELAPNDRCCRVCGKIGHYMKDCPKRKSSLLFRLKKKDSEEEKEGNEEEK
DSRDVLDPRDLHDTRDFRDPRDLRCFICGDAGHVRRECPEVKLARQRNSSVAAAQLVRNLVNAQQVAGSAQQQGDQSIRTRQSSECSES
PSYSPQPQPFPQNSSQSAAITQPSSQPGSQPKLGPPQQGAQPPHQVQMPLYNFPQSPPAQYSPMHNMGLLPMHPLQIPAPSWPIHGPVI
HSAPGSAPSNIGLNDPSIIFAQPAARPVAIPNTSHDGHWPRTVAPNSLVNSGAVGNSEPGFRGLTPPIPWEHAPRPHFPLVPASWPYGL
HQNFMHQGNARFQPNKPFYTQDRCATRRCRERCPHPPRGNVSE

TUT7/ZCCHC6; 3 $CX_2CX_3GHX_4C$

MGDTAKPYFVKRTKDRGTMDDDDFRRGHPQQDYLIIDDHAKGHGSKMEKGLQKKKITPGNYGNTPRKGPCAVSSNPYAFKNPIYSQPAW
MNDSHKDQSKRWLSDEHTGNSDNWREFKPGPRIPVINRQRKDSFQENEDGYRWQDTRGCRTVRRLFHKDLTSLETTSEMEAGSPENKKQ
RSRPRKPRKTRNEENEQDGDLEGPVIDESVLSTKELLGLQQAEERLKRDCIDRLKRRPRNYPTAKYTCRLCDVLIESIAFAHKHIKEKR
HKKNIKEKQEEELLTTLPPPTPSQINAVGIAIDKVVQEFGLHNENLEQRLEIKRIMENVFQHKLPDCSLRLYGSSCSRLGFKNSDVNID
IQFPAIMSQPDVLLLVQECLKNSDSFIDVDADFHARVPVVVCREKQSGLLCKVSAGNENACLTTKHLTALGKLEPKLVPLVIAFRYWAK
LCSIDRPEEGGLPPYVFALMAIFFLQQRKEPLLPVYLGSWIEGFSLSKLGNFNLQDIEKDVVIWEHTDSAAGDTGITKEEAPRETPIKR
GQVSLILDVKHQPSVPVGQLWVELLRFYALEFNLADLVISIRVKELVSRELKDWPKKRIAIEDPYSVKRNVARTLNSQPVFEYILHCLR
TTYKYFALPHKITKSSLLKPLNAITCISEHSKEVINHHPDVQTKDDKLKNSVLAQGPGATSSAANTCKVQPLTLKETAESFGSPPKEEM
GNEHISVHPENSDCIQADVNSDDYKGDKVYHPETGRKNEKEKVGRKGKHLLTVDQKRGEHVVCGSTRNNESESTLDLEGFQNPTAKECE
GLATLDNKADLDGESTEGTEELEDSLNHFTHSVQGQTSEMIPSDEEEEDDEEEEEEEEPRLTINQREDEDGMANEDELDNTYTGSGDED
ALSEEDDELGEAAKYEDVKECGKHVERALLVELNKISLKEENVCEEKNSPVDQSDFFYEFSKLIFTKGKSPTVVCSLCKREGHLKKDCP
EDFKRIQLEPLPPLTPKFLNILDQVCIQCYKDFSPTIIEDQAREHIRQNLESFIRQDFPGTKLSLFGSSKNGFGFKQSDLDVCMTINGL
ETAEGLDCVRTIEELARVLRKHSGLRNILPITTAKVPIVKFFHLRSGLEVDISLYNTLALHNTRLLSAYSAIDPRVKYLCYTMKVFTKM
CDIGDASRGSLSSYAYTLMVLYFLQQRNPPVIPVLQEIYKGEKKPEIFVDGWNIYFFDQIDELPTYWSECGKNTESVGQLWLGLLRFYT
EEFDFKEHVISIRRKSLLTTFKKQWTSKYIVIEDPFDLNHNLGAGLSRKMTNFIMKAFINGRRVFGIPVKGFPKDYPSKMEYFFDPDVL
TEGELAPNDRCCRICGKIGHFMKDCPMRRKVRRRRDQEDALNQRYPENKEKRSKEDKEIHNKYTEREVSTKEDKPIQCTPQKAKPMRAA
ADLGREKILRPPVEKWKRQDDKDLREKRCFICGREGHIKKECPQFKGSSGSLSSKYMT
QGKASAKRTQQES

XRN2; 1 $CX_2CX_3GHX_4C$

MGVPAFFRWLSRKYPSIIVNCVEEKPKECNGVKIPVDASKPNPNDVEFDNLYLDMNGIIHPCTHPEDKPAPKNEDEMMVAIFEYIDRLF
SIVRPRRLLYMAIDGVAPRAKMNQQRSRRFRASKEGMEAAVEKQRVREEILAKGGFLPPEEIKERFDSNCITPGTEFMDNLAKCLRYYI
ADRLNNDPGWKNLTVILSDASAPGEGEHKIMDYIRRQRAQPNHDPNTHHCLCGADADLIMLGLATHEPNFTIIREEFKPNKPKPCGLCN
QFGHEVKDCEGLPREKKGKHDELADSLPCAEGEFIFLRLNVLREYLERELTMASLPFTFDVERSIDDWVFMCFFVGNDFLPHLPSLEIR
ENAIDRLVNIYKNVVHKTGGYLTESGYVNLQRVQMIMLAVGEVEDSIFKKRKDDEDSFRRRQKEKRKRMKRDQPAFTPSGILTPHALGS
RNSPGSQVASNPRQAAYEMRMQNNSSPSISPNTSFTSDGSPSPLGGIKRKAEDSDSEPEPEDNVRLWEAGWKQRYYKNKFDVDAADEKF
RRKVVQSYVEGLCWVLRYYYQGCASWKWYYPFHYAPFASDFEGIADMPSDFEKGTKPFKPLEQLMGVFPAASGNFLPPSWRKLMSDPDS
SIIDFYPEDFAIDLNGKKYAWQGVALLPFVDERRLRAALEEVYPDLTPEETRRNSLGGDVLFVGKHHPLHDFILELYQTGSTEPVEVPP
ELCHGIQGKFSLDEEAILPDQIVCSPVPMLRDLTQNTVVSINFKDPQFAEDYIFKAVMLPGARKPAAVLKPSDWEKSSNGRQWKPQLGF
NRDRRPVHLDQAAFRTLGHVMPRGSGTGIYSNAAPPPVTYQGNLYRPLLRGQAQIPKLMSNMRPQDSWRGPPPLFQQQRFDRGVGAEPL
LPWNRMLQTQNAAFQPNQYQMLAGPGGYPPRRDDRGGRQGYPREGRKYPLPPPSGRYNWN

ZCRB1/ZCCHC19/(U11/U12-31K); 1 $CX_2CX_3GHX_4C$

MSGGLAPSKSTVYVSNLPFSLTNNDLYRIFSKYGKVVKVTIMKDKDTRKSKGVAFILFLDKDSAQNCTRAINNKQLFGRVIKASIAIDN
GRAAEFIRRRNYFDKSKCYECGESGHLSYACPKNMLGEREPPKKKEKKKKKKAPEPEEEIEEVEESEDEGEDPALDSLSQAIAFQQAKI
EEEQKKWKPSSGVPSTSDDSRRPRIKKSTYFSDEEELSD

ZCCHC2; 1 $CX_2CX_3GHX_4C$

MLRMKLPLKPTHPAEPPPEAEEPEADARPGAKAPSRRRRDCRPPPPPPPPAGPSRGPLPPPPPPRGLGPPVAGGAAAGAGMPGGGGGPS
AALREQERVYEWFGLVLGSAQRLEFMCGLLDLCNPLELRFLGSCLEDLARKDYHYLRDSEAKANGLSDPGPLADFREPAVRSRLIVYLA
LLGSENREAAGRLHRLLPQVDSVLKSLRAARGEGSRGGAEDERGEDGDGEQDAEKDGSGPEGGIVEPRVGGGLGSRAQEELLLLFTMAS
LHPAFSFHQRVTLREHLERLRAALRGGPEDAEVEVEPCKFAGPRAQNNSAHGDYMQNNESSLIEQAPIPQDGLTVAPHRAQREAVHIEK
IMLKGVQRKRADKYWEYTFKVNWSDLSVTTVTKTHQELQEFLLKLPKELSSETFDKTILRALNQGSLKREERRHPDLEPILRQLFSSSS
QAFLQSQKVHSFFQSISSDSLHSINNLQSSLKTSKILEHLKEDSSEASSQEEDVLQHAIIHKKHTGKSPIVNNIGTSCSPLDGLTMQYS
EQNGIVDWRKQSCTTIQHPEHCVTSADQHSAEKRSLSSINKKKGKPQTEKEKIKKTDNRLNSRINGIRLSTPQHAHGGTVKDVNLDIGS
GHDTCGETSSESYSSPSSPRHDGRESFESEEEKDRDTDSNSEDSGNPSTTRFTGYGSVNQTVTVKPPVQIASLGNENGNLLEDPLNSPK
YQHISFMPTLHCVMHNGAQKSEVVVPAPKPADGKTIGMLVPSPVAISAIRESANSTPVGILGPTACTGESEKHLELLASPLPIPSTFLP
HSSTPALHLTVQRLKLPPPQGSSESCTVNIPQQPPGSLSIASPNTAFIPIHNPGSFPGSPVATTDPITKSASQVVGLNQMVPQIEGNTG
TVPQPTNVKVVLPAAGLSAAQPPASYPLPGSPLAAGVLPSQNSSVLSTAATSPQPASAGISQAQATVPPAVPTHTPGPAPSPSPALTHS
TAQSDSTSYISAVGNTNANGTVVPPQQMGSGPCGSCGRRCSCGTNGNLQLNSYYYPNPMPGPMYRVPSFFTLPSICNGSYLNQAHQSNG
NQLPFFLPQTPYANGLVHDPVMGSQANYGMQQMAGFGRFYPVYPAPNVVANTSGSGPKKNGNVSCYNCGVSGHYAQDCKQSSMEANQQG
TYRLRYAPPLPPSNDTLDSAD

ZCCHC3; 1 $CX_2CX_4HX_4C$ and 2 $CX_2CX_3GHX_4C$

MATGGGAEEERKRGRPQLLPPARPAARGEEADGGREKMGWAQVVKNLAEKKGEFREPRPPRREEESGGGGGSAGLGGPAGLAAPDLGDF
PPAGRGDPKGRRRDPAGEAVDPRKKKGAAEAGRRKKAEAAAAAAMATPARPGEAEDAAERPLQDEPAAAAAGPGKGRFLVRICFQGDEGA
CPTRDFVVGALILRSIGMDPSDIYAVIQIPGSREFDVSFRSAEKLALFLRVYEEKREQEDCWENFVVLGRSKSSLKTLFILFRNETVDV
EDIVTWLKRHCDVLAVPVKVTDRFGIWTGEYKCEIELRQGEGGVRHLPGAFFLGAERGYSWYKGQPKTCFKCGSRTHMSGSCTQDRCFR
CGEEGHLSPYCRKGIVCNLCGKRGHAFAQCPKAVHNSVAAQLTGVAGH

ZCCHC9; 2 $CX_2CX_4HX_4C$ and 2 $CX_2CX_3GHX_4C$

MTRWARVSTTYNKRPLPATSWEDMKKGSFEGTSQNLPKRKQLEANRLSLKNDAPQAKHKKNKKKKEYLNEDVNGFMEYLRQNSQMVHNG
QIIATDSEEVREEIAVALKKDSRREGRRLKRQAAKKNAMVCFHCRKPGHGIADCPAALENQDMGTGICYRCGSTEHEITKCKAKVDPAL
GEFPFAKCFVCGEMGHLSRSCPDNPKGLYADGGCKLCGSVEHLKKDCPESQNSERMVTVGRWAKGMSADYEEILDVPKPQKPKTKIPK
VVNF

ZCCHC10; 1 $CX_2CX_3GHX_4C$

MATPMHRLIARRQAFDTELQPVKTFWILIQPSIVISEANKQHVRCQKCLEFGHWTYECTGKRKYLHRPSRTAELKKALKEKENRLLLQQ
SIGETNVERKAKKKRSKSVTSSSSSSSDSSASDSSSESEETSTSSSSEDSDTDESSSSSSSSASSTTSSSSSDSDSDSSSSSSSSTSTD
SSSDDEPPKKKKKK

ZCCHC13; 4 $CX_2CX_3GHX_4C$

MSSKDFFACGHSGHWARGCPRGGAGGRRGGGHGRGSQCGSTTLSYTCYCCGESGRNAKNCVLLGNICYNCGRSGHIAKDCKDPKRERRQ
HCYTCGRLGHLARDCDRQKEQKCYSCGKLGHIQKDCAQVKCYRCGEIGHVAINCSKARPGQLLPLRQIPTSSQGMSQ

ZCCHC14; 1 $CX_2CX_3GHX_4C$

MASNHPAFSFHQKQVLRQELTQIQSSLNGGGGHGGKGAPGPGGALPTCPACHKITPRTEAPVSSVSNSLENALHTSAHSTEESLPKRPL
GKHSKVSVEKIDLKGLSHTKNDRNVECSFEVLWSDSSITSVTKSSSEVTEFISKLCQLYPEENLEKLIPCLAGPDAFYVERNHVDLDSG
LRYLASLPSHVLKNDHVRRFLSTSSPPQQLQSPSPGNPSLSKVGTVMGVSGRPVCGVAGIPSSQSGAQHHGQHPAGSAAPLPHCSHAGS
AGSALAYRTQMDTSPAILMPSSLQTPQTQEQNGILDWLRKLRLHKYYPVFKQLSMEKFLSLTEEDLNKFESLTMGAKKKLKTQLELEKE
KSERRCLNPSAPPLVTSSGVARVPPTSHVGPVQSGRGSHAAELRVEVEQPHHQLPREGSSSEYSSSSSSPMGVQAREESSDSAEENDRR
VEIHLESSDKEKPVMLLNHFTSSSARPTAQVLPVQNEASSNPSGHHPLPPQMLSAASHITPIRMLNSVHKPERGSADMKLLSSSVHSLL
SLEERNKGSGPRSSMKVDKSFGSAMMDVLPASAPHQPVQVLSGLSESSSMSPTVSFGPRTKVVHASTLDRVLKTAQQPALVVETSTAAT
GTPSTVLHAARPPIKLLLSSSVPADSAISGQTSCPNNVQISVPPAIINPRTALYTANTKVAFSAMSSMPVGPLQGGFCANSNTASPSSH
PSTSFANMATLPSCPAPSSSPALSSVPESSFYSSSGGGGSTGNIPASNPNHHHHHHHQQPPAPPQPAPPPPGCIVCTSCGCSGSCGSSG
LTVSYANYFQHPFSGPSVFTFPFLPFSPMCSSGYVSAQQYGGGSTFPVVHAPYSSSGTPDPVLSGQSTFAVPPMQNFMAGTAGVYQTQG
LVGSSNGSSHKKSGNLSCYNCGATGHRAQDCKQPSMDFNRPGTFRLKYAPPAESLDSTD

ZCCHC24; 1 $CX_2CX_3GHX_4C$

MSLLSAIDTSAASVYQPAQLLNWVYLSLQDTHQASAFDAFRPEPTAGAAPPELAFGKGRPEQLGSPLHSSYLNSFFQLQRGEALSNSVY
KGASPYGSLNNIADGLSSLTEHFSDLTLTSEARKPSKRPPPNYLCHLCFNKGHYIKDCPQARPKGEGLTPYQGKKRCFGEYKCPKCKRK
WMSGNSWANMGQECIKCHINVYPHKQRPLEKPDGLDVSDQSKEHPQHLCEKCKVLGYYCRRVQ

PEG10; 1 $CX_2CX_3GHX_4C$

MTERRRDELSEEINNLREKVMKQSEENNNLQSQVQKLTEENTTLREQVEPTPEDEDDDIELRGAAAAAAPPPPIEEECPEDLPEKFDGN
PDMLAPFMAQCQIFMEKSTRDFSVDRVRVCFVTSMMTGRAARWASAKLERSHYLMHNYPAFMMEMKHVFEDPQRREVAKRKIRRLRQGM
GSVIDYSNAFQMIAQDLDWNEPALIDQYHEGLSDHIQEELSHLEVAKSLSALIGQCIHIERRLARAAAARKPRSPPRALVLPHIASHHQ

VDPTEPVGGARMRLTQEEKERRRKLNLCLYCGTGGHYADNCPAKASKSSPAGKLPGPAVEGPSATGPEIIRSPQDDASSPHLQVMLQIH
LPGRHTLFVRAMIDSGASGNFIDHEYVAQNGIPLRIKDWPILVEAIDGRPIASGPVVHETHDLIVDLGDHREVLSFDVTQSPFFPVVLG
VRWLSTHDPNITWSTRSIVFDSEYCRYHCRMYSPIPPSLPPPAPQPPLYYPVDGYRVYQPVRYYYVQNVYTPVDEHVYPDHRLVDPHIE
MIPGAHSIPSGHVYSLSEPEMAALRDFVARNVKDGLITPTIAPNGAQVLQVKRGWKLQVSYDCRAPNNFTIQNQYPRLSIPNLEDQAHL
ATYTEFVPQIPGYQTYPTYAAYPTYPVGFAWYPVGRDGQGRSLYVPVMITWNPHWYRQPPVPQYPPPQPPPPPPPPPPPSYSTL

RTL3/ZCCHC5; 1 $CX_2CX_3GHX_4C$
MVEDLAASYIVLKLENEIRQAQVQWLMEENAALQAQIPELQKSQAAKEYDLLRKSSEAKEPQKLPEHMNPPAAWEAQKTPEFKEPQKPP
EPQDLLPWEPPAAWELQEAPAAPESLAPPATRESQKPPMAHEIPTVLEGQGPANTQDATIAQEPKNSEPQDPPNIEKPQEAPEYQETAA
QLEFLELPPPQEPLEPSNAQEFLELSAAQESLEGLIVVETSAASEFPQAPIGLEATDFPLQYTLTFSGDSQKLPEFLVQLYSYMRVRGH
LYPTEAALVSFVGNCFSGRAGWWFQLLLDIQSPLLEQCESFIPVLQDTFDNPENMKDANQCIHQLCQGEGHVATHFHLIAQELNWDEST
LWIQFQEGLASSIQDELSHTSPATNLSDLITQCISLEEKPDPNPLGKSSSAEGDGPESPPAENQPMQAAINCPHISEAEWVRWHKGRLC
LYCGYPGHFARDCPVKPHQALQAGNIQACQ

RTL4/ZCCHC16; 1 $CX_2CX_3GHX_4C$
MEKCTKSSSTMQVEPSFLQAENLILRLQMQHPTTENTAKRGQVMPALATTVMPVPYSLEHLTQFHGDPANCSEFLTQVTTYLTALQISN
PANDAQIKLFFDYLSQQLESCGIISGPDKSTLLKQYENLILEFQQSFGKPTKQEINPLMNAKFDKGDNSSQQDPATFHLLAQNLICNET
NQSGQFEKALADPNQDEESVTDMMDNLPDLITQCIQLDKKHSDRPELLQSETQLPLLASLIQHQALFSPTDPPPKKGPIQLREGQLPLT
PAKRARQQETQLCLYCSQSGHFTRDCLAKRSRAPATTNNTAHQ

PNMA3; 1 $CX_2CX_3GHX_4C$
MPLTLLQDWCRGEHLNTRRCMLILGIPEDCGEDEFEETLQEACRHLGRYRVIGRMFRREENAQAILLELAQDIDYALLPREIPGKGGPW
EVIVKPRNSDGEFLNRLNRFLEEERRTVSDMNRVLGSDTNCSAPRVTISPEFWTWAQTLGAAVQPLLEQMLYRELRVFSGNTISIPGAL
AFDAWLEHTTEMLQMWQVPEGEKRRRLMECLRGPALQVVSGLRASNASITVEECLAALQQVFGPVESHKIAQVKLCKAYQEAGEKVSSF
VLRLEPLLQRAVENNVVSRRNVNQTRLKRVLSGATLPDKLRDKLKLMKQRRKPPGFLALVKLLREEEEWEATLGPDRESLEGLEVAPRP
PARITGVGAVPLPASGNSFDVRPSQGYRRRRGRGQHRRGGVARAGSRGSRKRKRHTFCYSCGEDGHIRVQCINPSNLLLVKQKKQAAVE
SGNGNWAWDKSHPKSKAK

PNMA7A/ZCCHC12/SIZN1; 1 $CX_2CX_3GHX_4C$
MASIIARVGNSRRLNAPLPPWAHSMLRSLGRSLGPIMASMADRNMKLFSGRVVPAQGEETFENWLTQVNGVLPDWNMSEEEKLKRLMKT
LRGPAREVMRVLQATNPNLSVADFLRAMKLVFGESESSVTAHGKFFNTLQAQGEKASLYVIRLEVQLQNAIQAGIIAEKDANRTRLQQL
LLGGELSRDLRLRLKDFLRMYANEQERLPNFLELIRMVREEEDWDDAFIKRKRPKRSESMVERAVSPVAFQGSPPIVIGSADCNVIEID
DTLDDSDEDVILVESQDPPLPSWGAPPLRDRARPQDEVLVIDSPHNSRAQFPSTSGGSGYKNNGPGEMRRARKRKHTIRCSYCGEEGHS
KETCDNESDKAQVFENLIITLQELTHTEMERSRVAPGEYNDFSEPL

PNMA7B/ZCCHC18; 1 $CX_2CX_3GHX_4C$
MASITACVGNSRQQNAPLPPWAHSMLRSLGRSLCPLVVKMAERNMKLFSGRVVPAQGKETFENWLIQVNEVLPDWSMSEEEKLKRLMKT
LRGPAREVMRLLQAANPNLSVADFLRAMKLVFGESESSVTAHGKFFNTLQAQGEKASLYVIRLEVQLQNAIQAGILAEKDANQTRLQQL
LLGAELNRDLRFRLKHLLRMYANKQERLPNFLELIKMIREEEDWDDAFIKRKRPKRSEPIMERAASPVAFQGAQPIAISSADCNCNVIE
IDDTLDDSDEDVILVVSLYPSLTPTGAPPFRGRARPLDQVLVIDSPNNSGAQSLSTSGGSGYKNDGPGNIRRARKRKYTTRCSYCGEEG
HSKETCDNESNKAQVFENLIITLQELTHTEERSKEVPGEHSDASEPQ

Supplementary Figure 3. Human ZCCHC domain-containing proteins. The number of $CX_2CX_4HX_4C$ or $CX_2CX_3GHX_4C$ sequences in each protein is shown; these sequences have been highlighted. A slash separates alternative names given by different authors to the same protein.

Supplementary Figure 4. Amino acid frequencies in ZCCHC domains.

		Amino acid																			
	Position	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Saccharomyces cerevisiae	1	0	0	0	0	0	0	0	4	2	1	0	0	2	2	3	1	0	4	0	1
	2	1	0	0	0	3	0	0	1	3	2	1	0	1	3	1	1	0	1	1	1
	3	0	**20**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	4	0	0	0	1	1	0	0	0	1	0	0	2	1	0	1	1	3	0	0	**9**
	5	0	0	0	0	0	0	1	2	2	2	0	**10**	0	0	1	0	1	1	0	0
	6	0	**20**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	7	0	0	2	0	0	**10**	0	0	1	0	0	5	0	0	0	2	0	0	0	0
	8	1	0	0	4	0	3	0	0	2	1	0	1	0	5	0	2	1	0	0	0
	9	3	0	0	0	0	0	0	1	4	0	1	1	2	0	3	0	4	1	0	0
	10	0	0	2	0	0	**15**	0	0	1	0	0	1	0	0	1	0	0	0	0	0
	11	0	0	0	0	0	0	**20**	0	0	0	0	0	0	0	0	0	0	0	0	0
	12	0	0	0	0	3	0	0	2	2	3	2	1	0	0	0	2	0	2	1	2
	13	2	0	0	1	0	2	0	1	4	0	0	0	0	1	4	4	1	0	0	0
	14	1	0	2	3	0	0	0	0	5	0	0	0	0	0	4	3	1	0	0	1
	15	0	0	**11**	3	0	0	0	0	0	0	0	2	0	2	2	0	0	0	0	0
	16	0	**20**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	17	1	0	1	0	0	0	0	0	1	0	2	1	**11**	1	0	0	2	0	0	0
	18	0	0	0	4	0	0	5	0	2	0	1	2	0	0	0	4	1	1	0	0
Homo sapiens	1	3	0	6	0	0	3	1	4	6	2	1	3	2	5	5	3	4	7	0	2
	2	2	3	0	1	1	6	2	4	8	4	0	1	1	1	10	2	4	5	0	2
	3	0	**57**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	4	1	0	1	2	11	1	3	0	3	3	0	1	0	1	3	4	1	0	0	**22**
	5	0	0	0	2	2	0	2	6	7	6	0	10	0	0	10	3	1	2	0	6
	6	0	**57**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	7	1	0	1	1	1	**43**	1	0	2	1	0	1	0	2	1	2	0	0	0	0
	8	2	1	1	15	0	4	0	0	11	0	2	1	0	3	7	6	1	1	0	2
	9	2	0	2	8	4	2	0	5	5	7	2	0	5	0	4	7	3	1	0	0
	10	0	0	3	3	0	**47**	0	0	0	0	1	0	0	0	0	1	2	0	0	0
	11	0	0	0	0	0	0	**57**	0	0	0	0	0	0	0	0	0	0	0	0	0
	12	1	0	0	2	4	1	2	7	2	12	3	0	0	1	2	3	0	3	5	9
	13	**20**	0	0	0	1	1	0	5	7	1	3	0	0	2	3	7	4	3	0	0
	14	4	0	2	2	1	1	1	2	**21**	0	0	2	1	2	8	2	3	1	0	4
	15	2	0	**25**	11	0	0	0	0	1	0	0	7	0	2	1	4	3	0	0	1
	16	0	**57**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	17	1	0	5	1	2	0	1	1	7	1	0	1	**26**	0	2	3	6	0	0	0
	18	5	0	3	8	1	2	2	2	9	5	3	5	0	6	2	0	2	2	0	0
Arabidopsi thaliana	1	6	0	5	6	2	21	2	4	13	6	1	9	6	0	17	10	5	5	0	3
	2	8	1	7	5	5	11	0	9	14	7	0	3	6	1	12	14	6	11	0	1
	3	0	**121**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	4	2	0	1	3	30	3	5	3	6	4	0	6	0	3	2	7	3	0	1	**42**
	5	0	1	0	10	2	1	6	15	16	4	2	24	0	3	14	10	4	5	0	4
	6	0	**121**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	7	1	2	0	1	4	**87**	2	0	9	0	0	4	0	1	6	4	0	0	0	0
	8	4	0	10	24	0	18	4	3	15	1	3	3	0	13	12	6	1	4	0	0
	9	3	0	3	10	4	1	0	1	9	13	7	4	14	4	6	12	9	17	0	4
	10	1	0	4	2	0	**98**	0	0	0	0	0	2	0	0	1	6	7	0	0	0
	11	0	0	0	0	0	0	**121**	0	0	0	0	0	0	0	0	0	0	0	0	0
	12	1	1	5	0	23	2	1	14	1	15	6	8	3	1	10	6	4	6	13	1
	13	**58**	1	0	3	1	2	0	6	0	5	2	0	1	1	4	23	8	6	0	0
	14	6	0	0	1	2	1	3	2	18	3	4	5	1	4	**51**	10	5	1	0	4
	15	10	0	**40**	19	0	2	6	1	1	3	0	12	0	5	3	4	9	1	0	5
	16	0	**121**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	17	2	0	3	2	1	3	3	0	5	7	4	0	**53**	1	9	10	16	1	1	0
	18	4	0	5	7	2	10	1	5	10	2	0	21	0	6	8	18	9	3	0	10
All together	1	9	0	11	6	2	24	3	12	21	9	2	12	10	7	25	14	9	16	0	6
	2	11	4	7	6	9	17	2	14	25	13	1	4	8	5	23	17	10	17	1	4
	3	0	**198**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	4	3	0	2	6	42	4	8	3	10	7	0	9	1	4	6	12	7	0	1	**73**
	5	0	1	0	12	4	1	9	23	25	12	2	44	0	3	25	13	6	8	0	10
	6	0	**198**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	7	2	2	3	2	5	**140**	3	0	12	1	0	10	0	3	7	8	0	0	0	0
	8	7	1	11	43	0	25	4	3	28	2	5	5	0	21	19	14	3	5	0	2
	9	8	0	5	18	8	3	0	7	18	20	10	5	21	4	13	19	16	19	0	4
	10	1	0	9	5	0	**160**	0	0	1	0	1	3	0	0	2	7	9	0	0	0
	11	0	0	0	0	0	0	**198**	0	0	0	0	0	0	0	0	0	0	0	0	0
	12	2	1	5	2	30	3	3	23	5	30	11	9	3	2	12	11	4	11	19	12
	13	**80**	1	0	4	2	5	0	12	11	6	5	0	1	4	11	34	13	9	0	0
	14	11	0	4	6	3	2	4	4	44	3	4	7	2	6	**63**	15	9	2	0	9
	15	12	0	76	33	0	2	6	1	2	3	0	21	0	9	6	8	12	1	0	6
	16	0	**198**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	17	4	0	9	3	3	3	4	1	13	8	6	2	**90**	2	11	13	24	1	1	0
	18	9	0	8	19	3	12	8	7	21	7	4	28	0	12	10	22	12	6	0	10

Values indicate the number of cases in which each amino acid is present in a given position (from the first to the eightieth; 1-18 in the column headed as “Position”) within 20 ZCCCH domains from 7 yeast factors, 57 from 34 human factors, and 121 from 69 Arabidopsis factors. Numbers are also shown for all the 198 sequences from the 110 factors taken together. Transposon-derived and uncertain ZCCHC domains have been omitted from the analysis. The C and H residues that characterize the ZCCHC domain are highlighted in red, the highly conserved G at the seventh and tenth positions in green, and residues occupying positions with a frequency higher than 30% in blue.

Plant Physiology, 2025, **197**, kiae641
https://doi.org/10.1093/plphys/kiae641
Advance access publication 5 December 2024
Research Article

CAX-INTERACTING PROTEIN4 depletion causes early lethality and pre-mRNA missplicing in Arabidopsis

Uri Aceituno-Valenzuela,[1,†,‡] Sara Fontcuberta-Cervera,[1,‡] Rosa Micol-Ponce,[1] Raquel Sarmiento-Mañús,[1] Alejandro Ruiz-Bayón,[1] María Rosa Ponce[1,*]

[1]Instituto de Bioingeniería, Universidad Miguel Hernández, Campus de Elche, 03202 Elche, Alicante, Spain
*Author for correspondence: mrponce@umh.es
† Present address: Universidad de O'Higgins, Centro UOH de Biología de Sistemas para la Sanidad Vegetal (BioSaV). Ruta I-90 s/n, San Fernando, Chile.
‡ These authors contributed equally to this work.
The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (https://academic.oup.com/plphys/pages/General-Instructions) is María Rosa Ponce (mrponce@umh.es).

Abstract
Zinc knuckle (ZCCHC) motif-containing proteins are present in unicellular and multicellular eukaryotes, and most ZCCHC proteins with known functions participate in the metabolism of various classes of RNA, such as mRNAs, ribosomal RNAs, and microRNAs. The Arabidopsis (*Arabidopsis thaliana*) genome encodes 69 ZCCHC-containing proteins; however, the functions of most remain unclear. One of these proteins, CAX-INTERACTING PROTEIN 4 (CXIP4, encoded by AT2G28910), has been classified as a PTHR31437 family member. This family includes human Splicing regulatory glutamine/lysine-rich protein 1 (SREK1)-interacting protein 1 (SREK1IP1), which is thought to function in pre-mRNA splicing and RNA methylation. Metazoan SREK1IP1-like and plant CXIP4-like proteins only share a ZCCHC motif, and their functions remain almost entirely unknown. Here, we studied two loss-of-function alleles of Arabidopsis *CXIP4*: *cxip4-1* is likely null and shows early lethality, and *cxip4-2* is hypomorphic and viable, with pleiotropic morphological defects. The *cxip4-2* mutant exhibited deregulation of defense genes and upregulation of transcription factor genes, some of which might explain its developmental defects. The *cxip4-2* mutant also exhibited increased intron retention events, being more evident in *cxip4-1*. The specific functions of misspliced genes, such as those involved in "gene silencing by DNA methylation" and "mRNA polyadenylation factor" suggest that CXIP4 has additional functions. In *cxip4-2* plants, polyadenylated RNAs accumulate in the nucleus; these could be misspliced mRNAs. The CXIP4 protein localizes to the nucleus in a pattern resembling nuclear speckles rich in splicing factors. Therefore, *CXIP4* is required for plant development and survival and mRNA maturation.

Introduction

Zinc knuckles are 18-residue CCHC-type zinc finger (ZCCHC) motifs with the conserved sequence $CX_2CX_4HX_4C$. These motifs are present in proteins of unicellular and multicellular eukaryotes, from yeast to plants and animals, but are not found in bacteria. Most ZCCHC-containing proteins with known functions are involved in the metabolism of different classes of RNAs, including messenger RNAs (mRNAs), ribosomal RNAs (rRNAs), and microRNAs (miRNAs). The Arabidopsis (*Arabidopsis thaliana*) genome encodes 69 ZCCHC-containing proteins, only a few of which have been functionally characterized at some level (reviewed in Aceituno-Valenzuela et al. 2020).

The AT2G28910 Arabidopsis gene encodes the ZCCHC-containing protein CAX-INTERACTING PROTEIN 4 (CXIP4). Arabidopsis CXIP4 activated the H^+/Ca^{2+} antiporter CAX1 when expressed in yeast (Cheng et al. 2004). CXIP4 has been classified as a member of the Splicing regulatory glutamine/lysine-rich protein 1 (SREK1)-interacting protein 1 (SREK1IP1) family, which has rarely been studied. Human SREK1IP1 (also known as P18SRP or SFRS12IP1) interacts with the serine-arginine (SR)-rich splicing regulatory protein SREK1, as revealed in yeast two-hybrid (Y2H) assays. In turn, SREK1 interacts with other SR proteins, modulating splice site (SS) selection during alternative splicing (AS) events (Heese et al. 2004). Downregulation of human *SREK1IP1* by RNA interference promoted cell proliferation, migration, and invasion in tumor cell lines (Akiyama et al. 2016), pointing to its role in suppressing tumor growth. Human SREK1IP1 has also been identified as an interactor of METHYLTRANSFERASE LIKE 16 (METTL16; Covelo-Molares et al. 2021), one of the two enzymes that catalyze the deposition of N^6-methyladenosine (m^6A) epitranscriptomic marks onto mRNAs and the U6 small nuclear RNA (snRNA) (Warda et al. 2017). Dominant frameshift alleles of *SREK1IP1* have been associated with congenital anosmia (Kamarck et al. 2024).

We previously identified Arabidopsis CXIP4 in a Y2H-based screen for interactors of MORPHOLOGY OF ARGONAUTE1-52 SUPPRESSED2 (MAS2; Sánchez-García et al. 2015), which is the putative Arabidopsis ortholog of human NF-kappa-B-activating protein (NKAP; Chen et al. 2003). The multifunctional protein NKAP was recently identified as an exon ligation factor in the post-catalytic (C*) spliceosome (Fica et al. 2019), and as an m^6A epitranscriptomic mark reader during miRNA and mRNA maturation (Zhang et al. 2019; Sun et al. 2022).

Received June 6, 2024. Accepted November 8, 2024.

Overexpression of Arabidopsis *CXIP4* conferred resistance to dithiothreitol (DTT) and tunicamycin (TM), two drugs that induce endoplasmic reticulum stress and trigger the unfolded protein response (UPR; Hossain et al. 2016). The wheat (*Triticum aestivum*) putative CXIP4 ortholog (TaCAXIP4) has been proposed to participate in calcium-mediated plant immune responses during Fusarium infection (Chen et al. 2022). More recently, TaCAXIP4 has been found in a complex with several factors involved in the control of AS, including SR-rich proteins (He et al. 2024). Despite these findings, the precise functions of plant CXIP4-like and metazoan SREK1IP1-like proteins remain elusive, and loss-of-function mutants have not been characterized in any eukaryote.

Here, we performed functional analysis of Arabidopsis CXIP4 based on two T-DNA insertional alleles: *cxip4-1* and *cxip4-2*. Our findings provide molecular and genetic evidence for the essential role of CXIP4 in Arabidopsis development and survival, as well as its involvement in pre-mRNA splicing.

Results

CXIP4 is an essential gene, required for plant survival and proper development

Most ZCCHC-containing proteins studied to date participate in RNA metabolism (Aceituno-Valenzuela et al. 2020). Arabidopsis *CXIP4* (AT2G28910) is a single-copy gene with two exons (Fig. 1A), the second of which encodes a protein of 332 amino acids (aa) containing a ZCCHC motif (residues 81 to 98; Supplementary Fig. S1). As annotated in the UniProtKB database (https://www.uniprot.org/uniprotkb), CXIP4 also contains two intrinsically disordered regions (IDRs) flanking the ZCCHC motif which occupy a large part of the protein (residues 33 to 59 and 124 to 332); the latter IDR being part of a region rich in arginine (R; residues 198 to 289; Supplementary Fig. S1).

To date, only a single functional analysis of the *CXIP4* gene in Arabidopsis has been performed, which was based on transgenic lines with β-estradiol-inducible overexpression of the full-length cDNA (Coego et al. 2014). To investigate the effects of the loss of function of *CXIP4*, we obtained two T-DNA insertional lines with disruptions in the coding region of AT2G28910: GK-537C02 and SALK_044245 (Fig. 1A). GK-537C02 seeds either did not germinate or produced phenotypically wild-type plants or callus-like seedlings with early arrested development (Fig. 1, B, C, E, F, H, and I). The phenotypically wild-type plants were found to be homozygous for the wild-type allele of *CXIP4* or hemizygous for the T-DNA insertion; siliques of the latter plants exhibited some light-green, immature seeds that contained early arrested embryos (Fig. 2, A to D). We sowed 3,626 additional seeds from hemizygous plants, finding that 15.6% exhibited a wrinkled appearance and failed to germinate (12.6%) or produced dark callus-like seedlings (3%), while the remaining 84.4% gave rise to phenotypically wild-type plants (Fig. 2, E and F). We genotyped 30 callus-like seedlings and 170 phenotypically wild-type plants. All aberrant seedlings were homozygous for the T-DNA insertion, whereas those with a wild-type phenotype were hemizygous or homozygous for the wild-type allele of *CXIP4*. We named the insertional allele of *CXIP4* carried by

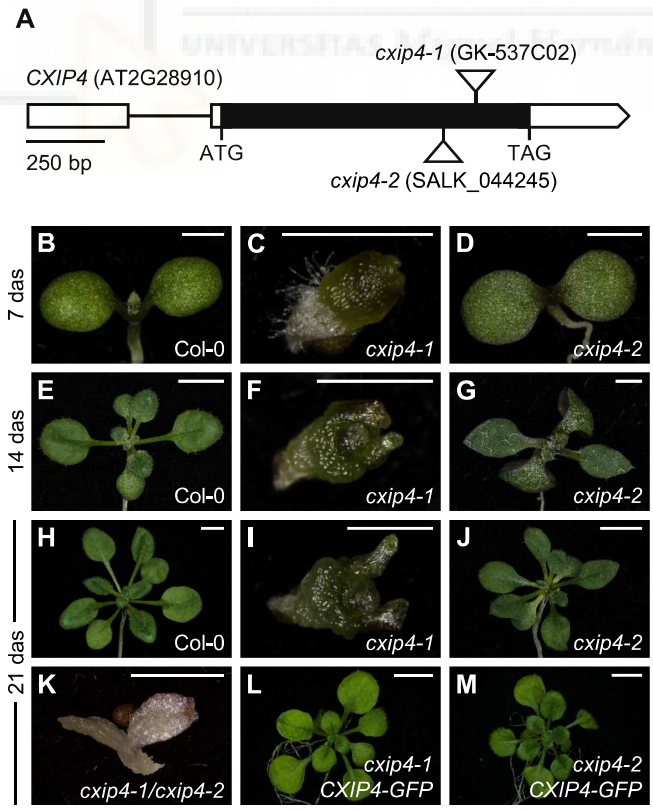


Figure 1. Structure of the *CXIP4* gene and developmental phenotypes of *cxip4* insertional alleles. **A)** Schematic representation of the *CXIP4* gene, indicating the positions of its start (ATG) and stop (TAG) codons and the T-DNA insertions (triangles) in the *cxip4* alleles studied in this work. Boxes represent exons, with untranslated (UTRs) and coding regions shown in white and black, respectively. **B–M)** Morphological vegetative phenotypes of (B, E, H) Col-0, (C, F, I) *cxip4-1*, (D, G, J) *cxip4-2*, (K) *cxip4-1/cxip4-2*, (L) *cxip4-1 CXIP4*$_{pro}$*:CXIP4:GFP*, and (M) *cxip4-2 CXIP4*$_{pro}$*:CXIP4:GFP* plants. Photographs were taken (B–D) 7, (E–G) 14, and (H–M) 21 days after stratification (das). Scale bars: (B, C, D, F, G, I, and K) 1 mm, and (E, H, J, L, and M) 5 mm.

the GK-537C02 line *cxip4-1*. The *cxip4-1/cxip4-1* seedlings did not develop true leaves or any other recognizable organ (Fig. 1, C, F, and I). The ungerminated seeds were assumed to also be homozygous for the *cxip4-1* allele, and we did not genotype them.

Some seeds of the SALK_044245 line yielded plants with slowed growth and defects in both vegetative and reproductive development (Fig. 1, D, G, and J; Fig. 3; Supplementary Fig. S2). Seven days after stratification (das), plants homozygous for the T-DNA insertion were smaller than Col-0 plants, and some had three cotyledons (8.8% of 979 seedlings; Fig. 1, B and D). 14 to 21 das, the mutant plants showed pointed and reticulated leaves with a reduced density of trichomes, which varied considerably among leaves of different individuals, and whose branching was not affected; rosette leaves also produced high levels of anthocyanins (Fig. 1, E, G, H, and J; Fig. 3, A to J). These plants flowered late, developed numerous shoots with reduced apical dominance, and remained green for more than 80 days (Supplementary Fig. S2). In addition, their flowers produced short siliques containing likely undeveloped ovules, along with a few mature seeds that were larger than those of Col-0 (Fig. 3, K to O).

The SALK_044245 line contains two additional annotated T-DNA insertions besides the one disrupting the *CXIP4* gene: one in AT2G23640 and another downstream of the 3′ untranslated region (3′-UTR) of AT3G50030 (Supplementary Fig. S3). We did not eliminate these insertions because (i) the ones in AT2G23640 and *CXIP4* (AT2G28910) are linked, being 2.36 megabases (Mb) apart. The other T-DNA insertion is located in an intergenic region, 88 bp downstream of AT3G50030, as we reconfirmed by Sanger sequencing, in which we also identified a duplicated T-DNA 58 bp upstream of the annotated one. (ii) Both genes, AT2G23640 and AT3G50030, are not expressed in the wild-type plants, except AT2G23640 that expresses in seeds, as described in eFP Browser (Klepikova et al. 2016). (iii) We obtained T-DNA insertional lines disrupting the coding regions of AT2G23640 and AT3G50030 genes, and confirmed that these lines do not exhibit a mutant phenotype in any vegetative or reproductive developmental step (Supplementary Fig. S3). We named the insertional allele of *CXIP4* carried by the SALK_044245 line *cxip4-2*.

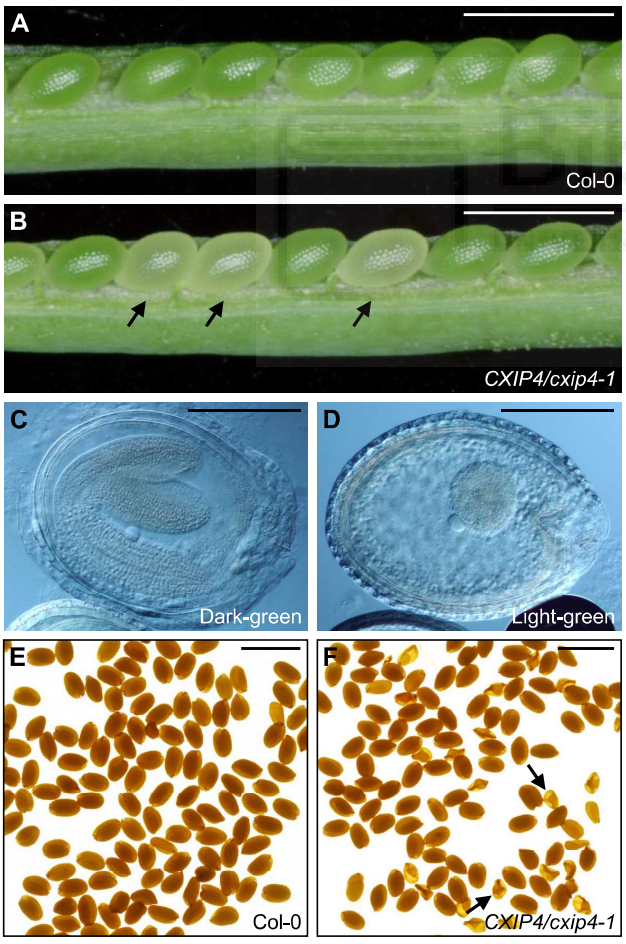


Figure 2. Developmental defects caused by the *cxip4-1* allele. **A, B)** Dissected immature siliques from (A) Col-0 and (B) *CXIP4/cxip4-1* plants; the latter exhibit light-green seeds (black arrows). **C, D)** Embryos in (C) dark- and (B) light-green seeds from siliques of *CXIP4/cxip4-1* plants. **E, F)** Seeds from (E) Col-0 and (F) *CXIP4/cxip4-1* plants; the latter developed some wrinkled seeds (black arrows), whose genotype was found to be *cxip4-1/cxip4-1*. Photographs in (A–D) were taken 60 das. Scale bars: (A, B, E, and F) 1 mm, and (C, D) 100 μm.

The T-DNA insertions in *cxip4-1* and *cxip4-2* are at similar locations (Fig. 1A), but their mutant phenotypes are quite different (Fig. 1, C, D, F, G, I, and J). Therefore, to further examine if *cxip4-1* and *cxip4-2* are null, we used several combinations of primers, flanking the insertions or hybridizing within the T-DNA and the *CXIP4* gene (Supplementary Fig. S4A and Table S1), to perform reverse transcription PCR (RT-PCR) analyses. We detected transcripts upstream of the insertions in both *cxip4-1* and *cxip4-2*, although the levels were low in *cxip4-2*. As expected, we did not detect transcripts using primers flanking the insertions, and only *cxip4-2* showed slight transcription downstream of the insertion (Supplementary Fig. S4B).

We then sequenced the PCR products obtained using primers that hybridize within the T-DNA and the *CXIP4* gene, which resulted in chimeric sequences combining *CXIP4* and T-DNA (Supplementary Fig. S4C). The mRNA from *cxip4-1* was predicted to encode a mutant CXIP4-1 protein four aa longer than the wild-type CXIP4, with only 271 aa shared between the mutant and wild type. Translation of the chimeric *cxip4-2* mRNA would result in a protein of 240 aa missing part of the C-terminal region (Supplementary Figs. S1 and S4C). The aberrant CXIP4-1 protein appears to lack activity, but the truncated CXIP4-2 protein may be partially functional. These results might explain the stronger phenotype observed in *cxip4-1* plants compared with *cxip4-2*. Since *cxip4-1* presents a much stronger phenotype than *cxip4-2*, and considering the results obtained through RT-PCR, we concluded that *cxip4-1* is likely to be a null allele of *CXIP4* whereas *cxip4-2* is hypomorphic.

We carried out a complementation test by crossing *CXIP4/cxip4-1* to *cxip4-2/cxip4-2* plants. We obtained four *cxip4-1/cxip4-2* F_1 plants from two different crosses, one of which displayed three cotyledons, as did some *cxip4-2/cxip4-2* plants. The four *cxip4-1/cxip4-2* heterozygous plants displayed a strong mutant phenotype, more similar to that of *cxip4-1/cxip4-1* than to *cxip4-2/cxip4-2* homozygotes, and their development was arrested after the emergence of cotyledons and one or two small leaves (Fig. 1K). These results confirm the allelism of *cxip4-1* and *cxip4-2*, and with the recessive nature of both, suggest that they are null and hypomorphic alleles of *CXIP4*, respectively.

To confirm that the disruption of *CXIP4* by the T-DNA insertions in *cxip4-1* and *cxip4-2* was the cause of their mutant phenotypes, we transferred a copy of the wild-type *CXIP4* gene fused to green fluorescent protein (GFP) into *CXIP4/cxip4-1* and *cxip4-2/cxip4-2* plants. Indeed, homozygous *cxip4-1* $CXIP4_{pro}$:*CXIP4:GFP* and *cxip4-2* $CXIP4_{pro}$:*CXIP4:GFP* plants were phenotypically wild type (Fig. 1, L and M), confirming in the last plants that vegetative and reproductive mutant traits of *cxip4-2* (growth retardation, leaf and rosette morphology and size, trichome density, anthocyanin levels, flowering time, plant architecture, silique development,

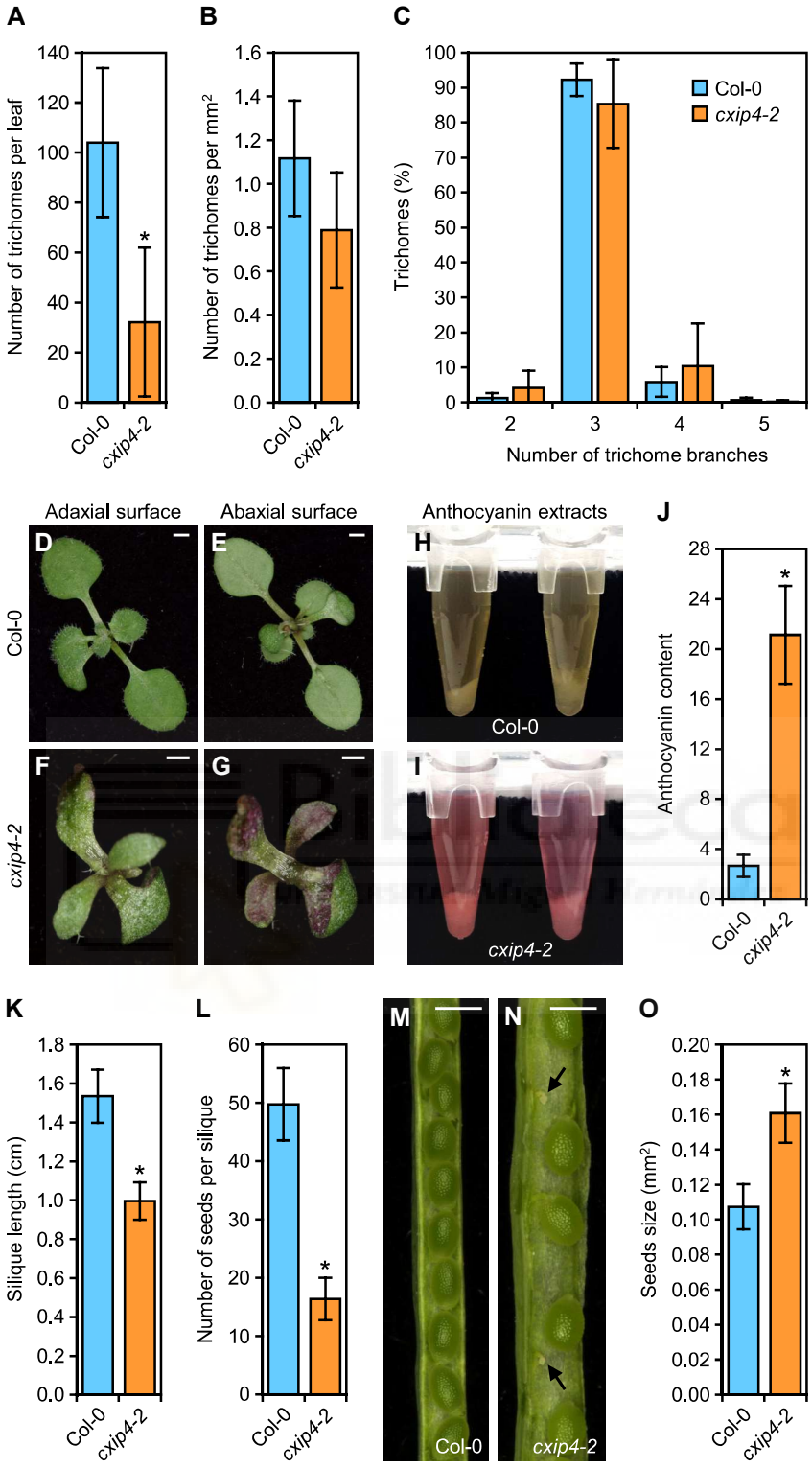


Figure 3. Pleiotropy of the morphological phenotype of *cxip4-2* plants. **A–C)** Trichome density and branching on the third leaves of the wild-type Col-0 and the *cxip4-2* mutant. **A)** Total number of trichomes. **B)** Trichome density. **C)** Percentage of trichomes with different levels of branching. The trichomes of ten leaves per genotype were collected 21 (Col-0) or 29 (*cxip4-2*) das. **D–J)** Anthocyanin accumulation in the leaves of *cxip4-2*. **D–G)** Adaxial and abaxial surfaces of rosettes of Col-0 and *cxip4-2* collected 12 das. **H, I)** Anthocyanin extracts from (H) Col-0 and (I) *cxip4-2*. **J)** Anthocyanin contents of ten rosettes per genotype, each collected 12 das. **K, L)** The (K) silique length, and (L) fertility of Col-0 and *cxip4-2* plants. Ten siliques collected from five plants per genotype were analyzed. **M, N)** Dissected immature siliques from (M) Col-0 and (N) *cxip4-2* plants, with the latter exhibiting undeveloped ovules (black arrows). **O)** Sizes of Col-0 and *cxip4-2* seeds, determined from 60 seeds per genotype. Error bars in (A–C), (J–L), and (O) represent standard deviations. Asterisks in (A), (J–L), and (O) indicate values significantly different from the wild type, determined using Student's *t*-test ($^*P < 0.0001$). Scale bars: (D–G) 1 mm and (M, N) 0.5 mm.

and size and number of seeds; Figs. 1 and 3; Supplementary Fig. S2) were normalized.

CXIP4 is ubiquitously expressed in Arabidopsis

Cis-regulatory elements in Arabidopsis are usually located within the 500-bp region upstream of the transcription start site (TSS), with the majority of these elements (86%) located proximal to the TSS (+1 bp position), spanning from −1,000 bp to +200 bp, and a low density of regulatory sequences at more than 1.5 kb from the TSS (Yu et al. 2016). AT2G28910 (*CXIP4*) and AT2G28920 are oriented head-to-head, with their respective 5′ ends toward each other and their translation start codons 1,026 bp away from each other, sharing an intergenic region of 401 bp that could potentially contain a bidirectional promoter. No 5′-UTR is annotated for AT2G28920, but three different splice variants have been found for *CXIP4*, differing only in the length of the 5′-UTR, due to the AS of its single intron that interrupts this region. Indeed, the coding region of *CXIP4* begins in its second exon, and we supposed that the first exon and its single intron could harbor regulatory sequences.

Based on this information, we constructed two transgenes by cloning the 2,104- and 971-bp genomic regions upstream of the translation start codon of *CXIP4* into the pMDC164 vector, including 1,479 ($CXIP4_{proI}$) or 346 ($CXIP4_{proII}$) bp from the TSS of *CXIP4* and that in the $CXIP4_{proI}$ included the entire AT2G28920 gene. The sizes of these regions were also conditioned by the presence of several stretches of repetitive nucleotides, which were avoided in the primer design. These promoters were fused to the β-glucuronidase (*GUS*) gene to drive its expression.

In the $CXIP4_{proI}$:*GUS* and $CXIP4_{proII}$:*GUS* transgenic plants, the highest GUS activity was detected in the emerging leaves and the vasculature of the cotyledons, leaves, and roots, especially in the root apex, but it was much lower in flowers (Supplementary Fig. S5, A to D), which is consistent with results previously described using an RNA gel blot analysis of RNA from the leaves, roots, stems, and flowers (Cheng et al. 2004). We also detected GUS activity during embryogenesis in seeds at the green mature embryo stage, which is consistent with the early lethality found in *cxip4-1* (Supplementary Fig. S5E). We did not observe any differences between plants expressing the $CXIP4_{proI}$:*GUS* or $CXIP4_{proII}$:*GUS* transgenes, and the expression of the $CXIP4_{proII}$:*CXIP4*:*GFP* transgene rescued all traits of the mutant phenotypes of *cxip4-1* and *cxip4-2* plants (Fig. 1, L and M), suggesting that the 971-bp region upstream of the translation start codon of *CXIP4* contains all the regulatory elements needed for *CXIP4* expression.

Phylogenetic analysis of CXIP4 and other PTHR31437 family members

CXIP4 appears to be plant-specific, although it is classified as a member of the SREK1IP1 protein family (PTHR31437; Supplementary Table S2) according to the PANTHER database (https://www.pantherdb.org; Thomas et al. 2022). Human SREK1IP1 and Arabidopsis CXIP4 share only a 21-aa region, which harbors the 18-aa ZCCHC motif (Supplementary Fig. S6). It is of note that the ZCCHC motifs of CXIP4 and SREK1IP1 clearly grouped in a branch of the unrooted tree that we obtained from the alignment of 198 ZCCHCs from 110 proteins of yeast, human, and Arabidopsis (Aceituno-Valenzuela et al. 2020). This grouping reflects the remarkably high similarity (13 identical aa) of the ZCCHC motifs of these two proteins (Supplementary Fig. S6).

The PTHR31437 family, as described in the PANTHER database, comprises 96 members from the animal (SREK1IP1-like), plant (CXIP4-like), protozoa, and chromista kingdoms, with proteins of this family expected to be present in all organisms of these four kingdoms. While all animal and most plant genomes studied to date encode only one SREK1IP1/CXIP4 protein (such as Arabidopsis), some plants contain two to four co-orthologs. This is the case in wheat, whose hexaploid genome encodes four CXIP4 proteins, three of which are almost identical and originate from its three subgenomes (Supplementary Table S2).

We studied the amino acid sequences of the 96 proteins classified as members of the PTHR31437 family in the UniProtKB database. One of the two proteins of the ancestral angiosperm *Amborella trichopoda* and those from the chromistan *Phytophthora ramorum* and *Thalassiosira pseudonana* lacked the ZCCHC motif (Supplementary Table S2). We aligned the sequences of five ZCCHC-containing proteins from five distant organisms: *Chlamydomonas reinhardtii* (UniProtKB accession number A0A2K3DE08), *Arabidopsis thaliana* (Q84Y18), *Physcomitrium* (*Physcomitrella*) *patens* (A0A2K1K6W5), *Drosophila melanogaster* (Q9W3Z5), and *Homo sapiens* (Q8N9Q2). The alignment revealed high conservation only in the region containing the ZCCHC motif and in three isolated basic residues of K or R, located downstream of the ZCCHC motif and at the C-terminus of the proteins (Supplementary Fig. S6). We found five additional conserved amino acids, in addition to the three canonical C and one H of the ZCXCHC motif, which may be relevant for the functions of PTHR31437 family proteins. The alignment also showed that the *Drosophila melanogaster* and human proteins lack the N-terminal extension found in the CXIP4-like proteins of photosynthetic organisms. In addition, the ZCCHC motifs in *Drosophila melanogaster* and humans are located at the N-terminal regions of these proteins (Supplementary Fig. S6), as also detected in all metazoan SREK1IP1 orthologs, whose lengths were half those of plant CXIP4-like proteins (Supplementary Table S2).

We also aligned the sequences of putative CXIP4-like proteins from different angiosperm lineages, selecting species encoding a single ortholog: *Hordeum vulgare*, *Arabidopsis thaliana*, *Capsicum annuum*, *Vitis vinifera*, and *Populus trichocarpa*. This second alignment revealed higher conservation than that shown in Supplementary Fig. S6, as expected, including a 48-aa region containing the ZCCHC motif (Supplementary Fig. S7) and the full conservation of the N-terminal 39 aa, which are absent in animals and partially conserved in the single-celled green alga *Chlamydomonas reinhardtii* (Supplementary Figs. S6 and S7). In fact, when we performed a BLASTP search using that 39-aa conserved sequence as a query, we only found the exact sequence in CXIP4 proteins from divergent land plant lineages of five major clades: bryophytes (such as *Physcomitrium patens*), lycophytes, monilophytes, gymnosperms, and angiosperms (such as Arabidopsis).

CXIP4 is a nucleoplasmic protein

When transiently expressed in tobacco (*Nicotiana tabacum*) plants, Arabidopsis CXIP4 localized to the nucleus in a diffuse pattern; however, when expressed in yeast and tobacco BY-2 cells, it was also found in the cytoplasm, forming discrete spots that did not correspond to mitochondria (Cheng et al. 2004). More recently, Arabidopsis CXIP4 was found to exclusively localize to the nuclei of chickpea (*Cicer arietinum*) protoplasts (Cheng and Nakata 2020).

As an initial assessment of the localization of Arabidopsis CXIP4, we used the MULocDeep web server (https://www.mu-loc.org/; Jiang et al. 2021, 2023), which predicts the localization of any eukaryotic protein in 44 suborganellar compartments based on its primary sequence. We included human SREK1IP1 in the analysis. MULocDeep predicted that both proteins localize

to the nucleus, with scores of 0.840 and 0.996, respectively (Supplementary Fig. S8A). Within the nucleus, both proteins are predicted to predominantly localize to the nucleoplasm, nucleolus, and nuclear speckles (Supplementary Fig. S8B), which are known to be enriched in pre-mRNA splicing factors, snRNAs, and polyadenylated [poly(A)$^+$] RNAs (Belmont 2022).

Using the LOCALIZER (https://localizer.csiro.au/; Sperschneider et al. 2017) and NoD (https://www.compbio.dundee.ac.uk/www-nod/index.jsp; Scott et al. 2010, 2011) web tools, we predicted nuclear and nucleolar localization signals (NLSs and NoLSs, respectively) in the C-terminal half of CXIP4. These NLSs and NoLSs signals were identified in two or three separate stretches, respectively, each composed of positively charged aa, mainly K and R (Supplementary Fig. S1).

To determine the localization of Arabidopsis CXIP4, we analyzed Col-0 plants carrying the *CXIP4*$_{pro}$*:CXIP4:GFP* transgene, which successfully rescued the mutant phenotypes of *cxip4-1* and *cxip4-2* plants (Fig. 1, L and M), revealing that the CXIP4-GFP fusion protein is functionally equivalent to the endogenous CXIP4. The CXIP4-GFP fusion protein appeared in a speckled pattern within the nucleus (Fig. 4), which may correspond to nuclear speckles, as predicted by MULocDeep (Supplementary Fig. S8B).

Very recently, Arabidopsis CXIP4 and its orthologs in wheat, tomato, and rice have been detected in multiple nuclear foci during a transient expression assay in wheat protoplasts (He et al. 2024). These foci appear very similar to those found in *CXIP4*$_{pro}$*:CXIP4:GFP* plants in a Col-0 background.

cxip4-2 globally affects the expression of genes involved in pathogen defense

To gain insight into the biological processes affected by the loss of function of *CXIP4*, we performed RNA-seq analysis using poly(A)$^+$ RNAs from Col-0 and *cxip4-2* plants collected 14 das. We identified 2,160 deregulated genes in *cxip4-2* compared to the wild type, with a similar number of up- (1,094) and downregulated (1,066) genes (Fig. 5A; Supplementary Data Set 1). *CXIP4* was the tenth most downregulated gene, suggesting that the T-DNA insertion disrupting *CXIP4* in the *cxip4-2* mutant strongly reduces its transcript levels (Fig. 5, A and B; Supplementary Data Set 1), as observed in the abovementioned RT-PCR analysis (Supplementary Fig. S4B). However, compared with the early lethality exhibited by *cxip4-1*, the low expression of *CXIP4* in *cxip4-2* suggests that it is sufficient for viability, but strongly compromises fertility, in addition to causing a highly pleiotropic developmental phenotype (Figs. 1, B to J, 2 and 3; Supplementary Fig. S2).

Gene Ontology (GO) enrichment analysis of the deregulated genes in *cxip4-2* plants revealed an overrepresentation of terms primarily associated with pathogen defense (Fig. 5C). In a separate analysis of downregulated genes, we also found an overrepresentation of terms related to indole glucosinolate, camalexin, and phytoalexin metabolism, as well as salicylic acid synthesis and perception, among others. By contrast, upregulated genes were mainly associated with responses to sulfur starvation and other stress conditions, as well as flavonoid, glucosinolate, and jasmonic acid metabolism (Supplementary Data Set 1).

Taken together, these results suggest that the deregulation of *CXIP4* causes a global defense response to biotic and abiotic stress. Interestingly, the upregulated genes were also enriched in DNA-binding transcription factor activity, including 80 out of 1,208 genes in the reference database, mainly with the C_2H_2 zinc finger (eight genes) and homeobox (32 genes) DNA-binding domains (Supplementary Data Set 1).

Several upregulated genes in *cxip4-2* may explain its mutant phenotype

The *cxip4-2* mutant is viable but exhibits a highly pleiotropic phenotype, suggesting that many developmental pathways or key genes are deregulated in the mutant. Therefore, we inspected the RNA-seq data, looking for specific deregulated genes that may explain some of the phenotypes exhibited by *cxip4-2* plants. Interestingly, *GLABROUS1* (*GL1*), encoding a positive regulator of trichome development (Larkin et al. 1994), was downregulated

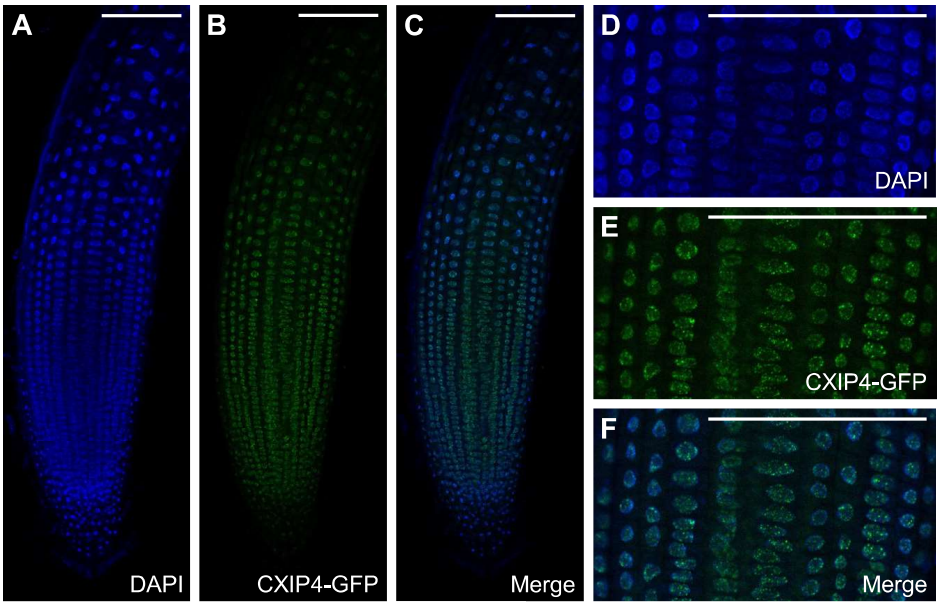


Figure 4. Subcellular localization of CXIP4 in Arabidopsis roots collected five das. **A–F)** Confocal laser-scanning micrographs of Col-0 plants homozygous for the *CXIP4*$_{pro}$*:CXIP4:GFP* transgene. Fluorescent signals correspond to (A, D) 4′,6-diamidino-2-phenylindole (DAPI) nuclear staining (in blue), (B, E) CXIP4-GFP (in green), and (C, F) their overlap. Scale bars: 100 μm.

(Supplementary Data Set 1), which could explain the low density of trichomes found in *cxip4-2* leaves, whereas trichome branching, which is not controlled by this gene, remained unaffected (Fig. 3, A to C).

Among the upregulated genes in *cxip4-2*, we also identified *ENHANCED RNA INTERFERENCE-1-LIKE-1* (*ERIL1*; Supplementary Data Set 1), which encodes a ribonuclease H-like protein involved in the processing of chloroplast pre-rRNAs (Mermigka et al. 2016). Transgenic plants overexpressing *ERIL1* are defective in chloroplast rRNA maturation, resulting in the accumulation of several rRNA biogenesis intermediates. These precursors include a 1.7-kb transcript that may correspond to the 16S rRNA containing the intergenic region that separates this gene from the *trnI* gene, as well as the 23S-4.5S dicistronic precursor (3.2 kb), and the 2.4 and 2.9 kb incompletely processed 23S rRNAs (Mermigka et al. 2016; Supplementary Fig. S9A). The upregulation of *ERIL1* in *cxip4-2* could explain the presence of two additional peaks in the profiles of total RNA samples, which we obtained before their use in the RNA-seq assay. Both peaks are located between those corresponding to the cytosolic 18S and 25S rRNAs, whose sizes in Arabidopsis are approximately 1.8 and 3.4 kb, respectively, and could correspond to the incompletely processed chloroplast 23S rRNAs mentioned above (Supplementary Fig. S9).

PRODUCTION OF ANTHOCYANIN PIGMENT 1 (*PAP1*) was also found among the upregulated genes in *cxip4-2* (Supplementary Data Set 1). PAP1 is a transcription factor that directly induces the transcription of genes involved in anthocyanin biosynthesis. *PAP1*-overexpressing plants accumulate high amounts of anthocyanins when grown in the presence of high sucrose concentrations, which cause stress (Teng et al. 2005). The upregulation of *PAP1*, along with the enriched GO terms "anthocyanin-containing compound biosynthetic process" and "flavonoid metabolic process" (Supplementary Data Set 1), may also account for the high anthocyanin levels observed in the *cxip4-2* mutant (Fig. 3, D to J), suggesting that *cxip4-2* plants are constitutively stressed.

We validate by RT-qPCR the deregulation of *GL1*, *ERIL1*, *PAP1*, and *CXIP4*, using cDNA from Col-0, *cxip4-2*, and *cxip4-2 CXIP4*$_{pro}$:*CXIP4:GFP* plants (Supplementary Fig. S10). We included *LIPOXYGENASE 2* (*LOX2*) in the validation as one of the most upregulated genes in *cxip4-2* plants belonging to one of the most enriched GO terms: "jasmonic acid metabolic process" (Supplementary Data Set 1 and Fig. S10).

We did not find enriched GO terms related to the UPR, as might have been anticipated given the resistance to DTT and TM observed in plants overexpressing *CXIP4* (Hossain et al. 2016). Nevertheless, among the downregulated genes in *cxip4-2* plants was AT1G42990, which encodes the BASIC REGION/LEUCINE ZIPPER MOTIF 60 (bZIP60) transcription factor. bZIP60 activates the transcription of key genes involved in UPR signaling (Iwata and Koizumi 2005).

Pre-mRNA splicing is defective in *cxip4-2* plants

As previously mentioned, the GO annotations of human SREK1IP1 and its interactions associate this protein with pre-mRNA splicing. Thus, we analyzed the effects of *cxip4-2* on pre-mRNA splicing, finding 939 differential AS events in *cxip4-2* plants compared with Col-0. The most frequent were intron retentions (IR), with 543 events (57.8% of the total differential AS events), mainly resulting in the retention of the alternatively spliced intron (Fig. 6A; Supplementary Data Set 2).

The second most frequent differential AS events identified in *cxip4-2* plants compared with Col-0 were alternative 3′ splice site (A3′SS) events, totaling 255 (27.2%). Using Integrative Genomics Viewer (IGV; Robinson et al. 2011), we determined that 85 out of these 255 events (33.3% of the differential A3′SS events) involved tandem 3′SSs (acceptor sites) with NAGNAG sequences (Supplementary Data Set 2), in which A3′SSs were exactly three nucleotides apart from each other (in-frame); their alternative usage would result in proteins differing by one aa, which might not affect their function. The tandem 3′SS motif NAGNAG is overrepresented in genes encoding RNA-recognition motifs (RRMs), including SR proteins (Schindler et al. 2008). However, we did not find enrichment in any protein class for these 85 A3′SS events, which included genes related to pre-mRNA splicing.

Subsets of misspliced genes in *cxip4-2* are related to different steps of mRNA metabolism

We analyzed the overrepresented GO terms for misspliced genes in *cxip4-2*, finding that they were mainly related to the regulation of cuticular wax biosynthesis, RNA-mediated gene silencing, mRNA splicing, chromatin remodeling, and DNA repair, among others (Fig. 6B; Supplementary Data Set 2). Five out of 15 genes in the reference database were categorized under the "ta-siRNA processing" term: *SUPPRESSOR OF GENE SILENCING 1* (*SGS1*; also known as *ANAC052*), *SGS2* (also known as *RNA DEPENDENT RNA POLYMERASE 6* [*RDR6*]), and *SGS3*, with differential IR events; and *DICER-LIKE 3* (*DCL3*) and *DOUBLE-STRANDED-RNA-BINDING PROTEIN 4* (*DRB4*), with differential A3′SS events. These five genes, whose expression was not deregulated in *cxip4-2* plants, were also classified under the term "siRNA processing", which also included the nonderegulated genes *NUCLEAR RNA POLYMERASE D1B* (*NRPD1B*) and *TOUGH* (*TGH*), exhibiting differential IR and A3′SS events, respectively (Supplementary Data Set 2). Other overrepresented GO terms with several overlapping genes included "post-transcriptional gene silencing", "regulatory ncRNA processing", "RNA-mediated gene silencing", and "negative regulation of gene expression, epigenetic" (Supplementary Data Set 2).

We also analyzed the PANTHER protein classes that were overrepresented among the misspliced genes in *cxip4-2*, which included the "mRNA polyadenylation factor" class, with six out of 39 genes in the reference database. These six genes encode NOT1 (the scaffold protein of the CCR4-NOT complex), the deadenylases CATABOLITE REPRESSOR 4C (CCR4C) and CCR4D, POLY(A) POLYMERASE 2 (PAPS2) and PAPS4 (also known as *NPAP*), and MATERNAL EFFECT EMBRYO ARREST 44 (MEE44). The differential AS events included IR events in *CCR4D*, *PAPS2*, *PAPS4*, and *MEE44*, and A3′SS events in *NOT1* and *CCR4C*, although these genes were not deregulated in *cxip4-2*. These proteins are thought to function in polyadenylation-assisted RNA degradation through the exonucleolytic cleavage of the poly(A) tail, thereby contributing to mRNA export, nuclear quality control, and translation (Passmore and Coller 2022).

We aimed to validate some of these events by RT-PCR. We ruled out validating A5′SS and A3′SS events because, in most cases, the length difference when using the canonical SS was only a few nucleotides. Additionally, we excluded genes that exhibited combinations of different AS events at the same coordinates, as we assumed they would be difficult to analyze. Consequently, we selected five genes from different functional categories, each of which had only one single significant intron retention event detected. These genes were *SGS1* (*ANAC052*) with 30.5% of IR relative to Col-0; *SGS3*, with 21.6%; *NRPD1B*, with 47.93%, representing the second highest percentage of IR from a gene with a single AS event detected; *CCR4D*, with 10.07%; and *MEE44*, with 21.77% of IR. We

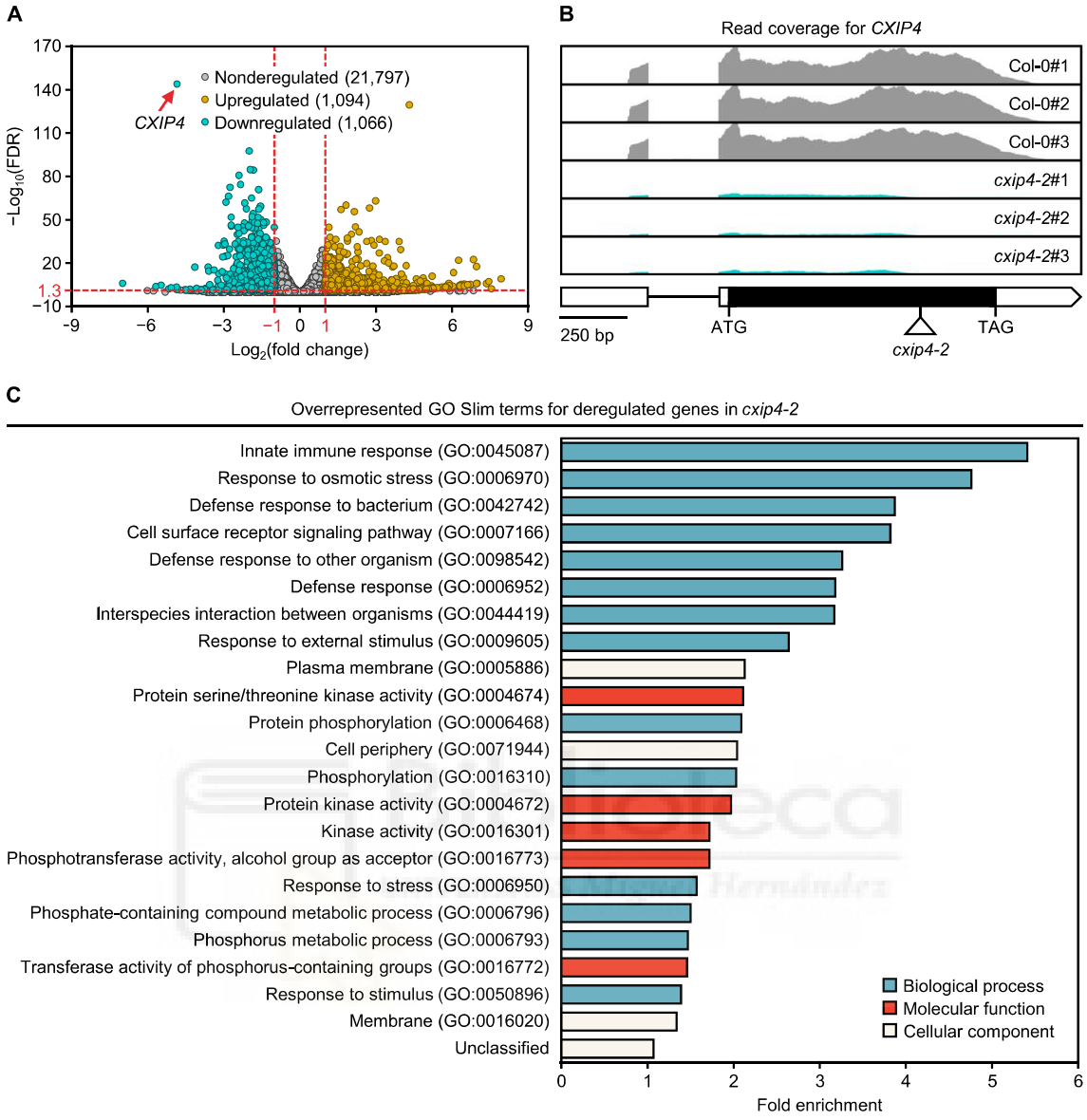


Figure 5. Differential gene expression analysis between Col-0 and *cxip4-2* plants. **A)** Volcano plot of differentially expressed genes in *cxip4-2* compared to Col-0 plants, collected 14 das. Upregulated genes (fold change ≥ 2) and downregulated genes (fold change ≤ 0.5) and with FDR < 0.05 (Benjamini–Hochberg false discovery rate) are shown as yellow and cyan dots, respectively; nonderegulated genes appear as gray dots. Horizontal and vertical red dashed lines mark the cutoffs of the negative decimal logarithm of FDR and binary logarithm of fold changes. The dot marked with a red arrow represents *CXIP4*, which was downregulated in *cxip4-2*, as expected. **B)** Plot of aligned reads for *CXIP4* in Col-0 and *cxip4-2*, obtained with Integrative Genomics Viewer (IGV) software (http://software.broadinstitute.org/software/igv/). Gene structure is represented as described in Fig. 1 legend. **C)** Bar plot representation of the overrepresented Gene Ontology (GO) Slim terms for deregulated genes in *cxip4-2*. These terms were identified through statistical overrepresentation tests using PANTHER (https://pantherdb.org/) GO Slim annotation sets for biological processes, molecular functions, and cellular components (fold enrichment > 1 and FDR < 0.05). Redundant terms, with an identical set of genes to others that were already included, are not plotted. For a separate analysis of upregulated and downregulated genes in *cxip4-2*, see Supplementary Data Set 1.

also included AT2G01100 in the validation, with 47.6% of IR relative to Col-0 (Fig. 7, A to G; Supplementary Data Set 2).

We initially used cDNAs from three biological replicates of Col-0, *cxip4-2*, and *cxip4-2 CXIP4*$_{pro}$*:CXIP4:GFP*. However, changes in the relative mRNA isoform proportions compared to the wild type were not evident for some of these genes. Subsequently, we repeated the RT-PCR experiments, including a sample of cDNA from *cxip4-1* escaper plants, which we isolated from several sowings, as we had observed that the number of these plants decreased in successive generations of *CXIP4/cxip4-1* heterozygous individuals. The results obtained with *cxip4-1* cDNA were highly noticeable (Fig. 7, A to G), confirming the involvement of *CXIP4* in pre-mRNA splicing and highlighting the different functional nature of the *cxip4* alleles, with *cxip4-2* being hypomorphic and *cxip4-1* likely a null allele.

cxip4-2 plants accumulate poly(A)+ RNAs

Eukaryotic cells prevent the translation of pre-mRNAs by different mechanisms. For example, pre-mRNAs can be retained in

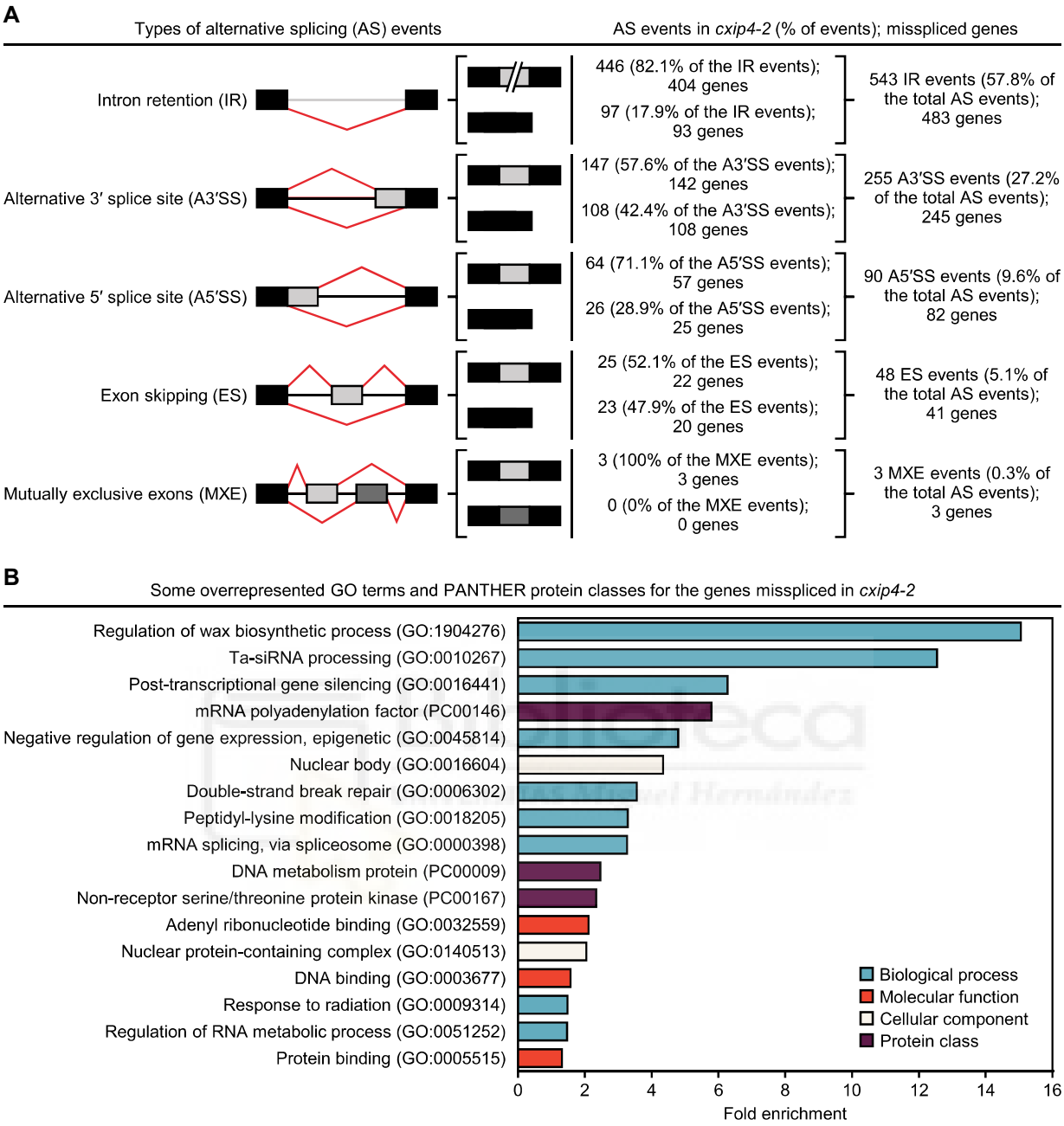


Figure 6. Analysis of differential AS between Col-0 and *cxip4-2* plants. **A)** Summary of differential AS events and missplicedgenes identified in *cxip4-2* compared to Col-0 18 das. Schematic representations depict five different types of AS events, with black or gray rectangles and lines representing constitutive or alternatively spliced exons and introns, respectively, and red lines marking the splice junctions. **B)** Bar plot of overrepresented GO terms and PANTHER protein classes for the missplicedgenes in *cxip4-2*. These terms were identified as described in Fig. 5 legend, but using GO Complete and PANTHER protein class annotation sets. Only the most specific hierarchically organized GO terms and PANTHER protein classes were plotted; for a complete list, see Supplementary Data Set 2.

the nucleus until they are fully or correctly spliced (Palazzo and Lee 2018; Wegener and Müller-McNicoll 2018; Rudzka et al. 2022). This mechanism also ensures efficient pre-mRNA splicing by factors that are confined to the nucleus. In addition, mRNA export factors are preferentially recruited to spliced mRNAs (reviewed in Reed 2003; Kelly and Corbett 2009), and pre-mRNAs or missplicedRNAs that still leak into the cytoplasm are degraded by nonsense-mediated mRNA decay (NMD; reviewed in Bhuvanagiri et al. 2010).

The *prp8-7* hypomorphic allele of *PRE-MRNA PROCESSING 8* (*PRP8*), encoding a central factor of the spliceosome, causes global missplicing, and we previously detected the nuclear accumulation of poly(A)+ RNAs in *prp8-7* plants (Cabezas-Fuster et al. 2022). We aimed to determine whether

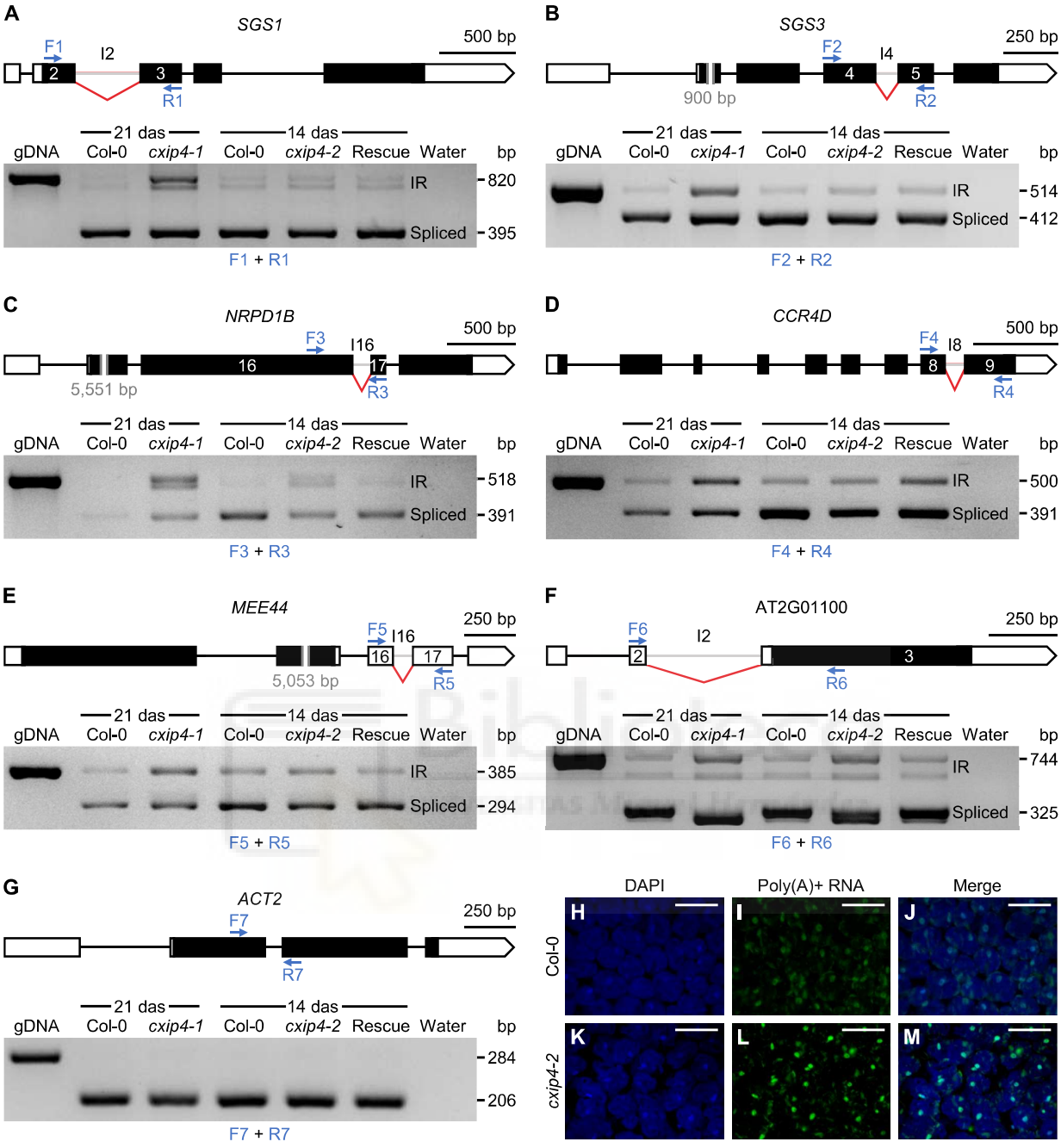


Figure 7. Validation of differential intron retention (IR) events in the *cxip4* mutants and accumulation of poly(A)+ RNAs in the nuclei of *cxip4-2* leaf cells. **A–G)** RT-PCR analysis of six differentially spliced introns in the Col-0, *cxip4-1*, *cxip4-2*, and *cxip4-2 CXIP4*$_{pro}$*:CXIP4:GFP* (rescue) plants. Schematic representations illustrate the structures of the (A) *SGS1*, (B) *SGS3*, (C) *NRPD1B*, (D) *CCR4D*, (E) *MEE44*, (F) AT2G01100, and (G) *ACT2* genes. Numbers indicate the positions in the pre-mRNA of the flanking exons corresponding to the analyzed introns (shown in gray) affected by the alternative IR events, with red lines marking the splice junctions. Arrows represent the primers used (not drawn to scale) and are labeled as follows: F1 (corresponding to SGS1-F in Supplementary Table S1), R1 (SGS1-R), F2 (SGS3-F), R2 (SGS3-R), F3 (NRPD1B-F), R3 (NRPD1B-R), F4 (CCR4D-F), R4 (CCR4D-R), F5 (MEE44-F), R5 (MEE44-R), F6 (AT2G01100-F), R6 (AT2G01100-R), F7 (ACT2-F), and R7 (ACT2-R). Genomic DNA (gDNA) from Col-0 and water were used as positive and negative controls for the PCR amplifications, respectively, while the expression of the housekeeping gene *ACT2* acted as an internal control to normalize the cDNA quantity used. **H–M)** Poly(A)+ RNA-FISH assays in palisade mesophyll cells of the first-node leaves of (H–J) Col-0 and (K–M) *cxip4-2*. Fluorescent signals correspond to (H, K) DAPI nuclear staining, (I, L) fluorescein from an oligo-dT probe targeting poly(A)+ RNAs, and (J, M) their merged images. Confocal laser-scanning micrographs were taken from five leaves per genotype of plants collected 14 das for Col-0 and 22 das for *cxip4-2*. The gain of the 515/30 nm detector was kept constant to allow for direct comparisons of fluorescein intensity between samples. Scale bars: 50 μm.

cxip4-2 plants accumulate mRNAs in the nucleus and we performed poly(A)+ RNA fluorescence in situ hybridization (RNA-FISH) assays using a fluorescein-labeled oligo(dT) probe, detecting poly(A)+ RNAs accumulated to high levels within the nuclei of *cxip4-2* leaf palisade mesophyll cells (Fig. 7, H to M). Accumulation of poly(A)+ RNAs in the nucleus of *cxip4-2* can be due to the presence of misspliced mRNAs that are retained.

Discussion

CXIP4 is an essential gene whose partial loss of function triggers pronounced pleiotropic changes

No mutational analyses of *CXIP4-like* genes have been described, and only a single functional analysis of the *CXIP4* gene in Arabidopsis has been performed, which used transgenic lines with β-estradiol-inducible overexpression of full-length cDNAs (Coego et al. 2014). Subsequent analysis of these lines revealed the resistance of *CXIP4* overexpressing plants to DTT and TM, suggesting a role for CXIP4 in the UPR (Hossain et al. 2016). However, gene overexpression can cause artifacts which do not occur in loss-of-function analysis. Moreover, loss-of-function analysis is usually more informative in the absence of functional redundancy. This would be the case for *CXIP4*, since it is a single-copy gene in Arabidopsis (Supplementary Table S2).

The functional analysis of two T-DNA insertional lines of *CXIP4* harboring alleles with different degrees of loss of function allowed us to conclude that *CXIP4* is an essential gene for embryogenesis but is also required for postembryonic development (Figs. 1 to 3; Supplementary Fig. S2). *cxip4-1* is likely a null allele of *CXIP4* that mainly causes early embryonic lethality, and the few embryos that escape this lethality produce callus-like plants with arrested development (Fig. 1, A, B, C, E, F, H, I, and L; Fig. 2). The *cxip4-2* allele, characterized by significantly reduced *CXIP4* expression, as identified by RNA-seq and RT-qPCR analyses (Fig. 5, A and B; Supplementary Fig. S10), is fully viable but causes a striking pleiotropic phenotype when homozygous. This phenotype includes severe growth retardation leading to delayed flowering, abnormal cotyledon and leaf morphology characterized by sparse trichome density, the loss of apical dominance, and markedly reduced fertility (Fig. 1, A, B, D, E, G, H, J, and M; Fig. 3; Supplementary Fig. S2). Our analysis revealed the deregulation of several key genes, predominantly encoding transcription factors, in *cxip4-2* plants, providing insights into the underlying mechanisms of its pleiotropic phenotype (Supplementary Data Set 1 and Fig. S10). Furthermore, our findings suggest that the C-terminal region of CXIP4, which is absent from the CXIP4-2 mutant protein, is dispensable for plant survival but crucial for proper development (Supplementary Fig. S1). The pleiotropy exhibited by *cxip4-2* and the embryonic lethality of *cxip4-1* match the broad GUS activity detected in almost all tissues, including embryos, where expression was very high (Supplementary Fig. S5).

Besides their aberrant morphological phenotype, *cxip4-2* plants accumulate high levels of anthocyanins, suggesting that they are constitutively stressed. This is consistent with the finding that the deregulated genes identified by RNA-seq analysis were enriched in GO terms involved in global defense responses (Fig. 3, D to J; Fig. 5C), and with the elevated overexpression of *PAP1* (Supplementary Data Set 1 and Fig. S10), encoding a well-known transcription factor that regulates anthocyanin biosynthesis (Teng et al. 2005). These results also agree with the recently proposed function of TaCAXIP4, one of the four putative co-orthologs of CXIP4 in wheat (Supplementary Table S2), in the calcium-mediated plant immune response against pathogens. This role of TaCAXIP4 results from its interaction with TaHRC, a histidine-rich calcium-binding nuclear protein conferring susceptibility to Fusarium head blight or scab, a destructive disease in wheat worldwide caused by *Fusarium graminearum* (Su et al. 2019). It has been proposed that the interaction between TaHRC and TaCAXIP4 leads to the hijacking of TaCAXIP4, resulting in the suppression of calcium-mediated immune responses (Chen et al. 2022).

Arabidopsis CXIP4 and human SREK1IP1 are highly divergent orthologs

The PANTHER database contains the complete proteomes of 143 organisms, including model organisms and others used in biomedical and biotechnological research. A search in PANTHER revealed that metazoan SREK1IP1-like and plant CXIP4-like proteins form the PTHR31437 family, but their sequences are highly divergent. In fact, they are classified as LDOs (least diverged orthologs), which are the most nearly "equivalent" gene pairs between different organisms based on phylogenetic analysis. Metazoan SREK1IP1-like proteins are approximately half the length of plant CXIP4-like proteins, sharing only some aa from the ZCCHC motif (Supplementary Table S2 and Fig. S6). However, it is of note that the ZCCHC motifs of Arabidopsis CXIP4 and human SREK1IP1, which share 13 to the 18 aa, clearly grouped in the same branch of an unrooted tree generated by phylogenetic analysis based on sequence alignment of 198 ZCCHC motifs from 110 ZCCHC-containing proteins in yeast (7), Arabidopsis (69), and humans (34) that we identified in a systematic search (Aceituno-Valenzuela et al. 2020). This grouping was likely due to the high number of conserved amino acids, in addition to the three canonical cysteines (C) and one histidine (H) that define the ZCCHC motif.

Despite the low similarity between Arabidopsis CXIP4 and human SREK1IP1 outside the ZCCHC domain, their subcellular and suborganellar localizations appear to be highly similar, as predicted by MULocDeep (Supplementary Fig. S8). MULocDeep uses a machine learning method based on primary sequence information (https://www.mu-loc.org/; Jiang et al. 2021, 2023) to identify common signals or patterns in two proteins. In fact, the NLS and NoLS signals were present in the C-terminal half of CXIP4, which is rich in basic aa, similar to human SREK1IP1. These findings support the notion that these proteins share similar functions.

The specific categories of misspliced genes in *cxip4-2*, rather than their number, may explain its mutant phenotype

The *prp8-7* hypomorphic allele of *PRP8*, encoding a central factor of the spliceosome, causes global missplicing, as revealed by its 8,124 increased IR events compared to Col-0 (Sasaki et al. 2015), which is 18 times higher than in the *cxip4-2* plants (446 increased IR events; Fig. 6A; Supplementary Data Set 2). However, the morphological phenotype of *cxip4-2* is much more severe than that of *prp8-7* with both mutants exhibiting a partial loss of function of the corresponding essential genes (*CXIP4* and *PRP8*). Therefore, the phenotype of *cxip4-2* plants is not caused by the number of missplicing events found in the RNA-seq analysis, which suggested a minor role for CXIP4 in general splicing. However, RT-PCR validation of six misspliced genes showing IR events, including cDNA from *cxip4-1* escaper plants, revealed a more prominent role of CXIP4 in pre-mRNA splicing, as the ratio of mRNAs with retained introns to fully spliced mRNAs was much higher in *cxip4-1* compared with *cxip4-2* plants (Fig. 7, A to G).

We also found that most of the detected differential AS events do not deregulate the expression of the affected genes, as is the case with the *prp8a-14* mutant of the *PRP8* gene (Llinas et al. 2022). It is possible that the missplicing events that we found in *cxip4-2* plants cause a reduction of functional mRNAs and alter the relative concentrations of the resulting protein isoforms, which may have diverse functional consequences. However, the overrepresentation of specific GO terms, mainly "gene silencing by DNA methylation" or the protein class "mRNA

polyadenylation factor" among the misspliced genes in *cxip4-2* plants (Supplementary Data Set 2) also suggests a role for CXIP4 in transcriptional and post-transcriptional regulation of gene expression, which should be studied further.

CXIP4 could play a key role in the modulation of the AS

In GeneCards in the Human Gene Database (https://www.genecards.org/), SREK1IP1 is described as having GO annotations related to "Possible splicing regulator involved in the control of cellular survival". We found relevant information on human SREK1IP1 in the HitPredict database (http://www.hitpredict.org), which compiles physical protein–protein interactions from six different databases and 124 species, all experimentally identified in both small-scale and high-throughput assays (López et al. 2015). An interaction with NKAP obtained through an affinity purification mass spectrometry (AP-MS) high-throughput assay (Huttlin et al. 2017) is described in HitPredict, along with several interactions with splicing factors (Supplementary Table S3). These results confirm the conservation of interactions between SREK1IP1/CXIP4 and NKAP/MAS2 in humans and Arabidopsis, supporting a role for human SREK1IP1 in pre-mRNA splicing.

Additionally, TaCAXIP4, one of the four CXIP4-like co-orthologs in wheat, has been proposed to be part of a nuclear speckled complex with FHB, based on Y2H assays using FHB as bait. This complex is thought to consist of these two proteins along with others whose orthologs participate in AS regulation, including RNA-BINDING PROTEIN LUC7-LIKE and two SR-rich proteins (He et al. 2024). The Arabidopsis CXIP4 could also play a key role in the formation of nuclear complexes that regulate AS and/or mRNA export. We observed that most of the differentially spliced introns identified in our RNA-seq analysis also show some retention in Col-0, with higher levels in the *cxip4-2* mutant (Supplementary Data Set 2), and even more pronounced retention in *cxip4-1*. However, we can only confirm the latter for the six IR events analyzed by RT-PCR (Fig. 7, A to G). These results suggest that the 5′SSs are weak in these misspliced introns, and that CXIP4 facilitates their recognition.

This hypothesis aligns with the proposed composition of the complex in wheat, based on interactions with HRC, as previously mentioned, where CXIP4 was identified alongside LUC7 (He et al. 2024). In Arabidopsis, LUC7 is encoded by three genes, whose products are involved in both general and AS. Similar to *cxip4-2* plants, RNA-seq analysis of the Arabidopsis *luc7* triple mutant revealed only 640 differential splicing events compared with the wild type, with only 17 of these events (2.7%) affecting transcript levels. Additionally, in the *luc7* triple mutant, the export of mRNAs with retained introns to the cytoplasm is blocked, suggesting that these transcripts evade the NMD mechanism, which normally operates in the cytoplasm (de Francisco Amorim et al. 2018). The nuclear retention of misspliced mRNAs in the *luc7* triple mutant is consistent with our RNA-FISH assay (Fig. 7, H to M), where we found an accumulation of poly(A)+ RNAs in the nuclei of *cxip4-2* leaf cells, likely corresponding to misspliced mRNAs. Furthermore, in yeast, Luc7p is involved in 5′SS recognition, and in *luc7* loss-of-function mutants, splicing of introns with nonconsensus 5′SS or branchpoint sequences is more defective than in the wild type (Fortes et al. 1999). Since SR proteins act as positive regulators of AS at weak splice sites, CXIP4 may collaborate with them in splice site recognition.

Our conclusions regarding the function of Arabidopsis *CXIP4* are based on the developmental and molecular phenotypes associated with its *cxip4-1* and *cxip4-2* loss-of-function alleles. However, the predicted CXIP4.1 and CXIP4.2 mutant proteins retain the ZCCHC motif, which is the sole motif identified in CXIP4-like and SREK1IP1-like proteins and represents their primary conserved feature. This motif is commonly found in proteins involved in mRNA and noncoding RNA metabolism, mediating interactions with other proteins and RNAs (reviewed in Aceituno-Valenzuela et al. 2020). Furthermore, the ZCCHC motif is the fourth most common RNA-binding motif in human proteins, highlighting its functional importance, and supporting its classification as an RNA-binding motif (Ray et al. 2023). Generating Arabidopsis CXIP4 mutants specifically lacking the ZCCHC motif, or introducing similar mutations in other CXIP4-like or SREK1IP1-like proteins, would be invaluable for clarifying the functional significance of this motif and for understanding its evolutionary conservation across the plant and animal kingdoms.

Materials and methods

Plant material, growth conditions, and genotyping

All *Arabidopsis thaliana* (L) Heynh. lines used in this work were in the Columbia-0 (Col-0) genetic background. The Col-0 wild type and the T-DNA insertional lines GABI_537C02 (*cxip4-1*), SALK_044245 (*cxip4-2*), SALK_097293, and SALK_017259 (Alonso et al. 2003; Kleinboelting et al. 2012) were obtained from the Nottingham Arabidopsis Stock Center (NASC; Nottingham, United Kingdom) and propagated in our laboratory for further analysis.

Seed sterilization and sowing, plant culture, and crosses were performed as previously described (Ponce et al. 1998; Berná et al. 1999), except that plant agar was replaced with 6 g·L^{-1} of Gelrite (Duchefa Biochemie). When required, culture media were supplemented with hygromycin (15 μg·mL^{-1}) and kanamycin (50 μg·mL^{-1}).

To genotype the T-DNA insertional lines, genomic DNA was extracted from the samples as described in Ponce et al. (2006), and the presence of the T-DNA insertions was verified by PCR amplification using the primers shown in Supplementary Table S1.

Most Sanger sequencing reactions and electrophoreses were carried out in our laboratory with ABI PRISM BigDye Terminator Cycle Sequencing kits and an ABI PRISM 3130xl Genetic Analyzer (Applied Biosystems). Some sequencing reactions were carried out at Stab Vida (Caparica, Portugal).

RT-PCR and RT-qPCR analysis

Total RNA was isolated with TRIzol Reagent (Invitrogen) from three biological replicates per genotype, each consisting of 100 (Col-0, *cxip4-2*, and *cxip4-2 CXIP4*$_{pro}$*:CXIP4:GFP*) or 20 (*cxip4-1*) mg of plants collected 14 or 21 das. Prior to cDNA synthesis, the RNA was treated with TURBO DNase (Invitrogen). RT-PCR amplifications were performed as described in Casanova-Sáez et al. (2014). The mRNAs produced by the *cxip4-1* and *cxip4-2* mutant alleles were analyzed by Sanger sequencing of cDNAs. qPCR amplifications were carried out in a Step-One Real-Time PCR System (Applied Biosystems) using three technical replicates per biological replicate. The primers used are shown in Supplementary Table S1. The housekeeping gene *ACTIN2* (*ACT2*) served as an internal control for normalization and relative quantification of gene expression. The C_T values obtained were normalized using the $2^{-\Delta\Delta C_T}$ method (Livak and Schmittgen 2001). Statistical analyses of the ΔC_T values between Col-0 and *cxip4-2*, and between

Col-0 and *cxip4-2* *CXIP4*$_{pro}$:*CXIP4:GFP*, were performed using unpaired Student's *t*-tests with the GraphPad *t*-test web calculator (https://www.graphpad.com/quickcalcs/ttest1/).

Construction of transgenes and analysis of transgenic lines

Transgenes were generated by Gateway cloning as described in Sánchez-García et al. (2015), using the pGEM-T Easy221 entry vector (provided by B. Scheres) and the pMDC107 and pMDC164 (Curtis and Grossniklaus 2003) destination vectors.

To assess the temporal and spatial expression patterns of *CXIP4*, two constructs, *CXIP4*$_{proI}$:*GUS* and *CXIP4*$_{proII}$:*GUS*, were generated by PCR amplification of the 2,104- and 971-bp genomic regions upstream of the translation start codon of *CXIP4*. The PCR products were subcloned into the pMDC164 destination vector. GUS enzymatic activity was observed throughout the development of Col-0 plants homozygous for the *CXIP4*$_{proI}$:*GUS* and *CXIP4*$_{proII}$:*GUS* transgenes. All parts of the plants were imaged using bright-field microscopy under a Nikon D-Eclipse C1 laser-scanning confocal microscope. GUS staining was performed as previously described (Donnelly et al. 1999). The samples were incubated in X-Gluc buffer overnight at 37 °C. The seed coat and embryo were separated before incubating the seeds in GUS staining solution.

To determine the subcellular localization of CXIP4, the *CXIP4*$_{proII}$:*CXIP4:GFP* construct was generated by PCR amplification of the 971-bp genomic region upstream of the translation start codon and the full-length coding sequence of *CXIP4* (without its stop codon), which was then subcloned into the pMDC107 destination vector.

The structural integrity of all constructs was verified by sequencing prior to their transfer into Arabidopsis plants by *Agrobacterium tumefaciens*-mediated transformation via the floral dip method (Clough and Bent 1998). Primers used to obtain these constructs and for Sanger sequencing are described in Supplementary Table S1.

Analysis of plant morphology

Rosette photographs were taken under a Nikon SMZ1500 stereomicroscope equipped with a Nikon DXM1200F or DS-Ri2 digital camera. For large rosettes, high-resolution images were obtained by taking multiple pictures of the same plant and assembling them with the Photomerge tool of Adobe Photoshop CS3 software. Adult plants were photographed with a Canon PowerShot SX200 IS camera.

Trichome density and branch number were determined using the third leaves of ten Col-0 and *cxip4-2* plants collected 21 and 29 das, respectively. The leaves were cleared with ethanol and chloral hydrate, mounted on slides, and photographed under a Leica DMRB microscope equipped with a Nikon DXM1200 digital camera (Nadi et al. 2023). The number of trichomes with two to five branches was scored for each leaf and expressed as a percentage of the total number of trichomes.

For the analysis of siliques and seeds, photographs were taken under a Nikon SMZ1500 stereomicroscope equipped with either a Nikon DXM1200F or DS-Ri2 digital camera. Silique length and number of seeds per silique were measured from ten siliques collected from five plants per genotype. Leaf and seed silhouettes (ten and 60 per genotype, respectively) were drawn on a Cintiq 18SX Interactive Pen Display (Wacom) using Adobe Photoshop CS3, and their areas were measured with the NIS Elements AR 3.1 image analysis software (Nikon).

Statistical analyses of the trichome density and branching, silique length, number of seeds per silique, and seeds size in Col-0 and *cxip4-2* were performed using unpaired Student's *t*-tests with the GraphPad *t*-test web calculator.

Anthocyanin extraction and measurement

To quantify anthocyanin contents, ten biological replicates per genotype were used, each consisting of a rosette collected 12 das and weighed. Anthocyanins were extracted and homogenized in 45% (v/v) methanol and 5% (v/v) acetic acid buffer using glass beads and a MixerMill 400 (Retsch) automatic mixer. The samples were centrifuged twice for 5 min at 13,500 × *g* to remove cellular debris. The absorbance of each clarified extract was measured spectrophotometrically at 530 and 657 nm for anthocyanin and contaminating chlorophyl in acidic extraction buffer, respectively (Mancinelli 1990). Anthocyanin content was calculated as $A_{530} - (0.25 \cdot A_{657}) \cdot g^{-1}$ of rosette fresh weight, where 25% of the A_{657} reading was subtracted to account for chlorophyl degradation products. A statistical analysis of the anthocyanin content in Col-0 and *cxip4-2* plants was performed using unpaired Student's *t*-test with the GraphPad *t*-test web calculator.

RNA-seq and analysis of AS

Total RNA was extracted using TRIzol Reagent (Invitrogen) from three biological replicates per genotype, each consisting of 40 to 80 mg of rosettes collected 14 das. At least 6 μg of RNA was sent to Novogene (Cambridge, United Kingdom) for sequencing and subsequent gene expression profiling. An Agilent 2100 bioanalyzer with an RNA 6000 Nano Kit (Agilent Technologies) was used to assess RNA concentration and integrity. Total RNA samples underwent mRNA enrichment prior to random fragmentation. Libraries were produced with a NEBNext Ultra RNA Library Prep Kit for Illumina (New England Biolabs) and sequenced on an Illumina NovaSeq 6000 platform using a 2 × 150 bp run protocol. More than 50 million nonstranded 150 bp paired-end reads were generated from each library. All FASTQ files were submitted to the Sequence Read Archive (SRA) database of the National Center for Biotechnology Information (NCBI) under the BioProject accession number PRJNA1090680 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1090680).

Differential gene expression analysis was conducted by Novogene. Briefly, raw data were processed through in-house Perl scripts to remove low-quality reads and reads containing adapters or more than 10% uncertain nucleotides. Clean reads were aligned to the Arabidopsis Col-0 genome (TAIR10) with HISAT2 2.0.5 (Kim et al. 2019) and assembled with StringTie 1.3.3b (Pertea et al. 2015), using default parameters. The number of reads (counts) mapped to each gene was determined using FeatureCounts 1.5.0-p3 (Liao et al. 2014) with default parameters, giving the number of fragments per kilobase of transcript per million mapped reads (FPKM). The identification of differentially expressed genes between Col-0 and *cxip4-2* plants was performed using the DESeq2 R package 1.20.0, with the selection criteria set at a fold change ≥ 2 or ≤ 0.5, for up- or downregulated genes, and a Benjamini–Hochberg false discovery rate (FDR) < 0.05.

Differential AS analysis was carried out at the Bioinformatics for Genomics and Proteomics Unit of the Centro Nacional de Biotecnología (CNB, Madrid). Briefly, the quality and purity of the raw reads were determined using FastQC 0.11.9 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and FastQ Screen 0.14.1 (Wingett and Andrews 2018), respectively. Reads were aligned to the Arabidopsis Col-0 genome (TAIR10) using STAR

2.7.10a (Dobin et al. 2013) with default parameters. Differential AS events between Col-0 and *cxip4-2* plants were assessed with rMATS-turbo 4.1.2 (Shen et al. 2014). Only those events with an FDR < 0.05, an absolute Delta (percent spliced-in, PSI) > 0.1 (10%), and an average number of reads covering the event > 5 were considered to be statistically significant.

Statistical analyses for the overrepresented GO terms and PANTHER protein classes among the differentially expressed or spliced genes between Col-0 and *cxip4-2* plants were performed using Fisher's Exact tests with FDR correction, using the GO database released 2023-01-05 (DOI: 10.5281/zenodo.7942786) and PANTHER 18.0 (released 2023-08-01; https://pantherdb.org/; Thomas et al. 2022). The selection criteria were set at a fold enrichment > 1 and an FDR < 0.05.

RNA-FISH and confocal microscopy

Poly(A)$^+$ RNA-FISH assays were carried out as previously described (Parry et al. 2006; Micol-Ponce et al. 2020), using first-node leaves from Col-0 and *cxip4-2* plants collected 14 or 22 das, respectively, and a 40-mer fluorescein-labeled oligo(dT) probe (synthesized by Eurofins Genomics) at a concentration of 0.5 $\mu g \cdot mL^{-1}$ in PerfectHyb Plus Hybridization Buffer (Sigma-Aldrich).

To visualize fluorescent signals from the RNA-FISH assays and from the CXIP4:GFP fusion protein, samples were mounted on slides with Vectashield antifade mounting medium (Vector Laboratories) or water containing 0.5 $\mu g \cdot mL^{-1}$ of 4′,6-diamidino-2-phenylindole (DAPI). All images were taken under a Nikon D-Eclipse C1 confocal microscope with Nikon EZ-C1 operation software, or a Leica Stellaris eight STED microscope with Leica Application Suite X (LAS X) software. DAPI, fluorescein, and GFP were excited at 408, 488, and 543 nm and their emissions detected at 450/35 nm, 515/30 nm, and 650 LP nm, respectively.

Bioinformatic analysis using protein sequences

The multiple sequence alignments shown in Supplementary Figs. S6 and S7 were obtained using MUSCLE V3.8 (Edgar 2004) from the EMBL-EBI bioinformatic web server (Madeira et al. 2022), with default parameters.

To assess the subcellular and suborganellar localizations of the Arabidopsis CXIP4 protein, the MULocDeep web server (https://www.mu-loc.org/; Jiang et al. 2021, 2023) was utilized.

To predict the effect of eliminating part of the C-terminal half of CXIP4-2, the LOCALIZER (https://localizer.csiro.au/; Sperschneider et al. 2017) and NoD (https://www.compbio.dundee.ac.uk/www-nod/index.jsp; Scott et al. 2010, 2011) web tools were utilized.

Accession numbers

Sequence data from this article can be found at The Arabidopsis Information Resource (TAIR; https://www.arabidopsis.org) under the following accession numbers: *CXIP4* (AT2G28910), AT2G23640, AT3G50030, *GL1* (AT3G27920), *ERIL1* (AT3G15140), *PAP1* (AT1G56650), *LOX2* (AT3G45140), *SGS1* (AT3G10490), *SGS3* (AT5G23570), *NRPD1B* (AT2G40030), *CCR4D* (AT1G31500), *MEE44* (AT4G00060), AT2G01100, and *ACT2* (AT3G18780).

Acknowledgments

The authors would like to thank J.A. García-Martín for the differential AS analyses; J. Castelló, D. Navarro, and M. Gomariz for their excellent technical assistance; and J.L. Micol for useful discussions and comments on the manuscript, as well as for the use of his facilities.

Author contributions

M.R.P. obtained funding and conceived, designed, and supervised research. R.M.-P and R.S.-M. partially co-supervised this work. U.A.-V., S.F.-C., R.M.-P., R.S.-M., and A.R.-B. conducted the experiments. U.A.-V., S.F.-C., R.M.-P., and M.R.P. analyzed the data, created the figures, tables, and datasets and wrote the manuscript. All authors revised and approved the manuscript.

Supplementary data

The following materials are available in the online version of this article.

Supplementary Figure S1. Localization of the ZCCHC motif, R-rich region, and nuclear and nucleolar localization signals in the predicted wild-type CXIP4 and mutant CXIP4-1 and CXIP4-2 proteins.

Supplementary Figure S2. Reproductive development of Col-0 and *cxip4-2* plants over time.

Supplementary Figure S3. Structure of the AT2G23640 and AT3G50030 genes and developmental phenotypes of individuals from the SALK_097293 and SALK_017259 lines.

Supplementary Figure S4. Molecular effects of the *cxip4* mutations on the expression of the *CXIP4* gene.

Supplementary Figure S5. Spatial expression of *CXIP4* during Arabidopsis development.

Supplementary Figure S6. Sequence conservation among putative eukaryotic CXIP4 orthologs.

Supplementary Figure S7. Sequence conservation among putative CXIP4 orthologs in angiosperms.

Supplementary Figure S8. Predicted localizations of Arabidopsis CXIP4 and human SREK1P1 proteins.

Supplementary Figure S9. Defects in 23S rRNA maturation in *cxip4-2* plants.

Supplementary Figure S10. RT-qPCR analysis of the relative expression of *CXIP4*, *GL1*, *ERIL1*, *PAP1*, and *LOX2* in Col-0, *cxip4-2*, and *cxip4-2 CXIP4*$_{pro}$*:CXIP4:GFP* plants.

Supplementary Table S1. Oligonucleotides used in this work.

Supplementary Table S2. Summary of protein-coding genes assigned to the PTHR31437 family according to the PANTHER database.

Supplementary Table S3. Summary of high-confidence physical protein–protein interactions for human SREK1IP1, identified from high-throughput experiments according to the HitPredict database.

Supplementary Data Set 1. Deregulated genes in *cxip4-2* plants.

Supplementary Data Set 2. Alternative splicing events in *cxip4-2* plants.

Funding

This work was supported by grants from the Ministerio de Ciencia, Innovación y Universidades of Spain (PID2020-117125RB-I00 [MCI/AEI/FEDER, UE]) and the Generalitat Valenciana (PROMETEO CIPROM/2022/2) to M.R.P. R.M.-P. held a María Zambrano distinguished researcher contract funded by the Next-Generation EU programs and administered by the Universidad Miguel Hernández. U.I.A.-V. held a predoctoral fellowship from the Generalitat Valenciana (GRISOLIAP/2016/134).

Conflict of interest statement. None declared.

Data availability

The data underlying this article are available in the article and in its online supplementary material except the FASTQ files that were submitted to the Sequence Read Archive (SRA) database of the National Center for Biotechnology Information (NCBI) under the BioProject accession number PRJNA1090680 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1090680).

References

Aceituno-Valenzuela U, Micol-Ponce R, Ponce MR. Genome-wide analysis of CCHC-type zinc finger (ZCCHC) proteins in yeast, Arabidopsis, and humans. *Cell Mol Life Sci*. 2020:77(20):3991–4014. https://doi.org/10.1007/s00018-020-03518-7

Akiyama Y, Koda Y, Byeon SJ, Shimada S, Nishikawaji T, Sakamoto A, Chen Y, Kojima K, Kawano T, Eishi Y, *et al*. Reduced expression of SET7/9, a histone mono-methyltransferase, is associated with gastric cancer progression. *Oncotarget*. 2016:7(4):3966–3983. https://doi.org/10.18632/oncotarget.6681

Alonso JM, Stepanova AN, Leisse TJ, Kim CJ, Chen H, Shinn P, Stevenson DK, Zimmerman J, Barajas P, Cheuk R, *et al*. Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. *Science*. 2003:301(5633):653–657. https://doi.org/10.1126/science.1086391

Belmont AS. Nuclear compartments: an incomplete primer to nuclear compartments, bodies, and genome organization relative to nuclear architecture. *Cold Spring Harb Perspect Biol*. 2022:14(7):a041268. https://doi.org/10.1101/cshperspect.a041268

Berná G, Robles P, Micol JL. A mutational analysis of leaf morphogenesis in *Arabidopsis thaliana*. *Genetics*. 1999:152(2):729–742. https://doi.org/10.1093/genetics/152.2.729

Bhuvanagiri M, Schlitter AM, Hentze MW, Kulozik AE. NMD: RNA biology meets human genetic medicine. *Biochem J*. 2010:430(3):365–377. https://doi.org/10.1042/BJ20100699

Cabezas-Fuster A, Micol-Ponce R, Fontcuberta-Cervera S, Ponce MR. Missplicing suppressor alleles of Arabidopsis *PRE-MRNA PROCESSING FACTOR 8* increase splicing fidelity by reducing the use of novel splice sites. *Nucleic Acids Res*. 2022:50(10):5513–5527. https://doi.org/10.1093/nar/gkac338

Casanova-Sáez R, Mateo-Bonmatí E, Kangasjärvi S, Candela H, Micol JL. *Arabidopsis* ANGULATA10 is required for thylakoid biogenesis and mesophyll development. *J Exp Bot*. 2014:65(9):2391–2404. https://doi.org/10.1093/jxb/eru131

Clough SJ, Bent AF. Floral dip: a simplified method for Agrobacterium-mediated transformation of *Arabidopsis thaliana*. *Plant J*. 1998:16(6):735–743. https://doi.org/10.1046/j.1365-313x.1998.00343.x

Coego A, Brizuela E, Castillejo P, Ruíz S, Koncz C, del Pozo JC, Piñeiro M, Jarillo JA, Paz-Ares J, León J, *et al*. The TRANSPLANTA collection of Arabidopsis lines: a resource for functional analysis of transcription factors based on their conditional overexpression. *Plant J*. 2014:77(6):944–953. https://doi.org/10.1111/tpj.12443

Covelo-Molares H, Obrdlik A, Poštulková I, Dohnálková M, Gregorová P, Ganji R, Potěšil D, Gawriyski L, Varjosalo M, Vaňáčová S. The comprehensive interactomes of human adenosine RNA methyltransferases and demethylases reveal distinct functional and regulatory features. *Nucleic Acids Res*. 2021:49(19):10895–-10910. https://doi.org/10.1093/nar/gkab900

Curtis MD, Grossniklaus U. A gateway cloning vector set for high-throughput functional analysis of genes in planta. *Plant Physiol*. 2003:133(2):462–469. https://doi.org/10.1104/pp.103.027979

Chen D, Li Z, Yang Q, Zhang J, Zhai Z, Shu HB. Identification of a nuclear protein that promotes NF-κB activation. *Biochem Biophys Res Commun*. 2003:310(3):720–724. https://doi.org/10.1016/j.bbrc.2003.09.074

Chen H, Su Z, Tian B, Hao G, Trick HN, Bai G. *TaHRC* suppresses the calcium-mediated immune response and triggers wheat Fusarium head blight susceptibility. *Plant Physiol*. 2022:190(3):1566–1569. https://doi.org/10.1093/plphys/kiac352

Cheng N, Nakata PA. Development of a rapid and efficient protoplast isolation and transfection method for chickpea (*Cicer arietinum*). *MethodsX*. 2020:7:101025. https://doi.org/10.1016/j.mex.2020.101025

Cheng NH, Liu JZ, Nelson RS, Hirschi KD. Characterization of CXIP4, a novel *Arabidopsis* protein that activates the H^+/Ca^{2+} antiporter, CAX1. *FEBS Lett*. 2004:559(1–3):99–106. https://doi.org/10.1016/S0014-5793(04)00036-5

de Francisco Amorim M, Willing EM, Szabo EX, Francisco-Mangilet AG, Droste-Borel I, Maček B, Schneeberger K, Laubinger S. The U1 snRNP subunit LUC7 modulates plant development and stress responses via regulation of alternative splicing. *Plant Cell*. 2018:30(11):2838–2854. https://doi.org/10.1105/tpc.18.00244

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

Donnelly PM, Bonetta D, Tsukaya H, Dengler RE, Dengler NG. Cell cycling and cell enlargement in developing leaves of *Arabidopsis*. *Dev Biol*. 1999:215(2):407–419. https://doi.org/10.1006/dbio.1999.9443

Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004:32(5):1792–1797. https://doi.org/10.1093/nar/gkh340

Fica SM, Oubridge C, Wilkinson ME, Newman AJ, Nagai K. A human post-catalytic spliceosome structure reveals essential roles of metazoan factors for exon ligation. *Science*. 2019:363(6428):710–714. https://doi.org/10.1126/science.aaw5569

Fortes P, Bilbao-Cortés D, Fornerod M, Rigaut G, Raymond W, Séraphin B, Mattaj IW. Luc7p, a novel yeast U1 snRNP protein with a role in 5′ splice site recognition. *Genes Dev*. 1999:13(18):2425–2438. https://doi.org/10.1101/gad.13.18.2425

He Y, Yang X, Xia X, Wang Y, Dong Y, Wu L, Jiang P, Zhang X, Jiang C, Ma H, *et al*. A phase-separated protein hub modulates resistance to Fusarium head blight in wheat. *Cell Host Microbe*. 2024:32(5):710–726.e710. https://doi.org/10.1016/j.chom.2024.04.002

Heese K, Fujita M, Akatsu H, Yamamoto T, Kosaka K, Nagai Y, Sawada T. The splicing regulatory protein p18SRP is down-regulated in Alzheimer's disease brain. *J Mol Neurosci*. 2004:24(2):269–276. https://doi.org/10.1385/JMN:24:2:269

Hossain MA, Henríquez-Valencia C, Gómez-Páez M, Medina J, Orellana A, Vicente-Carbajosa J, Zouhar J. Identification of novel components of the unfolded protein response in Arabidopsis. *Front Plant Sci*. 2016:7:650. https://doi.org/10.3389/fpls.2016.00650

Huttlin EL, Bruckner RJ, Paulo JA, Cannon JR, Ting L, Baltier K, ColbyG, Gebreab F, Gygi MP, Parzen H, *et al*. Architecture of the human interactome defines protein communities and disease networks. *Nature*. 2017:545(7655):505–509. https://doi.org/10.1038/nature22366

Iwata Y, Koizumi N. An *Arabidopsis* transcription factor, AtbZIP60, regulates the endoplasmic reticulum stress response in a manner unique to plants. *Proc Natl Acad Sci U S A*. 2005:102(14):5280–5285. https://doi.org/10.1073/pnas.0408941102

Jiang Y, Jiang L, Akhil CS, Wang D, Zhang Z, Zhang W, Xu D. MULocDeep web service for protein localization prediction and visualization at subcellular and suborganellar levels. *Nucleic*

Acids Res. 2023:51(W1):343–349. https://doi.org/10.1093/nar/gkad374

Jiang Y, Wang D, Yao Y, Eubel H, Künzler P, Møller IM, Xu D. MULocDeep: a deep-learning framework for protein subcellular and suborganellar localization prediction with residue-level interpretation. *Comput Struct Biotechnol J*. 2021:19:4825–4839. https://doi.org/10.1016/j.csbj.2021.08.027

Kamarck ML, Trimmer C, Murphy NR, Gregory KM, Manoel D, Logan DW, Saraiva LR, Mainland JD. Identifying candidate genes underlying isolated congenital anosmia. *Clin Genet*. 2024:105(4):376–385. https://doi.org/10.1111/cge.14470

Kelly SM, Corbett AH. Messenger RNA export from the nucleus: a series of molecular wardrobe changes. *Traffic*. 2009:10(9):1199–1208. https://doi.org/10.1111/j.1600-0854.2009.00944.x

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol*. 2019:37(8):907–915. https://doi.org/10.1038/s41587-019-0201-4

Kleinboelting N, Huep G, Kloetgen A, Viehoever P, Weisshaar B. GABI-Kat SimpleSearch: new features of the *Arabidopsis thaliana* T-DNA mutant database. *Nucleic Acids Res*. 2012:40(D1):D1211–D1215. https://doi.org/10.1093/nar/gkr1047

Klepikova AV, Kasianov AS, Gerasimov ES, Logacheva MD, Penin AA. A high resolution map of the *Arabidopsis thaliana* developmental transcriptome based on RNA-seq profiling. *Plant J*. 2016:88(6):1058–1070. https://doi.org/10.1111/tpj.13312

Larkin JC, Oppenheimer DG, Lloyd AM, Paparozzi ET, Marks MD. Roles of the *GLABROUS1* and *TRANSPARENT TESTA GLABRA* genes in Arabidopsis trichome development. *Plant Cell*. 1994:6(8):1065–1076. https://doi.org/10.2307/3869885

Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*. 2014:30(7):923–930. https://doi.org/10.1093/bioinformatics/btt656

Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) method. *Methods*. 2001:25(4):402–408. https://doi.org/10.1006/meth.2001.1262

Llinas RJ, Xiong JQ, Clark NM, Burkhart SE, Bartel B. An Arabidopsis *pre-RNA processing8a* (*prp8a*) missense allele restores splicing of a subset of mis-spliced mRNAs. *Plant Physiol*. 2022:189(4):2175–2192. https://doi.org/10.1093/plphys/kiac221

López Y, Nakai K, Patil A. HitPredict version 4: comprehensive reliability scoring of physical protein–protein interactions from more than 100 species. *Database*. 2015:2015:bav117. https://doi.org/10.1093/database/bav117

Madeira F, Pearce M, Tivey ARN, Basutkar P, Lee J, Edbali O, Madhusoodanan N, Kolesnikov A, Lopez R. Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Res*. 2022:50(W1):W276–W279. https://doi.org/10.1093/nar/gkac240

Mancinelli AL. Interaction between light quality and light quantity in the photoregulation of anthocyanin production. *Plant Physiol*. 1990:92(4):1191–1195. https://doi.org/10.1104/pp.92.4.1191

Mermigka G, Helm JM, Vlatakis I, Schumacher HT, Vamvaka E, Kalantidis K. ERIL1, the plant homologue of ERI-1, is involved in the processing of chloroplastic rRNAs. *Plant J*. 2016:88(5):839–853. https://doi.org/10.1111/tpj.13304

Micol-Ponce R, Sarmiento-Mañús R, Fontcuberta-Cervera S, Cabezas-Fuster A, de Bures A, Sáez-Vásquez J, Ponce MR. SMALL ORGAN4 is a ribosome biogenesis factor involved in 5.8S ribosomal RNA maturation. *Plant Physiol*. 2020:184(4):2022–2039. https://doi.org/10.1104/pp.19.01540

Nadi R, Juan-Vicente L, Mateo-Bonmatí E, Micol JL. The unequal functional redundancy of Arabidopsis *INCURVATA11* and *CUPULIFORMIS2* is not dependent on genetic background. *Front Plant Sci*. 2023:14:1239093. https://doi.org/10.3389/fpls.2023.1239093

Palazzo AF, Lee ES. Sequence determinants for nuclear retention and cytoplasmic export of mRNAs and lncRNAs. *Front Genet*. 2018:9:440. https://doi.org/10.3389/fgene.2018.00440

Parry G, Ward S, Cernac A, Dharmasiri S, Estelle M. The *Arabidopsis* SUPPRESSOR OF AUXIN RESISTANCE proteins are nucleoporins with an important role in hormone signaling and development. *Plant Cell*. 2006:18(7):1590–1603. https://doi.org/10.1105/tpc.106.041566

Passmore LA, Coller J. Roles of mRNA poly(A) tails in regulation of eukaryotic gene expression. *Nat Rev Mol Cell Biol*. 2022:23(2):93–106. https://doi.org/10.1038/s41580-021-00417-y

Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol*. 2015:33(3):290–295. https://doi.org/10.1038/nbt.3122

Ponce MR, Quesada V, Micol JL. Rapid discrimination of sequences flanking and within T-DNA insertions in the *Arabidopsis* genome. *Plant J*. 1998:14(4):497–501. https://doi.org/10.1046/j.1365-313X.1998.00146.x

Ponce MR, Robles P, Lozano FM, Brotóns MA, Micol JL. Low-resolution mapping of untagged mutations. *Methods Mol Biol*. 2006:323:105–113. https://doi.org/10.1385/1-59745-003-0:105

Ray D, Laverty KU, Jolma A, Nie K, Samson R, Pour SE, Tam CL, von Krosigk N, Nabeel-Shah S, Albu M, *et al*. RNA-binding proteins that lack canonical RNA-binding domains are rarely sequence-specific. *Sci Rep*. 2023:13(1):5238. https://doi.org/10.1038/s41598-023-32245-9

Reed R. Coupling transcription, splicing and mRNA export. *Curr Opin Cell Biol*. 2003:15(3):326–331. https://doi.org/10.1016/s0955-0674(03)00048-6

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. *Nat Biotechnol*. 2011:29(1):24–26. https://doi.org/10.1038/nbt.1754

Rudzka M, Wróblewska-Ankiewicz P, Majewska K, Hyjek-Składanowska M, Gołebiewski M, Sikora M, Smoliński DJ, Kołowerzo-Lubnau A. Functional nuclear retention of pre-mRNA involving Cajal bodies during meiotic prophase in European larch (*Larix decidua*). *Plant Cell*. 2022:34(6):2404–2423. https://doi.org/10.1093/plcell/koac091

Sánchez-García AB, Aguilera V, Micol-Ponce R, Jover-Gil S, Ponce MR. Arabidopsis MAS2, an essential gene that encodes a homolog of animal NF-k B activating protein, is involved in 45S ribosomal DNA silencing. *Plant Cell*. 2015:27(7):1999–2015. https://doi.org/10.1105/tpc.15.00135

Sasaki T, Kanno T, Liang SC, Chen PY, Liao WW, Lin WD, Matzke AJ, Matzke M. An rtf2 domain-containing protein influences pre-mRNA splicing and is essential for embryonic development in *Arabidopsis thaliana*. *Genetics*. 2015:200(2):523–535. https://doi.org/10.1534/genetics.115.176438

Schindler S, Szafranski K, Hiller M, Ali GS, Palusa SG, Backofen R, Platzer M, Reddy AS. Alternative splicing at NAGNAG acceptors in *Arabidopsis thaliana* SR and SR-related protein-coding genes. *BMC Genomics*. 2008:9(1):159. https://doi.org/10.1186/1471-2164-9-159

Scott MS, Boisvert FM, McDowall MD, Lamond AI, Barton GJ. Characterization and prediction of protein nucleolar localization sequences. *Nucleic Acids Res*. 2010:38(21):7388–7399. https://doi.org/10.1093/nar/gkq653

Scott MS, Troshin PV, Barton GJ. Nod: a nucleolar localization sequence detector for eukaryotic and viral proteins. *BMC Bioinformatics*. 2011:12(1):317. https://doi.org/10.1186/1471-2105-12-317

Shen S, Park JW, Lu ZX, Lin L, Henry MD, Wu YN, Zhou Q, Xing Y. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc Natl Acad Sci U S A*. 2014:111(51):E5593–E5601. https://doi.org/10.1073/pnas.1419161111

Sperschneider J, Catanzariti AM, DeBoer K, Petre B, Gardiner DM, Singh KB, Dodds PN, Taylor JM. LOCALIZER: subcellular localization prediction of both plant and effector proteins in the plant cell. *Sci Rep*. 2017:7(1):44598. https://doi.org/10.1038/srep44598

Su Z, Bernardo A, Tian B, Chen H, Wang S, Ma H, Cai S, Liu D, Zhang D, Li T, *et al*. A deletion mutation in *TaHRC* confers *Fhb1* resistance to Fusarium head blight in wheat. *Nat Genet*. 2019:51(7):1099–1105. https://doi.org/10.1038/s41588-019-0425-8

Sun S, Gao T, Pang B, Su X, Guo C, Zhang R, Pang Q. RNA binding protein NKAP protects glioblastoma cells from ferroptosis by promoting SLC7A11 mRNA splicing in an m^6A-dependent manner. *Cell Death Dis*. 2022:13(1):73. https://doi.org/10.1038/s41419-022-04524-2

Teng S, Keurentjes J, Bentsink L, Koornneef M, Smeekens S. Sucrose-specific induction of anthocyanin biosynthesis in Arabidopsis requires the *MYB75/PAP1* gene. *Plant Physiol*. 2005:139(4):1840–1852. https://doi.org/10.1104/pp.105.066688

Thomas PD, Ebert D, Muruganujan A, Mushayahama T, Albou LP, MiH. PANTHER: making genome-scale phylogenetics accessible to all. *Protein Sci*. 2022:31(1):8–22. https://doi.org/10.1002/pro.4218

Warda AS, Kretschmer J, Hackert P, Lenz C, Urlaub H, Höbartner C, Sloan KE, Bohnsack MT. Human METTL16 is a N^6-methyladenosine (m^6A) methyltransferase that targets pre-mRNAs and various non-coding RNAs. *EMBO Rep*. 2017:18(11):2004–2014. https://doi.org/10.15252/embr.201744940

Wegener M, Müller-McNicoll M. Nuclear retention of mRNAs—quality control, gene regulation and human disease. *Semin Cell Dev Biol*. 2018:79:131–142. https://doi.org/10.1016/j.semcdb.2017.11.001

Wingett SW, Andrews S. Fastq screen: a tool for multi-genome mapping and quality control. *F1000Res*. 2018:7:1338. https://doi.org/10.12688/f1000research.15931.1

Yu CP, Lin JJ, Li WH. Positional distribution of transcription factor binding sites in *Arabidopsis thaliana*. *Sci Rep*. 2016:6(1):25164. https://doi.org/10.1038/srep25164

Zhang J, Bai R, Li M, Ye H, Wu C, Wang C, Li S, Tan L, Mai D, Li G, *et al*. Excessive miR-25-3p maturation via N^6-methyladenosine stimulated by cigarette smoke promotes pancreatic cancer progression. *Nat Commun*. 2019:10(1):1858. https://doi.org/10.1038/s41467-019-09712-x

CXIP4 depletion causes early lethality and pre-mRNA missplicing in Arabidopsis

Uri Israel Aceituno-Valenzuela[1,2,†], Sara Fontcuberta-Cervera[1,†], Rosa Micol-Ponce[1], Raquel Sarmiento-Mañús[1], Alejandro Ruiz-Bayón[1], and María Rosa Ponce[1,*]

[1]Instituto de Bioingeniería, Universidad Miguel Hernández, Campus de Elche, 03202 Elche, Alicante, Spain

[2]Present address: Universidad de O'Higgins, Centro UOH de Biología de Sistemas para la Sanidad Vegetal (BioSaV). Ruta I-90 s/n, San Fernando, Chile

[†]These authors contributed equally to this work

*Author for correspondence: mrponce@umh.es

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (https://academic.oup.com/plphys/pages/General-Instructions) is María Rosa Ponce

Supplementary Figures and Tables

Supplementary Material included in this file:

Supplementary Figures S1 to S10

Supplementary Table S1

Supplementary Material not included in this file:

Supplementary Tables S2 and S3

Supplementary Data Sets S1 and S2

CXIP4

MPATAGRVRMPANNRVHSSAALQTHGIWQSAIGYDPYAPTSKEEPKTTQQKTEDPENSYASFQGLLALARITGSNNDEARGS
CKKCGRVGHLTFQCRNFLSTKEDKEKDPGAIEAAVLSGLEKIRRGVGKGEVEEVSSEEEEESESSDSDVDSEMERIIAERFG
KKKGGSSVKKTSSVRKKKKRVSDESDSDSDSGDRKRRRRSMKKRSSHKRRSLSESEDEEEGRSKRRKERRGRKRDEDDSDES
EDEDDRRVKRKSRKEKRRRRSRRNHSDDSDSESSEDDRRQKRRNKVAASSDSEANVSGDDVSRVGRGSSKRSEKKSRKRHHR
KERE

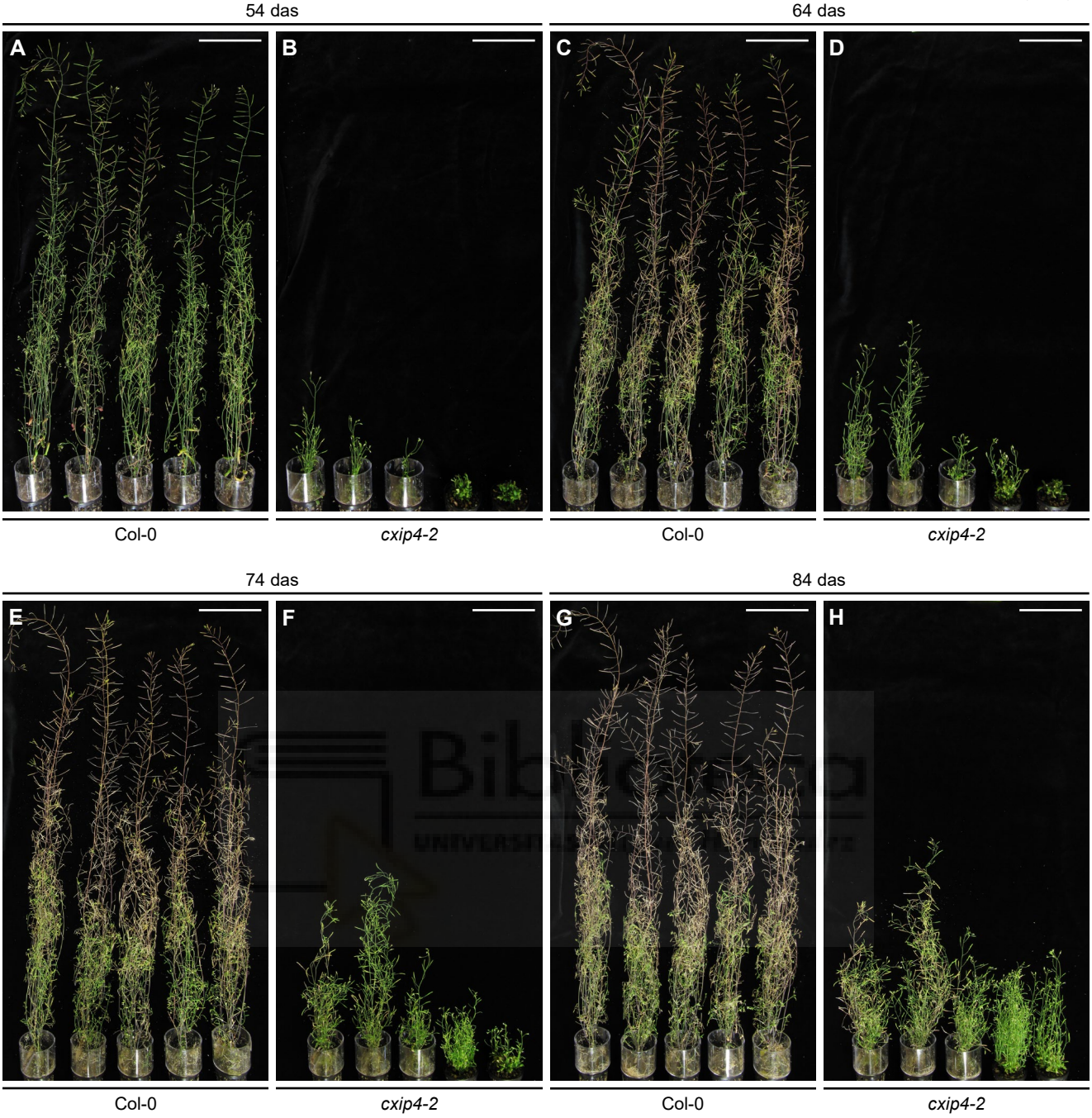
CXIP4-1

MPATAGRVRMPANNRVHSSAALQTHGIWQSAIGYDPYAPTSKEEPKTTQQKTEDPENSYASFQGLLALARITGSNNDEARGS
CKKCGRVGHLTFQCRNFLSTKEDKEKDPGAIEAAVLSGLEKIRRGVGKGEVEEVSSEEEEESESSDSDVDSEMERIIAERFG
KKKGGSSVKKTSSVRKKKKRVSDESDSDSDSGDRKRRRRSMKKRSSHKRRSLSESEDEEEGRSKRRKERRGRKRDEDDSDES
EDEDDRRVKRKSRKEKRRRRSRRNHRIYSIVNGFMSGKSTWISNEYDGQYGEKERVITNFFSIQKCRCPQRYYKMKVHFDKT
TNYDPSYL

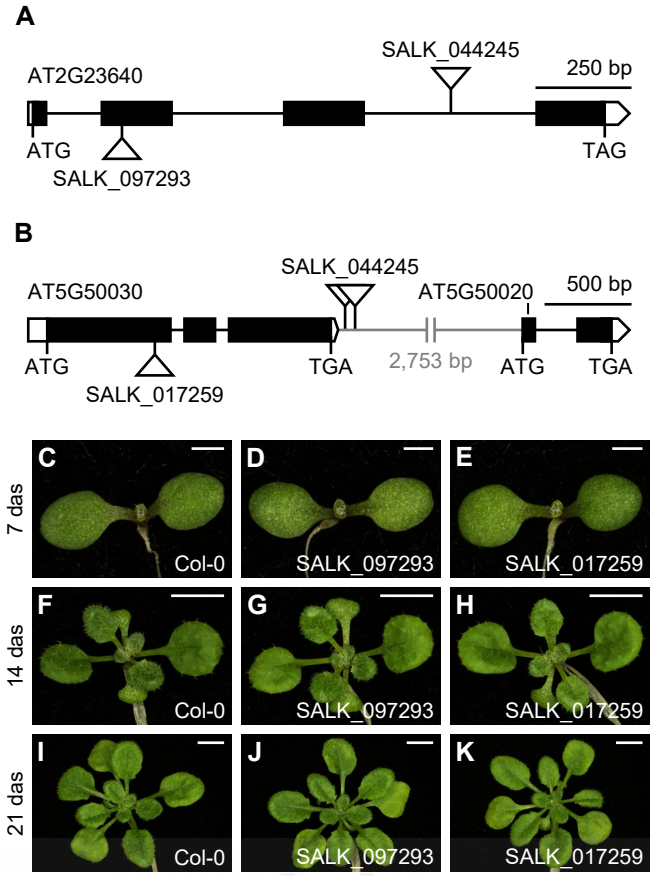
CXIP4-2

MPATAGRVRMPANNRVHSSAALQTHGIWQSAIGYDPYAPTSKEEPKTTQQKTEDPENSYASFQGLLALARITGSNNDEARGS
CKKCGRVGHLTFQCRNFLSTKEDKEKDPGAIEAAVLSGLEKIRRGVGKGEVEEVSSEEEEESESSDSDVDSEMERIIAERFG
KKKGGSSVKKTSSVRKKKKRVSDESDSDSDSGDRKRRRRSMKKRSSHKRRSLSESEDEEEGRSKRRKERRGRKMKP

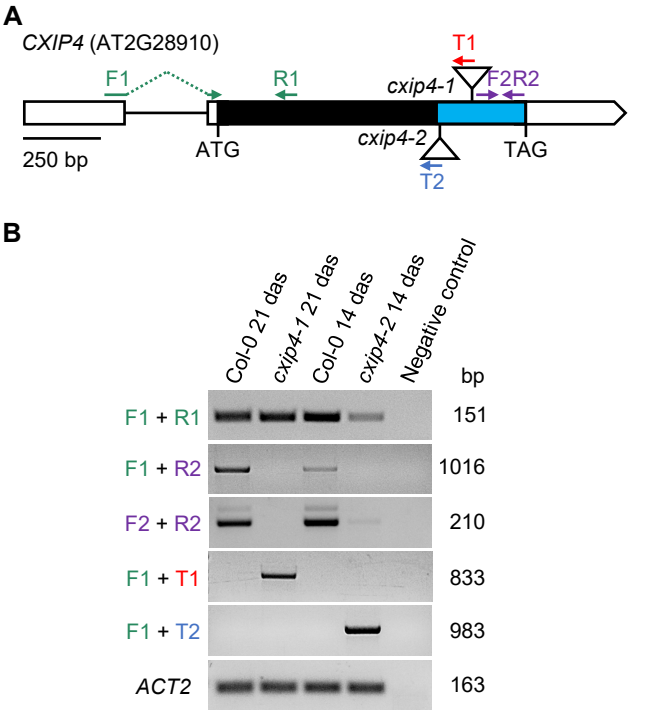
Supplementary Figure S1. Localization of the ZCCHC motif, R-rich region, and nuclear and nucleolar localization signals in the predicted wild-type CXIP4 and mutant CXIP4-1 and CXIP4-2 proteins. The ZCCHC motif is highlighted in green, with the conserved cysteine and histidine residues ($CX_2CX_4HX_4C$) shown in red letters. The R-rich region, as annotated in UniProtKB (https://www.uniprot.org/uniprotkb), is underlined. Nuclear localization signals (NLSs) are denoted with red letters, and nucleolar localization signals (NoLSs) are highlighted in yellow. The 65 amino acids (aa) at the C-terminus of CXIP4-1 (blue letters) replace the 61 aa of the corresponding region of CXIP4. CXIP4-2 contains 240 aa, with only the last three absent from CXIP4 (blue letters).



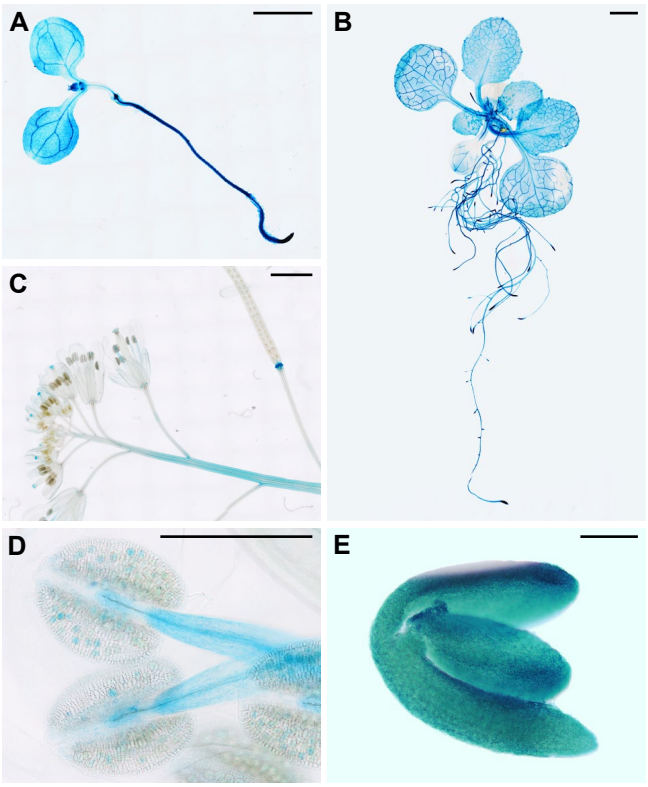
Supplementary Figure S2. Reproductive development of Col-0 and *cxip4-2* plants over time. Photographs were taken (A, B) 54, (C, D) 64, (E, F) 74, and (G, H) 84 days after stratification (das). Scale bars: 10 cm.



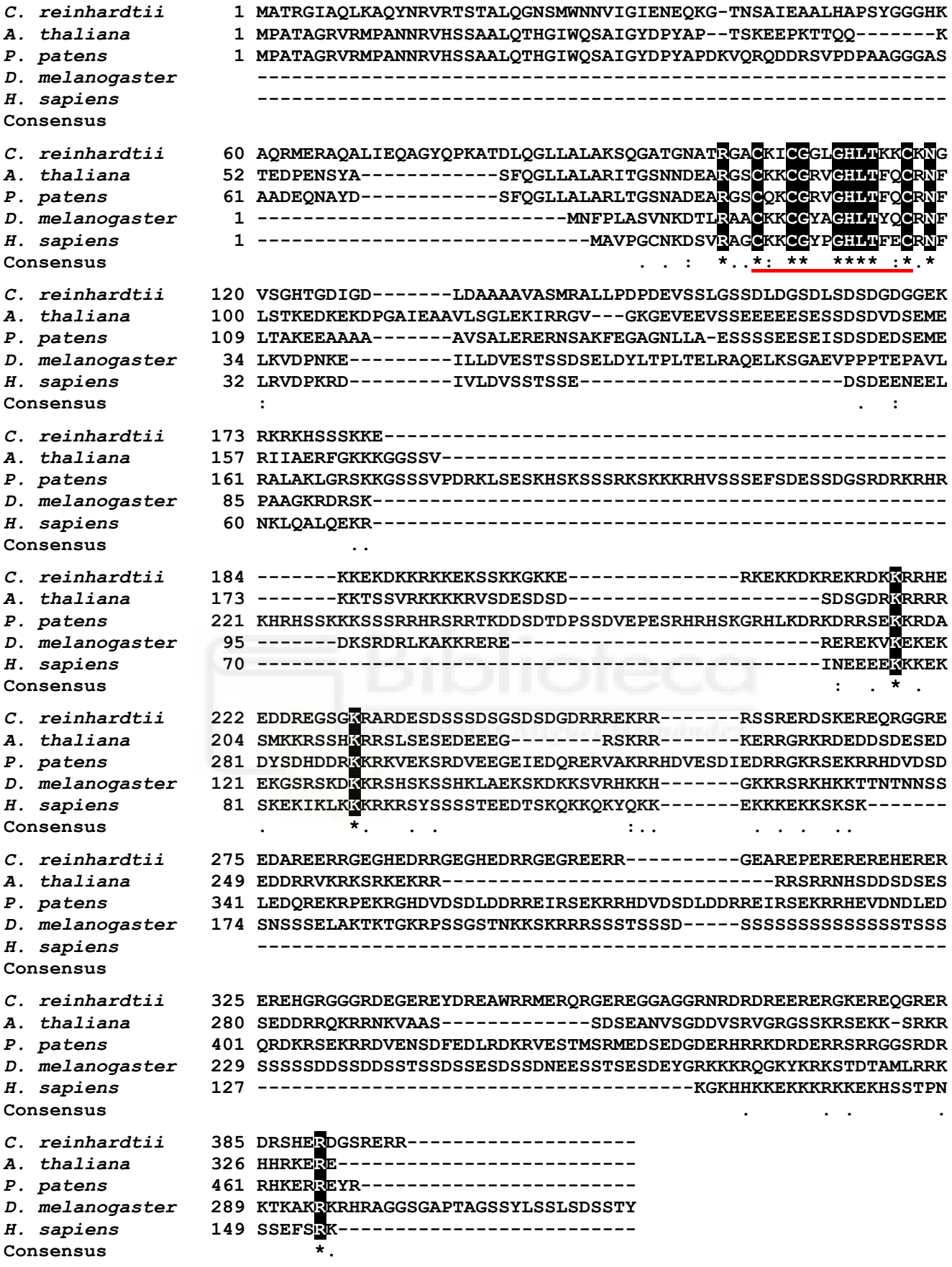
Supplementary Figure S3. Structure of the AT2G23640 and AT3G50030 genes and developmental phenotypes of individuals from the SALK_097293 and SALK_017259 lines. (A, B) Schematic representation of (A) AT2G23640, and (B) AT3G50030 and AT3G50020, including the 3,753-bp intergenic region between both genes. The positions of the start (ATG) and stop (TAG or TGA) codons, and the T-DNA insertions (triangles) in the SALK_044245, SALK_097293, and SALK_017259 lines studied in this work are indicated. Boxes represent exons, with untranslated (UTRs) and coding regions shown in white and black, respectively. The gray line represents the intergenic region between AT3G50030 and AT3G50020. (C-K) Morphological vegetative phenotypes of individuals from (C, F, I) Col-0, (D, G, J) SALK_097293, and (E, H, K) SALK_017259 lines. Photographs were taken (C-E) 7, (F-H) 14, and (I-K) 21 das. Scale bars: (C-E) 1 mm and (F-K) 5 mm.



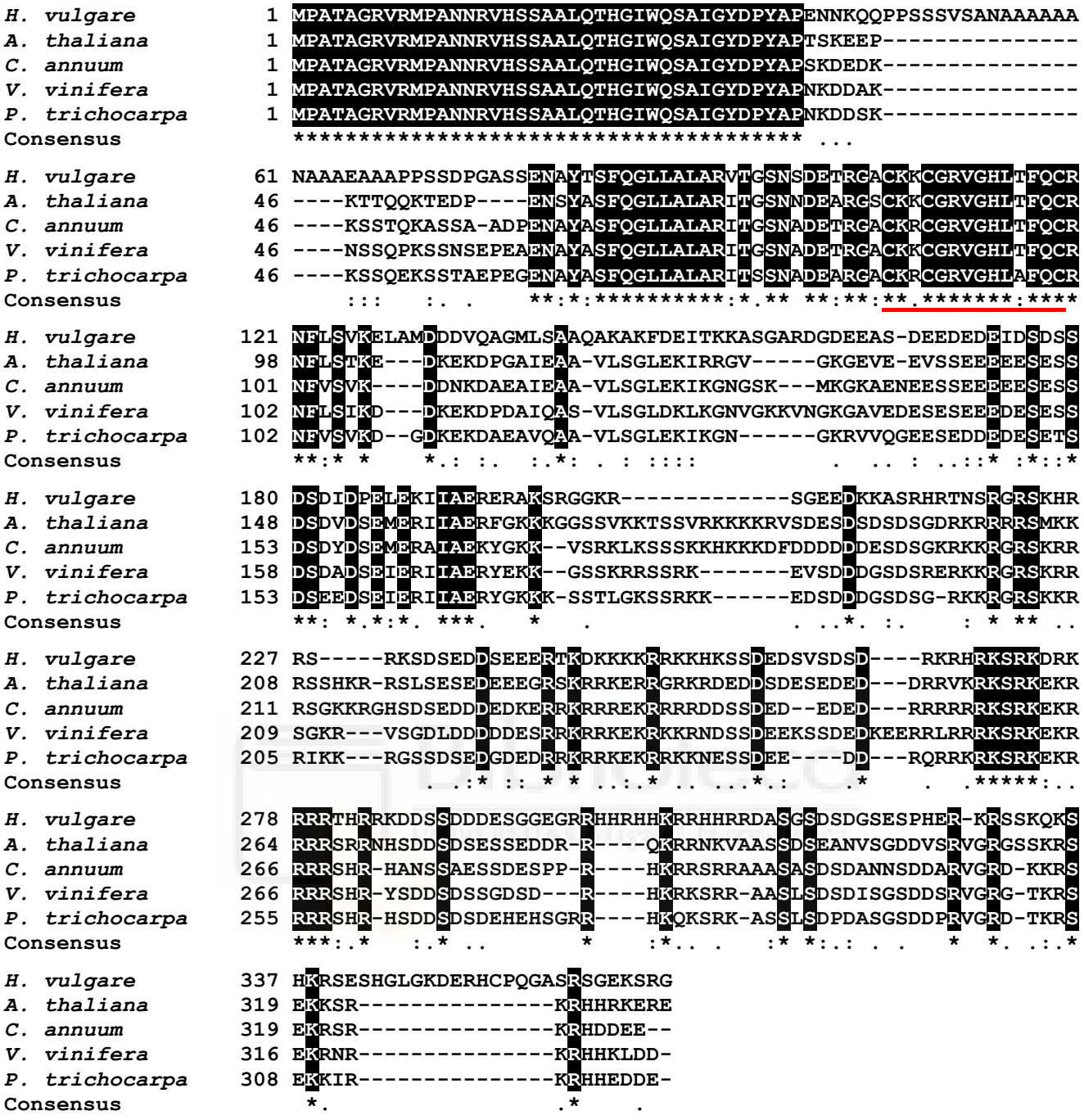
Supplementary Figure S4. Molecular effects of the *cxip4* mutations on the expression of the *CXIP4* gene. (A) Schematic representation of the *CXIP4* gene (as described in the legend of Figure 1), illustrating the positions of the primers used to analyze the effects of T-DNA insertions on *CXIP4* transcription. The primers are not drawn to scale and are indicated in green (F1 and R1), purple (F2 and R2), red (T1), and blue (T2). They are labeled as F1 (corresponding to CXIP4-F2 in the Supplementary Table S3), R1 (CXIP4-R2), F2 (CXIP4-F3), R2 (CXIP4-R3), T1 (o8409), and T2 (LBb1.3). (B) RT-PCR analysis of *CXIP4* expression in Col-0, *cxip4-1*, and *cxip4-2* plants. Water was used as a negative control, while the expression of the *ACTIN2* (*ACT2*) housekeeping gene acted as an internal control. (C) Sequences of the chimeric *cxip4-1* and *cxip4-2* cDNAs and their predicted translation products, CXIP4-1 and CXIP4-2. Nucleotide sequences corresponding to T-DNAs and their in silico translation to amino acids are highlighted in red. The blue rectangles in panels A and C indicate the same region. Asterisks denote translation stop codons.



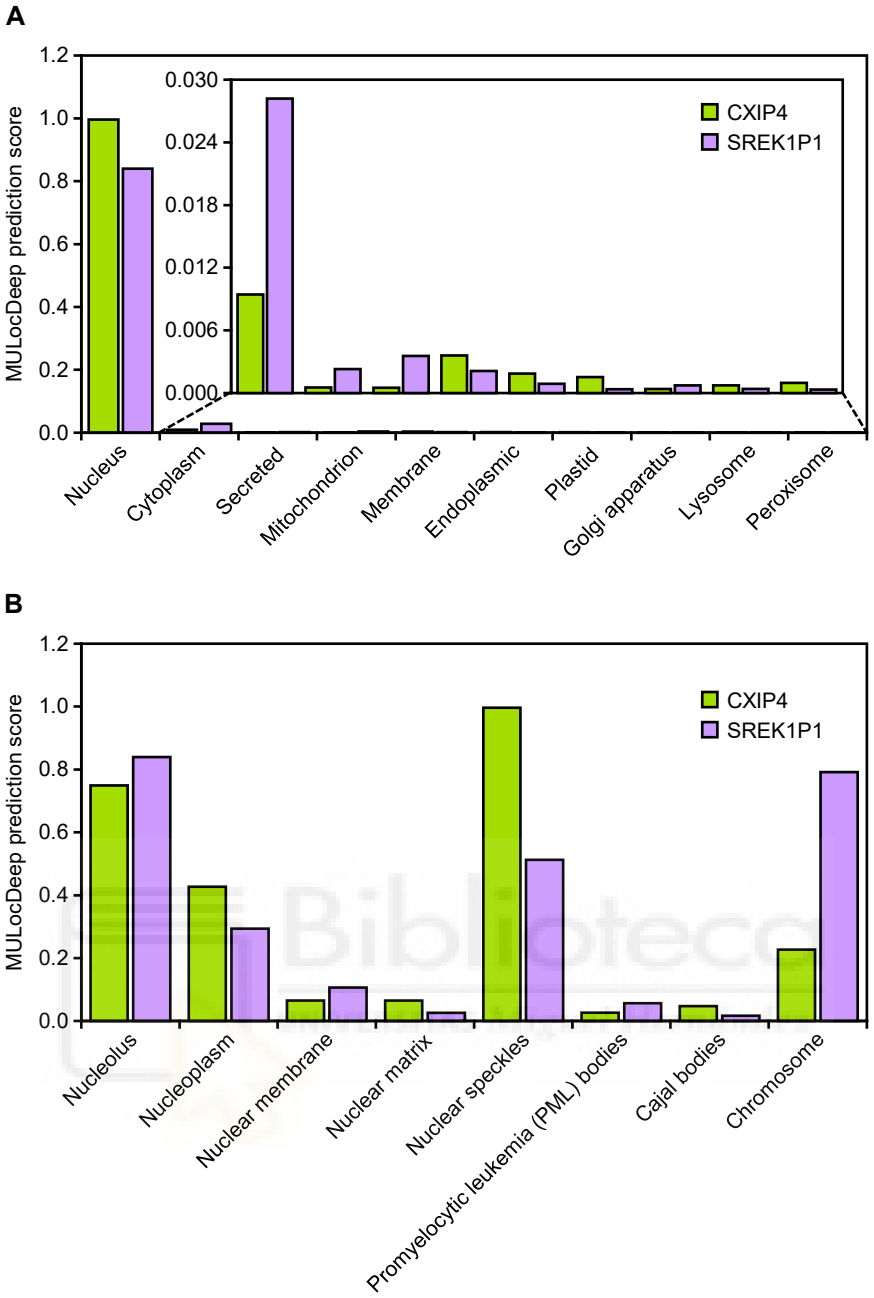
Supplementary Figure S5. Spatial expression of *CXIP4* during Arabidopsis development. GUS activity in Col-0 *CXIP4*$_{proII}$*:GUS* transgenic plants carrying the short region of the *CXIP4* promoter during (A, B) vegetative and (C-E) reproductive development. GUS staining was performed using (A, B) whole plants collected (A) 7 and (B) 14 das; (C, D) inflorescences, siliques, and anthers from 50 das-old plants; and (E) embryos dissected from green mature seeds of 52 das-old plants. Scale bars: (A-D) 2 mm, and (E) 100 µm. Equivalent GUS staining results were obtained from Col-0 *CXIP4*$_{proI}$*:GUS* plants.



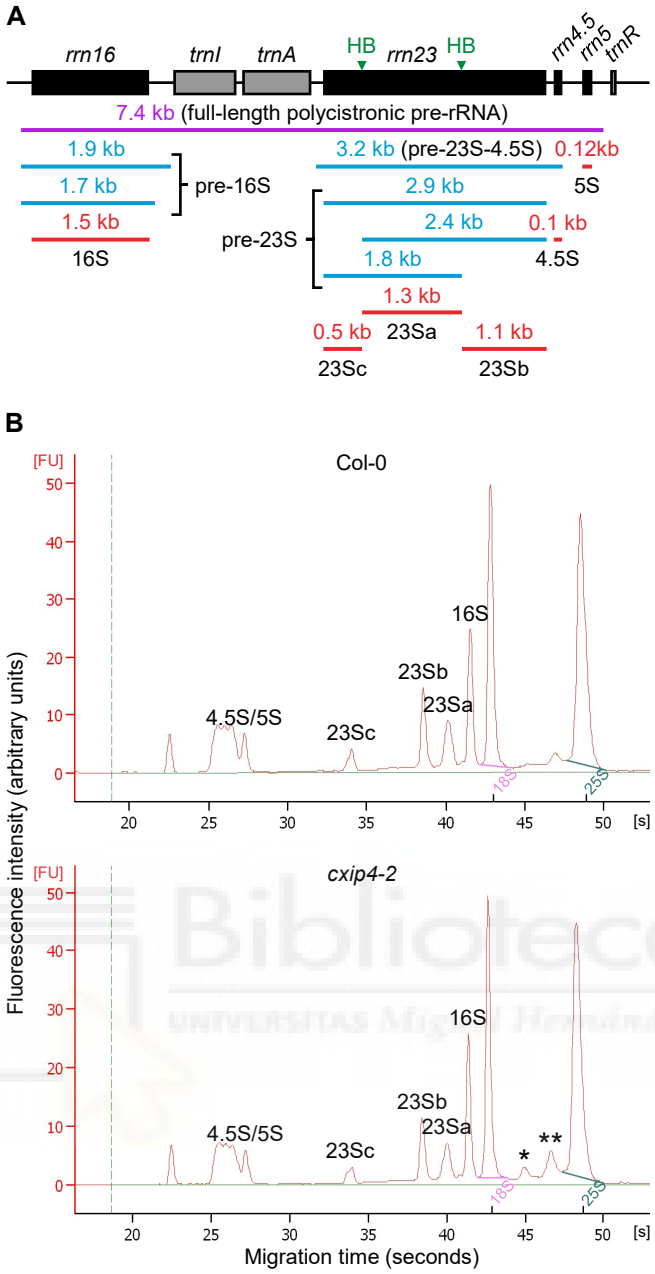
Supplementary Figure S6. Sequence conservation among putative eukaryotic CXIP4 orthologs. Alignment of full-length putative CXIP4 orthologs from *Chlamydomonas reinhardtii* (UniProtKB accession number A0A2K3DE08), *Arabidopsis thaliana* (Q84Y18), *Physcomitrium patens* (A0A2K1K6W5), *Drosophila melanogaster* (Q9W3Z5), and *Homo sapiens* (Q8N9Q2). Asterisks and dots in the consensus line indicate identical (also shaded in black) and conserved residues, respectively. The consensus sequence of the ZCCHC motif is underlined in red.



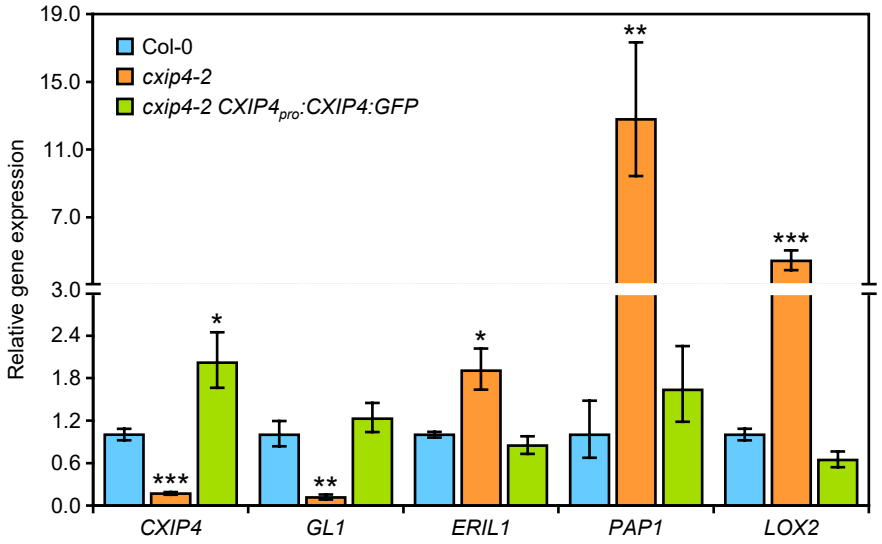
Supplementary Figure S7. Sequence conservation among CXIP4 orthologs in angiosperms. Alignment of full-length CXIP4 orthologs from different angiosperm lineages: *Hordeum vulgare* (UniProtKB accession number A0A287R5P6) representing Poales, *Arabidopsis thaliana* (Q84Y18) for Malvids, *Capsicum annuum* (A0A2G3ANM6) for Asterids, *Vitis vinifera* (F6GV10) for Vitales, and *Populus trichocarpa* (A0A2K2BXT0) for Malpighiales. Only species with a single ZCCHC motif, whose $CX_2CX_4HX_4C$ consensus sequence is underlined in red, were selected for analysis. Asterisks and dots indicate identical (also shaded in black) and conserved residues, respectively.



Supplementary Figure S8. Predicted localizations of Arabidopsis CXIP4 and human SREK1P1 proteins. (A, B) Predicted (A) subcellular and (B) suborganellar localizations of CXIP4 and SREK1P1, using the MULocDeep web server (https://mu-loc.org/).



Supplementary Figure S9. Defects in 23S rRNA maturation in *cxip4-2* plants. (A) Schematic representation of the chloroplast *rrn* operon, indicating the lengths of the polycistronic primary transcript (purple), various processing intermediates (blue), and the 4.5S, 5S, 16S, and 23Sa/b/c mature rRNAs (red). Black and gray boxes represent chloroplast rRNA and tRNA genes, respectively, while lines indicate the intergenic sequences. The positions of the internal cleavage sites or hidden breaks (HB) in the 23S rRNA are shown with green arrowheads above the *rrn23* gene. The diagram was modified from Li *et al.* (2021). (B) Agilent 2100 bioanalyzer electropherogram profiles of total RNA extracted from Col-0 and *cxip4-2* plants collected 14 das. Peaks corresponding to chloroplast 4.5S/5S, 16S, and 23Sa/b/c rRNAs and cytoplasmic 18S and 25S rRNAs are labeled. Asterisks in the electropherogram profile of *cxip4-2* highlight the accumulation of incompletely processed 23S rRNAs of 2.4 (*) and 2.9 (**) kb.



Supplementary Figure S10. RT-qPCR analysis of the relative expression of *CXIP4*, *GL1*, *ERIL1*, *PAP1*, and *LOX2* in Col-0, *cxip4-2*, and *cxip4-2 CXIP4*$_{pro}$*:CXIP4:GFP* plants. Error bars indicate the interval delimited by $2^{-\Delta\Delta CT} \pm$ SD, where SD is the standard deviation of the $\Delta\Delta C_T$ values. Three biological replicates were analyzed per genotype, with three technical replicates in each experiment. The *ACT2* gene was used as an internal control for the normalization and relative quantification of gene expression. Asterisks indicate ΔC_T values that are significantly different from the wild type, as determined using Student's *t*-test (**p* < 0.01, ***p* < 0.001, and ****p* < 0.0001).

Supplementary Table S1. Oligonucleotides used in this work

Purpose	Names	Sequences (5′ → 3′)
Genotyping of T-DNA insertional lines	CXIP4-F1[a]	GTCGGGGTTGGAGAAGATTAG
	CXIP4-R1[a]	TATGGTTCAACGGTGGAGAAG
	AT2G23640-F1[b]	ACGTCACCACAACCTTACCAC
	AT2G23640-R1[b]	CCGATGAAGAGACAAGTAATGG
	AT2G23640-F2[c]	TAGATGGATGTTTCGAGTCGG
	AT2G23640-R2[c]	GACGGACCCTAATATCTTCGC
	AT3G50030-F1[d]	CCGTCACTAGGTTTGTTCGAG
	AT3G50030-R1[d]	TCCAACGGACACAAATCAATC
	AT3G50030-F2[e]	GAAGATGACGACCGAGGAAG
	AT3G50030-R2[e]	CCCAGTTTAAATAAGCAGCCC
	o8409[a,f]	ATATTGACCATCATACTCATTGC
	LBb1.3[a-e,g]	ATTTTGCCGATTTCGGAAC
Gateway cloning and verification of constructs	CXIP4proI-F[h]	*GGGGACAAGTTTGTACAAAAAAGCAGGCT*AAAGTTCCACATTGCCCTCTAC
	CXIP4proII-F[h]	*GGGGACAAGTTTGTACAAAAAAGCAGGCT*GAGTTTATTGAATTGTGGGGATG
	CXIP4proI/II-R[h]	*GGGGACCACTTTGTACAAGAAAGCTGGGTG*CTTTTCGAAACCCTAGAAAAAAC
	CXIP4proII:CXIP4:GFP-R[h]	*GGGGACCACTTTGTACAAGAAAGCTGGGTG*CTCTCGCTCTTTCCTGTGATG
	M13-F	GTTGTAAAACGACGGCCAGTG
	GUS-R	CACAAACGGTGATACGTACACT
	GFP-R	CTTGTACAGCTCGTCCATGC
RT-PCR	CXIP4-F2	CCTTCACTCGAATCAGGTTTGT
	CXIP4-R2	ATAACCAATCGCACTCTGCC
	CXIP4-F3	GTGTGAAGAGAAAGAGTAGAAAGG
	CXIP4-R3	CTGCTCTTCTTCTCAGACCG
	SGS1-F	TTCATCCGACTGATGAGGAAC
	SGS1-R	GTTTCCGTTTTTCTCGGTTTC

Supplementary Table S1. Oligonucleotides used in this work

Purpose	Names	Sequences (5′ → 3′)
RT-PCR (continued)	SGS3-F[i]	GCAAAACAAGGCTGAAATTCG
	SGS3-R[i]	CTCTTTCGGCAATCGTCATTGC
	NRPD1B-F	TGATAGAAATCCGTGGGTTTCC
	NRPD1B-R	TGTCTCTTTTTGGGGATGGAAG
	CCR4D-F	AATTGGCTGATGTCAAGCTTG
	CCR4D-R	TGATGATTCGGTAAGAAACCG
	MEE44-F	TCTGGAAGGTAGAAGCTGCTG
	MEE44-R	CCCATGCCTAGTAAACAATGC
	AT2G01100-F	TGCCTCTCCTCTTTTGGATTAG
	AT2G01100-R	GCATGTCTTCGGTGCTTATTG
RT-qPCR	GL1-F[j]	CGCATCGTCAGAAAAACTGGGCTA
	GL2-R[j]	CCGAGGAGCTTGTGGAGACGAATA
	ERIL1-F	GTCGTAGACTTATTCCACAGGTTTG
	ERIL1-R	CTTCAAACTCCTCAACAACTTGCT
	PAP1-F	CAAGTTCCTGTAAGAGCTGGGCTA
	PAP1-R	AGAAGCCTATGAAGGCGAAGAAGAAG
	LOX2-F[k]	GGGATACATAACGGCCCAAGAAG
	LOX2-R[k]	GTAGTCTTCTACCGTAATCCGCTG
	ACT2-F[l]	GCACCCTGTTCTTCTTACCG
	ACT2-R[l]	AACCCTCGTAGATTGGCACA

The oligonucleotides used as forward and reverse primers in PCR amplifications and/or Sanger sequencings are indicated as -F or -R, respectively. [a]The *CXIP4* wild-type allele was amplified with the [a]CXIP4-F1/R1 primers, flanking the T-DNA insertions of *cxip4-1* and *cxip4-2*. The *cxip4-1* and *cxip4-2* insertional alleles were amplified using the CXIP4-F1 + o8409 and CXIP4-F1 + LBb1.3 primer pairs, respectively. [b,c]The AT2G23640 wild-type allele was amplified with the [b]AT2G23640-F1/R1 (flanking the T-DNA insertion of SALK_097293) and [c]AT2G23640-F2/R2 (flanking one of the T-DNA insertions of SALK_044245) primers. The SALK_097293 and SALK_044245 T-DNA insertions were amplified using the AT2G23640-F1 + LBb1.3 and AT2G23640-

F2 + LBb1.3 primer pairs, respectively. [d,e]The AT3G50030 wild-type allele was amplified with the [d]AT3G50030-F1/R1 (flanking the T-DNA insertion of SALK_017259) and [e]AT3G50030-F2/R2 (flanking two of the T-DNA insertions of SALK_044245) primers. The SALK_017259 and SALK_044245 T-DNA insertions were amplified using the AT3G50030-F1 + LBb1.3, and AT3G50030-F2 or R2 + LBb1.3 primer pairs, respectively. [f,g,i,j]The sequences of these primers were taken from [f]https://www.gabi-kat.de/faq/vector-a-primer-info.html, [g]http://signal.salk.edu/tdnaprimers.2.html, [i]Mourrain *et al.* (2000), and [j]Czechowski *et al.* (2004). [h]The *att*B sequences used for the Gateway cloning of the PCR products are shown in italics. [l]The ACT2-F/R primers were also used to amplify the *ACT2* gene as a control of genomic DNA and cDNA integrity for the RT-PCR analyses.

Supplementary Table S2. Summary of protein-coding genes assigned to the PTHR31437 family according to the PANTHER database

Web link: https://www.pantherdb.org/panther/family.do?clsAccession=PTHR31437

Description

Column	Name	Meaning
A	Species	Scientific names of the species with protein-coding genes assigned to the PTHR31437 family
B	Kingdom	Kingdom to which the species belongs
C	Gene count	Number of protein-coding genes assigned to the PTHR31437 family (93 in total)
D	UniProtKB protein ID	Protein identifier (ID) according to the UniProtKB database. Proteins lacking the ZCCHC motif are highlighted in red
E	Protein size (aa)	Protein size in amino acids (aa)
F	ZCCHC motif	Sequence of the ZCCHC motif
G	Position of the $CX_2CX_4HX_4C$ sequence	Position of the conserved $CX_2CX_4HX_4C$ signature in the protein

Species	Kingdom	Gene count	UniProtKB protein ID	Protein size (aa)	ZCCHC motif	Position of the $CX_2CX_4HX_4C$ sequence
Anolis carolinensis	Animalia	1	G1KRB7	158	AGCKKCGYPGHLTFECRN	13-30
Anopheles gambiae	Animalia	1	Q7QJY9	178	AACKKCGFPGHLTYQCRN	15-32
Bos taurus	Animalia	1	F1MYF3	155	AGCKKCGYPGHLTFECRN	13-30
Branchiostoma floridae	Animalia	1	C3ZR60	149	VGCAKCGYTGHLTFECRN	24-41
Caenorhabditis briggsae	Animalia	1	A8XUT4	217	GACKRCGYPGHLYFQCRN	74-91
Caenorhabditis elegans	Animalia	1	O01489	222	GACKRCGYPGHLYFQCRN	75-92
Canis lupus familiaris	Animalia	1	F1PRF2	155	AGCKKCGYPGHLTFECRN	14-30
Ciona intestinalis	Animalia	1	F6XD32	149	IGCKKCGYPGHLTFECRN	16-33
Danio rerio	Animalia	1	Q3B7G7	158	AGCKRCGYPGHLTFECRN	13-30
Drosophila melanogaster	Animalia	1	Q9W3Z5	321	AACKKCGYAGHLTYQCRN	15-32
Equus caballus	Animalia	1	A0A5F5PJV8	155	AGCKKCGYPGHLTFECRN	13-30
Felis catus	Animalia	1	M3X420	155	AGCKKCGYPGHLTFECRN	13-30
Gallus gallus	Animalia	1	R4GL42	156	AGCKKCGYPGHLTFECRN	13-30
Gorilla gorilla gorilla	Animalia	1	A0A2I2ZQ27	155	AGCKKCGYPGHLTFECRN	13-30
Helobdella robusta	Animalia	1	T1FQR6	218	TSCRKCGYAGHLTFQCRN	20-37
Homo sapiens	**Animalia**	**1**	**Q8N9Q2 (SREK1IP1)**	**155**	**AGCKKCGYPGHLTFECRN**	**13-30**
Ixodes scapularis	Animalia	1	B7PCI0	171	PACKKCGYPGHFTYQCRN	22-39
Lepisosteus oculatus	Animalia	1	W5MXC9	176	AGCKKCGYPGHLTFECRN	13-30
Macaca mulatta	Animalia	1	A0A1D5QM33	154	AGCKKCGYPGHLTFECRN	13-30
Monodelphis domestica	Animalia	2	F7G5J1;F6QC96	154;158	AGCKKCGYPGHLTFECRN;AGCKKCGYPGHLTFECRN	13-30;13-30
Mus musculus	Animalia	1	Q4V9W2	153	AGCRKCGYPGHLTFECRN	13-30
Nematostella vectensis	Animalia	2	A7TCJ9;A7RQN8	150;198	VGCKKCGYPGHLTFQCRN;VGCKKCGYPGHLTFQCRN	23-40;25-40
Ornithorhynchus anatinus	Animalia	1	A0A6I8P3B4	158	AGCKKCGYPGHLTFECRN	13-30
Oryzias latipes	Animalia	1	H2LRU4	175	AGCKKCGYAGHLTFECRN	13-30
Pan troglodytes	Animalia	1	A0A2I3TFD6	155	AGCKKCGYPGHLTFECRN	13-30
Pristionchus pacificus	Animalia	1	A0A8R1YHE9	589	GACRRCGFPGHHTFQCYN	467-480
Rattus norvegicus	Animalia	1	Q5RJP9	153	AGCRKCGYPGHLTFECRN	13-30
Strongylocentrotus purpuratus	Animalia	1	A0A7M7RCP4	223	IACKRCGYTGHLTFQCRN	16-31
Sus scrofa	Animalia	1	A0A5G2R5N7	155	AGCKKCGYPGHLTFECRN	13-30
Tribolium castaneum	Animalia	1	D6WIH9	132	AACKKCGYAGHLTYQCRN	18-35
Xenopus tropicalis	Animalia	1	Q28EE8	155	AGCKKCGYPGHLTFECRN	11-30
Amborella trichopoda	Plantae	2	U5D559;W1P711	301;275	GSCKKCGRVGHLTFQCRN	85-102
Arabidopsis thaliana	**Plantae**	**1**	**Q84Y18 (CXIP4)**	**332**	**GSCKKCGRVGHLTFQCRN**	**81-98**
Brachypodium distachyon	Plantae	1	I1IIQ0	372	GACKKCGRVGHLTFQCRN	104-121
Brassica napus	Plantae	2	A0A078HM46;A0A078IDA7	322;328	GSCKKCGRVGHLTYQCRN;GSCKKCGRVGHLTYQCRN	81-98;81-98
Brassica rapa subsp. pekinensis	Plantae	2	M4FIJ9;M4F3L1	319;295	GSCKKCGRVGHLTYQCRN;GSCKKCGRVGHLTYQCRN	72-89;74-91
Capsicum annuum	Plantae	1	A0A2G3ANM6	330	GACKRCGRVGHLTFQCRN	84-101
Chlamydomonas reinhardtii	Plantae	1	A0A2K3DE08	397	GACKICGGLGHLTKKCKN	100-118
Cucumis sativus	Plantae	1	A0A0A0KHA7	310	GACKRCGRVGHLSFQCRN	85-102
Erythranthe guttata	Plantae	1	A0A022PVZ2	319	GCCKKCGRVGHLTFQCRN	88-105
Eucalyptus grandis	Plantae	4	A0A059C6D8;A0A059C4V0;A0A059C4Y3;A0A059AFZ9	285;338;235;331	GACKKCGRVGHLKFQCRN;GSCKKCGRVGHLTFQCRN;GSCKKCGRVGHLTFQCRN;GACKRCGRVGHLTFQCRN	52-69;85-102;85-102;85-102
Glycine max	Plantae	2	I1LF76;I1NF68	317; 314	GACKKCGRVGHLKFQCKN;GACKKCGRVGHLKFQCKN	84-101;84-101
Gossypium hirsutum	Plantae	1	A0A1U8L280	328	GACKKCGRVGHLTFQCKN	85-102
Helianthus annuus	Plantae	1	A0A251VLY8	296	GSCKKCGRVGHLTFQCRN	85-102
Hordeum vulgare subsp. vulgare	Plantae	1	A0A287R5P6	365	GACKKCGRVGHLTFQCRN	104-122
Juglans regia	Plantae	2	A0A2I4EV08;A0A2I4FWM3	321;318	GACKKCGRVGHLTFQCRN;GACKKCGRVGHLTFQCRN	85-102;85-102
Lactuca sativa	Plantae	1	A0A2J6L4U4	330	GSCKKCGRVGHLTFQCRN	86-103
Manihot esculenta	Plantae	2	A0A251LHL8;A0A2C9WLQ8	330;430	AGCKRCGRVGHLAYQCRN;GACKRCGRVGHLSYQCRN	85-102;172-189
Marchantia polymorpha	Plantae	1	A0A2R6XPG7	366	GACKKCGRVGHLTFQCRN	93-108
Musa acuminata subsp. malaccensis	Plantae	2	A0A804K238;A0A804KT39	358;333	GACNKCGRVGHLTFQCRN;GSCRKCGRVGHLTFQCRN	88-103;86-101
Nelumbo nucifera	Plantae	2	A0A1U8BEW9;A0A1U8AC80	316;316	GSCKKCGRVGHLTFQCRN;GACKKCGRVGHLTIQCRN	85-102;83-100
Nicotiana tabacum	Plantae	3	A0A1S4CR75;A0A1S3ZTG7;A0A1S4AM92	322;3193;22	GACKRCGRVGHLTFQCRN;GACKRCGRVGHLTFQCRN;GACKRCGRVGHLTFQCRN	82-99;82-99;82-99
Oryza sativa	Plantae	2	B9FK38;Q69K05	297;377	GGCRKCGRVGHLTFQCRN;GACRKCGRVGHLTFQCRN	90-107;106-123

Physcomitrella patens	Plantae	1	A0A2K1K6W5	469	GSCQKCGRVGHLTFQCRN	90-107
Populus trichocarpa	Plantae	1	A0A2K2BXT0	320	GACKRCGRVGHLAFQCRN	85-102
Prunus persica	Plantae	1	M5WBW2	323	GACKRCGRVGHLNFQCRN	85-102
Ricinus communis	Plantae	1	B9T3N1	321	GACKRCGRVGHLAFQCRN	88-105
Selaginella moellendorffii	Plantae	2	D8RE01;D8SU28	265;272	GACRRCKRVGHLPFQCRN;GACRRCKRVGHLPFQCRN	85-102;85-102
Setaria italica	Plantae	2	K3ZUL6;K3YI63	356;378	GACKKCGRVGHLTFQCRN;GSCKRCGRVGHLTFQCRN	107-124;110-127
Solanum lycopersicum	Plantae	1	A0A3Q7HGR1	293	GACKRCGRVGHLTFQCRN	84-101
Solanum tuberosum	Plantae	1	M1BL10	328	GACKRCGRVGHLTFQCRN	84-101
Sorghum bicolor	Plantae	2	C5X6Q4;C5YFL9	377;368	GACKKCGRVGHLTFQCRN;GACKKCGRVGHLTFQCRN	107-124;107-124
Spinacia oleracea	Plantae	1	A0A0K9RYF6	318	GACKKCGRVGHLTFQCRN	82-99
Theobroma cacao	Plantae	2	A0A061GW00;A0A061GX82	323;241	GACKRCGRVGHLTFQCKN;GSCKKCGRVGHLTFQCKN	85-102 85-102
Triticum aestivum	Plantae	4	A0A3B6KGM7;A0A3B6MR41;A0A3B6LKL1;A0A3B6B5X8	369;370;369;310	GACKKCGRVGHLTFQCRN;GACKKCGRVGHLTFQCRN;GACKKCGRVGHLTFQCRN;GACKKCGRVGHLTFQCRN	106-121;105-122;106-121;93-110
Vitis vinifera	Plantae	1	F6GV10	328	GACKKCGRVGHLTFQCRN	85-102
Zea mays	Plantae	2	A0A096QKW6;A0A1D6ELP7	371;397	GACKKCGRVGHLTFQCRN;GACKKCGRVGHLTFQCRN	107-124;107-124
Zostera marina	Plantae	1	A0A0K9NRK4	337	GACKKCGRVGHLTYQCRN	92-109
Dictyostelium purpureum	Protozoa	1	F1A2S1	318	SLCKKCGGNGHLTYQCIA	68-85
Plasmodium falciparum	Protozoa	1	A0A5K1K8H8	262	GACTVCNHIGHLPYQCRN	77-94
Phytophthora ramorum	Chromista	1	H3GFH1	372		
Thalassiosira pseudonana	Chromista	1	B8BR89	262		

Supplementary Table S3. Summary of high-confidence physical protein–protein interactions for human SREK1IP1, identified from high-throughput experiments according to the HitPredict database

Web link: http://www.hitpredict.org/htp_int.php?Value=661

Description

Column	Name	Meaning
A	Interaction ID	Interaction identifier (ID)
B	Interactor symbol	Symbols of the proteins physically interacting with human SREK1IP1. Those mentioned in the manuscript are highlighted in bold
C	Interactor full name	Complete names of the proteins physically interacting with human SREK1IP1. Those mentioned in the manuscript are highlighted in bold
D	Assays	Methods used for the identification of physical protein-protein interactions. Y2H: yeast two-hybrid assay; AP-MS: affinity purification-mass spectrometry
E	Experiments	Number of experiments supporting the interactions
F	Homologs	Number of homologous interactions
G	Method score	This score is derived from the available experimental information for interactions and is calculated as the mean of three scores: the publication score (based on the number of publications or experiments supporting the interaction), the method score (specified by the HUPO PSI-MI consortium for various interaction identificaction methods, including biophysical, protein complementation assay, post transcriptional inference, biochemical, imaging technique, and their subtypes), and the type score (specified by the HUPO PSI-MI consortium for each interaction type and subtype, encompassing association, physical association, and direct interaction). Method scores equal to or greater than 0.485 are considered indicative of high confidence.
H	Annotation score	This score is calculated as a likelihood ratio using naive Bayesian networks, relaying on structurally known interacting Pfam domains present in the interactors, shared Gene Ontology (GO) terms between the interacting proteins, and the number of homologous interactions. Likelihood ratios greater than 1 are scaled to give annotation scores between 0.5 and 1, indicating high-confidence interactions. The annotation score value increases with the evidence supporting the interaction
I	Interaction score	This score denotes the reliability of the interaction and is calculated as the geometric mean of the annotation-based score (column G) and the method-based score (column H). It considers both the experimental support for the interaction and the genomic features of the interacting proteins, having demostrated a better performance compared to either of the two previous scores
J	References	Publications supporting the interactions

Interaction ID	Interactor symbol	Interactor full name	Assays	Experiments	Homologs	Method score	Annotation score	Interaction score	References
126237	SDCBP1/ST-1	Syndecan binding protein 1/Syntenin-1	Y2H	3	1	0,760	0,500	0.616	Rolland *et al.* (2014); Luck *et al.* (2020)
1142717	SDCBP2/ST-2	Syndecan binding protein 2/Syntenin-2	Y2H	3	1	0,760	0,500	0.616	Rolland *et al.* (2014); Luck *et al.* (2020)
1142706	**RP9/PAP-1**	**Retinitis pigmentosa 9 protein/Pim-1 kinase associated protein**	Y2H	2	1	0,610	0,600	0.606	Luck *et al.* (2020)
1011852	RN151	RING finger protein 151	Y2H	2	0	0,610	0,550	0.580	Luck *et al.* (2020)
1142720	THAP1	THAP domain-containing protein 1	Y2H	2	0	0,610	0,550	0.580	Luck *et al.* (2020)
160181	**IPOA6/KPNA5**	**Importin subunit alpha 6/Karyopherin subunit alpha 5**	AP-MS	2	3	0,620	0,500	0.558	Huttlin *et al.* (2017), (2021)
225593	**IPOA7/KPNA6**	**Importin subunit alpha 7/Karyopherin subunit alpha 6**	AP-MS	2	3	0,620	0,500	0.558	Huttlin *et al.* (2017), (2021)
544291	CSK22	Casein kinase II subunit alpha'	AP-MS	2	1	0,620	0,500	0.558	Huttlin *et al.* (2017), (2021)
762561	**NUP153**	**Nucleoporin 153**	AP-MS	2	2	0,620	0,500	0.558	Huttlin *et al.* (2017), (2021)
854758	CSK21	Casein kinase II subunit alpha	AP-MS	2	1	0,620	0,500	0.558	Huttlin *et al.* (2017), (2021)
864861	MAP1A	Microtubule-associated protein 1A	AP-MS	2	1	0,620	0,500	0.558	Huttlin *et al.* (2017), (2021)
901895	EF1A2	Elongation factor 1-alpha 2	AP-MS	2	1	0,620	0,500	0.558	Huttlin *et al.* (2017), (2021)
248789	**PRPF40A**	**Pre-mRNA-processing factor 40 homolog A**	Y2H	2	1	0,610	0,500	0.553	Luck *et al.* (2020)
1038423	NKAPL	NKAP-like protein	Y2H	2	1	0,610	0,500	0.553	Luck *et al.* (2020)
967959	**RBM39**	**RNA-binding protein 39**	Y2H	1	0	0,550	0,550	0.549	Rolland *et al.* (2014)
663	A0A024RA76	HCG1744368, isoform CRA_a	Y2H	1	1	0,440	0,600	0.515	Luck *et al.* (2020)
780661	**IPOA5/KPNA1**	**Importin subunit alpha 5/Karyopherin subunit alpha 1**	AP-MS	1	3	0,480	0,500	0.489	Huttlin *et al.* (2017)
1138675	**NKAP**	**NF-kappa-B-activating protein**	AP-MS	1	1	0,480	0,500	0.489	Huttlin *et al.* (2017)
1142725	**SRRM2**	**Serine/arginine repetitive matrix 2**	AP-MS	1	1	0,480	0,500	0.489	Huttlin *et al.* (2017)
1142728	U3KQK0	Histone H2B	Y2H	1	1	0,440	0,500	0.470	Lafleur (1975)
753270	NR2C2	Nuclear receptor subfamily 2 group C member 2	AP-MS	1	0	0,380	0,550	0.458	Huttlin *et al.* (2017)
1142707	GATD1	GATA zinc finger domain-containing protein 1	AP-MS	1	0	0,380	0,550	0.458	Varier *et al.* (2016)
1142714	H2B1N	Histone H2B type 1-N	Y2H	1	1	0,390	0,500	0.439	Luck *et al.* (2020)
327700	MYC	Myc proto-oncogene protein	AP-MS	1	1	0,380	0,500	0.437	Wang *et al.* (2022)
357828	MYCN	N-myc proto-oncogene protein	AP-MS	1	1	0,380	0,500	0.437	Wang *et al.* (2022)
780417	**IPOA1/KPNA2**	**Importin subunit alpha 1/Karyopherin subunit alpha 2**	AP-MS	1	3	0,380	0,500	0.437	Huttlin *et al.* (2021)
1142723	ZRAN1	Ubiquitin thioesterase ZRANB1	AP-MS	1	1	0,380	0,500	0.437	Chen and Zhang (2022)
1142713	STAC3	SH3 and cysteine-rich domain-containing protein 3	Y2H	3	0	0,760	0,160	0.352	Rolland *et al.* (2014); Luck *et al.* (2020); Lafleur (1975)
1142722	PRR13	Proline-rich protein 13	Y2H	3	0	0,700	0,160	0.339	Luck *et al.* (2020); Slaughter (1976); Lafleur (1975)
1142715	DHB14	17-beta-hydroxysteroid dehydrogenase 14	Y2H	2	0	0,660	0,160	0.327	Rolland *et al.* (2014); Fragoza *et al.* (2019)
239142	CAPON	Carboxyl-terminal PDZ ligand of neuronal nitric oxide synthase protein	AP-MS	2	0	0,620	0,160	0.318	Huttlin *et al.* (2017), (2021)
343791	FRIL	Ferritin light chain	AP-MS	2	0	0,620	0,160	0.318	Huttlin *et al.* (2017), (2021)
1142726	TCAF1	TRPM8 channel-associated factor 1	AP-MS	2	0	0,620	0,160	0.318	Huttlin *et al.* (2017), (2021)
172500	FGF10	Fibroblast growth factor 10	Y2H	2	0	0,610	0,160	0.316	Luck *et al.* (2020); Lafleur (1975)
246100	H2AY	Core histone macro-H2A.1	Y2H	2	0	0,610	0,160	0.316	Luck *et al.* (2020); Lafleur (1975)
318329	HPRT	Hypoxanthine-guanine phosphoribosyltransferase	Y2H	2	0	0,610	0,160	0.316	Luck *et al.* (2020); Lafleur (1975)
644760	RL9	Large ribosomal subunit protein uL6	Y2H	2	0	0,610	0,160	0.316	Luck *et al.* (2020); Lafleur (1975)
1035803	CT451	Cancer/testis antigen family 45 member A1	Y2H	2	0	0,610	0,160	0.316	Luck *et al.* (2020); Lafleur (1975)
1079953	NKAP1	Uncharacterized protein NKAPD1	Y2H	2	0	0,610	0,160	0.316	Luck *et al.* (2020); Lafleur (1975)
1127695	ELOA2	Elongin-A2	Y2H	2	0	0,610	0,160	0.316	Luck *et al.* (2020); Lafleur (1975)
1142708	TF2H4	General transcription factor IIH subunit 4	Y2H	2	0	0,610	0,160	0.316	Luck *et al.* (2020); Lafleur (1975)
1142710	DDT4L	DNA damage-inducible transcript 4-like protein	Y2H	2	0	0,610	0,160	0.316	Luck *et al.* (2020); Lafleur (1975)
1142712	CD045	Uncharacterized protein C4orf45	Y2H	2	0	0,610	0,160	0.316	Luck *et al.* (2020); Lafleur (1975)
1142719	PLCC	1-acyl-sn-glycerol-3-phosphate acyltransferase gamma	Y2H	2	0	0,610	0,160	0.316	Luck *et al.* (2020); Lafleur (1975)

1142721	LYAR	Cell growth-regulating nucleolar protein	Y2H	2	0	0,610	0,160	0.316	Luck *et al.* (2020); Lafleur (1975)
1142724	CCNL1	Cyclin-L1	Y2H	2	0	0,610	0,160	0.316	Luck *et al.* (2020); Lafleur (1975)
511686	NDKA	Nucleoside diphosphate kinase A	Y2H	1	0	0,550	0,160	0.299	Rolland *et al.* (2014)
1142709	COAC	Phosphopantothenoylcysteine decarboxylase	Y2H	1	0	0,550	0,160	0.299	Rolland *et al.* (2014)
257228	OFD1	Centriole and centriolar satellite protein OFD1	AP-MS	2	0	0,540	0,160	0.298	Gupta *et al.* (2015); Brygier *et al.* (1975)
981219	PCM1	Pericentriolar material 1 protein	AP-MS	2	0	0,540	0,160	0.298	Gupta *et al.* (2015); Brygier *et al.* (1975)
1054270	CP135	Centrosomal protein of 135 kDa	AP-MS	2	0	0,540	0,160	0.298	Gupta *et al.* (2015); Brygier *et al.* (1975)

SUPPLEMENTARY REFERENCES

Czechowski T, Bari RP, Stitt M, Scheible WR, Udvardi MK. Real-time RT-PCR profiling of over 1400 *Arabidopsis* transcription factors: unprecedented sensitivity reveals novel root- and shoot-specific genes. Plant J. 2004:**38**(2): 366-379. https://doi.org/10.1111/j.1365-313X.2004.02051.x

Li M, Ruwe H, Melzer M, Junker A, Hensel G, Tschiersch H, Schwenkert S, Chamas S, Schmitz-Linneweber C, Börner T, Stein N. The Arabidopsis AAC proteins CIL and CIA2 are sub-functionalized paralogs involved in chloroplast development. Front. Plant Sci. 2021:**12**:681375. https://doi.org/10.3389/fpls.2021.681375

Mourrain P, Béclin C, Elmayan T, Feuerbach F, Godon C, Morel JB, Jouette D, Lacombe AM, Nikic S, Picault N, Rémoué K, Sanial M, Vo TA, Vaucheret H. *Arabidopsis SGS2* and *SGS3* genes are required for posttranscriptional gene silencing and natural virus resistance. Cell 2000:**101**(5): 533-542. https://doi.org/10.1016/s0092-8674(00)80863-6

X.- AGRADECIMIENTOS

XI.- AGRADECIMIENTOS

La realización de esta Tesis ha sido posible gracias a la financiación del trabajo que se realiza en el laboratorio de María Rosa Ponce por la Generalitat Valenciana (GRISOLIA/2016/134), el Ministerio de Economía y Competitividad (BIO2017-89728-R) y el Ministerio de Ciencia e Innovación (PID2020-117125RB-I00).

Gracias en primer lugar a mi directora de Tesis, María Rosa Ponce, por permitirme realizar esta Tesis en su laboratorio y quien ha liderado mi formación, gracias por transmitirme su pasión por la ciencia y la investigación siendo un ejemplo a seguir.

A José Luis Micol, por su interés en mi trabajo con sus preguntas y discusiones, y por aportar siempre un punto de vista crítico y constructivo.

A mi esposa Paulina por acompañarme en este camino de la vida y apoyándome en cada momento junto a mis hijos Lía y Vito. Los amo a los tres (pero a ti desde mucho antes).

A mis padres quienes me enseñaron a ser perseverante a perseguir los sueños, que con esfuerzo y trabajo puedes conseguir cosas y sobre todo a depositar nuestra fe en Dios.

A mis hermanos Dan, Jair y Debbie quienes son un ejemplo a seguir tanto en perseverancia para salir adelante y la pasión de querer ayudar siempre al resto.

A mis abuelos quienes con sus sabios consejos he podido estar donde estoy un beso al cielo abuelita Susana y tata Jhonny.

A mis amigos Riad y Karima con quienes he compartidos momentos de alegría que atesoro en mi corazón.

A Edu y José Manuel Serrano quienes nos extendieron sus brazos y nos recibieron como uno de los suyos.

A cada uno de mis compañeros del laboratorio Carla, Lucía, Adrián, Alejandro, Sergio y Samuel que pude compartir risas y alegrías mientras estaba en curso del programa de doctorado que han sido como mi segunda familia.

A Raquel, Rosa y Sara que me han apoyado siempre y sobre todo en las etapas finales de esta Tesis, quienes me tuvieron mucha paciencia y siempre tenían disposición a ayudarme y responder mis inquietudes, muchas gracias de verdad.

A Juan Castelló, María José, Diana y todos los que me apoyaron en la parte técnica.

¡Gracias totales!